November 2005

Risk Assessment of Military Populations to Predict Health Care Cost and Utilization

Final Report

Prepared for

Thomas Williams, Ph.D.
Center for Health Care Management Studies
TRICARE Management Activity; HPA&E
5111 Leesburg Pike, Suite 510
Falls Church, VA 22041

Prepared by

Arlene S. Ash, Ph.D., Boston University
Nancy McCall, Sc.D., RTI International

Participating Investigators
Jenn Fonda, M.A., Boston University
Amresh Hanchate, Boston University
Jeanne Speckman, M.Sc., Boston University

RTI International
1615 M Street, NW
Washington, DC 20036
and
Boston University School of Medicine
715 Albany Street
Boston, MA 02118

RTI Project Number 08490.006

RTI Project Number
08490.006

Risk Assessment of Military Populations to Predict Health Care Cost and Utilization

Final Report

November 2005

Prepared for
Thomas Williams, Ph.D.
Center for Health Care Management Studies
TRICARE Management Activity; HPA&E
5111 Leesburg Pike, Suite 510
Falls Church, VA 22041

Prepared by
Arlene S. Ash,[1] Ph.D., Boston University
Nancy McCall, Sc.D., RTI International
Participating Investigators
Jenn Fonda, M.A., Boston University
Amresh Hanchate, Boston University
Jeanne Speckman, M.Sc., Boston University

RTI International[2]
1615 M Street, NW
Washington, DC 20036
and
Boston University School of Medicine
715 Albany Street
Boston, MA 02118

[1] Dr. Ash is a developer of the Diagnostic Cost Group (DCG) models and a founder of the company (DxCG, Inc.) that maintains and licenses DCG models, one of the models evaluated in this project. Although the company has been sold, Dr. Ash continues to work one day a week for the company, as a senior scientist. Her primary affiliation, and the one through which she is involved in this project, is as a Research Professor at Boston University School of Medicine.

[2] RTI International is a trade name of Research Triangle Institute.

Table of Contents

Chapter **Page**

Appendixes

List of Figures

Number **Page**

List of Tables

Number **Page**

Executive Summary

The Military Health System is examining new approaches to financing the care of covered beneficiaries. Since population health care costs depend on illness burden (case-mix), efficient financing should recognize variations in the illness burden of populations seen by different providers. Several mature risk adjustment systems that extract illness burden profiles from computerized encounter records are now used by Medicare, Medicaid, the Veteran's Health Administration, commercial insurers in the United States, and international stakeholders to understand and manage health care delivery systems. These systems are designed for "population-based" management; "units" are person-years of medical care, and the key outcome is the cost, for each person, of a year of care. The expected cost for a provider for a year is the sum of individual-level person-year estimates for the population served.

We conducted side-by-side testing of a simple age-sex method similar to what is currently used in TRICARE Prime and four claims-based risk adjustment models: Adjusted Clinical Groups (ACGs), Chronic Disease and Disability Payment System (CDPS), Clinical Risk Groups (CRGs), and Diagnostic Cost Groups (DCGs). All four risk adjustment models use a person's age, sex, and the morbidities recorded (in ICD-9-CM codes) during a year to predict total costs (inpatient + outpatient + pharmacy) for the next year. In addition, the CRG model used limited information on dates of service, place of service, and procedures. We applied each model to the TRICARE Prime population, measuring its overall ability to predict future costs and the concordance between model-measured needs and actual health care spending in policy-relevant subgroups of TRICARE Prime enrollees. We also examined the data quality and credibility for use in monitoring and managing cost of care.

We used administrative data on enrollment and claims for TRICARE Prime enrollees in fiscal years (FY) 2001 and 2002. Our study sample included (all 2.3 million) TRICARE Prime enrollees under the age of 65 years who were continuously enrolled during for the 24 months of FY 2001 through FY 2002 and residing in the continental United States, Hawaii, or Alaska. We simulated the real-world experience of building a model that is applied to new data by fitting the models to an estimation sample of 1.8 million and testing how closely their predictions matched up with actual costs in the remaining 500,000, the validation sample.

Population Costs and Data Quality

The validation sample contained data for two 2 years on 500,000 people. Mean costs in the second year (FY 2002) were highly variable, with a mean (SD) of $1,796 ($5,705); 15 percent had no costs, and 0.6 percent had costs over $25,000. The volume and distribution of records were reassuringly similar to what is seen in good quality commercial data.

Model Comparisons

Each model was compared on its ability to predict all costs, as well as costs "top-coded" at $25,000 and $50,000. For example, a person whose total claims for FY 2002 were $40,000 would have

that figure used (unaltered) in analyses to predict the "all cost" or "top-coded-at-$50,000" outcome, but would have a cost of $25,000 substituted in the "top-coded-at-$25,000" analysis. Top-coded analyses are well suited to prospective payment systems with individual stop loss. We examined the performance of all four models on a wide range of commonly used statistical measures of model accuracy when predicting next year's cost for a population from this year's costs. These included two global measures of goodness-of-fit between model-predicted and actual costs for all people in the validation data: R-squared (R^2) and Cumming's Prediction Measure (CPM). R^2 and CPM are generally between 0 and 1, with 1 indicating a perfect match between predicted and actual costs. In Table ES-1, each measure was multiplied by 100 to provide a direct estimate of the percent of variation in predicted costs that was explained at the group level by each model. Table ES-1 shows that all four models performed similarly, and all vastly outperformed a simple age-sex model; the DCG model performed a little better than the other models on each global measure.

Table ES-1. Overall Measures of Predictive Accuracy

	All cost					Cost top-coded at $25,000				
	Age-sex	ACG	CDPS	CRG	DCG	Age-sex	ACG	CDPS	CRG	DCG
R^2×100	3.42	14.75	14.71	**15.17**	**15.97**	8.26	25.38	23.71	23.41	**26.78**
Cumming's Prediction Measure (CPM×100)	6.93	18.10	16.21	17.32	**19.21**	7.43	19.66	17.65	18.31	**20.92**

Better models have a higher R^2 and CPM. In each row-block, each model that performs best, or comes within 1 percent of best, is **bolded**.

A third measure of predictive accuracy, the predictive ratio (PR), was estimated for key subgroups of the TRICARE Prime population as defined by various demographic, geographic, and service-related characteristics; medical costs; and the presence of medical conditions. A PR is computed by dividing the model-predicted costs for everybody in each of the subgroups by their actual costs; a model whose predictions most closely match the group's actual costs has a PR closest to 1.0. For example, a PR of 0.9 means that the model-predicted cost for that subgroup was 90 percent as large as its actual cost. In terms of a prospective payment system, a PR of 0.9 indicates that the provider would be paid 10 percent less than actual costs.

Across a wide range of demographic, geographic, and service-related subgroups, all four models predicted differences in average costs across such subgroups well; the CRG model's predictions were closest to actual. If we have adequately adjusted for differences in medical risk, then PRs that differ from 1.0 across subgroups defined by service system factors (such as, patients who use different facilities) can provide important information for system monitoring. For these subgroups, PRs ranged from a low of 0.63 to a high of 1.21. We also examined how these risk models performed, separately, for people with prior low, intermediate, or high costs. The risk models somewhat overpredicted costs for the lowest cost subgroups and underpredicted costs for the highest cost ones, but much less than the age-sex model. At

the ends of the cost spectrum, the DCG model was closest to perfect accuracy; whereas the CDPS model was closest in the middle of the spectrum.

A critical characteristic of a risk adjustment model is its ability to accurately predict the costs for groups of people with expensive and/or highly prevalent diseases. We examined PRs for nine such conditions in TRICARE Prime. All four models came far closer to predicting true costs than the age-sex model, which always underpredicted costs for groups of sick people. The age-sex model predicted costs for people with chronic renal failure that were only 13 percent of their actual costs, and its predicted costs for people with osteoarthritis were only 58 percent of their actual costs. In stark contrast, the four risk adjustment models had PRs ranging from 0.75 (chronic renal failure) to 1.11 (congestive heart failure). The DCG and ACG predictions were most accurate here.

Conclusions

Predictive power for all four "off-the shelf" risk models is far better than what can be achieved with the age-sex models, and compares well to their performance in other populations. Thus, any of the risk adjustment models could be used by the Military Health System to make more meaningful and useful health comparisons across subpopulations of its enrollees and to predict health care resource utilization and cost in various subpopulations. Further, the risk adjustment models can properly "price" at the person level, based on disease burden, enabling the Military Health System to move away from payment based on utilization. DCGs have a slight edge in predictive accuracy, but any of these models will facilitate medical and financial management of the Military Health System.

This study focused on the "financial" applications of risk adjustment. However, there are many other important applications of risk adjustment within health care organizations and for payers such as TRICARE. The methods evaluated in this study may also be used to (1) compare morbidity, disease burden, and/or prevalence rates across subpopulations; (2) profile or monitor performance at various organizational units, i.e., military treatment facility, TRICARE region, or service; and (3) help identify high-risk cases for disease management by applying "predictive modeling" at the individual level.

Additional considerations are important when selecting a risk adjustment method. Drawing on the framework recommended by the Society of Actuaries, we note ease of use of the software; availability of standard reports; licensing fees; access to quality data that support the modeling; the suitability of the underlying logic or perspective of the model for a specific application; whether the model provides both useful clinical and financial information; whether the model will be used mostly for payment to providers/plans or for underwriting/rating/case management; where the model is currently in use in the market or organization; and the susceptibility of the model to gaming or upcoding. A qualitative evaluation to systematically examine the relative importance of these other considerations may be a reasonable extension of this study.

1

Introduction

The purpose of this project is to conduct side-by-side testing of the ability of various risk adjustment methods to

1. improve health comparisons of Military Health System (MHS) beneficiaries enrolled with military versus civilian primary care managers (PCMs) under TRICARE Prime and

2. predict health resource utilization and cost among TRICARE Prime enrollees.

We focus on how well a risk adjustment model utilizes demographic and diagnosis information in one year to predict health care costs in the following year. The key step in obtaining model predictions is estimating a statistical relationship between fiscal year (FY) 2002 costs and the model-specific determining factors (demographic and diagnosis-related condition categories in FY 2001) using fitting sample data. For each risk adjustment model, this step gives a set of risk weights, one for each demographic and diagnosis-related condition category. Each risk weight denotes the predicted dollar amount associated with a specific demographic or diagnosis category. The predicted cost (for FY 2002) from each model is obtained by aggregating all the risk weights from that model (using the validation sample data). A series of statistical comparisons are used to assess the predictive performance of each model.

It is important to note that the scope off work for this project focused strictly upon a quantitative analysis to identify the tool that has the greatest predictive power for making risk-adjusted health comparisons across subpopulations. Thus, this study tended to focus more specifically upon the "financial" applications of risk adjustment. However, it is important to note that there are a myriad of other applications of risk adjustment within health care organizations and payers such as TRICARE. The risk adjustment methods evaluated in this study may also be used to compare morbidity, disease burden, and/or prevalence rates across subpopulations; profiling or performance monitoring at alternative organizational units, i.e., military treatment facility, TRICARE region, or service; case and disease management applications; and high risk case identification through the application of "predictive modeling" at the individual level.

Further, this study was limited to a purely quantitative analysis of the alternative risk adjustment methods. Additional key considerations may be warranted when selecting a risk adjustment method for purposes of implementing a risk adjustment prospective payment methodology or when using risk scores for other population health management purposes.

2

Background

2.1 The TRICARE Program

TRICARE is the Department of Defense's (DoD) regionally administered managed care program for delivering health care services to active duty members of the Armed Services (Stoloff, Lurie, Goldberg, & Almendarez, 2001). Spouses and children of active duty members of the Uniformed Services, retirees and their family members, some former spouses, and spouses and children of deceased active duty members or deceased retirees are also entitled to health care coverage through TRICARE.

The goals of TRICARE are to (1) improve access among military health system beneficiaries, (2) provide faster and more convenient access to civilian health care services, (3) create a more efficient way to receive health care, (4) provide choices for health care, and (5) control health care costs (U.S. Department of Defense [DoD], 2005). To achieve these goals, Congress mandated that TRICARE be modeled on HMO plans offered in the private sector and other similar government health insurance programs. Congress also stipulated that the costs incurred by the DoD should be no greater than the costs that would otherwise have been incurred under the traditional benefit of direct care and CHAMPUS (Stoloff et al, 2001). In FY 2001, TRICARE spent roughly $3.5 billion on direct health care services (out of a total budget of $6.2 billion) for over five million eligible beneficiaries (TRICARE Management Activity [TMA], 2002). Over the past few years, TRICARE's health care expenditures and budget have grown by almost one-quarter, while the number of military health system beneficiaries has remained fairly constant.

In addition to being one of the largest health insurance programs in the country, TRICARE serves an extremely diverse population with a wide range of health care needs and resources. TRICARE beneficiaries vary in many respects, including eligibility status, geographic region, demographic characteristics (age, gender, race/ethnicity, and education), proximity to health care services, access to private supplemental health insurance coverage, health status, and medical conditions—all with important implications for risk adjustment. Of the 3.9 million enrolled beneficiaries as of July 2000, almost 25 percent were in active duty, nearly 50 percent were family members of active duty personnel, and 25 percent were retirees or family members of retirees. Based on a 2000 DoD survey of 8 of TRICARE's 11 health service regions, 25 percent of beneficiaries had supplemental private insurance.[1] Over 25 percent had completed 4 years of postsecondary education and, 66 percent lived within a health service catchment area.[2] Included among the 10 diagnosis-related categories with the highest share of total expenditures were such disparate diagnoses as mental disorders, deliveries, tracheostomies, rehabilitation, and heart bypass surgery.

[1] These and other statistics on enrollee characteristics are taken from the *2002 Chart Book of Statistics* (TMA, 2002).

[2] A catchment area is an approximately 40-mile radius around a military hospital, allowing for natural geographic boundaries and transportation accessibility.

Enrollee characteristics vary widely across the health service regions as well. The Mid-Atlantic and Pacific regions have the highest proportion of active duty personnel and their family members (about 89 percent); the Golden Gate region has the lowest (60 percent). Moreover, according to the 2000 survey of the eight selected health service regions, the proportion of African American active duty personnel varied from 21 percent in Region 3 to 7 percent in Region 10. The share of Hispanic personnel varied from 14 percent in Region 9 to 5 percent in Region 11. The proportion of female active duty personnel was highest in Region 6 (21 percent) and lowest in Region 9 (10 percent). Furthermore, nearly three-quarters of all uniformed personnel in Region 9 completed at least 4 years of college education, compared with only two-thirds in Region 4. The proportion with private supplemental coverage ranged from 13 percent in Region 10 to 5 percent in Regions 6 and 12. Finally, all active duty personnel in Region 12 lived within the service catchment area, compared with only 86 percent in Region 4. Differences in the military status and demographic, access, and health status profiles of TRICARE beneficiaries between and within the program's administrative health service regions carry important implications for efficient plan administration and appropriate payment for health care services.

To meet the needs of its large and diverse consumer-beneficiary population, TRICARE has developed three principal health benefit options for military health system beneficiaries: TRICARE Prime, TRICARE Standard, and TRICARE Extra.[3] These options vary in eligibility, cost sharing, choice of provider or treatment facility, covered services, primary care case management, paperwork and claims submission, and geographic availability. TRICARE Prime is comparable to a civilian health maintenance organization (HMO), and is the focus of this study. All active-duty personnel are automatically enrolled in TRICARE Prime at their nearest military treatment facility. Enrollment is open at all times with no restrictions on enrollment based on pre-existing conditions. The enrollee's chosen or assigned Primary Care Manager (PCM) coordinates all health care services under TRICARE Prime, serves as the enrollee's primary care provider and approves all referrals. The PCM, together with the contracted Health Care Finder, assists beneficiaries in locating specialty care in the civilian community when the needs of the enrollee cannot be met by the military treatment facility. Minimal waiting times for appointments, which vary by type of service, are enforced. TRICARE Prime is free to active duty personnel, but retirees and family members pay an annual fee plus nominal co-payments. TRICARE retirees pay only a small enrollment fee to maintain their health benefits. Non-active-duty TRICARE Prime enrollees can seek care from out-of-network providers through a point-of-service option, but are subject to an even higher cost share than under TRICARE Standard. TRICARE Prime enrollees are also covered for an expanded set of preventive services, such as eye and ear examinations, immunizations, mammography, Pap smears, prostate examinations, and other cancer-prevention and early-detection examinations.

In contrast, TRICARE Standard (formerly DoD's indemnity plan called CHAMPUS) allows retirees and family members to choose any civilian physician they want with the government payment varying by the beneficiary's military status and the type of service (Army, Navy, etc.). However, beneficiaries residing in a catchment area must first seek care from a military hospital for inpatient care

[3] In recent years, the DoD has introduced several new programs and demonstrations, adding to the array of available plans and options. These include TRICARE Senior Demonstration, TRICARE Senior Supplement Demonstration, TRICARE Dental Program, National Mail Order Pharmacy Program, Federal Employees Health Benefits Program Demonstration, TRICARE Prime Remote, and Pharmacy Redesign Pilot Program.

and for selected outpatient procedures. In contrast, TRICARE Extra restricts the access of retirees and active duty families to a network of civilian preferred providers who have agreed to charge a discounted rate for medical treatment and procedures. In return, the government pays a higher share of the costs. Since there is no formal enrollment process in TRICARE Standard or TRICARE Extra, eligible beneficiaries are free to access services and switch between network and nonnetwork providers as they prefer, subject to each plan's cost-sharing requirements. Active duty personnel and most Medicare-eligible beneficiaries are ineligible for TRICARE Standard or TRICARE Extra.

2.2 Overview of Risk Adjustment Models

A fundamental challenge to efficient health care management is to identify necessary and appropriate services and expected health outcomes, both of which are primarily affected by the health status of individual beneficiaries. Thus, if the patients seen within one health care system cost more, or fare less well, than average, it is not clear whether this is due to system inefficiencies or simply to its patients being sicker than average (adverse selection). In the simplest case, we would expect to pay more (and have worse health outcomes) for a group of 60-year-olds than for the same number of 20-year-olds. As important as age is for health, however, health status matters more—the expected health care costs and health outcomes for a healthy 60-year-old are lower than for a 20-year-old with renal failure. Monitoring and managing any health care delivery system (especially one as diverse as TRICARE's) in which providers (and subsystems) differ substantially in their patient mix requires a credible methodology for accounting for such differences.

Starting in the early 1980s, the Health Care Financing Administration (HCFA)—now the Centers for Medicare & Medicaid Services (CMS)—funded several teams of researchers to develop so-called "population-based" risk adjusters, to enable "health-based" payments to Medicare HMOs. A population-based risk adjuster uses health-relevant data (such as demographics and the medical problems recorded among patients seen during the current year) to predict total health care spending for a specified period (such as "next year"). CMS needed to be able to measure and pay for differences between the health care needs of Medicare beneficiaries who remain in the traditional fee-for-service (FFS) sector and those who chose to receive their benefit in the managed care sector. For individual HMOs, the government was legally required to base its capitated payment on an assessment of what its Medicare enrollees would have cost had they remained in FFS (Tax Equity and Fiscal Responsibility Act of 1982 [TEFRA]). In the context of payment systems, the distinction is often made between "risk assessment," in which differences in expected health care costs among populations are quantified, and "risk adjustment," which refers to a protocol for adjusting payments based on those assessments. Here, we will use the term risk adjustment broadly, to refer to techniques for calculating health-based estimates of patient's service needs and/or outcomes and for enabling comparisons of actual versus expected outcomes for any patient subgroup. As such, risk adjustment is crucial for both detecting and describing differences in health status among the patients of different providers and for comparing the efficiency and effectiveness of providers whose patients are differentially sick.

CMS' funding encouraged the development of several systems for population-based risk adjustment, most notably the Adjusted Clinical Groups (ACGs, associated with Jonathan Weiner and Barbara Starfield at Johns Hopkins), the Global Risk Adjustment Models (GRAM, Mark Hornbrook at

the Center for Health Research, Kaiser), CD-RISK (Grace Carter at RAND), the Chronic Disability Payment System (CDPS, Rick Kronick at the University of California, San Diego), and Diagnostic Cost Groups (DCGs, Arlene Ash and Randall Ellis at Boston University and the Boston Medical Center and Gregory Pope at RTI). Such systems use the data available on standard claims or encounter records (basically, the diagnoses recorded by clinicians during a year of patient care) plus simple demographics (mainly age and sex) to predict total health care expenditures in Medicare, Medicaid, and commercially insured populations. In response to a Balanced Budget Act of 1997 (BBA) mandate for health-based payments to Medicare+Choice health plans, CMS conducted a comparative evaluation of systems. At that time, the diagnostic data available for risk adjusting was from inpatient claims only, and the DCG team had identified the only model, the Principal Inpatient Diagnostic Cost Group (PIP-DCG) model, specifically designed to balance concerns about providers "gaming" the system (with unnecessary hospitalizations that could boost risk-adjusted payments) with accurately predicting costs from such data. In 2000, Medicare payments to HMOs started to reflect PIP-DCG risk assessments.

In choosing among risk adjustment methods, there is a substantial literature regarding quantitative performance measures that may inform this choice and qualitative features that should also be considered. A path-breaking paper describing useful quantitative measures is "How Well Do Models Work? Predicting Health Care Costs" by Arlene Ash and Susan Byrne-Logan (1998). Robert Cumming's 2002 report for the Society of Actuaries entitled "A Comparative Analysis of Claims-Based Methods of Health Risk Assessment" largely followed and somewhat extended the methods proposed by Ash and Byrne-Logan in comparing the predictive ability of several of the previously listed population-based risk adjustment methods plus Episode Risk Groups (ERGs, by Symmetry Health Data Systems) (Cumming, Knutson, Cameron, & Derrick, 2002).

In 2002, in preparation for moving to all-encounter risk-adjustment models (that is, models that extract information, principally diagnoses, from ambulatory as well as inpatient care), CMS reviewed the available risk adjustment systems (including most of those mentioned above plus Clinical Risk Groups, or CRGs, from 3M Health Information Systems) and again selected a DCG model (DCG/HCC) as the basis for Medicare+Choice payments to begin in January 2004.

2.2.1 The Johns Hopkins ACG Case-Mix System

Developed by: Jonathan Weiner, Barbara Starfield, and Chris Forrest, Johns Hopkins University

Accessibility: Available commercially from DST Health Solutions and for research directly from Johns Hopkins University.

Classification Method (in brief): All ICD-9 (or -10) inpatient and ambulatory diagnoses are assigned to 32 Adjusted Diagnosis Groups (ADGs), based on clinical criteria of severity, duration, diagnostic certainty, etiology, and specialty care (Weiner, Starfield, Steinwachs, & Mumford, 1991; Weiner, Starfield, & Lieberman, 1992; Weiner et al., 1996; Starfield et al., 1991). ADGs are combined with age and gender to produce mutually exclusive Adjusted Clinical Groups (ACGs) based on clusters of ADGs using a branching algorithm. An individual may have many ADGs, but only one ACG. National reference weights as well as locally calibrated cost weights are available as part of ACGs for Windows

software. The system also includes Expanded Diagnosis Clusters (EDCs), a categorization scheme that collapses all ICD codes into approximately 260 disease markers, providing a tool for easily identifying people with specific diseases or symptoms. The ACG Predictive Model (ACG-PM) combines ACGs, selected EDCs, and markers for "high likelihood of hospitalization" and "frailty." The ACG-PM predicts persons likely to become high risk users of services and incorporates weights for populations both under and over age 65. To find out more, see http://www.acg.jhsph.edu.

2.2.2 Chronic Disease and Disability Payment System (CDPS)

Developed by: Richard Kronick, Todd Gilmer, Tony Dreyfus, and Lora Lee, U.C, San Diego

Accessibility: Available at http://www.medicine.ucsd.edu/fpm/cdps

Classification Method (in brief): Individuals are assigned to one or more of 67 medical condition categories using diagnoses and to one of 16 age-gender categories (Kronick , Zhou, & Dreyfus, 1995; Kronick, Dreyfus, Lee, & Zhou, 1996; Kronick & Dreyfus, 1997; Kronick, Gilmer, Dreyfus, & Lee, 2000). "Diagnoses are initially assigned to chronic condition categories … [which] are arranged into hierarchies. Only the highest cost category in a disease hierarchy is used to produce an individual's total risk score … [which] is computed by adding the weights for the age and gender category and any medical categories across the hierarchies. Within the hierarchies only the highest cost category identified is used to assess risk." (Cumming, et al., 2002, p. 48). To find out more, see http://www.medicine.ucsd.edu/fpm/cdps.

2.2.3 Clinical Risk Groups (CRG)

Developed by: 3M Health Information Systems (HIS) and NACHRI

Accessibility: Available commercially from 3M HIS.

Classification Method: CRGs are a classification tool that focuses on identifying individuals with chronic conditions (Muldoon, Neff, & Gay, 1997). Its primary data are diagnosis codes; however, procedures may be used to adjust severity levels within groups (e.g., amputation for diabetics) or where they are so profound that they redefine the clinical classification of the enrollee (e.g., history of a major organ transplant). Codes are assigned to categories. Diagnostic data are further identified as acute or chronic. The system creates three alternative classification models. These primarily differ in their treatment of people without a history of chronic conditions and in their use of internally unvalidated data but are otherwise virtually the same. Individuals are assigned to one of nine core health status groups, then into specific risk adjustment groups, which are further refined with two, four, or six severity levels for chronic disease groups. The most detailed classifications may be rolled up in a series of discrete steps, which reduces the number of groups. The standard model has 1,073 groups, which are rolled up in steps to 415, 150, and 37 groups. The comparable numbers for the prospective model are 1,098, 441, 176, and 49. The comparable numbers for concurrent model are 1,104, 446, 185, and 52. Weights are calculated using normalized averages. Demographic and other adjustments are available. To find out more, contact Julie G. Beard, phone: 801-265-4552; e-mail: jgbeard@mmm.com.

2.2.4 Diagnostic Cost Groups (DCG)

Developed by: Arlene Ash & Randall Ellis, Boston University; Greg Pope, RTI

Accessibility: Available commercially from DxCG Inc, Boston

Classification Method (in brief): Diagnoses are mapped to DxGroup categories, which map into 184 Hierarchical Condition Categories (HCCs) (Ash, Porell, Gruenberg, et al., 1986; Ash et al., 2000; Pope et al., 2000; Ellis & Ash, 1995-96; Ellis et al., 1996). Condition categories are arranged in hierarchies of similar type and financial impact, generally by body system or disease. Within a disease hierarchy, only the most severe condition impacts an individual's risk assessment. Each individual may be assigned to one of 32 age-gender categories and anywhere from zero to multiple HCC's, reflecting his or her full spectrum of medical conditions. The demographic and HCC assignments are used (based on a regression model in a national benchmark population) to assign relative risk scores (RRS). In the prospective model used here, the RRS is proportional to next year's expected cost, that is, a person with an RRS score of 2 is expected to cost twice as much as a person with an RRS of 1. To find out more, visit http://www.dxcg.com.

2.3 Selected Groups for Model Comparison

One approach for comparing the predictive performance of the risk adjustment models is in terms of costs for subgroups of the sample population, with subgroups defined on the basis of cost, diagnoses, or other characteristics in the first year (FY 2001). We make comparisons for the following groups.

Demographic and Service Characteristics: It is useful to know if models fit particularly well, or poorly, on different subgroups of the population. We have shown the model fit on the characteristics of combat status, service, rank, beneficiary category, catchment area, primary care manager type (as of September 2001), region of residence, and urbanicity of residence. Models that fit well on large segments of the population, but perform less well on small subgroups may be considered superior to those that fit moderately well on all segments.

Medical Condition Groups: We are interested in comparing the predictive performance for important medical conditions. It is useful in choosing a risk adjustment method to know if there are certain disease groups that will be poorly estimated by a model. These groups may be separately reimbursed by a prospective payment system or otherwise compensated. Furthermore, it may be advantageous to choose a model that does well for common medical conditions among the population being modeled, to reduce the need for special payment schemes. We selected these conditions as follows. We used the three-digit ICD-9 diagnostic codes found in the inpatient (SIDR, HCSR-I) and outpatient (SADR, HCSR-N) files for FY 2001, excluding diagnoses from telephone consults. Three variables were created for each of the three-digit ICD-9 codes: the average total cost per person with the condition, the total cost of all individuals with the condition, and the prevalence of the condition in the half million-beneficiary validation sample. The ICD-9 codes were then sorted by each of these three variables and then a subset of three-digit codes was chosen. The subset included all codes found in at least one of the following categories:

1. Top 25 most prevalent.
2. Top 25 with highest average cost of treatment.
3. Top 25 with highest total treatment cost.

Furthermore, we included the ICD-9 code only if it had a prevalence of at least 400 and a total cost of at least 10 million dollars. A table showing each of these groups is in Appendix E. Transient clinical conditions were excluded because we were interested in exploring the predictive capabilities of the risk adjustment models. Since different manifestations of the same disease could be classified across more than one three-digit ICD-9 code, we used published medical literature to group related codes. From this list of grouped ICD-9 codes, or chronic medical conditions, in consultation with the DoD sponsor, we chose nine conditions for analysis: diabetes mellitus, female breast cancer, major mental health disorders, ischemic heart disease, congestive heart failure, chronic obstructive pulmonary disease (COPD), asthma, chronic renal failure, and osteoarthritis.

Prior Cost Percentiles: As the focus of this study was on health costs, we assessed model predictive performance based directly on groups formed by their average annual costs in the first year (FY 2001). Our analyses used the following groups:

1. Lowest 20 percent cost percentile.
2. Next highest 30 percent (i.e., 21 to 50 percentile).
3. Next highest 30 percent (i.e., 51 to 80 percentile).
4. Next highest 10 percent (i.e., 81 to 90 percentile).
5. Next highest 5 percent (i.e., 91 to 95 percentile).
6. Second highest 4 percent (i.e., 96 to 99 percentile).
7. Highest 1 percent.

50 Prior Cost 2-Percentiles: A variation of the above cost groups, this is a finer partitioning of the validation sample. The validation subset was sorted by total FY 2001 cost per person and then divided into 50 groups, each comprising 2 percent of the population, each group representing individuals with increasingly higher costs in the base year.

2.4 Model Comparison Criteria

As recent comparative studies have pointed out, no single measure is comprehensive and sensitive in distinguishing differences in predictive performance of different risk models (Ash and Byrne-Logan 1998; Cumming, Knutson et al., 2002). Several measures have been proposed, both at the individual level and at levels of group aggregation. In the results section, we examine several summary measures of model performance, including R-squared (R^2), Cummings Prediction Measure (CPM), and

Mean Absolute Prediction Error (MAPE). Also, we have calculated prediction ratios and grouped R^2 values in subgroups of the validation sample.

Individual Measures of Predictive Accuracy

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(a_i - \hat{a}_i)^2}{\sum_{i=1}^{N}(a_i - \bar{a})^2} \quad (1)$$

where

a_i is the actual claim dollars per person *i*

$\hat{a}_i$ is the predicted claim dollars per person *i* (based on the regression model)

$\bar{a}$ is the average actual claim dollars per person *i*

i goes from 1 to N (the total number of people)

R^2 indicates the proportion of the variation in a_i that is accounted for by the prediction model; so R^2 is a fraction ranging from 0 to 1, which represents between 0 and 100 percent of explained variation. If $R^2 = 1$ then all the variation in the cost per person can be explained by the model. When used for comparison, the model with the R^2 closest to 1 yields predictions whose averages per individual are closest to actual outcomes. One weakness of R^2 is that it gives more weight to large errors, and therefore is more likely to be influenced by persons with large claims. As the numbers and nature of large outliers vary from dataset to dataset, R^2 measures may also be unstable. Also, R^2 for diagnoses-based risk adjustment models typically vary from 10 percent to 20 percent for *prospective* applications, giving the impression of rather poor performance. This is primarily due to the fact that, in a large diverse population, actual costs are zero for a majority, and among those with positive costs, unexpected acute conditions account for another large share.

$$\textit{Cumming's Prediction Measure (CPM)} = 1 - \frac{\sum_{i=1}^{N}\left|a_i - \hat{a}_i\right|}{\sum_{i=1}^{N}\left|a_i - \bar{a}\right|} \quad (2)$$

where

a_i is the actual claim dollars per person *i*

$\hat{a}_i$ is the predicted claim dollars per person *i* (based on the regression model)

$\bar{a}$ is the average actual claim dollars per person *i*

i goes from 1 to N (the total number of people)

Cumming's Prediction Measure (CPM) is advantageous in that it is a standardized measure similar to R^2, but also gives equal weight to small and large errors. Note that CPM is similar in definition to R^2, with values ranging from 0 to 1—this gives the proportion of the sum of absolute deviations from mean in individual costs that is explained by the risk model

$$\textit{Mean Absolute Prediction Error (MAPE)} = \frac{\sum_{i=1}^{N} |a_i - \hat{a}|}{N} \quad (3)$$

where

a_i is the actual claim dollars per person *i*

$\hat{a}_i$ is the predicted claim dollars per person *i* (based on the regression model)

N is the total number of people in the sample

The advantage of MAPE is that persons with large claims do not influence it because equal weight is given to small and large errors. However, one disadvantage of this measure is that the final estimate is not a standardized measure; making comparisons across studies harder to interpret.

Group Measures of Predictive Accuracy

Here the focus is on predicting costs for certain subgroups of the population. Group measures give a richer sense of how models will perform for specific tasks. If the task at hand matches the method of the grouped test reported, then these scores might be the most useful for deciding among the various risk adjustment models.

$$\textit{Predictive ratio} = \frac{\text{Total predicted costs for group}}{\text{Total actual costs for group}} \quad (4)$$

The predictive ratio for a subgroup of the population is its predicted Year 2 costs divided by its actual costs; thus predictive ratios less than 1 indicate underpayment, and predictive ratios greater than 1 indicate overpayment.

$$\textit{Grouped } R^2 = 1 - \frac{\sum_{b=1}^{B} w_b * (AveY_b - Ave\hat{Y}_b)^2}{\sum_{b=1}^{B} w_b * (AveY_b - \bar{Y})^2} \quad (5)$$

where b = 1, 2, …, B are partitions of the sample

$AveY_b$ = the mean of actual costs for each partition b

$Ave\hat{Y}_b$ = the mean of predicted costs for each partition b

$\bar{Y}$ denotes the mean for the entire sample

This measure is analogous to the standard R^2, where each partition is treated as a distinct unit. Also note that unlike in the standard R^2, here each partition is weighted in accordance with its "importance"—in the present analysis this weight is determined by the partition size (i.e., number of individuals in each partition). As with the standard R^2, this measure gives the proportion of the variation across partitions accounted for by the prediction model. Typically, it ranges from 0 to 1. For illustration, a grouped R^2 of 0.5 indicates that 50 percent of the variation in partition-level (mean) costs is accounted for by the risk adjustment model.

3

Data

As the focus of this study is on TRICARE Prime, the data we requested covered all TRICARE Prime enrollees during the fiscal years 2001 and 2002 who resided in the continental United States, plus Alaska and Hawaii. For each year, there is one beneficiary information file, one enrollment file, and five health care utilization files. A unique identification number makes it possible to match data for each individual across files.

We used the following data source filenames and extracts from the MHS Data Repository for individual- and claim-/encounter-level information as follows:

- *Point in Time Extract (PITE) and TRICARE Enrollment File (TEF)*: These two files have data on demographic, enrollment and other individual characteristics as of the last month of each fiscal year.
  - Demographic: age, sex, race
  - Geographic location: five-digit zip code, residence region, urbanicity,[4] catchment area
  - Enrollment: monthly Alternative Care Value (Active Duty Prime, TRICARE Senior Prime, Non-Active Duty Prime)
  - Other: DoD occupation code (combat / non-combat), sponsor service (Army, Navy/Marines, Air Force, Other), service rank, beneficiary category (active duty/guard, dependent, retired, dependent of retired)
- *Standard Inpatient Data Record (SIDR)*: Diagnosis codes (8), procedure codes (8), completed full cost, completed incremental cost, treatment Military Treatment Facility (MTF,) and dates of service.
- *Standard Ambulatory Data Record (SADR)*: Diagnosis codes (4), procedure codes (5), completed full cost, completed variable cost, appointment status (appointment/walk-in, sick call/telephone consult), provider specialty, type of provider (military/civilian) and dates of service.
- *Health Care Service Record—Institutional (HCSR-I)*: Diagnosis codes (8), procedure codes (6), completed allowed amount and completed amount paid, dates of service, site of service.
- *Health Care Service Record—Non-Institutional (HCSR-N)*: Diagnosis codes (5), procedure codes (5), completed amount allowed, completed amount paid, lab procedure flag, denial of claim, type of provider (military/civilian), dates of service. This data file has space for up to 25 procedure codes per claim. However 97 percent of the claims have five or fewer codes. To economize on data size and processing, we requested data from only the first five procedure codes per claim.

[4] This is a derived indicator based on a three-digit zip code.

- *National Mail Order Pharmacy (NMOP)*: Total prescription price (used in FY 2001)
- *Pharmacy Data Transaction Service (PDTS)*: Net amount due (used in FY 2002)

Both the PDTS and the NMOP contain only a portion of all the pharmacy costs in the TRICARE system. Some pharmaceutical costs are rolled-up with ambulatory claims, thus appearing to be part of a claim for ambulatory care. There are similar issues with laboratory claims, which cannot be identified within the military sector (SADR), although they are identifiable in the civilian sector (HCSRN). To characterize illness burden more accurately, it is best to exclude diagnoses that appear only in laboratory records. As these could not consistently be identified, all diagnoses, whether from known laboratory visits or not, were included in the data modeled here. (Appendix Table F6 compares the effect of this inclusion.) More extensive detail on the data extraction and data inputs for the risk adjustment models may be found in Appendices C and D.

Analysis

4.1 Study Population

These analyses required information from TRICARE Prime enrollees in FY 2001 to predict outcomes for these enrollees in FY 2002. To insure complete accounting of diagnostic data in FY 2001 and costs in FY 2002, a sample of continuously enrolled subjects was required. From the approximately 4.5 million enrollees in TRICARE Prime in 2002, we selected those enrolled in all 24 months of FY 2001 and FY 2002 and residing in the continental United States, Hawaii, or Alaska. Additionally, subjects had to be present in the September 2001 beneficiary file, from which important demographic characteristics were assembled. Finally, only those individuals age 64 or younger as of September 2001 were included, to eliminate those with overlapping enrollment in Medicare. The enrollment files contained data on 4.5 million enrollees in FY 2001 and 4.6 million in FY 2002, with 3.9 million people appearing in the files for both years. However, only 2.6 million of them were enrolled during all 24 months of these 2 years. After applying our selection criteria, 2.3 million individuals were left in our study population. Transfers across bases within the United States did not affect inclusion as long as subjects continued to be enrolled in TRICARE Prime. Note that our focus is on TRICARE Prime utilization within the United States. Some individuals excluded were likely due to transfers to overseas bases or combat duty status. Appendix Table B1 compares subjects who were not continuously enrolled (~1.1 million) with those who were. These groups are similar with respect to most characteristics.

To get information about age, sex, and other demographic factors, we linked our data with the September 2001 beneficiary files (PITE & TEF), and then selected the 2.3 million enrollees who were also in these files. Finally we required that all subjects be age 64 or younger as of September 2001, so that we would not potentially lose track of subjects also enrolled in Medicare. The chart below gives the exact counts for each step of this process[5]

Prime enrollees in FY 2001	4,486,060
Prime enrollees in FY 2002	4,645,181
Prime enrollees in both years	3,951,194
Prime enrollees for all 24 months	2,610,041
Prime enrollees for all 24 months & in September 2001 beneficiary file	2,307,187
Prime enrollees for all 24 months, in September 2001 beneficiary file & aged 64 or younger as of September 2001	***2,304,926***

[5] Programming details in task1_sample.sas in Appendix D).

4.2 Fitting and Validation Samples

To make unbiased evaluations of the different risk models, the sample population used in obtaining the fitted estimates (fitted sample) was different from the sample population used for obtaining cost predictions (validation sample). From the 2.3 million enrollees in the overall study sample of TRICARE Prime beneficiaries, 0.5 million were randomly selected for a validation sample. The remaining 1.8 million enrollees served as the estimation (or fitting) sample. The estimation sample was used to obtain prospective risk weights—using sex, age, and diagnoses for the base year (FY 2001) and actual expenditures for FY 2002. These risk weights (along with base year information on sex, age, and diagnoses) were then used to obtain predicted prospective expenditures for the validation sample in FY 2002. Predictive accuracy was based on comparing predicted expenditures to actual FY 2002 expenditures for the validation sample.

4.3 Calculating Cost

Our interest is in the total health care costs incurred by each TRICARE Prime beneficiary in each fiscal year. The total includes all inpatient, outpatient, and prescription medication costs over the year. In the case of inpatient and outpatient care provided in the MHS, we aggregated the Completed Full Cost as obtained in SIDR and SADR. The costs of prescriptions dispensed (outpatient) in MHS are also covered in SADR. Note that these costs for encounters in MHS are computed *ex post*. In the case of inpatient care, the costing algorithm (called Patient Level Cost Allocation [PLCA]) uses a mix of patient-level information, including occupied bed days, Diagnosis-Related Groups (DRGs), and surgical DRGs to allocate ward-level resource costs (categorized as direct, clinician salaries, support, surgical and ICU, and ancillary).[6] The mail order filled prescription costs from MHS for FY 2001 are in NMOP and for FY 2002 are in PDTS.

Costs of inpatient and outpatient health care received from private providers are obtained as allowed amounts from claims records in HCSR-I and HCSR-NI. The prescription costs from non-mail order fills are included in HCSR-NI. The costs of mail order prescriptions from private providers for FY 2001 are in NMOP and for FY 2002 are in PDTS.

As with any health care utilization data, a tiny fraction of beneficiaries had very large costs. At the extreme, one person in the population (a child hemophiliac) cost $3.8 million in FY 2001, and a few others in each year cost between $500,000 and $1 million. To ensure that our analyses were not distorted by these very high-cost cases, we used each model to predict "all costs" and two "top-coded" cost variables, one top-coded at $25,000 and the other at $50,000. For example, if $25,000 is the top-coding threshold, everyone with higher total costs had their total cost reset to $25,000. This is a standard practice to limit the potentially significant influence on risk estimates of this small group of high-cost cases—data fit to top-coded costs are more stable than those fit to all costs. Top-coded models are also consistent with a prospective payment system that includes reinsurance for cases that exceed the top-coding threshold.

[6] Additional details on PLCA are in *Patient Level Cost Allocation (PLCA), Detailed Description with Examples* (TMA, 2000).

4.4 Diagnostic Information

All diagnoses for the first year (FY 2001) from the inpatient (SIDR, HCSR-I) and outpatient (SADR, HCSR-N) files were included. Repeated diagnoses were retained, although neither the ACGs, CDPS, nor DCG model treats more than one appearance of the same diagnostic code any differently than a single appearance. CRGs use multiple occurrences of a diagnosis for the purposes of validating the presence of a clinical condition, as is done when identifying diabetes and asthma in the construction of Health Plan Employer Data and Information Set (HEDIS®) measures; alternatively, in some contexts, the CRG risk group assignment may be refined by the recency, persistence, or recurrence of clinically significant comorbid conditions.

Appointment Status: SADR files include walk-ins as well as telephone consults; 11.6 percent of the SADR records are telephone consults. Diagnoses from the telephone consults were excluded in the analysis, although the costs were included.

Laboratory Services: Diagnoses related to laboratory services are often excluded from risk adjustment modeling as these diagnoses are often rule out or provisional in nature and do not necessarily reflect an underlying clinical condition or morbidity. Risk adjustment may be improved by excluding diagnoses codes from claims submitted by laboratories; thereby increasing the likelihood of including only confirmed diagnoses. Unfortunately, excluding laboratory claims was not feasible as such encounters are not identifiable in military sector data (SADR), although they are identified in the civilian sector data (HCSRN). Therefore, for consistency, diagnoses codes from all laboratory services were also included from both military and civilian care sectors. Appendix Table F6 calculates the difference in predictive accuracy resulting from whether or not lab-only records are excluded. All the other results in this report have laboratory records included.

Telephone Consults and Intern/Resident Workloads: These services provided in MHS are included in SADR encounters, although they do not have diagnoses or costs assigned. Because such services are not included in private provider claims (i.e., in HCSR-NI), we excluded these services from our analyses.

Denied Claims in HCSR-NI: Although a sizable portion of HCSR-NI records (5.3 percent) had denied claims, we included all the costs and all the diagnoses in the analyses. This is because, to the extent possible, we wish to capture the resources needed to treat each person in FY 2002, regardless of who paid for it. If some group of people had, for example, 80 percent of their costs paid by another system, we would rather capture all costs and apply a 0.2 multiplier to their expected total cost (to predict what they would cost the DoD) than view them as being much less medically needy than they really are.

4.5 Other New Variables

4.5.1 Primary Care Manager: Military or Civilian

We determined whether the primary care manager (PCM) was military or civilian using the enrollment DMISID field from the TEF. Those with an M (managed care) DMISID affiliation were defined as having a civilian PCM, and all others as having a military PCM. Note that this was determined

as of September 2001, and changes in PCM type were not made in the assignment of military or civilian status.

Table 4-1 gives the list of the key variables in the analytical dataset.

Table 4-1. Key Analytical Variables

Variable	Description
cost2001, cost2002	Total of health care costs incurred in each fiscal year (as described above).
diagnostic condition categories	Each risk adjustment model provides a distinct set of these dichotomous (present/absent) variables.
age and sex	Age on September 30, 2001. Each risk adjustment model creates different age and sex groupings.
beneficiary	Beneficiary category (active duty or guard, retired, dependent of active duty or guard, dependent of retired or survivor).
catchment	A dichotomous indicator (inside or outside catchment area). A catchment area is an approximately 40-mile radius around a military hospital, allowing for natural geographic boundaries and transportation accessibility.
occupation code	Indicates combat or non-combat status.
PCM type	Indicates whether primary care manager is military or civilian, as of September 2001.
race	Race groupings are White, Asian or Pacific Islander, Black, American Indian or Alaskan Native, Other, Unknown. No ethnicity variables are available in these data.
rank	Jr Enlisted, Sr Enlisted, Jr Officer, Sr Officer, Warrant, Other
region	Region of the United States in which the beneficiary resides; there are 13 in total.
service	Army, Navy or Marines, Air Force, Other
urbanicity	Rural, Suburban, Urban (using Beale Codes)

4.6 Model Implementation

In implementing the three models, we sought to ensure uniformity in all the steps. The implementation steps are as follows:

1. The same input datasets were used for the ACG, CDPS, and DCG models. Data on age, gender, and diagnoses for FY 2001 for the entire subpopulation were used to obtain diagnosis-based group indicators for these models. To run the CRG model, 3M Health Information Services took advantage of additional available data that included procedures provided at the time of service, date of diagnosis or procedure, site of diagnosis or procedure, and provider associated with the diagnosis.

2. The subpopulation was randomly split into a validation sample of 500,000 and a fitting sample of 1.8 million. The fitting sample data included age, gender, and model-specific

diagnosis groups for FY 2001 and health care expenditures for FY 2002 (and for the CRG model the additional variables mentioned above). The validation sample data had the same information but without health care expenditures for FY 2001.

3. Risk scores for each person for each model were obtained from the fitting sample by regressing prospective expenditures (FY 2002) on current age, gender, and model-specific diagnoses groups. The age and gender categories were identical across ACG, CDPS, and DCG models (age groups 0 to 14, 15 to 24, 25 to 44 and 45 to 64 years). The CRG model used slightly different age categories (<10, 10 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 years); however, the CRGs can readily accommodate any demographic categorization scheme without altering the basic model. The Boston University investigators on this project performed the regressions for the ACG, DCPS, and DCG models. The CRG software vendor (3M Health Information Systems) chose to develop normalized averages with their own copy of the data and provided the CRG risk scores that were merged onto the dataset used in producing this report.

4. The key step in obtaining model predictions is to estimate a statistical relationship between FY 2002 costs and the model-specific determining factors in FY 2001 (demographic and diagnosis-related condition categories) using the fitting sample. For each risk adjustment model, this step gives a set of risk weights, one for each demographic and diagnosis-related condition category. Each risk weight denotes the predicted dollar amount associated with that demographic or diagnosis category. We summed the model's risk weights to obtain the predicted cost for FY 2002 for the validation sample.

5. Predictive accuracy from each model was evaluated by comparing predicted expenditures for the FY 2001 validation sample with actual FY 2002 expenditures

Details on the data extraction, including the separate data supplied to 3M Health Information Systems, are in Appendix C. Details on the data processing for data input, regressions run by each of the software programs, model outputs, and programs used in the analyses are in Appendix D.

In addition to the risk adjustment models, a simple model, using age and sex alone to predict costs, is also shown in several of the results tables. This is a useful comparison to illustrate the significant gain in accuracy from using any of these more complex health risk adjustors.

While CDPS and DCG risk models are SAS programs, the other risk models are stand-alone software. All four models run were fast (all runs completed in less than 30 minutes). Further, all models are relatively easy to recalibrate and re-specify, if so desired. The primary output from each model is a vector of dichotomous variables denoting the presence or absence of each of the conditions grouped by the risk model. Implementing each model required specifying certain settings. In the case of CDPS and ACG, the Boston University team consulted with the developers on the selection of appropriate settings. The models and software also differ in the nature and extent of additional output, such as data quality reports and summary tables, produced.

5

Results

5.1 Description of the Study Population and Cost Characteristics

5.1.1 Cost Characteristics

For each individual in FY 2001 and FY 2002, six files from the TRICARE Prime claims database were used to construct the cost of care. Table 5-1 shows the characteristics of the files used, by claim (or record of each medical service or encounter). The Standard Ambulatory Data Record (SADR) and the Health Care Service Record: Non-Institutional (HCSR-NI) contained the most records. Mean cost per claim for these ambulatory encounters ranged from $115 to $184; mean inpatient claims in the Health Care Service Record: Institutional (HCSR-I) or the Standard Inpatient Data Record (SIDR) ranged from about $4,500 to $6,500. Note that the growth in the average inpatient claim cost between FY 2001 and FY 2002 in both files ranged from about $500 to $800, or 9 to 11 percent. The growth in ambulatory costs per claim was between $1 and $9 (from < 1 to 5 percent) in the same time period.

Pharmacy costs recorded in National Mail Order Pharmacy (NMOP) and Pharmacy Data Transaction Service (PDTS) are really only a fraction of the true pharmacy costs. In our analysis file, only 2 percent of the study population had pharmacy claims. Some of the costs of pharmaceuticals were included in ambulatory claims; however, further investigation of this issue is required. Therefore, the numbers listed as *pharmacy* in the table below are not a measure of all pharmacy costs, just those in these databases.

The smallest cost per claim type was for pharmacy items, $90 to $98 per claim in the two datasets. The SADR was the only file to have negative cost records and had the most claims with no costs, or missing costs. Negative cost, no cost, and missing cost records may be due to poor data quality, but may also reflect adjustments made in the claims payment process. SADR was the largest file. Only 1 percent of its many claims were missing cost information, whereas 3 percent of the SIDR file records had no cost data. The SADR file also accounted for, by far, the most total money in the claims system, with $2.2 billion worth of claims each year. The total cost of all claims in FY 2001 was $3.9 billion, growing to $4.1 billion (5.7 percent growth) for this same group of 2.3 million continuing TRICARE Prime enrollees in FY 2002.

Table 5-2 shows the characteristics of these databases by individual-level cost for the same 2.3 million continuing enrollees. That is, the unit of analysis is now a person-year, and all claims in each file for each beneficiary are aggregated. *Average individual cost* shows the average cost across all people, not just for those individuals with any claim of the indicated type. *Maximum individual* cost is the largest total year's claims costs for a single person. The penultimate column shows the proportion of the population with any claim in each file; 2 percent of the beneficiaries in this population had an inpatient stay, which resulted in a HCSR-I claim, and 85 to 86 percent of the population never had any type of claim filed in each year.

Table 5-1. Claim-level Costs by Data Source (N = 2.3 million)

Data file	Fiscal year	Number of records	Cost per claim record ($)			Number of records with $		Total cost ($ millions)
			Mean	Maximum	SD	Zero	Missing	
Inpatient								
HCSR-I*	2001	60,355	4,574	857,839	11,570	11	0	276.1
	2002	64,277	5,068	610,622	9,884	11	0	325.7
SIDR**	2001	78,119	5,726	815,825	7,305	0	232	446.0
	2002	72,331	6,525	650,667	8,150	0	236	470.4
Ambulatory								
HCSR-N*	2001	8,352,518	115	151,620	451	6,606	0	956.4
	2002	9,457,022	116	99,172	427	2,978[§]	0	1,100.3
SADR**	2001	12,776,227	175	72,337	655	71,595[§§]	129,553	2,211.4
	2002	12,153,452	184	24,082	221	52,251	144,471	2,209.2
Pharmacy								
NMOP†	2001	244,851	90	269,627	590	0	0	22.2
PDTS‡	2002	295,897	98	10,007	233	3	0	29.0
Total ($)	2001							3,912.1
	2002							4,134.7

TRICARE Prime beneficiaries age 64 or younger in September 2001, enrolled for 24 continuous months following and residing in continental United States, Hawaii, or Alaska. HCSR-I: Health Care Service Record Institutional; HCSR-NI: Health Care Service Record Non-Institutional; SADR: Standard Ambulatory Data Record; SIDR: Standard Inpatient Data Record; NMOP: National Mail Order Pharmacy; PDTS: Pharmacy Data Transaction Service.

* Completed amount allowed (variable name *allowed* or *tallowed*)

** Completed full cost (variable name *fcost* or *fullcost*)

†Total price (*rxtotal*)

‡Net amount due (variable name *netamt*)

[§]Two records with negative costs

[§§]98 records with negative costs

Table 5-2. Person-level Costs by Data Source (N = 2.3 million)

Data file	Fiscal year	Individual annual cost ($)			Percentage of people with any cost	Total cost ($ millions)	Percentage growth FY 2001 to 2002
		Mean	Maximum	SD			
Inpatient							
HCSR-I	2001	120	1,110,876	2,520	2.0	276.1	–
	2002	141	649,006	2,403	2.1	325.7	18.0
SIDR	2001	194	815,825	2,038	2.9	446.0	–
	2002	204	668,064	2,285	2.6	470.4	5.5
Ambulatory							
HCSR-N	2001	415	3,784,023[†]	3,414	38.5	956.4	–
	2002	477	906,026	2,685	40.0	1,100.3	15.0
SADR	2001	959	274,325	2,368	74.1	2,211.4	–
	2002	958	136,152	1,722	72.5	2,209.2	–0.1
Pharmacy							
NMOP	2001	10	270,374	249	1.6	22.2	–
PDTS	2002	13	44,204	211	1.7	29.0	30.6
Total	2001	1,697	3,820,360	5,943	85.7	3,912.1	–
	2002	1,794	942,990	5,482	85.2	4,134.7	5.7

TRICARE Prime beneficiaries age 64 or younger in September 2001, enrolled for 24 continuous months following and residing in continental United States, Hawaii, or Alaska. Variable names are the same as for Table 5-1.

[†]Unusually high costs were checked, the case with a $3.7 million yearly cost was a subject in the fitting dataset with ICD-9 code 286.0 (hemophilia and antihemophilic globulin [AHG] deficiency disorder) and was the result of 39 claims, each at least $30,000.

Table 5-3 shows costs by inpatient and ambulatory spending. Costs from both inpatient files are aggregated and show 10 percent growth in total cost between FY 2001 and FY 2002. Smaller growth, 4.5 percent, is seen in the aggregated ambulatory costs in these years.

Table 5-4 divides the beneficiaries by where they received their care: those who received care in the military sector only (46.8 percent in FY 2001, decreasing to 44.9 percent in FY 2002), those receiving care only in the civilian sector (11.6 percent in FY 2001, increasing to 12.7 percent in FY 2002), those who received care in both sectors (27 percent, increasing to 28 percent), and in neither (14.3 percent, increasing to 14.8 percent). Because costs per person went up (even though only 0.5 percent), the total dollars spent by people who used the military sector exclusively decreased somewhat less (by 3.7 percent as opposed to 4.2 percent) than the numbers of beneficiaries did; costs in the civilian sector increased much faster than the growth in beneficiaries cared for, as did costs for those cared for in both sectors. It is surprising to see any continuing population in which the number of people receiving no care increases (here, by 3.3 percent) as the population ages one full year.

Table 5-3. Person-level Costs by Service Setting (N = 2.3 million)

Type of service	Fiscal year	Individual annual cost ($)			Percentage of people with Any Cost	Total cost ($ millions)	Percentage of change in $
		Mean	Maximum	SD			
Inpatient	2001	313	1,146,285	3,337	4.8	722	–
	2002	345	668,064	3,419	4.6	796	10.2
Ambulatory*	2001	1,374	3,791,666	4,206	85.7	3,168	–
	2002	1,436	917,786	3,269	85.2	3,310	4.5

TRICARE Prime beneficiaries age 64 or younger in September 2001, enrolled for 24 continuous months following and residing in continental United States, Hawaii, or Alaska.

*Including pharmacy costs that appear in the ambulatory files, but not those that are separately listed in NMOP and PDTS.

Table 5-4. Person-level Costs by Service Sector Utilization (N = 2.3 million)

Service sector utilization	Fiscal year	N	Percentage of change in N	Individual annual cost ($)			Total cost ($ millions)	Percentage of change in $
				Mean	SD	Percentage of change in mean		
Military only	2001	1,079,253		1,404	3,612	–	1,515	
	2002	1,034,009	–4.2	1,411	2,948	0.5	1,459	–3.7
Civilian only	2001	267,282		1,695	5,745	–	453	
	2002	293,357	9.8	1,888	6,436	11.4	554	22.3
Both	2001	629,209		3,090	9,453	–	1,944	
	2002	637,374	1.3	3,330	8,438	7.8	2,122	9.2
Neither	2001	329,182		0	0	–	0	
	2002	340,186	3.3	0	0	–	0	

TRICARE Prime beneficiaries age 64 or younger in Sept. 2001, enrolled for 24 continuous months following and residing in continental United States, Hawaii, or Alaska.

5.1.2 Study Population

As explained in the analysis section, we randomly selected a half million beneficiaries from these 2.3 million TRICARE Prime enrollees to create a validation sample. The remaining 1.8 million enrollees comprised our fitting sample, so that all risk adjustment models could be fit to this very large, representative subset of enrollees. The validation sample of a half million was more than adequate for producing validated tests of each model's performance. Appendix Table B2 compares characteristics of

the fitting and validation populations as of September 2001. The two samples are hardly distinguishable from each other, verifying that the validation set represents the same population as the fitting data.

The characteristics of the 2.3 million-member sample are shown in Table 5-5. In addition, the mean total individual cost in FY 2002 and its coefficient of variation are given for the fitting and validation samples. The coefficient of variation (CV) is the standard deviation (SD) of a variable in a sample divided by its mean and multiplied by 100. The CVs shown here are typically in the range 300-400, which indicates SDs are 3 to 4 times larger than the mean, which is typical for health care costs (Ash et al., 2000). This shows that the random sample of a half million cases used in the validation subset were not substantially different from those in the fitting subset. It is interesting to note the cost differences by some of these characteristics, however it should be kept in mind that some groups represent few people. There are no important differences in cost between the two subsets.

5.2 Comparing Models: Individual Measures of Predictive Accuracy

We compared the models in their ability to predict future expenditures for a sample of a half million TRICARE beneficiaries. As described in the methods section, each model was fitted on a 1.8 million person sample using costs in FY 2001 and FY 2002. This produced prospective risk weights. These risk weights (along with FY 2001 information on sex, age, and diagnoses) were then fed back into the model to obtain predicted expenditures for the half million-person validation sample in FY 2002. Predictive accuracy is based on comparing predicted expenditures to actual expenditures in FY 2002. Tables 5-6 and 5-7 show the results of this exercise. Five methods of risk adjusting were compared in this analysis: ACG, CDPS, CRG, DCG, and a simple age and sex model. The age-sex model is included to show the potential value of using these risk models as opposed to not including any health data in predicting costs.

Three outcomes were predicted: all costs, and two top-coded cost variables. We examined top-coded costs because in any given year a few individuals have extremely high costs, and such cases are typically very difficult to predict. A model that predicts well in the bulk of cases with more normal costs may be best for many purposes, even if it predicts the few high cost cases less well than other models. Thus, we computed goodness-of-fit statistics comparing results when predicting *all* (unaltered or raw) FY 2002 expenditures and when expenditures were capped (or *top-coded*) at $25,000 or $50,000. For example, in the case of a person whose total claims for FY 2002 were $40,000, the comparisons based on all expenditures and with a $50,000 cap would both use the actual $40,000 figure. In the comparison with a $25,000 cap, the actual $40,000 cost would be replaced in the data with the threshold limit of $25,000. All models will predict FY 2002 top-coded costs better than raw costs; what interests us is comparison across models, for each of the three outcomes. In each block of cells, the model that performs the best is in **bold** face type.

Table 5-5. Demographic Characteristics and Total Individual Cost of Fitted and Validation Samples in FY 2002

	Total sample (N = 2.3 million)	Fitting sample (N = 1.8 million)		Validation sample (N = 0.5 million)	
	Percentage of N	Mean ($)	CV* ($)	Mean ($)	CV* ($)
Gender					
Female	47.9	2,191	263	2,197	274
Male	52.1	1,428	354	1,427	376
Age					
0 to 14	28.6	943	406	955	538
15 to 24	16.5	1,438	329	1,436	343
25 to 44	36.9	1,817	242	1,808	230
45 to 64	18.0	3,425	254	3,454	255
Race					
White	26.9	1,646	271	1,647	276
Black	1.0	1,578	332	1,508	250
Asian or Pacific Islander	7.1	1,742	235	1,745	212
American Indian or Alaskan native	0.2	1,692	215	1,799	251
Other	2.0	1,530	224	1,513	218
Unknown	62.7	2,249	330	2,264	334
DoD Occupation Code					
Combat	15.1	1,365	274	1,380	281
Non-combat	84.9	1,870	303	1,870	319
Beneficiary Category					
Dependent Active Duty/Guard	45.1	1,511	308	1,515	350
Retired	9.3	2,924	301	2,897	289
Dependent Retired or Survivor, Other, Unknown	17.8	2,518	289	2,538	303
Active Duty and Guard	27.8	1,411	221	1,413	215
Rank					
Jr Enlist	11.3	1,690	269	1,696	251
Sr Enlist	7.2	1,829	309	1,826	317
Jr Officer	0.2	1,535	264	1,577	397
Sr Officer	65.7	1,828	300	1,838	329
Warrant	13.1	1,870	316	1,819	298
Other	2.4	2,011	267	2,254	388

(continued)

Table 5-5. Demographic Characteristics and Total Individual Cost of Fitted and Validation Samples in FY 2002 (continued)

	Total sample (N = 2.3 million)	Fitting sample (N = 1.8 million)		Validation sample (N = 0.5 million)	
	% of N	Mean ($)	CV* ($)	Mean ($)	CV* ($)
Catchment area					
Yes	32.6	1,821	304	1,826	330
No	67.4	1,736	298	1,732	288
Resident region					
Northeast (1)	10.5	1,795	283	1,833	418
Mid-Atlantic (2)	11.3	1,495	311	1,491	326
Southeast (3)	13.6	1,784	283	1,781	273
Gulf south (4)	7.5	1,925	278	1,944	281
Heartland (5)	6.5	1,723	290	1,735	331
Southwest (6)	13.6	2,000	309	1,996	275
TRICARE Central (7)	5.7	2,005	345	1,946	281
TRICARE Central (8)	9.5	1,712	280	1,737	280
Southern California (9)	8.7	1,725	330	1,694	401
Golden Gate (10)	3.0	2,068	312	2,013	259
Northwest (11)	5.4	1,761	313	1,805	361
Hawaii Pacific (12)	3.1	1,751	288	1,757	270
Alaska (AK)	1.6	1,672	277	1,704	229
Sponsor service					
Army	35.3	1,757	307	1,769	312
Navy, Marines, Navy Afloat	32.0	1,909	282	1,745	366
Air Force	30.6	1,742	317	1,902	278
Other	2.1	1,373	349	1,347	295
Urbanicity					
Central counties of metro. areas of 1 million pop. or more	30.9	1,877	313	1,848	356
Fringe counties of metro. areas of 1 million pop. or more	3.4	1,832	312	1,894	333
Counties in metro. areas of 250,000 to 1,000,000 pop.	35.0	1,783	300	1,803	289
Counties in metro. areas of less than 250,000 pop.	13.6	1,639	283	1,648	303

(continued)

Table 5-5. Demographic Characteristics and Total Individual Cost of Fitted and Validation Samples in FY 2002 (continued)

	Total sample (N = 2.3 million)	Fitting sample (N = 1.8 million)		Validation sample (N = 0.5 million)	
	% of N	Mean ($)	CV* ($)	Mean ($)	CV* ($)
Urban pop. of 20,000 or more, adjacent to a metro. area	4.2	1,775	270	1,718	253
Urban pop. of 20,000 or more, not adjacent to a metro. area	5.1	1,621	275	1,678	344
Urban pop. of 2,500 to 19,999, adjacent to a metro. area	3.6	1,945	333	1,968	323
Urban pop. of 2,500 to 19,999, not adjacent to a metro. area	1.6	1,896	287	1,833	242
Completely rural (no places with a pop. of 2,500 or more) adjacent to a metro. area	0.3	2,091	261	2,193	268
Completely rural (no places with a pop. of 2,500 or more) not adjacent to a metro. area	0.3	2,267	282	2,291	235

TRICARE Prime beneficiaries age 64 or younger in Sept. 2001, enrolled for 24 continuous months following and residing in continental United States, Hawaii, or Alaska.

*CV = Coefficient of Variation = 100*SD/Mean.

In the validation subset, 0.5 percent of cases had costs greater than $25,000, and 0.15 percent had costs over $50,000; only 0.03 percent of cases cost more than $100,000. The proportions in the fitting subset were the same.

In Table 5-6, we report the three global statistical measures commonly used to evaluate predictive accuracy: R^2, MAPE, and CPM. MAPE and CPM are simple (monotonic) functions of each other (discussed in detail in Section 2.4) and therefore, by definition, they rank models identically. Thus, there are really only two independent measures of global model performance. MAPE is the average number of dollars that is mispredicted per person. There is no standard against which to judge a "good value" for MAPE; however, because all models are compared on the same sample, the smaller the MAPE, the better the model. The R^2 and CPM are generally between 0 and 1 with a 1 indicating that all of the variation in cost is explained by the model (that is, the model's predicted cost equals actual cost for each person) and 0 indicating that no variation is explained (that is, the model is no better than a constant prediction, equal to the population average, for each person). All four models performed similarly, and vastly better than a simple age-sex model, with the DCG model performing a little better than the other models on each global measure. The DCG system always had the lowest average error ($1,573 error for all expenditures, for example), but was only slightly better than the other risk adjustment models (whose MAPEs for unadjusted expenditures ranged from $1,594 for ACGs to a maximum of only $1,632 for CDPS). Similarly, the DCG model's R^2 was always highest (explaining 15.97 percent of the variation in

Table 5-6. Overall Measures of Predictive Accuracy, by Model*

	All expenditures					Expenditures top-coded at $50,000					Expenditures Top-Coded at $25,000				
	Age-Sex	ACG	CDPS	CRG	DCG	Age-Sex	ACG	CDPS	CRG	DCG	Age-Sex	ACG	CDPS	CRG	DCG
$R^2 \times 100$	3.42	14.75	14.71	15.17	**15.97**	6.64	22.85	21.41	21.61	**24.00**	8.26	25.38	23.71	23.41	**26.78**
Cumming's Prediction Measure (CPM×100)	6.93	18.10	16.21	17.32	**19.21**	7.17	19.02	17.10	17.90	**20.23**	7.43	19.66	17.65	18.31	**20.92**
Mean Absolute Prediction Error (MAPE) ($)	1,812	1,594	1,632	1,610	**1,573**	1,717	1,498	1,533	1,519	**1,476**	1,610	1,397	1,433	1,421	**1,376**

Comparisons of actual to predicted FY 2002 cost in the validation subset (N = 500,000) using risk weights developed on the fitting data (N = 1.8 million).

*Better models have higher R^2s and CPMs. In each row-block, the model that performs best is bolded. Since CPM = 1 – constant × MAPE, the higher the CPM the lower the MAPE. Thus, CPM and MAPE are different ways of looking at the same thing, not independent measures of model performance.

all costs, for example), but the other models ranged from 15.17 percent for CRGs (a near-tie with DCGs), down to 14.71—still very respectable—for CDPS. Although we did not formally test the statistical significance of differences in model performance, examining performance under the three cost outcomes provides some insight into the stability of these rankings.

Summary measures of model performance, such as R^2 and MAPE (or CPM) are broad, crude ways to compare model results, and R^2 may be particularly sensitive to a few cases with large prediction errors. It is often more useful to compare predictions for particular groups of interest to see how each model performs. Table 5-7 shows the predictive ratio for each model for many demographic characteristics. The predictive ratio for a model for a group that we report here is the model-predicted cost in FY 2002 for the group divided by its actual cost in the 0.5 million validation data. The average cost per person in FY 2002 was $1,796, and all four models produced an expected expenditure of $1,796; therefore, each of them shows a ratio of 1.0 across top row of the table.

However, the average expenditure was less among the 76,005 beneficiaries with combat status ($1,380), compared to the 423,995 beneficiaries with non-combat status ($1,870). Here, we can see better what the column "ratio of total vs. group mean" refers to. This column shows a ratio of the overall average cost (in the numerator) to the average cost in each group. So, for combat status individuals, this ratio is 1796/1380 or 1.3, which means that, in FY 2002, the average enrollee cost 30 percent more than one with combat status. This ratio is 0.96 for non-combat status, as these enrollees cost 4 percent more than the average. The predictive ratio shows that, for all expenditures, the ACG model correctly predicted this difference between combat and non-combat groups (at least within the 1 percent margin shown). The CDPS and DCG models both overpredicted the cost of combat status individuals by 1 percent. In this case, an overprediction of 1 percent would result in an overpayment to a provider of about $14 per person per year. Note that the models used age, sex, diagnosis, and, in the case of CRGs, some information on timing, nature and place of service; none of the model predictions used information on combat status, or any of the other demographic variables.

The actual costs among beneficiaries in the Army were slightly less than the overall average, and in the Navy were slightly more than average, but all four models did well at predicting costs among those beneficiaries. The costs among Air Force beneficiaries were slightly less than average, and in the "other" group were appreciably less, and each of the models did worse at predicting these groups. It is interesting to note that retired beneficiaries and their dependents had the highest costs, and yet all four models still did remarkably well at predicting them.

The last two rows of the table show the *mean deviation score* and the *weighted mean deviation score*. The *mean deviation score* is the sum of the absolute amount that each model deviates from a perfect predictive ratio of 1.0 in each row, divided by the number of categories in the table (in Table 5-7 there are seven categories). The *weighted mean deviation score* is similar. It also sums the absolute deviation from 1.0, but after it has been multiplied by the proportion of the subjects in each row of the total population, this sum is divided by the number of categories in the table. These measures will then be an average of numbers that are typically in the range 0.0 to 0.08, one weighted by the proportion of the sample to which each number applies. They can be interpreted as a typical-sized percent deviation

Table 5-7. Predictive Ratio for Demographic and Service-Related Subgroups, by Model

	N	Mean of actual expendi-ture ($)	Ratio of total vs. group mean	All costs				Costs top-coded at $50,000				Costs top-coded at $25,000			
				ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG
Total	500,000	1,796	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Combat status															
Combat	76,005	1,380	1.30	1.00	1.01	1.03	1.01	1.00	1.00	1.03	1.00	1.00	1.01	1.03	1.00
Non-combat	423,995	1,870	0.96	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Service															
Army	176,613	1,769	1.02	1.01	1.01	1.03	1.02	1.01	1.00	1.03	1.02	1.01	1.01	1.03	1.02
Navy or Marines	152,389	1,902	0.94	1.00	1.01	0.99	0.99	1.00	0.97	1.00	1.00	1.01	0.96	1.00	1.00
Air Force	160,609	1,745	1.03	0.98	0.97	0.97	0.97	0.97	1.01	0.97	0.97	0.97	1.02	0.97	0.97
Other	10,389	1,347	1.33	1.18	1.21	1.10	1.14	1.16	1.19	1.09	1.11	1.17	1.20	1.09	1.12
Rank															
Jr Enlist	56,408	1,696	1.06	0.94	0.93	0.92	0.96	0.92	0.91	0.92	0.95	0.92	0.91	0.91	0.94
Jr Officer	35,851	1,577	1.14	0.96	0.97	0.99	0.95	0.98	0.98	1.01	0.97	0.98	0.99	1.01	0.97
Other	935	2,254	0.80	0.63	0.64	0.67	0.68	0.69	0.69	0.73	0.73	0.70	0.69	0.74	0.73
Sr Enlist	328,919	1,826	0.98	1.01	1.01	1.00	1.01	1.01	1.01	1.00	1.01	1.01	1.01	1.01	1.01
Sr Officer	65,638	1,838	0.98	1.01	1.02	1.04	1.01	1.01	1.02	1.03	1.00	1.00	1.01	1.02	1.00
Warrant	12,249	1,819	0.99	1.07	1.06	1.06	1.07	1.09	1.06	1.06	1.07	1.07	1.06	1.06	1.06
Beneficiary category															
Active duty or guard	138,830	1,413	1.27	0.95	1.01	0.99	0.95	0.94	1.01	0.98	0.95	0.94	1.01	0.98	0.95
Retired	46,042	2,897	0.62	1.03	1.03	1.00	1.03	1.02	1.02	1.00	1.02	1.02	1.03	1.00	1.02
Dep. active duty/grd	226,327	1,515	1.19	1.00	1.01	1.01	1.00	1.01	1.02	1.01	1.00	1.00	1.03	1.01	1.00
Dep. retired/ survivor	88,801	2,538	0.71	1.02	0.94	0.99	1.01	1.03	0.93	1.00	1.02	1.04	0.93	1.01	1.02
Catchment area															
Inside	337,431	1,826	0.98	0.96	0.97	1.00	0.97	0.97	0.97	1.00	0.97	0.97	0.97	1.00	0.97
Outside	162,569	1,732	1.04	1.08	1.06	1.01	1.06	1.07	1.06	1.00	1.05	1.07	1.05	1.00	1.05

(continued)

Table 5-7. Predictive Ratio for Demographic and Service-Related Subgroups, by Model (continued)

	N	Mean of actual expendi-ture ($)	Ratio of total vs. group mean	All costs				Costs top-coded at $50,000				Costs top-coded at $25,000			
				ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG
Primary Care Manager															
Military	423,826	1,784	1.01	0.96	0.97	1.00	0.97	0.97	0.97	1.00	0.97	0.97	0.97	1.00	0.97
Civilian	76,174	1,864	0.96	1.19	1.16	0.99	1.16	1.18	1.15	0.99	1.15	1.18	1.16	1.00	1.16
Region															
Northeast (1)	52,522	1,833	0.98	0.89	0.90	0.94	0.89	0.91	0.92	0.96	0.91	0.92	0.93	0.96	0.91
Mid-Atlantic (2)	57,043	1,491	1.20	1.07	1.08	1.09	1.07	1.08	1.09	1.09	1.08	1.08	1.08	1.09	1.07
Southeast (3)	67,703	1,781	1.01	1.06	1.06	1.03	1.06	1.04	1.04	1.02	1.04	1.04	1.04	1.02	1.04
Gulf south (4)	37,405	1,944	0.92	1.00	0.99	0.95	1.00	0.99	0.98	0.95	0.99	0.99	0.98	0.95	0.99
Heartland (5)	32,899	1,735	1.04	1.00	0.98	0.98	0.99	1.01	1.00	0.99	1.00	1.01	1.00	0.99	1.01
Southwest (6)	67,923	1,996	0.90	0.97	0.95	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.95	0.96	0.97
TRICARE Central (7)	28,981	1,946	0.92	0.99	0.99	0.99	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.99	0.98
TRICARE Central (8)	47,670	1,737	1.03	1.01	1.00	1.02	1.02	1.00	1.00	1.01	1.01	1.01	1.00	1.01	1.01
S. California (9)	43,036	1,694	1.06	1.02	1.03	1.02	1.01	1.03	1.05	1.03	1.03	1.03	1.05	1.03	1.03
Golden Gate (10)	14,705	2,013	0.89	1.03	1.03	0.98	1.01	1.01	1.01	0.96	0.99	1.01	1.01	0.96	0.99
Northwest (11)	26,747	1,805	1.00	1.03	1.04	1.05	1.03	1.04	1.05	1.05	1.04	1.04	1.05	1.05	1.04
Hawaii Pacific (12)	15,320	1,757	1.02	0.91	0.93	1.00	0.93	0.91	0.92	0.99	0.93	0.91	0.92	0.99	0.93
Alaska (AK)	8,046	1,704	1.05	0.87	0.90	0.96	0.87	0.88	0.89	0.95	0.87	0.88	0.89	0.95	0.87

(continued)

Table 5-7. Predictive Ratio for Demographic and Service-Related Subgroups, by Model (continued)

	N	Mean of actual expenditure ($)	Ratio of total vs. group mean	All costs				Costs top-coded at $50,000				Costs top-coded at $25,000			
				ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG
Urbanicity															
Central counties of metro. areas of 1 million pop. or more	154,431	1,848	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.97
Fringe counties of metro. areas of 1 million pop. or more	17,308	1,894	0.95	0.96	0.95	0.96	0.95	0.99	0.98	0.98	0.97	1.00	0.99	1.00	0.99
Counties in metro. areas of 250,000–1,000,000 pop.	174,839	1,803	1.00	1.00	1.00	1.01	1.01	1.00	1.00	1.01	1.00	1.00	1.00	1.01	1.00
Counties in metro. areas of less than 250,000 pop.	68,160	1,648	1.09	1.04	1.04	1.03	1.04	1.04	1.04	1.03	1.04	1.04	1.04	1.03	1.04
Urban pop. of 20,000 or more, adjacent to a metro. area	20,906	1,718	1.05	1.04	1.02	1.02	1.03	1.02	1.01	1.01	1.02	1.01	1.00	1.01	1.01
Urban pop. of 20,000 or more, not adjacent to a metro. area	25,466	1,678	1.07	0.95	0.94	0.96	0.96	0.97	0.96	0.98	0.97	0.97	0.96	0.98	0.97
Urban pop. of 2,500-19,999, adjacent to a metro. area	18,243	1,968	0.91	1.01	1.01	0.98	1.02	1.01	1.01	0.99	1.02	1.02	1.02	0.99	1.02
Urban pop. of 2,500-19,999, not adjacent to a metro. area	7,794	1,833	0.98	1.09	1.07	1.05	1.09	1.06	1.04	1.02	1.06	1.05	1.03	1.01	1.04
Completely rural (no places with a pop. of 2,500 or more) adjacent to a metro. area	1,547	2,193	0.82	1.03	1.05	1.00	1.03	1.02	1.03	0.98	1.01	1.03	1.05	0.99	1.02

(continued)

Table 5-7. Predictive Ratio for Demographic and Service-Related Subgroups, by Model (continued)

	N	Mean of actual expendi-ture ($)	Ratio of total vs. group mean	All costs				Costs top-coded at $50,000				Costs top-coded at $25,000			
				ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG
Completely rural (no places with a pop. of 2,500 or more) not adjacent to a metro. area	1,237	2,291	0.78	1.00	1.01	0.96	1.01	0.96	0.96	0.92	0.96	0.95	0.95	0.92	0.94
Mean Deviation Score*				5.0	5.1	**3.4**	4.7	4.6	4.7	**3.2**	4.2	4.6	4.7	**3.1**	4.2
Weighted Mean Deviation Score **				3.0	2.9	**1.5**	2.8	2.8	2.8	**1.4**	2.6	2.8	2.8	**1.5**	2.6

Comparisons of actual to predicted FY 2002 cost in the validation subset (N = 500,000) using risk weights developed on the fitting data (N = 1.8 million).

Bolded numbers have the best (lowest) deviation scores in each block, or have scores that are within 1.0 percent of the best score.

*For each column: $100 \times (\sum |1 - \text{predictive ratio}|) / \text{total number of rows (53)}$. A mean deviation score is the "typical" size of the percentage error that a model makes in predicting the costs of the groups portrayed in the rows of the table.

**For each column: $100 \times (\sum N_i \times |1 - \text{predictive ratio}|) / \sum N_i$ where N_i = the number of cases in the ith row.

between the predicted and actual costs of the groups studied. The models with the smallest deviation have performed the best; overall, the best performer is consistently the CRG model.

5.3 Comparing Models: Predictive Ratios for Disease Cohorts

Being able to accurately predict the costs for groups of people with high-cost or highly prevalent diseases is an important characteristic of a risk adjustment model. All patient populations vary, and models that do particularly well at predicting the diseases associated with aging may not do as well when applied to populations with a significant proportion of younger individuals, and vice versa. Therefore, we examined nine conditions/diseases that were deemed to be significant in the TRICARE Prime population.

We determined which conditions to examine by identifying disease cohorts that were either most prevalent, had the highest average cost, or had the highest total cost as described in the methods section. We then excluded transient clinical conditions because our focus was on exploring the ability to predict future costs. Since different manifestations of the same disease could be classified across more than one three-digit ICD-9 code, we used published medical literature to group related codes. From this list of medical conditions, in consultation with TMA, nine clinical conditions were chosen for analysis.

Our general approach was to require that an individual have a single claim within a broad definition of the disease (i.e., three-digit ICD-9 code) to be included in our analyses. It should be noted that the CRG risk adjustment methodology, which uses a different approach to defining disease groups (i.e., requiring that diagnoses be validated by multiple occurrences), may perform differently when compared to the other three risk adjustment methodologies (i.e., ACGs, CDPS, and DCGs) that use an approach that requires a single diagnosis. However, none of the latter referenced risk adjustment methods of disease classification coincide purely with the three digit ICD-9 designations used in this study. Thus, our analysis by disease category could partially reflect adherence to a set of disease identification methodological rules, and may not be a pure test of predictive power.

Table 5-8 shows predictive ratios (model predicted cost divided by actual cost) for each risk adjustment model, and the age-sex model, for each of nine important medical conditions. Note that the mean annual cost of the cohort of people with diabetes, for example, is not the cost of treating their diabetes—it is the cost of treating all their medical problems. The condition affecting the largest number (~26,000) of TRICARE Prime beneficiaries was Major Mental Health Disorders. Total annual costs for people with this condition were more than double the population average. Both ACG and DCG models did reasonably well in predicting the cost in this group, with predictive ratios of 0.97 and 0.98 respectively, among all costs. The CDPS was not quite as accurate at 0.90, nor was CRG at 0.81. The highest cost category was chronic renal failure, affecting only 467 individuals in the 0.5 million validation dataset, but costing over $25,000 per affected enrollee. Within these disease cohorts, top-coding strongly affects model performance. In predicting all costs, CRG performed worst at 0.75 and CDPS best at 0.87, when predicting costs top-coded at $25,000, CRG was still poorest, but at 0.88, and CDPS best at 0.99. In these nine disease-defined population cohorts, each examined for three outcomes, no single risk adjustment model consistently made the most accurate prediction. However, all four models always came far closer to predicting true costs than the age-sex model, which always underpredicted costs for groups of sick people. For example, the age-sex model predicted costs for people with chronic renal failure at

Table 5-8. Predictive Ratios for Disease-Based Cohorts, by Model

	N	Actual average cost 2002 ($)	All Costs					Costs top-coded at $50,000					Costs top-coded at $25,000				
			Age-Sex	ACG	CDPS	CRG	DCG	Age-Sex	ACG	CDPS	CRG	DCG	Age-Sex	ACG	CDPS	CRG	DCG
Total	500,000	1,796	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Diabetes Mellitus	12,112	6,513	0.52	0.97	0.97	0.97	0.97	0.54	0.98	0.98	0.98	0.98	0.57	0.98	0.99	0.98	0.98
Female breast cancer	1,282	7,880	0.47	1.07	0.91	1.05	1.00	0.49	1.06	0.86	1.03	0.94	0.51	1.03	0.85	1.00	0.91
Major Mental Health Disorders	26,201	4,342	0.51	0.98	0.90	0.81	0.97	0.52	0.99	0.91	0.81	0.98	0.53	0.99	0.90	0.81	0.98
Ischemic Heart Disease	5,643	8,089	0.45	1.02	0.94	0.94	1.02	0.47	1.01	0.93	0.93	1.02	0.50	1.01	0.94	0.93	1.01
Congestive Heart Failure	1,182	13,897	0.27	1.11	0.98	0.97	1.05	0.30	1.03	0.93	0.92	0.99	0.34	1.00	0.93	0.92	0.97
COPD	3,976	7,957	0.42	0.97	0.92	0.90	0.98	0.46	0.99	0.92	0.91	1.00	0.49	0.99	0.92	0.92	0.99
Asthma	21,752	3,184	0.54	0.99	0.99	0.86	0.97	0.55	1.00	0.98	0.86	0.98	0.56	1.01	0.97	0.86	0.99
Chronic Renal Failure	467	25,038	0.13	0.82	0.87	0.75	0.81	0.17	0.90	0.96	0.83	0.86	0.23	0.94	0.99	0.88	0.90
Osteoarthritis	9,369	5,547	0.58	0.86	0.93	0.84	0.94	0.58	0.85	0.93	0.83	0.93	0.58	0.84	0.92	0.82	0.92
Mean Deviation Score*			56.8	6.8	6.6	11.5	**4.8**	54.6	**4.3**	6.7	10.6	**4.0**	52.0	**3.4**	6.6	9.7	**4.1**
Weighted Mean Deviation Score**			48.8	**3.6**	5.8	13.5	**3.3**	47.3	**2.6**	5.7	13.1	**2.6**	46.0	**2.9**	6.2	12.9	**2.5**

Bolded numbers have the best (lowest) deviation scores in each block, or have scores that are within 1.0% of the best score.

*For each column: $100 \times (\sum |1 - \text{predictive ratio}|) /$ total number of rows (9). A mean deviation score is the "typical" size of the percentage error that a model makes in predicting the costs of the groups portrayed in the rows of the table.

**For each column: $100 \times (\sum N_i \times |1 - \text{predictive ratio}|) / \sum N_i$, where N_i = the number of cases in the ith row.

only 13 to 23 percent of their actual costs, and costs for people with osteoarthritis at only 58 percent of their actual costs. The DCG model performed at or near best for both summary measures and all three cost outcomes (with typical-sized errors in the range of 3 to 5 percent), while the ACG model was at or near best in 5 of the 6 tests. In marked contrast, the CRG model, which had predicted demographic-based subgroups most accurately, was least accurate in cost predictions for disease-based cohorts (with typical-sized errors ranging from 10 to 13 percent), although far more accurate than the age-sex model, where the typical amount of underprediction was approximately 50 percent.

5.4 Comparing Models: Prior Cost Groups

In addition to understanding how well these risk models fare when comparing by demographic characteristics and by medical condition groups, we sought to understand how well they performed in cases of low cost, high cost, and the vast range of intermediate costs. We sorted the half million members of the validation population by their FY 2001 costs. Many individuals had very low health care costs per year, or no cost. We selected the least expensive 20 percent for our lowest cost group. In FY 2002, their mean cost was $553 per year, less than a third of the average TRICARE Prime enrollee in our population (Table 5-9). The DCG model predictions were closest to actual in this group; for all costs the predictive ratio was 0.94, meaning it underpredicted the cost by 6 percent. The CRG model overpredicted by 20 percent, the ACG model by 33 percent, and the CDPS by 56 percent.

We then used the 30 percent of enrollees with next highest FY 2001 cost from our sorted list. These individuals had mean costs in FY 2002 of $937, still significantly lower than the average enrollee. Again, DCG had predictive ratios closest to 1.0, but by a lesser margin. Next, we tested a group of 30 percent of enrollees with the most “average” cost. Here, the model predictions were more accurate, with CDPS performing best for this group. Then, we tested progressively smaller groups of very high cost individuals; as these groups got more costly, the models started to underpredict costs. The DCG model predictions came closest to actual in all these categories, although not always by very large percentages of the actual amounts. It is worth remembering that a five-percent error in a low cost group is a much less money (per person) than a five-percent error in the higher cost categories, although far fewer people are in these categories. For example, if a model had a predictive ratio of 0.95 in both the lowest cost group (mean cost $553 among 100,000 people) and in the highest cost group (mean cost $15,384 among 5,000 people), it would underpay a total of $2.77 million in the low cost group and $3.85 million in the high cost group.

In making Figure 5-1, the half million validation population was sorted by FY 2001 actual costs and split into 50 groups, each containing two percent of the population. To highlight the features of the “central” part of the data, the lowest five groups (which cost nothing, or almost nothing in FY 2001) and the top two groups are not plotted. For each of the remaining 43 groups, four points are plotted—one for each of the four models—and all at the same height, since the vertical axis represents (all) actual cost in FY 2002. The horizontal position of each mark is determined by the predicted cost for the group under one of the four models. The solid line represents perfect accuracy, where the predicted and actual costs are equal. Figure 5-1 was based on models that predicted all costs vs. all actual FY 2002 costs (the extended results in Appendix F contain figures based on top-coded costs). Figure 5-1 reveals patterns similar to those in Table 5-9. At the ends of the spectrum, the DCG model is closest to perfect accuracy; but the CDPS

Table 5-9. Predictive Ratios for FY 2001 Cost-based Subgroups, by Model

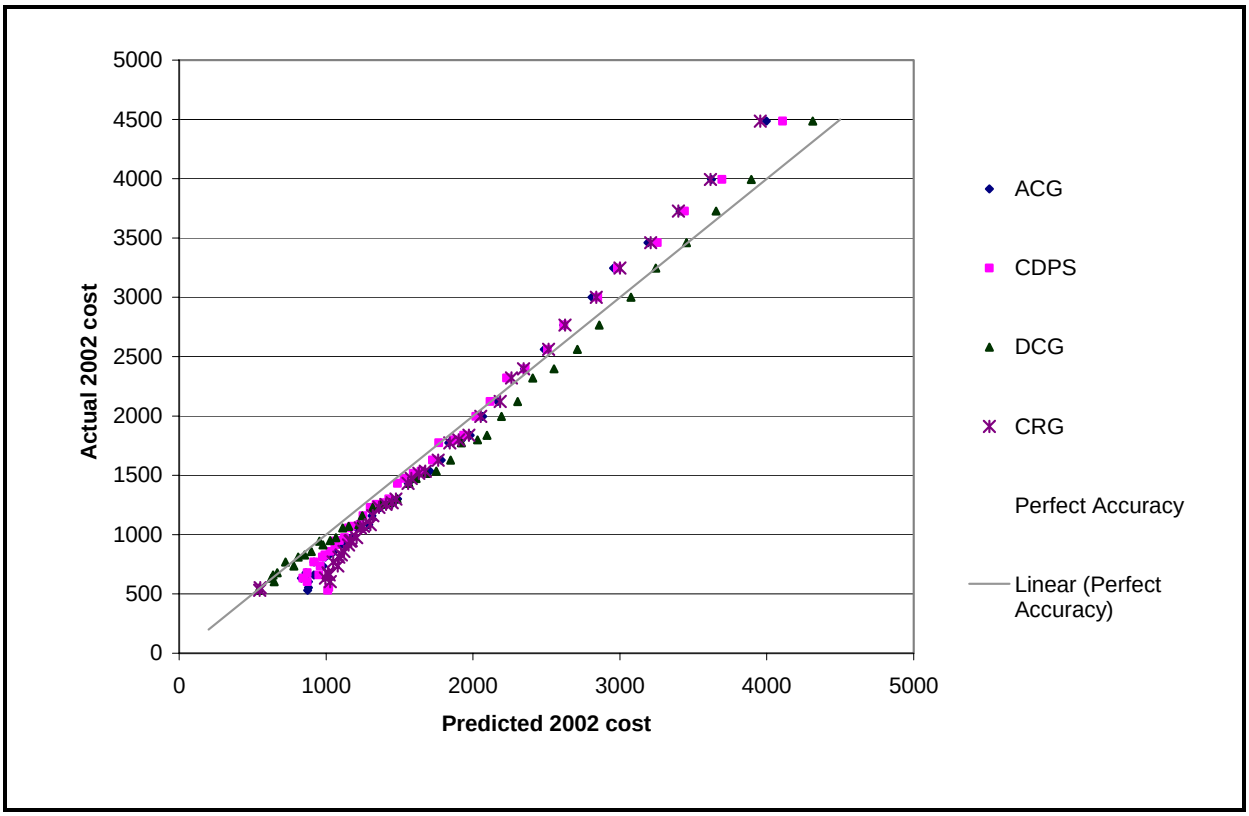
	N	Mean of actual FY 2002 expenditure ($)	All expenditures				Expenditures top-coded at $50,000				Expenditures top-coded at $25,000			
			ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG	ACG	CDPS	CRG	DCG
Total	500,000	1,796	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Lowest 20% of 2001 cost	100,000	553	1.33	1.56	1.20	0.94	1.35	1.65	1.19	0.95	1.33	1.6	1.19	0.93
Next highest 30% (21–50%)	150,000	938	1.19	1.15	1.25	1.05	1.21	1.19	1.26	1.09	1.20	1.17	1.26	1.07
Next highest 30% (51–80%)	150,000	1,710	1.06	1.03	1.06	1.10	1.06	1.01	1.04	1.10	1.06	1.02	1.05	1.10
Next highest 10% (81–90%)	50,000	3,007	0.94	0.95	0.94	1.02	0.91	0.90	0.92	0.99	0.92	0.92	0.93	1.01
Next highest 5% (91–95%)	25,000	3,959	0.91	0.92	0.90	0.97	0.86	0.86	0.88	0.93	0.88	0.88	0.89	0.94
Second highest 4% (96–99%)	20,000	5,957	0.83	0.83	0.82	0.89	0.78	0.79	0.79	0.84	0.80	0.80	0.80	0.86
Highest 1% of 2001 cost (100%)	5,000	15,384	0.70	0.65	0.71	0.75	0.76	0.71	0.77	0.79	0.74	0.68	0.75	0.77
Mean Deviation Score*			17.1	19.9	16.3	8.9	18.7	22.7	16.1	9.9	17.9	21.6	16.1	9.7
Weighted Mean Deviation Score**			16.1	18.5	15.4	6.7	17.8	21.8	15.3	8.0	16.9	20.2	15.4	7.7

Bolded numbers have the best (lowest) deviation scores in each block, or have scores that are within 1.0% of the best score.

*For each column: $100 \times (\sum |1 - \text{predictive ratio}|) / \text{total number of rows (7)}$. A mean deviation score is the "typical" size of the percentage error that a model makes in predicting the costs of the groups portrayed in the rows of the table.

**For each column: $100 \times (\sum N_i \times |1 - \text{predictive ratio}|)/500000$, since here $\sum N_i = 500{,}000$.

Figure 5-1. Actual versus Predicted FY 2002 Costs by 43 "Central" Two-Percentile Groups of FY 2001 Actual (All) Cost*



*To improve legibility, the lowest five cost groups (or 10% of the sample) and two highest cost groups (4% of the sample) are not shown. Actual costs in the two most expensive groups were $5,698 and $11,597. Predicted costs for these groups were ACG, $4,713 and $8,525; CDPS, $4,783 and $8,046; CRG, $4,683 and $8,489; and DCG, $5,047 and $9,072.

model is closest in the middle ranges. This pattern continues in the two highest cost groups, which are not shown on the figure. The ACG, CDPS, CRG, and DCG predictions for the first of these groups were $4,713, $4,783, $4,683 and $5,047, respectively, compared to an actual cost of $5,698; their respective predictions for the very highest group were $8,525, $8,046, $8,489, and $9,072, compared to an actual cost of $11,597.

5.5 Comparing Models: Grouped R-squared (R^2)

Another method to compare models is the *grouped* R^2. This computation is similar to the R^2 as described in the methods section, however, it produces a summary figure for how accurately the predicted means match actual means for costs within the subgroups of a partition of the validation sample. A partition is a division of the population into categories such that every person belongs to one and only one category. Note that the classification of people into disease-based groups, as shown in Table 5-8, is not a partition, since some people have none of these diseases and some have several.

Table 5-10 shows the results of a series of grouped R^2 calculations. The first partition of the population was into 50 groups each comprising two percent of the population, as sorted by FY 2001 total

Table 5-10. Grouped R^2 Values for Various Partitions of the Population

Grouping variable	# Groups	ACG	CDPS	CRG	DCG
Two-percentile partition based on prior year costs	50	0.92	0.91	0.92	**0.95**
Service	4	**0.74**	0.68	**0.74**	0.72
Service (excluding other)	3	**0.81**	0.75	0.59	0.64
Rank	6	0.23	0.38	0.36	**0.61**
Bencat	4	**0.99**	**0.99**	**0.99**	**0.99**
Catchment	2	–4.8	–2.5	**0.96**	–2.3
PCM Type	2	–27.7	–19.8	**0.98**	–19.7
Region	13	0.56	0.55	**0.74**	0.58
Urbanicity	11	0.67	0.71	**0.78**	0.60

Numbers are bolded when they are best (largest), or within 0.01 of best for the indicated partition.

cost per person. (Basically, the same information that was used to produce Figure 5-1.) When the summary statistics for the R^2 calculations over these 50 groups were compared, the DCG model performed best, indicating that the model explains 95 percent of variation in average costs among these 50 groups. When the service type (Army, Navy/Marine, Air Force, and Other) was compared, none of the groups did so well, with the ACG and CRG models performing best. Note that the Other group comprises only two percent of the sample and has generally lower costs (as shown earlier in section A). When this calculation was performed and the Other category was excluded (therefore three partitions), the grouped R^2 performance order changed: CRG and DCG models performed worse, and ACG and CDPS models did better. The grouping based on rank saw a dramatic separation of grouped R^2 results—the DCG model performed significantly better than the other two. Differences by enrollee category were small and not detectably different by model. The CRG model performed best by regional grouping and urbanicity grouping.

The results by catchment area and primary care manager (PCM) type are unusual. A negative value for R^2 means that the model's predictions fit the data less well than simply guessing the mean (here $1,796) for each observation; here three of the four models have negative R^2 values. For example, the "within catchment" group spent more than those outside (as shown in Table 5-7), yet were considered healthier (had lower expected costs) than average by three of the risk adjustment models. Similarly, these three models predicted that individuals with military PCMs would cost less than they actually did; that is, they appeared healthier than their actual costs suggest. (Detailed grouped R^2 calculations are in Appendix Tables F4 and F5.) It is possible that civilian doctors list diagnoses more comprehensively than their military counterparts, thereby generating higher illness scores (and the appearance that their patients are sicker); or, civilian doctors may not be as accessible as military doctors (i.e., there may be undercare in the civilian sector or overcare in the military one). In contrast, the CRG model predicted costs for each of these groups very near the actual amount.

6

Conclusions/Discussion

Understanding the health of a population and predicting its health care costs is a complex task that requires many components. One important piece is a measurement of the overall illness burden. In 1984, no methods existed to measure this burden, but now there are several mature systems that extract illness burden profiles from computerized encounter records. The purpose of this project was to conduct side-by-side testing of four risk adjustment methods (ACG, CDPS, CRG, and DCG) using TRICARE Prime data. In addition, we examined the quality of the DoD data, this report comments on its credibility for use in monitoring and managing TRICARE Prime.

We obtained administrative data on enrollment and claims (including diagnoses) for all TRICARE Prime enrollees in fiscal years 2001 and 2002. Our study sample included 2.3 million TRICARE Prime enrollees under the age of 65 and continuously enrolled during 24 months of FY 2001 through FY 2002 and residing in the continental United States, Alaska, or Hawaii. Our focus was on how well each risk adjustment model uses demographic and diagnosis information in one year (FY 2001) to predict health care costs in the following year (FY 2002). The key step in obtaining model predictions is estimating a statistical relationship between FY 2002 costs and the model-specific determining factors (demographic and diagnosis-related condition categories) using fitting sample data.

Three risk adjustment models (ACGs, CDPS, and DCGs) were used separately with FY 2001 diagnoses (ICD-9-CM codes) to obtain health-based predicted expenditures for FY 2002 for the validation sample. We also tested a fourth model, CRGs, that in this study chose to take advantage of additional information such as dates of service, place of service, and procedure codes to make predictions. We examined the performance of all four models on a wide range of commonly used statistical measures of model accuracy when predicting next year's cost for a population from this year's data.

We found that overall each of the models performs well. When considering global measures of predictive accuracy, the DCG model slightly outperformed the others, although all of the models were rather similar on most measures, and all did much better than a simple age- and sex-adjusted prediction. In comparing predictive ratios for groups defined by demographic and service characteristics, the CRG model predictions were most accurate, although again the differences between models were small. Predictive ratios were produced for nine different diseases/conditions, ranging from asthma and osteoarthritis to chronic renal failure and congestive heart failure. All the models predicted costs that were reasonably close to actual costs for most conditions, with the exception of chronic renal failure, and all came much closer than a simple age-sex model. When considering how well these models predict high-cost, low-cost, and midrange-cost cases the DCG model outperformed the others overall, and in most specific cost areas. Only when considering the high middle range of costs (51st to 80th percentile, averaging $1,700) was the DCG model outperformed by the CDPS model. Finally, in considering the grouped R^2 measurements, the results varied. ACGs did best at predicting differences in costs by service;

DCGs best predicted costs by rank and by prior year cost groups; and CRGs did best by region and urbanicity.

The grouped R^2 measurements revealed that the three diagnosis-based models (ACG, CDPS, and DCG) agree with each other and differ from CRGs in how they "view" the expected versus actual costs of people who either reside near a military facility (within an MTF catchment area) or who have a military (non-civilian) primary care manager.

There are some important data limitations. One is dual eligibility. As we have included allowed amounts for care from private providers, the presence of secondary coverage does not affect our analyses. Those 65 years old or older who are Medicare eligible would not be eligible for TRICARE Prime. However, dual eligibility in TRICARE Prime and the Veteran's Health Administration cannot be ruled out, and the TRICARE Prime costs may be an underestimate. As there is no systematic documentation of such dual care, its extent is unknown. The second limitation is the difficulty of assessing pharmaceutical costs. Costs of most pharmaceuticals are embedded in the ambulatory care cost files, so that we cannot readily distinguish among pharmaceutical and nonpharmaceutical health care utilization and costs. Finally, we are limited by difficulties identifying laboratory-only claims. To characterize illness burden more accurately, our original intention (as recommended by most software vendors) was to exclude diagnoses that appeared only in laboratory records. This is because such diagnoses may be rule outs—for example, a heart attack diagnosis is written on a test done to see if there was a heart attack—although we do not know whether a heart attack was confirmed. However, since such distinctions could be made only in records from the civilian sector (HCSR-N), for consistency we included diagnoses from all laboratory records. This should not affect the model comparisons, however, because all models were fit to the same data.

It is worth noting that between FY 2001 and FY 2002, costs at sites of care were shifting fairly substantially in the TRICARE Prime population. Among our 2.3 million total study population, all of whom were continuously enrolled for these 24 months, the proportion of patients receiving care in the military sector dropped by only 4 percent, although costs for this sector dropped slightly less. At the same time, the proportion of people receiving care in the civilian sector increased by almost 10 percent, but costs in this sector increased 22 percent, with the mean individual cost increasing 11 percent. There were more modest increases in those receiving care in both sectors (1.3 percent) for whom mean individual costs increased 8 percent, and those receiving no recorded care (3.3 percent) in these years. Overall, inpatient costs (aggregated between civilian and military) increased 10 percent, and ambulatory costs increased 4.5 percent.

Despite this changing environment, the risk adjustment models all performed well on virtually all tests performed. There is no one perfect test for model fit, and most differences in model performance were modest. Our tests, while extensive and performed on a validation sample of 500,000, were only conducted on a single dataset. Decisions about which risk adjustment model to use in future work on developing a prospective payment system or for a myriad of other possible uses—i.e., high-risk case identification, profiling, etc.—will depend on the relative importance of various features of the models. TMA may choose to use different models for different purposes. Our empirical findings focus on differences in model accuracy. Other factors are important. Specifically, drawing on the framework

recommended by the Society of Actuaries (Cumming et al., 2002), additional factors worthy of consideration by the TMA are

- ease of use of the software.
- availability of standard reports.
- cost of the software.
- access to data of sufficient quality.
- the underlying logic or perspective of a model that makes it best for a specific application.
- whether the model provides both useful clinical and financial information.
- whether the model will be used mostly for payment to providers/plans or for underwriting/rating/case management.
- whether the model is currently in use in the market or organization.
- the susceptibility of the model to gaming or upcoding.

A qualitative evaluation of these other considerations may be a reasonable next step for the TMA.

The overall message from this work is positive. Despite some obvious deficiencies in the TRICARE Prime data, and the fact that none of these models was developed for TRICARE's population, these off-the-shelf methods predict next year's costs as well or better here as they have in the populations on which they were developed. These data are clearly adequate to support credible management activities for TRICARE Prime.

7

References

Ash, A.S., & Byrne-Logan, S. (1998). How Well Do Models Work? Predicting Health Care Costs. In *Proceedings of the Section on Statistics in Epidemiology,* American Statistical Association.

Ash, A.S., Ellis, R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin,H., Iezzoni, L.I., MacKay, E., & Yu, W. (2000). Using Diagnoses to Describe Populations and Predict Costs. *Health Care Financing Review*, 21(3), pp. 7-28.

Ash, A.S., Porell, F., Gruenberg, L., et al. (1986). *An Analysis of Alternative AAPCC Models Using Data From the Continuous Medicare History Sample.* Waltham, MA: HCFA University Health Policy Research Consortium, Brandeis and Boston University.

Cumming, R.B., Knutson, D, Cameron, B.A., & Derrick., B. (2002). *A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations*. Society of Actuaries.

Ellis, R.P., Pope, G.C., Iezzoni, L., Ayanian, J.Z., Bates, D.W., Burstin, H., & Ash, A.S. (1996 Spring). Diagnosis-Based Risk Adjustment for Medicare Capitation Payments. *Health Care Financing Review*, 17(3), pp. 101-28.

Ellis, RP, & Ash, A. (1995-96 Winter). Refinements to the Diagnostic Cost Group (DCG) Model. *Inquiry*, 32(4), pp. 418-29.

Kronick, R., & Dreyfus, T. (1997). *The Challenge of Risk Adjustment for People with Disabilities: Health-Based Payment for Medicaid Programs*. Princeton, NJ: Center for Health Care Strategies.

Kronick, R., Dreyfus, T., Lee, L., & Zhou, Z. (1996). Diagnostic Risk Adjustment for Medicaid: The Disability Payment System. *Health Care Financing Review*, 17(3), pp. 7-33.

Kronick, R., Gilmer, T., Dreyfus, T., & Lee, L. (2000 Spring). Improving Health-Based Payment for Medicaid Beneficiaries: CDPS. *Health Care Financing Review*, 21(3), pp. 29-64.

Kronick, R., Zhou, Z., & Dreyfus, T. (1995). Making Risk Adjustment Work for Everyone. *Inquiry*, 32(1), pp. 41-55.

Muldoon, J.H., Neff, J.M., & Gay, J.C. (1997). Profiling the Health Service Needs of Populations Using Diagnosis-Based Classification Systems. *Journal of Ambulatory Care Management*, 20(3), pp. 1-18.

Pope, G.C., Ellis, R.P., Ash, A.S., Liu, C.F., Ayanian, J.Z., Bates, D.W., Burstin, H., Iezzoni, L.I., & Ingber, M.J. (2000 Spring). Principal Inpatient Diagnostic Cost Group Model for Medicare Risk Adjustment. *Health Care Financing Review*, 21(3), pp. 93-118.

Starfield, B., Weiner, J., Mumford, L., & Steinwachs, D. (1991). Ambulatory Care Groups: A Categorization of Diagnoses for Research and Management. *Health Services Research*, 26(1), pp. 53-74.

Stoloff, P.H., Lurie, P.M., Goldberg, L., & Almendarez, M. (2001). *Evaluation of the TRICARE Program: FY2000 Report to Congress*, U.S. Department of Defense.

TRICARE Management Activity (TMA) (2000). *Patient Level Cost Allocation (PLCA) Detailed Description with Examples*. TRICARE Management Activity, Health Program Analysis and Evaluation.

TRICARE Management Activity (TMA) (2002 August). *2002 Chart Book of Statistics*, TRICARE Management Activity, Health Program Analysis and Evaluation, Data Quality and Functional Proponency Office. Available at http://199.211.83.250/Reports/Chartbook/

U.S. Department of Defense (2005). Military Health System Web site. Accessed November 28, 2005, at http://www.tricare.osd.mil

Weiner, J, Starfield, B., Steinwachs, D. & Mumford, L. (1991). Development and Application of a Population-Oriented Measure of Ambulatory Care Case-Mix. *Medical Care*, 29(5), pp. 452-472.

Weiner, J., Dobson, A., Maxwell, A.S., Coleman, K., Starfield, B., & Anderson, G. (1996). Risk-Adjusted Medicare Capitation Rates Using Ambulatory and Inpatient Diagnoses. *Health Care Financing Review*, 17(3), pp. 77-99.

Weiner, J.P., Starfield, B.H., & Lieberman, R.N. (1992). Johns Hopkins Ambulatory Care Groups: A Case-Mix System for UR, QA, and Capitation Adjustment. *HMO Practice*, 6(1), pp 13-19.

Appendix A: Bibliographical Materials

Appendix A
Bibliographical Materials

1. A literature review performed by RTI and BU for the Department of Defense

2. How Well do Models Work? Predicting Health Care Costs" by Arlene Ash and Susan Byrne-Logan. In *Proceedings of the Section on Statistics in Epidemiology*, American Statistical Association, 1998.

A Comparative Survey of Diagnosis-Based Risk Adjustment Models

This concise survey covers only diagnosis-based risk adjustment models, comparing their intent, logic, and construction as well as their performance in side-by-side comparisons. Few studies have compared several models using common methods on a common data set. Here we summarize findings from five papers for six models: Adjusted Clinical Groups (ACGs), Diagnostic Cost Groups (DCGs), Chronic Disease and Disability Payment System (CDPS), Global Risk Adjustment Model (GRAM) and Clinical Risk Groups (CRGs).

In cases where multiple model versions were run with limited diagnoses (e.g., diagnoses from inpatient events only) or all diagnoses (e.g., from ambulatory as well as inpatient encounters) we present the all-encounter models only. Also, we do not report findings relating to models that require the use of procedures and/or pharmacy data.

When choosing a model, non-quantitative issues, as well as quantitative measures are important. These include ease of use, cost and availability, options for handling imperfect data, and the extent to which the model developers continue to regularly update the software (for example, incorporating changes in ICD-9 codes), and the types of outputs that are readily available without special programming. For example, we previously used the Clinical Classifications for Health Policy Research (CCHPR) model by Elixhauser and colleagues to predict 1-year mortality following a hospital admission for acute myocardial infarction (Elixhauser 1996). While it ultimately performed well, the free code that could be downloaded from the Web was extremely difficult to implement and ultimately could only be run on a subset of our large database.

Table 1 provides a brief review of the most current versions of these five models, followed by a concise summary of the comparison papers reviewed. In this section, we also briefly define the statistical scores used. More detail may be found about each study in the appendix. Finally, we summarize the findings.

Although these reviews were all published in the last one to four years, some of the software versions studied are as much as 10 years old now. Most of these models have continued to evolve and therefore these results are applicable, but with reservation. What has been tested in some of these papers are earlier versions which are now more refined, for example the version of the DCG model reported in the Madden paper was quite early (1995), and a more intermediate version was tested by Cumming. A current version of the DCG software now uses 184 DxGroups rather than 112. The Adjusted Clinical Groups (formerly, Ambulatory Care Groups) now have 93, rather than 81 clinical categories, including “pregnancy” and “low birthweight neonates.”

Table 1. Overview of the Risk Adjustment Models

Model	**Adjusted Clinical Group (ACG)**	**Diagnostic Cost Group (DCG)**	**Chronic Disease and Disability Payment System (CDPS)**
Developers	Jonathan Weiner and Barbara Starfield, Johns Hopkins University	Arlene Ash and Randall Ellis, Boston University	Richard Kronick and Tony Dreyfus, UC–San Diego
Accessibility	Commercially available Computer Science Corporation, Inc	Commercially available DxCG, Inc, Boston	Available "for essentially a no cost license by contacting the developers at UC–SD" [(Cumming, Knutson et al. 2002), pp. 49]
Classification method	All ICD-9 inpatient and ambulatory diagnoses are assigned to 32 Adjusted Diagnosis Groups (ADGs), based on clinical criteria of severity, duration, diagnostic certainty, etiology and specialty care. ADGs are combined with age and gender to produce 93 mutually exclusive Adjusted Cost Groups (ACGs) based on clusters of ADGs using a branching algorithm. An individual may have many ADGs, but only one ACG. As the ACGs incorporate age/gender characterisics these are not separately included when modeled. No standard risk weights are provided. (Cumming, Knutson et al. 2002; ACG 2004)	Diagnoses are mapped to DxGroup categories which are combined into 136 Hierarchical Condition Categories (HCCs). Categories are arranged in hierarchies of diseases of similar type, generally body system. Within hierarchies, only the weight of the highest cost category is used to assess risk. Individuals are assigned to one HCC and one of 32 age/gender categories to predict costs. (Cumming, Knutson et al. 2002)	Individuals are assigned to one or more of 67 medical condition categories using diagnoses, and are assigned to one of 16 age/gender categories. "Diagnoses are initially assigned to chronic condition categories… [which] are arranged into hierarchies. Only the highest cost category in a disease hierarchy is used to produce an individual's total risk score...[which] is computed by adding the weights for the age and gender category and any medical categories across the hierarchies. Within the hierarchies only the highest cost category identified is used to assess risk." (Cumming, Knutson et al. 2002)

Table 1. Overview of the Risk Adjustment Models (continued)

Model	**Global Risk Adjustment Model (GRAM)**	**Clinical Risk Groups (CRG)**
Developers	Mark Hornbrook, Richard T. Meenan, et al. Center for Health Research, Kaiser Permanente, Northwest Division	James Gay, John Muldoon, John Neff, and 3M Health Information Systems
Accessibility		Commercially available
Classification method	"GRAM classifies patients based on ICD-9-CM major diagnosis categories and a separate 'clinical resource intensity (CRI)' dimension." Eligible diagnoses are classified into "18 major diagnosis groups, which were subdivided into 66 subgroups based on dominant diseases, body systems, and etiology, and farther subdivided into 130 combinations of diagnosis and CRI" (Meenan, Goodman et al. 2003)	This system was primarily designed to identify people with chronic conditions but has multiple uses. It uses both diagnostic and procedure data. Diagnosis codes are assigned to a category (acute/chronic) and body system, procedure codes to a procedure category. Individuals are put into one of nine core health status groups then into specific risk adjustment categories (269), and severity levels (2–6 for chronic conditions, 2 for acute). There are 1,061 possible cells. The software will aggregate these into smaller groups of 380 (body system), 146 (super body system), 79 (core status with separate mental & physical health), 37 (core status), 29 (core status with rolled up catestrophic, metastatic malignancy & chronic triplet). (Madden, Mackay et al. 2001; Shenkman and Breiner 2001)

(Cumming, Knutson et al. 2002)

This study performs an "independent comparison of currently available risk adjusters" (pp. 2). The authors compared pharmacy claim-based risk adjusters as well as those that are diagnosis based. The diagnosis-based models they compared are ACG, DCG, CDPS and ERG. The ERG model requires the use of pharmacy drug code data as well as diagnosis and procedure codes—therefore the results for this model are not presented here.

The data for this study are claim and enrollment records from an employer covered population of 749,145 managed care members in 1998 and 1999. Actual payments to health care providers were used. The study uses *offered* risk weights that were provided with the model software as well as *recalibrated* risk weights that were estimated using the study data. As the ACG model does not come with offered weights, only the recalibrated weights were used for this case.

This study used four methods of comparing the models; individual R^2, Cumming's Prediction Measure (CPM), mean absolute prediction error (MAPE), and a group predictive ratio. The individual R^2 is a standard statistical measurement of model performance. It indicates the proportion of variation in the dependent variable (cost) explained by the model, ranging from a theoretical low of zero if there is no relationship to a high of 1 if all variation is explained. It weights large errors heavily and therefore may be driven more by catastrophic medical events with large claim amounts. The Cumming's Prediction Measure (CPM) is a standardized measure similar to R^2, and gives equal weight to small and large errors as it uses the absolute value of the prediction errors rather than the square of them. It has values ranging from 0 to 1, with 1 indicating perfect model fit. It gives the proportion of the sum of absolute deviations from mean in individual costs that is explained by the risk model. The Mean Absolute Prediction Error (MAPE) is also reported, as the CPM is a simple linear function of the MAPE and therefore is reporting the same result we have not included it here. The group predictive ratio compares the predicted cost for a sub-group of individuals to their actual cost, it will be closer to 1 if the actual and predicted costs are similar. The groups used here are based on either one of five medical conditions; breast cancer, congestive heart failure (CHF), asthma, depression and HIV or based on quintiles of total medical costs in 1999 (the prediction year). When based on medical conditions the performance of this measure will vary with the medical conditions chosen and will do better if these groups are similar to groups created by the risk adjustment method being evaluated.

Table 2 shows the results of the individual measures and the group predictive measures for these four risk adjustment models. It is worth noting that for the group predictive measure by cost quintile each model overpredicts the people with below average costs and underpredicts those with above average costs. The ratios are notably well above one for all models in predicting the group with the lowest quintile of cost.

Table 2. Comparison of the Predictive Accuracy of Risk Adjustment Models (Cumming, Knutson et al. 2002)

Type of model run	***Predictive measure***	***ACG***	***CDPS***	***DCG***
Prospective model—offered weights; claims truncated at $50K	R^2	NA	0.134	**0.195**
	CPM	NA	0.127	**0.172**
Prospective model—recalibrated weights; claims truncated at $50K	R^2	0.172	0.208	**0.224**
	CPM	0.179	0.190	**0.208**
Concurrent model—recalibrated weights; claims truncated at $50K	R^2	0.429	0.440	**0.564**
	CPM	0.381	0.343	**0.419**
Prospective model—recalibrated weights; untruncated claims	R^2	0.099	0.149	**0.154**
	CPM	0.167	0.178	**0.190**
Results of grouped analyses Prospective models, recalibrated weights, untruncated claims				
Grouped by **1998** medical conditions				
Breast cancer	Predictive Ratio	0.81	0.84	**0.92**
CHF	Predictive Ratio	0.51	0.79	**0.85**
Asthma	Predictive Ratio	**0.96**	0.95	**0.96**
Depression	Predictive Ratio	0.85	0.91	**0.97**
HIV	Predictive Ratio	0.34	**0.95**	0.91
Grouped by **1999** medical conditions				
Breast cancer	Predictive Ratio	0.52	0.54	**0.60**
CHF	Predictive Ratio	0.27	0.39	**0.42**
Asthma	Predictive Ratio	**0.72**	0.69	0.71
Depression	Predictive Ratio	0.65	0.66	**0.69**
HIV	Predictive Ratio	0.26	**0.59**	0.56
Grouped by 1999 claim dollar quintiles				
1st quintile	Predictive Ratio	92.16	98.16	**80.26**
2nd quintile	Predictive Ratio	6.92	6.38	**6.04**
3rd quintile	Predictive Ratio	3.10	**2.91**	2.94
4th quintile	Predictive Ratio	1.73	**1.68**	1.73
5th quintile	Predictive Ratio	0.48	0.51	**0.53**

(AdvanceMed 2003)

This study evaluated five commercially available risk adjustment models in their applicability for Military Health System (MHS) data. Four of the five groupers were variations on the DxCG model (all encounter, pharmacy, inpatient + pharmacy and inpatient). We have limited the review to comparing the all encounter versions of ACG and DCG models. The groupers were evaluated in their ability to predict future health care costs among TRICARE Prime enrollees—overall as well as selected subpopulations.

The study included a sample of 200,000 enrollees selected from TRICARE Prime/Senior Prime in three regions (Southwest, Southern California, Northwest) enrolled for an aggregate of at least 9 months during FY00 and FY01, including active duty, retired personnel, and their dependents.

Although much of the study focuses on how the different models explain risk and expenditure differences by regions, we find limited direct tests of model comparisons. More often the focus is on qualitative differences across models in how risk differences are characterized.

The most useful information on model comparison of predictive accuracy is in Table 22. Limiting only to medical dollars (excluding pharmacy costs due to "inconsistencies in the financial data" (pp. 47)), this compares the predicted FY2001 expenditures by different models with the actual expenditures. This is done by the three regions. While both the all encounter models (ACG and DCG) predicted more than 86% of the actual costs for each region they differed by 2% to 4% (with DCG predictions being better in two of the three regions).

Table 3. Comparison of Predicted and Actual Expenditure/Enrollee for FY2001 by Region

	Region 6		Region 9		Region 11	
	Predicted	% Actual ($1,967)	Predicted	% Actual ($1,549)	Predicted	% Actual ($1,680)
ACG	$1,734	88.2%	$1,340	86.5%	$1,523	90.1%
DCG	$1,703	86.6%	$1,372	88.6%	$1,577	93.9%

The study reports model differences in relative *current* risk scores across regions and age groups. Both ACG and DCG models have the very similar current risk scores by regions, those by age groups have sizable differences. But the authors conclude: "All 5 grouping methods yielded consistent, nearly identical results for each population studied." (pp. 26). One curious difference is that while *current* and *prospective* risk scores are calculated for all four DCG models, only *current* risk scores are calculated for the ACG model, with no obvious explanation offered, besides the statement "ACG developers maintain that a 'current' risk score is accurate predictor of the next 6–12 months" (pp. 26).

The section titled "Predictive Accuracy—Risk Scores" purports to compare ACG and DCG models in their accuracy in prospective risk prediction (i.e., using one year demographics and diagnoses to predict following year health care utilization). This is actually limited to a comparison of diabetes cases—that is, "the ability of DCG and ACG All Encounter methods to identify true positives and true negatives among individual study group members with diabetes mellitus" (pp. 34). The study reports that DCG prediction accuracy was 62% and that of ACG was 68%, with the qualification that "the different levels of accuracy are unlikely to be statistically significant" (pp. 35).

(Ash and Byrne-Logan 1998)

The study is focused on criteria used for comparatively assessing cost predictions from different risk adjustment models. Arguing that regression-based R^2 criterion is often indeterminate, this study suggests measures focused on group-level predictability rather than an individual-level focus on R^2. For a majority of the population, health costs are zero or minimal in any given year and most of these are a result of accidents or acute conditions with little predictive content for later years; consequently overall R^2 for all models are quite low. Therefore, comparative assessment of prospective costs from different models is better conducted by focusing on subpopulations for which diagnostic information has a predictive element—that is, diagnoses in one year signal higher likelihood of utilization in subsequent years.

In this study the models compared are ACG and DCG all encounter models. A model using costs from the previous year, or an indicator of no cost, was also included and is reported here simply for comparison.

Models are fitted using a data set of 1.4 million persons who were enrolled during 1992 and 1993 in various commercially insured, under-age-65, nationally disbursed plans. The fitted models are applied to a validating data set of 192,000 Massachusetts State employees and results from this are reported in model comparisons.

Table 4. Comparison Results (Ash & Byrne-Logan, 1998)

	ACG	DCG
Individual R^2	.07	.09
Grouped R^2		
Partitioned by DCG-Predicted categories	.92	.97
Partitioned by ACG-Predicted categories	.69	.79
Partitioned as 50 2-percentile groups of year one costs	.79	.83

(Meenan, Goodman et al. 2003)

This study compares risk adjustment models in predicting high-cost healthcare expenditures at the individual level for a commercial multi-HMO population. This was same data set used in the development of the GRAM model. GRAM was evaluated against DCG, ACG, RxRisk and a prior cost model—only the first three are reported here. The authors justify the focus on high-cost usage by the skewness of the distribution of total costs across the population. The top 0.5% of the study sample by annual medical expenditures accounted for 20.1% of the total population costs, while the top 1% accounted for 28.7% of the total costs.

The authors measured the fit of the models using the area under the receiver operating characteristic curve (AUC) and using the proportion of high cost dollars accurately predicted. These measures were applied for the whole sample as well as specific demographic and diagnostic subgroups. The following tables apply for the whole sample.

Table 5a. Estimated Area under Receiver Operating Characteristic Curve (AUC)

	1% high-cost users threshold	*0.5% high-cost users threshold*
ACG	0.83	0.84
DCG	0.85	0.86
GRAM	0.85	0.85

Table 5b. Proportion (%) of Actual High-Cost Dollars Correctly Predicted

	1% high-cost users threshold	*0.5% high-cost users threshold*
ACG	0.21	0.11
DCG	0.26	0.14
GRAM	0.26	0.14

Comparison by subgroups yielded somewhat mixed results in terms of proportion of actual high-cost dollars predicted. As shown in Table 5a, the AUC results were found to be similar across all models. (Note that the results quoted are as reported in the study.) In the subgroup of Medicaid enrollees, the prior-expense model captured 45% of total high-cost dollars, while ACGs captured 43%. (No result was reported for DCG or GRAM.) In subjects age 64 and older, “ACGs, DCGs and GRAM performed equally well”. Among children under 13, the GRAM performance (28%) was “slightly higher than the DCG and prior cost models. ACGs and RxRisk performed worst (15%).”

When reporting on analyses of important diagnostic subgroups of asthma, diabetes and depression subgroup results are not provided; however, the authors provide a narrative summary of findings, “In terms of correctly predicting high-cost dollars, DCGs consistently captured the highest proportion across diseases. The performance of the other models varied considerably with RxRisk least accurate among enrollees with asthma or depression and ACGs least accurate among enrollees with diabetes.”

(Madden, Mackay et al. 2001)

This study compared risk adjustment models in two populations; Washington State Medicaid SSI enrollees (between 1994–1995 and 1992–1993) and Washington State Medicaid non-SSI enrollees (1992–1993). They examined two versions (3 and 4) of the ACG system, CRG, CDPS, and two versions of the DCG system. The model which this report calls HCC is most similar to what is now generally called the DCG/HCC model, or, elsewhere in this report, simply DCG. In this summary, we refer to results only from the later ACG version (4) reported on in this report and the HCC version of the DCG software.

The authors reported overall goodness-of-fit results, using an R^2 for both populations. Two statistical models were used, an ordinary least square and a two part GLM model; typically, the GLM model performed best. In the 1994-5 SSI population, the R^2 for the GLM model was best for the DCG (0.45), intermediate for CDPS (0.36) and ACG (0.35) and worst for CRG (0.31). In the 1992-93 non-SSI population, the ACG model performed best (0.73), then CRG (0.67), followed by CDPS (0.52) and DCG (0.37). In addition the authors report prediction ratios, using actual expense deciles and selected sub-groups, which are reported in Table 6.

Table 6. Prediction Ratio Split Sample GLM Model

	1994-95 SSI population				1992-93 non SSI population			
Actual Expense Decile	CDPS	ACG	DCG	CRG	CDPS	ACG	DCG	CRG
1	22.2	21.6	22.7	20.0	15.2	18.9	15.3	14.5
2	8.9	9.5	9.1	8.3	6.2	7.3	6.6	6.0
3	5.5	6.1	5.5	5.3	3.8	4.3	4.0	3.7
4	3.8	4.3	3.8	3.8	2.7	3.0	2.8	2.6
5	2.8	3.0	2.8	2.8	2.0	2.3	2.1	2.0
6	2.1	2.2	2.0	2.1	1.6	1.8	1.6	1.6
7	1.5	1.5	1.4	1.4	1.3	1.4	1.2	1.3
8	0.9	0.9	0.9	0.9	0.9	1.0	0.8	0.9
9	0.4	0.3	0.3	0.4	0.4	0.3	0.3	0.3
Sub-groups								
Asthma	0.89	0.95	0.97	0.84	1.44	0.94	1.08	0.89
Congestive Heart failure	0.96	0.68	0.90	0.98	2.05	0.81	2.61	1.16
Hypertension	0.93	0.94	0.89	0.98	1.24	1.01	0.95	0.28
Diabetes	1.01	0.85	0.99	1.02	1.27	0.77	0.94	1.03
HIV	1.00	0.65	1.16	0.89	1.36	1.01	1.34	1.99
Stroke	0.94	1.10	1.01	1.02	0.82	0.49	0.79	0.69
Cardiac arrest/shock	0.82	0.88	0.86	0.89	1.92	0.57	0.94	1.08
Septicemia	0.78	0.69	0.88	0.77	1.13	0.82	1.11	0.59
Major pneumonia	1.00	0.83	0.91	0.87	1.96	0.76	1.03	0.79
Kidney infection	0.96	0.81	0.81	0.79	1.43	0.81	0.71	0.86

Summary

Each statistical measure used to summarize model performance has strengths and limitations. The most common method reported in these papers was an individual R^2. When an R^2 is reported the model with the smallest individual R^2 yields predictions whose averages in groups are closest to actual outcomes. This measure, however, may be overly sensitive to a few cases with large prediction errors. As the numbers and nature of large outliers vary from dataset to dataset, R^2 measures may also be unstable. Other measures were reported by some papers including the Cumming Prediction Measure (CPM) and the area under the receiver-operating curve (AUC). Group measures give a richer sense of how models will perform for specific tasks. If the task at hand matches the method of the grouped test reported then these scores may be the most useful for deciding among the various risk adjustment models. Ash and Byrne Logan point out that other groups that might be useful to consider, depending on the anticipated use of the model, could be whether or not there was a hospital stay in the prior year, the number of days of an inpatient stay in a prior year, or some constructed measure of "frailty." "Each binning [method of grouping] provides additional insight into the kinds of people who are well or poorly priced by each candidate model" (pp. 8).

In Cumming, Knutson et al., the concurrent models outperform prospective models (by a lot) in all cases, and the DRG model outperform ACGs and CDPS in most but not all tests. When grouped by 1998 medical conditions, the DCGs perform best for CHF, breast cancer and depression, ACGs and DCGs are equal for asthma, and CDPS does best for HIV. ACGs do as well as DCGs for asthma costs, but do much worse than the other two for HIV and CHF—perhaps reflecting that the ACGs use broader, less clinically specific, categories. In general, all the methods do less well predicting costs for the 1999 medical conditions as compared to 1998 medical conditions: DCGs perform best for the same three conditions (CHF, breast cancer, depression), the ACG model wins asthma and again DCPS does best for HIV. Finally, when grouped by quintiles of 1999 cost, the DCG method outperforms the others in the lowest and highest quintiles (1, 2 and 5) with ACGs doing the best in the mid-range. All the models have the same pattern of over-predicting low cost cases and under-predicting high cost cases. In the highest 40% of costs, the differences between models are not great. All of the methods tested by Cumming, Knutson et al. showed improvement in performance when risk weights were recalibrated. As expected, truncating the claims at $50,000 substantially improves the R^2 measured performance and the CPM measured performance less so.

The AdvanceMed paper had very limited head-to-head comparisons of all-encounter ACG and DCG methods. In comparing actual to predicted costs, they found the methods to be similarly accurate, with DCGs slightly outperforming ACGs in two of three regions. In comparing risk scores in the entire population and disease detection in a subpopulation (diabetes), the authors found no important differences between the models.

The Ash and Byrne-Logan paper found that the DCG model outperformed the ACG model in both individual and group R^2 tests, with a 4% difference when compared for fifty two-percentile groups of the base year cost.

Meenan and colleagues used the area under the ROC curve statistical method and found model performance essentially identical across competing models. Considering only the highest cost cases, the DCG and GRAM models performed slightly better than the rest. The limited reporting

of subgroup analyses makes meaningful conclusions difficult. Among older subjects, the models appear to be roughly equivalent, while among children GRAM might be best.

Similarities were found in performance among the models when Madden, Mackay, et al used an SSI population, with the CDPS and DCG models doing somewhat better. These authors hypothesize that this is due to the fact that these methods were designed to put more emphasis on correctly predicting high cost conditions in populations where serious illness is common. In the non-SSI population, the ACG model performed best. This population has a much lower mean yearly expense ($882 compared to $4,111), and is largely (75%) children. More variation among the models is seen when looking at the prediction ratios among sub-populations. The ACG model strongly underpredicted CHF and HIV costs in the SSI population; in the non-SSI population, DCGs produced predictions that were far too high for CHF and CDPSs produced predictions that were far too high for HIV.

In conclusion, each of the models reviewed here have some areas of comparative strength. Due to the variation in the vintage of the models and data used in these assessments, we cannot with confidence predict which model will predict best for the military. Based upon our assessment of the strength of prior reviews and the knowledge that we have gained over the years in working with or attempting to work with other risk adjustment models, we suggest the following four models for evaluation in this project:

Adjusted Clinical Group (ACG)
Diagnostic Cost Group (DCG)
Clinical Risk Groups (CRG)
Chronic Disease and Disability Payment System (CDPS)

The first three methods are commercially available; while the last method is available at no cost. However, anecdotal stories suggest that scoring data using the CRG algorithm may be very difficult. Thus, we may have a need for 3M Health Information Systems to do the scoring. This could result in project delays and/or additional costs.

We do not recommend evaluating the Episode Risk Groups (ERGs) method nor the Global Risk Assessment Model (GRAM). ERGs requires the creation of episodes of care, which will be a significant challenge for the enlisted population of the MHS, and requires full pharmacy data. The GRAM method has undergone the lowest level of comparisons across other methods and did not outperform the DCG model in the study conducted by Meenan, Gooodman et al.

We look forward to discussing these recommendations with you in the near future.

APPENDIX

(Cumming, Knutson et al. 2002)

Data

The data for this study are claim and enrollment records from an employer covered ("commercial group") population of 749,145 members enrolled in a "nationwide mix of both PPO and HMO business" from Jan 1, 1998 to Dec 31, 1999. The costs studied are the actual payments to health care providers (i.e., including patient copays). "The data used permits up to 15 diagnoses per inpatient admission and up to 2 diagnoses per outpatient claim. For this analysis, all of the reported diagnoses are used." (pp. 8)

Statistical Measures

Several measures of predictive accuracy are used in this paper, both at individual and group levels. At the individual level, the measures are

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(a_i - \hat{a}_i)^2}{\sum_{i=1}^{N}(a_i - \bar{a})^2}$$

$$\textit{Mean absolute prediction error (MAPE)} = \frac{\sum_{i=1}^{N}\left|a_i - \hat{a}\right|}{N}$$

$$\textit{Cumming's Prediction Measure (CPM)} = 1 - \frac{\sum_{i=1}^{N}\left|a_i - \hat{a}_i\right|}{\sum_{i=1}^{N}\left|a_i - \bar{a}\right|}$$

where
a_i is the actual claim dollars per person i
$\hat{a}_i$ is the predicted claim dollars per person I (based on the regression model)
$\bar{a}$ is the actual claim dollars per person i
i goes from 1 to N (the total number of people)

Group measures of predictive accuracy:

Here the focus is on predicting costs for certain subgroups of the population. The measure of predictive accuracy for each group is

$$\textit{Predictive ratio} = \frac{\text{Total predicted costs for group}}{\text{Total actual costs for group}}$$

In this study, the following subgroups are considered:

1. Those with breast cancer, congestive heart failure, asthma, depression and HIV in 1998,
2. Those with breast cancer, congestive heart failure, asthma, depression and HIV in 1999,
3. Population grouped by quintiles based on medical claim dollars for 1999, and
4. Ranges of medical claim dollars for 1999.

(AdvanceMed 2003)

In detail, the five groupers studied were:

1. CSC/Johns Hopkins' ACG "all encounter" (V5.0)
2. DxCG's DCG "all encounter" (V5.2)
3. DxCG's RxGroups pharmacy (V2.0)
4. DxCG's RxGroups+Inpatient pharmacy plus inpatient (V2.0)
5. DxCG's Inpatient (V2.0)

The groupers were evaluated in their ability to predict future health care costs among TRICARE Prime enrollees—overall as well as selected subpopulations.

Data:

Demographic, diagnosis and health care utilization data were obtained from the following MHS sources: TRICARE Enrollment File, SIDR, SADR, HCSR-I, HCSR-NI (Pharmacy and Non-Pharmacy), CHCS Ad Hoc Pharmacy Report and NMOP.

Note:

1. Pharmacy claims for FY01 were not available—so predictions of the Rx based models were compared with that of the DCG all encounter model.
2. HCSR-NI did not contain NDCs, so the Rx models excluded these claims.
3. The study had "no information available to estimate the number of study members in Active Duty status in the Marines or Navy who may have been afloat for some portion of each of the study years." Given that in one of the regions (Southern California) the Marine and Navy personnel and their dependents constituted 81 of the study population, the authors point out that the impact on relative risk scores will be a "probable minor effect" (pp. 24).

<u>Statistical Measures</u>

The measure of predictive accuracy used in this paper is simply the percentage of average actual expenditures per enrollee explained by the average predicted expenditure per enrollee. This is reported as a percentage.

(Ash and Byrne-Logan 1998)

Data:

Models were fitted using a data set of 1.4 million persons who were enrolled during 1992 and 1993 in various commercially insured, under-age-65, nationally disbursed plans. The fitted models are applied to a validating data set of 192,000 Massachusetts State employees and results from this are reported in model comparisons.

Statistical Measures

Measures of predictive accuracy:

1. Individual-based criterion: R^2 from ordinary least squares regression
2. Group-based criteria:
    a. Predictive Ratio (predicted versus actual costs) for selected medical conditions.
    b. Predictive Ratio for partitions of sample population by model-predicted costs.
    c. Predictive Ratio for 50 2-percentile group partitions of sample population by year 1 costs (and by age).
    d. Grouped R^2: This is a partition-level analogue (i.e., using partition-level sum of squares) using partition count as weights. That is,

$$1 - \frac{\sum_{b=1}^{B} w_b * (AveY_b - Ave\hat{Y}_b)^2}{\sum_{b=1}^{B} w_b * (AveY_b - \overline{Y})^2}$$, where b=1,2,…,B are the partitions,

$\overline{Y}$ denotes the population average, $AveY_b$ the partition average of actual costs and $Ave\hat{Y}_b$ the partition average of predicted costs.

(Meenan, Goodman et al. 2003)

Study sample

Used data on health care utilization of 1.53 million enrollees in five HMOs (Colorado, northeastern Ohio, eastern NY/western New England, Minneapolis/St.Paul and western Washington state) in 1995 and 1996. In the prediction exercise, this population was split into a data fitting subsample of 1.43 million and a validation subsample of the remaiing 106,000 enrollees.

Statistical Measures

Measures of predictive accuracy

"Risk models produce individual-level predictions of annual healthcare expense, which are the sum of relevant coefficients. Predicted expense totals are the sum of relevant coefficients, e.g., GRAM predicted that a 55-year-old woman receiving a 1995 asthma diagnosis would generate a 1996 expense of $1586 = $976 (female 50-62) = $610 (asthma). Expense data produce a distribution of individual-level actual annual expenses. Each risk model then produces a distribution of individual-level predicted annual expenses. For each model, actual and predicted distributions are ordered from highest to lowest, and an a priori percentage threshold separating high- and low- cost individuals is set within each distribution. For actual expense, this threshold establishes high-cost 'prevalence'. Sensitivity is the proportion of high-cost cases correctly forecasted; specificity is the analogous proportion of low-cost cases."

"We used the data to generate a series of receiver operating characteristic (ROC) curves, one for each model… We also estimated the area under each ROC curve (AUC), a quantitative measure of overall discrimination that represents the percentage of randomly selected pairs of actual high- and low-cost patients for which the high-cost patient has the higher predicted expense, i.e., the more 'abnormal' test result. Discrimination rises with the number of random pairs for which this is true. In other words, higher AUC values represent superior overall discrimination. We used a non-parametric trapezoidal method to approximate the AUC and calculate standard errors. Note that the trapezoidal method generally underestimates discrimination. We present results for the most accurate version of each model tested based on highest estimated AUC." [Meenan, Goodman et al. 2003, pp. 1305]

References

ACG (2004). Adjusted Clinical Groups.

AdvanceMed (2003). Applicability of predictive risk groupers to MHS data model: Final report, model testing and validation, AdvanceMed.

Ash, A. and S. Byrne-Logan (1998). "How well do models work? Predicting health care costs." Proceedings of the Section on Statistics in Epidemiology of the American Statistical Association: 42-49.

Cumming, R., D. Knutson, et al. (2002). A comparative analysis of claims-based methods of health risk assessment for commercial populations, Society of Actuaries.

Elixhauser, A. (1996). Clinical classifications for health policy research, version 2: Software and user's guide. Healthcare Cost and Utilization Project (HCUP-3) Research Note 2. Rockville, MD, Agency for Health Care Policy and Research.

Madden, C. W., B. P. Mackay, et al. (2001). Measuring Health Status for Risk Adjusting Capitation Payments. Working Paper, Center for Health Care Strategies, Inc.

Meenan, R. T., M. J. Goodman, et al. (2003). "Using risk-adjustment models to identify high-cost risks." Medical Care **41**(11): 1301-1312.

Shenkman, E. A. and J. R. Breiner (2001). Characteristics of risk adjustment systems, Institute for Child Health Policy, University of Florida.

HOW WELL DO MODELS WORK? PREDICTING HEALTH CARE COSTS

Arlene S. Ash, Susan Byrne-Logan, Department of General Internal Medicine, Boston University School of Medicine
Arlene S. Ash, 720 Harrison Avenue #1108, Boston, MA 02118

Key Words: Grouped R^2, summary performance measures, DCG, ACG

Introduction

A common problem in statistical modeling is determining how well models "work." For predicting a continuous outcome, the traditional R^2 is often reported. But, especially when the largest achievable R^2s are small and implementing a more powerful model may be costly, potential users need to assess the practical implications of small differences in R^2s.

In this paper we explore alternative measures and graphical methods for describing and comparing models that predict expected costs of people who sign up for health plans (such as HMOs). The intended application is in calculating payments to plans that adjust for the health care costs of the particular people they enroll. While models based on age and sex can reliably distinguish population subgroups whose costs differ by as much as 10 to 1 (for, say, people over 65 versus 10 year olds), R^2s for age-sex models rarely exceed 0.02. Health-based predictions explain more of the variation, detect subgroups whose costs differ more, and have much larger, but still small, R^2s.

We develop an exemplary range of numerical summaries and graphical displays which can be used to create rich pictures of model performance. These ideas are useful at the most basic level of understanding the strengths and weaknesses of any imperfect model. They may be particularly important when choosing which of two rather similar models should be used for a particular purpose.

Background

When health care purchasers pay the same for each person, plans make (undeserved) profits for signing up healthy people and predictably lose money on each sick enrollee. If, on the other hand, health care purchasers (such as government agencies or employers) pay the right amount to cover sick people's costs, each HMO should "be happy" to compete with other plans to care for sick people. For fairness and economic efficiency, payments to plans should approximate actual expenses within "policy-relevant" subgroups of populations, such as persons with particular chronic illnesses that usually require high-cost care, or persons who have been expensive to care for in the past.

In a broad-based, under-65-year-old population most people incur few expenses in a year and the highest costs are often due to unpredictable events, such as accidents. To encourage healthy competition among plans, we do not need to be able to identify exactly who will cost nothing next year and who will experience a random, catastrophic event. We need "merely" pay proper valuations based on people's "expected costs."

Thus, we seek to predict year 2 health care costs from available year 1 data, including age and sex and other risk factors, such as information about the medical problems seen in year 1 (in particular, the diagnoses listed on encounter records). Age and sex are usually available, but do not discriminate very well. Year 1 health care costs may be available and are substantially more powerful predictors of future expenditures. However, health care purchasers do not like to base payments on prior costs because they reflect provider practice style as well as the illnesses being treated. The goal is to pay more when a person's illnesses are generally known to require more care, rather than simply because more expensive care was given.

In this paper we develop and illustrate the use of multiple performance criteria in comparing two diagnosis-based models for risk adjusting: Ambulatory Care Group's (ACG) diagnosis-group model and Diagnostic Cost Group's (DCG) multi-condition model. Potential users with particular intended applications can use such methods to evaluate and compare the most pertinent strengths and weaknesses of models. The "heritage," logic and most common applications of the two systems differ substantially.[1] ACGs were originally developed using binary-splits algorithms on exclusively ambulatory data to estimate expected numbers of ambulatory clinic visits during the same period in which various diagnoses were recorded (concurrent modeling). In contrast, DCGs were developed using ordinary least squares regression to predict next year's costs for over-age-65 Medicare patients based on the single worst medical problem that led to a hospitalization during the current year (prospective modeling). However, in subsequent developments each methodology has been refined, adapted and applied to other populations and more comprehensive data sources, using both concurrent and prospective modeling, and for a wide range of potential applications.[2, 3] For our purposes, the important fact is that each model can be used to predict next year's total health care costs based on age, sex and diagnoses in under-65 commercially-insured populations.

Potential users of either model need to be convinced that estimates based on age and sex alone (which is much easier to implement) are importantly inferior to any diagnosis-based modeling. What is

"important" depends on the user. For example, a health care purchaser, such as the Massachusetts General Insurance Commission whose state-employee health care data we examine in this paper, wants to use a model that estimates future costs for people who have been high-cost, or who have particular medical problems, fairly. If the pricing is not fair, providers who care for such people face unwarranted financial jeopardy, which can lead to serious access problems for sick people.

Methods

Fitting the models

A key principal is using validated performance measures. This is always desirable when assessing a model for use in a new setting, and is particularly important for making fair comparisons across models with different levels of complexity (degrees of freedom) or which were developed on data with different characteristics. Thus we fit each model by regressing Y, year 2 annual spending, on various explanatory variables, always including dummy variables to distinguish among age/sex categories. Our FITTING database contains 1.4 million persons who were enrolled during 1992 and 1993 in various commercially insured, under-age-65, nationally disbursed plans. Each case is a person who is covered (for at least part of the time) in each year. These data are described in detail elsewhere.[4]

Specifically, we regress Y (that is, year 2 cost) on predictors that define four models in the FITTING data. The models are: AGE/SEX only (in 16 categories), PRIOR COST (with two predictors relating to year 1: total cost and an indicator for NOCOST), ACGs (with 32 indicators for Ambulatory Diagnosis Groups) and DCGs (with indicators for 84 Condition Categories). †

We apply our fitted models to a VALIDATING data set consisting of 192,000 Massachusetts State employees.[4] This yields a $\hat{y}$ (that is, a predicted year 2 cost) for each person for each model. Since there is no a priori reason that any of these $\hat{y}$s will have average values close to the average Y in the new data, we recalibrate as follows. For each model, we regress Y on a common set of relevant demographic variables and a single additional variable, the model-computed $\hat{y}$. This yields a new $\hat{y}$, whose mean value in the VALIDATING data matches that of Y. In this way we account for additional factors that may have special relevance in the new population. In this example, we add two indicator variables to the age/sex markers: one that distinguishes a state employee from other persons (dependents) covered under the same contract and a second that identifies under-age-65 retirees.

Examining the data

We briefly describe the distributions of actual and predicted costs in the VALIDATING data, reporting means, standard deviations (SD), coefficients of variation (CV), medians, minima and maxima.

Evaluating model fit

First, for each model we produce the traditional R^2, defined as 1 – (SSE/SST), where SSE and SST are each sums, over all observations in the VALIDATING data, of terms of the form $(Y - \hat{y})^2$ and $(Y - \bar{y})^2$, respectively.

Then we produce other, richer descriptions of model performance, all based on assessing how well average $\hat{y}$s match average Ys within subgroups of interest. The subgroups, or "bins," that we use in this paper illustrate the kinds of categorizations that may be useful to policy makers. They are as follows:

1) People with particular diseases. These groups are identified using diagnostic codes. To facilitate fair and meaningful comparisons we use definitions constructed by an outside panel of evaluators who do not know how the diagnostic information is coded and used in the competing models. The 11 specified disease groups used here are illustrative. They are: persons with Lymphoma, Inflammatory bowel disease, Asthma, Cystic fibrosis, HIV/AIDS, Arthritis, Acute myocardial infarction, Diabetes with complications, Benign/unspecified hypertension, Alcohol/drug dependence, and Depression. In addition, we examine an "Any chronic disease" subgroup as defined by the same panel of outside evaluators.
2) We also form three distinct sets of subgroups, each of which "partition" the population (in that each person belongs to one and only one subset of the partition). The partitions we examine are based on:
   - up to 19 levels of model-predicted cost (less than \$250, \$250-\$499, \$500-\$749, \$750-\$999, \$1000-\$1500… up to \$40000 and higher). We form two such partitions, using: i) predictions under the DCG model to bin people, and ii) predictions under the ACG model to bin, and,

† Each of ACGs and DCGs provide more than one model structure, and the model names have changed over time. The two specific models that we use here are the currently most powerful models produced by each development group. What we call ACG and DCG models in this paper are known more specifically in the risk adjustment literature as the Ambulatory Diagnostic Group (or ADG) and Hierarchical Condition Category (or DCG/HCC) models.

- 50 2-percentile groups of year 1 costs. To break the 25% of individuals with no expenses in year 1 into 2%-sized subgroups, we use a second-level sort on increasing age.

For each of the illness-defined subgroups we draw bar graphs of actual average costs versus average predicted costs under each of the ACG and DCG models. We also display the same information as Predictive Ratios (PRs) for each of ACGs and DCGs. A PR for a group is defined as its predicted year 2 costs divided by its actual costs; thus, PRs less than 1.00 are evidence of underpricing.

We compute, for each bin in each of the two model-predicted-cost partitions, the following: numbers of cases, average actual cost, and average predicted costs. We plot each average cost progression (actual and model-predicted) across bins of increasing predicted cost. Better models will track actual costs better across the bins. When comparing two models, we bin "both ways" because we expect that each model has an advantage in tracking costs better within groups defined by its own levels of predicted risk. Importantly, we retain the same bins and scales on the two plots, to enhance our ability to visually compare the models.

For the 2-percentile partition based on year 1 costs, we produce a scatterplot with 2 sets of fifty points each, one set depicting DCG-predicted cost versus actual and the other, ACG-predicted cost versus actual. We also mark the 45-degree line on the plot. Since the Y-axis represents actual cost, each 2-percentile subgroup leads to one DCG- and one ACG-based point at the same height; whichever is (horizontally) closer to the 45-degree line represents a better fit of that model for that subgroup.

Bin	Count	Actual	Predicted	Squared Error
1	n_1	$AveY_1$	Ave $\hat{y}_1$	$(AveY_1 - Ave\ \hat{y}_1)^2$
2	n_2	$AveY_2$	Ave $\hat{y}_2$	$(AveY_2 - Ave\ \hat{y}_2)^2$
3	n_3	$AveY_3$	Ave $\hat{y}_3$	$(AveY_3 - Ave\ \hat{y}_3)^2$
…				
B	n_B	$AveY_B$	Ave $\hat{y}_B$	$(AveY_B - Ave\ \hat{y}_B)^2$
All	n	$\Sigma n_b * AveY_b$	$\Sigma n_b *$ Ave $\hat{y}_b$	$\Sigma n_b * (AveY_b - Ave\ \hat{y}_b)^2$

Finally, for each partition, we produce a single, summary number, called a grouped R^2, defined as an appropriately weighted version of the traditional 1 – SSE/SST definition.

Note that $\bar{y}$, the population average, equals $(\Sigma n_b * AveY_b)/n$ or $\Sigma w_b * AveY_b$, with $w_b = n_b/n$. When the $\hat{y}$s are calibrated to the data, $\bar{y}$ also equals $\Sigma w_b * Ave\ \hat{y}_b$. In this notation, then, for a given partition, with bins 1,2, .. , B, we have:

Grouped R^2 (a partition) $= 1 - SSE/SST$

$= 1 - \Sigma w_b * (AveY_b - Ave\ \hat{y}_b)^2 / \Sigma w_b * (AveY_b - \bar{y})^2$.

We compute Grouped R^2s for each of the three methods of partitioning, for each of ACGs and DCGs, and for the age/sex-only and prior-cost models.

Results

Annual health care costs, in our validating (State) data, as elsewhere, are highly skewed (Table 1). A substantial fraction (here, about 25%) of a general population incurs no health care costs in a year. These people's next year's costs are substantially less than (here, only 25% as great as) average. The 2% whose year 1 costs are highest use up about 40% of that year's dollars and 17% of the next year's dollars; the highest actual costs in large populations typically exceed $1,000,000.

Table 1. Year 2 Cost Distributions

	Mean	SD	CV	Median	Min	Max
Actual $	$1,818	$7,971	438	$ 320	$-656	$2,003,124
Predictions:						
Age/Sex	1,818	1,018	56	2,035	551	4,061
Prior cost	1,818	1,660	91	1,697	1,298	662,043
DCG	1,818	2,414	133	1,295	422	114,690
ACG	1,818	2,053	113	1,069	74	28,542

Predicted costs should be "skewed" as well, but less than actual costs. (No person's expected costs are as low as 0, nor are any cases "expected' to cost 500 times as much as average.) However, if there are classes of people (identifiable in year 1) whose costs next year will be 25 times average while others will be 0.25 times average, then our models should be able to identify them. If we cannot pay the right amount for very sick people, then HMOs will systematically and predictably lose money for each such person they enroll.

In these data, year 2 costs average $1,818 with a standard deviation (SD) of $7,971 (that is, the SD is 4.4 times larger than the mean).

Table 2 shows that each of the diagnosis-based models had substantially larger validated R^2s than the age/sex model, and even larger R^2s than the prior cost model.

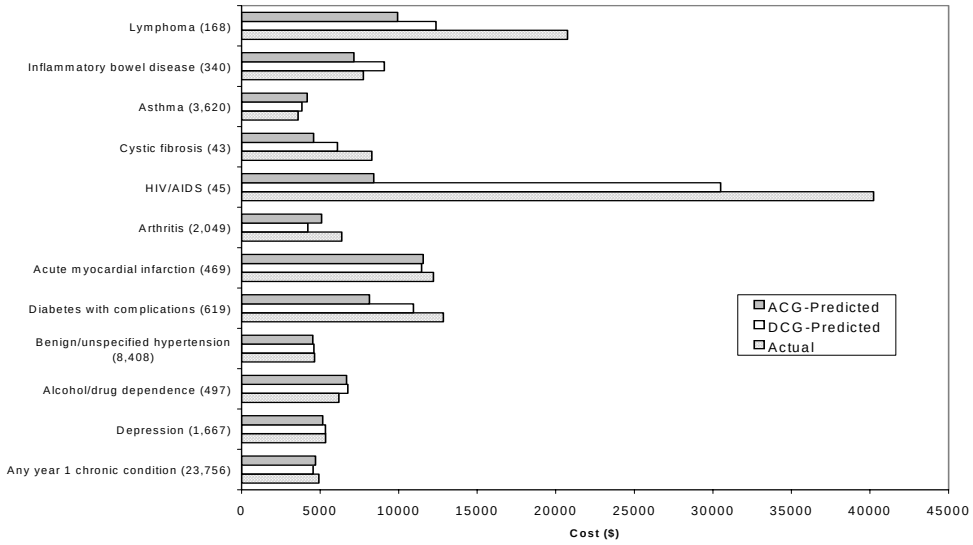
Figure 1 enables visual comparisons of how well the ACG and DCG models' cost predictions match actual year 2 expenses in each of 12 illness-defined subgroups within the State data. Both of the models are moderately effective at predicting the costs for most of these subgroups. However, in the two highest cost subgroups (persons with lymphoma and HIV/AIDS) DCGs underpredict substantially and ACGs underpredict even more. Because the scale on such a figure is made to accommodate the most expensive group's costs, it is not so easy to see what may be large <u>percentage</u> differences between actual and predicted costs in lower-cost groups, such as those defined by Asthma, Arthritis, Hypertension, Depression, and the Any chronic condition groups.

Table 2. The Ordinary R^2

Model	R^2
Age/Sex	1.6%
Prior Cost	6.3
ACG	6.6
DCG	9.2

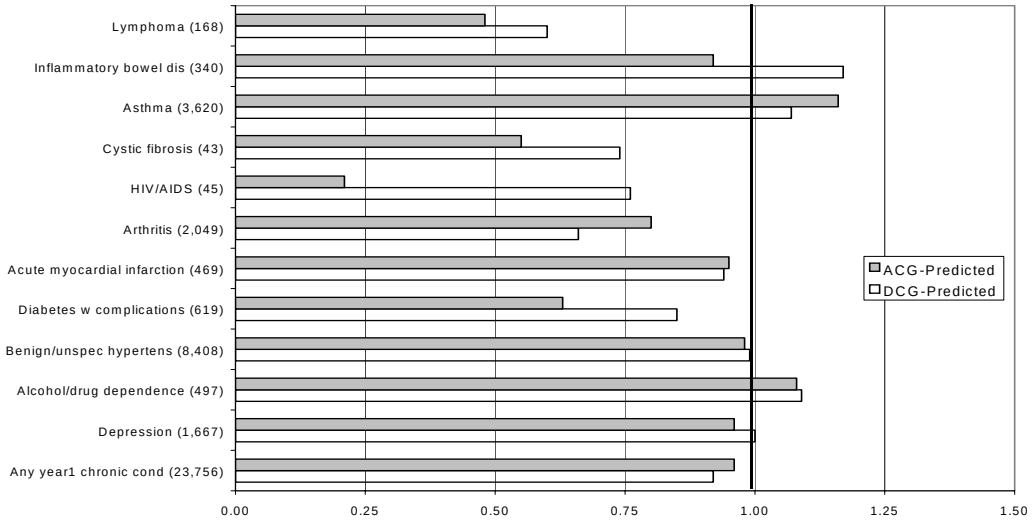
Figure 2 is based on the same information as Figure 1, but by showing predictive ratios it

Figure 1. DCG and ACG-Predicted versus Actual Costs, Selected Medical Conditions*



*Groups based on presence of year 1 diagnoses

Figure 2. Predictive Ratios* for DCGs and ACGs, Selected Medical Conditions**



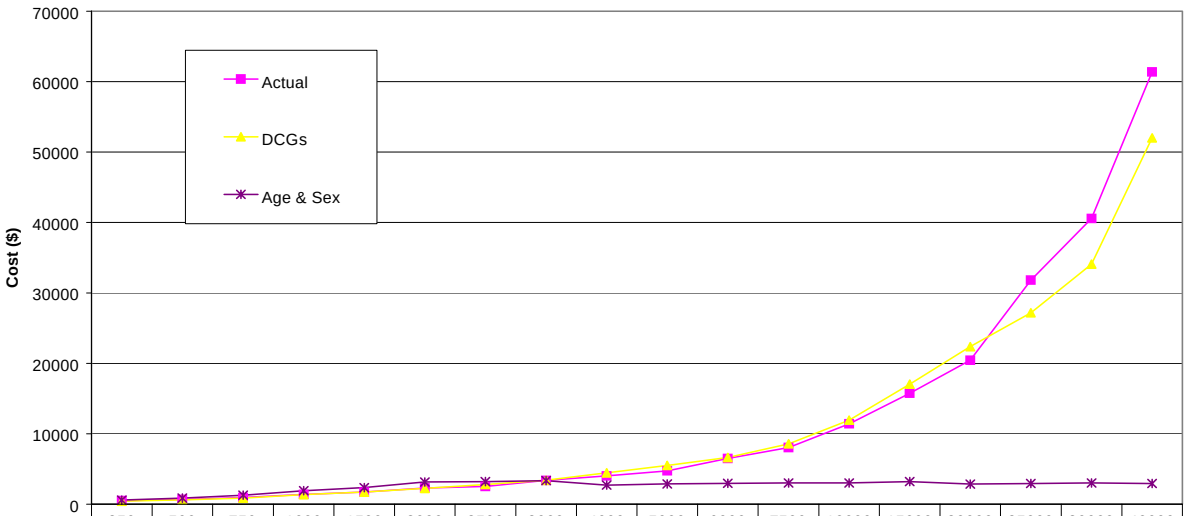
* Predictive Ratio = Actual year 2 cost / Predicted year 2 cost

** Groups based on presence of year1 diagnoses

clarifies the head-to-head comparison between the two competitor models. By normalizing each group to 1.00, the figure makes discrepancies in lower-cost disease groups easier to see. Here, for example, it is evident that both models underpay quite substantially (on a percentage basis) for people with Arthritis.

Figure 3 compares the DCG model to an age/sex model in properly pricing individuals across a wide-spectrum of expected costs. This picture dramatizes several things: the DCG model identifies subgroups whose costs turn out to be much higher (and much lower) than average; the DCG model predicts prices at all ranges of the prediction-spectrum reasonably well; and, age-sex models drastically underpay those whose DCG-predicted costs exceed $4,000. However, to fairly and clearly compare ACGs and DCGs requires a pair of similarly-scaled pictures, as provided by the

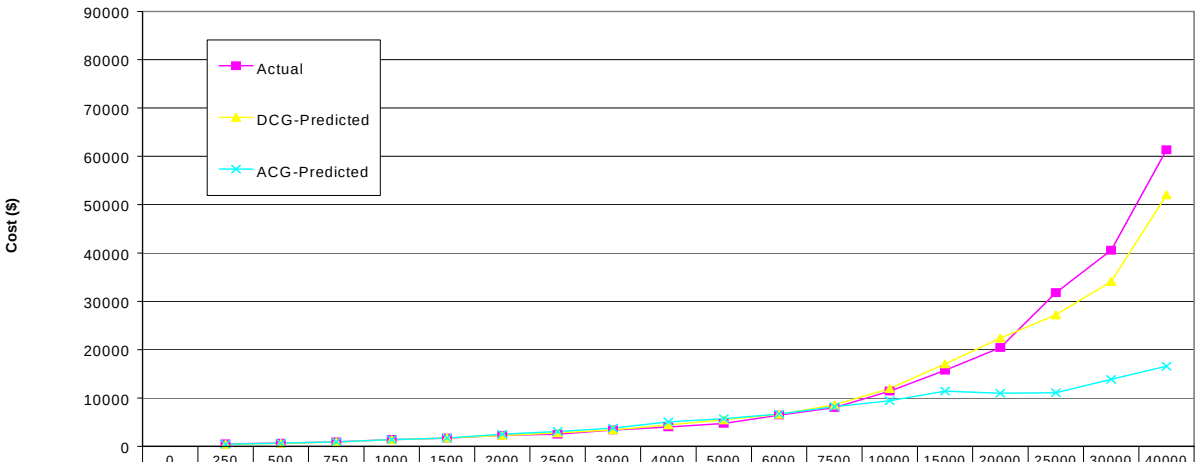
Figure 3. Means of Actual and Predicted Cost by DCG-Prediction Group



	250	500	750	1000	1500	2000	2500	3000	4000	5000	6000	7500	10000	15000	20000	25000	30000	40000
Actual	510	672	931	1391	1707	2295	2510	3373	3993	4734	6478	8025	11415	15741	20426	31804	40559	61380
DCGs	417	620	867	1347	1714	2242	2779	3406	4451	5485	6624	8557	11939	17042	22377	27181	34087	52026
Age & Sex	570	855	1262	1915	2335	3140	3195	3332	2695	2859	2951	3012	3006	3217	2853	2929	3010	2926

Predicted Cost Group ($)

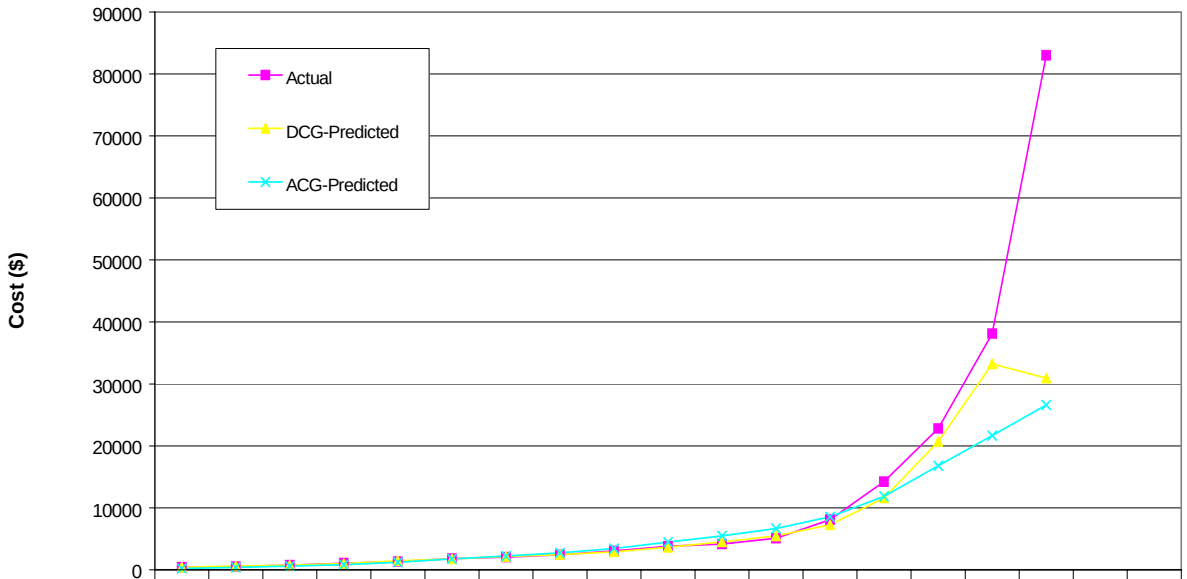
Figure 4. Means of Actual and Predicted Cost by DCG-Prediction Group (2)



	0	250	500	750	1000	1500	2000	2500	3000	4000	5000	6000	7500	10000	15000	20000	25000	30000	40000
Actual		510	672	931	1391	1707	2295	2510	3373	3993	4734	6478	8025	11415	15741	20426	31804	40559	61380
DCG-Predicted		417	620	867	1347	1714	2242	2779	3406	4451	5485	6624	8557	11939	17042	22377	27181	34087	52026
ACG-Predicted		427	643	904	1382	1741	2504	3083	3798	5089	5719	6702	8184	9458	11414	10972	11091	13850	16565
N	0	21,650	41,384	22,649	28,782	30,786	16,100	5,828	9,086	4,944	3,474	2,747	1,980	1,314	441	257	227	154	74

Predicted Cost Group ($)

Figure 5. Means of Actual and Predicted Cost by ACG-Prediction Group



	0	250	500	750	1000	1500	2000	2500	3000	4000	5000	6000	7500	10000	15000	20000	25000	30000	40000
Actual	430	507	746	1058	1323	1794	2047	2430	3047	3791	4144	5081	8089	14222	22811	38117	83027		
DCG-Predicted	418	585	765	1004	1463	1748	2141	2462	2916	3646	4507	5481	7277	11592	20673	33228	30929		
ACG-Predicted	213	380	597	838	1198	1752	2229	2734	3428	4467	5481	6681	8534	11849	16786	21652	26607		
N	15,769	27,381	23,276	13,566	42,493	15,915	13,738	8,506	10,935	6,017	4,145	4,217	3,452	2,027	381	56	3	0	0

Predicted Cost Group ($)

next two figures. To make each graphic more informative and to enhance the comparison, we add N's for each bin.

Figures 4 and 5 plot actual, DCG-predicted and ACG-predicted costs across subgroups binned by levels of predicted cost. In Figure 4, we use DCG-predictions to bin; in Figure 5, ACGs. Notice that the DCGs never predict that a person will cost less than $250 and that ACGs never make predictions as high as $30,000. Also, we should not put too much weight on the points at the far-right of Figure 5, since they represent the actual and predicted experience of just 3 people (these are the only people for whom ACG-predicted costs exceed $25,000).

Interestingly, DCGs predict better than ACGs not only within DCG-predicted cost groups (as in Figure 4), but also within ACG-predicted groups (Figure 5). Of particular note is the fact that average actual costs within the lowest group defined by ACGs are not as low as the ACG model predicts ($430 for actual costs, versus only $213 in ACG-predicted costs). Actual costs for this group, in fact, look much more like the expenses predicted by the DCG model ($418).

Figure 6 provides yet another perspective on the data. It compares predicted to actual year 2 costs within 2-percentile bins based on actual year 1 costs. A good model should predict well within these groups, because HMOs can know which of their enrollees have been expensive in the past, and should be assured that, on average, future payments for these people will not differ too much from their costs. In the most expensive of these groups (which averages $15,608 in year 2 costs), both models "know" that these people will be far more expensive than average ($1,818), but still underpredict substantially. However, DCGs predict $9,657 (for a PR of 0.62), while ACGs do even worse ($8,200, with a PR of .53). However, in the other 2-percentile groups shown in Figure 6, ACG predictions are almost always closer to actual costs than are the DCG predictions; ACGs "win" in most of these subgroups.

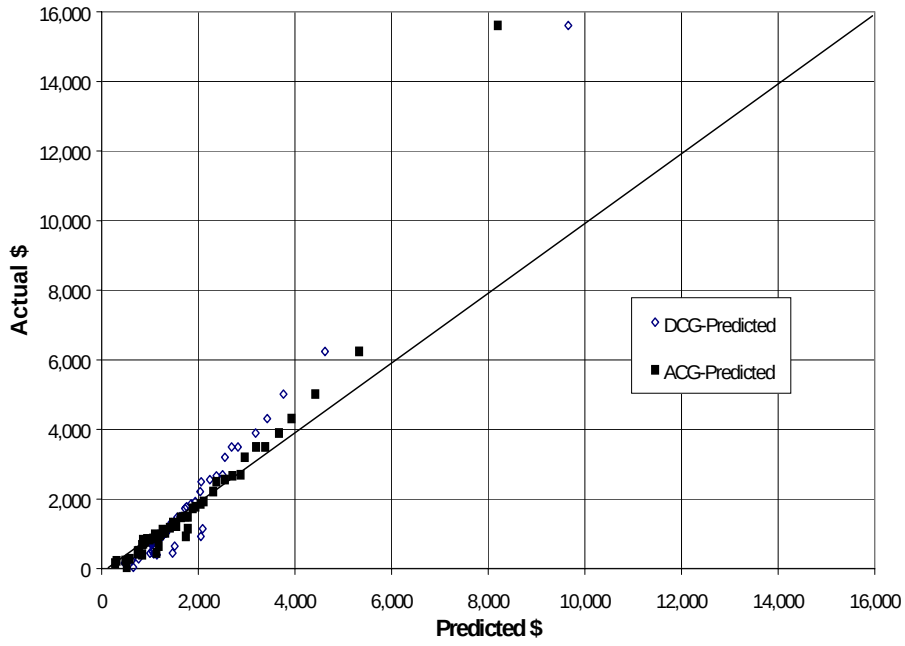
Finally, we summarize the comparisons of the two models' performance that we have just been making visually, by computing grouped R^2s for the three partitions that underlay Figures 4 through 6 (see Table 3). These numbers summarize and quantify the impressions that previous calculations and pictures have conveyed. These are: age and sex alone do not predict well within subgroups defined by any of the three partitions; a prior cost model does much better, but not as well as either diagnosis-based model; and, DCGs tend to predict more accurately than ACGs. One surprising thing in this table is that DCGs exceed ACGs in the grouped R^2 comparison based on 2-percentile

year 1 cost groups, even though ACGs predict better than DCGs in most of the 50 subgroups. The answer to this apparent anomaly lies in the squared error "penalty function" implicit in the grouped R^2 definition, which gives enormous weight to the large prediction error contributed by the single highest cost subgroup. If a few "large mistakes" in pricing are, in fact, much more serious than a cluster of smaller errors, then squared error loss may be appropriate. However, had we measured loss using "absolute error" rather than "squared error," ACGs would look better on this measure for that partition.

Table 3. Grouped-R^2s for 4 Models using 3 Partitions

Partition method	DCG-Pred$	ACG-Pred$	$2%-ile Yr1$
# of bins	18 17	50	
Models			
1. Age/Sex	29.6%	21.3%	17.3%
2. Prior Cost	52.9	44.2	50.5
3. ACG	91.6	69.4	78.6
4. DCG	96.7	98.2	81.9

Figure 6. Predicted versus Actual Year 2 Costs for 2-Percentile Groups Based on Year 1 Cost



Discussion

Looking at predictive performance, several ways provides a rich picture of the strengths and weaknesses of models. Each of the two diagnosis-based models predict much better (according to all measures) than age-sex models, are reasonably good at predicting within most of the illness-defined subgroups which we examined, and generally predict somewhat better than prior costs. However, we have seen that each fails to predict costs as high as actual expenses for several extremely high-cost subgroups, including people with lymphoma, with HIV or AIDS, and those with the highest 2 percent of costs last year. While DCGs estimate actual costs better within each of these subgroups, they still underpredict by amounts that could lead to access problems. These insights can be used to inform the developers of where their models could be improved or to inform users where inequities may arise if the models are implemented without adjustments.

ACGs predict more accurately within most groups defined by levels of prior cost, with the notable exception of those 2 percent of cases whose costs were highest last year. However, DCGs' greater accuracy in pricing the very highest-cost people makes it "better overall", when viewed through the "lens" of the squared error loss function that underlies a grouped R^2 summary statistic. Those who only examine a summary measure will fail to see the interesting and possibly important subtleties that underlie it.

Particular failings of models, once identified, can be used to improve them. For instance, the ACG developers do not dictate how their model must be used to make predictions. After finding that our use of ACGs underprices people whose ACG-prediction is below $250 per year, any of several easy mathematical adjustments could be used to eliminate this problem. For example, the regression in the state data could also include squared and cubed Yhat predictors. The point is that we won't know which problems need fixing unless we have examined model performance in multiple ways.

Conclusion

We computed standard, validated R^2 values to compare the overall ability of models to predict accurately. In addition, we "binned" cases in ways that are likely to be important to those who want to predict next year's health care expenses from this year's data. Other binnings might also be of interest to such users, such as by pre-specified levels (rather than quantiles) of year 1 cost, by hospital use last year (any versus none or number of visits or number of days of inpatient stay), or by some measure of "frailty." Levels of frailty might either be inferred from data in these files or tested in a new data set that captures this health dimension more reliably. Each binning provides additional insight into the kinds of people who are well or poorly priced by each candidate model. However, if different bins contain very different numbers of people, n's for each bin should also be conveyed. When very high-cost groups have extremely small n's, not only are actual means very unstable, but also pictures which give equal visual weight to each point implicitly exaggerate a model's true discriminatory power. As a general principal, actual n's should be indicated for any "small" group. When dealing with data as variable as health care costs, " small" may mean groups of up to 1,000.

We have illustrated how to examine and display the levels of concordance between actual costs and predictions of cost within each subgroup, by directly comparing those averages or through predictive ratios. Grouped R^2s provide summary numbers (measures) which enable uni-dimensional comparisons of models based on how well they fit across all the bins provided by any partition of the data. Ideally, we seek to bin using a "neutral" but strong cost discriminator that is not "in the competition" for use as a predictive model. "Prior cost" serves this role here.

As natural a concept as squared-error-loss is to mathematical statisticians, it is only one of many possible loss functions. It may or may not seem reasonable to a user, to count, say, a single group mispriced by $5,000 as being much worse than 10 equally sized groups mispriced by $500 each. Note: $(5000)^2$ is 10 times larger than $10x(500)^2$. Weighting penalties on larger errors disproportionately heavily is arguably appropriate in this context, since many small, scattered, non-systematic mispricings may not trigger "bad behaviors" from health plans in the way that a single, readily-identified group that is very much underpriced is likely to. However, squared-error loss is not the only, and possibly not the most appropriate loss function. Mean absolute error, for example, could easily be used in the "grouped" summary-measure context, and may prove useful in addition to grouped R^2.

In short, no single picture or number suffices to characterize the various ways in which individual models may perform more or less well. We urge statisticians and potential real-world users of models to work together to construct rich descriptions of model performance of the sort that we have produced here.

---

[1] Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor, Michigan. Health Administration Press. 2nd Edition. Chapter 1. 1997.

[2] Fowles J, Weiner J, Knutson D, Fowler E, Tucker A, Ireland M. Taking health status into account when setting capitation rates: a comparison of risk adjustment methods. JAMA. 1996;276(16):1316-1321.

[3] Ellis R, Pope GC, Iezzoni LI, Ayanian JZ, Bates DW, Burstin H, Ash A. Diagnosis-based risk adjustment for Medicare capitation payments. Health Care Financing Review. 1996;17(3):101-128.

[4] "Risk Adjustment for the Non-Elderly." AS Ash, RP Ellis, W Yu, EA MacKay, LI Iezzoni, JZ Ayanian, DW Bates, HR Burstin, S Byrne-Logan, GC Pope. Final report for the Health Care Financing Administration, contract # 18-C-90462/1-02. June 1998.

Appendix B: Extended Population Description Tables

Appendix B
Extended Population Description Tables

We selected a sample of beneficiaries enrolled for all 24 months from all the TRICARE Prime beneficiaries enrolled for some duration during fiscal years 2001 and 2002. Appendix Table B1 compares the characteristics for beneficiaries enrolled all 24 months to those not enrolled the entire time.

We randomly split the 2.3 million beneficiaries enrolled for all 24 months into an estimation sub-sample of 1.8 million and a validation sub-sample of 0.5 million. The demographic and cost characteristics for these sub-samples are described in Tables B2 and B3.

Table B1. Characteristics of TRICARE PRIME beneficiaries enrolled for 24 months vs. those not enrolled for all 24 months (as of September 2001)

	Beneficiaries enrolled for all 24 months*	Beneficiaries not enrolled for all 24 months*
Number of beneficiaries	2,304,926	1,079,613
% female	47.86	47.83
Age (%)		
0 to 14	28.6	26.3
15 to 24	16.5	28.7
25 to 44	36.9	28.9
45 to 64	18.0	16.2
Race (%)		
White	26.87	26.60
Black	7.13	7.68
Asian or Pacific Islander	0.96	1.06
American Indian or Alaskan native	0.25	0.36
Other	2.04	1.96
Unknown	62.75	62.33
DoD Occupation code (%)		
Combat	15.16	13.52
Non-combat	84.84	86.48
Beneficiary category (%)		
Dependent Active Duty/Guard	45.16	41.12
Retired	9.24	10.16
Dependent Retired or Survivor, Other, Unknown	17.80	21.34
Active Duty and Guard	27.81	27.39
Rank (%)		
Jr Enlist	11.33	26.54
Sr Enlist	65.75	55.03
Jr Officer	7.18	6.71
Sr Officer	13.11	9.28
Warrant	2.43	1.86
Other	0.19	0.59

	Beneficiaries enrolled for all 24 months*	Beneficiaries not enrolled for all 24 months*
Catchment area (%)		
Yes	67.41	67.90
No	32.59	32.10
Resident region (%)		
Northeast (1)	10.53	13.44
Mid-Atlantic (2)	11.35	14.42
Southeast (3)	13.59	12.70
Gulf south (4)	7.50	6.70
Heartland (5)	6.55	7.54
Southwest (6)	13.57	13.46
TRICARE Central (7)	5.75	4.47
TRICARE Central (8)	9.54	7.46
Southern California (9)	8.66	9.27
Golden Gate (10)	2.95	2.62
Northwest (11)	5.35	4.41
Hawaii Pacific (12)	3.06	2.27
Alaska (AK)	1.60	1.23
Sponsor service (%)		
Army	35.30	37.08
Navy, Marines, Navy Afloat	30.57	35.98
Air Force	32.05	24.72
Other	2.07	2.23
Urbanicity (%)		
Central countries of metro areas of 1 million pop. or more	30.87	34.29
Fringe counties of metro areas of 1 million pop. or more	3.44	3.59
Counties in metro areas of 250,000 - 1,000,000 pop.	35.01	31.83
Counties in metro areas of less than 250,000 pop.	13.60	12.65
Urban pop. of 20,000 or more, adjacent to a metro area	4.19	4.32
Urban pop. of 20,000 or more, not adjacent to a metro area	5.11	5.13
Urban pop. of 2,500-19,999, adjacent to a metro area	3.64	3.65
Urban pop. of 2,500-19,999, not adjacent to a metro area	1.54	1.64
Completely rural (no places with a pop. of 2,500 or more) adjacent to a metro area	0.31	0.35
Completely rural (no places with a pop. of 2,500 or more) not adjacent to a metro area	0.25	0.33
Missing/Unknown	1.87	2.06

* FY2001 & FY2002

Table B2. Characteristics of fitted and validation samples as of September 2001

	Fitting sample	Validation sample
Number of beneficiaries	1804926	500000
% female	47.9	47.9
Age (%)		
0 to 14	28.6	28.7
15 to 24	16.5	16.6
25 to 44	36.9	36.9
45 to 64	18.0	17.9
Race (%)		
White	26.9	26.8
Black	1.0	0.9
Asian or Pacific Islander	7.1	7.1
American Indian or Alaskan native	0.2	0.2
Other	2.0	2.1
Unknown	62.7	62.8
DoD Occupation code (%)		
Combat	15.1	15.2
Non-combat	84.9	84.8
Beneficiary category (%)		
Dependent Active Duty/Guard	45.1	45.3
Retired	9.3	9.2
Dependent Retired or Survivor, Other, Unknown	17.8	17.8
Active Duty and Guard	27.8	27.8
Rank (%)		
Jr Enlist	11.3	11.3
Sr Enlist	7.2	7.2
Jr Officer	0.2	0.2
Sr Officer	65.7	65.8
Warrant	13.1	13.1
Other	2.4	2.5
Catchment area (%)		
Yes	32.6	32.5
No	67.4	67.5
Primary Care Manager Type (%)		
Military	84.7	84.8
Civilian	15.3	15.2
Resident region (%)		
Northeast (1)	10.5	10.5
Mid-Atlantic (2)	11.3	11.4
Southeast (3)	13.6	13.5
Gulf south (4)	7.5	7.5
Heartland (5)	6.5	6.6
Southwest (6)	13.6	13.6
TRICARE Central (7)	5.7	5.8
TRICARE Central (8)	9.5	9.5
Southern California (9)	8.7	8.6
Golden Gate (10)	3.0	2.9

	Fitting sample	Validation sample
Northwest (11)	5.4	5.4
Hawaii Pacific (12)	3.1	3.1
Alaska (AK)	1.6	1.6
Sponsor service (%)		
Army	35.3	35.3
Navy, Marines, Navy Afloat	32.0	32.1
Air Force	30.6	30.5
Other	2.1	2.1
Urbanicity (%)		
Central countries of metropolitan areas of 1 million population or more	30.9	31.5
Fringe counties of metropolitan areas of 1 million population or more	3.4	3.5
Counties in metropolitan areas of 250,000 - 1,000,000 population	35.0	35.6
Counties in metropolitan areas of less than 250,000 population	13.6	13.9
Urban population of 20,000 or more, adjacent to a metropolitan area	4.2	4.3
Urban population of 20,000 or more, not adjacent to a metropolitan area	5.1	5.2
Urban population of 2,500-19,999, adjacent to a metropolitan area	3.6	3.7
Urban population of 2,500-19,999, not adjacent to a metropolitan area	1.6	1.6
Completely rural (no places with a population of 2,500 or more) adjacent to a metropolitan area	0.3	0.3
Completely rural (no places with a population of 2,500 or more) not adjacent to a metropolitan area	0.3	0.3
Missing/Unknown	2.0	0.1

Table B3. Total individual cost of fitted and validation samples in FY2001 by demographic characteristics

	Fitting		Valid	
	Mean	Coefficient of variance	Mean	Coefficient of variance
Gender				
Female	2,036	256	2,047	297
Male	1,383	467	1,391	435
Age				
0 to 14	965	690	988	627
15 to 24	1,440	307	1,450	389
25 to 44	1,767	278	1,754	259
45 to 64	2,944	248	2,992	280
Race				
White	1,557	309	1,577	338
Black	1,471	294	1,504	311
Asian or Pacific Islander	1,792	285	1,752	282
American Indian or Alaskan native	1,887	397	1,672	308
Other	1,582	299	1,620	343
Unknown	2,044	311	2,066	379
DoD Occupation code				
Combat	1,389	301	1,419	432
Non-combat	1,750	352	1,756	345
Beneficiary category				
Dependent Active Duty/Guard	1,477	414	1,503	416
Retired	2,557	313	2,593	346
Dependent Retired or Survivor, Other, Unknown	2,215	283	2,192	284
Active Duty and Guard	1,429	295	1,428	290
Rank				
Jr Enlist	1,832	267	1,896	399
Sr Enlist	1,703	371	1,699	325
Jr Officer	1,462	247	1,475	435
Sr Officer	1,650	339	1,685	395
Warrant	1,778	308	1,779	449
Other	1,611	376	1,722	312
Catchment area				
Yes	1,741	366	1,760	378
No	1,601	299	1,591	294
Resident region				
Northeast (1)	1,608	335	1,662	461
Mid-Atlantic (2)	1,753	604	1,734	376
Southeast (3)	1,607	283	1,634	292
Gulf south (4)	1,733	265	1,702	244
Heartland (5)	1,573	285	1,577	356
Southwest (6)	1,848	280	1,870	312

	Fitting		Valid	
	Mean	Coefficient of variance	Mean	Coefficient of variance
TRICARE Central (7)	1,797	297	1,759	250
TRICARE Central (8)	1,624	287	1,653	335
Southern California (9)	1,621	334	1,605	474
Golden Gate (10)	2,005	343	1,986	308
Northwest (11)	1,668	293	1,720	474
Hawaii Pacific (12)	1,698	266	1,724	317
Alaska (AK)	1,608	222	1,582	232
Sponsor service				
Army	1,749	316	1,768	357
Navy, Marines, Navy Afloat	1,615	450	1,630	376
Air Force	1,742	280	1,737	338
Other	1,245	321	1,241	309
Urbanicity				
Central countries of metropolitan areas of 1 million population or more	1,826	444	1,814	412
Fringe counties of metropolitan areas of 1 million population or more	1,690	296	1,714	360
Counties in metropolitan areas of 250,000 - 1,000,000 population	1,655	281	1,674	320
Counties in metropolitan areas of less than 250,000 population	1,549	272	1,547	289
Urban population of 20,000 or more, adjacent to a metropolitan area	1,666	269	1,661	267
Urban population of 20,000 or more, not adjacent to a metropolitan area	1,559	236	1,608	403
Urban population of 2,500-19,999, adjacent to a metropolitan area	1,788	305	1,837	334
Urban population of 2,500-19,999, not adjacent to a metropolitan area	1,689	338	1,790	375
Completely rural (no places with a population of 2,500 or more) adjacent to a metropolitan area	1,892	276	1,987	249
Completely rural (no places with a population of 2,500 or more) not adjacent to a metropolitan area	1,901	278	2,178	315

Appendix C: Data Extraction Materials

Appendix C
Data Extraction Materials

Data were pulled from the DOD dataset twice for this study. The study subjects were the same in both data pulls. Additional variables were necessary in order to run the CRG risk adjustment model (primarily dates of service, place of service, provider specialty and procedure codes) and these were added to the second pull of data. The materials below describe the second data pull and include:

Variables in Master Dataset (second data pull)

Full listing of variables in Master Dataset, separated by original datasets from which variables were drawn.

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Point-in-Time Extract (PITE)		
PID	SponSSN (9) DOB (34) Sex (43)	Derived from scrambling SponSSN, DOB, Sex
Age	34	Age=(End FY-DOB)/365.25 If Age >=90 then recode Age=90
Sex	43	F=Female M=Male Z=Unknown
Race	44	C=White M=Asian or Pacific Islander N=Black R=American Indian or Alaskan native X=Other Z=Unknown
DoD Occupation Code	142	Combat, Non-Combat Compute number of months beneficiary is eligible under each occupational code.
Patient Zip Code	347	Substring to 3 char
Residence Region	352	01-12, AK
Catchment Area	385	Y=All other not equal to 'N' N=078*, 09** Compute number of months beneficiary is in Catchment or Non-Catchment
Sponsor Service	412	A=Army N=Navy, Marines, Navy Afloat F=Air Force O=Other
Bencat	415	1=Dependent Active Duty/Guard 2=Retired 3=Dep of Retired or Survivor, Other, Unk 4=Active Duty and Guard Compute number of months beneficiary is eligible under each beneficiary category.

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Point-in-Time Extract (PITE) (cont.)		
Urbanicity	Derive from Patient Zip Code (File Provided by BU)	Rural Suburban Urban
Rank	Derive from Pay Plan Code (@135) and Pay Grade Code (@140)	**Pay Plan** / **Pay Grade** Jr Enlist ME 01-04 Sr Enlist ME 05-09 Jr Officer MO 01-03 Sr Officer MO 04-11 Warrant MW All Other Other All Other All Other
Death Date		
Death Code		
Enhanced Death Code (if available)		
Enhanced Death Date (if available)		

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
TRICARE Enrollment File (TEF)		
Alternate Care Value	ACVfycm	A=Active Duty Prime D=TRICARE Senior Prime E=Non-Active Duty Prime Longitudinal monthly ACV value
Enrollment DMISID	ENRfycm	MTF=MTF PCM MCSC=MCSC PCM Use mapping provided by Kennell Longitudinal monthly ACV value
Standard Inpatient Data Record (SIDR)		
Admission Date (YYYYMMDD)		
End Date of Care (YYYYMMDD)	dispdate	
Treatment DMISID		
Catchment Area		
PRISM Area		
Disposition Status Code		
Admission Source		
Diagnosis Codes	DX1 – DX8	Substring to 5 char
Procedure Codes	PROC1 – PROC8	Substring to 4 char
Treatment MTF	MTF	
Treatment MTF Service	MTFSVC	A=Army N=Navy, Marines, Navy Afloat F=Air Force O=Other
Full Cost, Completed	FULLCOST, Completed	Apply completion factor
Incremental Cost, Completed	INCCOST, Completed	Apply completion factor
Diagnosis/Procedure Code		Total field length of 5. Left justified. First 5 characters of diagnosis code. First 4 characters of procedure code.

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Standard Inpatient Data Record (SIDR) (cont.)		
Code type indicator		1 if record is created from diagnosis 1 (principal diagnosis); 2 if record is created from any other diagnosis code; 3 if record is created from any of the procedure codes
Type of Provider		Set = 1 (every record is from a hospital)
Site of Service		Set = 04 (inpatient)
Date of Service		Disposition Date
Person ID		Create per Risk Assessment Specs
Length of Stay	dmisdays	
Baseline RWP	baserwp	
Outlier RWP	outrwp	
Total RWP	totrwp	
Treatment DMISID		
MEPRS Code		
ICU Days	icudays	
Injury Codes	stanag	
DRG		
Medicare Flag		

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Standard Outpatient Data Record (SADR)		
Encounter Date (YYYYMMDD)		
Provider Specialty Code		
Provider ID		
Treatment DMISID		
Catchment Area		
PRISM Area		
Disposition Status Code		
Diagnosis Codes	ICD1 – ICD4	Substring to 5 char
CPT Codes (E&M, CPT1 - CPT4)	CPT CPT1-CPT4	
Appointment Status	APPTSTAT	1 = Appointment Schedule 3 = Walk in 4 = Sick Call 6 = Telephone Consult
Provider Specialty	SPC	See appended document
Treatment DMISID	DMISID	
Treatment DMISID Service	TXSVC	A=Army N=Navy, Marines, Navy Afloat F=Air Force O=Other
Full Cost, Completed	FCOST, Completed	Apply completion factor
Variable Cost, Completed	COST, Completed	Apply completion factor
Diagnosis/Procedure Code		Total field length of 7. Left justified. First 5 characters of diagnosis code or entire procedure code
Code type indicator		'1' if record is created from diagnosis 1; '2' if record is created from any other diagnosis code; '4' if record is created from any of the procedure codes (including the E&M procedure code)
Type of Provider		See mapping in appendix
Site of Service		07
Date of Service		Encounter Date
Person ID		Create per Risk Assessment Specs
APG1-APG4, APG E&M (5 fields)		
Procedure Codes (all)		

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Standard Outpatient Data Record (SADR) (cont.)		
Diagnosis Codes (all)		
Provider ID		
Treatment DMISID		
MEPRS Code		
Injury Codes		
Appt Status Code		
SDS Flag		
Medicare Flag		
Simple RVU		
Other Insurance Indicator		

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Health Care Service Record - Institutional (HCSR-I)		
Admission Date (YYYYMMDD)		
Begin Date of Care (YYYYMMDD)		
End Date of Care (YYYYMMDD)		
Provider ID		
Provider Zip/Country Code		
Multiple Provider Suffix		
Institution Type		
Admission Diagnosis		
Admission Type		
Type of Submission Code		
Bill Frequency Code		
Catchment Area		
PRISM Area		
Disposition Status Code		
Admission Source		
Amount Allowed, Completed	234	Apply completion factor
Amount Paid, Completed	243	Apply completion factor
Diagnosis Codes (Primary DX DX1-DX8)	358 - 406	Substring to 5 char
Procedure Codes (PProc, Proc1-Proc5)	515-540	
Diagnosis/Procedure Code		First 5 characters of diagnosis code or first 4 characters of procedure code
Code type indicator		1 if record is created from diagnosis 1; 2 if record is created from any other diagnosis code; 3 if record is created from any of the procedure codes
Type of Provider		Set equal to 1

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Health Care Service Record - Institutional (HCSR-I) (cont.)		
Site of Service		Mapping in appendix, based on institution type in HCSR. If institution type is blank, or not on list in appendix, set to 09
Date of Service		End Date of Care
Person ID		Create per Risk Assessment Specs
Length of Stay		
Total RWP		
Procedure Codes (all)		
Diagnosis Codes (all)		
Provider ID		
DRG		
# of Services		
Revenue Code		
Medicare Flag		
Other Insurance Indicator		If OHI paid >0 then set to 1, else 0

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Health Care Service Record – Non-Institutional (HCSR-NI)		
Begin Date of Care (YYYYMMDD)		
End Date of Care (YYYYMMDD)		
Provider Specialty Code		
Provider ID		
Provider Zip/Country Code		
Multiple Provider Suffix		
Place of Service		
Type of Service 1		
Type of Service 2 (Service Nature)		
Catchment Area		
PRISM Area		
Amount Allowed, Completed	234	Apply completion factor
Amount Paid, Completed	243	Apply completion factor
Diagnosis Codes (Primary DX DX1-DX4)	358 - 382	Substring to 5 digits
Lab Flag	586 Proc1-Proc5 (LI)	1=Only lab procedure in this claim 2=Lab procedure along with other procedures 3=No lab procedure
Proc1-Proc5	586 (LI)	
Denie1-Denie5	633 (LI)	No transformation. If Deny reason is non-blank then line item is denied.
Diagnosis/Procedure Code		First 5 characters of diagnosis code or procedure code

(Continued)

Variables	Name on Layout or Position in Original Data File	Transformation Rule
Health Care Service Record – Non-Institutional (HCSR-NI) (cont.)		
Code type indicator		'1' if record is created from diagnosis 1; '2' if record is created from any other diagnosis code; '4' if record is created from any of the procedure codes (including the E&M procedure code)
Type of Provider		Mapping in appendix B, based on provider specialty code. If provider specialty code is blank or not on list in appendix, set to 4.
Site of Service		Mapping in appendix, based on place of service variable in each line item. If place of service is blank or not on list in appendix, set to 4.
Date of Service		End Date of Care
Person ID		Create per Risk Assessment Specs
Procedure Codes (all)		
Diagnosis Codes (all)		
Provider ID		
# of Services		
Medicare Flag		If age ge 64 then medflag=1 else 0
Simple RVU		
Pharmacy Data Transaction Service (PDTS)		
Net Amount Due		
National Mail Order Pharmacy (NMOP		
Total Price	Rx_total_price	

Data Extraction Specification

Point-in-Time Extract (PITE)

Date: FY01-FY02, longitudinal file
Provide longitudinal PITE extraction which includes demographics variable based on year end data (Sept01 and Sept02) or latest month available and number of eligible months for each Occupational Code (Combat, Non-Combat) and Bencat (1, 2, 3, 4).

PITE Data filters:
@403 MHS Elig Indicator (Include if eligible = 1)
@413 Primary record (Include if primary record = 1)
@352 Residence Region (Include if CONUS = Regions 01-12, AK)

TRICARE Enrollment File (TEF)

Date: FY01-FY02 longitudinal file

TEF Data filters:
ACV (Include if beneficiary is Prime at any one point during the year=A, D, E)
Remove ineligibles (see notes below)

Notes:
If PITE monthly file is called Oct01 then its corresponding TEF monthly file is called Nov01. If PITE monthly file is called Nov01 then its corresponding TEF monthly file is called Dec01 and so on.

In order to remove ineligibles from TEF, merge corresponding PITE and TEF monthly files by PID (SponSSN, DOB, Sex). Remove beneficiaries in TEF that are not in its corresponding PITE file.

Keep longitudinal ACV and Enrollment DMISID variables (12 ACVs and 12 Enr DMISIDs for each FY file).

Standard Inpatient Data Record (SIDR)

Date: FY01-FY02

Data filters:
PATREGN (Include CONUS = 01-12, AK)
ACV=prime (Include if Prime=A, D or E)
DEERSENR (Exclude if TPR =79**)
MTF (Exclude if Newport Civilian Hosp=5401)

Apply Completion Factors:
Apply completion factors to Full Costs and Incremental Costs by merging appropriate completion factor tables by FY, FM, and Treatment DMISID. Compute Completed Costs based on formula below. If the Completion Factor = 0 then recode Completion Factor = 1. Output completed costs.

Completed Cost = Raw Cost/CF

Standard Outpatient Data Record (SADR)

Date: FY01-FY02

Data filters:
PATREGN (Include CONUS = 01-12, AK)
ACV=prime (Include if Prime = A, D or E)
ENRDMIS (Exclude if TPR =79**)
MEPRSCD (Include 'B' or 'FBN' codes)

Apply Completion Factors:
Apply completion factors to Full Costs and Variable Costs by merging appropriate completion factor tables by FY, FM, Treatment DMISID and MEPRS Code. Compute Completed Costs based on formula below. If the Completion Factor = 0 then recode Completion Factor = 1. Output completed costs.

Completed Cost = Raw Cost/CF

Health Care Service Record - Institutional (HCSR-I)

Date: FY01-FY02

Data filters:
@198 Enrollment Status (Include if Prime = BB, U, Z)
@317 Submission Code (Exclude if cancelled/denied = C, D, E)
@463 Health Service Region Code (Include if CONUS = 01-12, AK)

Apply Completion Factors:
Apply completion factors to Amount Allowed and Amount Paid. Kennell will provide detail documentation and SAS codes to STI. Output completed costs.

Health Care Service Record – Non-Institutional (HCSR-NI)

Date: FY01-FY02

Data filters:
@198 Enrollment Status (Include if Prime = BB, U, Z)
@317 Submission Code (Exclude cancelled/denied if = C, D, E)
@463 Health Service Region Code (Include if CONUS = 01-12, AK)
@629 Plsvc1-Plsvc25 (Exclude resource sharing if = 26). If the line item is resource sharing then subtract that line item allowed amount from Total Allowed Amount (Position 234),
Apply Completion Factors:
Apply completion factors to Amount Allowed and Amount Paid. Kennell will provide detail documentation and SAS codes to STI. Output completed costs.

Pharmacy Data Transaction Service (PDTS)

Date: FY02

Data filters:
@313 Fill Location (Include if Mail Order = T)
@340 Region (Include if CONUS = 01-12, AK)
@353 ACV (Include if Prime = A, D or E)

National Mail Order Pharmacy (NMOP)

Date: FY01

Data filters:
Input file is '!' delimited
ACV=prime (Include if Prime = A, D or E)
Tricare_region_cd (Include if CONUS = 01-12, AK)

These following four tables were included in the second round of data draws to supplement the 1st run.

Mapping of SADR Provider Specialty Codes to "Type of Provider"

Code	Description	CRG Prov Type
000	General Medical Officer	2
001	Family Practice Physician	2
002	Contract Physician	2
003	Family Practice Physician Resident	2
004	Emergency Physician	2
005	Emergency Physician Resident	2
011	Internist	2
012	Allergist	2
013	Oncologist	2
014	Cardiologist	2
015	Cardiopulmonary Laboratory Physician	2
016	Endocrinologist	2
017	Geriatrician	2
018	Gastroenterologist	2
019	Hematolgoist	2
020	Rheumatolgoist	2
021	Pulmonary Disease Physician	2
022	Infectious Disease Physician	2
023	Metabolic Disease Physician	2
024	Nephrologist	2
025	Medical Geneticist	2
026	Tropical Medicine Physician	2
027	Nuclear Medicine Physician	2
028	Internal Medicine Resident	2
040	Pediatrician	2
041	Pediatric Allergist	2
042	Adolescent Medicine Physician	2
043	Pediatric Cardiologist	2
044	Pediatric Dermatologist	2
045	Pediatric Endocrinologist	2
046	Perinatolgoist	2
047	Pediatric Metabolic Disease Physician	2
048	Pediatric Hematologist	2
049	Pediatric Neurologist	2
050	Pediatric Pulmonary Disease Physician	2
051	Pediatric Infectious Disease Physician	2
052	Pediatric Resident	2
053	Pediatric Gastroenterologist	2
054	Pediatric Nephrologist	2
060	Neurologist	2
061	Neurologist Resident	2
070	Psychiatrist	2
071	Child Psychiatrist	2
072	Psychoanalyst	2
073	Psychiatric Resident	2
074	Alcohol Abuse Counselor	3
075	Drug Abuse Counselor	3
080	Dermatologist	2
081	Dermatologist Resident	2

Code	Description	CRG Prov Type
090	Physical Medicine Physician	2
091	Special Weapons Defense Physician	2
092	Anesthesiologist	2
093	Anesthesiology Resident	2
094	Anesthetist	3
100	General Surgeon	2
101	Thoracic Surgeon	2
102	Colon & Rectal Surgeon	2
103	Cardiac Surgeon	2
104	Pediatric Surgeon	2
105	Peripheral Vascular Surgeon	2
106	Neurological Surgeon	2
107	Plastic Surgeon	2
108	Resident Surgeon	2
109	Burn Therapist	2
110	Urologist	2
111	Urology Resident	2
115	Plastic Surgery Resident	2
120	Ophthalmologist	2
121	Ophthalmology Resident	2
130	Otorhinolaryngologist	2
131	Otorhinolaryngology Resident	2
140	Orthopedic Surgeon	2
141	Hand Surgeon	2
142	Orthopedic Resident	2
150	Obstetrician/Gynecologist (OB/GYN)	2
151	Endocrinologist, OB/GYN	2
152	Oncologist, OB/GYN	2
153	Pathologist, OB/GYN	2
154	OB/GYN Resident	2
200	Pathologist	2
202	Medical Chemist	3
203	Medical Microbiologist	3
204	Forensic Pathologist	3
205	Neuropathologist	2
206	Nuclear Medicine Pathologist	2
207	Pathology Resident	2
208	Histopathologist	2
210	Biomedical Lab Officer	3
211	Biomedical Lab Science Officer	3
212	Microbiology Lab Officer	3
213	Chemistry Lab Officer	3
214	Blood Bank Officer	3
215	Clinical Lab Officer, Other	3
300	Aerospace Medicine Physician	2
301	Aerospace Medicine Resident	2
302	Aerospace Med Flight Surgeon/Family Practice	2
320	Preventive Medicine Physician	2
321	Occupational Medicine Physician	2
322	Hyperbaric/Underseas Medicine Physician	2
400	Radiologist	2
401	Radiation Therapist	3
402	Neuro-Radiologist	2

Code	Description	CRG Prov Type
403	Nuclear Medicine Radiologist	2
404	Diagnostic Radiologist	2
405	Special Procedures Radiologist	2
406	Radiology Resident	2
407	Radiophysicist	3
500	Senior Staff Physician	2
501	Anesthesiology Consultant	2
502	Internal Medicine Consultant	2
503	Pediatric Medicine Consultant	2
504	Neurology Consultant	2
505	Psychology Consultant	2
506	Dermatology Consultant	2
507	Physical Medicine Consultant	2
508	Surgery Consultant	2
509	Urology Consultant	2
510	Ophthalmology Consultant	2
511	Otorhinolarynology Consultant	2
512	Orthopedic Surgery Consultant	2
513	OB/GYN Consultant	2
514	Aerospace Medicine Consultant	2
515	Preventive Medicine Consultant	2
516	Radiology Consultant	2
517	Dental Consultant	2
518	Other Consultant	2
600	Nurse, General Duty	3
601	Mental Health Nurse	3
602	OB/GYN Nurse Practitioner	3
603	Pediatric Nurse Practitioner	3
604	Primary Care Nurse Practitioner Qualified	3
605	Primary Care Nurse Practitioner – Entry	3
606	Aerospace Nurse	3
607	Community Health Nurse	3
608	Certified Nurse Midwife	3
609	Nurse Midwife – Entry Level	3
610	Clinical Nurse- Entry Level for Nurse Practitioner	3
611	Psychiatric Nurse Practitioner	3
612	Nurse Anesthetist	3
700	Other Provider (Officer)	3
701	Aerospace Physiologist	3
702	Clinical Psychologist	2
703	Psychology Worker	3
704	Dietician – Nutritionist	3
705	Occupational Therapist	3
706	Physical Therapist	3
707	Podiatrist	3
708	Optometrist	3
709	Audiologist	3
710	Speech Therapist	3
711	Other Biomedical Specialist	3
713	Contract Chiropractor	3
800	Oral Surgeon	2
801	Oral Surgery Resident	2
802	Periodontist	2
803	Periodontic Resident	2

Code	Description	CRG Prov Type
804	Prosthodontist	2
805	Prosthodontic Resident	2
806	Orthodontist	2
807	Orthodontic Resident	2
808	Oral Pathologist	2
809	Oral Pathology Resident	2
810	Endodontist	2
811	Endodontic Resident	2
812	Dental Officer General	2
813	Dental Officer Resident	2
814	Dental Staff Officer	3
815	Pedodontist	2
816	Pedodontic Resident	2
900	Corpsman/Technician	3
901	Physician Assistant	3
902	Dental Assistant	3
905	Cardiopulmonary Lab Technician	3
910	Adolescent Medicine	3
911	Aerospace Medicine	3
912	Allergy	3
913	Anesthesiology	3
914	Audiology	3
915	Cardiology	3
916	Community Health	3
917	Critical Care Medicine	3
918	Dental	3
919	Dermatology	3
920	Dietetics	3
921	Emergency Medicine	3
922	Endocrinology	3
923	Family Practice/Primary Care	3
924	Gastroenterology	3
925	General Medicine	3
926	Gerontology/Geriatrics	3
927	Gynecology	3
928	Health Benefits	3
929	Hematology	3
930	Immunology	3
931	Infectious Disease	3
932	Internal Medicine	3
933	Laboratory/Pathology	3
934	Medical Genetics	3
935	Metabolic Disease	3
936	Nephrology	3
937	Neonatal/Perinatal Medicine	3
938	Neurology	3
939	Nuclear Medicine	3
940	Nursing	3
941	Nutrition	3
942	OB/GYN	3
943	Ocuupational Health	3
944	Oncology	3
945	Ophthalmology	3
946	Optometry	3

Code	Description	CRG Prov Type
947	Orthopedics	3
948	Otorhinolaryngology	3
949	Pediatrics	3
950	Physical Medicine and Rehabilitation	3
951	Podiatry	3
952	Preventive Medicine	3
953	Psychiatry	3
954	Psychology	3
955	Pulmonary Disease	3
956	Radiology	3
957	Rheumatology	3
958	Social Work	3
959	Surgery	3
960	Physical Therapy	3
961	Radiation Therapy	3
962	Speech Language Pathology Therapy	3
963	Urology	3
964	Obstetrics	3
965	Sleep Disorders	3
966	Occupational Therapy	3
967	Developmental Pediatrics	3
968	Hyperbaric Medicine	3
969	Respiratory Therapy	3
970	Peripheral Vascular Medicine	3
971	Proctology	3
972	Thoracic Surgery	3
999	Unknown	3

Mapping of Purchased Care (HCSR) provider specialty codes to CRG Provider Type

Code	Specialty	CRG Prov Type
01	General Practice	2
02	General Surgery	2
03	Allergy	2
04	Otology, Laryngology, Rhinology	2
05	Anesthesiology	2
06	Cardiovascular Disease	2
07	Dermatology	2
08	Family Practice	2
10	Gastroenterology	2
11	Internal Medicine	2
13	Neurology	2
14	Neurosurgery	2
16	Obstetrics/Gynecology	2
18	Ophthalmology	2
19	Oral Surgery (Dentists only)	2
20	Orthopedic Surgery	2
22	Pathology	2
24	Plastic Surgery	2
25	Physical Medicine and Rehabilitation	2
26	Psychiatry	2
28	Proctology	2
29	Pulmonary Diseases	2
30	Radiology	2
33	Thoracic Surgery	2
34	Urology	2
35	Chiropractor, licensed	3
36	Nuclear Medicine	3
37	Pediatrics	2
38	Geriatrics	2
39	Nephrology	2
40	Neonatology	2
42	Nurses (RN)	3
43	Nurses (LPN)	3
44	Occupational Therapy (OTR)	3
45	Speech Pathologist/Speech Therapist	3
47	Endocrinology	2
48	Podiatry - Surgical Chiropody	3
49	Miscellaneous	4
50	Proctology and Rectal Surgery	2
51	Medical Supply Co	4
57	Certified Prosthetist - Orthotist	2
59	Ambulance Service Supplier	4

Code	Specialty	CRG Prov Type
60	Public Health or Welfare Agencies	4
61	Voluntary Health or Charitable Agencies	4
62	Clinical Psychologist (Billing Independently)	2
64	Audiologists (Billing Independently)	3
65	Physical Therapist (Independent Practice)	3
69	Independent Laboratory (Billing Independently)	4
70	Clinic or other group practice	2
80	Anesthetist	3
81	Dietitian (Deleted 10/25/98)	3
82	Education Specialist	3
83	Nurse, Private Duty	3
84	Physician's Assistant	3
85	Certified Clinical Social Worker	2
86	Christian Science	4
88	Pharmacy	4
90	Nurse Practitioner	3
91	Clinical Psychiatric Nurse Specialist	3
92	Certified Nurse Midwife	3
93	Mental Health Counselor	3
94	Certified Marriage and Family Therapist	3
95	Pastoral Counselor	3
96	Marriage and Family Therapist (Only valid for Connecticut, Massachusetts, New Jersey and New York) (Deleted 10/25/94)	3
97	M.S.W., A.C.S.W. (Deleted 10/25/94)	3
98	Optometrist	3
99	Facility charges - use for facility charges for outpatient services, (e.g.,ambulatory surgery, hospital services)	4
AA	Corporate Provider - Radiation Therapy	4
AB	Corporate Provider - Cardiac Catheterization Clinics	4
AC	Corporate Provider - Freestanding Sleep Disorder Diagnostic Centers	4
AD	Corporate Provider - Independent Physiological Laboratories	4
AE	Corporate Provider - Freestanding Kidney Dialysis Centers	4
AF	Corporate Provider - Freestanding Magnetic Resonance Imaging Centers	4
AG	Corporate Provider - Comprehensive Outpatient Rehabilitation Facilities (CORFs)	4

Code	Specialty	CRG Prov Type
AH	Corporate Provider - Home Health Agencies (HHAs)	4
AI	Corporate Provider - Freestanding Bone Marrow Transplant Centers (American Association of Blood Banks (AABB) Accredited)	4
AJ	Corporate Provider - Home Infusion (Accreditation Commission for Health Care, Inc. (ACHC) Accredited)	4
AK	Corporate Provider - Diabetic Output Self Management Education Program (ADA Accredited)	4
BC	Freestanding Birthing Center	4
CT	Corporate Provider - Temporary	4
GY	Gynecology (GYN)	2
HB	Hospital Outpatient Birthing Room	4
HA	Home Health Care Agency	4
HH	Home Health Aide/Homemaker	4
LS	ALS Mandated	4
TS	Transportation Services (Privately Owned Vehicle)	4

Institution Types in HCSR-I and "Site of Service"

Code	Description	Site
10	General medical and surgical	04
11	Hospital unit of an institution (prison hospital, college infirmary etc.)	04
12	Hospital unit within an institution for the mentally retarded	04
22	Psychiatric hospital or unit of	04
33	Tuberculosis and other respiratory disease	04
44	Obstetrics and gynecology	04
45	Eye, ear, nose and throat	04
46	Rehabilitation	04
47	Orthopedic	04
48	Chronic disease	04
49	Other specialty[1]	04
50	Children's general	04
51	Children's hospital unit of an institution	04
52	Children's psychiatric hospital or unit of	04
53	Children's tuberculosis and other respiratory diseases	04
55	Children's eye, ear, nose, and throat	04
56	Children's rehabilitation	04
57	Children's orthopedic	04
58	Children's chronic	04
59	Children's other specialty[1]	04
62	Institution for mental retardation	04
70	Home Health Care Agency	02
71	Specialized Treatment Facility	04
72	Residential Treatment Center	04
73	Extended Care Facility	04
74	Christian Science Facility	04
75	Hospital based Ambulatory Surgery Center	07
76	Skilled Nursing Facility	08
78	Non-hospital based hospice	03
79	Hospital based hospice	03
82	Substance Use Disorders Rehabilitation Facility (SUDRF)	04
90	Cancer	04
91	Sole community	04
92	Freestanding Ambulatory Surgery Center	07

HCSR-N Place of Service mapping to site of service

Place of Service Code	Site of Service code
11=Office	06
12=Home	02
21=Inpatient Hospital	04
22=Outpatient Hospital	07
23= Emergency Room-Hospital	07
24=Ambulatory Surgery Center	07
25=Birthing Center	04
26=Military Treatment Facility	07
31=Skilled Nursing Facility	08
32=Nursing Facility	05
33=Custodial Care Facility	05
34=Hospice	03
41=Ambulance-Land	09
42=Ambulance-Air or Water	09
51=Inpatient Psych Facility	04
52=Psych Facility Partial Hospitalization	07
53=Community Mental Health Center	09
54=Intermediate Care Fac/Mentally Retarded	05
55=Residential Substance Abuse	05
56=Psych Res Treatment Center	05
61=Comp Inpatient Rehab Facility	04
62=Comp Outpatient Rehab Facility	07
65=End Stage Renal Disease Trt Fac	07
71=State/Local Public Health Clinic	01
72=Rural Health Clinic	01
81=Independent Lab	09
99=Other Unlisted Facility	09

Appendix D: Data Processing—Overview and SAS Logs

Appendix D
Data Processing—Overview and SAS Logs

Most of the data processing was done using SAS 8.2, and this appendix contains the log files from the SAS programs.

To assist understanding of these programs, along with their sequence, the following overview presents a mapping of the SAS programs. A list of page numbers for each attachment may be found on page 9.

Overview

Data processing were broken down into the following steps.

1. Raw data files
2. Select sample for analysis
3. TRICARE Prime utilization costs
4. Input data files for running risk-adjustment models
5. Running risk adjustment models
    a. CDPS Model
    b. ACG Model
    c. DCG Model
    d. CRG Model
6. Obtaining individual expenditure predictions from the risk adjustment models
7. Miscellanous

1. Raw data files

We obtained a total of 7 data files for FY 2001 and another 7 for FY 2002 (all in SAS xpt format). For each year there is 1 beneficiary information file (pben), 1 enrollment file (lenr) and five health care utilization files (sidr, sadr, hcsri, hcsrni and nmop01/pdts02)[1].

2. Select sample for analysis

Out of approximately 4.5 million TRICARE Prime enrollees in each of FY 2001 and FY 2002, we selected 2.3 million who were continuously enrolled during both years and who were aged 64 or younger as of September 2001.

Attachment A1 is the program log for this sample selection (task1_sample.log). This produces the base version of the main individual-level analytic data set (dod1.sas7bdat) with several individual-level character variables from the enrollment data file (age, gender, rank, region, service, etc). The corresponding data file for the excluded population is called excluded.sas7bdat.

[1] Costs of prescriptions from military providers are broken into mail order and non-mail order sources. The non-mail order fill costs are also included in SADR claims (FY01, FY02), while the mail order fills for FY01 are in NMOP and for FY02 they will be in PDTS. And for prescriptions from civilian providers, the non-mail order fill costs are already in HCSR-NI (FY01, FY02), while the mail order fills for FY01 are in NMOP and for FY02 they will be in PDTS.

The number of Prime enrollees in the raw data files (TRICARE Enrollment Files) for FY2001 and FY2002 and the sub-sample that is selected for our analysis are provided below:

Prime enrollees in FY 2001	4,486,060
Prime enrollees in FY 2002	4,645,181
Prime enrollees in both years	3,951,194
Prime enrollees for all 24 months	2,610,041
Prime enrollees for all 24 months & are in Sept-2001 beneficiary file	2,307,187
Prime enrollees for all 24 months, are in Sept-2001 beneficiary file & are aged 64 or younger as of Sept 2001	***2,304,926***

3. TRICARE Prime utilization costs

In this step the health care utilization files are processed for the selected sample to obtain individual level annual TRICARE expenditures for FY2001 and FY 2002—variables giving breakdown of the total cost by file-specific total costs (SIDR, SADR, HCSRI and HCSRN) are also created.

Attachment A2 has the log of the SAS program used for these computations. These cost variables are added to the main analytic data file (dod2.sas7bdat).

4. Input data files for running risk-adjustment models

The four risk models selected for this study vary in specific format of the input data files. Software for three of these models (CDPS, ACG and DCG) were obtained and implemented by the research team at Boston University. Another similarity in these three models is that they are identical in the type of information used—individual annual costs for FY 2002, and individual age, gender and diagnoses codes (ICD-9) for FY 2001. The fourth model (CRG) was implemented by the software producer (3M Health Information Systems)—this model in addition uses information on procedure codes as well as dates of care (admission/discharge).

CDPS

Attachment A3_CDPS (task4b_cdpsdata.log) is the log of the SAS program used to create the two input data files used to run the CDPS model.

ACG

Attachment A3_ACG (task4c_acgdata.log) is the log of the SAS program used to create the two input data files used to run the ACG model.

DCG

Attachment A3_DCG (dcgfiles.log) is the log of the SAS program used to create the two input data files used to run the ACG model.

CRG

Attachment A3_CRG_1 is the CRG data specification that was sent to Kennell for cutting the data properly. The objectives of this data run were two-fold; to provide data to 3M to be used in their CRG grouping software, and to provide the Risk Assessment team with the same data, but without the specific formatting details required by the groupers. Essentially, what this meant was that several data elements were added to the Risk Assessment Data, and that a special output was prepared from the new Risk Assessment files for CRG grouping purposes.

Attachment A3_CRG_2 is the SAS log file for the creation of the finder file sent to Kennell for selecting the 1.8 million fitting sample and 0.5 million validation sample.

5. Running risk adjustment models

While CDPS and DCG risk models are SAS programs, the other risk models are stand alone software. All the three models run at Boston University were fast (all runs completed in less than 30 minutes). The primary output from each model is a vector of dichotomous variables denoting the presence/absence of each of the conditions grouped by the risk model. For instance, the DCG model produces a vector of 184 variables corresponding to the 184 HCCs.

Each model involved specification of certain settings. In the case of CDPS and DCG, the Boston University team consulted with the developers on the selection of appropriate settings. The models also differ also in the extent of secondary results produced. These details relating to each model are included in this section.

The CRG model was implemented by 3M Health Information Systems. The description of the technical details of model implementation provided by 3M is included in the CRG section below.

CDPS

The latest model of the CDPS software (Version 2.0) was downloaded from the developer website (http://medicine.ucsd.edu/fpm/cdps/) in June 2004. The software is a set of SAS programs. We experienced some difficulties in running the program based on documentation instructions. But these were resolved in consultation with one of the software authors, Dr. Todd Gilmer. **Attachment A4_CDPS_1** (task5b_cdps.log) is the SAS program that was run. **Attachment A4_CDPS_2** (task5b_cdps.lst) is the SAS secondary results file that gives the CDPS condition dichotomous variables and their frequency across the sample population.

This model program does not give any details on the proportion of all diagnoses that were actually utilized. Nor does it give any details on potentially suspicious diagnoses (male pregnancies, etc).

ACG

The Boston University team obtained Version 6.0 of the ACG Model software (Windows version) from developers at Johns Hopkins University. We consulted with Chad Abrams (Johns Hopkins University) at various stages—especially in selecting the appropriate model settings. On his advice we ran the acgPM (Prediction Model) version of the software. Results from the model were shared with him. **Attachment A4_ACG_1** (task5c_acg.txt) is the command file (text format) that details the model options selected.

This model produces a lot of useful summary information on the count of all diagnoses in the input file, what proportion were accepted by the model and also a count of the suspicious diagnoses (along with descriptions). These details are in **Attachment A4_ACG_2** (dod2acgbprn.txt).

DCG

The Boston University team obtained Version 6.1 (Windows version) from the developers of DxCG. This software provides a template for the SAS program necessary to run the program. The DxCG manual provides clear and straightforward instructions. We consulted with Elizabeth Bertuccini at DxCG with the few questions which arose at various stages.

Specific models can be fit using command options described in the manual. Some of these options included the level of diagnostic detail (184 disease cohorts), the use of a hierarchy for the disease cohorts, the model purpose (explanatory or payment model), the inclusion of inpatient or all-encounter data, and the type of medical expenditures. Further information about the options we chose can be found in **Attachment A4_DCG_1.** There are additional options that we did not use for this project which may be helpful in future including different levels of top-coded expenditures.

Furthermore, several tables are produced from DCG to summarize information such as the diagnosis rates for the 30 and 184 disease group categories. The “report” and “appendix” text files created from the SAS program are used to create these Excel tables. The “appendix” text file also provides information about the number of diagnoses in the input file as well as any suspicious or numerically invalid diagnoses. More specific details about the Excel tables are found in the DCG manual. A copy of the tables produced during this project may be found in **Attachment A4_DCG_2.**

CRG[2]

Data

The goal of this project is to take clinically relevant claim data (diagnoses and procedures) incurred in one year (FY 2001) and, supplemented with demographic information, predict costs in the following year (FY 2002). The population consisted of two components, development (calibration) and validation. According to the documentation, the population consisted of 2,304,926 individuals split between these two groups (1,804,926 for development and 500,000 for validation). These individuals had two continuous years of enrollment and were less than 65 years old. Individuals assigned to the development and validation groups were identified with a pair of finder files.

The file received included data on 5,179,977 individuals or just under 3,000,000 more cases than expected. In addition, depending on how age is calculated, about 1% of the population was older than 64. The data also included diagnosis and procedure data from both FY2001 and FY2002.[3]

The data were edited as follows:

- All individuals not identified in the two finder files were excluded.
- All diagnoses and procedures from FY 2002 were excluded from the analysis as that would change the analysis from prospective to concurrent.
- The issue of age was not addressed beyond recoding birth dates to ensure that they fell into the proper range if the individual was too young.

Risk Group Assignment

Risk groups were assigned using 3M Clinical Risk Grouping Software (CRGs), Version 1.3. Version 1.3 is the latest release of CRGs. It was formally released in June, 2005.

The CRG software assigns each individual to a single mutually exclusive group. These groups, CRGs, are aggregated by steps into broader groups, the ACRG1, ACRG2, ACRG3, and Status for reporting and other purposes. Version 1.3 has three grouping schema. The schema are referred to as the CRGs, PCRGs, and QCRGs. The first , the CRGs, is the standard grouping methodology. The second and third schema are variants of that methodology. These variants are optimized for different applications. The PCRGs are optimized for prospective applications, such as this project. The QCRGs are optimized for retrospective or concurrent applications. Like the CRGs, the PCRGs and QCRGs have their own ACRGs.

As this project is a prospective analysis, the PCRGs were used (See PCRG_List.xls for a list of PCRGs). There are 1,098 PCRGs. The PCRGs are aggregated

[2] The information about the CRG model was provided to the authors by 3M Health Information Systems
[3] Upon receipt of the dataset from 3M Health Information Systems (VALIDATION_DATA_OUT.TXT) the unique subject identification numbers were checked against the input datset held at B.U. and found to be identical.

in steps into 440 groups (ACRG1 or PACRG1), then 176 groups (ACRG2 or PACRG2), then 48 Groups (ACRG3 or PACRG3), and finally 9 groups (Status). The nine statuses are:

Status	Description
1	Healthy
2	History Of Significant Acute Disease
3	Single Minor Chronic Disease
4	Minor Chronic Disease In Multiple Organ Systems
5	Single Dominant Or Moderate Chronic Disease
6	Significant Chronic Disease In Multiple Organ Systems
7	Dominant Chronic Disease In Three Or More Organ Systems
8	Dominant, Metastatic, And Complicated Malignancies
9	Catastrophic Conditions

Individuals were assigned to the following demographic categories based on their age and sex:

Demographic Categories		
Age	Male	Female
age <10	M02	F02
age 10 - 17	M03	F03
age 18 - 24	M04	F04
age 25 - 34	M05	F05
age 35 - 44	M06	F06
age 45 - 54	M07	F07
age 55 - 64	M08	F08
age 65 +	M09	F09

Please note that categories M01 and F01 are usually assigned to individuals less than one year old at the end of the period used for assigning groups. For the purposes of this analysis these M01 and F01 were merged Mo2 and FO2 (age 1 – 9) to form categories for individuals < 10 years old.

Weight and Cost Estimate Calculation

The project design required that projections be made using three sets of total costs; uncapped, capped at $25,000, and capped at $50,000. A set of weights was calculated for each of the three sets of costs using the calibration data. In addition, three sets of demographic adjustments intended to be used with the weight sets were also calculated.

For CRGs, regardless of grouping schema, weights are normalized averages, albeit subject to adjustment if cell volume is too low and to ensure monotonicty. A normalized average is simply the average cost per group divided by the overall average cost of the population. Therefore, the average weight for the population equals 1.0.

Demographic adjustments for a given demographic schema are calculated after the group weights are calculated. The demographic adjustments measure the impact of demographic factors net of the effect of risk group assignment. Demographic adjustments are calculated independently for each status. Therefore, the adjustments for Status 1 are not the same as those used for Status 5. Typically, as was done here, demographic factors are not calculated for Statuses 7, 8, and 9. This is done by setting the demographic factors to 1.0. The resultant weights are then normalized so that the average weight of the population equals 1.0.

To convert weights into cost estimates, individuals are assigned weights based on their CRG and, if appropriate, their demographic group. These weights are renormalized so that the average weight of the population for which the cost estimate is being calculated equals 1.0. The individuals weights are then multiplied by an estimate of the average cost of the population to obtain a predicted cost. For this analysis, the average cost was defined as the average cost of the calibration data set for each of the three cap levels (none, $25,000, and $50,000).

Output Data Set

Each individual in the validation data set was assigned a PCRG and demographic category. Based on these assignments, the individual was assigned PCRG weights and demographic factors. These were then normalized. Cost estimates were calculated by multiplying the weights by the projected group average as discussed above. Each individual has six costs estimates (one for each of the three caps, with and without demographic adjustments). The results were placed into a fixed format file, **VALIDATION_DATA_OUT.TXT**. This file has the following format:

@1	ID	$18.
@20	DOB	YYMMDD10.
@33	sex	$1.
@36	PCRG	$5.
@44	demo_cat	$3.
@50	ptot_weight	14.10
@65	p25_weight	14.10
@80	p50_weight	14.10
@95	ptot_weight_demo	14.10
@110	p25_weight_demo	14.10
@125	p50_weight_demo	14.10
@140	ptot_cost_multiplier	9.2
@150	p25_cost_multiplier	9.2
@160	p50_cost_multiplier	9.2
@170	eptot	9.2
@180	ep25	9.2
@190	ep50	9.2
@200	eptot_demo	9.2
@210	ep25_demo	9.2
@220	ep50_demo	9.2

The fields are defined as follows:

Field	Description
ID	Unique identifier
DOB	Date of birth
Sex	Sex (1 = male, 2 = female)
PCRG	PCRG
demo_cat	Demographic category
ptot_weight	PCRG weight (uncapped)
p25_weight	PCRG weight ($25,000 cap)
p50_weight	PCRG weight ($50,000 cap)
ptot_weight_demo	PCRG weight (uncapped) adjusted for demographic characteristics
p25_weight_demo	PCRG weight ($25,000 cap) adjusted for demographic characteristics
p50_weight_demo	PCRG weight ($50,000 cap) adjusted for demographic characteristics
ptot_cost_multiplier	Average total cost (uncapped) of calibration sample
p25_cost_multiplier	Average total cost ($25,000 cap) of calibration sample
p50_cost_multiplier	Average total cost ($50,000 cap) of calibration sample
Eptot	Expected cost based on PCRG - uncapped
ep25	Expected cost based on PCRG - $25,000 cap
ep50	Expected cost based on PCRG - $50,000 cap
eptot_demo	Expected cost based on PCRG and demographic category – uncapped
ep25_demo	Expected cost based on PCRG and demographic category - $25,000 cap
ep50_demo	Expected cost based on PCRG and demographic category - $50,000 cap

Attachment A4_CRG is the table of PCRG numbers and their description provided by 3M Health Information Systems.

6. Obtaining individual expenditure predictions from the risk adjustment models

The group of dichotomous variables created by group risk model are then summarized into a single summary score. As explained in the Methods section, this score is the projected TRICARE expenditures for FY2002 given gender, age and FY2001 diagnoses. The exception is the CRG model where in addition to gender, age and FY2001 diagnoses, data on procedure codes as well as procedure dates are used. Following gives the SAS log of this step for each model:

- CDPS – **Attachment A5_CDPS** (task6b_cdpspred.log)
- ACG – **Attachment A5_ACG** (task6c_acgpred.log)
- DCG – **Attachment A5_DCG** (task6_dcgpred.log)
- CRG – **Attachment A5_CRG** (task7_dod3valid5.log)

7. Miscellaneous

An assortment of other processing jobs were performed to create the final analytic dataset.

- **Attachment A6_1** (task6d_dod3valid.log): This program consolidates all the variables created thus far for the 0.5 million validation sample. This dataset is called dod3valid.sas7bdat.
- **Attachment A6_2** (task6e_dod3dxgr.log): For analysis by subgroups with specific diagnoses, this program creates 9 diagnoses condition indicators. The updated dataset is called dod3valid2.sas7bdat.
- **Attachment A6_3** (task6f_agesexpred.log): For comparison with predictions from risk adjustment models, this program obtains prediction of a model based only on gender and age (i.e., no adjustment for illness risk). The updated dataset is called dod3valid3.sas7bdat.
- **Attachment A6_4** (task6f_pcmtype.log): This job creates a field that identifies whether the Primary Care Manager (PCM) is in military or civilian sector. The updated dataset is called dod3valid4.sas7bdat.
- **Attachment A6_5** (task6g_pcmtype_exclude.log): This program creates the aforementioned field (PCM military or civilian) for the excluded sample (excluded.sas7bdat).

List of the attachments follwing and the pages they may be found on.

Attachment	Appendix Page
A7_2 (task6e_dod3dxgr.log)	260-265
A7_3 (task6f_agesexpred.log)	266-273
A7_4 (task6f_pcmtype.log)	274-277
A7_5 (task6g_pcmtype_exclude.log)	278-290

Attachment A1
(task1_sample.log)

```
1    The SAS System                                22:41 Tuesday, July 13, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.99 seconds
      cpu time            0.07 seconds

1          /*
2          /data/dod1/saswork/task1_sample.sas
3          Amresh Hanchate
4          DOD1 Project
5          June 29, 2004
6
7          Selects the sample of analytical interest, namely, those enrolled in
TRICARE Prime for all the 24 months (FY 2001 and
7        ! 2002) & and those aged 64 or younger; produces two data sets
8
9          dod1 - selected
10         excluded - excluded
11         */
12
13         options ps=52 ls=72 nocenter;
14         libname rawdata '/data/dod1/rawdata/';
NOTE: Libref RAWDATA was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/rawdata
15         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
16         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: The format name '$REBENCATF' exceeds 8 characters. Only the first
      8 characters will be used.
NOTE: Format $REBENCA has been output.
NOTE: Format $RESREG has been output.
NOTE: Format $RECSVC has been output.
```

```
NOTE: PROCEDURE FORMAT used:
      real time           0.31 seconds
      cpu time            0.02 seconds


43         footnote '/data/dod1/saswork/task1_sample.sas';
44
45         *using enrollment for FY2001 and FY2002 create subset of
45       ! those with at least one month of enrollment each year;
46         libname xpt1 xport '/data/dod1/rawdata/lenr01.xpt';
NOTE: Libref XPT1 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/lenr01.xpt
47         proc sort data=xpt1.lenr01 out=lenr01; by pid; run;

NOTE: There were 4486060 observations read from the data set
      XPT1.LENR01.
NOTE: The data set WORK.LENR01 has 4486060 observations and 25
      variables.
NOTE: PROCEDURE SORT used:
      real time           7:43.31
      cpu time            2:50.63


48
49         libname xpt2 xport '/data/dod1/rawdata/lenr02.xpt';
NOTE: Libref XPT2 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/lenr02.xpt
50         proc sort data=xpt2.lenr02 out=lenr02; by pid; run;

NOTE: There were 4645181 observations read from the data set
      XPT2.LENR02.
NOTE: The data set WORK.LENR02 has 4645181 observations and 25
      variables.
NOTE: PROCEDURE SORT used:
      real time           8:08.05
      cpu time            2:56.48


51
52         data lenr;
53           merge lenr01 (in=one) lenr02 (in=two);
54           by pid;
55           if one & two;
56         run;

NOTE: There were 4486060 observations read from the data set
      WORK.LENR01.
```

```
3 The SAS System                              22:41 Tuesday, July 13, 2004

NOTE: There were 4645181 observations read from the data set
      WORK.LENR02.
NOTE: The data set WORK.LENR has 3951194 observations and 49 variables.
NOTE: DATA statement used:
      real time           5:01.67
      cpu time            1:38.11


57
58         *calculate months of Prime enrollment in FY2001 and in FY2002
58       ! ;
59         data lenr;
60           set lenr;
61           prime01=acv0101||acv0102||acv0103||acv0104||acv0105||acv0106
61       ! ||
62           acv0107||acv0108||acv0109||acv0110||acv0111||acv0112;
63           * months enrolled in A or D or E - by removing others (G, L,
63       !  U or blank);
64           ckprime1=compress(prime01,"GLU ");
65           primemths1 = length(ckprime1);
66           *make sure no non-ADE entries - should be all 0s;
67           ifnonade1 = verify(ckprime1, "ADE ");
68
69           prime02=acv0201||acv0202||acv0203||acv0204||acv0205||acv0206
69       ! ||
70           acv0207||acv0208||acv0209||acv0210||acv0211||acv0212;
71           ckprime2=compress(prime02,"GLU ");
72           primemths2 = length(ckprime2);
73           ifnonade2 = verify(ckprime2, "ADE ");
74         run;

NOTE: There were 3951194 observations read from the data set WORK.LENR.
NOTE: The data set WORK.LENR has 3951194 observations and 57 variables.
NOTE: DATA statement used:
      real time           8:36.25
      cpu time            2:07.62


75         *nubmer of months of enrollment each year;
76         proc freq;
77         tables ifnonade1 ifnonade2 primemths1 primemths2
77       ! primemths1*primemths2 / missing;
78         run;

NOTE: There were 3951194 observations read from the data set WORK.LENR.
NOTE: The PROCEDURE FREQ printed pages 1-5.
NOTE: PROCEDURE FREQ used:
      real time           34.83 seconds
      cpu time            30.57 seconds
```

```
79
80          *select those enrolled 12 months each year -- also create a
80        ! file for the rest;
81          data sample sample2;
82            set lenr (keep = pid primemths1 primemths2);
83            if primemths1=12 & primemths2=12 then output sample;
84            else output sample2;
85          run;

NOTE: There were 3951194 observations read from the data set WORK.LENR.
NOTE: The data set WORK.SAMPLE has 2610041 observations and 3 variables.
NOTE: The data set WORK.SAMPLE2 has 1341153 observations and 3
      variables.
NOTE: DATA statement used:
      real time           1:56.49
      cpu time            29.69 seconds


86
87          *add individual characteristics from beneficiary (pben) file
87        ! for FY2001 (i.e., as of Sept 2001);
88          libname xpt3 xport '/data/dod1/rawdata/pben01.xpt';
NOTE: Libref XPT3 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/pben01.xpt
89          proc sort data=xpt3.pben01 (keep=pid zip3 rebencat catcode
89        ! recsvc rank recocc resreg sex race age urban) out=pben01; by
89        ! pid; run;

NOTE: There were 7649132 observations read from the data set
      XPT3.PBEN01.
NOTE: The data set WORK.PBEN01 has 7649132 observations and 12
      variables.
NOTE: PROCEDURE SORT used:
      real time           11:07.75
      cpu time            5:03.93


90          proc sort data=sample; by pid; run;

NOTE: There were 2610041 observations read from the data set
      WORK.SAMPLE.
NOTE: The data set WORK.SAMPLE has 2610041 observations and 3 variables.
NOTE: PROCEDURE SORT used:
      real time           2:23.55
      cpu time            35.06 seconds
```

```
91         proc sort data=sample2; by pid; run;

NOTE: There were 1341153 observations read from the data set
      WORK.SAMPLE2.
NOTE: The data set WORK.SAMPLE2 has 1341153 observations and 3
      variables.
NOTE: PROCEDURE SORT used:
      real time           1:25.55
      cpu time            17.56 seconds


92         * first match for the selected sample;
93         data sample;
94           merge sample (in=one) pben01 (in=two);
95           by pid;
96           drop primemths1 primemths2;
97           *create grouped age variable (age2);
98           if (age LT 15) & (age NE .) then age2 = "14 Or under";
99           else if (age GE 15) & (age LT 25) then age2 = "15 to 24";
100          else if (age GE 25) & (age LT 45) then age2 = "25 to 44";
101          else if (age GE 45) & (age LT 65) then age2 = "45 to 64";
102          if one & two;
103        run;

NOTE: There were 2610041 observations read from the data set
      WORK.SAMPLE.
NOTE: There were 7649132 observations read from the data set
      WORK.PBEN01.
NOTE: The data set WORK.SAMPLE has 2307187 observations and 13
      variables.
NOTE: DATA statement used:
      real time           1:44.41
      cpu time            1:26.80


104        *now match those in sample2 with pben01;
105        data sample2;
106          merge sample2 (in=one) pben01 (in=two);
107          by pid;
108          *create grouped age variable (age2);
109          if (age LT 15) & (age NE .) then age2 = "14 Or under";
110          else if (age GE 15) & (age LT 25) then age2 = "15 to 24";
111          else if (age GE 25) & (age LT 45) then age2 = "25 to 44";
112          else if (age GE 45) & (age LT 65) then age2 = "45 to 64";
113          if one & two;
114        run;

NOTE: There were 1341153 observations read from the data set
```

```
      WORK.SAMPLE2.
NOTE: There were 7649132 observations read from the data set
      WORK.PBEN01.
NOTE: The data set WORK.SAMPLE2 has 1118039 observations and 15
      variables.
NOTE: DATA statement used:
      real time           1:42.43
      cpu time            1:19.09


115
116        *now select those aged 64 or younger -- for both sample and
116      ! sample2;
117        data saswork.dod1; set sample; if age GE 0 & age LT 65; run;

NOTE: There were 2307187 observations read from the data set
      WORK.SAMPLE.
NOTE: The data set SASWORK.DOD1 has 2304926 observations and 13
      variables.
NOTE: DATA statement used:
      real time           1:14.11
      cpu time            17.23 seconds


118        data saswork.excluded; set sample2; if age GE 0 & age LT 65;
118      ! run;

NOTE: There were 1118039 observations read from the data set
      WORK.SAMPLE2.
NOTE: The data set SASWORK.EXCLUDED has 1079613 observations and 15
      variables.
NOTE: DATA statement used:
      real time           43.79 seconds
      cpu time            9.59 seconds


119
120
121
122
123

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           52:24.77
      cpu time            19:42.53
```

Attachment A2
(task2_costs.log)

```
1    The SAS System                                  08:45 Friday, July 16, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.33 seconds
      cpu time            0.08 seconds

1          /* task2_costs.sas
2          Amresh Hanchate
3          DOD1 Project
4
5          Creates individual-level costs -- total as well as by each file type
(SIDR, SADR, etc), and for each year (FY2001 and
5        ! FY2002).
6          This is done only for the selected sample population in
dod1.sas7bdat (those continuously enrolled in Prime for 24 months
6        !  & aged 64 or younger). The new data file is called dod2.
7          NOTES:
8            1) if a person does not appear in a file at all then it is assumed
that no health care expenditures were incurred; these
8        !  are identified by missing entries;
9            2) persons who enter utilization files but the records have
missing expenditures, then these are set to zero; these too
9        ! are identified by missing entries; NOTE that this assumption mainly
affects the SADR files (about 1%) -- Wendy believes
9        ! that most of th
10           3) there are very few negative expenditure records; these are
treated as zero costs when calculating individual level
10       ! costs;
11
12         */
13         options ps=52 ls=80 mprint;
14
15         libname rawdata '/data/dod1/rawdata/';
NOTE: Libref RAWDATA was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/rawdata
16         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
17
18         *following macro takes record level data set and creates a subset
18       ! with only the PIDs of interest -- the PIDs of interest are in
PIDFILE
```

```
18       ! ;
19         %MACRO SUBSET(INSASDATA, PIDFILE, OUTSASDATA);
20               PROC SORT DATA=&INSASDATA; BY PID; RUN;
21               PROC SORT DATA=&PIDFILE; BY PID; RUN;
22               DATA &OUTSASDATA;
23                  MERGE &INSASDATA (IN=INFIRST) &PIDFILE (IN=INSECOND);
24                  BY PID;
25                  IF INFIRST & INSECOND;
26               RUN;
27         %MEND SUBSET;
28
29         *assumes insasdata1 has only the subset of interest - creates indiv
29       ! level outvar and adds it to existing insasdata2 - outfile called
29       ! outsasdata;
30         %MACRO INDIVCOST(INSASDATA1, INSASDATA2,OUTSASDATA,INVAR,OUTVAR);
31               DATA TEMP;
32                  SET &INSASDATA1 (KEEP=PID &INVAR);
33                  BY PID;
34                  RETAIN &OUTVAR;
35                  IF FIRST.PID THEN &OUTVAR = MAX(0,&INVAR);
36                  ELSE &OUTVAR = SUM(&OUTVAR, MAX(0,&INVAR));
37                  IF LAST.PID;
38               RUN;
39               *attach temp to insasdata2 with output called outsasdata;
40               PROC SORT DATA=&INSASDATA2; BY PID; RUN;
41               PROC SORT DATA=TEMP (KEEP=PID &OUTVAR); BY PID; RUN;
42               DATA &OUTSASDATA;
43                  MERGE &INSASDATA2 (IN=INFIRST) TEMP (IN=INSECOND);
44                  BY PID;
45                  IF INFIRST;
46               RUN;
47         %MEND INDIVCOST;
48
49         * define macro that counts negatives, zeros, positives and missing;
50         %MACRO COUNTSIGN(INSASDATA,NUMVAR);
51               DATA WORK.TEMP;
52                     SET &INSASDATA (KEEP=&NUMVAR);
53                     length vartype $10.;
54                     if &numvar=. then vartype='.=Missing';
55                     else if &numvar<0 then vartype='1=Negative';
56                     else if &numvar=0 then vartype='2=Zero';
57                     else if &numvar>0 then vartype='3=Positive';
58               PROC FREQ;
59                     TABLES VARTYPE;
60               RUN;
61         %MEND COUNTSIGN;
62
63
64         *pid file of population of interest;
65         proc sort data=saswork.dod1 (keep=pid) out=pidfile; by pid; run;
```

```
NOTE: There were 2304926 observations read from the data set SASWORK.DOD1.
NOTE: The data set WORK.PIDFILE has 2304926 observations and 1 variables.
NOTE: PROCEDURE SORT used:
      real time           1:11.85
      cpu time            26.64 seconds


66
67         *get sidr01 file;
68         libname xpt1 xport '/data/dod1/rawdata/sidr01.xpt';
NOTE: Libref XPT1 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sidr01.xpt
69         data sidr01;
70           set xpt1.sidr01 (keep=pid fullcost);
71         *calculate record level cost univariate;
72         %subset(sidr01,pidfile,sidr01s);

NOTE: There were 141226 observations read from the data set XPT1.SIDR01.
NOTE: The data set WORK.SIDR01 has 141226 observations and 2 variables.
NOTE: DATA statement used:
      real time           4.76 seconds
      cpu time            1.78 seconds


MPRINT(SUBSET):   PROC SORT DATA=sidr01;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 141226 observations read from the data set WORK.SIDR01.
NOTE: The data set WORK.SIDR01 has 141226 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           3.32 seconds
      cpu time            1.93 seconds


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.04 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA sidr01s;
MPRINT(SUBSET):   MERGE sidr01 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
```

```
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 141226 observations read from the data set WORK.SIDR01.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.SIDR01S has 78119 observations and 2 variables.
NOTE: DATA statement used:
      real time           14.47 seconds
      cpu time            14.10 seconds


73         proc univariate data=sidr01s;
74           var fullcost;
75         run;

NOTE: There were 78119 observations read from the data set WORK.SIDR01S.
NOTE: The PROCEDURE UNIVARIATE printed pages 1-2.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.78 seconds
      cpu time            0.53 seconds


76         %countsign(sidr01s, fullcost);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET sidr01s (KEEP=fullcost);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if fullcost=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if fullcost<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if fullcost=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if fullcost>0 then vartype='3=Positive';

NOTE: There were 78119 observations read from the data set WORK.SIDR01S.
NOTE: The data set WORK.TEMP has 78119 observations and 2 variables.
NOTE: DATA statement used:
      real time           1.35 seconds
      cpu time            0.33 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 78119 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 3.
NOTE: PROCEDURE FREQ used:
      real time           0.32 seconds
      cpu time            0.17 seconds


77         *calculate indiv level cost univariate;
```

```
78         %indivcost(sidr01s, pidfile, costfile, fullcost, cost_sidr01);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET sidr01s (KEEP=PID fullcost);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_sidr01;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_sidr01 = MAX(0,fullcost);
MPRINT(INDIVCOST):   ELSE cost_sidr01 = SUM(cost_sidr01, MAX(0,fullcost));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 78119 observations read from the data set WORK.SIDR01S.
NOTE: The data set WORK.TEMP has 66453 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.79 seconds
      cpu time            0.48 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=pidfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_sidr01);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 66453 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 66453 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.85 seconds
      cpu time            0.58 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE pidfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: There were 66453 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 2 variables.
NOTE: DATA statement used:
      real time           33.27 seconds
```

```
      cpu time            19.27 seconds


79
80         *get sadr01 file;
81         libname xpt2 xport '/data/dod1/rawdata/sadr01.xpt';
NOTE: Libref XPT2 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sadr01.xpt
82         data sadr01;
83           set xpt2.sadr01 (keep=pid fcost);
84         *calculate record level cost univariate;
85         %subset(sadr01,pidfile,sadr01s);

NOTE: There were 20691247 observations read from the data set XPT2.SADR01.
NOTE: The data set WORK.SADR01 has 20691247 observations and 2 variables.
NOTE: DATA statement used:
      real time           10:42.96
      cpu time            3:31.60


MPRINT(SUBSET):   PROC SORT DATA=sadr01;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 20691247 observations read from the data set WORK.SADR01.
NOTE: The data set WORK.SADR01 has 20691247 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           18:03.35
      cpu time            7:17.38


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.04 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA sadr01s;
MPRINT(SUBSET):   MERGE sadr01 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 20691247 observations read from the data set WORK.SADR01.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
```

```
NOTE: The data set WORK.SADR01S has 12776227 observations and 2 variables.
NOTE: DATA statement used:
      real time           3:57.15
      cpu time            2:36.80


86         proc univariate data=sadr01s;
87           var fcost;
88         run;

NOTE: There were 12776227 observations read from the data set WORK.SADR01S.
NOTE: The PROCEDURE UNIVARIATE printed pages 4-5.
NOTE: PROCEDURE UNIVARIATE used:
      real time           1:17.70
      cpu time            1:17.50


89         %countsign(sadr01s, fcost);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET sadr01s (KEEP=fcost);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if fcost=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if fcost<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if fcost=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if fcost>0 then vartype='3=Positive';

NOTE: There were 12776227 observations read from the data set WORK.SADR01S.
NOTE: The data set WORK.TEMP has 12776227 observations and 2 variables.
NOTE: DATA statement used:
      real time           2:36.77
      cpu time            50.80 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 12776227 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 6.
NOTE: PROCEDURE FREQ used:
      real time           28.25 seconds
      cpu time            28.06 seconds


90         *calculate indiv level cost univariate;
91         %indivcost(sadr01s, costfile, costfile, fcost, cost_sadr01);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET sadr01s (KEEP=PID fcost);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_sadr01;
```

```
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_sadr01 = MAX(0,fcost);
MPRINT(INDIVCOST):   ELSE cost_sadr01 = SUM(cost_sadr01, MAX(0,fcost));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 12776227 observations read from the data set WORK.SADR01S.
NOTE: The data set WORK.TEMP has 1718520 observations and 3 variables.
NOTE: DATA statement used:
      real time           2:08.84
      cpu time            56.77 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1:41.11
      cpu time            29.91 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_sadr01);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 1718520 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 1718520 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1:06.36
      cpu time            21.72 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 1718520 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 3 variables.
NOTE: DATA statement used:
      real time           44.71 seconds
      cpu time            26.10 seconds


92
```

```
9    The SAS System                              08:45 Friday, July 16, 2004

93         *get hcsri01 file;
94         libname xpt3 xport '/data/dod1/rawdata/hcsri01.xpt';
NOTE: Libref XPT3 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/hcsri01.xpt
95         data hcsri01;
96           set xpt3.hcsri01 (keep=pid allowed);
97         *calculate record level cost univariate;
98         %subset(hcsri01,pidfile,hcsri01s);

NOTE: There were 144094 observations read from the data set XPT3.HCSRI01.
NOTE: The data set WORK.HCSRI01 has 144094 observations and 2 variables.
NOTE: DATA statement used:
      real time           4.57 seconds
      cpu time            1.65 seconds


MPRINT(SUBSET):   PROC SORT DATA=hcsri01;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 144094 observations read from the data set WORK.HCSRI01.
NOTE: The data set WORK.HCSRI01 has 144094 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           3.50 seconds
      cpu time            1.87 seconds


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.01 seconds


MPRINT(SUBSET):   DATA hcsri01s;
MPRINT(SUBSET):   MERGE hcsri01 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 144094 observations read from the data set WORK.HCSRI01.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.HCSRI01S has 60355 observations and 2 variables.
NOTE: DATA statement used:
      real time           16.95 seconds
      cpu time            14.45 seconds
```

```
99         proc univariate data=hcsri01s;
100          var allowed;
101        run;

NOTE: There were 60355 observations read from the data set WORK.HCSRI01S.
NOTE: The PROCEDURE UNIVARIATE printed pages 7-8.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.46 seconds
      cpu time            0.45 seconds


102        %countsign(hcsri01s, allowed);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET hcsri01s (KEEP=allowed);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if allowed=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if allowed<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if allowed=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if allowed>0 then vartype='3=Positive';

NOTE: There were 60355 observations read from the data set WORK.HCSRI01S.
NOTE: The data set WORK.TEMP has 60355 observations and 2 variables.
NOTE: DATA statement used:
      real time           1.04 seconds
      cpu time            0.32 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 60355 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 9.
NOTE: PROCEDURE FREQ used:
      real time           0.22 seconds
      cpu time            0.12 seconds


103        *calculate indiv level cost univariate;
104        %indivcost(hcsri01s, costfile, costfile, allowed, cost_hcsri01);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET hcsri01s (KEEP=PID allowed);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_hcsri01;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_hcsri01 = MAX(0,allowed);
MPRINT(INDIVCOST):   ELSE cost_hcsri01 = SUM(cost_hcsri01, MAX(0,allowed));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;
```

```
NOTE: There were 60355 observations read from the data set WORK.HCSRI01S.
NOTE: The data set WORK.TEMP has 47208 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.38 seconds
      cpu time            0.39 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 3 variables.
NOTE: PROCEDURE SORT used:
      real time           1:55.06
      cpu time            31.90 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_hcsri01);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 47208 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 47208 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.47 seconds
      cpu time            0.42 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 47208 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 4 variables.
NOTE: DATA statement used:
      real time           49.38 seconds
      cpu time            24.11 seconds


105
106        *get hcsrn01 file;
107        libname xpt4 xport '/data/dod1/rawdata/hcsrn01.xpt';
NOTE: Libref XPT4 was successfully assigned as follows:
      Engine:        XPORT
```

```
      Physical Name: /data/dod1/rawdata/hcsrn01.xpt
108         data hcsrn01;
109           set xpt4.hcsrn01 (keep=pid tallowed);
110         *calculate record level cost univariate;
111         %subset(hcsrn01,pidfile,hcsrn01s);

NOTE: There were 13189028 observations read from the data set XPT4.HCSRN01.
NOTE: The data set WORK.HCSRN01 has 13189028 observations and 2 variables.
NOTE: DATA statement used:
      real time           6:57.58
      cpu time            2:17.80


MPRINT(SUBSET):   PROC SORT DATA=hcsrn01;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 13189028 observations read from the data set WORK.HCSRN01.
NOTE: The data set WORK.HCSRN01 has 13189028 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           9:31.10
      cpu time            4:35.83


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.03 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA hcsrn01s;
MPRINT(SUBSET):   MERGE hcsrn01 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 13189028 observations read from the data set WORK.HCSRN01.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.HCSRN01S has 8352518 observations and 2 variables.
NOTE: DATA statement used:
      real time           2:16.37
      cpu time            1:47.90


112         proc univariate data=hcsrn01s;
113           var tallowed;
```

```
13  The SAS System                                08:45 Friday, July 16, 2004

114        run;

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01S.
NOTE: The PROCEDURE UNIVARIATE printed pages 10-11.
NOTE: PROCEDURE UNIVARIATE used:
      real time           59.11 seconds
      cpu time            58.97 seconds


115        %countsign(hcsrn01s, tallowed);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET hcsrn01s (KEEP=tallowed);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if tallowed=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if tallowed<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if tallowed=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if tallowed>0 then vartype='3=Positive';

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01S.
NOTE: The data set WORK.TEMP has 8352518 observations and 2 variables.
NOTE: DATA statement used:
      real time           1:38.14
      cpu time            33.76 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 8352518 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 12.
NOTE: PROCEDURE FREQ used:
      real time           16.95 seconds
      cpu time            16.84 seconds


116        *calculate indiv level cost univariate;
117        %indivcost(hcsrn01s, costfile, costfile, tallowed, cost_hcsrn01);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET hcsrn01s (KEEP=PID tallowed);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_hcsrn01;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_hcsrn01 = MAX(0,tallowed);
MPRINT(INDIVCOST):   ELSE cost_hcsrn01 = SUM(cost_hcsrn01, MAX(0,tallowed));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01S.
NOTE: The data set WORK.TEMP has 887198 observations and 3 variables.
NOTE: DATA statement used:
```

```
      real time           1:15.00
      cpu time            34.05 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 4 variables.
NOTE: PROCEDURE SORT used:
      real time           2:19.58
      cpu time            34.20 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_hcsrn01);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 887198 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 887198 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           16.94 seconds
      cpu time            8.36 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 887198 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 5 variables.
NOTE: DATA statement used:
      real time           58.16 seconds
      cpu time            27.25 seconds


118
119        *get nmop01 file;
120        libname xpt5 xport '/data/dod1/rawdata/nmop01.xpt';
NOTE: Libref XPT5 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/nmop01.xpt
121        data nmop01;
122          set xpt5.nmop01 (keep=pid rxtotal);
123        *calculate record level cost univariate;
```

```
15    The SAS System                              08:45 Friday, July 16, 2004

124        %subset(nmop01,pidfile,nmop01s);

NOTE: There were 366140 observations read from the data set XPT5.NMOP01.
NOTE: The data set WORK.NMOP01 has 366140 observations and 2 variables.
NOTE: DATA statement used:
      real time           7.45 seconds
      cpu time            2.76 seconds


MPRINT(SUBSET):   PROC SORT DATA=nmop01;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 366140 observations read from the data set WORK.NMOP01.
NOTE: The data set WORK.NMOP01 has 366140 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           7.36 seconds
      cpu time            5.95 seconds


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA nmop01s;
MPRINT(SUBSET):   MERGE nmop01 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 366140 observations read from the data set WORK.NMOP01.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.NMOP01S has 244851 observations and 2 variables.
NOTE: DATA statement used:
      real time           16.78 seconds
      cpu time            16.12 seconds


125        proc univariate data=nmop01s;
126          var rxtotal;
127        run;

NOTE: There were 244851 observations read from the data set WORK.NMOP01S.
NOTE: The PROCEDURE UNIVARIATE printed pages 13-14.
```

```
NOTE: PROCEDURE UNIVARIATE used:
      real time           1.13 seconds
      cpu time            1.13 seconds


128        %countsign(nmop01s, rxtotal);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET nmop01s (KEEP=rxtotal);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if rxtotal=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if rxtotal<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if rxtotal=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if rxtotal>0 then vartype='3=Positive';

NOTE: There were 244851 observations read from the data set WORK.NMOP01S.
NOTE: The data set WORK.TEMP has 244851 observations and 2 variables.
NOTE: DATA statement used:
      real time           3.01 seconds
      cpu time            0.98 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 244851 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 15.
NOTE: PROCEDURE FREQ used:
      real time           0.56 seconds
      cpu time            0.48 seconds


129        *calculate indiv level cost univariate;
130        %indivcost(nmop01s, costfile, costfile, rxtotal, cost_nmop01);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET nmop01s (KEEP=PID rxtotal);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_nmop01;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_nmop01 = MAX(0,rxtotal);
MPRINT(INDIVCOST):   ELSE cost_nmop01 = SUM(cost_nmop01, MAX(0,rxtotal));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 244851 observations read from the data set WORK.NMOP01S.
NOTE: The data set WORK.TEMP has 37442 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.38 seconds
      cpu time            0.96 seconds
```

```
MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 5 variables.
NOTE: PROCEDURE SORT used:
      real time           2:22.48
      cpu time            36.48 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_nmop01);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 37442 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 37442 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.11 seconds
      cpu time            0.35 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 37442 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 6 variables.
NOTE: DATA statement used:
      real time           1:02.29
      cpu time            26.19 seconds


131
132        *get sidr02 file;
133        libname xpt6 xport '/data/dod1/rawdata/sidr02.xpt';
NOTE: Libref XPT6 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sidr02.xpt
134        data sidr02;
135          set xpt6.sidr02 (keep=pid fullcost);
136        *calculate record level cost univariate;
137        %subset(sidr02,pidfile,sidr02s);

NOTE: There were 133754 observations read from the data set XPT6.SIDR02.
NOTE: The data set WORK.SIDR02 has 133754 observations and 2 variables.
```

```
NOTE: DATA statement used:
      real time           5.72 seconds
      cpu time            1.64 seconds


MPRINT(SUBSET):   PROC SORT DATA=sidr02;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 133754 observations read from the data set WORK.SIDR02.
NOTE: The data set WORK.SIDR02 has 133754 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           2.83 seconds
      cpu time            1.76 seconds


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA sidr02s;
MPRINT(SUBSET):   MERGE sidr02 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 133754 observations read from the data set WORK.SIDR02.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.SIDR02S has 72331 observations and 2 variables.
NOTE: DATA statement used:
      real time           14.97 seconds
      cpu time            14.19 seconds


138        proc univariate data=sidr02s;
139          var fullcost;
140        run;

NOTE: There were 72331 observations read from the data set WORK.SIDR02S.
NOTE: The PROCEDURE UNIVARIATE printed pages 16-17.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.47 seconds
      cpu time            0.47 seconds
```

```
141        %countsign(sidr02s, fullcost);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET sidr02s (KEEP=fullcost);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if fullcost=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if fullcost<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if fullcost=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if fullcost>0 then vartype='3=Positive';

NOTE: There were 72331 observations read from the data set WORK.SIDR02S.
NOTE: The data set WORK.TEMP has 72331 observations and 2 variables.
NOTE: DATA statement used:
      real time           1.19 seconds
      cpu time            0.28 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 72331 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 18.
NOTE: PROCEDURE FREQ used:
      real time           0.23 seconds
      cpu time            0.16 seconds


142        *calculate indiv level cost univariate;
143        %indivcost(sidr02s, costfile, costfile, fullcost, cost_sidr02);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET sidr02s (KEEP=PID fullcost);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_sidr02;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_sidr02 = MAX(0,fullcost);
MPRINT(INDIVCOST):   ELSE cost_sidr02 = SUM(cost_sidr02, MAX(0,fullcost));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 72331 observations read from the data set WORK.SIDR02S.
NOTE: The data set WORK.TEMP has 60764 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.59 seconds
      cpu time            0.48 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;
```

```
NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 6 variables.
NOTE: PROCEDURE SORT used:
      real time           2:29.71
      cpu time            38.44 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_sidr02);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 60764 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 60764 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.42 seconds
      cpu time            0.54 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 60764 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 7 variables.
NOTE: DATA statement used:
      real time           1:07.82
      cpu time            27.82 seconds


144
145        *get sadr02 file;
146        libname xpt7 xport '/data/dod1/rawdata/sadr02.xpt';
NOTE: Libref XPT7 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sadr02.xpt
147        data sadr02;
148          set xpt7.sadr02 (keep=pid fcost);
149        *calculate record level cost univariate;
150        %subset(sadr02,pidfile,sadr02s);

NOTE: There were 20349549 observations read from the data set XPT7.SADR02.
NOTE: The data set WORK.SADR02 has 20349549 observations and 2 variables.
NOTE: DATA statement used:
      real time           11:09.70
      cpu time            3:28.37
```

```
MPRINT(SUBSET):   PROC SORT DATA=sadr02;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 20349549 observations read from the data set WORK.SADR02.
NOTE: The data set WORK.SADR02 has 20349549 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           18:37.14
      cpu time            7:06.33


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.03 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA sadr02s;
MPRINT(SUBSET):   MERGE sadr02 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 20349549 observations read from the data set WORK.SADR02.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.SADR02S has 12153452 observations and 2 variables.
NOTE: DATA statement used:
      real time           3:43.26
      cpu time            2:33.04


151        proc univariate data=sadr02s;
152          var fcost;
153        run;

NOTE: There were 12153452 observations read from the data set WORK.SADR02S.
NOTE: The PROCEDURE UNIVARIATE printed pages 19-20.
NOTE: PROCEDURE UNIVARIATE used:
      real time           1:09.46
      cpu time            1:09.17


154        %countsign(sadr02s, fcost);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET sadr02s (KEEP=fcost);
```

```
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if fcost=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if fcost<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if fcost=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if fcost>0 then vartype='3=Positive';

NOTE: There were 12153452 observations read from the data set WORK.SADR02S.
NOTE: The data set WORK.TEMP has 12153452 observations and 2 variables.
NOTE: DATA statement used:
      real time           2:35.50
      cpu time            53.72 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 12153452 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 21.
NOTE: PROCEDURE FREQ used:
      real time           26.86 seconds
      cpu time            26.70 seconds


155        *calculate indiv level cost univariate;
156        %indivcost(sadr02s, costfile, costfile, fcost, cost_sadr02);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET sadr02s (KEEP=PID fcost);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_sadr02;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_sadr02 = MAX(0,fcost);
MPRINT(INDIVCOST):   ELSE cost_sadr02 = SUM(cost_sadr02, MAX(0,fcost));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 12153452 observations read from the data set WORK.SADR02S.
NOTE: The data set WORK.TEMP has 1682912 observations and 3 variables.
NOTE: DATA statement used:
      real time           1:58.11
      cpu time            52.88 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 7 variables.
NOTE: PROCEDURE SORT used:
```

```
      real time           2:58.52
      cpu time            41.52 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_sadr02);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 1682912 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 1682912 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1:06.97
      cpu time            20.68 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 1682912 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 8 variables.
NOTE: DATA statement used:
      real time           1:21.17
      cpu time            34.61 seconds


157
158        *get hcsri02 file;
159        libname xpt8 xport '/data/dod1/rawdata/hcsri02.xpt';
NOTE: Libref XPT8 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/hcsri02.xpt
160        data hcsri02;
161          set xpt8.hcsri02 (keep=pid allowed);
162        *calculate record level cost univariate;
163        %subset(hcsri02,pidfile,hcsri02s);

NOTE: There were 156210 observations read from the data set XPT8.HCSRI02.
NOTE: The data set WORK.HCSRI02 has 156210 observations and 2 variables.
NOTE: DATA statement used:
      real time           5.63 seconds
      cpu time            1.75 seconds


MPRINT(SUBSET):   PROC SORT DATA=hcsri02;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;
```

```
NOTE: There were 156210 observations read from the data set WORK.HCSRI02.
NOTE: The data set WORK.HCSRI02 has 156210 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           3.35 seconds
      cpu time            2.05 seconds


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


MPRINT(SUBSET):   DATA hcsri02s;
MPRINT(SUBSET):   MERGE hcsri02 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 156210 observations read from the data set WORK.HCSRI02.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.HCSRI02S has 64277 observations and 2 variables.
NOTE: DATA statement used:
      real time           17.39 seconds
      cpu time            14.61 seconds


164        proc univariate data=hcsri02s;
165          var allowed;
166        run;

NOTE: There were 64277 observations read from the data set WORK.HCSRI02S.
NOTE: The PROCEDURE UNIVARIATE printed pages 22-23.
NOTE: PROCEDURE UNIVARIATE used:
      real time           0.51 seconds
      cpu time            0.52 seconds


167        %countsign(hcsri02s, allowed);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET hcsri02s (KEEP=allowed);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if allowed=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if allowed<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if allowed=0 then vartype='2=Zero';
```

```
MPRINT(COUNTSIGN):   else if allowed>0 then vartype='3=Positive';

NOTE: There were 64277 observations read from the data set WORK.HCSRI02S.
NOTE: The data set WORK.TEMP has 64277 observations and 2 variables.
NOTE: DATA statement used:
      real time           1.27 seconds
      cpu time            0.31 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 64277 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 24.
NOTE: PROCEDURE FREQ used:
      real time           0.87 seconds
      cpu time            0.14 seconds


168        *calculate indiv level cost univariate;
169        %indivcost(hcsri02s, costfile, costfile, allowed, cost_hcsri02);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET hcsri02s (KEEP=PID allowed);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_hcsri02;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_hcsri02 = MAX(0,allowed);
MPRINT(INDIVCOST):   ELSE cost_hcsri02 = SUM(cost_hcsri02, MAX(0,allowed));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 64277 observations read from the data set WORK.HCSRI02S.
NOTE: The data set WORK.TEMP has 48596 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.37 seconds
      cpu time            0.40 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           3:20.05
      cpu time            42.73 seconds
```

```
MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_hcsri02);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 48596 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 48596 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.55 seconds
      cpu time            0.47 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 48596 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 9 variables.
NOTE: DATA statement used:
      real time           1:28.79
      cpu time            30.15 seconds


170
171        *get hcsrn02 file;
172        libname xpt9 xport '/data/dod1/rawdata/hcsrn02.xpt';
NOTE: Libref XPT9 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/hcsrn02.xpt
173        data hcsrn02;
174          set xpt9.hcsrn02 (keep=pid tallowed);
175        *calculate record level cost univariate;
176        %subset(hcsrn02,pidfile,hcsrn02s);

NOTE: There were 15734412 observations read from the data set XPT9.HCSRN02.
NOTE: The data set WORK.HCSRN02 has 15734412 observations and 2 variables.
NOTE: DATA statement used:
      real time           8:04.59
      cpu time            2:39.40


MPRINT(SUBSET):   PROC SORT DATA=hcsrn02;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 15734412 observations read from the data set WORK.HCSRN02.
NOTE: The data set WORK.HCSRN02 has 15734412 observations and 2 variables.
NOTE: PROCEDURE SORT used:
```

```
27    The SAS System                              08:45 Friday, July 16, 2004

      real time           13:13.23
      cpu time            5:31.37


MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.03 seconds
      cpu time            0.01 seconds


MPRINT(SUBSET):   DATA hcsrn02s;
MPRINT(SUBSET):   MERGE hcsrn02 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 15734412 observations read from the data set WORK.HCSRN02.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.HCSRN02S has 9457022 observations and 2 variables.
NOTE: DATA statement used:
      real time           2:32.08
      cpu time            2:02.66


177        proc univariate data=hcsrn02s;
178          var tallowed;
179        run;

NOTE: There were 9457022 observations read from the data set WORK.HCSRN02S.
NOTE: The PROCEDURE UNIVARIATE printed pages 25-26.
NOTE: PROCEDURE UNIVARIATE used:
      real time           1:13.85
      cpu time            1:13.71


180        %countsign(hcsrn02s, tallowed);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET hcsrn02s (KEEP=tallowed);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if tallowed=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if tallowed<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if tallowed=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if tallowed>0 then vartype='3=Positive';

NOTE: There were 9457022 observations read from the data set WORK.HCSRN02S.
NOTE: The data set WORK.TEMP has 9457022 observations and 2 variables.
```

```
28   The SAS System                          08:45 Friday, July 16, 2004

NOTE: DATA statement used:
      real time           2:10.57
      cpu time            37.76 seconds


MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 9457022 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 27.
NOTE: PROCEDURE FREQ used:
      real time           20.81 seconds
      cpu time            20.68 seconds


181        *calculate indiv level cost univariate;
182        %indivcost(hcsrn02s, costfile, costfile, tallowed, cost_hcsrn02);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET hcsrn02s (KEEP=PID tallowed);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_hcsrn02;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_hcsrn02 = MAX(0,tallowed);
MPRINT(INDIVCOST):   ELSE cost_hcsrn02 = SUM(cost_hcsrn02, MAX(0,tallowed));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 9457022 observations read from the data set WORK.HCSRN02S.
NOTE: The data set WORK.TEMP has 923117 observations and 3 variables.
NOTE: DATA statement used:
      real time           1:24.48
      cpu time            39.28 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 9 variables.
NOTE: PROCEDURE SORT used:
      real time           3:38.45
      cpu time            46.33 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_hcsrn02);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;
```

```
NOTE: There were 923117 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 923117 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           17.79 seconds
      cpu time            8.87 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 923117 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 10 variables.
NOTE: DATA statement used:
      real time           1:35.82
      cpu time            35.33 seconds


183
184        *get pdts02 file;
185        libname xpt10 xport '/data/dod1/rawdata/pdts02.xpt';
NOTE: Libref XPT10 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/pdts02.xpt
186        data pdts02;
187          set xpt10.pdts02 (keep=pid netamt);
188        *calculate record level cost univariate;
189        %subset(pdts02,pidfile,pdts02s);

NOTE: There were 453115 observations read from the data set XPT10.PDTS02.
NOTE: The data set WORK.PDTS02 has 453115 observations and 2 variables.
NOTE: DATA statement used:
      real time           8.92 seconds
      cpu time            3.71 seconds


MPRINT(SUBSET):   PROC SORT DATA=pdts02;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: There were 453115 observations read from the data set WORK.PDTS02.
NOTE: The data set WORK.PDTS02 has 453115 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           9.18 seconds
      cpu time            6.38 seconds
```

```
MPRINT(SUBSET):   PROC SORT DATA=pidfile;
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   RUN;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.01 seconds


MPRINT(SUBSET):   DATA pdts02s;
MPRINT(SUBSET):   MERGE pdts02 (IN=INFIRST) pidfile (IN=INSECOND);
MPRINT(SUBSET):   BY PID;
MPRINT(SUBSET):   IF INFIRST & INSECOND;
MPRINT(SUBSET):   RUN;

NOTE: There were 453115 observations read from the data set WORK.PDTS02.
NOTE: There were 2304926 observations read from the data set WORK.PIDFILE.
NOTE: The data set WORK.PDTS02S has 295897 observations and 2 variables.
NOTE: DATA statement used:
      real time           18.56 seconds
      cpu time            16.78 seconds


190        proc univariate data=pdts02s;
191          var netamt;
192        run;

NOTE: There were 295897 observations read from the data set WORK.PDTS02S.
NOTE: The PROCEDURE UNIVARIATE printed pages 28-29.
NOTE: PROCEDURE UNIVARIATE used:
      real time           1.30 seconds
      cpu time            1.30 seconds


193        %countsign(pdts02s, netamt);
MPRINT(COUNTSIGN):   DATA WORK.TEMP;
MPRINT(COUNTSIGN):   SET pdts02s (KEEP=netamt);
MPRINT(COUNTSIGN):   length vartype $10.;
MPRINT(COUNTSIGN):   if netamt=. then vartype='.=Missing';
MPRINT(COUNTSIGN):   else if netamt<0 then vartype='1=Negative';
MPRINT(COUNTSIGN):   else if netamt=0 then vartype='2=Zero';
MPRINT(COUNTSIGN):   else if netamt>0 then vartype='3=Positive';

NOTE: There were 295897 observations read from the data set WORK.PDTS02S.
NOTE: The data set WORK.TEMP has 295897 observations and 2 variables.
NOTE: DATA statement used:
      real time           3.49 seconds
      cpu time            1.17 seconds
```

```
MPRINT(COUNTSIGN):   PROC FREQ;
MPRINT(COUNTSIGN):   TABLES VARTYPE;
MPRINT(COUNTSIGN):   RUN;

NOTE: There were 295897 observations read from the data set WORK.TEMP.
NOTE: The PROCEDURE FREQ printed page 30.
NOTE: PROCEDURE FREQ used:
      real time           0.65 seconds
      cpu time            0.56 seconds


194        *calculate indiv level cost univariate;
195        %indivcost(pdts02s, costfile, costfile, netamt, cost_pdts02);
MPRINT(INDIVCOST):   DATA TEMP;
MPRINT(INDIVCOST):   SET pdts02s (KEEP=PID netamt);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RETAIN cost_pdts02;
MPRINT(INDIVCOST):   IF FIRST.PID THEN cost_pdts02 = MAX(0,netamt);
MPRINT(INDIVCOST):   ELSE cost_pdts02 = SUM(cost_pdts02, MAX(0,netamt));
MPRINT(INDIVCOST):   IF LAST.PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 295897 observations read from the data set WORK.PDTS02S.
NOTE: The data set WORK.TEMP has 38304 observations and 3 variables.
NOTE: DATA statement used:
      real time           1.51 seconds
      cpu time            1.15 seconds


MPRINT(INDIVCOST):   *attach temp to insasdata2 with output called outsasdata;
MPRINT(INDIVCOST):   PROC SORT DATA=costfile;
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 10 variables.
NOTE: PROCEDURE SORT used:
      real time           3:52.56
      cpu time            47.84 seconds


MPRINT(INDIVCOST):   PROC SORT DATA=TEMP (KEEP=PID cost_pdts02);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 38304 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.TEMP has 38304 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           1.16 seconds
```

```
      cpu time            0.39 seconds


MPRINT(INDIVCOST):   DATA costfile;
MPRINT(INDIVCOST):   MERGE costfile (IN=INFIRST) TEMP (IN=INSECOND);
MPRINT(INDIVCOST):   BY PID;
MPRINT(INDIVCOST):   IF INFIRST;
MPRINT(INDIVCOST):   RUN;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: There were 38304 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.COSTFILE has 2304926 observations and 11 variables.
NOTE: DATA statement used:
      real time           1:45.63
      cpu time            32.21 seconds


196
197        data saswork.costfile;
198          set costfile;
199          if cost_hcsri01<0 then cost_hcsri01=0;
200          if cost_hcsri02<0 then cost_hcsri02=0;
201          if cost_hcsrn01<0 then cost_hcsrn01=0;
202          if cost_hcsrn02<0 then cost_hcsrn02=0;
203          if cost_nmop01<0 then cost_nmop01=0;
204          if cost_pdts02<0 then cost_pdts02=0;
205          if cost_sadr01<0 then cost_sadr01=0;
206          if cost_sadr02<0 then cost_sadr02=0;
207          if cost_sidr01<0 then cost_sidr01=0;
208          if cost_sidr02<0 then cost_sidr02=0;
209          cost_tot01 = sum(cost_hcsri01, cost_hcsrn01, cost_nmop01,
209      ! cost_sadr01, cost_sidr01);
210          cost_tot02 = sum(cost_hcsri02, cost_hcsrn02, cost_pdts02,
210      ! cost_sadr02, cost_sidr02);
211          label cost_hcsri01='HCSR-I expenditures,individual, FY2001'
212          label cost_hcsri02='HCSR-I expenditures,individual, FY2002'
213          label cost_hcsrn01='HCSR-N expenditures,individual, FY2001'
214          label cost_hcsrn02='HCSR-N expenditures,individual, FY2002'
215          label cost_nmop01='NMOP expenditures,individual, FY2001'
216          label cost_pdts02='PDTS expenditures,individual, FY2002'
217          label cost_sadr01='SADR expenditures,individual, FY2001'
218          label cost_sadr02='SADR expenditures,individual, FY2002'
219          label cost_sidr01='SIDR expenditures,individual, FY2001'
220          label cost_sidr02='SIDR expenditures,individual, FY2002'
221          label cost_tot01='Total expenditures,individual, FY2001'
222          label cost_tot02='Total expenditures,individual, FY2002';
223        run;

NOTE: There were 2304926 observations read from the data set WORK.COSTFILE.
NOTE: The data set SASWORK.COSTFILE has 2304926 observations and 13 variables.
```

```
33    The SAS System                              08:45 Friday, July 16, 2004

NOTE: DATA statement used:
      real time           2:04.72
      cpu time            25.22 seconds


224
225        proc sort data=saswork.dod1; by pid; run;

NOTE: There were 2304926 observations read from the data set SASWORK.DOD1.
NOTE: The data set SASWORK.DOD1 has 2304926 observations and 13 variables.
NOTE: PROCEDURE SORT used:
      real time           3:03.43
      cpu time            41.35 seconds


226        proc sort data=saswork.costfile; by pid; run;

NOTE: There were 2304926 observations read from the data set SASWORK.COSTFILE.
NOTE: The data set SASWORK.COSTFILE has 2304926 observations and 13 variables.
NOTE: PROCEDURE SORT used:
      real time           4:43.48
      cpu time            54.10 seconds


227        data saswork.dod2;
228          merge saswork.dod1 (in=first) saswork.costfile (in=second);
229          by pid;
230          if first & second;
231        run;

NOTE: There were 2304926 observations read from the data set SASWORK.DOD1.
NOTE: There were 2304926 observations read from the data set SASWORK.COSTFILE.
NOTE: The data set SASWORK.DOD2 has 2304926 observations and 25 variables.
NOTE: DATA statement used:
      real time           2:51.66
      cpu time            51.45 seconds


232

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           3:07:08.88
      cpu time            1:14:50.05
```

Attachment A3_CDPS
(task4b_cdpsdata.log)

```
1  The SAS System                          17:08 Thursday, September 30, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.

NOTE: SAS initialization used:
      real time           0.32 seconds
      cpu time            0.07 seconds
1
2          /*
3          task4b_cdpsdata.sas
4          Amresh Hanchate
5          DOD1 Project
6          Sept 20, 2004
7
8          THIS IS ALMOST IDENTICAL TO TASK4A_CDPSDATA.SAS WITH ONE CHANGE;
HERE LABFLAG=1 RECORDS (RECORDS WITH ONLY LAB
8        ! PROCEDURES) FROM HCSRN FILE ARE INCLUDED WHEREAS IN TASK4A THEY WERE
EXCLUDED.
9
10         This program creates the data sets required for running CDPS and
other risk adjustment models. Two variants of this data
10       ! are created (a and b) -- this program producing the "b" variant.
While the "a" variant excludes the diagnoses found in
10       ! lab-only records
11
12         Two data sets are created, one containing individual characteristics
and prospective expenditures (dod2charb) and another
12       !  containing diagnoses (dod2diagb).
13         The data file dod2charb also contains a flag (datatype) that splits
the total sample into estimation (1.8 million) and
13       ! validation (500K) subsamples -- datatype= 1(fitting) and
2(validation).
14
15         Note that both these datasets (dod2charb, dod2diagb) follow the
layout specifications in the document "CDPS Instruction
15       ! Manual Version 2.0 July 1, 2002" downloaded along with the CDPS
software. It turns out that this layout also works for
15       ! running the DCG m
16
17         What diagnoses are excluded:
18         a) Diagnoses from records indicating telephone consults (apptstat=6)
are deleted -- they are 11.6% of the records in
18       ! SADR01. Only this file has telephone consults.
19
20         */
21
22         options ps=60 ls=80 nocenter;
23         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
```

```
24          footnote '/data/dod1/saswork/task4a_cdpsdata.sas';
25
26          * first all manipulations are done for the entire sample, afterwhich
26        ! they are split into validation and estimation samples;
27          * following data set has elig and exp data for the whole sample;
```

```
28          * NOTE: THE EXPENDITURE UNIT IS ANNUAL;
29          data combdata;
30            *rename cost_tot02 by the requisite name, expyr;
31            set saswork.dod2 (keep=pid sex age cost_tot02
31        ! rename=(cost_tot02=expyr));
32            if sex="M" then male = 1;
33            else if sex="F" then male = 0;
34            *create aidcat as specified in CDPS manual;
35            if age GE 18 then aidcat = 1;
36            else if age NE . then aidcat = 2;
37            *create elig variable as specified in CDPS manual (this gives the
37        ! number of months of eligibility) -- since we selected only persons
37        ! enrolled for the entire duration it is set to 12;
38            elig = 12;
39            *topcoded expenditures;
40            expyrt25 = expyr;
41            if expyr GT 25000 then expyrt25 = 25000;
42            expyrt50 = expyr;
43            if expyr GT 50000 then expyrt50 = 50000;
44            keep pid aidcat age male elig expyr expyrt25 expyrt50;
45          run;

NOTE: There were 2304926 observations read from the data set SASWORK.DOD2.
NOTE: The data set WORK.COMBDATA has 2304926 observations and 8 variables.
NOTE: DATA statement used:
      real time           2:29.00
      cpu time            23.03 seconds


46
47          /* create the combined diagnoses dataset, combdiag;
48          as diagnoses are in four different file, four corresponding files in
48        ! the desired format are first created and then merged */
49          * SIDR FY2001;
50          libname xpt1 xport '/data/dod1/rawdata/sidr01.xpt';
NOTE: Libref XPT1 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sidr01.xpt
51          data sidr01;
52            set xpt1.sidr01;
53            keep pid dx:;
54          run;

NOTE: There were 141226 observations read from the data set XPT1.SIDR01.
NOTE: The data set WORK.SIDR01 has 141226 observations and 9 variables.
NOTE: DATA statement used:
      real time           9.37 seconds
      cpu time            3.39 seconds


55          *keep only records relating to those in combdata;
56          proc sort data=combdata; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: The data set WORK.COMBDATA has 2304926 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           3:19.06
```

```
      cpu time            46.01 seconds


57         proc sort data=sidr01; by pid; run;

NOTE: There were 141226 observations read from the data set WORK.SIDR01.
NOTE: The data set WORK.SIDR01 has 141226 observations and 9 variables.
NOTE: PROCEDURE SORT used:
      real time           7.56 seconds
      cpu time            2.87 seconds


58         data sidr01;
59            merge combdata (in=infirst keep=pid) sidr01 (in=insecond);
60            by pid;
61            if infirst & insecond;
62         run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: There were 141226 observations read from the data set WORK.SIDR01.
NOTE: The data set WORK.SIDR01 has 78119 observations and 9 variables.
NOTE: DATA statement used:
      real time           20.76 seconds
      cpu time            20.31 seconds


63         /* PROC TRANSPOSE is used to create a 'skinny' file with one
63       ! diagnosis (along with pid) listed in each record; so each
beneficiary
63       !  has as many records as there are diagnoses;
64         since the raw datasets have multiple records per person, create
rowid
64       !  identifier;
65         */
66         data sidr01a;
67         set sidr01;
68         rowid = _N_;

NOTE: There were 78119 observations read from the data set WORK.SIDR01.
NOTE: The data set WORK.SIDR01A has 78119 observations and 10 variables.
NOTE: DATA statement used:
      real time           3.78 seconds
      cpu time            0.63 seconds


69         proc sort data=sidr01a; by pid rowid; run;

NOTE: There were 78119 observations read from the data set WORK.SIDR01A.
NOTE: The data set WORK.SIDR01A has 78119 observations and 10 variables.
NOTE: PROCEDURE SORT used:
      real time           4.47 seconds
      cpu time            1.20 seconds


70         proc transpose data=sidr01a out=sidr01b (rename=(col1=diag));
71         var dx1-dx8;
72         by pid rowid;
73         run;
```

```
4 The SAS System                                17:08 Thursday, September 30, 2004

NOTE: There were 78119 observations read from the data set WORK.SIDR01A.
NOTE: The data set WORK.SIDR01B has 624952 observations and 5 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           28.90 seconds
      cpu time            10.60 seconds


74         data saswork.combsidr;
75         set sidr01b (keep=pid diag);
76         if diag NE " ";
77         run;

NOTE: There were 624952 observations read from the data set WORK.SIDR01B.
NOTE: The data set SASWORK.COMBSIDR has 270836 observations and 2 variables.
NOTE: DATA statement used:
      real time           3.66 seconds
      cpu time            2.22 seconds


78         proc print data=saswork.combsidr (obs=10); title "Sample of
78       ! observations"; run;

NOTE: There were 10 observations read from the data set SASWORK.COMBSIDR.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used:
      real time           0.03 seconds
      cpu time            0.02 seconds


79         proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype   File Size  Last Modified
-----------------------------------------------------
1  COMBDATA  DATA      186966016   30SEP2004:17:14:00
2  SIDR01    DATA        6479872   30SEP2004:17:14:29
3  SIDR01A   DATA        7634944   30SEP2004:17:14:37
4  SIDR01B   DATA       55664640   30SEP2004:17:15:06
80         delete sidr01 sidr01a sidr01b;
81         run;

NOTE: Deleting WORK.SIDR01 (memtype=DATA).
NOTE: Deleting WORK.SIDR01A (memtype=DATA).
NOTE: Deleting WORK.SIDR01B (memtype=DATA).
82
83
84         * SADR FY2001;
```

```
5 The SAS System                              17:08 Thursday, September 30, 2004

85         * note that records from telephone consults (apptstat=6) are deleted
85       ! -- they are 11.6% of the records;
86         libname xpt2 xport '/data/dod1/rawdata/sadr01.xpt';
NOTE: Libref XPT2 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/sadr01.xpt

NOTE: PROCEDURE DATASETS used:
      real time           0.33 seconds
      cpu time            0.19 seconds


87         data sadr01;
88           set xpt2.sadr01 (keep = pid icd: apptstat);
89           if apptstat=6 then delete;
90         run;

NOTE: Character values have been converted to numeric
      values at the places given by: (Line):(Column).
      89:5
NOTE: There were 20691247 observations read from the data set XPT2.SADR01.
NOTE: The data set WORK.SADR01 has 18283115 observations and 6 variables.
NOTE: DATA statement used:
      real time           13:40.32
      cpu time            5:02.89


91         *keep only records relating to the selected subset;
92         proc sort data=sadr01 (keep = pid icd:); by pid; run;

NOTE: There were 18283115 observations read from the data set WORK.SADR01.
NOTE: The data set WORK.SADR01 has 18283115 observations and 5 variables.
NOTE: PROCEDURE SORT used:
      real time           25:00.66
      cpu time            7:56.57


93         data sadr01;
94            merge combdata (in=infirst keep=pid) sadr01 (in=insecond);
95            by pid;
96            if infirst & insecond;
97         run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: There were 18283115 observations read from the data set WORK.SADR01.
NOTE: The data set WORK.SADR01 has 11248919 observations and 5 variables.
NOTE: DATA statement used:
      real time           7:51.92
      cpu time            2:53.84


98         /* PROC TRANSPOSE is used to create a 'skinny' file with one
98       ! diagnosis (along with pid) listed in each record; so each
beneficiary
98       !  has as many records as there are diagnoses;
99         since the raw datasets have multiple records per person, create
rowid
99       !  identifier;
100        */
```

```
101        data sadr01a;
102        set sadr01;
103        rowid = _N_;

NOTE: There were 11248919 observations read from the data set WORK.SADR01.
NOTE: The data set WORK.SADR01A has 11248919 observations and 6 variables.
NOTE: DATA statement used:
      real time           5:18.88
      cpu time            1:16.98


104        proc sort data=sadr01a; by pid rowid; run;

NOTE: There were 11248919 observations read from the data set WORK.SADR01A.
NOTE: The data set WORK.SADR01A has 11248919 observations and 6 variables.
NOTE: PROCEDURE SORT used:
      real time           17:21.66
      cpu time            3:30.99


105        proc transpose data=sadr01a out=sadr01b (rename=(col1=diag));
106        var icd1-icd4;
107        by pid rowid;
108        run;

NOTE: There were 11248919 observations read from the data set WORK.SADR01A.
NOTE: The data set WORK.SADR01B has 44995676 observations and 5 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           37:57.05
      cpu time            20:46.51


109        data saswork.combsadr;
110        set sadr01b (keep=pid diag);
111        if diag NE " ";
112        run;

NOTE: There were 44995676 observations read from the data set WORK.SADR01B.
NOTE: The data set SASWORK.COMBSADR has 15163862 observations and 2 variables.
NOTE: DATA statement used:
      real time           11:16.93
      cpu time            3:04.03


113        proc print data=saswork.combsadr (obs=10); run;

NOTE: There were 10 observations read from the data set SASWORK.COMBSADR.
NOTE: The PROCEDURE PRINT printed page 2.
NOTE: PROCEDURE PRINT used:
      real time           0.06 seconds
      cpu time            0.00 seconds


114        proc datasets library=work;
```

```
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype    File Size  Last Modified
-----------------------------------------------------
1  COMBDATA  DATA       186966016   30SEP2004:17:14:00
2  SADR01    DATA       614350848   30SEP2004:18:01:43
3  SADR01A   DATA       725614592   30SEP2004:18:24:23
4  SADR01B   DATA      4006584320   30SEP2004:19:02:20
115        delete sadr01 sadr01a sadr01b;
116        run;

NOTE: Deleting WORK.SADR01 (memtype=DATA).
NOTE: Deleting WORK.SADR01A (memtype=DATA).
NOTE: Deleting WORK.SADR01B (memtype=DATA).
117
118        * HCSRI FY2001;
119        libname xpt3 xport '/data/dod1/rawdata/hcsri01.xpt';
NOTE: Libref XPT3 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/hcsri01.xpt

NOTE: PROCEDURE DATASETS used:
      real time           4.33 seconds
      cpu time            3.46 seconds


120        data hcsri01;
121          set xpt3.hcsri01 (rename=(pdx=dx9));
122          keep pid dx:;
123        run;

NOTE: There were 144094 observations read from the data set XPT3.HCSRI01.
NOTE: The data set WORK.HCSRI01 has 144094 observations and 10 variables.
NOTE: DATA statement used:
      real time           7.28 seconds
      cpu time            2.73 seconds


124        *keep only records relating to the selected subset;
125        proc sort data=hcsri01; by pid; run;

NOTE: There were 144094 observations read from the data set WORK.HCSRI01.
NOTE: The data set WORK.HCSRI01 has 144094 observations and 10 variables.
NOTE: PROCEDURE SORT used:
      real time           18.82 seconds
      cpu time            2.37 seconds
```

```
126        data hcsri01;
127           merge combdata (in=infirst keep=pid) hcsri01 (in=insecond);
128           by pid;
129           if infirst & insecond;
130        run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: There were 144094 observations read from the data set WORK.HCSRI01.
NOTE: The data set WORK.HCSRI01 has 60355 observations and 10 variables.
NOTE: DATA statement used:
      real time           22.63 seconds
      cpu time            17.90 seconds


131        /* PROC TRANSPOSE is used to create a 'skinny' file with one
131      ! diagnosis (along with pid) listed in each record; so each
beneficiary
131      !  has as many records as there are diagnoses;
132        since the raw datasets have multiple records per person, create
rowid
132      !  identifier;
133        */
134        data hcsri01a;
135        set hcsri01;
136        rowid = _N_;

NOTE: There were 60355 observations read from the data set WORK.HCSRI01.
NOTE: The data set WORK.HCSRI01A has 60355 observations and 11 variables.
NOTE: DATA statement used:
      real time           2.41 seconds
      cpu time            0.42 seconds


137        proc sort data=hcsri01a; by pid rowid; run;

NOTE: There were 60355 observations read from the data set WORK.HCSRI01A.
NOTE: The data set WORK.HCSRI01A has 60355 observations and 11 variables.
NOTE: PROCEDURE SORT used:
      real time           2.94 seconds
      cpu time            0.76 seconds


138        proc transpose data=hcsri01a out=hcsri01b (rename=(col1=diag));
139        var dx:;
140        by pid rowid;
141        run;

NOTE: There were 60355 observations read from the data set WORK.HCSRI01A.
NOTE: The data set WORK.HCSRI01B has 543195 observations and 4 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           12.57 seconds
      cpu time            6.99 seconds


142        data saswork.combhcsri;
143        set hcsri01b (keep=pid diag);
144        if diag NE " ";
145        run;
```

```
NOTE: There were 543195 observations read from the data set WORK.HCSRI01B.
NOTE: The data set SASWORK.COMBHCSRI has 220499 observations and 2 variables.
NOTE: DATA statement used:
      real time           2.94 seconds
      cpu time            1.51 seconds


146        proc print data=saswork.combhcsri (obs=10); run;

NOTE: There were 10 observations read from the data set SASWORK.COMBHCSRI.
NOTE: The PROCEDURE PRINT printed page 3.
NOTE: PROCEDURE PRINT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


147        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype   File Size  Last Modified
-----------------------------------------------------
1  COMBDATA  DATA      186966016   30SEP2004:17:14:00
2  HCSRI01   DATA        3850240   30SEP2004:19:14:32
3  HCSRI01A  DATA        4390912   30SEP2004:19:14:38
4  HCSRI01B  DATA       21938176   30SEP2004:19:14:50
148        delete hcsri01 hcsri01a hcsri01b;
149        run;

NOTE: Deleting WORK.HCSRI01 (memtype=DATA).
NOTE: Deleting WORK.HCSRI01A (memtype=DATA).
NOTE: Deleting WORK.HCSRI01B (memtype=DATA).
150
151        * HCSRN FY2001;
152        * lab-only records are NOT dropped;
153        libname xpt4 xport '/data/dod1/rawdata/hcsrn01.xpt';
NOTE: Libref XPT4 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/hcsrn01.xpt

NOTE: PROCEDURE DATASETS used:
      real time           0.28 seconds
      cpu time            0.07 seconds


154        data hcsrn01;
155          set xpt4.hcsrn01(rename=(pdx=dx5));
```

```
156          *if labflag NE 1;
157          keep pid dx:;
158        run;

NOTE: There were 13189028 observations read from the data set XPT4.HCSRN01.
NOTE: The data set WORK.HCSRN01 has 13189028 observations and 6 variables.
NOTE: DATA statement used:
      real time           9:10.86
      cpu time            3:53.56


159        *keep only records relating to the selected subset;
160        proc sort data=hcsrn01; by pid; run;

NOTE: There were 13189028 observations read from the data set WORK.HCSRN01.
NOTE: The data set WORK.HCSRN01 has 13189028 observations and 6 variables.
NOTE: PROCEDURE SORT used:
      real time           12:33.06
      cpu time            5:05.47


161        data hcsrn01;
162           merge combdata (in=infirst keep=pid) hcsrn01 (in=insecond);
163           by pid;
164           if infirst & insecond;
165        run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: There were 13189028 observations read from the data set WORK.HCSRN01.
NOTE: The data set WORK.HCSRN01 has 8352518 observations and 6 variables.
NOTE: DATA statement used:
      real time           3:38.00
      cpu time            2:54.38


166        /* PROC TRANSPOSE is used to create a 'skinny' file with one
166      ! diagnosis (along with pid) listed in each record; so each
beneficiary
166      !  has as many records as there are diagnoses;
167        since the raw datasets have multiple records per person, create
rowid
167      !  identifier;
168        */
169        data hcsrn01a;
170        set hcsrn01;
171        rowid = _N_;

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01.
NOTE: The data set WORK.HCSRN01A has 8352518 observations and 7 variables.
NOTE: DATA statement used:
      real time           3:25.11
      cpu time            51.28 seconds


172        proc sort data=hcsrn01a; by pid rowid; run;

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01A.
NOTE: The data set WORK.HCSRN01A has 8352518 observations and 7 variables.
NOTE: PROCEDURE SORT used:
```

```
      real time           9:28.10
      cpu time            2:26.76


173        proc print data=hcsrn01a (obs=20); run;

NOTE: There were 20 observations read from the data set WORK.HCSRN01A.
NOTE: The PROCEDURE PRINT printed page 4.
NOTE: PROCEDURE PRINT used:
      real time           0.06 seconds
      cpu time            0.01 seconds


174        proc transpose data=hcsrn01a out=hcsrn01b (rename=(col1=diag));
175        var dx:;
176        by pid rowid;
177        run;

NOTE: There were 8352518 observations read from the data set WORK.HCSRN01A.
NOTE: The data set WORK.HCSRN01B has 41762590 observations and 4 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           18:46.07
      cpu time            13:54.87


178        data saswork.combhcsrn;
179        set hcsrn01b (keep=pid diag);
180        if diag NE " ";
181        run;

NOTE: There were 41762590 observations read from the data set WORK.HCSRN01B.
NOTE: The data set SASWORK.COMBHCSRN has 10915624 observations and 2 variables.
NOTE: DATA statement used:
      real time           4:28.26
      cpu time            2:06.90


182        proc print data=saswork.combhcsrn (obs=10); run;

NOTE: There were 10 observations read from the data set SASWORK.COMBHCSRN.
NOTE: The PROCEDURE PRINT printed page 5.
NOTE: PROCEDURE PRINT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


183        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512
```

```
12 The SAS System                              17:08 Thursday, September 30, 2004

#  Name      Memtype    File Size  Last Modified
-----------------------------------------------------------
1  COMBDATA  DATA       186966016   30SEP2004:17:14:00
2  HCSRN01   DATA       362045440   30SEP2004:19:40:15
3  HCSRN01A  DATA       471900160   30SEP2004:19:53:09
4  HCSRN01B  DATA      1685331968   30SEP2004:20:11:56
184        delete hcsrn01 hcsrn01a hcsrn01b;
185        run;

NOTE: Deleting WORK.HCSRN01 (memtype=DATA).
NOTE: Deleting WORK.HCSRN01A (memtype=DATA).
NOTE: Deleting WORK.HCSRN01B (memtype=DATA).
186
187        *merge the four diagnoses data sets & rename pid as recipno;

NOTE: PROCEDURE DATASETS used:
      real time           2.23 seconds
      cpu time            1.90 seconds


188        data combdiag1;
189          set saswork.combsidr saswork.combsadr saswork.combhcsri
189      ! saswork.combhcsrn;
190          rename pid = recipno;

NOTE: There were 270836 observations read from the data set SASWORK.COMBSIDR.
NOTE: There were 15163862 observations read from the data set SASWORK.COMBSADR.
NOTE: There were 220499 observations read from the data set SASWORK.COMBHCSRI.
NOTE: There were 10915624 observations read from the data set
SASWORK.COMBHCSRN.
NOTE: The data set WORK.COMBDIAG1 has 26570821 observations and 2 variables.
NOTE: DATA statement used:
      real time           7:06.44
      cpu time            1:54.93


191        proc sort data=combdiag1; by recipno; run;

NOTE: There were 26570821 observations read from the data set WORK.COMBDIAG1.
NOTE: The data set WORK.COMBDIAG1 has 26570821 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           21:48.13
      cpu time            5:48.64


192
193        /* reshape combdiag1 to create dod2diag which is wider to conserve
193      ! space;
194        found that the median number of claims/episodes among those with a
194      ! claim is 6; so six diagnosis fields are allowed (diag1, diag2, ...,
194      ! diag6) per line;
195        proc transpose is used to do this -- to facilitate this first create
195      ! two indicators:
196
197        diagnum : diagnosis field to which this diagnosis is assigned (1 to
197      ! 6);
198
199        linenum : if there are 6 or fewer diagnoses then only one line is
```

```
13 The SAS System                              17:08 Thursday, September 30, 2004

199      ! needed, else multiple lines may be needed;
200        */
201        data combdiag2;
202          set combdiag1;
203        * rnum is the record number in the set file -- this essentially
gives
203      !  the number of diagnoses for each beneficiary;
204        retain rnum;
205        by recipno;
206        if first.recipno then rnum=1;
207        else rnum = rnum+1;
208        linenum =  int((rnum-1)/6) + 1;
209        diagnum = rnum - (linenum-1)*6;
210

NOTE: There were 26570821 observations read from the data set WORK.COMBDIAG1.
NOTE: The data set WORK.COMBDIAG2 has 26570821 observations and 5 variables.
NOTE: DATA statement used:
      real time           14:12.20
      cpu time            3:38.89


211        proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name       Memtype    File Size  Last Modified
-----------------------------------------------------------
1  COMBDATA   DATA       186966016   30SEP2004:17:14:00
2  COMBDIAG1  DATA       697671680   30SEP2004:20:45:21
3  COMBDIAG2  DATA      1501175808   30SEP2004:20:59:33
212        delete combdiag1;
213        run;

NOTE: Deleting WORK.COMBDIAG1 (memtype=DATA).

NOTE: PROCEDURE DATASETS used:
      real time           0.67 seconds
      cpu time            0.43 seconds


214        proc datasets library=saswork;
         -----Directory-----

Libref:            SASWORK
Engine:            V8
Physical Name:     /data/dod1/saswork
File Name:         /data/dod1/saswork
Inode Number:      675905
```

```
         -----Directory-----

Access Permission: rwxrwxrwx
Owner Name:        amresh
File Size (bytes): 3072


#  Name       Memtype   File Size  Last Modified
-----------------------------------------------------
1  COMBHCSRI  DATA        5128192   30SEP2004:19:14:54
2  COMBHCSRN  DATA      253329408   30SEP2004:20:16:24
3  COMBSADR   DATA      412712960   30SEP2004:19:13:38
4  COMBSIDR   DATA        7127040   30SEP2004:17:15:10
5  DOD2       DATA      389341184   24SEP2004:10:52:51
6  DOD2CHARB  DATA      205250560   30SEP2004:13:40:22
7  DOD2DIAGB  DATA      475840512   30SEP2004:13:19:07
8  DOD3CDPSB  DATA      136552448   24SEP2004:09:16:29
9  DOD3DCG    DATA      106471424   24SEP2004:12:09:46
215        delete combsidr combsadr combhcsri combhcsrn;
216        run;

NOTE: Deleting SASWORK.COMBSIDR (memtype=DATA).
NOTE: Deleting SASWORK.COMBSADR (memtype=DATA).
NOTE: Deleting SASWORK.COMBHCSRI (memtype=DATA).
NOTE: Deleting SASWORK.COMBHCSRN (memtype=DATA).
217

NOTE: PROCEDURE DATASETS used:
      real time           1.06 seconds
      cpu time            0.43 seconds


218        proc sort data= combdiag2; by recipno linenum diagnum; run;

NOTE: There were 26570821 observations read from the data set WORK.COMBDIAG2.
NOTE: The data set WORK.COMBDIAG2 has 26570821 observations and 5 variables.
NOTE: PROCEDURE SORT used:
      real time           39:59.03
      cpu time            8:21.45


219
220        proc transpose data=combdiag2 out=saswork.dod2diagb
220      ! (rename=(col1=diag1 col2=diag2 col3=diag3 col4=diag4 col5=diag5
220      ! col6=diag6));
221        var diag;
222        by recipno linenum;
223        run;

NOTE: There were 26570821 observations read from the data set WORK.COMBDIAG2.
NOTE: The data set SASWORK.DOD2DIAGB has 5343739 observations and 9 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           13:34.12
      cpu time            10:27.56


224
```

```
15 The SAS System                              17:08 Thursday, September 30, 2004

225        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name       Memtype    File Size  Last Modified
-----------------------------------------------------------
1  COMBDATA   DATA       186966016   30SEP2004:17:14:00
2  COMBDIAG2  DATA      1501175808   30SEP2004:21:39:33
226        delete combdiag2;
227        run;

NOTE: Deleting WORK.COMBDIAG2 (memtype=DATA).
228
229        /* split the datasets into estimation and validation samples;
230        validation sample is 0.5 million and the remainder is estimation
230      ! sample
231        */
232        * rename pid as recipno in combdata;

NOTE: PROCEDURE DATASETS used:
      real time           1.62 seconds
      cpu time            1.21 seconds


233        data combdata;
234          set combdata;
235          rename pid = recipno;
236
237        * select validation sample of 0.5 million;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: The data set WORK.COMBDATA has 2304926 observations and 8 variables.
NOTE: DATA statement used:
      real time           1:53.93
      cpu time            19.28 seconds


238        proc surveyselect data=combdata out=valdata seed=2983553
238      ! sampsize=500000; run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: The data set WORK.VALDATA has 500000 observations and 8 variables.
NOTE: The PROCEDURE SURVEYSELECT printed page 6.
NOTE: PROCEDURE SURVEYSELECT used:
      real time           23.31 seconds
      cpu time            11.14 seconds
```

```
16 The SAS System                              17:08 Thursday, September 30, 2004

239        * identify the estimation sample -- create flag (datatype) that
239      ! indicates estimation sample;
240        proc sort data=valdata; by recipno; run;

NOTE: There were 500000 observations read from the data set WORK.VALDATA.
NOTE: The data set WORK.VALDATA has 500000 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           44.00 seconds
      cpu time            9.16 seconds


241        proc sort data=combdata; by recipno; run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: The data set WORK.COMBDATA has 2304926 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           3:18.16
      cpu time            44.93 seconds


242        data estdata;
243          merge combdata (in=infirst) valdata (in=insecond);
244          by recipno;
245          if infirst=1 & insecond=0;
246          datatype=1;
247          label datatype='1=fitting sample, 2=Validation sample';
248        run;

NOTE: There were 2304926 observations read from the data set WORK.COMBDATA.
NOTE: There were 500000 observations read from the data set WORK.VALDATA.
NOTE: The data set WORK.ESTDATA has 1804926 observations and 9 variables.
NOTE: DATA statement used:
      real time           1:10.29
      cpu time            28.99 seconds


249        *combine valdata and estdata to get dod2charb -- modify datatype to
249      ! also identify valdata;
250        data saswork.dod2charb;
251          set estdata valdata;
252          if datatype=. then datatype=2;
253          keep recipno aidcat age male elig expyr expyrt25 expyrt50
datatype;
254        run;

NOTE: There were 1804926 observations read from the data set WORK.ESTDATA.
NOTE: There were 500000 observations read from the data set WORK.VALDATA.
NOTE: The data set SASWORK.DOD2CHARB has 2304926 observations and 9 variables.
NOTE: DATA statement used:
      real time           1:31.43
      cpu time            20.32 seconds


255        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
```

```
17 The SAS System                              17:08 Thursday, September 30, 2004

                     -----Directory-----

Physical Name:     /data/saswork/SAS_workFDBE00006DD8_genmed2
File Name:         /data/saswork/SAS_workFDBE00006DD8_genmed2
Inode Number:      309632
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype   File Size  Last Modified
-----------------------------------------------------
1  COMBDATA  DATA      186966016   30SEP2004:21:59:30
2  ESTDATA   DATA      160735232   30SEP2004:22:00:40
3  VALDATA   DATA       40566784   30SEP2004:21:56:12
256        delete combdata valdata estdata;
257        run;

NOTE: Deleting WORK.COMBDATA (memtype=DATA).
NOTE: Deleting WORK.VALDATA (memtype=DATA).
NOTE: Deleting WORK.ESTDATA (memtype=DATA).
258

NOTE: PROCEDURE DATASETS used:
      real time           0.56 seconds
      cpu time            0.39 seconds


259        proc contents data=saswork.dod2charb varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.23 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 7.

260        proc contents data=saswork.dod2diagb varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.10 seconds
      cpu time            0.02 seconds

NOTE: The PROCEDURE CONTENTS printed page 8.

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           4:54:10.81
      cpu time            1:50:41.67
```

Attachment A3_ACG
(task4c_acgdata.log)

```
1     SAS System                              17:09 Thursday, October 14, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.


NOTE: SAS initialization used:
      real time           1.56 seconds
      cpu time            0.08 seconds

1
2          /*
3          task4c_acgdata.sas
4          Amresh Hanchate
5          DOD1 Project
6          Oct 13, 2004
7
8          Creates the input files for running DCG for the 2.3 million sample.
9          It takes the input files created for running CDPS (dod2charb,
dod2diagb) and modifies these.
10         Note that recipno is actually the person id (pid), so changed for
conforming to CDPS requirements.
11
12         Input: dod2charb.sas7bdat, dod2diagb.sas7bdat
13         Output: dod2acgb1.txt, dod2acgb2.txt
14
15         */
16
17
18         /* IMPORTANT: to change to full-blown sample
19         1) in Part 1, sort dod2diagb by recipno & linenum to produce temp2
20         */
21
22         options ps=60 ls=80 nocenter;
23         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
24         footnote '/data/dod1/saswork/task4c_acgdata.sas';
25
26
27         /*       Part 1
28           1) the diagnoses file has duplicates included -- drop these
29           2) recalculate median # of diagnoses per person
30           3) rearrange diagnoses so that median # of diagnoses per record
31         */
32         proc sort data=saswork.dod2charb; by recipno; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.25 seconds
```

```
      cpu time            0.02 seconds


33         * to drop duplicate diagnoses first rearrange diagnoses so that one
33       ! diagnosis per record;
34         * to use proc transpose create record id as there could be multiple
34       ! records per person;
35         data temp2a;
36           set saswork.dod2diagb (keep=recipno diag:);
37           linenum = _N_;
38         run;

NOTE: There were 5343739 observations read from the data set SASWORK.DOD2DIAGB.
NOTE: The data set WORK.TEMP2A has 5343739 observations and 8 variables.
NOTE: DATA statement used:
      real time           4:19.75
      cpu time            45.98 seconds


39         proc sort data=temp2a (keep=recipno linenum diag:); by recipno
39       ! linenum; run;

NOTE: There were 5343739 observations read from the data set WORK.TEMP2A.
NOTE: The data set WORK.TEMP2A has 5343739 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           8:34.08
      cpu time            1:48.70


40         proc print data=temp2a (obs=30); run;

NOTE: There were 30 observations read from the data set WORK.TEMP2A.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used:
      real time           0.71 seconds
      cpu time            0.03 seconds


41         proc transpose data=temp2a out=temp2b (rename=(col1=diag));
42           var diag1-diag6;
43           by recipno linenum;
44         run;

NOTE: There were 5343739 observations read from the data set WORK.TEMP2A.
NOTE: The data set WORK.TEMP2B has 32062434 observations and 4 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           22:16.10
      cpu time            10:04.17


45         proc print data=temp2b (obs=50); run;

NOTE: There were 50 observations read from the data set WORK.TEMP2B.
NOTE: The PROCEDURE PRINT printed page 2.
NOTE: PROCEDURE PRINT used:
      real time           0.01 seconds
      cpu time            0.01 seconds
```

```
46         proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work264500000DDC_genmed2
File Name:         /data/saswork/SAS_work264500000DDC_genmed2
Inode Number:      551296
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name    Memtype    File Size  Last Modified
-----------------------------------------------------
1  TEMP2A  DATA       433438720   14OCT2004:17:22:11
2  TEMP2B  DATA      1554186240   14OCT2004:17:44:29
46       !                              delete temp2 temp2a; run;

NOTE: The file WORK.TEMP2 (memtype=DATA) was not found, but appears on a DELETE
      statement.
NOTE: Deleting WORK.TEMP2A (memtype=DATA).
47         * first get rid of blank diagnoses cells;

NOTE: PROCEDURE DATASETS used:
      real time           2.00 seconds
      cpu time            0.28 seconds


48         data temp2c;
49           set temp2b;
50           if diag NE "";
51         run;

NOTE: There were 32062434 observations read from the data set WORK.TEMP2B.
NOTE: The data set WORK.TEMP2C has 26570821 observations and 4 variables.
NOTE: DATA statement used:
      real time           22:41.68
      cpu time            3:00.23


52         * get rid of duplicate diagnoses per person;
53         proc sort data=temp2c (keep=recipno diag) out=temp2d nodupkey; by
53       ! recipno diag; run;

NOTE: 12347905 observations with duplicate key values were deleted.
NOTE: There were 26570821 observations read from the data set WORK.TEMP2C.
NOTE: The data set WORK.TEMP2D has 14222916 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           24:29.16
      cpu time            5:38.17


54         proc datasets library=work;
```

```
4 The SAS System                                    17:09 Thursday, October 14, 200

                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work264500000DDC_genmed2
File Name:         /data/saswork/SAS_work264500000DDC_genmed2
Inode Number:      551296
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name    Memtype    File Size  Last Modified
-----------------------------------------------------
1  TEMP2B  DATA      1554186240   14OCT2004:17:44:29
2  TEMP2C  DATA      1287987200   14OCT2004:18:07:14
3  TEMP2D  DATA       373456896   14OCT2004:18:31:42
54       !                              delete temp2b temp2c; run;

NOTE: Deleting WORK.TEMP2B (memtype=DATA).
NOTE: Deleting WORK.TEMP2C (memtype=DATA).
55         * calculate median number of diagnoses;

NOTE: PROCEDURE DATASETS used:
      real time           12.67 seconds
      cpu time            1.81 seconds


56         data temp3;
57           set temp2d;
58           by recipno;
59           retain numdiag;
60           if first.recipno then numdiag=1;
61           else numdiag = numdiag+1;
62           if last.recipno then output;
63           keep recipno numdiag;
64         run;

NOTE: There were 14222916 observations read from the data set WORK.TEMP2D.
NOTE: The data set WORK.TEMP3 has 1976274 observations and 2 variables.
NOTE: DATA statement used:
      real time           1:18.93
      cpu time            53.19 seconds


65         proc freq data=temp3; tables numdiag / missing; run;

NOTE: There were 1976274 observations read from the data set WORK.TEMP3.
NOTE: The PROCEDURE FREQ printed pages 3-5.
NOTE: PROCEDURE FREQ used:
      real time           7.23 seconds
      cpu time            4.59 seconds


66         proc datasets library=work;
```

```
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work264500000DDC_genmed2
File Name:         /data/saswork/SAS_work264500000DDC_genmed2
Inode Number:      551296
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name    Memtype   File Size  Last Modified
-----------------------------------------------
1  TEMP2D  DATA      373456896   14OCT2004:18:31:42
2  TEMP3   DATA       63750144   14OCT2004:18:33:16
66       !                            delete temp3; run;

NOTE: Deleting WORK.TEMP3 (memtype=DATA).
67         /* The median # is 5 -- so re-arrange temp2d so that 5 diagnoses per
67       ! record per person (diag1, diag2, ..., diag5);
68         proc transpose is used to do this -- to facilitate this first create
68       ! two indicators:
69
70         diagnum : diagnosis field to which this diagnosis is assigned (1 to
70       ! 5);
71
72         linenum : if there are 5 or fewer diagnoses then only one line is
72       ! needed, else multiple lines may be needed;
73         */

NOTE: PROCEDURE DATASETS used:
      real time           0.19 seconds
      cpu time            0.08 seconds


74         data temp2e;
75           set temp2d;
76         * rnum is the record number in the set file -- this essentially
gives
76       !  the number of diagnoses for each beneficiary;
77           retain rnum;
78           by recipno;
79           if first.recipno then rnum=1;
80           else rnum = rnum+1;
81           linenum =  int((rnum-1)/5) + 1;
82           diagnum = rnum - (linenum-1)*5;
83

NOTE: There were 14222916 observations read from the data set WORK.TEMP2D.
NOTE: The data set WORK.TEMP2E has 14222916 observations and 5 variables.
NOTE: DATA statement used:
      real time           16:09.14
      cpu time            2:06.22


84         proc datasets library=work;
```

```
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work264500000DDC_genmed2
File Name:         /data/saswork/SAS_work264500000DDC_genmed2
Inode Number:      551296
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name    Memtype   File Size  Last Modified
-----------------------------------------------
1  TEMP2D  DATA      373456896   14OCT2004:18:31:42
2  TEMP2E  DATA      803561472   14OCT2004:18:49:33
84       !                            delete temp2d; run;

NOTE: Deleting WORK.TEMP2D (memtype=DATA).
85

NOTE: PROCEDURE DATASETS used:
      real time           2.10 seconds
      cpu time            0.22 seconds


86         proc sort data= temp2e; by recipno linenum diagnum; run;

NOTE: There were 14222916 observations read from the data set WORK.TEMP2E.
NOTE: The data set WORK.TEMP2E has 14222916 observations and 5 variables.
NOTE: PROCEDURE SORT used:
      real time           37:58.08
      cpu time            4:34.90


87
88         proc transpose data=temp2e out=temp2f (rename=(col1=diag1 col2=diag2
88       ! col3=diag3 col4=diag4 col5=diag5));
89           var diag;
90           by recipno linenum;
91         run;

NOTE: There were 14222916 observations read from the data set WORK.TEMP2E.
NOTE: The data set WORK.TEMP2F has 3724179 observations and 8 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           24:58.53
      cpu time            7:28.73


92         proc print data=temp2e (obs=50); run;

NOTE: There were 50 observations read from the data set WORK.TEMP2E.
NOTE: The PROCEDURE PRINT printed page 6.
NOTE: PROCEDURE PRINT used:
      real time           0.04 seconds
      cpu time            0.02 seconds
```

```
93         proc print data=temp2f (obs=5); run;

NOTE: There were 5 observations read from the data set WORK.TEMP2F.
NOTE: The PROCEDURE PRINT printed page 7.
NOTE: PROCEDURE PRINT used:
      real time           0.00 seconds
      cpu time            0.01 seconds


94         proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work264500000DDC_genmed2
File Name:         /data/saswork/SAS_work264500000DDC_genmed2
Inode Number:      551296
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name    Memtype   File Size  Last Modified
-----------------------------------------------------
1  TEMP2E  DATA      803561472   14OCT2004:19:27:32
2  TEMP2F  DATA      302080000   14OCT2004:19:52:34
94       !                              delete temp2e; run;

NOTE: Deleting WORK.TEMP2E (memtype=DATA).
95
96         /*       Part 2
97         write the person-level file (dod2acgb1.txt); note that
98           recipno is 18-long char
99           age is 8-long numeric -- write it out as 2-digit right-justified
99       ! (max age being 64)
100          male is 8-long numeric (so when printed using put it is
100      ! automatically left-justified; 1=male 2=female)
101        */

NOTE: PROCEDURE DATASETS used:
      real time           2.56 seconds
      cpu time            0.53 seconds


102        data _NULL_;
103          set saswork.dod2charb;
104          file '/data/dod1/saswork/dod2acgb1.txt';
105          put @1 recipno @19 age 2. @21 male 1.;
106        run;

NOTE: The file '/data/dod1/saswork/dod2acgb1.txt' is:
      File Name=/data/dod1/saswork/dod2acgb1.txt,
      Owner Name=amresh,Group Name=sas,
      Access Permission=rw-rw----

NOTE: 2304926 records were written to the file
```

```
      '/data/dod1/saswork/dod2acgb1.txt'.
      The minimum record length was 21.
      The maximum record length was 21.
NOTE: There were 2304926 observations read from the data set SASWORK.DOD2CHARB.
NOTE: DATA statement used:
      real time           1:44.27
      cpu time            29.05 seconds


107
108        /*       Part 3
109        write the diagnoses file (dod2acgb2.txt); note that
110          recipno is 18-long char
111          each diagx is 8-char long
112
113        */
114        data _NULL_;
115          set temp2f;
116          file '/data/dod1/saswork/dod2acgb2.txt';
117          put @1 recipno @19 diag1 @27 diag2 @35 diag3 @43 diag4 @51 diag5;
118        run;

NOTE: The file '/data/dod1/saswork/dod2acgb2.txt' is:
      File Name=/data/dod1/saswork/dod2acgb2.txt,
      Owner Name=amresh,Group Name=sas,
      Access Permission=rw-rw----

NOTE: 3724179 records were written to the file
      '/data/dod1/saswork/dod2acgb2.txt'.
      The minimum record length was 51.
      The maximum record length was 55.
NOTE: There were 3724179 observations read from the data set WORK.TEMP2F.
NOTE: DATA statement used:
      real time           2:13.43
      cpu time            51.36 seconds


119

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           2:47:19.38
      cpu time            37:48.48
```

Attachment A3_DCG
(dcgfiles.log)

```
1   The SAS System                                14:28 Tuesday, August 17, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



WARNING: Your system is scheduled to expire on September 12, 2004, which is 26
days from now.  Please contact your SAS Software
         Representative to obtain your updated SETINIT information.  The SAS
System will no longer function on or after that date.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.81 seconds
      cpu time            0.06 seconds

1          /*************************************************/
2          /* Creating person & diagnostic file to use for     */
3          /* to use for the DCG model for DOD1 project.      */
4          /* These files are created using the cdps files    */
5          /* and making the appropriate adjustments to the   */
6          /* variables.                                   */
7          /* Jenn Fonda                                      */
8          /* Created: August 5, 2004                      */
9          /* Last modified: August 17, 2004                  */
10         /*************************************************/
11         libname dod1 '/data/dod1/saswork/';
NOTE: Libref DOD1 was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
12         libname out '/data/dod1/saswork/fonda';
NOTE: Libref OUT was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork/fonda
13         options ps=52 ls=80 compress=yes;
14         %include '/data/dod1/saswork/formats.sas';
WARNING: The Base Product product with which FORMAT is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
```

```
NOTE: PROCEDURE FORMAT used:
      real time           1.26 seconds
      cpu time            0.02 seconds


44         footnote '/data/dod1/saswork/dcgfiles.sas';
45
46
47         ** Use CDPS 'dod2char' dataset to create DCG person file;
48         data out.pers; set dod1.dod2char;
WARNING: The Base Product product with which DATASTEP is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.
49           ** Rename variables to fit DCG recommendations;
50           rename elig=ELIG1;
51           rename recipno = IDNO;
52           rename expyr = EXPEND2;
53           if male=1 then SEX='1'; else SEX='2';
54           drop male aidcat expyrt25 expyrt50;
55         run;

NOTE: There were 2304926 observations read from the data set DOD1.DOD2CHAR.
NOTE: The data set OUT.PERS has 2304926 observations and 6 variables.
NOTE: Compressing data set OUT.PERS decreased size by 0.84 percent.
      Compressed is 14375 pages; un-compressed would require 14497 pages.
NOTE: DATA statement used:
      real time           4:00.55
      cpu time            36.07 seconds


56
57         proc sort data=out.pers; by IDNO; run;
WARNING: The Base Product product with which SORT is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.

NOTE: There were 2304926 observations read from the data set OUT.PERS.
NOTE: The data set OUT.PERS has 2304926 observations and 6 variables.
NOTE: Compressing data set OUT.PERS decreased size by 0.84 percent.
      Compressed is 14375 pages; un-compressed would require 14497 pages.
NOTE: PROCEDURE SORT used:
      real time           4:53.42
      cpu time            1:03.77


58
59         proc contents data=out.pers;
WARNING: The Base Product product with which CONTENTS is associated will expire
         within 30 days. Please contact your SAS installation representative to
```

```
         have it renewed.
60
61         ** Check sex, elig1 and elig2 variables to see if transformed
61       ! correctly;

NOTE: The PROCEDURE CONTENTS printed page 1.
NOTE: PROCEDURE CONTENTS used:
      real time           0.79 seconds
      cpu time            0.03 seconds


62         proc freq data=out.pers;
WARNING: The Base Product product with which FREQ is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.
63           tables SEX ELIG1;
64         run;

NOTE: There were 2304926 observations read from the data set OUT.PERS.
NOTE: The PROCEDURE FREQ printed page 2.
NOTE: PROCEDURE FREQ used:
      real time           17.30 seconds
      cpu time            16.03 seconds


65
66
67         ** Make diagnoses left justified as specified in DCG manual;
68         data out.diag; set dod1.dod2diag;
WARNING: The Base Product product with which DATASTEP is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.
69           rename recipno=IDNO;
70           rename diag1=DIAG1;
71           rename diag2=DIAG2;
72           rename diag3=DIAG3;
73           rename diag4=DIAG4;
74           rename diag5=DIAG5;
75           rename diag6=DIAG6;
76
77           DIAG1=left(DIAG1);
78           DIAG2=left(DIAG2);
79           DIAG3=left(DIAG3);
80           DIAG4=left(DIAG4);
81           DIAG5=left(DIAG5);
82           DIAG6=left(DIAG6);
83         run;

NOTE: There were 5343739 observations read from the data set DOD1.DOD2DIAG.
NOTE: The data set OUT.DIAG has 5343739 observations and 7 variables.
```

```
NOTE: Compressing data set OUT.DIAG decreased size by 7.25 percent.
      Compressed is 40297 pages; un-compressed would require 43446 pages.
NOTE: DATA statement used:
      real time           7:54.41
      cpu time            1:59.86


84
85         proc sort data=out.diag; by IDNO; run;
WARNING: The Base Product product with which SORT is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.

NOTE: There were 5343739 observations read from the data set OUT.DIAG.
NOTE: The data set OUT.DIAG has 5343739 observations and 7 variables.
NOTE: Compressing data set OUT.DIAG decreased size by 7.25 percent.
      Compressed is 40297 pages; un-compressed would require 43446 pages.
NOTE: PROCEDURE SORT used:
      real time           12:11.94
      cpu time            2:48.13


86         proc contents data=out.diag; run;
WARNING: The Base Product product with which CONTENTS is associated will expire
         within 30 days. Please contact your SAS installation representative to
         have it renewed.

NOTE: PROCEDURE CONTENTS used:
      real time           0.27 seconds
      cpu time            0.02 seconds

NOTE: The PROCEDURE CONTENTS printed page 3.

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           29:22.69
      cpu time            6:44.02
```

Attachment A3_CRG_1
CRG Risk Grouping Data Specification
The objectives of this data run are two-fold; to provide data to 3M to be used in their CRG grouping software, and to provide the Risk Assessment team with the same data, but without the specific formatting details required by the groupers. Essentially, what this means is that several data elements will be added to the Risk Assessment Data, and that a special output will be prepared from the new Risk Assessment files for grouping purposes.

Update to Risk Assessment Data Files
The first step to the project will be to run the same code and the same files used in the last data run, but to incorporate the following additional data elements:

File Type	Data Element	Note
SIDR	Disposition Date	YYYYMMDD
SADR	Provider Specialty Code	No transformation
SADR	Encounter Date	YYYYMMDD
HCSR-I	Institution Type	No transformation
HCSR-I	End Date of Care	YYYYMMDD
HCSR-N	Provider Specialty Code	No transformation
HCSR-N	Place of Service	No transformation
HCSR-N	End Date of Care	No transformation

These data are to be provided to BU/RTI.

Preparation of files for CRG Grouper
There are 5 record types (A, B, C, E and M) that need to be provided to the CRG Grouper. Each record type is described below. Though the process for preparing files for the grouper is a bit cumbersome, in the end, **only two files are provided to the grouper**. These files are prepared by combining the various record types, as described below.

The first file contains the A, B and C records (only three rows in that file), and the second will contain all of the E and M records, sorted by person identifier and record type.

File 1:
A record
B record
C record

File 2:
E record for enrollee 1
M record(s) for enrollee 1 (if there are any)
E record for enrollee 2
M record(s) for enrollee 2 (if there are any)
Etc…

Step 1: Prepare A, B and C records. Three rows of data in one file.

A Record:
This record type contains information about the formats of important data elements in the data feed. The data provided in subsequent records need to use the formats specified. There is only one A record created, representing information about all the records sent to the grouper.

A Record			
Data Element	**Rule**	**Format**	**Position**
Record Type	Set to A	$1	1
Date format code of service dates on B record	1	$1	2
Date format of the birth date field on E record	1	$1	3
Date format of the diagnosis code on M record	1	$1	4
Enrollee ID Length	18	$2	5-6
Length of data sequencing field (?)	00	$2	7-8
Number of bytes of user patient data included on the E record.	000	$3	9-11
Number of bytes of user patient data included on the M record.	000	$3	12-14
Filler	Blank	$24	15-39

B Record:
This record defines the service period in the batch of records sent to the grouper. There is only one B record created, representing information about all the records sent to the grouper.

Record Type B:

B Record			
Data Element	**Rule**	**Format**	**Position**
Record Type	Set to B	$1	1
Service beginning date	Earliest date of service in M file	YYYYMMDD	2
Service cutoff date	Latest date of service in M file	YYYYMMDD	10
Filler	Blank	$21	18-39

C Record:
This record defines the type of output we'd like the grouper to send us. Will assume we want each record type.

Record Type C:

C Record			
Data Element	**Rule**	**Format**	**Position**
Record Type	Set to C	$1	1
How used code	01	$2	2
Record Types	Blank fill	$17	4
Filler	Blank fill	$18	21-39

Step 2: Prepare enrollee data.

Record Type E: This record contains enrollee information. This file should be created from the enrollment file prepared for the Risk Assessment Study. There should be one E record per enrollee.

E Record			
Data Element	**Rule**	**Format**	**Position**
Record Type	Set to E	$1	1
Date of Birth	YYYYMMDD	YYYYMMDD	2-9
Sex	If TEF code is M then set to 1, else if TEF code is F then set to 2, else set to 0	$1	10
Filler	Blank fill	$10	11
Unique Enrollee ID	Person ID from Risk Assessment Data	$18	21-39

Step 3: Prepare M records from the SIDR, SADR, HCSR-I and HCSR-N records.

M Records:
The M records contain the health care data for each of the enrollees. M records are provided for each separate (non-blank) diagnosis or procedure code. For example, if a patient had one record with 2 diagnoses and 3 procedures, that one record would translate to 5 M records. This can be readily accomplished using the "output" function in SAS, after checking the diagnosis or procedure codes to ensure they are populated.

The processing required is described below.

- SIDR Records: One record per dx/px.

SIDR Records			
Data Element	**Rule**	**Format**	**Position**
Record Type	M	$1	1
Diagnosis/Procedure Code	Total field length of 5. Left justified. First 5 characters of diagnosis code. First 4 characters of procedure code.	$7	2-8
Code type indicator	1 if record is created from diagnosis 1	$1	9

SIDR Records			
	(principal diagnosis); 2 if record is created from any other diagnosis code; 3 if record is created from any of the procedure codes		
Type of Provider	Set = 1 (every record is from a hospital)	$1	10
Site of Service	Set = 04 (inpatient)	$2	11-12
Date of Service	Disposition Date	YYYYMMDD	13-20
Person ID	Create per Risk Assessment Specs	$18	21-39

- SADR Records: One per dx/px/E&M

SADR Records			
Data Element	**Rule**	**Format**	**Position**
Record Type	M	$1	1
Diagnosis/Procedure Code	Total field length of 7. Left justified. First 5 characters of diagnosis code or entire procedure code	$7	2-8
Code type indicator	'1' if record is created from diagnosis 1; '2' if record is created from any other diagnosis code; '4' if record is created from any of the procedure codes (including the E&M procedure code)	$1	9
Type of Provider	See mapping in appendix	$1	10
Site of Service	07	$2	11-12
Date of Service	Encounter Date	YYYYMMDD	13-20
Person ID	Create per Risk Assessment Specs	$18	21-39

- HCSR-I Records:

HCSR-I Records			
Data Element	**Rule**	**Format**	**Position**
Record Type	M	$1	1
Diagnosis/Procedure Code	First 5 characters of diagnosis code or first 4 characters of procedure code	$7	2-8
Code type indicator	1 if record is created from diagnosis 1; 2 if record is created from any other diagnosis code; 3 if record is created from any of the procedure codes	$1	9
Type of Provider	Set equal to 1	$1	10
Site of Service	Mapping in appendix, based on institution type in HCSR. If institution type is blank, or not on list in appendix, set to 09	$2	11-12
Date of Service	End Date of Care	YYYYMMDD	13-20
Person ID	Create per Risk Assessment Specs	$18	21-39

- HCSR-N Records: First split HCSRs into one record per claim per line item. (Like M2). From there, one record per recorded (test if blank prior to outputting) diagnosis and procedure.

HCSR-N Records			
Data Element	**Rule**	**Format**	**Position**
Record Type	M	$1	1
Diagnosis/Procedure Code	First 5 characters of diagnosis code or procedure code	$7	2-8
Code type indicator	'1' if record is created from diagnosis 1; '2' if record is created from any other diagnosis code; '4' if record is created from any of the procedure codes (including the E&M procedure code)	$1	9
Type of Provider	Mapping in appendix, based on provider specialty code. If provider specialty code is blank or not on list in appendix, set to 4.	$1	10
Site of Service	Mapping in appendix, based on place of service variable in each line item. If place of service is blank or not on list in appendix, set to 4.	$	11-12
Date of Service	End Date of Care	YYYYMMDD	13-20
Person ID	Create per Risk Assessment Specs	$18	21-39

Step 4: Concatenate E records and M records and sort

Append the E records and the M records together and sort by person ID and record type.

Mapping of SADR Provider Specialty Codes to “Type of Provider” for Risk Grouper.

Code	Description	CRG Prov Type
000	General Medical Officer	2
001	Family Practice Physician	2
002	Contract Physician	2
003	**Family Practice Physician Resident**	2
004	Emergency Physician	2
005	Emergency Physician Resident	2
011	Internist	2
012	Allergist	2
013	Oncologist	2
014	Cardiologist	2
015	Cardiopulmonary Laboratory Physician	2
016	Endocrinologist	2
017	Geriatrician	2
018	Gastroenterologist	2
019	Hematolgoist	2
020	Rheumatolgoist	2
021	Pulmonary Disease Physician	2
022	Infectious Disease Physician	2
023	Metabolic Disease Physician	2
024	Nephrologist	2
025	Medical Geneticist	2
026	Tropical Medicine Physician	2
027	Nuclear Medicine Physician	2
028	Internal Medicine Resident	2
040	Pediatrician	2
041	Pediatric Allergist	2
042	Adolescent Medicine Physician	2
043	Pediatric Cardiologist	2
044	Pediatric Dermatologist	2
045	Pediatric Endocrinologist	2
046	Perinatolgoist	2
047	Pediatric Metabolic Disease Physician	2
048	Pediatric Hematologist	2
049	Pediatric Neurologist	2
050	Pediatric Pulmonary Disease Physician	2
051	Pediatric Infectious Disease Physician	2
052	Pediatric Resident	2
053	Pediatric Gastroenterologist	2
054	Pediatric Nephrologist	2
060	Neurologist	2
061	Neurologist Resident	2
070	Psychiatrist	2
071	Child Psychiatrist	2
072	**Psychoanalyst**	2
073	**Psychiatric Resident**	2
074	Alcohol Abuse Counselor	3
075	Drug Abuse Counselor	3
080	Dermatologist	2
081	Dermatologist Resident	2
090	Physical Medicine Physician	2

Code	Description	CRG Prov Type
091	Special Weapons Defense Physician	2
092	Anesthesiologist	2
093	Anesthesiology Resident	2
094	Anesthetist	3
100	General Surgeon	2
101	Thoracic Surgeon	2
102	Colon & Rectal Surgeon	2
103	Cardiac Surgeon	2
104	Pediatric Surgeon	2
105	Peripheral Vascular Surgeon	2
106	Neurological Surgeon	2
107	Plastic Surgeon	2
108	Resident Surgeon	2
109	Burn Therapist	2
110	Urologist	2
111	Urology Resident	2
115	Plastic Surgery Resident	2
120	Ophthalmologist	2
121	Ophthalmology Resident	2
130	Otorhinolaryngologist	2
131	Otorhinolaryngology Resident	2
140	Orthopedic Surgeon	2
141	Hand Surgeon	2
142	Orthopedic Resident	2
150	Obstetrician/Gynecologist (OB/GYN)	2
151	Endocrinologist, OB/GYN	2
152	Oncologist, OB/GYN	2
153	Pathologist, OB/GYN	2
154	OB/GYN Resident	2
200	Pathologist	2
202	Medical Chemist	3
203	Medical Microbiologist	3
204	Forensic Pathologist	3
205	Neuropathologist	2
206	Nuclear Medicine Pathologist	2
207	Pathology Resident	2
208	Histopathologist	2
210	Biomedical Lab Officer	3
211	Biomedical Lab Science Officer	3
212	Microbiology Lab Officer	3
213	Chemistry Lab Officer	3
214	Blood Bank Officer	3
215	Clinical Lab Officer, Other	3
300	Aerospace Medicine Physician	2
301	Aerospace Medicine Resident	2
302	Aerospace Med Flight Surgeon/Family Practice	2
320	Preventive Medicine Physician	2
321	Occupational Medicine Physician	2
322	Hyperbaric/Underseas Medicine Physician	2
400	Radiologist	2
401	Radiation Therapist	3
402	Neuro-Radiologist	2
403	Nuclear Medicine Radiologist	2

Code	Description	CRG Prov Type
404	Diagnostic Radiologist	2
405	Special Procedures Radiologist	2
406	Radiology Resident	2
407	Radiophysicist	3
500	Senior Staff Physician	2
501	Anesthesiology Consultant	2
502	Internal Medicine Consultant	2
503	Pediatric Medicine Consultant	2
504	Neurology Consultant	2
505	Psychology Consultant	2
506	Dermatology Consultant	2
507	Physical Medicine Consultant	2
508	Surgery Consultant	2
509	Urology Consultant	2
510	Ophthalmology Consultant	2
511	Otorhinolarynology Consultant	2
512	Orthopedic Surgery Consultant	2
513	OB/GYN Consultant	2
514	Aerospace Medicine Consultant	2
515	Preventive Medicine Consultant	2
516	Radiology Consultant	2
517	Dental Consultant	2
518	Other Consultant	2
600	Nurse, General Duty	3
601	Mental Health Nurse	3
602	OB/GYN Nurse Practitioner	3
603	Pediatric Nurse Practitioner	3
604	Primary Care Nurse Practitioner Qualified	3
605	Primary Care Nurse Practitioner – Entry	3
606	Aerospace Nurse	3
607	Community Health Nurse	3
608	Certified Nurse Midwife	3
609	Nurse Midwife – Entry Level	3
610	Clinical Nurse- Entry Level for Nurse Practitioner	3
611	Psychiatric Nurse Practitioner	3
612	Nurse Anesthetist	3
700	Other Provider (Officer)	3
701	Aerospace Physiologist	3
702	Clinical Psychologist	2
703	Psychology Worker	3
704	Dietician – Nutritionist	3
705	Occupational Therapist	3
706	Physical Therapist	3
707	Podiatrist	3
708	Optometrist	3
709	Audiologist	3
710	Speech Therapist	3
711	Other Biomedical Specialist	3
713	Contract Chiropractor	3
800	Oral Surgeon	2
801	Oral Surgery Resident	2
802	Periodontist	2
803	Periodontic Resident	2
804	Prosthodontist	2

Code	Description	CRG Prov Type
805	Prosthodontic Resident	2
806	Orthodontist	2
807	Orthodontic Resident	2
808	Oral Pathologist	2
809	Oral Pathology Resident	2
810	Endodontist	2
811	Endodontic Resident	2
812	Dental Officer General	2
813	Dental Officer Resident	2
814	Dental Staff Officer	3
815	Pedodontist	2
816	Pedodontic Resident	2
900	Corpsman/Technician	3
901	Physician Assistant	3
902	Dental Assistant	3
905	Cardiopulmonary Lab Technician	3
910	Adolescent Medicine	3
911	Aerospace Medicine	3
912	Allergy	3
913	Anesthesiology	3
914	Audiology	3
915	Cardiology	3
916	Community Health	3
917	Critical Care Medicine	3
918	Dental	3
919	Dermatology	3
920	Dietetics	3
921	Emergency Medicine	3
922	Endocrinology	3
923	Family Practice/Primary Care	3
924	Gastroenterology	3
925	General Medicine	3
926	Gerontology/Geriatrics	3
927	Gynecology	3
928	Health Benefits	3
929	Hematology	3
930	Immunology	3
931	Infectious Disease	3
932	Internal Medicine	3
933	Laboratory/Pathology	3
934	Medical Genetics	3
935	Metabolic Disease	3
936	Nephrology	3
937	Neonatal/Perinatal Medicine	3
938	Neurology	3
939	Nuclear Medicine	3
940	Nursing	3
941	Nutrition	3
942	OB/GYN	3
943	Ocuupational Health	3
944	Oncology	3
945	Ophthalmology	3
946	Optometry	3
947	Orthopedics	3

Code	Description	CRG Prov Type
948	Otorhinolaryngology	3
949	Pediatrics	3
950	Physical Medicine and Rehabilitation	3
951	Podiatry	3
952	Preventive Medicine	3
953	Psychiatry	3
954	Psychology	3
955	Pulmonary Disease	3
956	Radiology	3
957	Rheumatology	3
958	Social Work	3
959	Surgery	3
960	Physical Therapy	3
961	Radiation Therapy	3
962	Speech Language Pathology Therapy	3
963	Urology	3
964	Obstetrics	3
965	Sleep Disorders	3
966	Occupational Therapy	3
967	Developmental Pediatrics	3
968	Hyperbaric Medicine	3
969	Respiratory Therapy	3
970	Peripheral Vascular Medicine	3
971	Proctology	3
972	Thoracic Surgery	3
999	Unknown	3

Institution Types in HCSR-I and "Site of Service"

Code	Description	Site
10	General medical and surgical	04
11	Hospital unit of an institution (prison hospital, college infirmary etc.)	04
12	Hospital unit within an institution for the mentally retarded	04
22	Psychiatric hospital or unit of	04
33	Tuberculosis and other respiratory disease	04
44	Obstetrics and gynecology	04
45	Eye, ear, nose and throat	04
46	Rehabilitation	04
47	Orthopedic	04
48	Chronic disease	04
49	Other specialty[1]	04
50	Children's general	04
51	Children's hospital unit of an institution	04
52	Children's psychiatric hospital or unit of	04
53	Children's tuberculosis and other respiratory diseases	04
55	Children's eye, ear, nose, and throat	04
56	Children's rehabilitation	04
57	Children's orthopedic	04

Code	Description	Site
58	Children's chronic	04
59	Children's other specialty[1]	04
62	Institution for mental retardation	04
70	Home Health Care Agency	02
71	Specialized Treatment Facility	04
72	Residential Treatment Center	04
73	Extended Care Facility	04
74	Christian Science Facility	04
75	Hospital based Ambulatory Surgery Center	07
76	Skilled Nursing Facility	08
78	Non-hospital based hospice	03
79	Hospital based hospice	03
82	Substance Use Disorders Rehabilitation Facility (SUDRF)	04
90	Cancer	04
91	Sole community	04
92	Freestanding Ambulatory Surgery Center	07

HCSR-N Place of Service mapping to site of service

Place of Service Code	Site of Service code
11=Office	06
12=Home	02
21=Inpatient Hospital	04
22=Outpatient Hospital	07
23= Emergency Room-Hospital	07
24=Ambulatory Surgery Center	07
25=Birthing Center	04
26=Military Treatment Facility	07
31=Skilled Nursing Facility	08
32=Nursing Facility	05
33=Custodial Care Facility	05
34=Hospice	03
41=Ambulance-Land	09
42=Ambulance-Air or Water	09
51=Inpatient Psych Facility	04
52=Psych Facility Partial Hospitalization	07
53=Community Mental Health Center	09
54=Intermediate Care Fac/Mentally Retarded	05
55=Residential Substance Abuse	05
56=Psych Res Treatment Center	05
61=Comp Inpatient Rehab Facility	04
62=Comp Outpatient Rehab Facility	07
65=End Stage Renal Disease Trt Fac	07
71=State/Local Public Health Clinic	01
72=Rural Health Clinic	01
81=Independent Lab	09
99=Other Unlisted Facility	09

Attachment A3_CRG_2

```
1   The SAS System                          17:13 Wednesday, October 6, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.

NOTE: SAS initialization used:
      real time           0.37 seconds
      cpu time            0.09 seconds

1          /*

2          task8_finder.sas

3          Amresh Hanchate

4          DOD1 Project

5          Oct 6, 2004

6

7          This program produces two finder files to be sent to
Kennell/DOD for identifying the selected 2.3 million sample for CRG
7        ! analysis. The first one is the development (estimation)
sample of 1.8 million and the second one has the 0.5 million
7        ! validation sample.
8

9          Input: dod2charb

10         Output: dod2find_devp.xpt, dod2find_valid.xpt;

11

12         */

13         options ps=60 ls=80 nocenter;

14         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
14       !

15         footnote '/data/dod1/saswork/task8_finder.sas';

16
```

```
17         data dod2sub;

18          set saswork.dod2charb (keep=recipno expyr expyrt50 expyrt25
18       ! datatype);

19          rename recipno=pid;

20          rename expyr = totexp02;

21          rename expyrt50 = totexp02t50;

22          rename expyrt25 = totexp02t25;

23          label expyrt50='totexp02 top-coded at $50K';

24          label expyrt25='totexp02 top-coded at $25K';

25

26


NOTE: There were 2304926 observations read from the data set
SASWORK.DOD2CHARB.
NOTE: The data set WORK.DOD2SUB has 2304926 observations and 5
variables.
NOTE: DATA statement used:
      real time           1:35.42
      cpu time            16.60 seconds


2 The SAS System                        17:13 Wednesday, October 6, 2004


27         data dod2find_devp dod2find_valid;

28          set dod2sub;

29          if datatype=1 then output dod2find_devp;

30          else if datatype=2 then output dod2find_valid;

31          keep pid totexp02 totexp02t50 totexp02t25;

32         run;

NOTE: There were 2304926 observations read from the data set
WORK.DOD2SUB.
NOTE: The data set WORK.DOD2FIND_DEVP has 1804926 observations and 4
variables.
NOTE: The data set WORK.DOD2FIND_VALID has 500000 observations and 4
variables.
NOTE: DATA statement used:
      real time           48.11 seconds
      cpu time            13.78 seconds
```

```
32       !

33         data dod2find_valid;

34          set dod2find_valid (keep=pid);

35         run;

NOTE: There were 500000 observations read from the data set
WORK.DOD2FIND_VALID.
NOTE: The data set WORK.DOD2FIND_VALID has 500000 observations and 1
variables.
NOTE: DATA statement used:
      real time           4.45 seconds
      cpu time            1.85 seconds


35       !

36

37         proc contents data=dod2find_devp varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.13 seconds
      cpu time            0.05 seconds

NOTE: The PROCEDURE CONTENTS printed page 1.

37       !

38         proc contents data=dod2find_valid varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.09 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 2.

38       !

39

40         proc cport data=dod2find_devp
40       ! file='/data/dod1/saswork/dod2find_devp.xpt'; run;


NOTE: Proc CPORT begins to transport data set WORK.DOD2FIND_DEVP
NOTE: The data set contains 4 variables and 1804926 observations.
      Logical record length is 48.
```

```
3 The SAS System                          17:13 Wednesday, October 6, 2004

NOTE: PROCEDURE CPORT used:
      real time           1:15.12
      cpu time            48.44 seconds


41         proc cport data=dod2find_valid
41       ! file='/data/dod1/saswork/dod2find_valid.xpt'; run;


NOTE: Proc CPORT begins to transport data set WORK.DOD2FIND_VALID
NOTE: The data set contains 1 variables and 500000 observations.
      Logical record length is 18.
NOTE: PROCEDURE CPORT used:
      real time           17.09 seconds
      cpu time            10.81 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           4:01.35
      cpu time            1:31.67
```

Attachment A4_CDPS_1
(task5b_cdps.log)

```
1   The SAS System                                11:22 Friday, October 1, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.

NOTE: SAS initialization used:
      real time           0.32 seconds
      cpu time            0.06 seconds

1
*******************************************************************************;
2          * CDPS run for DOD1 project (this file is adaption of
driver.sas.xmpl) ;
3          * Amresh Hanchate ;
4          * Sept 22, 2004;
5
*******************************************************************************;
6
7          /*
8          NOTE:
9          1) this runs the "b" variant of the diagnoses (includes lab-only
records from HCSRN);
10         2) this program aims only to produce disease-group flags -- these
are not regressed on prospective costs to obtain risk
10       ! scores/weights (this step is done separately);
11
12         Inputs: dod2charb, dod2diagb
13         Output: cdpsbg
14
15         */
16
17
18
19         * File cdpsdir must be created before using this program;
20         *  then the filename for cdpsdir should be edited by the user;
21         *  to reflect the location of the files;
22         *  Include the entire path name along with the file name;
23
24         * REVISE THE NEXT LINE;
25         filename cdpsdir "/data/dod1/cdps/cdpsdir";
26         %include cdpsdir;
29
30         *debug;
31         *options mprint mtrace symbolgen source2;
32
33         options nonotes;
34
35         filename applywts "&cdpsdir.applywts.sas";
36         %include applywts;
START TIME: 11:22:14
NOTE: Libref CDPSLIB was successfully assigned as follows:
```

```
      Engine:        V8
      Physical Name: /data/dod1/cdps

2  The SAS System                                11:22 Friday, October 1, 2004

NOTE: Format $INDPS has been output.
NOTE: Format $INCDPS has been output.
NOTE: Format $INCONC has been output.
NOTE: Format $INPA has been output.
NOTE: Format AIDCAT has been output.
NOTE: Format AIDCOMB has been output.
NOTE: Format AA has been output.
NOTE: Format AC has been output.
NOTE: Format DA has been output.
NOTE: Format DC has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.26 seconds
      cpu time            0.04 seconds


beginning of macinit
end of macinit
1690
1691
*************************************************************************;
1692       * The user must set the following parameters in macros;
1693       * general, filelist and incvars;
1694
1695
*************************************************************************;
1696
1697       * REVISE THE NEXT LINE;
1698       %general(id=cdpsb, ndiags=6)
1699
1700
*************************************************************************;
1701
1702       * The following libnames are particular to the user environment;
1703       *   and must be edited before running the program;
1704       *   if these libnames are used in input parameters;
1705
1706       * REVISE THE NEXT LINES, using your own directory structure;
1707       libname da      "/data/dod1/saswork";
NOTE: Libref DA was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
1708       *libname tmpdir "/data/dod1/saswork";
1709
1710       * REVISE THE NEXT LINE;
1711       %filelist(inelig=da.dod2charb, indiag=da.dod2diagb, datadir=da)
NOTE: Libref CDPSLIB was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/cdps
1712
1713
*************************************************************************;
1714
1715       * REVISE THE NEXT LINE;
1716       %incvars()
1717
1718
*************************************************************************;
```

```
1719
1720
1721
*************************************************************************;
1722
1723       %execute
NOTE: Format SEX has been output.
NOTE: Format MALE has been output.
NOTE: Format AGECAT has been output.
NOTE: Format AGE has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.00 seconds
      cpu time            0.01 seconds


Before stmods and stextra
After stmods and stextra
Before mymods and myextra
After mymods and myextra


3   The SAS System                        11:22 Friday, October 1, 2004

NOTE: There were 1 observations read from the data set WORK.ALLVARS.
NOTE: The data set WORK.VARLIST has 78 observations and 3 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           0.20 seconds
      cpu time            0.01 seconds



NOTE: There were 78 observations read from the data set WORK.VARLIST.
NOTE: The data set WORK.ONAMES has 78 observations and 4 variables.
NOTE: DATA statement used:
      real time           0.21 seconds
      cpu time            0.00 seconds



NOTE: There were 78 observations read from the data set WORK.ONAMES.
NOTE: The data set WORK.ONAMES has 78 observations and 4 variables.
NOTE: PROCEDURE SORT used:
      real time           0.25 seconds
      cpu time            0.00 seconds



CDPS Parameter File for Job cdpsb
NOTE: 23 lines were written to file PRINT.
NOTE: The DATASTEP printed page 1.
NOTE: DATA statement used:
      real time           0.08 seconds
      cpu time            0.04 seconds



NOTE: There were 2304926 observations read from the data set DA.DOD2CHARB.
NOTE: The data set WORK.PFILE has 2304926 observations and 5 variables.
NOTE: DATA statement used:
      real time           1:03.67
      cpu time            16.60 seconds
```

```
NOTE: There were 2304926 observations read from the data set WORK.PFILE.
NOTE: The data set WORK.RECIPS has 2304926 observations and 26 variables.
NOTE: Compressing data set WORK.RECIPS decreased size by 72.64 percent.
      Compressed is 7690 pages; un-compressed would require 28110 pages.
NOTE: DATA statement used:
      real time           1:41.90
      cpu time            53.65 seconds

NOTE: There were 2304926 observations read from the data set WORK.RECIPS.
NOTE: The data set WORK.RECIPS has 2304926 observations and 26 variables.
NOTE: Compressing data set WORK.RECIPS decreased size by 72.64 percent.
      Compressed is 7690 pages; un-compressed would require 28110 pages.
NOTE: PROCEDURE SORT used:
      real time           5:52.10
      cpu time            1:39.36



NOTE: There were 2304926 observations read from the data set WORK.PFILE.
NOTE: The data set WORK.EFILE has 2304926 observations and 2 variables.

4     The SAS System                      11:22 Friday, October 1, 2004

NOTE: DATA statement used:
      real time           54.71 seconds
      cpu time            11.81 seconds



NOTE: There were 2304926 observations read from the data set WORK.EFILE.
NOTE: The data set WORK.EXP has 2304926 observations and 3 variables.
NOTE: DATA statement used:
      real time           40.37 seconds
      cpu time            10.89 seconds

NOTE: There were 2304926 observations read from the data set WORK.EXP.
NOTE: The data set WORK.EXP has 2304926 observations and 3 variables.
NOTE: PROCEDURE SORT used:
      real time           1:55.08
      cpu time            32.30 seconds


                              -----Directory-----

          Libref:            WORK
          Engine:            V8
          Physical Name:     /data/saswork/SAS_work7C1B000014D0_genmed2
          File Name:         /data/saswork/SAS_work7C1B000014D0_genmed2
          Inode Number:      320960
          Access Permission: rwxrwx---
          Owner Name:        amresh
          File Size (bytes): 512


              #  Name     Memtype   File Size  Last Modified
              ---------------------------------------------------
              1  ALLVARS  DATA          32768   01OCT2004:11:22:15
              2  EFILE    DATA       74350592   01OCT2004:11:31:49
              3  EXP      DATA       93028352   01OCT2004:11:34:24
              4  ONAMES   DATA          16384   01OCT2004:11:22:16
              5  PFILE    DATA      130236416   01OCT2004:11:23:20
              6  RECIPS   DATA      126001152   01OCT2004:11:30:54
```

```
                 7  VARLIST  DATA          16384   01OCT2004:11:22:16

NOTE: PROCEDURE DATASETS used:
      real time           0.11 seconds
      cpu time            0.01 seconds

NOTE: There were 2304926 observations read from the data set WORK.RECIPS.
NOTE: There were 2304926 observations read from the data set WORK.EXP.
NOTE: The data set WORK.ABDAA has 1529597 observations and 29 variables.
NOTE: Compressing data set WORK.ABDAA decreased size by 70.19 percent.
      Compressed is 5700 pages; un-compressed would require 19121 pages.
NOTE: The data set WORK.ABDAC has 775329 observations and 29 variables.
NOTE: Compressing data set WORK.ABDAC decreased size by 68.76 percent.
      Compressed is 3028 pages; un-compressed would require 9692 pages.
NOTE: The data set WORK.ABDDA has 0 observations and 29 variables.
NOTE: The data set WORK.ABDDC has 0 observations and 29 variables.
NOTE: The data set WORK.SUMS has 1 observations and 4 variables.
NOTE: DATA statement used:
      real time           2:30.41
      cpu time            1:37.58

5   The SAS System                      11:22 Friday, October 1, 2004



NOTE: Numeric values have been converted to character values at the places
given by:
      (Line):(Column).
      2191:8    2191:35   2191:62   2191:89

NAA=1,529,597 NAC=775,329 NDA=0 NDC=0
NOTE: There were 1 observations read from the data set WORK.SUMS.
NOTE: DATA statement used:
      real time           0.05 seconds
      cpu time            0.00 seconds



NOTE: There were 5343739 observations read from the data set DA.DOD2DIAGB.
NOTE: The data set WORK.DFILE has 5343739 observations and 7 variables.
NOTE: DATA statement used:
      real time           3:38.48
      cpu time            41.77 seconds


NOTE: There were 5343739 observations read from the data set WORK.DFILE.
NOTE: The data set WORK.OKDIAGS has 5343739 observations and 7 variables.
NOTE: DATA statement used:
      real time           2:33.06
      cpu time            36.39 seconds



NOTE: There were 5343739 observations read from the data set WORK.OKDIAGS.
NOTE: The data set WORK.OKDIAGS has 5343739 observations and 7 variables.
NOTE: PROCEDURE SORT used:
      real time           6:18.23
      cpu time            1:33.41



NOTE: There were 5343739 observations read from the data set WORK.OKDIAGS.
```

```
NOTE: The data set WORK.STEP00 has 13799798 observations and 3 variables.
NOTE: DATA statement used:
      real time           2:54.80
      cpu time            1:28.83


NOTE: There were 13799798 observations read from the data set WORK.STEP00.
NOTE: The data set WORK.STEP0 has 13799798 observations and 3 variables.
NOTE: DATA statement used:
      real time           2:45.77
      cpu time            58.07 seconds


NOTE: There were 1529597 observations read from the data set WORK.ABDAA.
NOTE: The data set WORK.SHORT has 1529597 observations and 1 variables.
NOTE: DATA statement used:
      real time           19.71 seconds
 cpu time            18.62 seconds


6    The SAS System                      11:22 Friday, October 1, 2004

NOTE: There were 1529597 observations read from the data set WORK.SHORT.
NOTE: There were 13799798 observations read from the data set WORK.STEP0.
NOTE: The data set WORK.STEP1 has 10738802 observations and 3 variables.
NOTE: DATA statement used:
      real time           2:27.87
      cpu time            1:44.26


                              -----Directory-----

              Libref:            WORK
              Engine:            V8
              Physical Name:     /data/saswork/SAS_work7C1B000014D0_genmed2
              File Name:         /data/saswork/SAS_work7C1B000014D0_genmed2
              Inode Number:      320960
              Access Permission: rwxrwx---
              Owner Name:        amresh
              File Size (bytes): 1024


                   #  Name     Memtype   File Size  Last Modified
                  ---------------------------------------------------
                   1  ABDAA    DATA       93396992   01OCT2004:11:36:54
                   2  ABDAC    DATA       49618944   01OCT2004:11:36:54
                   3  ABDDA    DATA          40960   01OCT2004:11:36:55
                   4  ABDDC    DATA          40960   01OCT2004:11:36:55
                   5  ALLVARS  DATA          32768   01OCT2004:11:22:15
                   6  DFILE    DATA      355917824   01OCT2004:11:40:34
                   7  EFILE    DATA       74350592   01OCT2004:11:31:49
                   8  EXP      DATA       93028352   01OCT2004:11:34:24
                   9  OKDIAGS  DATA      355917824   01OCT2004:11:49:25
                  10  ONAMES   DATA          16384   01OCT2004:11:22:16
                  11  PFILE    DATA      130236416   01OCT2004:11:23:20
                  12  RECIPS   DATA      126001152   01OCT2004:11:30:54
                  13  SHORT    DATA       27860992   01OCT2004:11:55:25
                  14  STEP0    DATA      389832704   01OCT2004:11:55:06
                  15  STEP00   DATA      389832704   01OCT2004:11:52:20
                  16  STEP1    DATA      303366144   01OCT2004:11:57:53
                  17  STEP2    DATA      516759552   01OCT2004:12:10:38
                  18  SUMS     DATA          16384   01OCT2004:11:36:55
```

```
                  19  VARLIST  DATA          16384   01OCT2004:11:22:16
NOTE: Deleting WORK.SHORT (memtype=DATA).
NOTE: Deleting WORK.STEP00 (memtype=DATA).
NOTE: Deleting WORK.STEP0 (memtype=DATA).
NOTE: Deleting WORK.STEP1 (memtype=DATA).

NOTE: PROCEDURE DATASETS used:
      real time           0.90 seconds
      cpu time            0.70 seconds



NOTE: There were 1529597 observations read from the data set WORK.ABDAA.
NOTE: There were 1529597 observations read from the data set WORK.STEP2.
NOTE: The data set WORK.STEP2AA has 1529597 observations and 98 variables.
NOTE: Compressing data set WORK.STEP2AA decreased size by 85.33 percent.
      Compressed is 4314 pages; un-compressed would require 29416 pages.
NOTE: DATA statement used:
      real time           6:27.23
      cpu time            6:16.47



7   The SAS System                        11:22 Friday, October 1, 2004

NOTE: There were 1529597 observations read from the data set WORK.STEP2AA.
NOTE: The data set WORK.SEE has 1 observations and 88 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           1:20.41
      cpu time            1:19.90



NOTE: There were 1 observations read from the data set WORK.SEE.
NOTE: The PROCEDURE PRINT printed page 2.
NOTE: PROCEDURE PRINT used:
      real time           0.07 seconds
      cpu time            0.07 seconds



NOTE: There were 5343739 observations read from the data set DA.DOD2DIAGB.
NOTE: The data set WORK.DFILE has 5343739 observations and 7 variables.
NOTE: DATA statement used:
      real time           3:50.54
      cpu time            42.18 seconds



NOTE: There were 5343739 observations read from the data set WORK.DFILE.
NOTE: The data set WORK.OKDIAGS has 5343739 observations and 7 variables.
NOTE: DATA statement used:
      real time           2:31.91
      cpu time            35.86 seconds

NOTE: There were 5343739 observations read from the data set WORK.OKDIAGS.
NOTE: The data set WORK.OKDIAGS has 5343739 observations and 7 variables.
NOTE: PROCEDURE SORT used:
      real time           6:27.68
      cpu time            1:34.26

NOTE: There were 5343739 observations read from the data set WORK.OKDIAGS.
NOTE: The data set WORK.STEP00 has 13799798 observations and 3 variables.
```

```
NOTE: DATA statement used:
      real time           2:58.99
      cpu time            1:23.57

NOTE: There were 13799798 observations read from the data set WORK.STEP00.
NOTE: The data set WORK.STEP0 has 13799798 observations and 3 variables.
NOTE: DATA statement used:
      real time           2:39.19
      cpu time            57.72 seconds



NOTE: There were 775329 observations read from the data set WORK.ABDAC.
NOTE: The data set WORK.SHORT has 775329 observations and 1 variables.
NOTE: DATA statement used:
      real time           10.33 seconds
      cpu time            9.85 seconds


8    The SAS System                      11:22 Friday, October 1, 2004

NOTE: There were 775329 observations read from the data set WORK.SHORT.
NOTE: There were 13799798 observations read from the data set WORK.STEP0.
NOTE: The data set WORK.STEP1 has 3389648 observations and 3 variables.
NOTE: DATA statement used:
      real time           1:25.94
      cpu time            1:22.97


                              -----Directory-----

             Libref:            WORK
             Engine:            V8
             Physical Name:     /data/saswork/SAS_work7C1B000014D0_genmed2
             File Name:         /data/saswork/SAS_work7C1B000014D0_genmed2
             Inode Number:      320960
             Access Permission: rwxrwx---
             Owner Name:        amresh
             File Size (bytes): 1024


              #  Name     Memtype   File Size  Last Modified
             ---------------------------------------------------
              1  ABDAA    DATA       93396992   01OCT2004:11:36:54
              2  ABDAC    DATA       49618944   01OCT2004:11:36:54
              3  ABDDA    DATA          40960   01OCT2004:11:36:55
              4  ABDDC    DATA          40960   01OCT2004:11:36:55
              5  ALLVARS  DATA          32768   01OCT2004:11:22:15
              6  DFILE    DATA      355917824   01OCT2004:12:22:17
              7  EFILE    DATA       74350592   01OCT2004:11:31:49
              8  EXP      DATA       93028352   01OCT2004:11:34:24
              9  OKDIAGS  DATA      355917824   01OCT2004:12:31:17
             10  ONAMES   DATA          16384   01OCT2004:11:22:16
             11  PFILE    DATA      130236416   01OCT2004:11:23:20
             12  RECIPS   DATA      126001152   01OCT2004:11:30:54
             13  SEE      DATA          65536   01OCT2004:12:18:27
             14  SHORT    DATA       14131200   01OCT2004:12:37:06
             15  STEP0    DATA      389832704   01OCT2004:12:36:56
             16  STEP00   DATA      389832704   01OCT2004:12:34:16
             17  STEP1    DATA       95764480   01OCT2004:12:38:32
             18  STEP2    DATA      256647168   01OCT2004:12:43:08
             19  STEP2AA  DATA      106029056   01OCT2004:12:17:06
             20  SUMS     DATA          16384   01OCT2004:11:36:55
```

```
                 21  VARLIST  DATA          16384   01OCT2004:11:22:16
NOTE: Deleting WORK.SHORT (memtype=DATA).
NOTE: Deleting WORK.STEP00 (memtype=DATA).
NOTE: Deleting WORK.STEP0 (memtype=DATA).
NOTE: Deleting WORK.STEP1 (memtype=DATA).

NOTE: PROCEDURE DATASETS used:
      real time           0.82 seconds
      cpu time            0.65 seconds

NOTE: There were 775329 observations read from the data set WORK.ABDAC.
NOTE: There were 775329 observations read from the data set WORK.STEP2.
NOTE: The data set WORK.STEP2AC has 775329 observations and 98 variables.
NOTE: Compressing data set WORK.STEP2AC decreased size by 85.66 percent.
      Compressed is 2224 pages; un-compressed would require 15508 pages.
NOTE: DATA statement used:
      real time           3:36.72
      cpu time            3:10.44

9   The SAS System                      11:22 Friday, October 1, 2004



NOTE: There were 775329 observations read from the data set WORK.STEP2AC.
NOTE: The data set WORK.SEE has 1 observations and 88 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           38.37 seconds
      cpu time            38.05 seconds



NOTE: There were 1 observations read from the data set WORK.SEE.
NOTE: The PROCEDURE PRINT printed page 3.
NOTE: PROCEDURE PRINT used:
      real time           0.07 seconds
      cpu time            0.07 seconds


                                -----Directory-----

              Libref:            WORK
              Engine:            V8
              Physical Name:     /data/saswork/SAS_work7C1B000014D0_genmed2
              File Name:         /data/saswork/SAS_work7C1B000014D0_genmed2
              Inode Number:      320960
              Access Permission: rwxrwx---
              Owner Name:        amresh
              File Size (bytes): 1024


                   #  Name     Memtype   File Size  Last Modified
                  ---------------------------------------------------
                   1  ABDAA    DATA       93396992   01OCT2004:11:36:54
                   2  ABDAC    DATA       49618944   01OCT2004:11:36:54
                   3  ABDDA    DATA          40960   01OCT2004:11:36:55
                   4  ABDDC    DATA          40960   01OCT2004:11:36:55
                   5  ALLVARS  DATA          32768   01OCT2004:11:22:15
                   6  DFILE    DATA      355917824   01OCT2004:12:22:17
                   7  EFILE    DATA       74350592   01OCT2004:11:31:49
                   8  EXP      DATA       93028352   01OCT2004:11:34:24
                   9  OKDIAGS  DATA      355917824   01OCT2004:12:31:17
                  10  ONAMES   DATA          16384   01OCT2004:11:22:16
                  11  PFILE    DATA      130236416   01OCT2004:11:23:20
```

```
                  12  RECIPS   DATA      126001152   01OCT2004:11:30:54
                  13  SEE      DATA          65536   01OCT2004:12:47:24
                  14  STEP2    DATA      256647168   01OCT2004:12:43:08
                  15  STEP2AA  DATA      106029056   01OCT2004:12:17:06
                  16  STEP2AC  DATA       54665216   01OCT2004:12:46:46
                  17  SUMS     DATA          16384   01OCT2004:11:36:55
                  18  VARLIST  DATA          16384   01OCT2004:11:22:16

NOTE: Deleting WORK.STEP2 (memtype=DATA).
NOTE: PROCEDURE DATASETS used:
      real time           0.40 seconds
      cpu time            0.32 seconds

NOTE: There were 78 observations read from the data set CDPSLIB.V20PROS.
NOTE: The data set WORK.DPSWTS has 78 observations and 7 variables.
NOTE: DATA statement used:
      real time           0.26 seconds

0    The SAS System                     11:22 Friday, October 1, 2004

      cpu time            0.02 seconds



NOTE: There were 78 observations read from the data set WORK.DPSWTS.
NOTE: The data set WORK.DPSWTS has 78 observations and 7 variables.
NOTE: PROCEDURE SORT used:
      real time           0.29 seconds
      cpu time            0.01 seconds



NOTE: There were 78 observations read from the data set WORK.DPSWTS.
NOTE: There were 78 observations read from the data set WORK.ONAMES.
NOTE: The data set WORK.CHOOSEOK has 78 observations and 9 variables.
NOTE: DATA statement used:
      real time           0.21 seconds
      cpu time            0.03 seconds

NOTE: There were 78 observations read from the data set WORK.CHOOSEOK.
NOTE: The data set WORK.CHOOSEOK has 78 observations and 9 variables.
NOTE: PROCEDURE SORT used:
      real time           0.27 seconds
      cpu time            0.00 seconds

NOTE: Character values have been converted to numeric values at the places
given by:
      (Line):(Column).
      34623:205
NOTE: There were 78 observations read from the data set WORK.CHOOSEOK.
NOTE: The data set WORK.GOODONES has 78 observations and 11 variables.
NOTE: DATA statement used:
      real time           0.27 seconds
      cpu time            0.02 seconds



NOTE: There were 78 observations read from the data set WORK.GOODONES.
NOTE: The PROCEDURE PRINT printed pages 4-5.
NOTE: PROCEDURE PRINT used:
      real time           0.01 seconds
      cpu time            0.02 seconds
```

```
NOTE: There were 78 observations read from the data set WORK.GOODONES.
NOTE: The data set WORK.LOOKUP has 78 observations and 9 variables.
NOTE: DATA statement used:
      real time           0.23 seconds
      cpu time            0.02 seconds


NOTE: There were 78 observations read from the data set WORK.LOOKUP.
NOTE: The data set WORK.TEMP has 78 observations and 2 variables.
NOTE: DATA statement used:
      real time           0.21 seconds
      cpu time            0.01 seconds


11   The SAS System                       11:22 Friday, October 1, 2004

NOTE: There were 78 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.SCORE has 1 observations and 80 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           0.31 seconds
      cpu time            0.02 seconds


d=78
NOTE: There were 1 observations read from the data set WORK.SCORE.
NOTE: The data set WORK.SCORET has 1 observations and 121 variables.
NOTE: DATA statement used:
      real time           0.27 seconds
      cpu time            0.04 seconds


NOTE: There were 1529597 observations read from the data set WORK.STEP2AA.
NOTE: The data set WORK.PERSONAL has 1529597 observations and 219 variables.
NOTE: DATA statement used:
      real time           15:54.94
      cpu time            3:16.36


NOTE: There were 1 observations read from the data set WORK.SCORET.
NOTE: There were 1529597 observations read from the data set WORK.PERSONAL.
NOTE: The data set WORK.AASCORE has 1529597 observations and 99 variables.
NOTE: DATA statement used:
      real time           11:00.61
      cpu time            2:21.53


NOTE: There were 1529597 observations read from the data set WORK.AASCORE.
NOTE: The data set WORK.MNPREDAA has 1 observations and 3 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           11.37 seconds
      cpu time            11.19 seconds


NOTE: There were 1529597 observations read from the data set WORK.AASCORE.
NOTE: There were 1529597 observations read from the data set WORK.ABDAA.
NOTE: The data set WORK.RESULTAA has 1529597 observations and 100 variables.
```

```
NOTE: Compressing data set WORK.RESULTAA decreased size by 83.65 percent.
      Compressed is 4903 pages; un-compressed would require 29993 pages.
NOTE: DATA statement used:
      real time           2:59.30
      cpu time            2:45.44


NOTE: There were 1 observations read from the data set WORK.MNPREDAA.
NOTE: There were 1529597 observations read from the data set WORK.RESULTAA.
NOTE: The data set WORK.RESULTAA has 1529597 observations and 102 variables.
NOTE: Compressing data set WORK.RESULTAA decreased size by 82.33 percent.
      Compressed is 5407 pages; un-compressed would require 30593 pages.
NOTE: DATA statement used:
      real time           1:36.27
      cpu time            1:19.96

12   The SAS System                    11:22 Friday, October 1, 2004


NOTE: There were 78 observations read from the data set WORK.LOOKUP.
NOTE: The data set WORK.TEMP has 78 observations and 2 variables.
NOTE: DATA statement used:
      real time           0.27 seconds
      cpu time            0.01 seconds


NOTE: There were 78 observations read from the data set WORK.TEMP.
NOTE: The data set WORK.SCORE has 1 observations and 80 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           0.30 seconds
      cpu time            0.01 seconds



d=78
NOTE: There were 1 observations read from the data set WORK.SCORE.
NOTE: The data set WORK.SCORET has 1 observations and 121 variables.
NOTE: DATA statement used:
      real time           0.38 seconds
      cpu time            0.03 seconds


NOTE: There were 775329 observations read from the data set WORK.STEP2AC.
NOTE: The data set WORK.PERSONAL has 775329 observations and 219 variables.
NOTE: DATA statement used:
      real time           8:38.93
      cpu time            1:39.54


NOTE: There were 1 observations read from the data set WORK.SCORET.
NOTE: There were 775329 observations read from the data set WORK.PERSONAL.
NOTE: The data set WORK.ACSCORE has 775329 observations and 99 variables.
NOTE: DATA statement used:
      real time           4:36.82
      cpu time            1:10.43


NOTE: There were 775329 observations read from the data set WORK.ACSCORE.
NOTE: The data set WORK.MNPREDAC has 1 observations and 3 variables.
NOTE: PROCEDURE SUMMARY used:
```

```
      real time           6.08 seconds
      cpu time            5.89 seconds

NOTE: There were 775329 observations read from the data set WORK.ACSCORE.
NOTE: There were 775329 observations read from the data set WORK.ABDAC.
NOTE: The data set WORK.RESULTAC has 775329 observations and 100 variables.
NOTE: Compressing data set WORK.RESULTAC decreased size by 83.88 percent.
      Compressed is 2551 pages; un-compressed would require 15824 pages.
NOTE: DATA statement used:
      real time           1:28.74
      cpu time            1:23.84



NOTE: There were 1 observations read from the data set WORK.MNPREDAC.
NOTE: There were 775329 observations read from the data set WORK.RESULTAC.


13   The SAS System                       11:22 Friday, October 1, 2004

NOTE: The data set WORK.RESULTAC has 775329 observations and 102 variables.
NOTE: Compressing data set WORK.RESULTAC decreased size by 82.63 percent.
      Compressed is 2806 pages; un-compressed would require 16154 pages.
NOTE: DATA statement used:
      real time           46.02 seconds
      cpu time            41.13 seconds



NOTE: There were 1529597 observations read from the data set WORK.RESULTAA.
NOTE: There were 775329 observations read from the data set WORK.RESULTAC.
NOTE: The data set DA.CDPSBG has 2304926 observations and 102 variables.
NOTE: The data set WORK.RESULT has 2304926 observations and 10 variables.
NOTE: DATA statement used:
      real time           10:05.15
      cpu time            3:15.42



NOTE: There were 2304926 observations read from the data set WORK.RESULT.
NOTE: The data set WORK.UNWGTED has 3 observations and 4 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           12.01 seconds
      cpu time            10.15 seconds



NOTE: There were 2304926 observations read from the data set WORK.RESULT.
NOTE: The data set WORK.WEIGHTED has 3 observations and 5 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           13.70 seconds
      cpu time            10.52 seconds



NOTE: There were 3 observations read from the data set WORK.UNWGTED.
NOTE: There were 3 observations read from the data set WORK.WEIGHTED.
NOTE: The data set DA.CDPSBS has 3 observations and 6 variables.
NOTE: DATA statement used:
      real time           0.25 seconds
      cpu time            0.02 seconds
```

```
NOTE: There were 3 observations read from the data set DA.CDPSBS.
NOTE: The PROCEDURE PRINT printed page 6.
NOTE: PROCEDURE PRINT used:
      real time           0.05 seconds
      cpu time            0.00 seconds


END TIME: 13:45:21
RUN TIME:  2:23:07
34789
34790      * Your own documentation of run time could be recorded here, for
reference;
34791      * Note: run time about 2 minutes;
34792
*************************************************************************;
34793
```

Attachment A4_CDPS_2
(task5b_cdps.lst)

```
                      Parameters for Job cdpsb   [CDPS Release 2.0]
1
                                             '     11:22 Friday, October 1, 2004

JOB ID (maximum of 7 characters)                        = id           = cdpsb

DATA SET OPTIONS

    Weighted Analysis? (No/Weighted)                    = wt           = No

    Minimum Number of Eligible Months in Base Year      = nbase        = 1

    Create File of CDPS Group Indicators? (n/y)         = calcg        = y
         Number of ICD-9 Diagnoses in Claims File       = ndiags       = 6

    Type of CDPS Groups to Use? (cdps/dps/conc/pa)      = type         = cdps

Prospective


LIST OF FILES (including libnames)

    The Default CDPS Directory is cdpslib   = /data/dod1/cdps/

    Input File of Eligibility (base year)        = inelig  = da.dod2charb

    Input File of Claims                         = indiag  = da.dod2diagb

    Output File of CDPS Group Indicators                   = da.cdpsbg

    Input File of CDPS Weights                   = inwgt   = cdpslib.v20pros

    Output File Summary Table                              = da.cdpsbs


INCLUDED VARIABLES

    Include class variables for reports?                = incrpt       = n
         (n/y  If Yes, the list follows)

    Include class variables?                            = inccls       = n
         (n/y  If Yes, the list follows)

    Include year variables?                             = incyr        = n
         (n/y  If Yes, the list follows)

    Include special formats? (n/y  If Yes, the list follows)
```

Program cdpsb [CDPS Release 2.0] Apply CDPS Weights from cdpslib.v20pros 2
using Files Elig=da.dod2charb, Diag=da.dod2diagb and Exp=None

' 11:22 Friday, October 1, 2004

Unweighted Counts for AFDC Adults

Obs	_FREQ_	CHILD	A_UNDER1	A_1_4	A_5_14F	A_5_14M	A_15_24F	A_15_24M	A_25_44F
1	1529597	0	0	0	0	0	111,971	153,311	399,151

Obs	A_25_44M	A_45_64F	A_45_64M	A_65_74F	A_65_74M	A_75_84F	A_75_84M	A_O85F	A_O85M
1	451,247	212,796	201,121	0	0	0	0	0	0

Obs	A_OVER64	NOCDPS	CARVH	CARM	CARL	CAREL	PSYH	PSYM	PSYL
1	0	878,206	984	8,779	49,867	140,465	2,684	5,418	90,874

Obs	SKCM	SKCL	SKCVL	SKCEL	CNSH	CNSM	CNSL	PULVH	PULH
1	609	20,523	80,656	62,834	303	5,385	76,532	0	2,455

Obs	PULM	PULL	GIH	GIM	GIL	DIA1H	DIA1M	DIA2M	DIA2L
1	3,908	71,696	1,050	10,930	107,282	527	10,469	4,634	38,903

Obs	SKNH	SKNL	SKNVL	RENVH	RENM	RENL	SUBL	SUBVL	CANH
1	551	1,940	43,049	2,313	14,473	31,216	2,249	10,469	4,415

Obs	CANM	CANL	DDM	DDL	GENEL	METH	METM	METVL	PRGCMP
1	18,874	7,239	0	738	53,776	4,975	2,992	10,721	52,445

Obs	PRGINC	EYEL	EYEVL	CERL	AIDSH	INFH	HIVM	INFM	INFL
1	26,484	5,470	22,910	6,306	1,071	269	117	1,593	8,844

Obs	HEMEH	HEMVH	HEMM	HEML
1	60	101	3,455	4,681

Program cdpsb [CDPS Release 2.0] Apply CDPS Weights from cdpslib.v20pros 3
using Files Elig=da.dod2charb, Diag=da.dod2diagb and Exp=None

' 11:22 Friday, October 1, 2004
Unweighted Counts for AFDC Children

Obs	_FREQ_	CHILD	A_UNDER1	A_1_4	A_5_14F	A_5_14M	A_15_24F	A_15_24M	A_25_44F
1	775,329	775,329	38,324	147,822	231,745	241,798	56,871	58,769	0

Obs	A_25_44M	A_45_64F	A_45_64M	A_65_74F	A_65_74M	A_75_84F	A_75_84M	A_O85F	A_O85M
1	0	0	0	0	0	0	0	0	0

Obs	A_OVER64	NOCDPS	CARVH	CARM	CARL	CAREL	PSYH	PSYM	PSYL
1	0	563,204	200	69	6,421	1,528	352	2,155	49,460

Obs	SKCM	SKCL	SKCVL	SKCEL	CNSH	CNSM	CNSL	PULVH	PULH
1	108	5,125	18,154	3,721	138	1,018	28,786	0	1,843

Obs	PULM	PULL	GIH	GIM	GIL	DIA1H	DIA1M	DIA2M	DIA2L
1	2,374	72,443	272	1,345	20,257	0	0	0	2,075

Obs	SKNH	SKNL	SKNVL	RENVH	RENM	RENL	SUBL	SUBVL	CANH
1	76	166	18,738	161	331	10,735	416	541	544

Obs	CANM	CANL	DDM	DDL	GENEL	METH	METM	METVL	PRGCMP
1	853	278	233	1,727	3,391	1,235	699	7,130	769

Obs	PRGINC	EYEL	EYEVL	CERL	AIDSH	INFH	HIVM	INFM	INFL
1	1,090	0	2,434	715	64	75	6	894	5,849

Obs	HEMEH	HEMVH	HEMM	HEML
1	99	689	806	1,381

Program cdpsb [CDPS Release 2.0] Apply CDPS Weights from cdpslib.v20pros 4
using Files Elig=da.dod2charb, Diag=da.dod2diagb and Exp=None

' 11:22 Friday, October 1, 2004
These weights (CDPSLIB.V20PROS) are applied to the data set

CDPS Group Name	Description	AFDC Adult	AFDC Child
INTERCEPT	Intercept	0.3615	0.6174
A_UNDER1	Child under Age 1	.	0.1716
A_1_4	Child Age 1-4	.	-0.0087
A_5_14F	Female Age 5-14	.	-0.0638
A_5_14M	Male Age 5-14	.	.
A_15_24F	Female Age 15-24	0.2847	0.8484
A_15_24M	Male Age 15-24	-0.1209	0.1963
A_25_44F	Female Age 25-44	0.1625	.
A_25_44M	Male Age 25-44	.	.
A_45_64F	Female Age 45-64	0.3155	.
A_45_64M	Male Age 45-64	0.2490	.
CARVH	Cardiovascular, very high	4.0785	17.4053
CARM	Cardiovascular, medium	1.3485	5.0832
CARL	Cardiovascular, low	0.5741	1.3818
CAREL	Cardiovascular, extra low	0.3956	0.9626
PSYH	Psychiatric, high	1.3172	9.2453
PSYM	Psychiatric, medium	1.3172	5.1377
PSYL	Psychiatric, low	0.5734	2.3029
SKCM	Skeletal, medium	2.1823	2.2790
SKCL	Skeletal, low	0.6282	0.9145
SKCVL	Skeletal, very low	0.4511	0.5373
SKCEL	Skeletal, extra low	0.4511	0.4984
CNSH	CNS, high	1.3238	14.5151
CNSM	CNS, medium	1.0329	5.5078
CNSL	CNS, low	0.5436	1.0063
PULVH	Pulmonary, very high	0.0000	0.0000
PULH	Pulmonary, high	1.1450	4.0983
PULM	Pulmonary, medium	1.3632	2.2512
PULL	Pulmonary, low	0.5272	0.7812
GIH	Gastro, high	1.3325	5.4414
GIM	Gastro, medium	1.3325	2.9228
GIL	Gastro, low	0.4670	0.4669
DIA1H	Diabetes, type 1 high	5.2227	0.0000
DIA1M	Diabetes, type 1 medium	1.5581	0.0000
DIA2M	Diabetes, type 2 medium	1.3899	0.0000
DIA2L	Diabetes, type 2 low	0.6249	2.1068
SKNH	Skin, high	1.3308	2.3644
SKNL	Skin, low	0.6896	1.1801
SKNVL	Skin, very low	0.2148	0.2652
RENVH	Renal, very high	5.4994	7.2309
RENM	Renal, medium	0.7792	0.9901
RENL	Renal, low	0.3752	0.7517
SUBL	Substance abuse, low	0.8323	3.0343
SUBVL	Substance abuse, very low	0.4316	1.4483
CANH	Cancer, high	2.3396	8.8078
CANM	Cancer, medium	1.0990	2.0915
CANL	Cancer, low	0.2269	1.5463
DDM	DD, medium	0.0000	7.8837
DDL	DD, low	0.3849	3.3277
GENEL	Genital, extra low	0.3001	0.8955
METH	Metabolic, high	1.1734	5.7936
METM	Metabolic, medium	1.1734	1.7558

METVL	Metabolic, very low	0.4712	0.9070
PRGCMP	Pregnancy, complete	0.2430	1.3196
PRGINC	Pregnancy, incomplete	1.1163	3.8357
EYEL	Eye, low	0.6229	0.0000

Program cdpsb [CDPS Release 2.0] Apply CDPS Weights from cdpslib.v20pros 5
using Files Elig=da.dod2charb, Diag=da.dod2diagb and Exp=None

' 11:22 Friday, October 1, 2004

These weights (CDPSLIB.V20PROS) are applied to the data set

CDPS Group Name	Description	AFDC Adult	AFDC Child
EYEVL	Eye, very low	0.3463	1.0139
CERL	Cerebrovascular, low	0.8322	2.2075
AIDSH	AIDS, high	2.3936	1.9780
INFH	Infectious, high	2.3936	1.9780
HIVM	HIV, medium	0.6965	1.3432
INFM	Infectious, medium	0.6965	1.3432
INFL	Infectious, low	0.1530	0.2213
HEMEH	Hematological, extra high	5.0977	20.2755
HEMVH	Hematological, very high	3.7447	5.4109
HEMM	Hematological, medium	0.7517	1.4402
HEML	Hematological, low	0.7517	1.0278
CCARVH	Childrens CARVH	.	.
CCNSM	Childrens CNSM	.	.
CPULVH	Childrens PULVH	.	.
CGIH	Childrens GIH	.	.
CGIM	Childrens GIM	.	.
CGIL	Childrens GIL	.	.
CDIA2L	Childrens DIA2L	.	.
CMETH	Childrens METH	.	.
CMETM	Childrens METM	.	.
CINFM	Childrens INFM	.	.
CHEMVH	Childrens HEMVH	.	.

Program cdpsb [CDPS Release 2.0] Apply CDPS Weights from cdpslib.v20pros 6
using Files Elig=da.dod2charb, Diag=da.dod2diagb and Exp=None

' 11:22 Friday, October 1, 2004
Means by AIDCAT
from File da.cdpsbs

Standard Aid Category	Frequency	Mean No CDPS Diag	Mean Months Eligible	Mean CDPS Case Mix
.	2304926	0.625	12.000	0.985
AFDC/TANF Adult	1529597	0.574	12.000	0.907
AFDC/TANF Child	775,329	0.726	12.000	1.139

Attachment A4_ACG_1
(task5c_acg.txt)

```
*************************************************************************;
* ACG run for DOD1 project ;
* Amresh Hanchate ;
* October, 2004;
*************************************************************************;

* NOTE:
*1) this runs the "b" variant of the diagnoses (includes lab-only records from
HCSRN);
*2) this program aims only to produce disease-group flags -- these are not
regressed on prospective costs to obtain risk scores/weights (this step is done
separately);
*3) the input files dod2acg1.txt and dod2acg2.txt are identical to dod2charb
and dod2diagb written to conform to ACG requirements (using
task4c_acgdata.sas);

*NOTES ON ACG-PM MODEL;
*Chad Abrams suggested that we run this variant of the model;
*Details of this are on pages 15-16 and 29-31 of the documentation file
(Version 6.0 Release Notes");
* To produce the indicator variables that will enable running ACG-PM the
following model asks for additional output (of note are the "edc..." line and
the "outrec..." line);

*Inputs: dod2acgb1.txt, dod2acgb2.txt;
*Output: dod2acgbout.txt, dod2acgbprn.txt, dod2acgbedc.txt;

*input and output files;
input1=dod2acgb1.txt
input2=dod2acgb2.txt
output=dod2acgbout.txt
print=dod2acgbprn.txt
*jhu=dod2acgbjhu.txt
edc=dod2acgbedc.txt

*branching;
delivered

*input file layout;
id=1,18
age=19,2
sex=21,1
female="0"
*tcost=22,8
icd=22,8,5

*output file layout;
outrec=id,acg,rub,hos,del,pre
```

Attachment A4_ACG_2
(dod2acgbprn.txt)

```
*******************************************************************************
*             Adjusted Clinical Groups (ACG) Assignment Software              *
*             (Including the Expanded Diagnosis Clusters - EDCs)              *
*                              PC-DOS Version 6.02                            *
*                                                                             *
*                          Thu Oct 28 11:39:39 2004                           *
*                                                                             *
*Copyright (c) The Johns Hopkins University 1990-2003.  All rights reserved.  *
*                                                                             *
*The Johns Hopkins University disclaims all warranties, implied or otherwise. *
*******************************************************************************


CONTROL CARDS PROCESSED:
  1:
*******************************************************************************
  2: * ACG run for DOD1 project
  3: * Amresh Hanchate
  4: * Oct 28, 2004
  5:
*******************************************************************************
  6:
  7: * NOTE:
  8: *1) this runs the "b" variant of the diagnoses (includes lab-only records
from HCSRN)
  9: *2) this program aims only to produce disease-group flags -- these are not
regressed on prospective costs to obtain risk scores/weights (this step is done
separately)
 10: *3) the input files dod2acg1.txt and dod2acg2.txt are identical to
dod2charb and dod2diagb written to conform to ACG requirements (using
task4c_acgdata.sas)
 11:
 12: *Inputs: dod2acgb1.txt, dod2acgb2.txt
 13: *Output: dod2acgbout.txt, dod2acgbprn.txt, dod2acgbedc.txt
 14:
 15: *input and output files
 16: input1=dod2acgb1.txt
 17: input2=dod2acgb2.txt
 18: output=dod2acgbout.txt
 19: print=dod2acgbprn.txt
 20: *jhu=dod2acgbjhu.txt
 21: edc=dod2acgbedc.txt
 22:
 23: *branching
 24: delivered
 25:
 26: *input file layout
 27: id=1,18
 28: age=19,2
 29: sex=21,1
 30: female="0"
 31: *tcost=22,8
 32: icd=22,8,5
 33:
 34: *output file layout
 35: outrec=id,acg,rub,hos,del,pre
 36:
 37:
 38: ÿ
```

```
FILE NAMES SPECIFIED:
Input 1    :dod2acgb1.txt
Input 2    :dod2acgb2.txt
Output     :dod2acgbout.txt
Control    :task5c_acg.txt
Nonmatch   :
EDC        :dod2acgbedc.txt
Print      :dod2acgbprn.txt
SUMMARY STATISTICS (See documentation for explanation)

Number of input records on file 1            :   2304926
Number of input records on file 2            :   3724179
Number of output records (people)            :   2304926
Sum of unique dx codes across IDs            :  14222914
Sum of unique non-grouped dx codes across IDs:    143682
Non-grouped dx code percentage               :   1.01021%
Number of people with non-grouped dx codes   :    106039
```

ACG DISTRIBUTION

ACG	Description	Frequency	Percent
0100	Acute Minor, Age 1	4100	0.18%
0200	Acute Minor, Age 2-5	27740	1.20%
0300	Acute Minor, Age 6+	158325	6.87%
0400	Acute: Major	67870	2.94%
0500	Likely To Recur, without Allergies	58411	2.53%
0600	Likely To Recur, with Allergies	12534	0.54%
0700	Asthma	2698	0.12%
0800	Chronic Medical, Unstable	3230	0.14%
0900	Chronic Medical, Stable	24160	1.05%
1000	Chronic Specialty	2873	0.12%
1100	Ophthalmological/Dental	43540	1.89%
1200	Chronic Specialty, Unstable	1930	0.08%
1300	Psychosocial, without Psychosocial Unstable	18301	0.79%
1400	Psychosocial, with Unstable, without Stable	1367	0.06%
1500	Psychosocial, with Unstable and Stable	800	0.03%
1600	Preventive/Administrative	149292	6.48%
1710	Pregnancy: 0-1 ADGs	0	0.00%
1711	..., Delivered	2155	0.09%
1712	..., Not Delivered	2779	0.12%
1720	Pregnancy: 2-3 ADGs, No Major ADGs	0	0.00%
1721	..., Delivered	8432	0.37%
1722	..., Not Delivered	7902	0.34%
1730	Pregnancy: 2-3 ADGs, 1+ Major ADGs	0	0.00%
1731	..., Delivered	1925	0.08%
1732	..., Not Delivered	881	0.04%
1740	Pregnancy: 4-5 ADGs, No Major ADGs	0	0.00%
1741	..., Delivered	7415	0.32%
1742	..., Not Delivered	7396	0.32%
1750	Pregnancy: 4-5 ADGs, 1+ Major ADGs	0	0.00%
1751	..., Delivered	4431	0.19%
1752	..., Not Delivered	2382	0.10%
1760	Pregnancy: 6+ ADGs, No Major ADGs	0	0.00%
1761	..., Delivered	5009	0.22%
1762	..., Not Delivered	6665	0.29%
1770	Pregnancy: 6+ ADGs, 1+ Major ADGs	0	0.00%
1771	..., Delivered	10198	0.44%
1772	..., Not Delivered	8476	0.37%
1800	Acute Minor and Acute Major	104104	4.52%
1900	Acute Minor and Likely To Recur, Age 1	7959	0.35%
2000	..., Age 2-5	31925	1.39%
2100	..., Age > 5,w/out Allergy	71306	3.09%
2200	..., Age > 5,with Allergy	23195	1.01%
2300	Acute Minor and Chronic Medical: Stable	20945	0.91%
2400	Acute Minor and Eye/Dental	26704	1.16%
2500	Acute Minor, Psychosocial, Without Unstable	14582	0.63%
2600	..., Unstable without Stable	771	0.03%
2700	..., with Unstable & Stable	596	0.03%
2800	Acute Major And likely To Recur	37396	1.62%
2900	Acute Minor and Major/Likely to Recur, Age 1	8410	0.36%
3000	..., Age 2-5	26726	1.16%
3100	..., Age 6-11	20952	0.91%
3200	..., Age > 12,w/out Allergies	63569	2.76%
3300	..., Age > 12, with Allergies	18542	0.80%
3400	Acute Minor/Likely To Recur/Eye & Dental	19793	0.86%
3500	Acute Minor/Likely To Recur/Psychosocial	13762	0.60%
3600	Acute Minor/Maj/Likely to Recur/Chronic Med:Stable	63591	2.76%
3700	Acute Minor & Major/Likely to Recur/Psychosocial	25576	1.11%
3800	2-3 Other ADG Combinations, Age 1-17	54099	2.35%
3900	..., Male, Age 18-34	33128	1.44%

4000	..., Female, Age 18-34	21338	0.93%
4100	..., Age >34	100567	4.36%
4210	4-5 Other ADG Combinations, Age 1-17, No Major ADG	41315	1.79%
4220	..., Age 1-17, 1 + Major ADGs	16056	0.70%
4310	..., Age 18-44, No Major ADGs	58467	2.54%
4320	..., Age 18-44, 1 Major ADG	33371	1.45%
4330	..., Age 18-44, 2 + Major ADGs	5770	0.25%
4410	..., Age >44, No Major ADGs	31890	1.38%
4420	..., Age >44, 1 Major ADGs	25000	1.08%
4430	..., Age >44, 2+ Major ADGs	5615	0.24%
4510	6-9 Other ADG Combinations, Age 1-5, No Major ADGs	10234	0.44%
4520	..., Age 1-5, 1+ Major ADGs	8289	0.36%
4610	..., Age 6-17, No Major ADGs	17018	0.74%
4620	..., Age 6-17, 1+ Major ADGs	13199	0.57%
4710	..., Male, Age 18-34, No Major ADGs	6021	0.26%
4720	..., Male, Age 18-34, 1 Major ADGs	8237	0.36%
4730	..., Male, Age 18-34 2+ Major ADGs	3816	0.17%
4810	..., Female, Age 18-34, No Major ADGs	14223	0.62%
4820	..., Female, Age 18-34, 1 Major ADG	11260	0.49%
4830	..., Female, Age 18-34, 2+ Major ADGs	4048	0.18%
4910	..., Age >34, 0-1 Major ADGs	94557	4.10%
4920	..., Age >34, 2 Major ADGs	27565	1.20%
4930	..., Age >34, 3 Major ADGs	7288	0.32%
4940	..., Age >34, 4+ Major ADGs	1209	0.05%
5010	10+ Other ADG Combinations, Age 1-17 No Major ADGs	1972	0.09%
5020	..., Age 1-17, 1 Major ADGs	3010	0.13%
5030	..., Age 1-17, 2+ Major ADGs	2776	0.12%
5040	..., Age 18+, 0-1 Major ADGs	21940	0.95%
5050	..., Age 18+, 2 Major ADGs	17936	0.78%
5060	..., Age 18+, 3 Major ADGs	11896	0.52%
5070	..., Age 18+, 4+ Major ADGs	8987	0.39%
5100	No or Only Unclassified Diagnoses & Non-Users	0	0.00%
5110	No or Only Unclassified Diagnoses (2 input files)	538	0.02%
5200	Non-Users (2 input files)	328652	14.26%
5310	Infants: 0-5 ADGs, No Major ADGs	9	0.00%
5311	..., Low Birth Weight	0	0.00%
5312	..., Normal Birth Weight	0	0.00%
5320	Infants: 0-5 ADGs, 1 + Major ADGs	3	0.00%
5321	..., Low Birth Weight	0	0.00%
5322	..., Normal Birth Weight	0	0.00%
5330	Infants: 6+ ADGs, No Major	2	0.00%
5331	..., Low Birth Weight	0	0.00%
5332	..., Normal Birth Weight	0	0.00%
5340	Infants: 6+ ADGs, 1+ Major ADG	4	0.00%
5341	..., Low Birth Weight	0	0.00%
5342	..., Normal Birth Weight	0	0.00%
9900	Invalid Age	0	0.00%

ADG DISTRIBUTION

ADG	Description	Frequency	Percent
1	Time Limited: Minor	464696	20.16%
2	Time Limited: Minor-Primary Infections	839343	36.42%
3	Time Limited: Major	74245	3.22%
4	Time Limited: Major-Primary Infections	118691	5.15%
5	Allergies	243489	10.56%
6	Asthma	100579	4.36%
7	Likely to Recur: Discrete	325632	14.13%
8	Likely to Recur: Discrete-Infections	427205	18.53%
9	Likely to Recur: Progressive	23044	1.00%
10	Chronic Medical: Stable	480702	20.86%
11	Chronic Medical: Unstable	164385	7.13%
12	Chronic Specialty: Stable-Orthopedic	69046	3.00%
13	Chronic Specialty: Stable-Ear,Nose,Throat	43760	1.90%
14	Chronic Specialty: Stable-Eye	329305	14.29%
15	No Longer in Use	0	0.00%
16	Chronic Specialty: Unstable-Orthopedic	32066	1.39%
17	Chronic Specialty: Unstable-Ear,Nose,Throat	4051	0.18%
18	Chronic Specialty: Unstable-Eye	54822	2.38%
19	No Longer in Use	0	0.00%
20	Dermatologic	257104	11.15%
21	Injuries/Adverse Effects: Minor	342305	14.85%
22	Injuries/Adverse Effects: Major	213914	9.28%
23	Psychosocial: Time Limited, Minor	146799	6.37%
24	Psychosocial:Recurrent or Persistent,Stable	180180	7.82%
25	Psychosocial:Recurrent or Persistent,Unstable	36981	1.60%
26	Signs/Symptoms: Minor	510600	22.15%
27	Signs/Symptoms: Uncertain	787943	34.19%
28	Signs/Symptoms: Major	314435	13.64%
29	Discretionary	208782	9.06%
30	See and Reassure	100077	4.34%
31	Prevention/Administrative	1434961	62.26%
32	Malignancy	38635	1.68%
33	Pregnancy	73689	3.20%
34	Dental	10131	0.44%

	Frequency	Percent
Number of people with 0 unique ADGs:	328944	14.27%
Number of people with 1 unique ADGs:	312731	13.57%
Number of people with 2 unique ADGs:	328682	14.26%
Number of people with 3 unique ADGs:	306207	13.28%
Number of people with 4 unique ADGs:	264802	11.49%
Number of people with 5 unique ADGs:	213481	9.26%
Number of people with 6 unique ADGs:	163476	7.09%
Number of people with 7 unique ADGs:	121710	5.28%
Number of people with 8 unique ADGs:	87252	3.79%
Number of people with 9 unique ADGs:	61120	2.65%
Number of people with 10+ unique ADGs:	116521	5.06%
Number of people with 1 Major ADGs :	339449	14.73%
Number of people with 2 Major ADGs :	83228	3.61%
Number of people with 3 Major ADGs :	23863	1.04%
Number of people with 4+ Major ADGs :	10964	0.48%

Average number of ADGs per person :	3.67
Average number major ADGs per person :	0.27
Of those with an ADG, average number of ADGs per person :	4.28
Of those with an ADG, average number of major ADGs per person :	0.32
Of those with a major, average number of ADGs per person :	6.98
Of those with a major, average number of major ADGs per person:	1.37

DIAGNOSIS DISTRIBUTION

Average number of unique Dx codes per person : 6.17
Of those with a Dx code, average number of unique codes per person : 7.20

	Frequency	Percent
Number of people with 0 unique diagnosis codes:	328652	14.26%
Number of people with 1 unique diagnosis codes:	219733	9.53%
Number of people with 2 unique diagnosis codes:	221067	9.59%
Number of people with 3 unique diagnosis codes:	209375	9.08%
Number of people with 4 unique diagnosis codes:	192119	8.34%
Number of people with 5 unique diagnosis codes:	169279	7.34%
Number of people with 6 unique diagnosis codes:	146925	6.37%
Number of people with 7 unique diagnosis codes:	125731	5.45%
Number of people with 8 unique diagnosis codes:	106545	4.62%
Number of people with 9 unique diagnosis codes:	90203	3.91%
Number of people with 10+ unique diagnosis codes:	495297	21.49%

AGE/SEX DISTRIBUTION

Age	Males	Percent	Females	Percent
<1	13	0.00%	5	0.00%
1	19625	0.85%	18681	0.81%
2-5	101088	4.39%	95991	4.16%
6-11	149376	6.48%	143582	6.23%
12-17	126017	5.47%	120951	5.25%
18-24	153311	6.65%	111971	4.86%
25-34	227430	9.87%	199646	8.66%
35-44	223817	9.71%	199505	8.66%
45-54	123021	5.34%	121790	5.28%
55-59	41162	1.79%	47438	2.06%
60-64	36938	1.60%	43568	1.89%
65-69	0	0.00%	0	0.00%
70-74	0	0.00%	0	0.00%
75-79	0	0.00%	0	0.00%
80-84	0	0.00%	0	0.00%
85+	0	0.00%	0	0.00%
Unknown	0	0.00%	0	0.00%

EDC DISTRIBUTION

Number and Prevalance per Thousand of Major Expanded Diagnosis Clusters and their Component Expanded Diagnosis Clusters

EDC Description	No. Persons	No. Persons per 1000 Population
ADM Administrative	1373435	595.87
ADM01 General medical exam	1336313	579.76
ADM02 Surgical aftercare	153713	66.69
ADM03 Transplant status	1301	0.56
ADM04 Complications of mechanical devices	5052	2.19
ALL Allergy	316622	137.37
ALL01 Allergic reactions	46400	20.13
ALL03 Allergic rhinitis	221516	96.11
ALL04 Asthma, w/o status asthmaticus	96484	41.86
ALL05 Asthma, with status asthmaticus	7768	3.37
ALL06 Disorders of the immune system	2707	1.17
CAR Cardiovascular	292554	126.93
CAR01 Cardiovascular signs and symptoms	51216	22.22
CAR03 Ischemic heart disease (excluding acute myocard	25637	11.12
CAR04 Congenital heart disease	5657	2.45
CAR05 Congestive heart failure	5294	2.30
CAR06 Cardiac valve disorders	11530	5.00
CAR07 Cardiomyopathy	3684	1.60
CAR08 Heart murmur	12146	5.27
CAR09 Cardiac arrhythmia	17442	7.57
CAR10 Generalized atherosclerosis	4130	1.79
CAR11 Disorders of lipoid metabolism	136529	59.23
CAR12 Acute myocardial infarction	4358	1.89
CAR13 Cardiac arrest, shock	383	0.17
CAR14 Hypertension, w/o major complications	170838	74.12
CAR15 Hypertension, with major complications	9752	4.23
DEN Dental	22901	9.94
DEN01 Disorders of mouth	7533	3.27
DEN02 Disorders of teeth	9542	4.14
DEN03 Gingivitis	1334	0.58
DEN04 Stomatitis	5348	2.32
EAR Ears, Nose, Throat	701030	304.14
EAR01 Otitis media	217664	94.43
EAR02 Tinnitus	5743	2.49
EAR03 Temporomandibular joint disease	7013	3.04
EAR04 Foreign body in ears, nose, or throat	4864	2.11
EAR05 Deviated nasal septum	6725	2.92
EAR06 Otitis externa	37891	16.44
EAR07 Wax in ear	22746	9.87
EAR08 Deafness, hearing loss	35054	15.21
EAR09 Chronic pharyngitis and tonsillitis	15052	6.53
EAR10 Epistaxis	9007	3.91
EAR11 Acute upper respiratory tract infection	560679	243.25
END Endocrine	125575	54.48
END02 Osteoporosis	10665	4.63
END03 Short stature	1150	0.50
END04 Thyroid disease	56502	24.51
END05 Other endocrine disorders	14914	6.47

END06 Type 2 diabetes, w/o complication	37125	16.11
END07 Type 2 diabetes w/complications	6677	2.90
END08 Type 1 diabetes, w/o complication	9239	4.01
END09 Type 1 diabetes w/complications	3332	1.45
EYE Eye.....	447886	194.32
EYE01 Ophthalmic signs and symptoms	21764	9.44
EYE02 Blindness	2671	1.16
EYE03 Retinal disorders (excluding diabetic retinopat	12999	5.64
EYE04 Disorders of the eyelid and lacrimal duct	15107	6.55
EYE05 Refractive errors	294372	127.71
EYE06 Cataract, aphakia	17292	7.50
EYE07 Conjunctivitis, keratitis	103544	44.92
EYE08 Glaucoma	20309	8.81
EYE09 Infections of eyelid	16375	7.10
EYE10 Foreign body in eye	4085	1.77
EYE11 Strabismus, amblyopia	17610	7.64
EYE12 Traumatic injuries of eye	12053	5.23
EYE13 Diabetic retinopathy	4399	1.91
FRE Female Reproductive.....	318612	138.23
FRE01 Pregnancy and delivery, uncomplicated	74646	32.39
FRE02 Female genital symptoms	49566	21.50
FRE03 Endometriosis	6701	2.91
FRE04 Pregnancy and delivery with complications	53853	23.36
FRE05 Female infertility	12386	5.37
FRE06 Abnormal pap smear	34778	15.09
FRE07 Ovarian cyst	11552	5.01
FRE08 Vaginitis, vulvitis, cervicitis	74268	32.22
FRE09 Menstrual disorders	63941	27.74
FRE10 Contraception	101500	44.04
FRE11 Menopausal symptoms	32789	14.23
FRE12 Utero-vaginal prolapse	9439	4.10
GAS Gastrointestinal/Hepatic.....	267699	116.14
GAS01 Gastrointestinal signs and symptoms	37200	16.14
GAS02 Inflammatory bowel disease	4785	2.08
GAS03 Constipation	28424	12.33
GAS04 Acute hepatitis	3574	1.55
GAS05 Chronic liver disease	4755	2.06
GAS06 Peptic ulcer disease	38616	16.75
GAS07 Diarrhea	123102	53.41
GAS08 Gastroesophageal reflux	73445	31.86
GAS09 Irritable bowel syndrome	13613	5.91
GAS10 Diverticular disease of colon	11733	5.09
GAS11 Acute pancreatitis	1728	0.75
GAS12 Chronic pancreatitis	1016	0.44
GSI General Signs and Symptoms.....	233098	101.13
GSI01 Nonspecific signs and symptoms	37408	16.23
GSI02 Chest pain	86291	37.44
GSI03 Fever	57235	24.83
GSI04 Syncope	12691	5.51
GSI05 Nausea, vomiting	46188	20.04
GSI06 Debility and undue fatigue	1329	0.58
GSI07 Lymphadenopathy	11688	5.07
GSI08 Edema	11631	5.05
GSU General Surgery.....	332404	144.21
GSU01 Anorectal conditions	38475	16.69
GSU02 Appendicitis	3460	1.50
GSU03 Benign and unspecified neoplasm	89552	38.85
GSU04 Cholelithiasis, cholecystitis	10458	4.54

GSU05 External abdominal hernias, hydroceles	14767	6.41
GSU06 Chronic cystic disease of the breast	17458	7.57
GSU07 Other breast disorders	29158	12.65
GSU08 Varicose veins of lower extremities	6017	2.61
GSU09 Nonfungal infections of skin and subcutaneous t	73602	31.93
GSU10 Abdominal pain	127143	55.16
GSU11 Peripheral vascular disease	5139	2.23
GSU12 Burns--1st degree	1277	0.55
GSU13 Aortic aneurysm	949	0.41
GSU14 Gastrointestinal obstruction/perforation	5699	2.47
GTC Genetic.....	1370	0.59
GTC01 Chromosomal anomalies	1370	0.59
GUR Genito-urinary.....	176886	76.74
GUR01 Vesicoureteral reflux	1598	0.69
GUR02 Undescended testes	958	0.42
GUR03 Hypospadias, other penile anomalies	1216	0.53
GUR04 Prostatic hypertrophy	2943	1.28
GUR05 Stricture of urethra	1514	0.66
GUR06 Urinary symptoms	57661	25.02
GUR07 Other male genital disease	24378	10.58
GUR08 Urinary tract infections	99198	43.04
GUR09 Renal calculi	12153	5.27
GUR10 Prostatitis	5357	2.32
HEM Hematologic.....	43093	18.70
HEM01 Hemolytic anemia	2404	1.04
HEM02 Iron deficiency, other deficiency anemias	34659	15.04
HEM03 Thrombophlebitis	3543	1.54
HEM04 Neonatal jaundice	257	0.11
HEM05 Aplastic anemia	629	0.27
HEM06 Deep vein thrombosis	2686	1.17
HEM07 Hemophilia, coagulation disorder	2309	1.00
INF Infections.....	177704	77.10
INF01 Tuberculosis infection	4461	1.94
INF02 Fungal infections	12874	5.59
INF03 Infectious mononucleosis	2954	1.28
INF04 HIV, AIDS	1121	0.49
INF05 Sexually transmitted diseases	28340	12.30
INF06 Viral syndromes	131949	57.25
INF07 Lyme disease	595	0.26
INF08 Septicemia	2898	1.26
MAL Malignancies.....	40157	17.42
MAL01 Malignant neoplasms of the skin	9378	4.07
MAL02 Low impact malignant neoplasms	13043	5.66
MAL03 High impact malignant neoplasms	4008	1.74
MAL04 Malignant neoplasms, breast	6022	2.61
MAL05 Malignant neoplasms, cervix, uterus	1870	0.81
MAL06 Malignant neoplasms, ovary	718	0.31
MAL07 Malignant neoplasms, esophagus	297	0.13
MAL08 Malignant neoplasms, kidney	556	0.24
MAL09 Malignant neoplasms, liver and biliary tract	155	0.07
MAL10 Malignant neoplasms, lung	939	0.41
MAL11 Malignant neoplasms, lymphomas	3023	1.31
MAL12 Malignant neoplasms, colorectal	1866	0.81
MAL13 Malignant neoplasms, pancreas	152	0.07
MAL14 Malignant neoplasms, prostate	2572	1.12
MAL15 Malignant neoplasms, stomach	148	0.06
MAL16 Acute leukemia	554	0.24
MAL18 Malignant neoplasms, bladder	731	0.32

MUS Musculoskeletal.....	616851	267.62
MUS01 Musculoskeletal signs and symptoms	261363	113.39
MUS02 Acute sprains and strains	205205	89.03
MUS03 Degenerative joint disease	43661	18.94
MUS04 Fractures (excluding digits)	59263	25.71
MUS05 Torticollis	1951	0.85
MUS06 Kyphoscoliosis	6781	2.94
MUS07 Congenital hip dislocation	305	0.13
MUS08 Fractures and dislocations/digits only	15208	6.60
MUS09 Joint disorders, trauma related	57300	24.86
MUS10 Fracture of neck of femur (hip)	2524	1.10
MUS11 Congenital anomalies of limbs, hands, and feet	9651	4.19
MUS12 Acquired foot deformities	25819	11.20
MUS13 Cervical pain syndromes	58444	25.36
MUS14 Low back pain	184140	79.89
MUS15 Bursitis, synovitis, tenosynovitis	114309	49.59
MUS16 Amputation status	1096	0.48
NUR Neurologic.....	232398	100.83
NUR01 Neurologic signs and symptoms	7823	3.39
NUR02 Headaches	110965	48.14
NUR03 Peripheral neuropathy, neuritis	51518	22.35
NUR04 Vertiginous syndromes	32810	14.23
NUR05 Cerebrovascular disease	9227	4.00
NUR06 Parkinson's disease	1727	0.75
NUR07 Seizure disorder	13210	5.73
NUR08 Multiple sclerosis	2727	1.18
NUR09 Muscular dystrophy	844	0.37
NUR10 Sleep problems	17523	7.60
NUR11 Dementia and delirium	1780	0.77
NUR12 Quadriplegia and paraplegia	905	0.39
NUR15 Head injury	6212	2.70
NUR16 Spinal cord injury/disorders	3761	1.63
NUR17 Paralytic syndromes, other	1728	0.75
NUR18 Cerebral palsy	2189	0.95
NUR19 Developmental disorder	15693	6.81
NUT Nutrition.....	66481	28.84
NUT01 Failure to thrive	8017	3.48
NUT02 Nutritional deficiencies	4958	2.15
NUT03 Obesity	54383	23.59
PSY Psychosocial.....	267033	115.85
PSY01 Anxiety, neuroses	140170	60.81
PSY02 Substance use	14958	6.49
PSY03 Tobacco abuse	34460	14.95
PSY04 Behavior problems	10330	4.48
PSY05 Attention deficit disorder	38078	16.52
PSY06 Family and social problems	58505	25.38
PSY07 Schizophrenia and affective psychosis	7757	3.37
PSY08 Personality disorders	4745	2.06
PSY09 Depression	50554	21.93
REC Reconstructive.....	77744	33.73
REC01 Cleft lip and palate	702	0.30
REC02 Lacerations	69699	30.24
REC03 Chronic ulcer of the skin	1676	0.73
REC04 Burns--2nd and 3rd degree	6598	2.86
REN Renal.....	31164	13.52
REN01 Chronic renal failure	2670	1.16
REN02 Fluid/electrolyte disturbances	27102	11.76

REN03 Acute renal failure	1083	0.47
REN04 Nephritis, nephrosis	2835	1.23
RES Respiratory.....	380726	165.18
RES01 Respiratory signs and symptoms	36021	15.63
RES02 Acute lower respiratory tract infection	154509	67.03
RES03 Cystic fibrosis	481	0.21
RES04 Emphysema, chronic bronchitis, COPD	18255	7.92
RES05 Cough	63596	27.59
RES06 Sleep apnea	16109	6.99
RES07 Sinusitis	187399	81.30
RES08 Pulmonary embolism	865	0.38
RES09 Tracheostomy	561	0.24
RES10 Respiratory arrest	7344	3.19
RHU Rheumatologic.....	44105	19.14
RHU01 Autoimmune and connective tissue diseases	16422	7.12
RHU02 Gout	7707	3.34
RHU03 Arthropathy	21798	9.46
RHU04 Raynaud's syndrome	1101	0.48
SKN Skin.....	549607	238.45
SKN01 Contusions and abrasions	111160	48.23
SKN02 Dermatitis and eczema	141492	61.39
SKN03 Keloid	4213	1.83
SKN04 Acne	56914	24.69
SKN05 Disorders of sebaceous glands	6893	2.99
SKN06 Sebaceous cyst	18769	8.14
SKN07 Viral warts and molluscum contagiosum	49861	21.63
SKN08 Other inflammatory conditions of skin	9824	4.26
SKN09 Exanthems	86741	37.63
SKN10 Skin keratoses	28497	12.36
SKN11 Dermatophytoses	59464	25.80
SKN12 Psoriasis	6912	3.00
SKN13 Disease of hair and hair follicles	19809	8.59
SKN14 Pigmented nevus	13310	5.77
SKN15 Scabies and pediculosis	6171	2.68
SKN16 Diseases of nail	19340	8.39
SKN17 Other skin disorders	54940	23.84
SKN18 Benign neoplasm of skin and subcutaneous tissue	54215	23.52
TOX Toxic Effects.....	18137	7.87
TOX01 Toxic effects of nonmedicinal agents	5915	2.57
TOX02 Adverse effects of medicinal agents	12563	5.45
UDC Unassigned.....	885250	384.07
UDC00 Unassigned diagnosis code	885250	384.07

Number of people with 0 unique EDCs:	352965	15.31%
Number of people with 1 unique EDCs:	321465	13.95%
Number of people with 2 unique EDCs:	317963	13.79%
Number of people with 3 unique EDCs:	281167	12.20%
Number of people with 4 unique EDCs:	236025	10.24%
Number of people with 5 unique EDCs:	191421	8.30%
Number of people with 6 unique EDCs:	149470	6.48%
Number of people with 7 unique EDCs:	113996	4.95%
Number of people with 8 unique EDCs:	85493	3.71%
Number of people with 9 unique EDCs:	64573	2.80%
Number of people with 10+ unique EDCs:	190388	8.26%

Average number of EDCs per person :	3.98
Of those with an EDC, average number of EDCs per person :	4.70

Attachment A4_DCG_1

```
1    /************************************/
2    /* DCG model:                       */
3    /* Jenn Fonda                       */
4    /* First created: Aug 16, 2004      */
5    /* Last modified: Dec  1, 2004      */
6    /************************************/
7
8    /* this program will provide the HCC's and Dx groups for the
500,000 in validation sample
8  ! */
9
10
11
12    Options
13      compress= yes
14      ls= 80
15      obs= max
16      nocenter
17      mstored
18      validvarname=upcase
19      sasmstore= SASAVE;
20
21    * SAS macro library for DxCG;
22    Libname SASAVE 'C:\Dxcg\61\sasave';
NOTE: Libref SASAVE was successfully assigned as follows:
      Engine:        V8
      Physical Name: C:\Dxcg\61\sasave
23
24    * Person file location;
25    Libname INPERS 'C:\Dxcg\DOD1';
NOTE: Libref INPERS was successfully assigned as follows:
      Engine:        V8
      Physical Name: C:\Dxcg\DOD1
26
27    * Diag file location;
28    Libname INDIAG 'C:\Dxcg\DOD1';
NOTE: Libname INDIAG refers to the same physical library as INPERS.
NOTE: Libref INDIAG was successfully assigned as follows:
      Engine:        V8
      Physical Name: C:\Dxcg\DOD1
29
30    * DXCG and DXCGS output file location;
31    Libname OUT    'C:\Dxcg\DOD1';
NOTE: Libname OUT refers to the same physical library as INDIAG.
NOTE: Libref OUT was successfully assigned as follows:
      Engine:        V8
      Physical Name: C:\Dxcg\DOD1
32
33    * Location where data quality appendix will be placed;
34    Filename APPENDIX 'C:\Dxcg\DOD1\dod1_append.txt';
35
36
37    %dxcg(
38     __indiag= INDIAG.diag,  /* Name of Diagnosis File input dataset
```

```
38 !  */
39     __inpers= INPERS.pers,  /* Name of Person File input dataset
39 !  */
40     __out   = OUT.dod_dxcg,     /* Libref.Name - Main Output dataset
40 !   */
41     __detail= 3,            /* Diagnostic detail to output
0,1,2,3,4,5
41 !  */
42     __hier=yes,     /* Heirarchy for CCs
42 !  */
43     __popgrp= C,    /* A= Medicare, B= Medicaid, C= Commercial
43 !  */
44     __modvar= 3,    /* 1= Use only inpatient data, 3= Use all-
encounter data
44 !  */
45     __purpos= E,    /* E= Explanation, P= Payment
45 !  */
46     __outcom= A,    /* A= Medical+Pharmacy costs,
46 !  */
47     __ndiag = 6,    /* Maximum number of ICD9 diags per record
47 !  */
48     __append= APPENDIX,      /* Fileref of text file for appendix
output
48 !   */
49     __sumexp= no          /* Specify "no" for expenditure info in
person file
49 !  */
50     );


************************************************
* DxCG - Promoting Fair and Efficient Health Care *
*                                              *
* DxCG Risk Adjustment Software Release 6.1       *
*                                              *
* Date:  2005-01-25                            *
*                                              *
* Time:  13:18:32                              *
************************************************



DxCG Note:    __NDIAG = 6
     6 DIAG variables expected on INDIAG.diag.

DxCG Note:    __NPROC = 0
     0 PROC variables expected on INDIAG.diag.


Checking the DIAGNOSIS data set INDIAG.diag ...

DxCG Warning:  SOURCE
     Optional variable not found on the input data set.
     It will be assigned a default value.

Warnings have been issued concerning INDIAG.diag.
```

```
The input requirements for INDIAG.diag have been satisfied.


**************************************************************************
*
* DxCG WARNING:  Warnings have been issued concerning INDIAG.diag.
*
*
*
* Please review the SAS log carefully to determine the nature of these
*
* warnings.
*
*
*
* DxCG Return Code: 4
*
**************************************************************************
*



Checking the PERSON data set INPERS.pers ...

DxCG Warning:  ELIG2
     Optional variable not found on the input data set.
     It will be assigned a default value and automatically
     added to the DxCG output data set.

DxCG Warning:  EXPEND1
     Optional variable not found on the input data set.
     It will be automatically added to the DxCG output data set.

Warnings have been issued concerning INPERS.pers.

The input requirements for INPERS.pers have been satisfied.


**************************************************************************
*
* DxCG WARNING:  Warnings have been issued concerning INPERS.pers.
*
*
*
* Please review the SAS log carefully to determine the nature of these
*
* warnings.
*
*
*
* DxCG Return Code: 5
*
**************************************************************************
*
```

```
DxCG Note:  user variables from the data set INPERS.pers:
===> DATATYPE
have been added to the DxCG output data set.


*******************************************************
* DxCG Release 6.1 dxcg application                   *
*                                                     *
* The application will evaluate the following model:  *
*                                                     *
* AllEnc/Comm/Expl/Med+Rx (3CEA)                      *
*                                                     *
* __MODVAR:  3                                        *
*                                                     *
* __POPGRP:  C                                        *
*                                                     *
* __PURPOS:  E                                        *
*                                                     *
* __OUTCOM:  A                                        *
*                                                     *
*******************************************************


NOTE: There were 5343739 observations read from the data set
INDIAG.DIAG.
NOTE: There were 2304926 observations read from the data set
INPERS.PERS.
NOTE: The data set OUT.DOD_DXCG has 2304926 observations and 1022
variables.
NOTE: Compressing data set OUT.DOD_DXCG decreased size by 94.94
percent.
      Compressed is 23316 pages; un-compressed would require 460994
pages.
NOTE: The data set OUT.DOD_DXCGS has 1 observations and 112 variables.
NOTE: Compressing data set OUT.DOD_DXCGS increased size by 100.00
percent.
      Compressed is 2 pages; un-compressed would require 1 pages.
NOTE: DATA statement used:
      real time           20:02.88
      cpu time            19:58.37




*********************************************************************
*********
* DxCG Note: The DxCG Data Quality Report has been output to the SAS
list file *
* and to an external file referenced by the filename APPENDIX.
*
*********************************************************************
*********

NOTE: Fileref= APPENDIX
      Physical Name= C:\Dxcg\DOD1\dod1_append.txt
```

```
51
52
53      * Location where report from TABLES() macro will be placed;
54     Filename REPORT 'C:\Dxcg\DOD1\dod1_report.txt';
55
56
57       %tables(
58       __dsn   = OUT.dod_dxcg,  /* Input SAS dataset name
58 !       */
59       __weight= 1,         /* Weighting by eligible months: Values 0-
2
59 !   */
60       __year  = 2,         /* 1= Concurrent, 2= Prospective
60 !   */
61       __report= REPORT     /* Output name for Summary Statistics file
61 !   */
62       );


************************************************
* DxCG - Promoting Fair and Efficient Health Care *
*                                              *
* DxCG Risk Adjustment Software Release 6.1    *
*                                              *
* Date:  2005-01-25                            *
*                                              *
* Time:  13:38:39                              *
************************************************



Checking parameter values ...

The input requirements for %tables have been satisfied.


************************************************************************
*********
* DxCG Release 6.1 tables application
*
*
*
* The application will generate a Summary Statistics File from the DxCG
output *
* data set: OUT.dod_dxcg for model: All-
encounter/Prospective/Comm/Expl/Med+Rx *
*
*
* Statistics are based on the following parameters:
*
*
*
* __GRPVAR:  None   (Totals will output)
*
*
*
```

```
* __GRPUDF:  $ALL.  (Totals will output)
*
*
*
* __GRPLAB:  Total
*
*
*
* __YEAR:    2 (Prospective)
*
*
*
* __WEIGHT:  Weighted by eligible months
*
*
*
* __NORM1:   2259
*
*
*
* __NORM2:   2184
*
*
*
*********************************************************************
*********



*********************************************************************
********
* DxCG Note: The DxCG Summary Statistics File has been output to the
external *
* file referenced by the filename REPORT.
*
*
*
* Format this text file with the Excel report template to create the
DxCG     *
* Standard Report
*
*********************************************************************
********

NOTE: Fileref= REPORT
      Physical Name= C:\Dxcg\DOD1\dod1_report.txt
```

Attachment A4_DCG_2

DxCG, Inc.
REPORT PARAMETERS AND USER OPTIONS

SYSTEM PARAMETERS:

Date:	January 25, 2005
Time:	13:39:00
Licensed to:	Arlene Ash - DOD Tricare Prime
DxCG Release:	6.1
Operating environment:	SAS 8.2 WIN
DxCG software serial number:	12312005
Software expiration date:	December 31, 2005
Maximum licensed population size:	2,300,000

INPUT FILES:

Main DxCG person-level file:	OUT.dod_dxcg
DxCG summary file:	OUT.dod_dxcgS
Population group:	Commercial
Model variant:	All-encounter, DCG/HCC
Model purpose:	Explanation
Model outcome:	Medical Expenses Including Pharmacy Spending
Level of clinical detail available:	ACCs, CCs and DxGroups
Hierarchies imposed:	YES
Over age 64 included:	NO
Number of people in main DxCG file:	2,304,926

OPTIONS:

Model year:	Prospective
Year description:	
Weighting:	Weighted by Eligible Months
Expenditure average for Year 1:	Not used
Expenditure average for Year 2:	2184
Abbreviation of model dimensions:	All-encounter/Prospective/Comm/Expl/Med+Rx
Group variable used:	None (Totals will output)
Format used with group variable:	$ALL.
Group variable name:	Total

DxCG, Inc.
Table 1.1: Summary of DCG Predictions
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx
Weighted by Eligible Months

	Benchmark	Current Sample
		Total
		Distribution of Individuals
Eligible year equivalents (Year 2)	2,255,992	2,304,926
Percent of Sample	100.00	100.00
		Relative Risk Scores
		Normalized to Benchmark Sample
Age/Sex Model	1.00	0.81
Prospective (Year 2)	1.00	1.01
		Relative Risk Scores
		Normalized to Current Sample
Age/Sex Model		1.00
Prospective (Year 2)		1.00

DxCG, Inc.
Table 2.1: Eligibility Information
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Current Sample Total
Number of People	2,304,926
Percent of Sample	100.00
Year 1	
People ever eligible in year	2,304,926
People eligible for all 12 months	2,304,926
Percent eligible for all 12 months	100.00
Eligible year equivalents	2,304,926
Percent of eligible years in group	100.00
Mean months of eligibility	12.00
Year 2	
People ever eligible in year	2,304,926
People eligible for all 12 months	2,304,926
Percent eligible for all 12 months	100.00
Eligible year equivalents	2,304,926
Percent of eligible years in group	100.00
Mean months of eligibility	12.00

DxCG, Inc.
Table 3.1: Summary of Age and Sex Distribution
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Benchmark	Current Sample Total
Total People	100.00	100.00
Female	51.14	47.86
Male	48.86	52.14
Child: Age 0 to 17	24.82	33.64
Young Adult: Age 18 to 44	41.39	48.40
Older Adult: Age 45 to 64	32.64	17.96
Senior: Age 65+	1.15	0.00
Mean Age	33.72	27.61

DxCG, Inc.
Table 3.2: Percent Distribution by Age and Sex
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Benchmark	Current Sample Total
Total People	**100.00**	**100.00**
Female, Age 0 to 5	3.41	4.98
Female, Age 6 to 12	4.86	7.18
Female, Age 13 to 17	3.81	4.29
Female, Age 18 to 24	4.97	4.86
Female, Age 25 to 34	6.92	8.66
Female, Age 35 to 44	9.45	8.66
Female, Age 45 to 54	9.96	5.28
Female, Age 55 to 59	4.00	2.06
Female, Age 60 to 64	3.20	1.89
Female, Age 65+	0.55	0.00
Total Female, All Ages	**51.14**	**47.86**
Male, Age 0 to 5	3.60	5.24
Male, Age 6 to 12	5.12	7.48
Male, Age 13 to 17	4.02	4.47
Male, Age 18 to 24	5.18	6.65
Male, Age 25 to 34	6.42	9.87
Male, Age 35 to 44	8.45	9.71
Male, Age 45 to 54	8.59	5.34
Male, Age 55 to 59	3.70	1.79
Male, Age 60 to 64	3.19	1.60
Male, Age 65+	0.60	0.00
Total Male, All Ages	**48.87**	**52.14**

DxCG, Inc.
Table 3.3: Frequency Distribution by Age and Sex
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Current Sample Total
Total People	**2,304,926**
Female, Age 0 to 5	114,677
Female, Age 6 to 12	165,587
Female, Age 13 to 17	98,946
Female, Age 18 to 24	111,971
Female, Age 25 to 34	199,646
Female, Age 35 to 44	199,505
Female, Age 45 to 54	121,790
Female, Age 55 to 59	47,438
Female, Age 60 to 64	43,568
Female, Age 65+	0
Total Female, All Ages	**1,103,128**
Male, Age 0 to 5	120,726
Male, Age 6 to 12	172,384
Male, Age 13 to 17	103,009
Male, Age 18 to 24	153,311
Male, Age 25 to 34	227,430
Male, Age 35 to 44	223,817
Male, Age 45 to 54	123,021
Male, Age 55 to 59	41,162
Male, Age 60 to 64	36,938
Male, Age 65+	0
Total Male, All Ages	**1,201,798**

DxCG, Inc.
Table 4.1: Number of Individuals by Aggregated Condition Category (ACC)
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

Aggregated Condition Category (ACC)	Current Sample Total
All People	2,304,926
No Claims	328,652
No Valid Diagnosis	866
01: Infectious and Parasitic	336,446
02: Malignant Neoplasm	41,593
03: Benign/In Situ/Uncertain Neoplasm	133,888
04: Diabetes	57,446
05: Nutritional and Metabolic	253,189
06: Liver	18,246
07: Gastrointestinal	296,674
08: Musculoskeletal and Connective Tissue	526,719
09: Hematological	50,649
10: Cognitive Disorders	8,580
11: Substance Abuse	47,532
12: Mental	167,402
13: Developmental Disability	54,634
14: Neurological	103,018
15: Cardio-Respiratory Arrest	4,859
16: Heart	203,334
17: Cerebro-Vascular	11,965
18: Vascular	43,887
19: Lung	256,706
20: Eyes	452,336
21: Ears, Nose and Throat	868,170
22: Urinary System	141,791
23: Genital System	249,558
24: Pregnancy-Related	78,904
25: Skin and Subcutaneous	380,176
26: Injury, Poisoning, Complications	474,272
27: Symptoms, Signs and Ill-Defined Conditions	880,074
28: Neonates	1,369
29: Transplants, Openings, Other V-Codes	3,235
30: Screening / History	1,430,537

DxCG, Inc.
Table 4.2: Rate per 10,000 Individuals by Aggregated Condition Category (ACC)
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

Aggregated Condition Category (ACC)	Benchmark	Current Sample Total
All People	10,000	10,000
No Claims	2984	1,426
No Valid Diagnosis	127	4
01: Infectious and Parasitic	884	1,460
02: Malignant Neoplasm	270	180
03: Benign/In Situ/Uncertain Neoplasm	890	581
04: Diabetes	267	249
05: Nutritional and Metabolic	1,013	1,098
06: Liver	91	79
07: Gastrointestinal	958	1,287
08: Musculoskeletal and Connective Tissue	1,939	2,285
09: Hematological	237	220
10: Cognitive Disorders	41	37
11: Substance Abuse	39	206
12: Mental	600	726
13: Developmental Disability	97	237
14: Neurological	343	447
15: Cardio-Respiratory Arrest	28	21
16: Heart	1,019	882
17: Cerebro-Vascular	70	52
18: Vascular	169	190
19: Lung	1,089	1,114
20: Eyes	669	1,962
21: Ears, Nose and Throat	2,722	3,767
22: Urinary System	556	615
23: Genital System	1,181	1,083
24: Pregnancy-Related	206	342
25: Skin and Subcutaneous	1,326	1,649
26: Injury, Poisoning, Complications	1,413	2,058
27: Symptoms, Signs and Ill-Defined Conditions	2,468	3,818
28: Neonates	67	6
29: Transplants, Openings, Other V-Codes	13	14
30: Screening / History	2,866	6,206

DxCG, Inc.
Table 4.3: Number of Individuals by Condition Category (CC)
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx
Hierarchies are Imposed

Condition Category (CC)	Current Sample Total
All People	2,304,926
No Claims	328,652
No Valid Diagnosis	866
ACC001: Infectious and Parasitic	336,446
001 HIV/AIDS	1,130
002 Septicemia/Shock	1,465
003 Central Nervous System Infection	2,717
004 Tuberculosis	1,475
005 Opportunistic Infections	481
006 Other Infectious Diseases	329,375
ACC002: Malignant Neoplasm	41,593
007 Metastatic Cancer and Acute Leukemia	2,338
008 Lung, Upper Digestive Tract, and Other Sev. Cancers	1,520
009 Lymphatic, Head and Neck, Brain, Other Maj. Cancers	5,081
010 Breast, Prostate, Colorectal, Other Cancers/Tumors	16,778
011 Other Respiratory and Heart Neoplasms	793
012 Other Digestive and Urinary Neoplasms	15,083
ACC003: Benign/In Situ/Uncertain Neoplasm	133,888
013 Other Neoplasms	43,542
014 Benign Neoplasms of Skin, Breast, Eye	78,904
ACC004: Diabetes	57,446
015 Diabetes with Renal Manifestation	1,702
016 Diabetes with Neurologic or Periph. Circ. Manifest.	3,057
017 Diabetes with Acute Complications	2,375
018 Diabetes with Ophthalmologic Manifestation	5,180
019 Diabetes with No or Unspecified Complications	45,132
020 Type I Diabetes Mellitus	12,573
ACC005: Nutritional and Metabolic	253,189
021 Protein-Calorie Malnutrition	874
022 Other Significant Endocrine and Metabolic Disorders	7,449
023 Disorders of Fluid/Electrolyte/Acid-Base Balance	26,836
024 Other Endocrine/Metabolic/Nutritional Disorders	219,066
ACC006: Liver	18,246
025 End-Stage Liver Disease	614
026 Cirrhosis of Liver	812
027 Chronic Hepatitis	2,591
028 Acute Liver Failure/Disease	744
029 Other Hepatitis and Liver Disease	6,103

030 Gallbladder and Biliary Tract Disorders	8,263
ACC007: Gastrointestinal	296,674
031 Intestinal Obstruction/Perforation	3,698
032 Pancreatic Disease	3,595
033 Inflammatory Bowel Disease	4,774
034 Peptic Ulcer, Hemorrhage, Other Specified GI Dis.	22,342
035 Appendicitis	3,368
036 Other Gastrointestinal Disorders	260,963
ACC008: Musculoskeletal and Connective Tissues	526,719
037 Bone/Joint/Muscle Infections/Necrosis	3,861
038 Rheumatoid Arthritis, Inflammatory Connect. Tissue Dis.	14,814
039 Disorders of the Vertebrae and Spinal Discs	39,725
040 Osteoarthritis of Hip or Knee	13,273
041 Osteoporosis and Other Bone/Cartilage Disorders	34,651
042 Congenital/Developmental Skeletal, Connect. Tissue Dis.	1,893
043 Other Musculoskeletal and Connective Tissue Disorders	423,416
ACC009: Hematological	50,649
044 Severe Hematological Disorders	1,969
045 Disorders of Immunity	2,963
046 Coagulation defd, Other Specified Hematological Dis.	6,739
047 Iron Deficiency, Other/Unspecified Anemias, Blood Dis.	39,499
ACC010: Cognitive Disorders	8,580
048 Delirium and Encephalopathy	2,776
049 Dementia	2,653
050 Senility, Nonpsychotic Organic Brain Syndromes/Cond.	3,262
ACC011: Substance Abuse	47,532
051 Drug/Alcohol Psychosis	1,738
052 Drug/Alcohol Dependence	6,864
053 Drug/Alcohol Abuse, Without Dependence	38,930
ACC012: Mental	167,402
054 Schizophrenia	1,679
055 Major Depressive, Bipolar, and Paranoid Disorders	37,363
056 Reactive and Unspecified Psychosis	784
057 Personality Disorders	2,287
058 Depression	53,652
059 Anxiety Disorders	12,825
060 Other Psychiatric Disorders	58,812
ACC013: Developmental Disability	54,634
061 Profound Mental Retardation/Developmental Disability	259
062 Severe Mental Retardation/Developmental Disability	222
063 Moderate Mental Retardation/Developmental Disability	199
064 Mild/Unspecified Mental Retardation/Develop. Disability	4,119
065 Other Developmental Disability	18,658
066 Attention Deficit Disorder	31,177
ACC014: Neurological	103,018
067 Quadriplegia, Other Extensive Paralysis	884
068 Paraplegia	596
069 Spinal Cord Disorders/Injuries	3,373

070 Muscular Dystrophy	293
071 Polyneuropathy	9,336
072 Multiple Sclerosis	2,643
073 Parkinson's and Huntington's Diseases	562
074 Seizure Disorders and Convulsions	12,287
075 Coma, Brain Compression/Anoxic Damage	573
076 Mononeuropathy, Other Neurological Conditions/Inj.	74,659
ACC015: Cardio-Respiratory Arrest	4,859
077 Respirator Dependence/Tracheostomy Status	536
078 Respiratory Arrest	129
079 Cardio-Respiratory Failure and Shock	4,194
ACC016: Heart	203,334
080 Congestive Heart Failure	8,416
081 Acute Myocardial Infarction	1,542
082 Unstable Angina, Other Acute Ischemic Heart Disease	4,624
083 Angina Pectoris/Old Myocardial Infarction	6,239
084 Coronary Atherosclerosis/Oth. Chron. Ischemic Heart Dis.	13,598
085 Heart Infection/Inflammation, Except Rheumatic	1,070
086 Valvular and Rheumatic Heart Disease	13,506
087 Major Congenital Cardiac/Circulatory Defect	754
088 Other Congenital Heart/Circulatory Disease	2,598
089 Hypertensive Heart and Renal Dis. or Encephalopathy	782
090 Hypertensive Heart Disease	3,123
091 Hypertension	163,405
092 Specified Heart Arrhythmias	8,477
093 Other Heart Rhythm and Conduction Disorders	8,856
094 Other and Unspecified Heart Disease	2,664
ACC017: Cerebro-Vascular	11,965
095 Cerebral Hemorrhage	712
096 Ischemic or Unspecified Stroke	3,088
097 Precerebral Arterial Occl., Transient Cerebral Ischemia	3,885
098 Cerebral Atherosclerosis and Aneurysm	498
099 Cerebrovascular Disease, Unspecified	376
100 Hemiplegia/Hemiparesis	1,234
101 Diplegia (Upper), Monoplegia, Other Paralytic Synd.	1,556
102 Speech, Language, Cognitive, Perceptual Deficits	918
103 Cerebrovascular Disease Late Effects, Unspecified	348
ACC018: Vascular	43,887
104 Vascular Disease with Complications	2,357
105 Vascular Disease	8,801
106 Other Circulatory Disease	32,729
ACC019: Lung	256,706
107 Cystic Fibrosis	460
108 Chronic Obstructive Pulmonary Disease	17,107
109 Fibrosis of Lung and Other Chronic Lung Disorders	5,693
110 Asthma	94,104
111 Aspiration and Specified Bacterial Pneumonias	1,442
112 Pneumococcal Pneumonia, Empyema, Lung Abscess	1,437

113 Viral and Unspecified Pneumonia, Pleurisy	36,818
114 Pleural Effusion/Pneumothorax	2,551
115 Other Lung Disorders	110,376
ACC020: Eyes	452,336
116 Legally Blind	251
117 Major Eye Infections/Inflammations	2,918
118 Retinal Detachment	1,344
119 Proliferative Diabetic Retinopathy, Vitreous Hemorrhage	1,279
120 Diabetic and Other Vascular Retinopathies	6,116
121 Retinal Dis., Exc. Detachment, Vascular Retinopathies	8,027
122 Glaucoma	20,240
123 Cataract	16,235
124 Other Eye Disorders	402,531
ACC021: Ears, Nose and Throat	868,170
125 Significant Ear, Nose, and Throat Disorders	7,032
126 Hearing Loss	32,653
127 Other Ear, Nose, Throat, and Mouth Disorders	829,899
ACC022: Urinary System	141,791
128 Kidney Transplant Status	504
129 End Stage Renal Disease	0
130 Dialysis Status	348
131 Renal Failure	2,887
132 Nephritis	2,047
133 Urinary Obstruction and Retention	18,893
134 Incontinence	14,807
135 Urinary Tract Infection	98,128
136 Other Urinary Tract Disorders	26,113
ACC023: Genital System	249,558
137 Female Infertility	12,053
138 Pelvic Inflammatory Dis., Oth. Spec. Female Genital Dis.	30,180
139 Other Female Genital Disorders	164,816
140 Male Genital Disorders	42,595
ACC024: Pregnancy Related	78,904
141 Ectopic Pregnancy	1,773
142 Miscarriage/Early Termination	7,023
143 Completed Pregnancy With Major Complications	5,728
144 Completed Pregnancy With Complications	30,292
145 Completed Pregnancy Without Comp. (Normal Delivery)	11,414
146 Uncompleted Pregnancy With Complications	3,894
147 Uncompleted Pregnancy With No or Minor Complications	20,620
ACC025: Skin and Subcutaneous	380,176
148 Decubitus Ulcer of Skin	597
149 Chronic Ulcer of Skin, Except Decubitus	1,869
150 Extensive Third-Degree Burns	27
151 Other Third-Degree and Extensive Burns	383
152 Cellulitis, Local Skin Infection	53,699
153 Other Dermatological Disorders	323,929
ACC026: Injury, Poisoning, Complications	474,272

154 Severe Head Injury	143
155 Major Head Injury	5,867
156 Concussion or Unspecified Head Injury	10,806
157 Vertebral Fractures	1,414
158 Hip Fracture/Dislocation	2,473
159 Major Fracture, Except of Skull, Vertebrae, or Hip	15,686
160 Internal Injuries	1,482
161 Traumatic Amputation	429
162 Other Injuries	396,293
163 Poisonings and Allegic Reactions	49,927
164 Major Complications of Medical Care and Trauma	7,984
165 Other Complications of Medical Care	11,590
ACC027: Symptoms, Signs and Ill-Defined Conditions	880,074
166 Major Symptoms, Abnormalities	325,014
167 Minor Symptoms, Signs, Findings	555,060
ACC028: Neonates	1,369
168 Extremely Low Birthweight Neonates	62
169 Very Low Birthweight Neonates	34
170 Serious Perinatal Problem Affecting Newborn	571
171 Other Perinatal Problems Affecting Newborn	678
172 Normal, Single Birth	24
ACC029: Transplants, Openings, Other V-Codes	3,235
173 Major Organ Transplant	0
174 Major Organ Transplant Status	551
175 Other Organ Transplant/Replacement	649
176 Artificial Openings for Feeding or Elimination	1,572
177 Amputation Status, Lower Limb/Amputation Comp.	281
178 Amputation Status, Upper Limb	62
ACC030: Screening / History	1,430,537
179 Post-Surgical States/Aftercare/Elective	115,543
180 Radiation Therapy	6,402
181 Chemotherapy	2,576
182 Rehabilitation	45,499
183 Screening/Observation/Special Exams	1,389,548
184 History of Disease	57,083

DxCG, Inc.
Table 4.4: Rate per 10,000 Individuals by Condition Category (CC)
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx
Hierarchies are Imposed

Condition Category (CC)	Benchmark	Current Sample Total
All People	10,000	10,000
No Claims	2,984	1,426
No Valid Diagnosis	127	4
ACC001: Infectious and Parasitic	884	1,460
001 HIV/AIDS	1	5
002 Septicemia/Shock	11	6
003 Central Nervous System Infection	9	12
004 Tuberculosis	4	6
005 Opportunistic Infections	3	2
006 Other Infectious Diseases	857	1,429
ACC002: Malignant Neoplasm	270	180
007 Metastatic Cancer and Acute Leukemia	23	10
008 Lung, Upper Digestive Tract, and Other Sev. Cancers	12	7
009 Lymphatic, Head and Neck, Brain, Other Maj. Cancers	32	22
010 Breast, Prostate, Colorectal, Other Cancers/Tumors	110	73
011 Other Respiratory and Heart Neoplasms	5	3
012 Other Digestive and Urinary Neoplasms	88	65
ACC003: Benign/In Situ/Uncertain Neoplasm	890	581
013 Other Neoplasms	245	189
014 Benign Neoplasms of Skin, Breast, Eye	553	342
ACC004: Diabetes	267	249
015 Diabetes with Renal Manifestation	6	7
016 Diabetes with Neurologic or Periph. Circ. Manifest.	15	13
017 Diabetes with Acute Complications	10	10
018 Diabetes with Ophthalmologic Manifestation	21	22
019 Diabetes with No or Unspecified Complications	215	196
020 Type I Diabetes Mellitus	63	55
ACC005: Nutritional and Metabolic	1,013	1,098
021 Protein-Calorie Malnutrition	4	4
022 Other Significant Endocrine and Metabolic Disorders	32	32
023 Disorders of Fluid/Electrolyte/Acid-Base Balance	69	116
024 Other Endocrine/Metabolic/Nutritional Disorders	910	950
ACC006: Liver	91	79
025 End-Stage Liver Disease	3	3
026 Cirrhosis of Liver	4	4
027 Chronic Hepatitis	5	11
028 Acute Liver Failure/Disease	3	3
029 Other Hepatitis and Liver Disease	38	26

030 Gallbladder and Biliary Tract Disorders	42	36
ACC007: Gastrointestinal	958	1,287
031 Intestinal Obstruction/Perforation	24	16
032 Pancreatic Disease	17	16
033 Inflammatory Bowel Disease	28	21
034 Peptic Ulcer, Hemorrhage, Other Specified GI Dis.	117	97
035 Appendicitis	12	15
036 Other Gastrointestinal Disorders	771	1,132
ACC008: Musculoskeletal and Connective Tissues	1,939	2,285
037 Bone/Joint/Muscle Infections/Necrosis	15	17
038 Rheumatoid Arthritis, Inflammatory Connect. Tissue Dis.	75	64
039 Disorders of the Vertebrae and Spinal Discs	211	172
040 Osteoarthritis of Hip or Knee	66	58
041 Osteoporosis and Other Bone/Cartilage Disorders	160	150
042 Congenital/Developmental Skeletal, Connect. Tissue Dis.	8	8
043 Other Musculoskeletal and Connective Tissue Disorders	1,432	1,837
ACC009: Hematological	237	220
044 Severe Hematological Disorders	9	9
045 Disorders of Immunity	18	13
046 Coagulation defd, Other Specified Hematological Dis.	35	29
047 Iron Deficiency, Other/Unspecified Anemias, Blood Dis.	177	171
ACC010: Cognitive Disorders	41	37
048 Delirium and Encephalopathy	13	12
049 Dementia	11	12
050 Senility, Nonpsychotic Organic Brain Syndromes/Cond.	16	14
ACC011: Substance Abuse	39	206
051 Drug/Alcohol Psychosis	4	8
052 Drug/Alcohol Dependence	22	30
053 Drug/Alcohol Abuse, Without Dependence	13	169
ACC012: Mental	600	726
054 Schizophrenia	8	7
055 Major Depressive, Bipolar, and Paranoid Disorders	168	162
056 Reactive and Unspecified Psychosis	4	3
057 Personality Disorders	6	10
058 Depression	148	233
059 Anxiety Disorders	57	56
060 Other Psychiatric Disorders	209	255
ACC013: Developmental Disability	97	237
061 Profound Mental Retardation/Developmental Disability	<1	1
062 Severe Mental Retardation/Developmental Disability	<1	1
063 Moderate Mental Retardation/Developmental Disability	<1	1
064 Mild/Unspecified Mental Retardation/Develop. Disability	6	18
065 Other Developmental Disability	24	81
066 Attention Deficit Disorder	67	135
ACC014: Neurological	343	447
067 Quadriplegia, Other Extensive Paralysis	3	4
068 Paraplegia	2	3
069 Spinal Cord Disorders/Injuries	16	15

070 Muscular Dystrophy	1	1
071 Polyneuropathy	28	41
072 Multiple Sclerosis	14	11
073 Parkinson's and Huntington's Diseases	4	2
074 Seizure Disorders and Convulsions	49	53
075 Coma, Brain Compression/Anoxic Damage	3	2
076 Mononeuropathy, Other Neurological Conditions/Inj.	230	324
ACC015: Cardio-Respiratory Arrest	28	21
077 Respirator Dependence/Tracheostomy Status	1	2
078 Respiratory Arrest	2	1
079 Cardio-Respiratory Failure and Shock	25	18
ACC016: Heart	1,019	882
080 Congestive Heart Failure	54	37
081 Acute Myocardial Infarction	11	7
082 Unstable Angina, Other Acute Ischemic Heart Disease	38	20
083 Angina Pectoris/Old Myocardial Infarction	38	27
084 Coronary Atherosclerosis/Oth. Chron. Ischemic Heart Dis.	110	59
085 Heart Infection/Inflammation, Except Rheumatic	6	5
086 Valvular and Rheumatic Heart Disease	95	59
087 Major Congenital Cardiac/Circulatory Defect	3	3
088 Other Congenital Heart/Circulatory Disease	11	11
089 Hypertensive Heart and Renal Dis. or Encephalopathy	5	3
090 Hypertensive Heart Disease	36	14
091 Hypertension	673	709
092 Specified Heart Arrhythmias	54	37
093 Other Heart Rhythm and Conduction Disorders	76	38
094 Other and Unspecified Heart Disease	20	12
ACC017: Cerebro-Vascular	70	52
095 Cerebral Hemorrhage	7	3
096 Ischemic or Unspecified Stroke	19	13
097 Precerebral Arterial Occl., Transient Cerebral Ischemia	28	17
098 Cerebral Atherosclerosis and Aneurysm	4	2
099 Cerebrovascular Disease, Unspecified	1	2
100 Hemiplegia/Hemiparesis	4	5
101 Diplegia (Upper), Monoplegia, Other Paralytic Synd.	5	7
102 Speech, Language, Cognitive, Perceptual Deficits	2	4
103 Cerebrovascular Disease Late Effects, Unspecified	3	2
ACC018: Vascular	169	190
104 Vascular Disease with Complications	18	10
105 Vascular Disease	50	38
106 Other Circulatory Disease	102	142
ACC019: Lung	1,089	1,114
107 Cystic Fibrosis	2	2
108 Chronic Obstructive Pulmonary Disease	93	74
109 Fibrosis of Lung and Other Chronic Lung Disorders	40	25
110 Asthma	245	408
111 Aspiration and Specified Bacterial Pneumonias	5	6
112 Pneumococcal Pneumonia, Empyema, Lung Abscess	7	6

113 Viral and Unspecified Pneumonia, Pleurisy	144	160
114 Pleural Effusion/Pneumothorax	19	11
115 Other Lung Disorders	583	479
ACC020: Eyes	669	1,962
116 Legally Blind	1	1
117 Major Eye Infections/Inflammations	8	13
118 Retinal Detachment	7	6
119 Proliferative Diabetic Retinopathy, Vitreous Hemorrhage	9	6
120 Diabetic and Other Vascular Retinopathies	26	27
121 Retinal Dis., Exc. Detachment, Vascular Retinopathies	24	35
122 Glaucoma	84	88
123 Cataract	71	70
124 Other Eye Disorders	462	1,746
ACC021: Ears, Nose and Throat	2,722	3,767
125 Significant Ear, Nose, and Throat Disorders	26	31
126 Hearing Loss	66	142
127 Other Ear, Nose, Throat, and Mouth Disorders	2,634	3,601
ACC022: Urinary System	556	615
128 Kidney Transplant Status	4	2
129 End Stage Renal Disease	0	0
130 Dialysis Status	2	2
131 Renal Failure	13	13
132 Nephritis	6	9
133 Urinary Obstruction and Retention	94	82
134 Incontinence	34	64
135 Urinary Tract Infection	367	426
136 Other Urinary Tract Disorders	135	113
ACC023: Genital System	1,181	1,083
137 Female Infertility	27	52
138 Pelvic Inflammatory Dis., Oth. Spec. Female Genital Dis.	148	131
139 Other Female Genital Disorders	755	715
140 Male Genital Disorders	251	185
ACC024: Pregnancy Related	206	342
141 Ectopic Pregnancy	4	8
142 Miscarriage/Early Termination	26	30
143 Completed Pregnancy With Major Complications	14	25
144 Completed Pregnancy With Complications	68	131
145 Completed Pregnancy Without Comp. (Normal Delivery)	35	50
146 Uncompleted Pregnancy With Complications	11	17
147 Uncompleted Pregnancy With No or Minor Complications	52	89
ACC025: Skin and Subcutaneous	1,326	1,649
148 Decubitus Ulcer of Skin	4	3
149 Chronic Ulcer of Skin, Except Decubitus	14	8
150 Extensive Third-Degree Burns	<1	0
151 Other Third-Degree and Extensive Burns	1	2
152 Cellulitis, Local Skin Infection	181	233
153 Other Dermatological Disorders	1,128	1,405
ACC026: Injury, Poisoning, Complications	1,413	2,058

154 Severe Head Injury	1	1
155 Major Head Injury	29	25
156 Concussion or Unspecified Head Injury	8	47
157 Vertebral Fractures	7	6
158 Hip Fracture/Dislocation	11	11
159 Major Fracture, Except of Skull, Vertebrae, or Hip	53	68
160 Internal Injuries	8	6
161 Traumatic Amputation	2	2
162 Other Injuries	1,140	1,719
163 Poisonings and Allegic Reactions	158	217
164 Major Complications of Medical Care and Trauma	39	35
165 Other Complications of Medical Care	30	50
ACC027: Symptoms, Signs and Ill-Defined Conditions	2,468	3,818
166 Major Symptoms, Abnormalities	1,280	1,410
167 Minor Symptoms, Signs, Findings	1,188	2,408
ACC028: Neonates	67	6
168 Extremely Low Birthweight Neonates	1	0
169 Very Low Birthweight Neonates	<1	0
170 Serious Perinatal Problem Affecting Newborn	11	2
171 Other Perinatal Problems Affecting Newborn	27	3
172 Normal, Single Birth	29	0
ACC029: Transplants, Openings, Other V-Codes	13	14
173 Major Organ Transplant	0	0
174 Major Organ Transplant Status	2	2
175 Other Organ Transplant/Replacement	4	3
176 Artificial Openings for Feeding or Elimination	6	7
177 Amputation Status, Lower Limb/Amputation Comp.	1	1
178 Amputation Status, Upper Limb	<1	0
ACC030: Screening / History	2,866	6,206
179 Post-Surgical States/Aftercare/Elective	124	501
180 Radiation Therapy	5	28
181 Chemotherapy	11	11
182 Rehabilitation	36	197
183 Screening/Observation/Special Exams	2,726	6,029
184 History of Disease	169	248

DxCG, Inc.
Table 5.1: Comparison of Actual and Predicted Expenditures
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx
Weighted by Eligible Months

	Current Sample Total
Eligible Year Equivalents (Year 2)	2,304,926
Actual Expenditures (Year 2)	$1,794
Actual Expenditure Scores (Year 2) (Normalized to Current Sample)	1.00
Relative Risk Scores	**(Normalized to Current Sample)**
Age/Sex Model	1.00
Prospective Model (Year 2)	1.00
Risk Adjusted Expenditures	**(Actual Expenditures Divided by Relative Risk Scores)**
Age/Sex Model	$1,794
Prospective Model (Year 2)	$1,794
Predicted Expenditures	**(Relative Risk Scores * Current Sample Mean)**
Age/Sex Model	$1,794
Prospective Model (Year 2)	$1,794
Efficiency Index	**(Actual Expenditures Divided by Predicted Expenditures - Observed / Expected)**
Age/Sex Model	1.00
Prospective Model (Year 2)	1.00

DxCG, Inc.
Table 6.1: Distribution by Aggregated DCG (ADCG) Category
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Benchmark		Current Sample			
	People	Dollars	People		Predicted Dollars	
ADCG Predicted Expenditure Range (in order of increasing severity)	Percent	Percent	Number	Percent	Mean	Percent
Total	100.00	100	2,304,926	100.00	$1,794	100.00
ADCG 0	42.07	10	984,419	42.71	$427	10.16
ADCG 1	48.44	49	1,091,667	47.36	$1,821	48.07
ADCG 5	7.26	22	175,804	7.63	$5,454	23.19
ADCG 10	1.88	12	45,210	1.96	$11,469	12.54
ADCG 25	0.35	6	7,826	0.34	$31,849	6.03

DxCG, Inc.
Table 6.2: Distribution by DCG Category
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Benchmark		Current Sample			
	People	Dollars	People		Predicted Dollars	
DCG Predicted Expenditure Range (in order of increasing severity)	Percent	Percent	Number	Percent	Mean	Percent
Total	100.00	100.00	2,304,926	100.00	$1,794	100.00
DCG 0.0	0.00	0.00	0	0.00	$0	0.00
DCG 0.1	7.41	0.62	150,176	6.52	$149	0.54
DCG 0.2	0.59	0.06	20,800	0.90	$199	0.10
DCG 0.3	8.17	1.24	175,225	7.60	$271	1.15
DCG 0.4	8.67	1.87	187,020	8.11	$379	1.71
DCG 0.5	5.50	1.52	176,571	7.66	$490	2.09
DCG 0.7	11.73	4.69	274,627	11.91	$691	4.59
DCG 1.0	15.62	8.67	349,492	15.16	$990	8.37
DCG 1.5	9.89	7.88	217,212	9.42	$1,406	7.39
DCG 2.0	6.58	6.71	152,265	6.61	$1,812	6.67
DCG 2.5	5.04	6.30	114,932	4.99	$2,225	6.18
DCG 3.0	7.00	11.05	158,836	6.89	$2,808	10.78
DCG 4.0	4.31	8.77	98,930	4.29	$3,626	8.67
DCG 5.0	2.87	7.15	67,982	2.95	$4,442	7.30
DCG 6.0	2.48	7.53	60,285	2.62	$5,412	7.89

DCG 7.5	1.91	7.45	47,537	2.06	$6,955	8.00
DCG 10.0	1.24	6.76	30,731	1.33	$9,677	7.19
DCG 15.0	0.41	3.21	9,788	0.42	$13,941	3.30
DCG 20.0	0.23	2.34	4,691	0.20	$18,051	2.05
DCG 25.0	0.12	1.53	2,752	0.12	$22,077	1.47
DCG 30.0	0.11	1.77	2,363	0.10	$27,907	1.59
DCG 40.0	0.06	1.15	1,259	0.05	$35,971	1.10
DCG 50.0	0.03	0.75	650	0.03	$44,140	0.69
DCG 60.0	0.01	0.43	381	0.02	$52,173	0.48
DCG 70.0	0.01	0.53	421	0.02	$68,161	0.69

DxCG, Inc.
Table 6.3: Percent Distribution by Aggregated DCG (ADCG) Category
By Total
Model Dimensions: All-encounter/Prospective/Comm/Expl/Med+Rx

	Benchmark	Current Sample
ADCG Predicted Expenditure Range (in order of increasing severity)	**Percent**	**Total**
Total	100.00	100.00
ADCG 0	42.07	42.71
ADCG 1	48.44	47.36
ADCG 5	7.26	7.63
ADCG 10	1.88	1.96
ADCG 25	0.35	0.34

DxCG Data Appendix
Table A-1: Summary of Input and Output Files

Date:	Tue, Jan 25, 2005
Time:	13:18:32
Licensed to:	Arlene Ash - DOD Tricare Prime
DxCG Release:	6.1
Operating environment:	SAS 8.2 WIN
DxCG software serial number:	12312005
Software expiration date:	31-Dec-05
Maximum licensed population size:	2,300,000

INPUT FILES:

Diagnosis file:	INDIAG.diag
Person file:	INPERS.pers

OUTPUT FILES:

Main DxCG output file:	OUT.dod_dxcg
DxCG summary file:	OUT.dod_dxcgS
Bad diagnosis file:	Not requested
Bad enrollment file:	Not requested
Appendix file:	APPENDIX

OPTIONS:

Population group:	Commercial
Model variant:	All-encounter, DCG/HCC
Model purpose:	Explanation
Model outcome:	Medical expenses including pharmacy spending
Maximum number of observations processed:	MAX
Level of detail output:	ACCs, CCs and DxGroups
Impose hierarchies:	Yes
Location of expenditure information:	None provided
Include over age 64:	No
Handle bad diagnoses:	Use any 2000-2002 legal codes
Handle bad enrollment:	Do not output invalid enrollments
Is source variable provided?	No
Are procedure codes provided?	No
Date for calculating age:	Not used

Maximum number of diagnoses per record:	6
Maximum number of procedures per record:	0
Rename input diagnosis file variables:	None
Rename input person file variables:	None

DxCG Data Appendix
Table A-2: Counts from Input Files

COUNTS FROM DIAGNOSIS FILE

IDs in diagnosis file:	1,976,274
IDs with no match in person file:	0
Number of records in diagnosis file:	5,343,739
Number of records with no match in person file:	0

COUNTS FROM DIAGNOSES MATCHED WITH PERSON FILE

Number of IDs:	1,976,274
Number of IDs of people who were ever in the hospital:	0
Number of records:	5,343,739
Number of diagnoses with allowed source:	26,570,821
Number of diagnoses with unallowed source:	0
Number of numerically invalid diagnoses:	358,975
Number of diagnoses incompatible with age or sex:	25,516
Number of principal inpatient diagnoses:	0
Number of procedure codes:	0
Average number of diagnosis records with allowed source per individual:	2.704
Average number of diagnoses with allowed source per individual:	13.445
Percent of diagnoses that are invalid or incompatible with age or sex:	1.447

COUNTS FROM PERSON FILE

Total IDs in either input file:	2,304,926
IDs in person file:	2,304,926
IDs in diagnosis file but not in person file:	0
IDs with invalid age:	0
IDs with invalid sex:	0
IDs with age < -1:	0

IDs with age 65 and over:	0
IDs in the person file not output to the DxCG file:	0
Total IDs output in DxCG file:	2,304,926

DxCG Data Appendix
Table A-3: Counts and Averages from the DxCG Output Files

COUNTS FROM DxCG OUTPUT FILE:

Total IDs output in DxCG file:	2,304,926
IDs eligible at all in year 1:	2,304,926
IDs eligible at all in year 2:	2,304,926
Eligible-year equivalents, year 1:	2,304,926
Eligible-year equivalents, year 2:	2,304,926
Average number of months eligible, year 1:	12
Average number of months eligible, year 2:	12
Number of IDs with at least one diagnosis:	1,976,274
IDs with no diagnosis records:	328,652
Number of people with:	
Year 1 expenditures >= 0:	0
Year 1 expenditures > 0:	0
Year 1 expenditures = 0:	0
Year 1 expenditures < 0:	0
Year 1 expenditures missing:	2,304,926
Year 2 expenditures >= 0:	2,304,926
Year 2 expenditures > 0:	1,964,740
Year 2 expenditures = 0:	340,186
Year 2 expenditures < 0:	0
Year 2 expenditures missing:	0
Number of IDs with any diagnosis assigned to an HCC:	1,975,408
Number of people with any hospitalization in DxCG file:	0
Number of hospitalizations for people in DxCG file:	0
Percent of individuals with any diagnosis:	85.74
Percent of individuals with a hospitalization:	0
Percent with positive health expenditures in year 1:	0
Percent with positive health expenditures in year 2:	85.24
Average year 1 expenditures per person:	0
Average year 2 expenditures per person:	1,794
Average year 1 expenditures per eligible year:	0
Average year 2 expenditures per eligible year:	1,794
Average number of year 1 hospitalizations when hosp > 0:	0

Sum of actual year 1 expenditures:	0
Sum of actual year 2 expenditures:	4,134,695,138
Sum of negative year 1 expenditures:	0
Sum of negative year 2 expenditures:	0

DxCG Data Appendix
Table A-4: Average Scores

AVERAGE RISK SCORES, WEIGHTED BY ELIGIBILITY:

SCORE02C Average risk, age/sex, year 2, Commercial:	0.808
SCORE31C Average risk, all-encounter, year 1, Commercial:	1.119
SCORE32C Average risk, all-encounter, year 2, Commercial:	1.012

AVERAGE RISK SCORES, NOT WEIGHTED BY ELIGIBILITY:

USCOR02C Unweighted risk, age/sex, year 2, Commercial:	0.808
USCOR31C Unweighted risk, all-encounter, year 1, Commercial:	1.119
USCOR32C Unweighted risk, all-encounter, year 2, Commercial:	1.012

Attachment A4_CRG

PCRG	PCRG_Description
10000	Healthy
10010	Healthy Non-User
10020	Delivery without Other Significant Illness
10040	Pregnancy without Delivery without Other Significant Illness
10050	Major Gynecological Diagnosis without Other Significant Illness
10060	Diagnosis of Major Congenital or Neonatal Problems without Other Significant Illness
10090	Catastrophic Diagnosis without Other Significant Illness
10100	Malignancy Diagnosis without Other Significant Illness
10110	Significant Neurological Diagnosis without Other Significant Illness
10120	Significant Cardiovascular, Pulmonary or other Vascular Diagnosis without Other Significant Illness
10130	Major Mental Illness or Substance Abuse Diagnosis without Other Significant Illness
10140	Significant Connective Tissue or Orthopedic Diagnosis without Other Significant Illness
10150	Significant Gastrointestinal, Hepatic, Renal or Hernia Diagnosis without Other Significant Illness
10160	Diabetic Diagnosis without Other Significant Illness
20100	2 or More Significant Acute Illnesses from Different MDCs Excluding ENT
20200	1 Significant Acute Illness - Span 90 Excluding ENT
20300	1 Significant Acute Illness Excluding ENT
20400	1 Significant Acute ENT Illness - Span 90
20500	1 Significant Acute ENT Illness
20600	1 Significant Acute Procedure
20720	Delivery with Other Significant Illness
20740	Pregnancy without Delivery with Other Significant Illness
20750	Major Gynecological Diagnosis with Other Significant Illness
20760	Diagnosis of Major Congenital or Neonatal Problems with Other Significant Illness

PCRG	PCRG_Description
20790	Catastrophic Diagnosis with Other Significant Illness
20800	Malignancy Diagnosis with Other Significant Illness
20810	Significant Neurological Diagnosis with Other Significant Illness
20820	Significant Cardiovascular, Pulmonary or other Vascular Diagnosis with Other Significant Illness
20830	Major Mental Illness or Substance Abuse Diagnosis with Other Significant Illness
20840	Significant Connective Tissue or Orthopedic Diagnosis with Other Significant Illness
20850	Significant Gastrointestinal, Hepatic, Renal or Hernia Diagnosis with Other Significant Illness
20860	Diabetic Diagnosis with Other Significant Illness
30171	Gait Abnormalities Level - 1
30172	Gait Abnormalities Level - 2
30181	Migraine Level - 1
30182	Migraine Level - 2
30191	Chronic Neuromuscular and Other Neurological Diagnoses - Minor Level - 1
30192	Chronic Neuromuscular and Other Neurological Diagnoses - Minor Level - 2
30781	Glaucoma Level - 1
30782	Glaucoma Level - 2
30791	Cataracts Level - 1
30792	Cataracts Level - 2
30801	Chronic Eye Diagnoses - Minor Level - 1
30802	Chronic Eye Diagnoses - Minor Level - 2
30991	Chronic Ear Diagnoses except Hearing Loss Level - 1
30992	Chronic Ear Diagnoses except Hearing Loss Level - 2
31001	Chronic Hearing Loss Level - 1
31002	Chronic Hearing Loss Level - 2
31011	Other Chronic Ear, Nose, and Throat Diagnoses Level - 1
31012	Other Chronic Ear, Nose, and Throat Diagnoses Level - 2

PCRG	PCRG_Description
31411	Chronic Bronchitis Level - 1
31412	Chronic Bronchitis Level - 2
31421	Other Chronic Pulmonary Diagnoses Level - 1
31422	Other Chronic Pulmonary Diagnoses Level - 2
31951	Ventricular and Atrial Septal Defects Level - 1
31952	Ventricular and Atrial Septal Defects Level - 2
31961	Chronic Cardiovascular Diagnoses - Minor Level - 1
31962	Chronic Cardiovascular Diagnoses - Minor Level - 2
32441	Chronic Disorders of Arteries and Veins - Minor Level - 1
32442	Chronic Disorders of Arteries and Veins - Minor Level - 2
32731	Chronic Ulcers Level - 1
32732	Chronic Ulcers Level - 2
32741	Chronic Gastrointestinal Diagnoses - Minor Level - 1
32742	Chronic Gastrointestinal Diagnoses - Minor Level - 2
33171	Gallbladder Disease Level - 1
33172	Gallbladder Disease Level - 2
33571	Osteoarthritis Level - 1
33572	Osteoarthritis Level - 2
33581	Chronic Joint and Musculoskeletal Diagnoses - Minor Level - 1
33582	Chronic Joint and Musculoskeletal Diagnoses - Minor Level - 2
34081	Skin Malignancy Level - 1
34082	Skin Malignancy Level - 2
34091	Psoriasis Level - 1
34092	Psoriasis Level - 2
34101	Chronic Skin Diagnoses - Minor Level - 1
34102	Chronic Skin Diagnoses - Minor Level - 2

PCRG	PCRG_Description
34451	Hyperlipidemia Level - 1
34452	Hyperlipidemia Level - 2
34461	Chronic Thyroid Disease Level - 1
34462	Chronic Thyroid Disease Level - 2
34821	Vesicoureteral Reflux Level - 1
34822	Vesicoureteral Reflux Level - 2
34831	Recurrent Urinary Tract Infections Level - 1
34832	Recurrent Urinary Tract Infections Level - 2
35101	Benign Prostatic Hyperplasia Level - 1
35102	Benign Prostatic Hyperplasia Level - 2
35111	Prostate Disease, Infertility,and Benign Neoplasms - Male Level - 1
35112	Prostate Disease, Infertility,and Benign Neoplasms - Male Level - 2
35231	Chronic Pelvic Inflammatory Disease Level - 1
35232	Chronic Pelvic Inflammatory Disease Level - 2
35241	Infertility - Female Level - 1
35242	Infertility - Female Level - 2
35251	Other Chronic Gynecological Diagnoses - Minor Level - 1
35252	Other Chronic Gynecological Diagnoses - Minor Level - 2
35571	Prematurity - Birthweight < 1000 Grams Level - 1
35572	Prematurity - Birthweight < 1000 Grams Level - 2
35921	Developmental Language Disorder Level - 1
35922	Developmental Language Disorder Level - 2
36131	Chronic Hematological Diagnoses - Minor Level - 1
36132	Chronic Hematological Diagnoses - Minor Level - 2
36981	Neoplasm of Uncertain Behavior Level - 1
36982	Neoplasm of Uncertain Behavior Level - 2

PCRG	PCRG_Description
37541	Attention Deficit / Hyperactivity Disorder Level - 1
37542	Attention Deficit / Hyperactivity Disorder Level - 2
37551	Depression Level - 1
37552	Depression Level - 2
37561	Chronic Mental Health Diagnoses - Minor Level - 1
37562	Chronic Mental Health Diagnoses - Minor Level - 2
37571	Chronic Stress and Anxiety Diagnoses Level - 1
37572	Chronic Stress and Anxiety Diagnoses Level - 2
37891	Drug Abuse Related Diagnoses Level - 1
37892	Drug Abuse Related Diagnoses Level - 2
40001	Multiple Minor Chronic PCDs Level - 1
40002	Multiple Minor Chronic PCDs Level - 2
40003	Multiple Minor Chronic PCDs Level - 3
40004	Multiple Minor Chronic PCDs Level - 4
50011	Progressive Neurological Diagnoses Level - 1
50012	Progressive Neurological Diagnoses Level - 2
50013	Progressive Neurological Diagnoses Level - 3
50014	Progressive Neurological Diagnoses Level - 4
50021	Extrapyramidal Diagnoses Level - 1
50022	Extrapyramidal Diagnoses Level - 2
50023	Extrapyramidal Diagnoses Level - 3
50024	Extrapyramidal Diagnoses Level - 4
50031	Acquired Hemiplegia Level - 1
50032	Acquired Hemiplegia Level - 2
50033	Acquired Hemiplegia Level - 3
50034	Acquired Hemiplegia Level - 4

PCRG	PCRG_Description
50041	Cerebrovascular Disease with Infarction or Intracranial Hemorrhage Level - 1
50042	Cerebrovascular Disease with Infarction or Intracranial Hemorrhage Level - 2
50043	Cerebrovascular Disease with Infarction or Intracranial Hemorrhage Level - 3
50044	Cerebrovascular Disease with Infarction or Intracranial Hemorrhage Level - 4
50051	Neurodegenerative Diagnoses Except Multiple Sclerosis and Parkinson's Level - 1
50052	Neurodegenerative Diagnoses Except Multiple Sclerosis and Parkinson's Level - 2
50053	Neurodegenerative Diagnoses Except Multiple Sclerosis and Parkinson's Level - 3
50054	Neurodegenerative Diagnoses Except Multiple Sclerosis and Parkinson's Level - 4
50061	Alzheimer's Disease and Other Dementias Level - 1
50062	Alzheimer's Disease and Other Dementias Level - 2
50063	Alzheimer's Disease and Other Dementias Level - 3
50064	Alzheimer's Disease and Other Dementias Level - 4
50071	Cerebral Palsy NOS Level - 1
50072	Cerebral Palsy NOS Level - 2
50073	Cerebral Palsy NOS Level - 3
50074	Cerebral Palsy NOS Level - 4
50101	Hydrocephalus and Other Brain Anomalies Level - 1
50102	Hydrocephalus and Other Brain Anomalies Level - 2
50103	Hydrocephalus and Other Brain Anomalies Level - 3
50104	Hydrocephalus and Other Brain Anomalies Level - 4
50111	Chronic Neuromuscular/Other Neurological Diagnoses - Moderate Level - 1
50112	Chronic Neuromuscular/Other Neurological Diagnoses - Moderate Level - 2
50113	Chronic Neuromuscular/Other Neurological Diagnoses - Moderate Level - 3
50114	Chronic Neuromuscular/Other Neurological Diagnoses - Moderate Level - 4
50121	History of Transient Ischemic Attack Level - 1
50122	History of Transient Ischemic Attack Level - 2

PCRG	PCRG_Description
50123	History of Transient Ischemic Attack Level - 3
50124	History of Transient Ischemic Attack Level - 4
50131	Cerebrovascular Disease without Infarction Level - 1
50132	Cerebrovascular Disease without Infarction Level - 2
50133	Cerebrovascular Disease without Infarction Level - 3
50134	Cerebrovascular Disease without Infarction Level - 4
50141	Epilepsy Level - 1
50142	Epilepsy Level - 2
50143	Epilepsy Level - 3
50144	Epilepsy Level - 4
50741	Macular Degeneration Level - 1
50742	Macular Degeneration Level - 2
50743	Macular Degeneration Level - 3
50744	Macular Degeneration Level - 4
50751	Blindness, Visual Loss, and Chronic Eye Diagnoses - Major / Moderate Level - 1
50752	Blindness, Visual Loss, and Chronic Eye Diagnoses - Major / Moderate Level - 2
50753	Blindness, Visual Loss, and Chronic Eye Diagnoses - Major / Moderate Level - 3
50754	Blindness, Visual Loss, and Chronic Eye Diagnoses - Major / Moderate Level - 4
51171	Anomaly Skull and Facial Bones Level - 1
51172	Anomaly Skull and Facial Bones Level - 2
51173	Anomaly Skull and Facial Bones Level - 3
51174	Anomaly Skull and Facial Bones Level - 4
51201	Cleft Lip and Palate Level - 1
51202	Cleft Lip and Palate Level - 2
51203	Cleft Lip and Palate Level - 3
51204	Cleft Lip and Palate Level - 4

PCRG	PCRG_Description
51321	Major Respiratory Anomalies Level - 1
51322	Major Respiratory Anomalies Level - 2
51323	Major Respiratory Anomalies Level - 3
51324	Major Respiratory Anomalies Level - 4
51331	Chronic Obstructive Pulmonary Disease and Bronchiectasis Level - 1
51332	Chronic Obstructive Pulmonary Disease and Bronchiectasis Level - 2
51333	Chronic Obstructive Pulmonary Disease and Bronchiectasis Level - 3
51334	Chronic Obstructive Pulmonary Disease and Bronchiectasis Level - 4
51341	Other Major Chronic Pulmonary Diagnoses Level - 1
51342	Other Major Chronic Pulmonary Diagnoses Level - 2
51343	Other Major Chronic Pulmonary Diagnoses Level - 3
51344	Other Major Chronic Pulmonary Diagnoses Level - 4
51351	Tracheostomy Status Level - 1
51352	Tracheostomy Status Level - 2
51353	Tracheostomy Status Level - 3
51354	Tracheostomy Status Level - 4
51381	Asthma Level - 1
51382	Asthma Level - 2
51383	Asthma Level - 3
51384	Asthma Level - 4
51771	Complex Cyanotic and Major Cardiac Septal Anomalies Level - 1
51772	Complex Cyanotic and Major Cardiac Septal Anomalies Level - 2
51773	Complex Cyanotic and Major Cardiac Septal Anomalies Level - 3
51774	Complex Cyanotic and Major Cardiac Septal Anomalies Level - 4
51781	Other Major Congenital Heart Diagnoses Except Valvular Level - 1
51782	Other Major Congenital Heart Diagnoses Except Valvular Level - 2

PCRG	PCRG_Description
51783	Other Major Congenital Heart Diagnoses Except Valvular Level - 3
51784	Other Major Congenital Heart Diagnoses Except Valvular Level - 4
51791	Congestive Heart Failure Level - 1
51792	Congestive Heart Failure Level - 2
51793	Congestive Heart Failure Level - 3
51794	Congestive Heart Failure Level - 4
51801	Other Cardiovascular Diagnoses - Major Level - 1
51802	Other Cardiovascular Diagnoses - Major Level - 2
51803	Other Cardiovascular Diagnoses - Major Level - 3
51804	Other Cardiovascular Diagnoses - Major Level - 4
51811	Valvular Disorders Level - 1
51812	Valvular Disorders Level - 2
51813	Valvular Disorders Level - 3
51814	Valvular Disorders Level - 4
51821	History of Myocardial Infarction Level - 1
51822	History of Myocardial Infarction Level - 2
51823	History of Myocardial Infarction Level - 3
51824	History of Myocardial Infarction Level - 4
51831	Angina and Ischemic Heart Disease Level - 1
51832	Angina and Ischemic Heart Disease Level - 2
51833	Angina and Ischemic Heart Disease Level - 3
51834	Angina and Ischemic Heart Disease Level - 4
51861	Atrial Fibrillation Level - 1
51862	Atrial Fibrillation Level - 2
51863	Atrial Fibrillation Level - 3
51864	Atrial Fibrillation Level - 4

PCRG	PCRG_Description
51871	Cardiac Dysrhythmia and Conduction Disorders Level - 1
51872	Cardiac Dysrhythmia and Conduction Disorders Level - 2
51873	Cardiac Dysrhythmia and Conduction Disorders Level - 3
51874	Cardiac Dysrhythmia and Conduction Disorders Level - 4
51881	History of Coronary Artery Bypass Graft Level - 1
51882	History of Coronary Artery Bypass Graft Level - 2
51883	History of Coronary Artery Bypass Graft Level - 3
51884	History of Coronary Artery Bypass Graft Level - 4
51891	History of Percutaneous Transluminal Coronary Angioplasty Level - 1
51892	History of Percutaneous Transluminal Coronary Angioplasty Level - 2
51893	History of Percutaneous Transluminal Coronary Angioplasty Level - 3
51894	History of Percutaneous Transluminal Coronary Angioplasty Level - 4
51901	Cardiac Device Status Level - 1
51902	Cardiac Device Status Level - 2
51903	Cardiac Device Status Level - 3
51904	Cardiac Device Status Level - 4
51911	Coronary Atherosclerosis Level - 1
51912	Coronary Atherosclerosis Level - 2
51913	Coronary Atherosclerosis Level - 3
51914	Coronary Atherosclerosis Level - 4
51921	Hypertension Level - 1
51922	Hypertension Level - 2
51923	Hypertension Level - 3
51924	Hypertension Level - 4
52371	Chronic Disorders of Arteries and Veins - Major Level - 1
52372	Chronic Disorders of Arteries and Veins - Major Level - 2

PCRG	PCRG_Description
52373	Chronic Disorders of Arteries and Veins - Major Level - 3
52374	Chronic Disorders of Arteries and Veins - Major Level - 4
52381	Peripheral Vascular Disease Level - 1
52382	Peripheral Vascular Disease Level - 2
52383	Peripheral Vascular Disease Level - 3
52384	Peripheral Vascular Disease Level - 4
52411	Leg Varicosities with Ulcers or Inflammation Level - 1
52412	Leg Varicosities with Ulcers or Inflammation Level - 2
52413	Leg Varicosities with Ulcers or Inflammation Level - 3
52414	Leg Varicosities with Ulcers or Inflammation Level - 4
52661	Inflammatory Bowel Disease Level - 1
52662	Inflammatory Bowel Disease Level - 2
52663	Inflammatory Bowel Disease Level - 3
52664	Inflammatory Bowel Disease Level - 4
52691	Gastrointestinal Anomalies Level - 1
52692	Gastrointestinal Anomalies Level - 2
52693	Gastrointestinal Anomalies Level - 3
52694	Gastrointestinal Anomalies Level - 4
52701	Chronic Gastrointestinal Diagnoses - Moderate Level - 1
52702	Chronic Gastrointestinal Diagnoses - Moderate Level - 2
52703	Chronic Gastrointestinal Diagnoses - Moderate Level - 3
52704	Chronic Gastrointestinal Diagnoses - Moderate Level - 4
53101	Alcoholic Liver Disease Level - 1
53102	Alcoholic Liver Disease Level - 2
53103	Alcoholic Liver Disease Level - 3
53104	Alcoholic Liver Disease Level - 4

PCRG	PCRG_Description
53111	Major Liver Disease except Alcoholic Level - 1
53112	Major Liver Disease except Alcoholic Level - 2
53113	Major Liver Disease except Alcoholic Level - 3
53114	Major Liver Disease except Alcoholic Level - 4
53141	Chronic Pancreatic and Liver Disorders - Moderate Level - 1
53142	Chronic Pancreatic and Liver Disorders - Moderate Level - 2
53143	Chronic Pancreatic and Liver Disorders - Moderate Level - 3
53144	Chronic Pancreatic and Liver Disorders - Moderate Level - 4
53431	Major Congenital Bone, Cartilage, and Muscle Diagnoses Level - 1
53432	Major Congenital Bone, Cartilage, and Muscle Diagnoses Level - 2
53433	Major Congenital Bone, Cartilage, and Muscle Diagnoses Level - 3
53434	Major Congenital Bone, Cartilage, and Muscle Diagnoses Level - 4
53441	History of Hip Fracture Age > 64 Years Level - 1
53442	History of Hip Fracture Age > 64 Years Level - 2
53443	History of Hip Fracture Age > 64 Years Level - 3
53444	History of Hip Fracture Age > 64 Years Level - 4
53451	Spinal Stenosis Level - 1
53452	Spinal Stenosis Level - 2
53453	Spinal Stenosis Level - 3
53454	Spinal Stenosis Level - 4
53481	Curvature or Anomaly of the Spine Level - 1
53482	Curvature or Anomaly of the Spine Level - 2
53483	Curvature or Anomaly of the Spine Level - 3
53484	Curvature or Anomaly of the Spine Level - 4
53491	Pelvis, Hip, and Femur Deformities Level - 1
53492	Pelvis, Hip, and Femur Deformities Level - 2

PCRG	PCRG_Description
53493	Pelvis, Hip, and Femur Deformities Level - 3
53494	Pelvis, Hip, and Femur Deformities Level - 4
53501	Amputation and Bone Disease Level - 1
53502	Amputation and Bone Disease Level - 2
53503	Amputation and Bone Disease Level - 3
53504	Amputation and Bone Disease Level - 4
53511	Disc Disease and Other Chronic Back Diagnoses Level - 1
53512	Disc Disease and Other Chronic Back Diagnoses Level - 2
53513	Disc Disease and Other Chronic Back Diagnoses Level - 3
53514	Disc Disease and Other Chronic Back Diagnoses Level - 4
53521	Crystal Arthropathy Level - 1
53522	Crystal Arthropathy Level - 2
53523	Crystal Arthropathy Level - 3
53524	Crystal Arthropathy Level - 4
53531	Joint Replacement Level - 1
53532	Joint Replacement Level - 2
53533	Joint Replacement Level - 3
53534	Joint Replacement Level - 4
53541	Osteoporosis Level - 1
53542	Osteoporosis Level - 2
53543	Osteoporosis Level - 3
53544	Osteoporosis Level - 4
53901	Connective Tissue Disease and Vasculitis Level - 1
53902	Connective Tissue Disease and Vasculitis Level - 2
53903	Connective Tissue Disease and Vasculitis Level - 3
53904	Connective Tissue Disease and Vasculitis Level - 4

PCRG	PCRG_Description
53911	Rheumatoid Arthritis Level - 1
53912	Rheumatoid Arthritis Level - 2
53913	Rheumatoid Arthritis Level - 3
53914	Rheumatoid Arthritis Level - 4
53941	Spondyloarthropathy and Other Inflammatory Arthropathies Level - 1
53942	Spondyloarthropathy and Other Inflammatory Arthropathies Level - 2
53943	Spondyloarthropathy and Other Inflammatory Arthropathies Level - 3
53944	Spondyloarthropathy and Other Inflammatory Arthropathies Level - 4
54041	Chronic Skin Ulcer Level - 1
54042	Chronic Skin Ulcer Level - 2
54043	Chronic Skin Ulcer Level - 3
54044	Chronic Skin Ulcer Level - 4
54051	Significant Skin and Subcutaneous Tissue Diagnoses Level - 1
54052	Significant Skin and Subcutaneous Tissue Diagnoses Level - 2
54053	Significant Skin and Subcutaneous Tissue Diagnoses Level - 3
54054	Significant Skin and Subcutaneous Tissue Diagnoses Level - 4
54241	Diabetes Level - 1
54242	Diabetes Level - 2
54243	Diabetes Level - 3
54244	Diabetes Level - 4
54391	Chronic Metabolic and Endocrine Diagnoses - Major Level - 1
54392	Chronic Metabolic and Endocrine Diagnoses - Major Level - 2
54393	Chronic Metabolic and Endocrine Diagnoses - Major Level - 3
54394	Chronic Metabolic and Endocrine Diagnoses - Major Level - 4
54421	Chronic Endocrine, Nutritional, Fluid, Electrolyte and Immune Diagnoses - Moderate Level - 1
54422	Chronic Endocrine, Nutritional, Fluid, Electrolyte and Immune Diagnoses - Moderate Level - 2

PCRG	PCRG_Description
54423	Chronic Endocrine, Nutritional, Fluid, Electrolyte and Immune Diagnoses - Moderate Level - 3
54424	Chronic Endocrine, Nutritional, Fluid, Electrolyte and Immune Diagnoses - Moderate Level - 4
54731	Chronic Renal Failure Level - 1
54732	Chronic Renal Failure Level - 2
54733	Chronic Renal Failure Level - 3
54734	Chronic Renal Failure Level - 4
54741	Kidney Transplant Status Level - 1
54742	Kidney Transplant Status Level - 2
54743	Kidney Transplant Status Level - 3
54744	Kidney Transplant Status Level - 4
54771	Nephritis Level - 1
54772	Nephritis Level - 2
54773	Nephritis Level - 3
54774	Nephritis Level - 4
54781	Anomalies of Kidney or Urinary Tract Level - 1
54782	Anomalies of Kidney or Urinary Tract Level - 2
54783	Anomalies of Kidney or Urinary Tract Level - 3
54784	Anomalies of Kidney or Urinary Tract Level - 4
54791	Chronic Genitourinary Diagnoses Level - 1
54792	Chronic Genitourinary Diagnoses Level - 2
54793	Chronic Genitourinary Diagnoses Level - 3
54794	Chronic Genitourinary Diagnoses Level - 4
55821	Down's Syndrome Level - 1
55822	Down's Syndrome Level - 2
55823	Down's Syndrome Level - 3
55824	Down's Syndrome Level - 4

PCRG	PCRG_Description
55831	Chromosomal Anomalies and Syndromes Except Down's Level - 1
55832	Chromosomal Anomalies and Syndromes Except Down's Level - 2
55833	Chromosomal Anomalies and Syndromes Except Down's Level - 3
55834	Chromosomal Anomalies and Syndromes Except Down's Level - 4
55841	Severe / Profound Mental Retardation Level - 1
55842	Severe / Profound Mental Retardation Level - 2
55843	Severe / Profound Mental Retardation Level - 3
55844	Severe / Profound Mental Retardation Level - 4
55851	Pervasive Development Disorder Level - 1
55852	Pervasive Development Disorder Level - 2
55853	Pervasive Development Disorder Level - 3
55854	Pervasive Development Disorder Level - 4
55881	Mild / Moderate Mental Retardation Level - 1
55882	Mild / Moderate Mental Retardation Level - 2
55883	Mild / Moderate Mental Retardation Level - 3
55884	Mild / Moderate Mental Retardation Level - 4
55891	Developmental Delay NOS / NEC / Mixed Level - 1
55892	Developmental Delay NOS / NEC / Mixed Level - 2
55893	Developmental Delay NOS / NEC / Mixed Level - 3
55894	Developmental Delay NOS / NEC / Mixed Level - 4
56051	Immune and Leukocyte Disorders Level - 1
56052	Immune and Leukocyte Disorders Level - 2
56053	Immune and Leukocyte Disorders Level - 3
56054	Immune and Leukocyte Disorders Level - 4
56061	Sickle Cell Anemia Level - 1
56062	Sickle Cell Anemia Level - 2

PCRG	PCRG_Description
56063	Sickle Cell Anemia Level - 3
56064	Sickle Cell Anemia Level - 4
56071	Coagulation Disorders Level - 1
56072	Coagulation Disorders Level - 2
56073	Coagulation Disorders Level - 3
56074	Coagulation Disorders Level - 4
56101	Chronic Hematological and Immune Diagnoses - Moderate Level - 1
56102	Chronic Hematological and Immune Diagnoses - Moderate Level - 2
56103	Chronic Hematological and Immune Diagnoses - Moderate Level - 3
56104	Chronic Hematological and Immune Diagnoses - Moderate Level - 4
56521	Chronic Lymphoid Leukemia Level - 1
56522	Chronic Lymphoid Leukemia Level - 2
56531	Chronic Non-Lymphoid Leukemia Level - 1
56532	Chronic Non-Lymphoid Leukemia Level - 2
56541	Multiple Myeloma Level - 1
56542	Multiple Myeloma Level - 2
56551	Acute Lymphoid Leukemia Level - 1
56552	Acute Lymphoid Leukemia Level - 2
56561	Acute Non-Lymphoid Leukemia Level - 1
56562	Acute Non-Lymphoid Leukemia Level - 2
56571	Colon Malignancy Level - 1
56572	Colon Malignancy Level - 2
56581	Other Malignancies Level - 1
56582	Other Malignancies Level - 2
56601	Hodgkin's Lymphoma Level - 1
56602	Hodgkin's Lymphoma Level - 2

PCRG	PCRG_Description
56611	Plasma Protein Malignancy Level - 1
56612	Plasma Protein Malignancy Level - 2
56621	Breast Malignancy Level - 1
56622	Breast Malignancy Level - 2
56631	Prostate Malignancy Level - 1
56632	Prostate Malignancy Level - 2
56641	Genitourinary Malignancy Level - 1
56642	Genitourinary Malignancy Level - 2
56651	Non-Hodgkin's Lymphoma Level - 1
56652	Non-Hodgkin's Lymphoma Level - 2
56951	Malignancy NOS/NEC Level - 1
56952	Malignancy NOS/NEC Level - 2
57061	Chronic Infections Except Tuberculosis Level - 1
57062	Chronic Infections Except Tuberculosis Level - 2
57063	Chronic Infections Except Tuberculosis Level - 3
57064	Chronic Infections Except Tuberculosis Level - 4
57071	Secondary Tuberculosis Level - 1
57072	Secondary Tuberculosis Level - 2
57073	Secondary Tuberculosis Level - 3
57074	Secondary Tuberculosis Level - 4
57431	Schizophrenia Level - 1
57432	Schizophrenia Level - 2
57433	Schizophrenia Level - 3
57434	Schizophrenia Level - 4
57441	Eating Disorder Level - 1
57442	Eating Disorder Level - 2

PCRG	PCRG_Description
57443	Eating Disorder Level - 3
57444	Eating Disorder Level - 4
57471	Bi-Polar Disorder Level - 1
57472	Bi-Polar Disorder Level - 2
57473	Bi-Polar Disorder Level - 3
57474	Bi-Polar Disorder Level - 4
57481	Conduct, Impulse Control, and Other Disruptive Behavior Disorders Level - 1
57482	Conduct, Impulse Control, and Other Disruptive Behavior Disorders Level - 2
57483	Conduct, Impulse Control, and Other Disruptive Behavior Disorders Level - 3
57484	Conduct, Impulse Control, and Other Disruptive Behavior Disorders Level - 4
57491	Depressive and Other Psychoses Level - 1
57492	Depressive and Other Psychoses Level - 2
57493	Depressive and Other Psychoses Level - 3
57494	Depressive and Other Psychoses Level - 4
57501	Major Personality Disorders Level - 1
57502	Major Personality Disorders Level - 2
57503	Major Personality Disorders Level - 3
57504	Major Personality Disorders Level - 4
57511	Chronic Mental Health Diagnoses - Moderate Level - 1
57512	Chronic Mental Health Diagnoses - Moderate Level - 2
57513	Chronic Mental Health Diagnoses - Moderate Level - 3
57514	Chronic Mental Health Diagnoses - Moderate Level - 4
57821	Cocaine Abuse Level - 1
57822	Cocaine Abuse Level - 2
57823	Cocaine Abuse Level - 3
57824	Cocaine Abuse Level - 4

PCRG	PCRG_Description
57831	Opioid Abuse Level - 1
57832	Opioid Abuse Level - 2
57833	Opioid Abuse Level - 3
57834	Opioid Abuse Level - 4
57841	Chronic Alcohol Abuse Level - 1
57842	Chronic Alcohol Abuse Level - 2
57843	Chronic Alcohol Abuse Level - 3
57844	Chronic Alcohol Abuse Level - 4
57851	Other Significant Drug Abuse Level - 1
57852	Other Significant Drug Abuse Level - 2
57853	Other Significant Drug Abuse Level - 3
57854	Other Significant Drug Abuse Level - 4
57861	Drug Abuse - Cannabis/NOS/NEC Level - 1
57862	Drug Abuse - Cannabis/NOS/NEC Level - 2
57863	Drug Abuse - Cannabis/NOS/NEC Level - 3
57864	Drug Abuse - Cannabis/NOS/NEC Level - 4
58201	Burns - Extreme Level - 1
58202	Burns - Extreme Level - 2
58203	Burns - Extreme Level - 3
58204	Burns - Extreme Level - 4
61001	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 1
61002	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 2
61003	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 3
61004	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 4
61005	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 5
61006	Chronic Renal Failure and Other Dominant or Moderate Chronic Disease Level - 6

PCRG	PCRG_Description
61101	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 1
61102	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 2
61103	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 3
61104	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 4
61105	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 5
61106	Chronic Obstructive Pulmonary Disease and Congestive Heart Failure Level - 6
61111	Congestive Heart Failure and Diabetes Level - 1
61112	Congestive Heart Failure and Diabetes Level - 2
61113	Congestive Heart Failure and Diabetes Level - 3
61114	Congestive Heart Failure and Diabetes Level - 4
61115	Congestive Heart Failure and Diabetes Level - 5
61116	Congestive Heart Failure and Diabetes Level - 6
61121	Congestive Heart Failure and Peripheral Vascular Disease Level - 1
61122	Congestive Heart Failure and Peripheral Vascular Disease Level - 2
61123	Congestive Heart Failure and Peripheral Vascular Disease Level - 3
61124	Congestive Heart Failure and Peripheral Vascular Disease Level - 4
61125	Congestive Heart Failure and Peripheral Vascular Disease Level - 5
61126	Congestive Heart Failure and Peripheral Vascular Disease Level - 6
61131	Congestive Heart Failure and Cerebrovascular Disease Level - 1
61132	Congestive Heart Failure and Cerebrovascular Disease Level - 2
61133	Congestive Heart Failure and Cerebrovascular Disease Level - 3
61134	Congestive Heart Failure and Cerebrovascular Disease Level - 4
61135	Congestive Heart Failure and Cerebrovascular Disease Level - 5
61136	Congestive Heart Failure and Cerebrovascular Disease Level - 6
61141	Congestive Heart Failure and Other Dominant Chronic Disease Level - 1
61142	Congestive Heart Failure and Other Dominant Chronic Disease Level - 2

PCRG	PCRG_Description
61143	Congestive Heart Failure and Other Dominant Chronic Disease Level - 3
61144	Congestive Heart Failure and Other Dominant Chronic Disease Level - 4
61145	Congestive Heart Failure and Other Dominant Chronic Disease Level - 5
61146	Congestive Heart Failure and Other Dominant Chronic Disease Level - 6
61151	Congestive Heart Failure and Dementing Disease Level - 1
61152	Congestive Heart Failure and Dementing Disease Level - 2
61153	Congestive Heart Failure and Dementing Disease Level - 3
61154	Congestive Heart Failure and Dementing Disease Level - 4
61155	Congestive Heart Failure and Dementing Disease Level - 5
61156	Congestive Heart Failure and Dementing Disease Level - 6
61161	Congestive Heart Failure and Other Moderate Chronic Disease Level - 1
61162	Congestive Heart Failure and Other Moderate Chronic Disease Level - 2
61163	Congestive Heart Failure and Other Moderate Chronic Disease Level - 3
61164	Congestive Heart Failure and Other Moderate Chronic Disease Level - 4
61165	Congestive Heart Failure and Other Moderate Chronic Disease Level - 5
61166	Congestive Heart Failure and Other Moderate Chronic Disease Level - 6
61171	Congestive Heart Failure and Other Chronic Disease Level 2 Level - 1
61172	Congestive Heart Failure and Other Chronic Disease Level 2 Level - 2
61173	Congestive Heart Failure and Other Chronic Disease Level 2 Level - 3
61174	Congestive Heart Failure and Other Chronic Disease Level 2 Level - 4
61201	Chronic Obstructive Pulmonary Disease and Diabetes Level - 1
61202	Chronic Obstructive Pulmonary Disease and Diabetes Level - 2
61203	Chronic Obstructive Pulmonary Disease and Diabetes Level - 3
61204	Chronic Obstructive Pulmonary Disease and Diabetes Level - 4
61205	Chronic Obstructive Pulmonary Disease and Diabetes Level - 5
61206	Chronic Obstructive Pulmonary Disease and Diabetes Level - 6

PCRG	PCRG_Description
61211	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 1
61212	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 2
61213	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 3
61214	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 4
61215	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 5
61216	Chronic Obstructive Pulmonary Disease and Advanced Coronary Artery Disease Level - 6
61221	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 1
61222	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 2
61223	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 3
61224	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 4
61225	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 5
61226	Chronic Obstructive Pulmonary Disease and Other Dominant Chronic Disease Level - 6
61231	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 1
61232	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 2
61233	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 3
61234	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 4
61235	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 5
61236	Chronic Obstructive Pulmonary Disease and Other Moderate Chronic Disease Level - 6
61241	Chronic Obstructive Pulmonary Disease and Hypertension Level - 1
61242	Chronic Obstructive Pulmonary Disease and Hypertension Level - 2
61243	Chronic Obstructive Pulmonary Disease and Hypertension Level - 3
61244	Chronic Obstructive Pulmonary Disease and Hypertension Level - 4
61245	Chronic Obstructive Pulmonary Disease and Hypertension Level - 5
61246	Chronic Obstructive Pulmonary Disease and Hypertension Level - 6
61251	Chronic Obstructive Pulmonary Disease and Other Chronic Disease Level 2 Level - 1
61252	Chronic Obstructive Pulmonary Disease and Other Chronic Disease Level 2 Level - 2

PCRG	PCRG_Description
61253	Chronic Obstructive Pulmonary Disease and Other Chronic Disease Level 2 Level - 3
61254	Chronic Obstructive Pulmonary Disease and Other Chronic Disease Level 2 Level - 4
61301	Cerebrovascular Disease and Diabetes Level - 1
61302	Cerebrovascular Disease and Diabetes Level - 2
61303	Cerebrovascular Disease and Diabetes Level - 3
61304	Cerebrovascular Disease and Diabetes Level - 4
61305	Cerebrovascular Disease and Diabetes Level - 5
61306	Cerebrovascular Disease and Diabetes Level - 6
61311	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 1
61312	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 2
61313	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 3
61314	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 4
61315	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 5
61316	Cerebrovascular Disease and Other Dominant Chronic Disease Level - 6
61321	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 1
61322	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 2
61323	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 3
61324	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 4
61325	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 5
61326	Cerebrovascular Disease and Other Moderate Chronic Disease Level - 6
61331	Cerebrovascular Disease and Other Chronic Disease Level 2 Level - 1
61332	Cerebrovascular Disease and Other Chronic Disease Level 2 Level - 2
61333	Cerebrovascular Disease and Other Chronic Disease Level 2 Level - 3
61334	Cerebrovascular Disease and Other Chronic Disease Level 2 Level - 4
61401	Diabetes and Advanced Coronary Artery Disease Level - 1
61402	Diabetes and Advanced Coronary Artery Disease Level - 2

PCRG	PCRG_Description
61403	Diabetes and Advanced Coronary Artery Disease Level - 3
61404	Diabetes and Advanced Coronary Artery Disease Level - 4
61405	Diabetes and Advanced Coronary Artery Disease Level - 5
61406	Diabetes and Advanced Coronary Artery Disease Level - 6
61411	Diabetes and Other Dominant Chronic Disease Level - 1
61412	Diabetes and Other Dominant Chronic Disease Level - 2
61413	Diabetes and Other Dominant Chronic Disease Level - 3
61414	Diabetes and Other Dominant Chronic Disease Level - 4
61415	Diabetes and Other Dominant Chronic Disease Level - 5
61416	Diabetes and Other Dominant Chronic Disease Level - 6
61421	Diabetes and Asthma Level - 1
61422	Diabetes and Asthma Level - 2
61423	Diabetes and Asthma Level - 3
61424	Diabetes and Asthma Level - 4
61425	Diabetes and Asthma Level - 5
61426	Diabetes and Asthma Level - 6
61431	Diabetes and Other Moderate Chronic Disease Level - 1
61432	Diabetes and Other Moderate Chronic Disease Level - 2
61433	Diabetes and Other Moderate Chronic Disease Level - 3
61434	Diabetes and Other Moderate Chronic Disease Level - 4
61435	Diabetes and Other Moderate Chronic Disease Level - 5
61436	Diabetes and Other Moderate Chronic Disease Level - 6
61441	Diabetes and Hypertension Level - 1
61442	Diabetes and Hypertension Level - 2
61443	Diabetes and Hypertension Level - 3
61444	Diabetes and Hypertension Level - 4

PCRG	PCRG_Description
61445	Diabetes and Hypertension Level - 5
61446	Diabetes and Hypertension Level - 6
61451	Diabetes and Other Chronic Disease Level 2 Level - 1
61452	Diabetes and Other Chronic Disease Level 2 Level - 2
61453	Diabetes and Other Chronic Disease Level 2 Level - 3
61454	Diabetes and Other Chronic Disease Level 2 Level - 4
61501	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 1
61502	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 2
61503	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 3
61504	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 4
61505	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 5
61506	Advanced Coronary Artery Disease and Other Dominant Chronic Disease Level - 6
61511	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 1
61512	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 2
61513	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 3
61514	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 4
61515	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 5
61516	Advanced Coronary Artery Disease and Other Moderate Chronic Disease Level - 6
61521	Advanced Coronary Artery Disease and Other Chronic Disease Level 2 Level - 1
61522	Advanced Coronary Artery Disease and Other Chronic Disease Level 2 Level - 2
61523	Advanced Coronary Artery Disease and Other Chronic Disease Level 2 Level - 3
61524	Advanced Coronary Artery Disease and Other Chronic Disease Level 2 Level - 4
61601	Dementing Disease and Other Dominant Chronic Disease Level - 1
61602	Dementing Disease and Other Dominant Chronic Disease Level - 2
61603	Dementing Disease and Other Dominant Chronic Disease Level - 3
61604	Dementing Disease and Other Dominant Chronic Disease Level - 4

PCRG	PCRG_Description
61605	Dementing Disease and Other Dominant Chronic Disease Level - 5
61606	Dementing Disease and Other Dominant Chronic Disease Level - 6
61611	Dementing Disease and Other Moderate Chronic Disease Level - 1
61612	Dementing Disease and Other Moderate Chronic Disease Level - 2
61613	Dementing Disease and Other Moderate Chronic Disease Level - 3
61614	Dementing Disease and Other Moderate Chronic Disease Level - 4
61615	Dementing Disease and Other Moderate Chronic Disease Level - 5
61616	Dementing Disease and Other Moderate Chronic Disease Level - 6
61621	Dementing Disease and Other Chronic Disease Level 2 Level - 1
61622	Dementing Disease and Other Chronic Disease Level 2 Level - 2
61623	Dementing Disease and Other Chronic Disease Level 2 Level - 3
61624	Dementing Disease and Other Chronic Disease Level 2 Level - 4
61701	Schizophrenia and Other Dominant Chronic Disease Level - 1
61702	Schizophrenia and Other Dominant Chronic Disease Level - 2
61703	Schizophrenia and Other Dominant Chronic Disease Level - 3
61704	Schizophrenia and Other Dominant Chronic Disease Level - 4
61705	Schizophrenia and Other Dominant Chronic Disease Level - 5
61706	Schizophrenia and Other Dominant Chronic Disease Level - 6
61711	Schizophrenia and Other Moderate Chronic Disease Level - 1
61712	Schizophrenia and Other Moderate Chronic Disease Level - 2
61713	Schizophrenia and Other Moderate Chronic Disease Level - 3
61714	Schizophrenia and Other Moderate Chronic Disease Level - 4
61715	Schizophrenia and Other Moderate Chronic Disease Level - 5
61716	Schizophrenia and Other Moderate Chronic Disease Level - 6
61721	Schizophrenia and Other Chronic Disease Level 2 Level - 1
61722	Schizophrenia and Other Chronic Disease Level 2 Level - 2

PCRG	PCRG_Description
61723	Schizophrenia and Other Chronic Disease Level 2 Level - 3
61724	Schizophrenia and Other Chronic Disease Level 2 Level - 4
61801	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 1
61802	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 2
61803	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 3
61804	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 4
61805	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 5
61806	History of Hip Fracture Age > 64 and Other Dominant Chronic Disease Level - 6
61811	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 1
61812	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 2
61813	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 3
61814	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 4
61815	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 5
61816	History of Hip Fracture Age > 64 and Other Moderate Chronic Disease Level - 6
61821	History of Hip Fracture Age > 64 and Other Chronic Disease Level 2 Level - 1
61822	History of Hip Fracture Age > 64 and Other Chronic Disease Level 2 Level - 2
61823	History of Hip Fracture Age > 64 and Other Chronic Disease Level 2 Level - 3
61824	History of Hip Fracture Age > 64 and Other Chronic Disease Level 2 Level - 4
61901	Two Other Dominant Chronic Diseases Level - 1
61902	Two Other Dominant Chronic Diseases Level - 2
61903	Two Other Dominant Chronic Diseases Level - 3
61904	Two Other Dominant Chronic Diseases Level - 4
61905	Two Other Dominant Chronic Diseases Level - 5
61906	Two Other Dominant Chronic Diseases Level - 6
62001	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 1
62002	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 2

PCRG	PCRG_Description
62003	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 3
62004	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 4
62005	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 5
62006	Other Dominant Chronic Disease and Psychiatric Disease (Except Schizophrenia) Level - 6
62011	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 1
62012	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 2
62013	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 3
62014	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 4
62015	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 5
62016	Psychiatric Disease (Except Schizophrenia) and Other Moderate Chronic Disease Level - 6
62101	Other Dominant Chronic Disease and Breast Malignancy Level - 1
62102	Other Dominant Chronic Disease and Breast Malignancy Level - 2
62103	Other Dominant Chronic Disease and Breast Malignancy Level - 3
62104	Other Dominant Chronic Disease and Breast Malignancy Level - 4
62105	Other Dominant Chronic Disease and Breast Malignancy Level - 5
62106	Other Dominant Chronic Disease and Breast Malignancy Level - 6
62111	Breast Malignancy and Other Moderate Chronic Disease Level - 1
62112	Breast Malignancy and Other Moderate Chronic Disease Level - 2
62113	Breast Malignancy and Other Moderate Chronic Disease Level - 3
62114	Breast Malignancy and Other Moderate Chronic Disease Level - 4
62115	Breast Malignancy and Other Moderate Chronic Disease Level - 5
62116	Breast Malignancy and Other Moderate Chronic Disease Level - 6
62201	Other Dominant Chronic Disease and Prostate Malignancy Level - 1
62202	Other Dominant Chronic Disease and Prostate Malignancy Level - 2
62203	Other Dominant Chronic Disease and Prostate Malignancy Level - 3
62204	Other Dominant Chronic Disease and Prostate Malignancy Level - 4

PCRG	PCRG_Description
62205	Other Dominant Chronic Disease and Prostate Malignancy Level - 5
62206	Other Dominant Chronic Disease and Prostate Malignancy Level - 6
62211	Prostate Malignancy and Other Moderate Chronic Disease Level - 1
62212	Prostate Malignancy and Other Moderate Chronic Disease Level - 2
62213	Prostate Malignancy and Other Moderate Chronic Disease Level - 3
62214	Prostate Malignancy and Other Moderate Chronic Disease Level - 4
62215	Prostate Malignancy and Other Moderate Chronic Disease Level - 5
62216	Prostate Malignancy and Other Moderate Chronic Disease Level - 6
62301	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 1
62302	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 2
62303	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 3
62304	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 4
62305	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 5
62306	Other Dominant Chronic Disease and Other Nondominant Malignancy Level - 6
62311	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 1
62312	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 2
62313	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 3
62314	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 4
62315	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 5
62316	Other Nondominant Malignancy and Other Moderate Chronic Disease Level - 6
62401	Other Dominant Chronic Disease and Asthma Level - 1
62402	Other Dominant Chronic Disease and Asthma Level - 2
62403	Other Dominant Chronic Disease and Asthma Level - 3
62404	Other Dominant Chronic Disease and Asthma Level - 4
62405	Other Dominant Chronic Disease and Asthma Level - 5
62406	Other Dominant Chronic Disease and Asthma Level - 6

PCRG	PCRG_Description
62411	Asthma and Other Moderate Chronic Disease Level - 1
62412	Asthma and Other Moderate Chronic Disease Level - 2
62413	Asthma and Other Moderate Chronic Disease Level - 3
62414	Asthma and Other Moderate Chronic Disease Level - 4
62415	Asthma and Other Moderate Chronic Disease Level - 5
62416	Asthma and Other Moderate Chronic Disease Level - 6
62421	Asthma and Hypertension Level - 1
62422	Asthma and Hypertension Level - 2
62423	Asthma and Hypertension Level - 3
62424	Asthma and Hypertension Level - 4
62425	Asthma and Hypertension Level - 5
62426	Asthma and Hypertension Level - 6
62501	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 1
62502	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 2
62503	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 3
62504	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 4
62505	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 5
62506	Other Dominant Chronic Disease and Moderate Chronic Substance Abuse Level - 6
62511	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 1
62512	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 2
62513	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 3
62514	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 4
62515	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 5
62516	Moderate Chronic Substance Abuse and Other Moderate Chronic Disease Level - 6
62601	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 1
62602	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 2

PCRG	PCRG_Description
62603	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 3
62604	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 4
62605	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 5
62606	One Other Dominant Chronic Disease and One or More Moderate Chronic Disease Level - 6
62611	One Other Dominant Chronic Disease and Other Chronic Disease Level 2 Level - 1
62612	One Other Dominant Chronic Disease and Other Chronic Disease Level 2 Level - 2
62613	One Other Dominant Chronic Disease and Other Chronic Disease Level 2 Level - 3
62614	One Other Dominant Chronic Disease and Other Chronic Disease Level 2 Level - 4
62701	Two Other Moderate Chronic Diseases Level - 1
62702	Two Other Moderate Chronic Diseases Level - 2
62703	Two Other Moderate Chronic Diseases Level - 3
62704	Two Other Moderate Chronic Diseases Level - 4
62705	Two Other Moderate Chronic Diseases Level - 5
62706	Two Other Moderate Chronic Diseases Level - 6
62801	Breast Malignancy and Other Chronic Disease Level 2 Level - 1
62802	Breast Malignancy and Other Chronic Disease Level 2 Level - 2
62811	Prostate Malignancy and Other Chronic Disease Level 2 Level - 1
62812	Prostate Malignancy and Other Chronic Disease Level 2 Level - 2
62821	Other Nondominant Malignancy and Other Chronic Disease Level 2 Level - 1
62822	Other Nondominant Malignancy and Other Chronic Disease Level 2 Level - 2
62901	Psychiatric Disease (Except Schizophrenia) and Other Chronic Disease Level 2 Level - 1
62902	Psychiatric Disease (Except Schizophrenia) and Other Chronic Disease Level 2 Level - 2
62903	Psychiatric Disease (Except Schizophrenia) and Other Chronic Disease Level 2 Level - 3
62904	Psychiatric Disease (Except Schizophrenia) and Other Chronic Disease Level 2 Level - 4
62911	Asthma and Other Chronic Disease Level 2 Level - 1
62912	Asthma and Other Chronic Disease Level 2 Level - 2

PCRG	PCRG_Description
62913	Asthma and Other Chronic Disease Level 2 Level - 3
62914	Asthma and Other Chronic Disease Level 2 Level - 4
62921	Moderate Chronic Substance Abuse and Other Chronic Disease Level 2 Level - 1
62922	Moderate Chronic Substance Abuse and Other Chronic Disease Level 2 Level - 2
62923	Moderate Chronic Substance Abuse and Other Chronic Disease Level 2 Level - 3
62924	Moderate Chronic Substance Abuse and Other Chronic Disease Level 2 Level - 4
62931	One Other Moderate Chronic Disease and Other Chronic Disease Level 2 Level - 1
62932	One Other Moderate Chronic Disease and Other Chronic Disease Level 2 Level - 2
62933	One Other Moderate Chronic Disease and Other Chronic Disease Level 2 Level - 3
62934	One Other Moderate Chronic Disease and Other Chronic Disease Level 2 Level - 4
70011	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 1
70012	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 2
70013	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 3
70014	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 4
70015	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 5
70016	Chronic Renal Failure - Diabetes - Other Dominant Chronic Disease Level - 6
70021	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 1
70022	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 2
70023	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 3
70024	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 4
70025	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 5
70026	Chronic Renal Failure - 2 or More Other Dominant Chronic Diseases Level - 6
70101	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 1
70102	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 2
70103	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 3
70104	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 4

PCRG	PCRG_Description
70105	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 5
70106	Congestive Heart Failure - Diabetes - Chronic Obstructive Pulmonary Disease Level - 6
70111	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 1
70112	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 2
70113	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 3
70114	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 4
70115	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 5
70116	Congestive Heart Failure - Diabetes - Cerebrovascular Disease Level - 6
70121	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 1
70122	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 2
70123	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 3
70124	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 4
70125	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 5
70126	Congestive Heart Failure - Diabetes - Other Dominant Chronic Disease Level - 6
70131	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 1
70132	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 2
70133	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 3
70134	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 4
70135	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 5
70136	Congestive Heart Failure - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 6
70141	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 1
70142	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 2
70143	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 3
70144	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 4
70145	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 5
70146	Congestive Heart Failure - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 6

PCRG	PCRG_Description
70151	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 1
70152	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 2
70153	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 3
70154	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 4
70155	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 5
70156	Congestive Heart Failure - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 6
70161	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 1
70162	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 2
70163	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 3
70164	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 4
70165	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 5
70166	Congestive Heart Failure - 2 or More Other Dominant Chronic Diseases Level - 6
70201	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 1
70202	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 2
70203	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 3
70204	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 4
70205	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 5
70206	Diabetes - Advanced Coronary Artery Disease - Other Dominant Chronic Disease Level - 6
70211	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 1
70212	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 2
70213	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 3
70214	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 4
70215	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 5
70216	Diabetes - Cerebrovascular Disease - Other Dominant Chronic Disease Level - 6
70221	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 1
70222	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 2

PCRG	PCRG_Description
70223	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 3
70224	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 4
70225	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 5
70226	Diabetes - Chronic Obstructive Pulmonary Disease - Other Dominant Chronic Disease Level - 6
70231	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 1
70232	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 2
70233	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 3
70234	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 4
70235	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 5
70236	Diabetes - 2 or More Other Dominant Chronic Diseases Level - 6
70301	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 1
70302	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 2
70303	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 3
70304	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 4
70305	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 5
70306	Chronic Obstructive Pulmonary Disease - Advanced Coronary Artery Disease - Other Dominant Chronic Level - 6
70311	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 1
70312	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 2
70313	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 3
70314	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 4
70315	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 5
70316	Chronic Obstructive Pulmonary Disease - 2 or More Other Dominant Chronic Diseases Level - 6
70401	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 1
70402	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 2
70403	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 3
70404	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 4

PCRG	PCRG_Description
70405	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 5
70406	Advanced Coronary Artery Disease - Peripheral Vascular Disease - Other Dominant Chronic Disease Level - 6
70411	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 1
70412	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 2
70413	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 3
70414	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 4
70415	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 5
70416	Advanced Coronary Artery Disease - 2 or More Other Dominant Chronic Diseases Level - 6
70501	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 1
70502	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 2
70503	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 3
70504	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 4
70505	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 5
70506	Cerebrovascular Disease - 2 or More Other Dominant Chronic Diseases Level - 6
70601	3 or More Other Dominant Chronic Diseases Level - 1
70602	3 or More Other Dominant Chronic Diseases Level - 2
70603	3 or More Other Dominant Chronic Diseases Level - 3
70604	3 or More Other Dominant Chronic Diseases Level - 4
70605	3 or More Other Dominant Chronic Diseases Level - 5
70606	3 or More Other Dominant Chronic Diseases Level - 6
70701	Diabetes - Cerebrovascular Disease - Hypertension Level - 1
70702	Diabetes - Cerebrovascular Disease - Hypertension Level - 2
70703	Diabetes - Cerebrovascular Disease - Hypertension Level - 3
70704	Diabetes - Cerebrovascular Disease - Hypertension Level - 4
70705	Diabetes - Cerebrovascular Disease - Hypertension Level - 5
70706	Diabetes - Cerebrovascular Disease - Hypertension Level - 6

PCRG	PCRG_Description
70711	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 1
70712	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 2
70713	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 3
70714	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 4
70715	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 5
70716	Diabetes - Hypertension - Other Dominant Chronic Disease Level - 6
80011	Multiple Dominant Primary Malignancies Level - 1
80012	Multiple Dominant Primary Malignancies Level - 2
80013	Multiple Dominant Primary Malignancies Level - 3
80014	Multiple Dominant Primary Malignancies Level - 4
80021	Multiple Non-Dominant Primary Malignancies Level - 1
80022	Multiple Non-Dominant Primary Malignancies Level - 2
80023	Multiple Non-Dominant Primary Malignancies Level - 3
80024	Multiple Non-Dominant Primary Malignancies Level - 4
86411	Secondary Malignancy Level - 1
86412	Secondary Malignancy Level - 2
86413	Secondary Malignancy Level - 3
86414	Secondary Malignancy Level - 4
86461	Brain and Central Nervous System Malignancies Level - 1
86462	Brain and Central Nervous System Malignancies Level - 2
86463	Brain and Central Nervous System Malignancies Level - 3
86464	Brain and Central Nervous System Malignancies Level - 4
86471	Lung Malignancy Level - 1
86472	Lung Malignancy Level - 2
86473	Lung Malignancy Level - 3
86474	Lung Malignancy Level - 4

PCRG	PCRG_Description
86481	Pancreatic Malignancy Level - 1
86482	Pancreatic Malignancy Level - 2
86483	Pancreatic Malignancy Level - 3
86484	Pancreatic Malignancy Level - 4
86491	Kidney Malignancy Level - 1
86492	Kidney Malignancy Level - 2
86493	Kidney Malignancy Level - 3
86494	Kidney Malignancy Level - 4
86501	Ovarian Malignancy Level - 1
86502	Ovarian Malignancy Level - 2
86503	Ovarian Malignancy Level - 3
86504	Ovarian Malignancy Level - 4
86511	Digestive Malignancy Level - 1
86512	Digestive Malignancy Level - 2
86513	Digestive Malignancy Level - 3
86514	Digestive Malignancy Level - 4
86521	Chronic Lymphoid Leukemia Level - 1
86522	Chronic Lymphoid Leukemia Level - 2
86523	Chronic Lymphoid Leukemia Level - 3
86524	Chronic Lymphoid Leukemia Level - 4
86531	Chronic Non-Lymphoid Leukemia Level - 1
86532	Chronic Non-Lymphoid Leukemia Level - 2
86533	Chronic Non-Lymphoid Leukemia Level - 3
86534	Chronic Non-Lymphoid Leukemia Level - 4
86541	Multiple Myeloma Level - 1
86542	Multiple Myeloma Level - 2

PCRG	PCRG_Description
86543	Multiple Myeloma Level - 3
86544	Multiple Myeloma Level - 4
86551	Acute Lymphoid Leukemia Level - 1
86552	Acute Lymphoid Leukemia Level - 2
86553	Acute Lymphoid Leukemia Level - 3
86554	Acute Lymphoid Leukemia Level - 4
86561	Acute Non-Lymphoid Leukemia Level - 1
86562	Acute Non-Lymphoid Leukemia Level - 2
86563	Acute Non-Lymphoid Leukemia Level - 3
86564	Acute Non-Lymphoid Leukemia Level - 4
86571	Colon Malignancy Level - 1
86572	Colon Malignancy Level - 2
86573	Colon Malignancy Level - 3
86574	Colon Malignancy Level - 4
86581	Other Malignancies Level - 1
86582	Other Malignancies Level - 2
86583	Other Malignancies Level - 3
86584	Other Malignancies Level - 4
86601	Hodgkin's Lymphoma Level - 1
86602	Hodgkin's Lymphoma Level - 2
86603	Hodgkin's Lymphoma Level - 3
86604	Hodgkin's Lymphoma Level - 4
86611	Plasma Protein Malignancy Level - 1
86612	Plasma Protein Malignancy Level - 2
86613	Plasma Protein Malignancy Level - 3
86614	Plasma Protein Malignancy Level - 4

PCRG	PCRG_Description
86621	Breast Malignancy Level - 1
86622	Breast Malignancy Level - 2
86623	Breast Malignancy Level - 3
86624	Breast Malignancy Level - 4
86631	Prostate Malignancy Level - 1
86632	Prostate Malignancy Level - 2
86633	Prostate Malignancy Level - 3
86634	Prostate Malignancy Level - 4
86641	Genitourinary Malignancy Level - 1
86642	Genitourinary Malignancy Level - 2
86643	Genitourinary Malignancy Level - 3
86644	Genitourinary Malignancy Level - 4
86651	Non-Hodgkin's Lymphoma Level - 1
86652	Non-Hodgkin's Lymphoma Level - 2
86653	Non-Hodgkin's Lymphoma Level - 3
86654	Non-Hodgkin's Lymphoma Level - 4
90101	Dialysis with Diabetes Level - 1
90102	Dialysis with Diabetes Level - 2
90103	Dialysis with Diabetes Level - 3
90104	Dialysis with Diabetes Level - 4
90201	Dialysis without Diabetes Level - 1
90202	Dialysis without Diabetes Level - 2
90203	Dialysis without Diabetes Level - 3
90204	Dialysis without Diabetes Level - 4
90301	HIV Disease Level - 1
90302	HIV Disease Level - 2

PCRG	PCRG_Description
90303	HIV Disease Level - 3
90304	HIV Disease Level - 4
90401	Total Parenteral Nutrition Level - 1
90402	Total Parenteral Nutrition Level - 2
90403	Total Parenteral Nutrition Level - 3
90404	Total Parenteral Nutrition Level - 4
90501	Dependence on a Mechanical Ventilator Level - 1
90502	Dependence on a Mechanical Ventilator Level - 2
90503	Dependence on a Mechanical Ventilator Level - 3
90504	Dependence on a Mechanical Ventilator Level - 4
90601	History of a Major Organ Transplant Level - 1
90602	History of a Major Organ Transplant Level - 2
90603	History of a Major Organ Transplant Level - 3
90604	History of a Major Organ Transplant Level - 4
90701	Congenital Quadriplegia, Diplegia or Hemiplegia Level - 1
90702	Congenital Quadriplegia, Diplegia or Hemiplegia Level - 2
90703	Congenital Quadriplegia, Diplegia or Hemiplegia Level - 3
90704	Congenital Quadriplegia, Diplegia or Hemiplegia Level - 4
90801	Acquired Quadriplegia or Permanent Vegetative State Level - 1
90802	Acquired Quadriplegia or Permanent Vegetative State Level - 2
90803	Acquired Quadriplegia or Permanent Vegetative State Level - 3
90804	Acquired Quadriplegia or Permanent Vegetative State Level - 4
90901	Spina Bifida Level - 1
90902	Spina Bifida Level - 2
90903	Spina Bifida Level - 3
90904	Spina Bifida Level - 4

PCRG	PCRG_Description
91001	Progressive Muscular Dystrophy or Spinal Muscle Atrophy Level - 1
91002	Progressive Muscular Dystrophy or Spinal Muscle Atrophy Level - 2
91003	Progressive Muscular Dystrophy or Spinal Muscle Atrophy Level - 3
91004	Progressive Muscular Dystrophy or Spinal Muscle Atrophy Level - 4
91101	Cystic Fibrosis Level - 1
91102	Cystic Fibrosis Level - 2
91103	Cystic Fibrosis Level - 3
91104	Cystic Fibrosis Level - 4

Attachment A5_CDPS
(task6b_cdpspred.log)

```
1    The SAS System                                17:02 Thursday, March 3, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           1.48 seconds
      cpu time            0.06 seconds

1          /*
2          task6b_cdpspred.sas
3          Amresh Hanchate
4          DOD1 Project
5          Last revision: March 3, 2005 (see note below on the revision
details)
6
7          Using the grouped diagnoses indicators from CDPS, this job regresses
second-year expenditures on the group indicators for
7        !  the estimation sample and obtains predicted expenditures for the
validation sample. These predicted amounts are attached
7        !  to dod2 to ge
8
9          The main input file is cdpsbg.sas7bdat which was produced by
/data/dod1/saswork/task5b_cdps.sas. Other input files are
9        ! dod2charb and dod2.
10         Output datafile is dod3cdpsb.sas7bdat.
11
12         March 3, 2005 revision: The only change made is that when running
the regressions previously all age/genders dummy
12       ! variables were included -- now we have dropped one as the reference
category.
13
14          */
15
16         options ps=60 ls=80 nocenter;
17         footnote '/data/dod1/saswork/task6b_cdpspred.sas';
18
19         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
19       !
20         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
```

```
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
NOTE: Format COSTCAT has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.58 seconds
      cpu time            0.01 seconds
```

```
53
54          * prepare data for running regression;
55          * varlist of CDPS diagnoses groups;
56          %let cdpslist = carvh carm carl carel psyh psym psyl skcm skcl skcvl
56        ! skcel cnsh cnsm cnsl pulvh
57          pulh pulm pull gih gim gil dia1h dia1m dia2m dia2l sknh sknl sknvl
57        ! renvh renm renl subl subvl
58          canh canm canl ddm ddl genel meth metm metvl prgcmp prginc eyel
eyevl
58        !  cerl aidsh infh hivm infm
59          infl hemeh hemvh hemm heml ccarvh ccnsm cpulvh cgih cgim cgil cdia2l
59        ! cmeth cmetm cinfm chemvh;
60
61          data cdps;
62            length &cdpslist 3;
63            set saswork.cdpsbg (keep=recipno age male &cdpslist);
64            *create age and gender indicators;
65            if male=1 & age LE 1 then m0_1=1; else m0_1=0;
66            if male=1 & (age GE 2) & (age LE 10) then m2_10=1; else m2_10=0;
67            if male=1 & (age GE 11) & (age LE 17) then m11_17=1; else
m11_17=0;
68            if male=1 & (age GE 18) & (age LE 25) then m18_25=1; else
m18_25=0;
69            if male=1 & (age GE 26) & (age LE 30) then m26_30=1; else
m26_30=0;
70            if male=1 & (age GE 31) & (age LE 35) then m31_35=1; else
m31_35=0;
71            if male=1 & (age GE 36) & (age LE 40) then m36_40=1; else
m36_40=0;
72            if male=1 & (age GE 41) & (age LE 45) then m41_45=1; else
m41_45=0;
73            if male=1 & (age GE 46) & (age LE 50) then m46_50=1; else
m46_50=0;
74            if male=1 & (age GE 51) & (age LE 55) then m51_55=1; else
m51_55=0;
75            if male=1 & (age GE 56) & (age LE 60) then m56_60=1; else
m56_60=0;
76            if male=1 & (age GE 61) & (age LE 64) then m61_64=1; else
m61_64=0;
77            if male=0 & age LE 1 then f0_1=1; else f0_1=0;
78            if male=0 & (age GE 2) & (age LE 10) then f2_10=1; else f2_10=0;
79            if male=0 & (age GE 11) & (age LE 17) then f11_17=1; else
f11_17=0;
80            if male=0 & (age GE 18) & (age LE 25) then f18_25=1; else
f18_25=0;
81            if male=0 & (age GE 26) & (age LE 30) then f26_30=1; else
f26_30=0;
82            if male=0 & (age GE 31) & (age LE 35) then f31_35=1; else
f31_35=0;
83            if male=0 & (age GE 36) & (age LE 40) then f36_40=1; else
f36_40=0;
84            if male=0 & (age GE 41) & (age LE 45) then f41_45=1; else
f41_45=0;
85            if male=0 & (age GE 46) & (age LE 50) then f46_50=1; else
f46_50=0;
86            if male=0 & (age GE 51) & (age LE 55) then f51_55=1; else
f51_55=0;
87            if male=0 & (age GE 56) & (age LE 60) then f56_60=1; else
f56_60=0;
```

```
88           if male=0 & (age GE 61) & (age LE 64) then f61_64=1; else
f61_64=0;
89         run;

NOTE: There were 2304926 observations read from the data set SASWORK.CDPSBG.
NOTE: The data set WORK.CDPS has 2304926 observations and 94 variables.
NOTE: DATA statement used:
      real time           16:30.06
      cpu time            1:34.87


89       !
90
91         * cdps does not contain the expenditure variables provided - so add
91       ! them here by merging from the file (dod2char) used for running cdps
91       ! -- also get datatype variable;
92         proc sort data=cdps; by recipno; run;
```

```
NOTE: There were 2304926 observations read from the data set WORK.CDPS.
NOTE: The data set WORK.CDPS has 2304926 observations and 94 variables.
NOTE: PROCEDURE SORT used:
      real time           25:33.55
      cpu time            2:24.27


92       !
93         proc sort data=saswork.dod2charb; by recipno; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.07 seconds
      cpu time            0.00 seconds


93       !
94         data cdps2;
95           merge cdps (in=infirst) saswork.dod2charb (in=insecond keep =
95       ! recipno datatype expyr expyrt25 expyrt50);
96           by recipno;
97           *create a revised version of expenditure variables (expyr,
expyrt25
97       ! and expyrt50) that have missing entries for validation sample;
98           if datatype=1 then expyrs = expyr;
99           if datatype=1 then expyrt25s = expyrt25;
100          if datatype=1 then expyrt50s = expyrt50;
101          label datatype='1=fitting sample, 2=Validation sample';
102          label expyr = 'Total health care expenditures, FY2002';
103          label expyrt25 = 'expyr topped at $25K';
104          label expyrt50 = 'expyr topped at $50K';
105        run;

NOTE: There were 2304926 observations read from the data set WORK.CDPS.
NOTE: There were 2304926 observations read from the data set SASWORK.DOD2CHARB.
NOTE: The data set WORK.CDPS2 has 2304926 observations and 101 variables.
NOTE: DATA statement used:
      real time           20:18.84
      cpu time            2:07.66


105      !
106
107        /* since some regressors (age groups or diagnoses) may take only one
107      ! value for all data, first identify these (note that regression is
107      ! only on the fitting sample; with separate regressions for adults and
107      ! children)
108        */
109        * list of all regressors -- define f61_64 as the reference group;
110        %let xlist = m0_1 m2_10 m11_17 m18_25 m26_30 m31_35 m36_40 m41_45
110      ! m46_50
111        m51_55 m56_60 m61_64 f0_1 f2_10 f11_17 f18_25 f26_30 f31_35 f36_40
111      ! f41_45
112        f46_50 f51_55 f56_60
113        carvh carm carl carel psyh psym psyl skcm skcl skcvl skcel cnsh cnsm
113      ! cnsl
114        pulvh pulh pulm pull gih gim gil dia1h dia1m dia2m dia2l sknh sknl
114      ! sknvl
115        renvh renm renl subl subvl canh canm canl ddm ddl genel meth metm
```

```
115      ! metvl
116        prgcmp prginc eyel eyevl cerl aidsh infh hivm infm infl hemeh hemvh
116      ! hemm
117        heml ccarvh ccnsm cpulvh cgih cgim cgil cdia2l cmeth cmetm cinfm
117      ! chemvh;
118
119        proc means data=cdps2 (where=(datatype=2)) min max mean;
119      !
120              var &xlist;
120      !
121              title "Means for validation sample";
121      !
122        run;

NOTE: There were 500000 observations read from the data set WORK.CDPS2.
      WHERE datatype=2;
NOTE: The PROCEDURE MEANS printed pages 1-2.
NOTE: PROCEDURE MEANS used:
      real time           4:41.68
      cpu time            42.72 seconds


122      !
123
124        *freq of different diagnoses;
124      !
125        proc freq data=cdps2 (where=(datatype=2));
125      !
126              tables &xlist;
126      !
127              title "Freq for validation sample";
127      !
128        run;

NOTE: There were 500000 observations read from the data set WORK.CDPS2.
      WHERE datatype=2;
NOTE: The PROCEDURE FREQ printed pages 3-17.
NOTE: PROCEDURE FREQ used:
      real time           8:59.34
      cpu time            50.91 seconds


128      !
129
130        * run regression for adults;
131        proc reg data=cdps2;
131      !
132           model expyrs expyrt25s expyrt50s = &xlist;
133           where age GE 18;
134           output out=cdps3a (keep=recipno datatype expyr expyrt25 expyrt50
134      ! pexp_cdps pexpt25_cdps pexpt50_cdps) p=pexp_cdps pexpt25_cdps
134      ! pexpt50_cdps;
135        run;

NOTE: 1529597 observations read.
NOTE: 331312 observations have missing values.
NOTE: 1198285 observations used in computations.
135      !
```

```
136
137        * run regression for children;

NOTE: There were 1529597 observations read from the data set WORK.CDPS2.
      WHERE age>=18;
NOTE: The data set WORK.CDPS3A has 1529597 observations and 8 variables.
NOTE: The PROCEDURE REG printed pages 18-32.
NOTE: PROCEDURE REG used:
      real time           18:49.68
      cpu time            2:58.90


138        proc reg data=cdps2;
138      !
139           model expyrs expyrt25s expyrt50s = &xlist;
140           where age LE 17;
141           output out=cdps3b (keep=recipno datatype expyr expyrt25 expyrt50
141      ! pexp_cdps pexpt25_cdps pexpt50_cdps) p=pexp_cdps pexpt25_cdps
141      ! pexpt50_cdps;
142        run;

NOTE: 775329 observations read.
NOTE: 168688 observations have missing values.
NOTE: 606641 observations used in computations.
142      !
143
144        * merge the adult and children files;
145        * keep only required variables;

NOTE: There were 775329 observations read from the data set WORK.CDPS2.
      WHERE age<=17;
NOTE: The data set WORK.CDPS3B has 775329 observations and 8 variables.
NOTE: The PROCEDURE REG printed pages 33-50.
NOTE: PROCEDURE REG used:
      real time           7:14.86
      cpu time            1:34.63


146        data cdps3;
147          set cdps3a cdps3b;
148        run;

NOTE: There were 1529597 observations read from the data set WORK.CDPS3A.
NOTE: There were 775329 observations read from the data set WORK.CDPS3B.
NOTE: The data set WORK.CDPS3 has 2304926 observations and 8 variables.
NOTE: DATA statement used:
      real time           2:04.71
      cpu time            17.15 seconds


148      !
149
150        * delete unnecessary files;
151        proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
```

```
                     -----Directory-----

Physical Name:     /data/saswork/SAS_work704600005BE0_genmed2
File Name:         /data/saswork/SAS_work704600005BE0_genmed2
Inode Number:      4142272
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype     File Size  Last Modified
-----------------------------------------------------
1  CDPS     DATA        983457792   03MAR2005:17:44:12
2  CDPS2    DATA       1110761472   03MAR2005:18:04:38
3  CDPS3    DATA        186966016   03MAR2005:18:46:30
4  CDPS3A   DATA        124076032   03MAR2005:18:37:10
5  CDPS3B   DATA         62898176   03MAR2005:18:44:25
6  FORMATS  CATALOG         28672   03MAR2005:17:02:15
7  REGSTRY  ITEMSTOR        32768   03MAR2005:17:02:14
151      !
152
152      !  delete cdps2 cdps3a cdps3b;
153        run;

NOTE: Deleting WORK.CDPS2 (memtype=DATA).
NOTE: Deleting WORK.CDPS3A (memtype=DATA).
NOTE: Deleting WORK.CDPS3B (memtype=DATA).
153      !
154
155        *save predicted expenditures from cdps in new file
saswork.dod3cdpsb;
155      !
156        *merge with dod2 (dod2 has individual characteristics);

NOTE: PROCEDURE DATASETS used:
      real time           4.52 seconds
      cpu time            1.17 seconds


157        proc sort data=cdps3 (rename=(recipno=pid)); by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.CDPS3.
NOTE: The data set WORK.CDPS3 has 2304926 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           4:59.45
      cpu time            42.49 seconds


157      !
158        proc sort data=saswork.dod2; by pid; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.01 seconds


158      !
159        data cdps3b;
```

```
160          merge cdps3 (in=infirst) saswork.dod2 (in=insecond);
161          by pid;
162          if infirst=1 & insecond=1;
163        run;

NOTE: There were 2304926 observations read from the data set WORK.CDPS3.
NOTE: There were 2304926 observations read from the data set SASWORK.DOD2.
NOTE: The data set WORK.CDPS3B has 2304926 observations and 32 variables.
NOTE: DATA statement used:
      real time           6:30.25
      cpu time            1:00.31


163      !
164        *post-estimation analysis;
165        *limit to validation sample;
166        data cdps4;
167          set cdps3b;
168          abserr = abs(expyr-pexp_cdps);
169          abserrt25 = abs(expyrt25-pexpt25_cdps);
170          abserrt50 = abs(expyrt50-pexpt50_cdps);
171          if datatype = 2;
172          label abserr = 'Absolute prediction error for expyr';
173          label abserrt25 = 'Absolute prediction error for expyrt25';
174          label abserrt50 = 'Absolute prediction error for expyrt50';
175        run;

NOTE: There were 2304926 observations read from the data set WORK.CDPS3B.
NOTE: The data set WORK.CDPS4 has 500000 observations and 35 variables.
NOTE: DATA statement used:
      real time           2:09.40
      cpu time            19.59 seconds


175      !
176        proc datasets library=work;
                    -----Directory-----

Libref:             WORK
Engine:             V8
Physical Name:      /data/saswork/SAS_work704600005BE0_genmed2
File Name:          /data/saswork/SAS_work704600005BE0_genmed2
Inode Number:       4142272
Access Permission: rwxrwx---
Owner Name:         amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  CDPS     DATA       983457792   03MAR2005:17:44:12
2  CDPS3    DATA       186966016   03MAR2005:18:51:33
3  CDPS3B   DATA       519716864   03MAR2005:18:58:03
4  CDPS4    DATA       125419520   03MAR2005:19:00:13
5  FORMATS  CATALOG        28672   03MAR2005:17:02:15
6  REGSTRY  ITEMSTOR       32768   03MAR2005:17:02:14
176      !
177
```

```
177      !  delete cdps3 cdps3b;
178        run;

NOTE: Deleting WORK.CDPS3 (memtype=DATA).
NOTE: Deleting WORK.CDPS3B (memtype=DATA).
178      !
179
180        * create absolute deviation from mean variable;

NOTE: PROCEDURE DATASETS used:
      real time           3.50 seconds
      cpu time            0.71 seconds


181        proc sql;
181      !
182          create table cdps5 as
182      !
183          select *,  abs(expyr - mean(expyr)) as absmdev,
184           abs(expyrt25 - mean(expyrt25)) as absmdevt25,
185           abs(expyrt50 - mean(expyrt50)) as absmdevt50
186          from cdps4;
NOTE: The query requires remerging summary statistics back with the original
      data.
NOTE: Table WORK.CDPS5 created, with 500000 rows and 38 columns.

186      !
187        quit;
NOTE: PROCEDURE SQL used:
      real time           2:59.90
      cpu time            24.71 seconds


187      !
188        data saswork.dod3cdpsb;
189          set cdps5;
190          label absmdev='Absolute mean deviation for expyr';
191          label absmdevt25='Absolute mean deviation for expyrt25';
192          label absmdevt50='Absolute mean deviation for expyrt50';
193        run;

NOTE: There were 500000 observations read from the data set WORK.CDPS5.
NOTE: The data set SASWORK.DOD3CDPSB has 500000 observations and 38 variables.
NOTE: DATA statement used:
      real time           1:12.04
      cpu time            9.18 seconds


193      !
194
195        proc contents data=saswork.dod3cdpsb varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.12 seconds
      cpu time            0.02 seconds

NOTE: The PROCEDURE CONTENTS printed pages 51-52.
```

```
195      !
196
197
198
199
200
201

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           2:02:17.11
      cpu time            15:09.40
```

Attachment A5_ACG
(task6c_acgpred.log)

```
1    The SAS System                                15:49 Friday, March 4, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.

NOTE: SAS initialization used:
      real time           0.68 seconds
      cpu time            0.09 seconds

1          /*
2          task6c_acgpred.sas
3          Amresh Hanchate
4          DOD1 Project
5
6          Last revision: March 3, 2005 (see below for changes)
7
8          Using the grouped diagnoses indicators from ACG, this job regresses
second-year expenditures on the group indicators for
8        ! the estimation sample and obtains predicted expenditures for the
validation sample. These predicted amounts are attached
8        ! to dod2 to get
9
10         Input: dod2acgbout.txt, dod2acgbedc.txt (both produced by the ACG
run task5c_acg.txt) and dod2charb;
11         Output: dod3acgb.v8x
12
13         Revisions on March 3  2005:
14         i)
15
16          */
17
18         options ps=60 ls=80 nocenter;
19
20         footnote '/data/dod1/saswork/task6c_acgpred.sas';
21         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
21       !
22         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
NOTE: Format COSTCAT has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.54 seconds
      cpu time            0.02 seconds
```

```
55
56          *for the ACG-PM model a lot of new indicator variables need to be
56        ! created (see pages 15. 29-31) of ACG Version 6.0 Release Notes for
56        ! what variables go into the PM model;
57          *first indicators of the selected ACGs;
58          data acgdat1;
59            infile '/data/dod1/saswork/dod2acgbout.txt';
60            input pid $ 1-18 acg $ 19-22 rub 23 hos $ 24 del 25 pre 26;
61          run;

NOTE: The infile '/data/dod1/saswork/dod2acgbout.txt' is:
      File Name=/data/dod1/saswork/dod2acgbout.txt,
      Owner Name=amresh,Group Name=sas,
      Access Permission=rw-r--r--,
      File Size (bytes)=62233002

NOTE: 2304926 records were read from the infile
      '/data/dod1/saswork/dod2acgbout.txt'.
      The minimum record length was 26.
      The maximum record length was 26.
NOTE: The data set WORK.ACGDAT1 has 2304926 observations and 6 variables.
NOTE: DATA statement used:
      real time           2:43.25
      cpu time            27.17 seconds


61        !
62          *following to make sure all values in valid range -- answer YES (acg
62        ! values checked against acg print file);
63          proc freq;
64            tables acg rub hos del pre / missing;
65            title 'Frequencies of the ACG produced variables';
66          run;

NOTE: There were 2304926 observations read from the data set WORK.ACGDAT1.
NOTE: The PROCEDURE FREQ printed pages 1-3.
NOTE: PROCEDURE FREQ used:
      real time           52.06 seconds
      cpu time            11.72 seconds


66        !
67
68          *define flags for selected ACGs;
69          data acgpred;
70            set acgdat1;
71            if acg='4220' then acg4220='1'; else acg4220='0';
72            if acg='4330' then acg4330='1'; else acg4330='0';
73            if acg='4420' then acg4420='1'; else acg4420='0';
74            if acg='4430' then acg4430='1'; else acg4430='0';
75            if acg='4510' then acg4510='1'; else acg4510='0';
76            if acg='4520' then acg4520='1'; else acg4520='0';
77            if acg='4610' then acg4610='1'; else acg4610='0';
78            if acg='4620' then acg4620='1'; else acg4620='0';
79            if acg='4730' then acg4730='1'; else acg4730='0';
80            if acg='4830' then acg4830='1'; else acg4830='0';
```

```
81           if acg='4910' then acg4910='1'; else acg4910='0';
82           if acg='4920' then acg4920='1'; else acg4920='0';
83           if acg='4930' then acg4930='1'; else acg4930='0';
84           if acg='4940' then acg4940='1'; else acg4940='0';
85           if acg='5010' then acg5010='1'; else acg5010='0';
86           if acg='5020' then acg5020='1'; else acg5020='0';
87           if acg='5030' then acg5030='1'; else acg5030='0';
88           if acg='5040' then acg5040='1'; else acg5040='0';
89           if acg='5050' then acg5050='1'; else acg5050='0';
90           if acg='5060' then acg5060='1'; else acg5060='0';
91           if acg='5070' then acg5070='1'; else acg5070='0';
92           if acg='5320' then acg5320='1'; else acg5320='0';
93           if acg='5330' then acg5330='1'; else acg5330='0';
94           if acg='5340' then acg5340='1'; else acg5340='0';
95         *define pregnancy without delivery;
96         *Revised March 3, 2005 -- previously the statement read,  if (pre=1
&
96       !  del=2)..... ;
97           if (pre=1 & del=0) then prenodel='1'; else prenodel='0';
98         *flags for first three RUB bands;
99         /* Revised March 3, 2005 -- previously ACG produced RUB variable was
99       ! used -- but Chad Abrams pointed out that these RUBs are concurrently
99       ! calibrated and not prospective -- he suggested creating RUBs using
99       ! the definition on page 31 of documentation -- for r
100          if rub=1 then rub1='1'; else rub1='0';
101          if rub=2 then rub2='1'; else rub2='0';
102          if rub=3 then rub3='1'; else rub3='0';
103        */
104        if acg in ('0100' '0200' '0300' '1100' '1200' '1600' '5100' '5110'
104      ! '5200' '9900') then rub1='1';
105        else rub1='0';
106        if acg in ('0400' '0500' '0600' '0900' '1000' '1300' '1800' '1900'
106      ! '2000' '2100'
107        '2200' '2300' '2400' '2500' '2800' '2900' '3000' '3100' '3400'
'3900'
107      !  '4000' '1711' '1721' '1731'
108        '1741') then rub2='1';
109        else rub2='0';
110        if acg in ('0700' '0800' '1400' '1500' '2600' '2700' '3200' '3300'
110      ! '3500' '3600'
111        '3700' '3800' '4100' '4210' '4310' '4320' '4410' '4710' '4720'
'4810'
111      !  '4820' '5310' '1751' '1761'
112        '1771') then rub3='1';
113        else rub3='0';
114        *hos variable already in correct form (1=person has conditions
likely
114      !  to lead to hospitalization, 0=no);
115          keep pid acg4: acg5: prenodel rub1-rub3 hos;
116        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGDAT1.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 30 variables.
NOTE: DATA statement used:
      real time           3:31.67
      cpu time            38.00 seconds

116      !
117
118        proc sort data=acgpred; by pid; run;
```

```
NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 30 variables.
NOTE: PROCEDURE SORT used:
      real time           4:06.79
      cpu time            33.29 seconds


118      !
119        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGDAT1  DATA       111738880   04MAR2005:15:52:00
2  ACGPRED  DATA       109158400   04MAR2005:16:00:31
3  FORMATS  CATALOG        28672   04MAR2005:15:49:17
4  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
119      !                              delete acgdat1; run;

NOTE: Deleting WORK.ACGDAT1 (memtype=DATA).
119      !
120
121        *define selected EDCs;
122        *the EDC input file format is "skinny" with one EDC per records --
so
122      !  potentially a large number of records per person;
123        *also there are 73 EDCs flags to create;
124        *first the split the EDC file into 3, each covering 28/22/23
distinct
124      !  EDCs;

NOTE: PROCEDURE DATASETS used:
      real time           1.31 seconds
      cpu time            0.08 seconds


125        data edc1 edc2 edc3;
126          infile '/data/dod1/saswork/dod2acgbedc.txt';
127          input edc $ 1-5 pid $ 12-29;
128          if edc in ('ADM04' 'ALL04' 'ALL05' 'ALL06' 'CAR03' 'CAR04' 'CAR05'
128      ! 'CAR06' 'CAR07' 'CAR09'
129            'CAR14' 'CAR15' 'END02' 'END09' 'END08' 'END07' 'END06' 'EYE13'
129      ! 'GAS02' 'GAS05' 'GAS06' 'GAS12'
130            'GSI08' 'GSU11' 'GSU13' 'GSU14' 'GTC01' 'GUR04') then output edc1;
131            else if edc in ('HEM01' 'HEM05' 'HEM06' 'HEM07' 'INF04' 'MAL02'
131      ! 'MAL03' 'MAL04' 'MAL06' 'MAL07'
132            'MAL08' 'MAL09' 'MAL10' 'MAL11' 'MAL12' 'MAL13' 'MAL14' 'MAL15'
132      ! 'MAL16' 'MAL18' 'MUS10' 'MUS14')
133            then output edc2;
134            else if edc in ('NUR05' 'NUR07' 'NUR08' 'NUR09' 'NUR12' 'NUR15'
```

```
134      ! 'NUR16' 'NUR17' 'NUR18' 'NUR19'
135           'NUT02' 'PSY01' 'PSY03' 'PSY05' 'PSY07' 'PSY08' 'PSY09' 'REC01'
135      ! 'REC03' 'REN01' 'RES03' 'RES04'
136           'RES09') then output edc3;
137        run;

NOTE: The infile '/data/dod1/saswork/dod2acgbedc.txt' is:
      File Name=/data/dod1/saswork/dod2acgbedc.txt,
      Owner Name=amresh,Group Name=sas,
      Access Permission=rw-r--r--,
      File Size (bytes)=301520220

NOTE: 10050674 records were read from the infile
      '/data/dod1/saswork/dod2acgbedc.txt'.
      The minimum record length was 29.
      The maximum record length was 29.
NOTE: The data set WORK.EDC1 has 510185 observations and 2 variables.
NOTE: The data set WORK.EDC2 has 230597 observations and 2 variables.
NOTE: The data set WORK.EDC3 has 361563 observations and 2 variables.
NOTE: DATA statement used:
      real time           15:19.98
      cpu time            3:52.47


137      !
138        *now create the edc flags in each of these files;
139        *edc1;
140        *turn 'long' skinny file into 'wide' format with one record per
140      ! person;
141        proc sort data=edc1; by pid edc;
142        *note that following command creates variables called col1, col2,
...
142      !  col12 each of which contains a distinct edc;

NOTE: There were 510185 observations read from the data set WORK.EDC1.
NOTE: The data set WORK.EDC1 has 510185 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           18.74 seconds
      cpu time            4.35 seconds


143        proc transpose data=edc1 out=edc1b;
143      !
144           var edc;
144      !
145           by pid;
145      !
146        run;

NOTE: There were 510185 observations read from the data set WORK.EDC1.
NOTE: The data set WORK.EDC1B has 360600 observations and 15 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           2:09.07
      cpu time            28.70 seconds


146      !
147        proc contents varnum; run;
```

```
NOTE: PROCEDURE CONTENTS used:
      real time           0.14 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 4.

147      !
148        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1     DATA        11853824   04MAR2005:16:16:12
3  EDC1B    DATA        33202176   04MAR2005:16:18:21
4  EDC2     DATA         5365760   04MAR2005:16:15:53
5  EDC3     DATA         8404992   04MAR2005:16:15:53
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
148      !                              delete edc1; run;

NOTE: Deleting WORK.EDC1 (memtype=DATA).
148      !

NOTE: PROCEDURE DATASETS used:
      real time           0.17 seconds
      cpu time            0.02 seconds


149        data edc1c;
150          set edc1b;
151          array cols{13} col1-col13;
152           do i=1 to 13;
153              if cols{i}='ADM04' then edcadm04='1';
154              if cols{i}='ALL04' then edcall04='1';
155              if cols{i}='ALL05' then edcall05='1';
156              if cols{i}='ALL06' then edcall06='1';
157              if cols{i}='CAR03' then edccar03='1';
158              if cols{i}='CAR04' then edccar04='1';
159              if cols{i}='CAR05' then edccar05='1';
160              if cols{i}='CAR06' then edccar06='1';
161              if cols{i}='CAR07' then edccar07='1';
162              if cols{i}='CAR09' then edccar09='1';
163              if cols{i}='CAR14' then edccar14='1';
164              if cols{i}='CAR15' then edccar15='1';
165              if cols{i}='END02' then edcend02='1';
166              if cols{i}='END09' then edcend09='1';
167              if cols{i}='END08' then edcend08='1';
```

```
168              if cols{i}='END07' then edcend07='1';
169              if cols{i}='END06' then edcend06='1';
170              if cols{i}='EYE13' then edceye13='1';
171              if cols{i}='GAS02' then edcgas02='1';
172              if cols{i}='GAS05' then edcgas05='1';
173              if cols{i}='GAS06' then edcgas06='1';
174              if cols{i}='GAS12' then edcgas12='1';
175              if cols{i}='GSI08' then edcgsi08='1';
176              if cols{i}='GSU11' then edcgsu11='1';
177              if cols{i}='GSU13' then edcgsu13='1';
178              if cols{i}='GSU14' then edcgsu14='1';
179              if cols{i}='GTC01' then edcgtc01='1';
180              if cols{i}='GUR04' then edcgur04='1';
181           end;
182           keep pid edc:;
183        run;

NOTE: There were 360600 observations read from the data set WORK.EDC1B.
NOTE: The data set WORK.EDC1C has 360600 observations and 29 variables.
NOTE: DATA statement used:
      real time           3:03.02
      cpu time            44.42 seconds


183      !
184        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1B    DATA        33202176   04MAR2005:16:18:21
3  EDC1C    DATA        16703488   04MAR2005:16:21:24
4  EDC2     DATA         5365760   04MAR2005:16:15:53
5  EDC3     DATA         8404992   04MAR2005:16:15:53
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
184      !                              delete edc1b; run;

NOTE: Deleting WORK.EDC1B (memtype=DATA).
184      !

NOTE: PROCEDURE DATASETS used:
      real time           1.14 seconds
      cpu time            0.07 seconds


185        proc sort data=edc1c; by pid; run;
```

```
NOTE: There were 360600 observations read from the data set WORK.EDC1C.
NOTE: The data set WORK.EDC1C has 360600 observations and 29 variables.
NOTE: PROCEDURE SORT used:
      real time           19.43 seconds
      cpu time            3.55 seconds


185      !
186        *edc2;
187        *turn 'long' skinny file into 'wide' format with one record per
187      ! person;
188        proc sort data=edc2; by pid edc;
189        *note that following command creates variables called col1, col2,
...
189      !  col8 each of which contains a distinct edc;

NOTE: There were 230597 observations read from the data set WORK.EDC2.
NOTE: The data set WORK.EDC2 has 230597 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           10.23 seconds
      cpu time            1.91 seconds


190        proc transpose data=edc2 out=edc2b;
190      !
191           var edc;
191      !
192           by pid;
192      !
193        run;

NOTE: There were 230597 observations read from the data set WORK.EDC2.
NOTE: The data set WORK.EDC2B has 217972 observations and 10 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           1:11.79
      cpu time            16.28 seconds


193      !
194        proc contents varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.09 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 5.

194      !
195        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
```

```
                    -----Directory-----

File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1C    DATA        16703488   04MAR2005:16:21:45
3  EDC2     DATA         5365760   04MAR2005:16:21:55
4  EDC2B    DATA        14532608   04MAR2005:16:23:08
5  EDC3     DATA         8404992   04MAR2005:16:15:53
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
195      !                           delete edc2; run;

NOTE: Deleting WORK.EDC2 (memtype=DATA).
195      !

NOTE: PROCEDURE DATASETS used:
      real time           0.17 seconds
      cpu time            0.02 seconds


196        data edc2c;
197          set edc2b;
198          array cols{8} col1-col8;
199           do i=1 to 8;
200              if cols{i}='HEM01' then edchem01='1';
201              if cols{i}='HEM05' then edchem05='1';
202              if cols{i}='HEM06' then edchem06='1';
203              if cols{i}='HEM07' then edchem07='1';
204              if cols{i}='INF04' then edcinf04='1';
205              if cols{i}='MAL02' then edcmal02='1';
206              if cols{i}='MAL03' then edcmal03='1';
207              if cols{i}='MAL04' then edcmal04='1';
208              if cols{i}='MAL06' then edcmal06='1';
209              if cols{i}='MAL07' then edcmal07='1';
210              if cols{i}='MAL08' then edcmal08='1';
211              if cols{i}='MAL09' then edcmal09='1';
212              if cols{i}='MAL10' then edcmal10='1';
213              if cols{i}='MAL11' then edcmal11='1';
214              if cols{i}='MAL12' then edcmal12='1';
215              if cols{i}='MAL13' then edcmal13='1';
216              if cols{i}='MAL14' then edcmal14='1';
217              if cols{i}='MAL15' then edcmal15='1';
218              if cols{i}='MAL16' then edcmal16='1';
219              if cols{i}='MAL18' then edcmal18='1';
220              if cols{i}='MUS10' then edcmus10='1';
221              if cols{i}='MUS14' then edcmus14='1';
222           end;
223           keep pid edc:;
224        run;

NOTE: There were 217972 observations read from the data set WORK.EDC2B.
NOTE: The data set WORK.EDC2C has 217972 observations and 23 variables.
NOTE: DATA statement used:
      real time           59.36 seconds
```

```
      cpu time            14.07 seconds


224      !
225        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1C    DATA        16703488   04MAR2005:16:21:45
3  EDC2B    DATA        14532608   04MAR2005:16:23:08
4  EDC2C    DATA         8814592   04MAR2005:16:24:08
5  EDC3     DATA         8404992   04MAR2005:16:15:53
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
225      !                           delete edc2b; run;

NOTE: Deleting WORK.EDC2B (memtype=DATA).
225      !

NOTE: PROCEDURE DATASETS used:
      real time           0.17 seconds
      cpu time            0.04 seconds


226        proc sort data=edc2c; by pid; run;

NOTE: There were 217972 observations read from the data set WORK.EDC2C.
NOTE: The data set WORK.EDC2C has 217972 observations and 23 variables.
NOTE: PROCEDURE SORT used:
      real time           9.58 seconds
      cpu time            1.99 seconds


226      !
227        *edc3;
228        *turn 'long' skinny file into 'wide' format with one record per
228      ! person;
229        proc sort data=edc3; by pid edc;
230        *note that following command creates variables called col1, col2,
...
230      !  col 11 each of which contains a distinct edc;

NOTE: There were 361563 observations read from the data set WORK.EDC3.
NOTE: The data set WORK.EDC3 has 361563 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           14.67 seconds
      cpu time            2.96 seconds
```

```
231        proc transpose data=edc3 out=edc3b;
231      !
232           var edc;
232      !
233           by pid;
233      !
234        run;

NOTE: There were 361563 observations read from the data set WORK.EDC3.
NOTE: The data set WORK.EDC3B has 270979 observations and 13 variables.
NOTE: PROCEDURE TRANSPOSE used:
      real time           1:33.99
      cpu time            21.34 seconds


234      !
235        proc contents varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.11 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 6.

235      !
236        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1C    DATA        16703488   04MAR2005:16:21:45
3  EDC2C    DATA         8814592   04MAR2005:16:24:17
4  EDC3     DATA         8404992   04MAR2005:16:24:32
5  EDC3B    DATA        22216704   04MAR2005:16:26:06
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
236      !                           delete edc3; run;

NOTE: Deleting WORK.EDC3 (memtype=DATA).
236      !

NOTE: PROCEDURE DATASETS used:
      real time           0.17 seconds
      cpu time            0.04 seconds
```

```
237        data edc3c;
238          set edc3b;
239          array cols{11} col1-col11;
240           do i=1 to 11;
241              if cols{i}='NUR05' then edcnur05='1';
242              if cols{i}='NUR07' then edcnur07='1';
243              if cols{i}='NUR08' then edcnur08='1';
244              if cols{i}='NUR09' then edcnur09='1';
245              if cols{i}='NUR12' then edcnur12='1';
246              if cols{i}='NUR15' then edcnur15='1';
247              if cols{i}='NUR16' then edcnur16='1';
248              if cols{i}='NUR17' then edcnur17='1';
249              if cols{i}='NUR18' then edcnur18='1';
250              if cols{i}='NUR19' then edcnur19='1';
251              if cols{i}='NUT02' then edcnut02='1';
252              if cols{i}='PSY01' then edcpsy01='1';
253              if cols{i}='PSY03' then edcpsy03='1';
254              if cols{i}='PSY05' then edcpsy05='1';
255              if cols{i}='PSY07' then edcpsy07='1';
256              if cols{i}='PSY08' then edcpsy08='1';
257              if cols{i}='PSY09' then edcpsy09='1';
258              if cols{i}='REC01' then edcrec01='1';
259              if cols{i}='REC03' then edcrec03='1';
260              if cols{i}='REN01' then edcren01='1';
261              if cols{i}='RES03' then edcres03='1';
262              if cols{i}='RES04' then edcres04='1';
263              if cols{i}='RES09' then edcres09='1';
264           end;
265           keep pid edc:;
266        run;

NOTE: There were 270979 observations read from the data set WORK.EDC3B.
NOTE: The data set WORK.EDC3C has 270979 observations and 24 variables.
NOTE: DATA statement used:
      real time           1:39.01
      cpu time            24.51 seconds


266      !
267        proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       109158400   04MAR2005:16:00:31
2  EDC1C    DATA        16703488   04MAR2005:16:21:45
3  EDC2C    DATA         8814592   04MAR2005:16:24:17
```

```
13 The SAS System                                    15:49 Friday, March 4, 2005

#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
4  EDC3B    DATA        22216704   04MAR2005:16:26:06
5  EDC3C    DATA        11223040   04MAR2005:16:27:46
6  FORMATS  CATALOG        28672   04MAR2005:15:49:17
7  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
267      !                           delete edc3b; run;

NOTE: Deleting WORK.EDC3B (memtype=DATA).
267      !

NOTE: PROCEDURE DATASETS used:
      real time           0.53 seconds
      cpu time            0.06 seconds


268        proc sort data=edc3c; by pid; run;

NOTE: There were 270979 observations read from the data set WORK.EDC3C.
NOTE: The data set WORK.EDC3C has 270979 observations and 24 variables.
NOTE: PROCEDURE SORT used:
      real time           13.19 seconds
      cpu time            2.49 seconds


268      !
269        *merge all EDCs into acgpred;
270        *merge edc1c;
271        data acgpred;
272          merge acgpred (in=infirst) edc1c (in=insecond);
273          by pid;
274          first = infirst;
275          second = insecond;
276        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: There were 360600 observations read from the data set WORK.EDC1C.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 60 variables.
NOTE: DATA statement used:
      real time           3:28.16
      cpu time            31.85 seconds


276      !
277        proc freq; tables first*second / missing; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The PROCEDURE FREQ printed page 7.
NOTE: PROCEDURE FREQ used:
      real time           40.30 seconds
      cpu time            9.46 seconds


277      !
278        data acgpred;
279          set acgpred;
280          drop first second;
281        run;
```

```
NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 58 variables.
NOTE: DATA statement used:
      real time           2:03.44
      cpu time            17.40 seconds


281      !
282        proc sort data=acgpred; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 58 variables.
NOTE: PROCEDURE SORT used:
      real time           4:29.61
      cpu time            39.68 seconds


282      !
283        *merge edc2c;
284        data acgpred;
285          merge acgpred (in=infirst) edc2c (in=insecond);
286          by pid;
287          first = infirst;
288          second = insecond;
289        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: There were 217972 observations read from the data set WORK.EDC2C.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 82 variables.
NOTE: DATA statement used:
      real time           3:58.37
      cpu time            34.61 seconds


289      !
290        proc freq; tables first*second / missing; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The PROCEDURE FREQ printed page 8.
NOTE: PROCEDURE FREQ used:
      real time           37.45 seconds
      cpu time            9.64 seconds


290      !
291        data acgpred;
292          set acgpred;
293          drop first second;
294        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 80 variables.
NOTE: DATA statement used:
      real time           2:23.40
      cpu time            19.94 seconds
```

```
294      !
295        proc sort data=acgpred; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 80 variables.
NOTE: PROCEDURE SORT used:
      real time           5:26.62
      cpu time            45.90 seconds


295      !
296        *merge edc3c;
297        proc sort data=acgpred; by pid; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


297      !
298        data acgpred;
299          merge acgpred (in=infirst) edc3c (in=insecond);
300          by pid;
301          first = infirst;
302          second = insecond;
303        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: There were 270979 observations read from the data set WORK.EDC3C.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 105 variables.
NOTE: DATA statement used:
      real time           4:21.13
      cpu time            39.18 seconds


303      !
304        proc freq; tables first*second / missing; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The PROCEDURE FREQ printed page 9.
NOTE: PROCEDURE FREQ used:
      real time           51.55 seconds
      cpu time            12.76 seconds


304      !
305        data acgpred;
306          set acgpred;
307          drop first second;
308        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 103 variables.
NOTE: DATA statement used:
      real time           2:43.38
      cpu time            22.87 seconds
```

```
308      !
309        proc sort data=acgpred; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED has 2304926 observations and 103 variables.
NOTE: PROCEDURE SORT used:
      real time           6:23.91
      cpu time            51.12 seconds


309      !
310        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED  DATA       277700608   04MAR2005:17:05:27
2  EDC1C    DATA        16703488   04MAR2005:16:21:45
3  EDC2C    DATA         8814592   04MAR2005:16:24:17
4  EDC3C    DATA        11223040   04MAR2005:16:28:00
5  FORMATS  CATALOG        28672   04MAR2005:15:49:17
6  REGSTRY  ITEMSTOR       32768   04MAR2005:15:49:16
310      !
311
311      !  delete edc1c edc2c edc3c;
312        run;

NOTE: Deleting WORK.EDC1C (memtype=DATA).
NOTE: Deleting WORK.EDC2C (memtype=DATA).
NOTE: Deleting WORK.EDC3C (memtype=DATA).
312      !
313
314        *now get the rest of the variables (age, gender, FY2002 expenditure)
314      ! from saswork.dod2charb;
315        * also get datatype variable to distinguish the estimation and
315      ! validation sample;

NOTE: PROCEDURE DATASETS used:
      real time           0.28 seconds
      cpu time            0.04 seconds


316        data dod2charbsub;
317          set saswork.dod2charb (keep = recipno age male datatype expyr
317      ! expyrt25 expyrt50 rename=(recipno=pid));
318          *create a revised version of expenditure variables (expyr,
expyrt25
318      ! and expyrt50) that have missing entries for validation sample;
```

```
319          if datatype=1 then expyrs = expyr;
320          if datatype=1 then expyrt25s = expyrt25;
321          if datatype=1 then expyrt50s = expyrt50;
322          label datatype='1=fitting sample, 2=Validation sample';
323          label expyr = 'Total health care expenditures, FY2002';
324          label expyrt25 = 'expyr topped at $25K';
325          label expyrt50 = 'expyr topped at $50K';
326          label expyrs = 'expyr defined only for estimation sample';
327          label expyrt25s = 'expyrt25 defined only for estimation sample';
328          label expyrt50s = 'expyrt50 defined only for estimation sample';
329          *create age and gender indicators;
330          if male=1 & age LE 1 then m0_1='1'; else m0_1='0';
331          if male=1 & (age GE 2) & (age LE 10) then m2_10='1'; else
m2_10='0';
331      !
332          if male=1 & (age GE 11) & (age LE 17) then m11_17='1'; else
332      ! m11_17='0';
333          if male=1 & (age GE 18) & (age LE 25) then m18_25='1'; else
333      ! m18_25='0';
334          if male=1 & (age GE 26) & (age LE 30) then m26_30='1'; else
334      ! m26_30='0';
335          if male=1 & (age GE 31) & (age LE 35) then m31_35='1'; else
335      ! m31_35='0';
336          if male=1 & (age GE 36) & (age LE 40) then m36_40='1'; else
336      ! m36_40='0';
337          if male=1 & (age GE 41) & (age LE 45) then m41_45='1'; else
337      ! m41_45='0';
338          if male=1 & (age GE 46) & (age LE 50) then m46_50='1'; else
338      ! m46_50='0';
339          if male=1 & (age GE 51) & (age LE 55) then m51_55='1'; else
339      ! m51_55='0';
340          if male=1 & (age GE 56) & (age LE 60) then m56_60='1'; else
340      ! m56_60='0';
341          if male=1 & (age GE 61) & (age LE 64) then m61_64='1'; else
341      ! m61_64='0';
342          if male=0 & age LE 1 then f0_1='1'; else f0_1='0';
343          if male=0 & (age GE 2) & (age LE 10) then f2_10='1'; else
f2_10='0';
343      !
344          if male=0 & (age GE 11) & (age LE 17) then f11_17='1'; else
344      ! f11_17='0';
345          if male=0 & (age GE 18) & (age LE 25) then f18_25='1'; else
345      ! f18_25='0';
346          if male=0 & (age GE 26) & (age LE 30) then f26_30='1'; else
346      ! f26_30='0';
347          if male=0 & (age GE 31) & (age LE 35) then f31_35='1'; else
347      ! f31_35='0';
348          if male=0 & (age GE 36) & (age LE 40) then f36_40='1'; else
348      ! f36_40='0';
349          if male=0 & (age GE 41) & (age LE 45) then f41_45='1'; else
349      ! f41_45='0';
350          if male=0 & (age GE 46) & (age LE 50) then f46_50='1'; else
350      ! f46_50='0';
351          if male=0 & (age GE 51) & (age LE 55) then f51_55='1'; else
351      ! f51_55='0';
352          if male=0 & (age GE 56) & (age LE 60) then f56_60='1'; else
352      ! f56_60='0';
353          if male=0 & (age GE 61) & (age LE 64) then f61_64='1'; else
353      ! f61_64='0';
354          drop male age;
```

```
355        run;

NOTE: There were 2304926 observations read from the data set SASWORK.DOD2CHARB.
NOTE: The data set WORK.DOD2CHARBSUB has 2304926 observations and 32 variables.
NOTE: DATA statement used:
      real time           2:55.99
      cpu time            25.85 seconds


355      !
356        proc sort data=dod2charbsub; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.DOD2CHARBSUB.
NOTE: The data set WORK.DOD2CHARBSUB has 2304926 observations and 32 variables.
NOTE: PROCEDURE SORT used:
      real time           5:42.62
      cpu time            46.24 seconds


356      !
357
358        *merge the two datasets;
359        data acgpred2;
360          merge dod2charbsub (in=infirst) acgpred (in=insecond);
361          by pid;
362          first = infirst;
363          second = insecond;
364        run;

NOTE: There were 2304926 observations read from the data set WORK.DOD2CHARBSUB.
NOTE: There were 2304926 observations read from the data set WORK.ACGPRED.
NOTE: The data set WORK.ACGPRED2 has 2304926 observations and 136 variables.
NOTE: DATA statement used:
      real time           6:51.07
      cpu time            59.64 seconds


364      !
365        proc freq data=acgpred2; tables first*second; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED2.
NOTE: The PROCEDURE FREQ printed page 10.
NOTE: PROCEDURE FREQ used:
      real time           46.00 seconds
      cpu time            12.41 seconds


365      !
366        proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
```

-----Directory-----

File Size (bytes): 512

#	Name	Memtype	File Size	Last Modified
1	ACGPRED	DATA	277700608	04MAR2005:17:05:27
2	ACGPRED2	DATA	501334016	04MAR2005:17:20:58
3	DOD2CHARBSUB	DATA	240558080	04MAR2005:17:14:07
4	FORMATS	CATALOG	28672	04MAR2005:15:49:17
5	REGSTRY	ITEMSTOR	32768	04MAR2005:15:49:16

```
366      !
367
367      !  delete dod2charbsub acgpred;
368        run;

NOTE: Deleting WORK.DOD2CHARBSUB (memtype=DATA).
NOTE: Deleting WORK.ACGPRED (memtype=DATA).
368      !
369
370        *list the variables and note their ordering;

NOTE: PROCEDURE DATASETS used:
      real time           3.85 seconds
      cpu time            0.39 seconds


371        proc contents data=acgpred2 varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.11 seconds
      cpu time            0.02 seconds

NOTE: The PROCEDURE CONTENTS printed pages 11-14.

371      !
372
373        *the above proc contents output indicates that all the regression
373      ! covariates are located in sequence, from m0_1 to edcres09;
374        *following converts missing to 0 among the covariates;
375        data acgpred3;
376          set acgpred2;
377          array varlist{126}  m0_1--edcres09;
378          do i=1 to 126;
379                *the edc variables need to have 0 in place of . (missing);
380                if varlist{i}='' then varlist{i}='0';
381          end;
382          drop first second i;
383        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED2.
NOTE: The data set WORK.ACGPRED3 has 2304926 observations and 134 variables.
NOTE: DATA statement used:
      real time           16:20.58
      cpu time            3:23.22
```

```
383      !
384
385        *Disease incidence rates & age/gender frequencies;
386        proc freq data=acgpred3;
387          tables m0_1--edcres09 / missing;
388          title "Disease incidence rates";
389        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED3.
NOTE: The PROCEDURE FREQ printed pages 15-32.
NOTE: PROCEDURE FREQ used:
      real time           16:20.29
      cpu time            4:14.65


389      !
390
391        *convert all indicator variables (char) to numeric;
392        *Reference age/gender and diagnoses groups -- we treat f61_64
(female
392      !  aged 61 to 64) and rub1 (lowest resource utilization band group) as
392      ! the reference categories in the prediction regression;
393        data acgpred4;
394          length indvar1-indvar124 3;
395          set acgpred3;
396          array oldvar{124}  m0_1--f56_60 hos--prenodel rub2--edcres09;
397          array newvar{124} indvar1-indvar124;
398          do i=1 to 124;
399                 if oldvar{i}='0' then newvar{i}=0;
400                 else if oldvar{i}='1' then newvar{i} = 1;
401          end;
402          *just to check if it worked following prints first 10 obs of
402      ! variables 13 (f0_1) and 103 (edcmus14);
403          if _N_<11 then put f0_1 indvar13 edcmus14 indvar103;
404          drop i m0_1--edcres09;
405        run;

0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0
NOTE: There were 2304926 observations read from the data set WORK.ACGPRED3.
NOTE: The data set WORK.ACGPRED4 has 2304926 observations and 132 variables.
NOTE: DATA statement used:
      real time           20:34.08
      cpu time            3:37.57


405      !
406
407        * list of all regressors;
408        %let xlist = indvar1-indvar124;
409
```

```
410        * run regression for adults;
411        proc reg data=acgpred4;
411      !
412           model expyrs expyrt25s expyrt50s = &xlist;
413           output out=acgpred5 (keep=pid datatype expyr expyrt25 expyrt50
413      ! pexp_acg pexpt25_acg pexpt50_acg) p=pexp_acg pexpt25_acg
pexpt50_acg;
413      !
414        run;

NOTE: 2304926 observations read.
NOTE: 500000 observations have missing values.
NOTE: 1804926 observations used in computations.
414      !

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED4.
NOTE: The data set WORK.ACGPRED5 has 2304926 observations and 8 variables.
NOTE: The PROCEDURE REG printed pages 33-53.
NOTE: PROCEDURE REG used:
      real time           14:16.82
      cpu time            3:14.56


415        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype     File Size  Last Modified
-----------------------------------------------------------
1  ACGPRED2  DATA        501334016   04MAR2005:17:20:58
2  ACGPRED3  DATA        466247680   04MAR2005:17:38:09
3  ACGPRED4  DATA       1037524992   04MAR2005:18:15:03
4  ACGPRED5  DATA        186966016   04MAR2005:18:29:20
5  FORMATS   CATALOG         28672   04MAR2005:15:49:17
6  REGSTRY   ITEMSTOR        32768   04MAR2005:15:49:16
415      !
416
416      !  delete acgpred2 acgpred3 acgpred4;
417        run;

NOTE: Deleting WORK.ACGPRED2 (memtype=DATA).
NOTE: Deleting WORK.ACGPRED3 (memtype=DATA).
NOTE: Deleting WORK.ACGPRED4 (memtype=DATA).
417      !
418
419        *save predicted expenditures from acg in new file saswork.dod3acgb;
420        *merge with dod2 (dod2 has individual characteristics);

NOTE: PROCEDURE DATASETS used:
      real time           7.04 seconds
```

```
      cpu time            1.56 seconds


421        proc sort data=acgpred5; by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED5.
NOTE: The data set WORK.ACGPRED5 has 2304926 observations and 8 variables.
NOTE: PROCEDURE SORT used:
      real time           4:47.27
      cpu time            41.95 seconds


421      !
422        proc sort data=saswork.dod2; by pid; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.09 seconds
      cpu time            0.00 seconds


422      !
423        data acgpred5b;
424          merge acgpred5 (in=infirst) saswork.dod2 (in=insecond keep=pid
424      ! recsvc rank rebencat catcode resreg urban recocc);
425          by pid;
426          if infirst=1 & insecond=1;
427        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED5.
NOTE: There were 2304926 observations read from the data set SASWORK.DOD2.
NOTE: The data set WORK.ACGPRED5B has 2304926 observations and 15 variables.
NOTE: DATA statement used:
      real time           4:27.85
      cpu time            55.23 seconds


427      !
428        proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name       Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED5   DATA       186966016   04MAR2005:18:34:15
2  ACGPRED5B  DATA       240558080   04MAR2005:18:38:43
3  FORMATS    CATALOG        28672   04MAR2005:15:49:17
4  REGSTRY    ITEMSTOR       32768   04MAR2005:15:49:16
428      !
```

```
23 The SAS System                                    15:49 Friday, March 4, 2005

429
429      !  delete acgpred5;
430        run;

NOTE: Deleting WORK.ACGPRED5 (memtype=DATA).
430      !
431
432        *need following new variables for predictive accuracy analysis;
433        *limit to validation sample;

NOTE: PROCEDURE DATASETS used:
      real time           2.28 seconds
      cpu time            0.24 seconds


434        data acgpred6;
435          set acgpred5b;
436          if datatype = 2;
437          abserr = abs(expyr-pexp_acg);
438          abserrt25 = abs(expyrt25-pexpt25_acg);
439          abserrt50 = abs(expyrt50-pexpt50_acg);
440          label abserr = 'Absolute prediction error for expyr';
441          label abserrt25 = 'Absolute prediction error for expyrt25';
442          label abserrt50 = 'Absolute prediction error for expyrt50';
443          label pexp_acg = 'ACG Predicted expenditures using expyr';
444          label pexpt25_acg = 'ACG Predicted expenditures using expyrt25';
445          label pexpt50_acg = 'ACG Predicted expenditures using expyrt50';
446        run;

NOTE: There were 2304926 observations read from the data set WORK.ACGPRED5B.
NOTE: The data set WORK.ACGPRED6 has 500000 observations and 18 variables.
NOTE: DATA statement used:
      real time           1:10.76
      cpu time            11.42 seconds


446      !
447        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB65100006404_genmed2
File Name:         /data/saswork/SAS_workB65100006404_genmed2
Inode Number:      4293312
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name       Memtype    File Size  Last Modified
-----------------------------------------------------
1  ACGPRED5B  DATA       240558080   04MAR2005:18:38:43
2  ACGPRED6   DATA        64528384   04MAR2005:18:39:56
3  FORMATS    CATALOG        28672   04MAR2005:15:49:17
4  REGSTRY    ITEMSTOR       32768   04MAR2005:15:49:16
447      !
448
```

```
448      !  delete acgpred5b;
449        run;

NOTE: Deleting WORK.ACGPRED5B (memtype=DATA).
449      !
450
451        * create absolute deviation from mean variable;

NOTE: PROCEDURE DATASETS used:
      real time           2.09 seconds
      cpu time            0.27 seconds


452        proc sql;
452      !
453          create table acgpred7 as
453      !
454          select *,  abs(expyr - mean(expyr)) as absmdev,
455           abs(expyrt25 - mean(expyrt25)) as absmdevt25,
456           abs(expyrt50 - mean(expyrt50)) as absmdevt50
457           from acgpred6;
NOTE: The query requires remerging summary statistics back with the original
      data.
NOTE: Table WORK.ACGPRED7 created, with 500000 rows and 21 columns.

457      !
458        quit;
NOTE: PROCEDURE SQL used:
      real time           1:50.90
      cpu time            16.10 seconds


458      !
459        data acgpred7;
460          set acgpred7;
461          label absmdev='Absolute mean deviation for expyr';
462          label absmdevt25='Absolute mean deviation for expyrt25';
463          label absmdevt50='Absolute mean deviation for expyrt50';
464        run;

NOTE: There were 500000 observations read from the data set WORK.ACGPRED7.
NOTE: The data set WORK.ACGPRED7 has 500000 observations and 21 variables.
NOTE: DATA statement used:
      real time           46.26 seconds
      cpu time            5.56 seconds


464      !
465
466        proc contents data=acgpred7 varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.11 seconds
      cpu time            0.01 seconds

NOTE: The PROCEDURE CONTENTS printed page 54.

466      !
```

```
25 The SAS System                                    15:49 Friday, March 4, 2005

467
468        proc cport data=acgpred7 file='/data/dod1/saswork/dod3acgb.v8x';
run;


NOTE: Proc CPORT begins to transport data set WORK.ACGPRED7
NOTE: The data set contains 21 variables and 500000 observations.
      Logical record length is 152.
NOTE: PROCEDURE CPORT used:
      real time           2:09.36
      cpu time            25.01 seconds


468      !
469
470
471
472
473

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           2:55:30.04
      cpu time            33:36.21
```

Attachment A5_DCG
(task6_dcgpred.log)

```
1    The SAS System                              11:26 Friday, September 24, 2004

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.66 seconds
      cpu time            0.06 seconds

1          /*************************************/
2          /* task6_dcgpred.sas                 */
3          /* DoD1: Creates dataset with dcgp   */
4          /* actual & predicted expenditures   */
5          /* for validation sample             */
6          /* Jenn Fonda                                  */
7          /* First created: September 24, 2004 */
8          /* Last modified: September 24, 2004 */
9          /*************************************/
10
11         /* Using the grouped diagnoses indicators from DCG, this job
regresses second-year
12         expenditures on the group indicators for the estimation sample and
obtains predicted
13         expenditures for the validation sample.  These predicted amounts are
attached to dod2
14         to get dod3dcg.  Note that dod3dcg only contains validation sample
(1/2 million)*/
15
16         /* The input file dxcg.sas7bdat was created by dod_dxcg which was
run on the PC version
17         of DCG.  The output file is dod3dcg.sas7bdat. */
18
19         options ps=60 ls=80 nocenter compress=yes;
20         footnote '/data/dod1/saswork/fonda/task6_dcgpred.sas';
21         libname indcg '/data/dod1/saswork/fonda';
NOTE: Libref INDCG was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork/fonda
22         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
23         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
```

```
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.31 seconds
      cpu time            0.02 seconds
```

```
53
54          /* Create dataset that contains only idno, expenditures for
54        ! 2002(EXPEND2),
55          and the values for agesex and HCC001-HCC184 */
56          data dcg; set indcg.dxcg (keep=IDNO AGE SEX HCC001-HCC184 EXPEND2
56        ! datatype);
57             **  Create topcoded expenditures;
58                   EXPENDt25 = EXPEND2;
59                   if EXPEND2 GT 25000 then EXPENDt25 = 25000;
60                   EXPENDt50 = EXPEND2;
61                   if EXPEND2 GT 50000 then EXPENDt50 = 50000;
62             ** Create age and gender indicators;
63             if SEX=1 & AGE LE 1 then m0_1=1; else m0_1=0;
64             if SEX=1 & (AGE GE 2) & (AGE LE 10) then m2_10=1; else m2_10=0;
65             if SEX=1 & (AGE GE 11) & (AGE LE 17) then m11_17=1; else m11_17=0;
66             if SEX=1 & (AGE GE 18) & (AGE LE 25) then m18_25=1; else m18_25=0;
67             if SEX=1 & (AGE GE 26) & (AGE LE 30) then m26_30=1; else m26_30=0;
68             if SEX=1 & (AGE GE 31) & (AGE LE 35) then m31_35=1; else m31_35=0;
69             if SEX=1 & (AGE GE 36) & (AGE LE 40) then m36_40=1; else m36_40=0;
70             if SEX=1 & (AGE GE 41) & (AGE LE 45) then m41_45=1; else m41_45=0;
71             if SEX=1 & (AGE GE 46) & (AGE LE 50) then m46_50=1; else m46_50=0;
72             if SEX=1 & (AGE GE 51) & (AGE LE 55) then m51_55=1; else m51_55=0;
73             if SEX=1 & (AGE GE 56) & (AGE LE 60) then m56_60=1; else m56_60=0;
74             if SEX=1 & (age GE 61) & (AGE LE 64) then m61_64=1; else m61_64=0;
75             if SEX=2 & AGE LE 1 then f0_1=1; else f0_1=0;
76             if SEX=2 & (AGE GE 2) & (AGE LE 10) then f2_10=1; else f2_10=0;
77             if SEX=2 & (AGE GE 11) & (AGE LE 17) then f11_17=1; else f11_17=0;
78             if SEX=2 & (AGE GE 18) & (AGE LE 25) then f18_25=1; else f18_25=0;
79             if SEX=2 & (AGE GE 26) & (AGE LE 30) then f26_30=1; else f26_30=0;
80             if SEX=2 & (AGE GE 31) & (AGE LE 35) then f31_35=1; else f31_35=0;
81             if SEX=2 & (AGE GE 36) & (AGE LE 40) then f36_40=1; else f36_40=0;
82             if SEX=2 & (AGE GE 41) & (AGE LE 45) then f41_45=1; else f41_45=0;
83             if SEX=2 & (AGE GE 46) & (AGE LE 50) then f46_50=1; else f46_50=0;
84             if SEX=2 & (AGE GE 51) & (AGE LE 55) then f51_55=1; else f51_55=0;
85             if SEX=2 & (AGE GE 56) & (AGE LE 60) then f56_60=1; else f56_60=0;
86             if SEX=2 & (AGE GE 61) & (AGE LE 64) then f61_64=1; else f61_64=0;
87

NOTE: Character values have been converted to numeric
      values at the places given by: (Line):(Column).
      63:5   64:5   65:5   66:5   67:5   68:5   69:5   70:5   71:5   72:5
      73:5   74:5   75:5   76:5   77:5   78:5   79:5   80:5   81:5   82:5
      83:5   84:5   85:5   86:5
NOTE: There were 2304926 observations read from the data set INDCG.DXCG.
NOTE: The data set WORK.DCG has 2304926 observations and 215 variables.
NOTE: Compressing data set WORK.DCG decreased size by 89.57 percent.
      Compressed is 4713 pages; un-compressed would require 45196 pages.
NOTE: DATA statement used:
      real time           11:45.11
      cpu time            11:26.18


88          data dcg2; set dcg;
89             ** Create new variable, expyr, where expyr=EXPEND2 if it's the
90               'fitting data' and missing if it's the 'validation data';
```

```
91           if datatype=1 then expyr=EXPEND2;
92           if datatype=1 then expyrt25=EXPENDt25;
93           if datatype=1 then expyrt50=EXPENDt50;
94         run;

NOTE: There were 2304926 observations read from the data set WORK.DCG.
NOTE: The data set WORK.DCG2 has 2304926 observations and 218 variables.
NOTE: Compressing data set WORK.DCG2 decreased size by 87.68 percent.
      Compressed is 5795 pages; un-compressed would require 47040 pages.
NOTE: DATA statement used:
      real time           3:33.85
      cpu time            2:54.71


95
96         /* since some regressors (age groups or diagnoses) may take only one
96       ! value for all data,
97         first identify these (note that regression is only on the fitting
97       ! sample) */
98         ** List of all regressors;
99         %let xlist = m0_1 m2_10 m11_17 m18_25 m26_30 m31_35 m36_40 m41_45
99       ! m46_50
100        m51_55 m56_60 m61_64 f0_1 f2_10 f11_17 f18_25 f26_30 f31_35 f36_40
100      ! f41_45
101        f46_50 f51_55 f56_60 f61_64 HCC001-HCC184;
102
103        proc means data=dcg2 (where=(datatype=2)) min max mean;
104          var &xlist;
105          title 'Means for validation sample';
106        run;

NOTE: There were 500000 observations read from the data set WORK.DCG2.
      WHERE datatype=2;
NOTE: The PROCEDURE MEANS printed pages 1-9.
NOTE: PROCEDURE MEANS used:
      real time           1:36.03
      cpu time            1:35.77


107
108        ** Freq of different diagnoses;
109        proc freq data=dcg2 (where=(datatype=2));
110          tables &xlist;
111          title 'Freq for validation sample';
112        run;

NOTE: There were 500000 observations read from the data set WORK.DCG2.
      WHERE datatype=2;
NOTE: The PROCEDURE FREQ printed pages 10-44.
NOTE: PROCEDURE FREQ used:
      real time           1:35.48
      cpu time            1:34.26


113
114        ** Run regression;
115        proc reg data=dcg2;
```

```
4 The SAS System                                11:26 Friday, September 24, 2004

116          model expyr expyrt25 expyrt50 = &xlist;
117          output out=dcg3 (keep=IDNO datatype EXPEND2 EXPENDt25 EXPENDt50
118          pexp_dcg pexpt25_dcg pexpt50_dcg) p=pexp_dcg pexpt25_dcg
pexpt50_dcg
118      ! ;
119        run;

NOTE: 2304926 observations read.
NOTE: 500000 observations have missing values.
NOTE: 1804926 observations used in computations.
120
121        ** Delete unnecessary files;

NOTE: There were 2304926 observations read from the data set WORK.DCG2.
NOTE: The data set WORK.DCG3 has 2304926 observations and 8 variables.
NOTE: Compressing data set WORK.DCG3 increased size by 8.37 percent.
      Compressed is 22709 pages; un-compressed would require 20955 pages.
NOTE: The PROCEDURE REG printed pages 45-95.
NOTE: PROCEDURE REG used:
      real time           6:55.44
      cpu time            5:55.61


122        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB3FF000000D6_genmed2
File Name:         /data/saswork/SAS_workB3FF000000D6_genmed2
Inode Number:      940224
Access Permission: rwxrwx---
Owner Name:        jfonda
File Size (bytes): 512


#  Name     Memtype   File Size  Last Modified
-----------------------------------------------------
1  DCG      DATA      193052672   24SEP2004:11:38:40
2  DCG2     DATA      237371392   24SEP2004:11:42:14
3  DCG3     DATA      186040320   24SEP2004:11:52:22
4  FORMATS  CATALOG       24576   24SEP2004:11:26:55
123
123      !  delete dcg2;
124        run;

NOTE: Deleting WORK.DCG2 (memtype=DATA).
125
126        ** Save predicted expenditures from dcg in new file saswork.dod3dcg;
127        ** Merge with dod2 (dod2 has individual characteristics);

NOTE: PROCEDURE DATASETS used:
      real time           0.30 seconds
      cpu time            0.20 seconds


128        proc sort data=dcg3 (rename=(IDNO=pid)); by pid; run;
```

```
NOTE: There were 2304926 observations read from the data set WORK.DCG3.
NOTE: The data set WORK.DCG3 has 2304926 observations and 8 variables.
NOTE: Compressing data set WORK.DCG3 increased size by 8.37 percent.
      Compressed is 22709 pages; un-compressed would require 20955 pages.
NOTE: PROCEDURE SORT used:
      real time           4:43.79
      cpu time            1:09.20


129        proc sort data=saswork.dod2; by pid; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.00 seconds
      cpu time            0.00 seconds


130
131        data dcg4;
132          merge dcg3 (in=infirst) saswork.dod2 (in=insecond);
133          by pid;
134          if infirst=1 & insecond=1;
135        run;

NOTE: There were 2304926 observations read from the data set WORK.DCG3.
NOTE: There were 2304926 observations read from the data set SASWORK.DOD2.
NOTE: The data set WORK.DCG4 has 2304926 observations and 32 variables.
NOTE: Compressing data set WORK.DCG4 decreased size by 26.65 percent.
      Compressed is 23161 pages; un-compressed would require 31575 pages.
NOTE: DATA statement used:
      real time           8:58.80
      cpu time            1:54.77


136
137        *post-estimation analysis;
138        *limit to validation sample;
139        data dcg5;
140          set dcg4;
141          abserr = abs(EXPEND2-pexp_dcg);
142          abserrt25 = abs(EXPENDt25-pexpt25_dcg);
143          abserrt50 = abs(EXPENDt50-pexpt50_dcg);
144          if datatype=2;
145        run;

NOTE: There were 2304926 observations read from the data set WORK.DCG4.
NOTE: The data set WORK.DCG5 has 500000 observations and 35 variables.
NOTE: Compressing data set WORK.DCG5 decreased size by 24.10 percent.
      Compressed is 5751 pages; un-compressed would require 7577 pages.
NOTE: DATA statement used:
      real time           1:25.69
      cpu time            44.92 seconds


146
147        * create absolute deviation from mean variable;
148        proc sql;
```

```
149          create table saswork.dod3dcg as
150          select *,  abs(EXPEND2 - mean(EXPEND2)) as absmdev,
151           abs(EXPENDt25 - mean(EXPENDt25)) as absmdevt25,
152           abs(EXPENDt50 - mean(EXPENDt50)) as absmdevt50
153          from dcg5;
NOTE: The query requires remerging summary statistics back with the original
      data.
NOTE: Compressing data set SASWORK.DOD3DCG decreased size by 22.03 percent.
      Compressed is 6498 pages; un-compressed would require 8334 pages.
NOTE: Table SASWORK.DOD3DCG created, with 500000 rows and 38 columns.

154        quit;
NOTE: PROCEDURE SQL used:
      real time           2:15.87
      cpu time            39.89 seconds


155
156        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_workB3FF000000D6_genmed2
File Name:         /data/saswork/SAS_workB3FF000000D6_genmed2
Inode Number:      940224
Access Permission: rwxrwx---
Owner Name:        jfonda
File Size (bytes): 512


#  Name     Memtype   File Size  Last Modified
-----------------------------------------------------
1  DCG      DATA      193052672   24SEP2004:11:38:40
2  DCG3     DATA      186040320   24SEP2004:11:57:06
3  DCG4     DATA      379478016   24SEP2004:12:06:05
4  DCG5     DATA       94232576   24SEP2004:12:07:31
5  FORMATS  CATALOG       24576   24SEP2004:11:26:55
157
157      !  delete dcg3 dcg4 dcg5;
158        run;

NOTE: Deleting WORK.DCG3 (memtype=DATA).
NOTE: Deleting WORK.DCG4 (memtype=DATA).
NOTE: Deleting WORK.DCG5 (memtype=DATA).
159

NOTE: PROCEDURE DATASETS used:
      real time           0.74 seconds
      cpu time            0.56 seconds


160        proc contents data=saswork.dod3dcg varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.10 seconds
      cpu time            0.03 seconds
```

NOTE: The PROCEDURE CONTENTS printed pages 96-97.

161

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
real time 42:53.58
cpu time 27:56.23

Attachment A5_CRG

```
1 The SAS System                          17:48 Thursday, June 23, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.

This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.


NOTE: SAS initialization used:
      real time           0.38 seconds
      cpu time            0.07 seconds

1          /*******************************/
2          /* task7_dod3valid5.sas         */
3          /* Project: DOD1                */
4          /* Created by: Jenn Fonda       */
5          /* First created: June 23, 2005 */
6          /* Last modified: June 23, 2005 */
7          /*******************************/
8
9          /* This program adds the CRG predicted costs to the dataset
that already contains predicted costs from DCG, ACG, and
10         CDPS as well as demographic variables.
11
12         Input files: dod3valid4.v8x, validation_3mdata.sas7bdat
13         Output file: dod3valid5
14         */
15
16         proc printto
print='/data/dod1/final_programs/task7_dod3valid5.log'; run;

NOTE: PROCEDURE PRINTTO used:
      real time           0.00 seconds
      cpu time            0.00 seconds

17
18         options ps=55 ls=80;
19         libname crg '/data/dod1/3m';
NOTE: Libref CRG was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/3m
20         libname dod '/data/dod1/saswork';
NOTE: Libref DOD was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
21
22         proc sort data=crg.validation_3mdata (keep=id eptot_demo
ep25_demo
```

```
22       ! ep50_demo
23         rename=(id=pid eptot_demo=pexp_crg ep25_demo=pexpt25_crg
23       ! ep50_demo=pexpt50_crg)) out=crgdata; by pid; run;

NOTE: There were 500000 observations read from the data set
      CRG.VALIDATION_3MDATA.
NOTE: The data set WORK.CRGDATA has 500000 observations and 4
variables.
NOTE: PROCEDURE SORT used:
      real time           16.87 seconds
      cpu time            7.13 seconds

24
25         proc cimport data=dod3valid4
25       ! infile='/data/dod1/saswork/dod3valid4.v8x'; run;

NOTE: Proc CIMPORT begins to create/update data set WORK.DOD3VALID4
NOTE: Data set contains 44 variables and 500000 observations.
      Logical record length is 320

NOTE: PROCEDURE CIMPORT used:
      real time           1:28.71
      cpu time            51.77 seconds

26
27         proc sort data=dod3valid4; by pid; run;

NOTE: There were 500000 observations read from the data set
WORK.DOD3VALID4.
NOTE: The data set WORK.DOD3VALID4 has 500000 observations and 44
variables.
NOTE: PROCEDURE SORT used:
      real time           2:41.95
      cpu time            23.96 seconds

28
29         ** Merge CRG data with dod3valid4;
30         data dod3valid5;
31          merge crgdata (in=one) dod3valid4 (in=two);
32          by pid;
33
34          incrg=one;
35          indod3valid4=two;
36         run;
NOTE: There were 500000 observations read from the data set
WORK.CRGDATA.
NOTE: There were 500000 observations read from the data set
WORK.DOD3VALID4.
NOTE: The data set WORK.DOD3VALID5 has 500000 observations and 49
variables.
NOTE: DATA statement used:
      real time           1:33.59
      cpu time            17.34 seconds
```

```
37
38         ** Check that all pids in CRG data match those in
dod3valid4--should
38       ! be a perfect match for 500,000 unique pid's;
39         proc freq data=dod3valid5;
40          title 'Check all pids match in crg & our data';
41          tables incrg*indod3valid4 / missing;
42         run;
```

Check all pids match in crg & our data

The FREQ Procedure

Table of incrg by indod3valid4

incrg	indod3valid4	
Frequency Percent Row Pct Col Pct	1	Total
1	500000 100.00 100.00 100.00	500000 100.00
Total	500000 100.00	500000 100.00

```
NOTE: There were 500000 observations read from the data set
WORK.DOD3VALID5.
NOTE: The PROCEDURE FREQ printed page 1.
NOTE: PROCEDURE FREQ used:
      real time           3.90 seconds
      cpu time            3.63 seconds


43
44         data dod.dod3valid5; set dod3valid5;
45          drop incrg indod3valid4;
46         run;

NOTE: There were 500000 observations read from the data set
WORK.DOD3VALID5.
NOTE: The data set DOD.DOD3VALID5 has 500000 observations and 47
variables.
NOTE: DATA statement used:
      real time           1:11.88
      cpu time            11.32 seconds
```

```
47

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           7:17.91
      cpu time            1:55.29
```

Attachment A6_1
(task6d_dod3valid.log)

```
1     The SAS System                                15:23 Tuesday, March 8, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.38 seconds
      cpu time            0.07 seconds

1          /*
2          task6d_dod3valid.sas
3          Amresh Hanchate / Jeanne Speckman
4          DOD1 Project
5          Dec 3, 2004
6          Last changed: Dec 20, 2004 (Amresh)
7
8          This program pools the results from the three risk adjustment models
that we have run (CDPS, ACG and DCG). It also
8        ! includes other variables that are used for cross-tabulation. It also
creates new expenditure category variables
8        ! (percentile expenditures) tha
9
10         NOTE THAT THIS DATASET IS FOR ONLY THE 0.5 MILLION VALIDATION
SAMPLE.
11
12         Input files: dod3cdpsb, dod3dcg, dod3acgb
13         Output: dod3valid
14
15          */
16
17         options ps=55 ls=80 nocenter formdlim=' ';
18
19         footnote '/data/dod1/saswork/analysis_groupr2.sas';
20         libname saswork '/data/dod1/saswork/';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
20       !
21         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
NOTE: Format COSTCAT has been output.
```

```
NOTE: PROCEDURE FORMAT used:
      real time           0.42 seconds
      cpu time            0.02 seconds


54
55         *dod3cdpsb is used as the base dataset as it already has most of the
55       ! covariates -- some unnecessary variables from this file are dropped;
56         data indata1;
57           set saswork.dod3cdpsb;
58           keep pid--age2 cost_tot01;
59         run;

NOTE: There were 500000 observations read from the data set SASWORK.DOD3CDPSB.
NOTE: The data set WORK.INDATA1 has 500000 observations and 21 variables.
NOTE: DATA statement used:
      real time           1:00.14
      cpu time            6.60 seconds


59       !
60         proc sort data=indata1; by pid; run;

NOTE: There were 500000 observations read from the data set WORK.INDATA1.
NOTE: The data set WORK.INDATA1 has 500000 observations and 21 variables.
NOTE: PROCEDURE SORT used:
      real time           1:30.88
      cpu time            12.35 seconds


60       !
61         proc means data=indata1 mean min max nmiss n;
62         var expyr;
63         run;

NOTE: There were 500000 observations read from the data set WORK.INDATA1.
NOTE: The PROCEDURE MEANS printed page 1.
NOTE: PROCEDURE MEANS used:
      real time           5.60 seconds
      cpu time            1.33 seconds


63       !
64
65         *add new variables from dod3dcg;
66         data indata2;
67           set saswork.dod3dcg;
68           keep pid pexp:;
69         run;

NOTE: There were 500000 observations read from the data set SASWORK.DOD3DCG.
NOTE: The data set WORK.INDATA2 has 500000 observations and 4 variables.
```

```
NOTE: DATA statement used:
      real time           39.00 seconds
      cpu time            8.81 seconds


69       !
70         proc sort data=indata2; by pid; run;

NOTE: There were 500000 observations read from the data set WORK.INDATA2.
NOTE: The data set WORK.INDATA2 has 500000 observations and 4 variables.
NOTE: PROCEDURE SORT used:
      real time           33.06 seconds
      cpu time            6.10 seconds


70       !
71         data outdata1;
72           merge indata1 (in=first) indata2 (in=second);
73           by pid;
74           infirst = first;
75           insecond = second;
76         run;

NOTE: There were 500000 observations read from the data set WORK.INDATA1.
NOTE: There were 500000 observations read from the data set WORK.INDATA2.
NOTE: The data set WORK.OUTDATA1 has 500000 observations and 26 variables.
NOTE: DATA statement used:
      real time           1:24.43
      cpu time            13.28 seconds


76       !
77         proc freq data=outdata1;
78           tables infirst*insecond;
79         run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA1.
NOTE: The PROCEDURE FREQ printed page 2.
NOTE: PROCEDURE FREQ used:
      real time           11.01 seconds
      cpu time            2.40 seconds


79       !
80         data outdata1;
81           set outdata1;
82           drop infirst insecond;
83         run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA1.
NOTE: The data set WORK.OUTDATA1 has 500000 observations and 24 variables.
NOTE: DATA statement used:
      real time           53.55 seconds
```

```
      cpu time            6.10 seconds


83       !
84         proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work0D1100000D40_genmed2
File Name:         /data/saswork/SAS_work0D1100000D40_genmed2
Inode Number:      4312192
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype  File Size  Last Modified
-----------------------------------------------------
1  FORMATS   CATALOG      28672   08MAR2005:15:23:10
2  INDATA1   DATA       68280320   08MAR2005:15:25:41
3  INDATA2   DATA       24248320   08MAR2005:15:26:59
4  OUTDATA1  DATA       80338944   08MAR2005:15:29:28
84       !                            delete indata1 indata2; run;

NOTE: Deleting WORK.INDATA1 (memtype=DATA).
NOTE: Deleting WORK.INDATA2 (memtype=DATA).
84       !

NOTE: PROCEDURE DATASETS used:
      real time           1.08 seconds
      cpu time            0.08 seconds


85         proc means data=outdata1 mean min max nmiss n;
86         var expyr;
87         run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA1.
NOTE: The PROCEDURE MEANS printed page 3.
NOTE: PROCEDURE MEANS used:
      real time           6.24 seconds
      cpu time            1.56 seconds


87       !
88
89
90         *add new variables from dod3acgb;
91         data indata3;
92           set saswork.dod3acgb;
93           keep pid pexp:;
94         run;
```

```
NOTE: There were 500000 observations read from the data set SASWORK.DOD3ACGB.
NOTE: The data set WORK.INDATA3 has 500000 observations and 4 variables.
NOTE: DATA statement used:
      real time           25.77 seconds
      cpu time            4.13 seconds


94       !
95         proc sort data=indata3; by pid; run;

NOTE: There were 500000 observations read from the data set WORK.INDATA3.
NOTE: The data set WORK.INDATA3 has 500000 observations and 4 variables.
NOTE: PROCEDURE SORT used:
      real time           30.29 seconds
      cpu time            5.95 seconds


95       !
96         data outdata2;
97           merge outdata1 (in=first) indata3 (in=second);
98           by pid;
99           infirst = first;
100          insecond = second;
101        run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA1.
NOTE: There were 500000 observations read from the data set WORK.INDATA3.
NOTE: The data set WORK.OUTDATA2 has 500000 observations and 29 variables.
NOTE: DATA statement used:
      real time           1:36.28
      cpu time            14.16 seconds


101      !
102        proc freq data=outdata2;
103          tables infirst*insecond;
104        run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA2.
NOTE: The PROCEDURE FREQ printed page 4.
NOTE: PROCEDURE FREQ used:
      real time           11.06 seconds
      cpu time            2.64 seconds


104      !
105        data outdata2;
106          set outdata2;
107          drop infirst insecond;
108        run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA2.
```

```
NOTE: The data set WORK.OUTDATA2 has 500000 observations and 27 variables.
NOTE: DATA statement used:
      real time           57.06 seconds
      cpu time            6.91 seconds


108      !
109        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work0D1100000D40_genmed2
File Name:         /data/saswork/SAS_work0D1100000D40_genmed2
Inode Number:      4312192
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


#  Name      Memtype  File Size  Last Modified
-----------------------------------------------------
1  FORMATS   CATALOG      28672   08MAR2005:15:23:10
2  INDATA3   DATA      24248320   08MAR2005:15:30:32
3  OUTDATA1  DATA      80338944   08MAR2005:15:29:28
4  OUTDATA2  DATA      93118464   08MAR2005:15:33:16
109      !                            delete indata3 outdata1; run;

NOTE: Deleting WORK.INDATA3 (memtype=DATA).
NOTE: Deleting WORK.OUTDATA1 (memtype=DATA).
109      !

NOTE: PROCEDURE DATASETS used:
      real time           0.67 seconds
      cpu time            0.09 seconds


110        proc contents data=outdata2 varnum; run;

NOTE: PROCEDURE CONTENTS used:
      real time           0.15 seconds
      cpu time            0.02 seconds

NOTE: The PROCEDURE CONTENTS printed pages 5-6.

110      !
111        proc means data=outdata2 mean min max nmiss n;
112        var expyr;
113        run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA2.
NOTE: The PROCEDURE MEANS printed page 7.
NOTE: PROCEDURE MEANS used:
```

```
      real time           6.98 seconds
      cpu time            1.80 seconds


113      !
114
115        *for our predictive accuracy analysis by subgroups we also need to
115      ! create subgroups by individual expenditures in the first year -- two
115      ! such variables are defined below;
116        *the first of this creates a 7-category variable, costcat1;
117        proc sort data=outdata2; by cost_tot01; run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.OUTDATA2 has 500000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           2:04.41
      cpu time            17.58 seconds


117      !
118        options obs=100000;
119
119      !  proc sort data=outdata2 out=new20; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 100000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW20 has 100000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           11.15 seconds
      cpu time            1.24 seconds


119      !
120        options firstobs=100001 obs=250000;
121
121      !  proc sort data=outdata2 out=new30; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 150000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW30 has 150000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           17.11 seconds
      cpu time            1.82 seconds


121      !
122        options firstobs=250001 obs=400000;
123
123      !  proc sort data=outdata2 out=newr30; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
```

```
      set.
NOTE: There were 150000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEWR30 has 150000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           14.75 seconds
      cpu time            1.81 seconds


123      !
124        options firstobs=400001 obs=450000;
125
125      !  proc sort data=outdata2 out=new10; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 50000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW10 has 50000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           5.51 seconds
      cpu time            0.65 seconds


125      !
126        options firstobs=450001 obs=475000;
127
127      !  proc sort data=outdata2 out=new5; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 25000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW5 has 25000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           3.09 seconds
      cpu time            0.34 seconds


127      !
128        options firstobs=475001 obs=495000;
129
129      !  proc sort data=outdata2 out=new4; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 20000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW4 has 20000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           1.74 seconds
      cpu time            0.29 seconds


129      !
130        options firstobs=495001 obs=500000;
131
```

```
131      !  proc sort data=outdata2 out=new1; by cost_tot01; run;

NOTE: Input data set is already sorted; it has been copied to the output data
      set.
NOTE: There were 5000 observations read from the data set WORK.OUTDATA2.
NOTE: The data set WORK.NEW1 has 5000 observations and 27 variables.
NOTE: PROCEDURE SORT used:
      real time           0.84 seconds
      cpu time            0.05 seconds


131      !
132        options firstobs=1;
133        data new20; set new20; costcat1=1; run;

NOTE: There were 100000 observations read from the data set WORK.NEW20.
NOTE: The data set WORK.NEW20 has 100000 observations and 28 variables.
NOTE: DATA statement used:
      real time           9.42 seconds
      cpu time            1.49 seconds


133      !
134        data new30; set new30; costcat1=2; run;

NOTE: There were 150000 observations read from the data set WORK.NEW30.
NOTE: The data set WORK.NEW30 has 150000 observations and 28 variables.
NOTE: DATA statement used:
      real time           16.03 seconds
      cpu time            2.03 seconds


134      !
135        data newr30; set newr30; costcat1=3; run;

NOTE: There were 150000 observations read from the data set WORK.NEWR30.
NOTE: The data set WORK.NEWR30 has 150000 observations and 28 variables.
NOTE: DATA statement used:
      real time           16.32 seconds
      cpu time            2.04 seconds


135      !
136        data new10; set new10; costcat1=4; run;

NOTE: There were 50000 observations read from the data set WORK.NEW10.
NOTE: The data set WORK.NEW10 has 50000 observations and 28 variables.
NOTE: DATA statement used:
      real time           5.97 seconds
      cpu time            0.69 seconds


136      !
```

```
137        data new5; set new5; costcat1=5; run;

NOTE: There were 25000 observations read from the data set WORK.NEW5.
NOTE: The data set WORK.NEW5 has 25000 observations and 28 variables.
NOTE: DATA statement used:
      real time           3.31 seconds
      cpu time            0.39 seconds


137      !
138        data new4; set new4; costcat1=6; run;

NOTE: There were 20000 observations read from the data set WORK.NEW4.
NOTE: The data set WORK.NEW4 has 20000 observations and 28 variables.
NOTE: DATA statement used:
      real time           3.12 seconds
      cpu time            0.34 seconds


138      !
139        data new1; set new1; costcat1=7; run;

NOTE: There were 5000 observations read from the data set WORK.NEW1.
NOTE: The data set WORK.NEW1 has 5000 observations and 28 variables.
NOTE: DATA statement used:
      real time           0.96 seconds
      cpu time            0.08 seconds


139      !
140        data outdata3;
141          set new20 new30 newr30 new10 new5 new4 new1;
142          label costcat1='Previous Year Cost Category (1 to 7)';
143        run;

NOTE: There were 100000 observations read from the data set WORK.NEW20.
NOTE: There were 150000 observations read from the data set WORK.NEW30.
NOTE: There were 150000 observations read from the data set WORK.NEWR30.
NOTE: There were 50000 observations read from the data set WORK.NEW10.
NOTE: There were 25000 observations read from the data set WORK.NEW5.
NOTE: There were 20000 observations read from the data set WORK.NEW4.
NOTE: There were 5000 observations read from the data set WORK.NEW1.
NOTE: The data set WORK.OUTDATA3 has 500000 observations and 28 variables.
NOTE: DATA statement used:
      real time           53.82 seconds
      cpu time            7.28 seconds


143      !
144        proc freq;
145          tables costcat1 / missing;
146          format costcat1 costcat.;
147        run;
```

```
NOTE: There were 500000 observations read from the data set WORK.OUTDATA3.
NOTE: The PROCEDURE FREQ printed page 8.
NOTE: PROCEDURE FREQ used:
      real time           10.75 seconds
      cpu time            2.44 seconds


147      !
148        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work0D1100000D40_genmed2
File Name:         /data/saswork/SAS_work0D1100000D40_genmed2
Inode Number:      4312192
Access Permission: rwxrwx---
Owner Name:        amresh
File Size (bytes): 512


 #  Name      Memtype  File Size  Last Modified
-----------------------------------------------------
 1  FORMATS   CATALOG      28672   08MAR2005:15:23:10
 2  NEW1      DATA        991232   08MAR2005:15:37:19
 3  NEW10     DATA       9658368   08MAR2005:15:37:12
 4  NEW20     DATA      19292160   08MAR2005:15:36:33
 5  NEW30     DATA      28925952   08MAR2005:15:36:49
 6  NEW4      DATA       3874816   08MAR2005:15:37:18
 7  NEW5      DATA       4841472   08MAR2005:15:37:15
 8  NEWR30    DATA      28925952   08MAR2005:15:37:06
 9  OUTDATA2  DATA      93118464   08MAR2005:15:35:29
10  OUTDATA3  DATA      96395264   08MAR2005:15:38:13
148      !                              delete outdata2 new20 new30 newr30 new10
148      ! new5 new4 new1; run;

NOTE: Deleting WORK.OUTDATA2 (memtype=DATA).
NOTE: Deleting WORK.NEW20 (memtype=DATA).
NOTE: Deleting WORK.NEW30 (memtype=DATA).
NOTE: Deleting WORK.NEWR30 (memtype=DATA).
NOTE: Deleting WORK.NEW10 (memtype=DATA).
NOTE: Deleting WORK.NEW5 (memtype=DATA).
NOTE: Deleting WORK.NEW4 (memtype=DATA).
NOTE: Deleting WORK.NEW1 (memtype=DATA).

NOTE: PROCEDURE DATASETS used:
      real time           2.07 seconds
      cpu time            0.17 seconds


149        proc means data=outdata3 mean min max nmiss n;
150        var expyr;
```

```
151        run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA3.
NOTE: The PROCEDURE MEANS printed page 9.
NOTE: PROCEDURE MEANS used:
      real time           6.60 seconds
      cpu time            2.00 seconds


151      !
152
153        *the second is a 50-category (by 2-percentiles), costcat2;
154        *sort by cost_tot01;
155        proc sort data=outdata3; by cost_tot01; run;

NOTE: There were 500000 observations read from the data set WORK.OUTDATA3.
NOTE: The data set WORK.OUTDATA3 has 500000 observations and 28 variables.
NOTE: PROCEDURE SORT used:
      real time           2:02.64
      cpu time            14.95 seconds


155      !
156        data saswork.dod3valid;
157          set outdata3;
158          *define a obs number counter (1,2,3,...,500K);
159          obsnum = _N_;
160          *define the desired 2-percentiles (first 2% called 1, second 2%
160      ! called 2,...upto 50);
161          costcat2 = int((obsnum-1)/10000)+1;
162          *print some trial numbers;
163          if obsnum in (1, 10000, 10001, 49000, 49001, 500000) then put
obsnum
163      !  costcat2;
164        run;

1 1
10000 1
10001 2
49000 5
49001 5
500000 50
NOTE: There were 500000 observations read from the data set WORK.OUTDATA3.
NOTE: The data set SASWORK.DOD3VALID has 500000 observations and 30 variables.
NOTE: DATA statement used:
      real time           56.03 seconds
      cpu time            8.60 seconds


164      !
165        proc freq data=saswork.dod3valid;
166          tables costcat2 / missing;
167        run;
```

```
NOTE: There were 500000 observations read from the data set SASWORK.DOD3VALID.
NOTE: The PROCEDURE FREQ printed pages 10-11.
NOTE: PROCEDURE FREQ used:
      real time           11.23 seconds
      cpu time            2.51 seconds


167      !
168        proc means data=saswork.dod3valid mean min max nmiss n;
169        var expyr;
170        run;

NOTE: There were 500000 observations read from the data set SASWORK.DOD3VALID.
NOTE: The PROCEDURE MEANS printed page 12.
NOTE: PROCEDURE MEANS used:
      real time           8.58 seconds
      cpu time            2.05 seconds


170      !
171
172
173
174
175
176
177
178

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           18:43.37
      cpu time            2:45.35
```

Attachment A6_2
(task6e_dod3dxgr.log)

```
1    The SAS System                                12:12 Tuesday, March 1, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.31 seconds
      cpu time            0.09 seconds

1          /*********************************/
2          /* task6e_dod3dxgr.sas       */
3          /* DOD1 Project                  */
4          /* Create dx groups of interest */
5          /* Jenn Fonda                    */
6          /* First created: Feb  1, 2005  */
7          /* Last modified: Feb 28, 2005  */
8          /*********************************/
9
10         /* This program first creates 9 diagnosis groups of interest from 3-
digit ICD9-9 diagnostic codes (Female breast cancer,
11         diabetes, major mental health disorders, ischemic heart disease,
chf, copd, asthma, chronic renal failure, &
12         osteoarthritis). Next, it adds these 9 diagnoses groups of interest
for the 500,000 in the validation sample.
13
14         We will later run a program to calculate the predictive ratios for
all three models (ACG, CDPS, and DCG) for untop-coded
15         and top-coded ($25K and $50K) expenditures for 2002.
16
17         Input files: dod2diagb, dod3valid
18         Output files: dod3valid2
19         */
20
21         options ps=55 ls=80;
22         libname dod '/data/dod1/saswork';
NOTE: Libref DOD was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
23         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
NOTE: Format COSTCAT has been output.
```

```
NOTE: PROCEDURE FORMAT used:
      real time           0.39 seconds
      cpu time            0.03 seconds


56
57
58         data dxgr; set dod.dod2diagb (keep=recipno diag1-diag6);
59           rename recipno=pid;
60           diabetes=0; cancer_fbrst=0; mentalh=0; ischhrt=0; chf=0; copd=0;
60       ! asthma=0; chrenfail=0; osteoart=0;
61           array dx (6) diag:;
62           do i=1 to 6;
63                        /* Diabetes Mellitus */
64                  if substr(dx(i),1,3)='250' then diabetes=1;
65                        /* Female breast cancer */
66                  else if substr(dx(i),1,3)='174' then cancer_fbrst=1;
67                        /* Major Mental Health Disorders */
68                  else if substr(dx(i),1,3) in('295' '296' '300' '308' '309')
then
68       ! mentalh=1;
69                        /* Ischemic Heart Disease */
70                  else if substr(dx(i),1,3) in('410' '411' '412' '413' '414')
then
70       ! ischhrt=1;
71                        /* Congestive Heart Failure */
72                  else if substr(dx(i),1,3)='428' then chf=1;
73                        /* Chronic Obstructive Pulmonary Disease (COPD) */
74                  else if substr(dx(i),1,3) in('491' '492' '494' '496') then
copd=1;
75                        /* Asthma */
76                  else if substr(dx(i),1,3)='493' then asthma=1;
77                        /* Chronic Renal Failure */
78                  else if substr(dx(i),1,3)='585' then chrenfail=1;
79                        /* Osteoarthritis */
80                  else if substr(dx(i),1,3)='715' then osteoart=1;
81           end;
82           drop i;
83           label diabetes='Diabetes Mellitus' cancer_fbrst='Female breast
83       ! cancer'
84           mentalh='Major Mental Health Disorders' ischhrt='Ischemic Heart
84       ! Disease'
85           chf='Congestive Heart Failure' copd='Chronic Obstructive Pulmonary
85       ! Disease (COPD)'
86           asthma='Asthma' chrenfail='Chronic Renal Failure'
86       ! osteoart='Osteoarthritis';
87         run;

NOTE: There were 5343739 observations read from the data set DOD.DOD2DIAGB.
NOTE: The data set WORK.DXGR has 5343739 observations and 16 variables.
NOTE: DATA statement used:
      real time           26:24.71
      cpu time            5:09.94


88
```

```
89         ** Summarize to one line per person;
90         proc sort data=dxgr; by pid; run;

NOTE: There were 5343739 observations read from the data set WORK.DXGR.
NOTE: The data set WORK.DXGR has 5343739 observations and 16 variables.
NOTE: PROCEDURE SORT used:
      real time           22:24.05
      cpu time            2:27.00


91         proc summary data=dxgr;
92           by pid;
93           var diabetes cancer_fbrst mentalh ischhrt chf copd asthma
chrenfail
93       ! osteoart;
94           output out=dxgr2 max(diabetes cancer_fbrst mentalh ischhrt chf
copd
94       ! asthma chrenfail osteoart)
95           = diabetes cancer_fbrst mentalh ischhrt chf copd asthma chrenfail
95       ! osteoart;
96         run;

NOTE: There were 5343739 observations read from the data set WORK.DXGR.
NOTE: The data set WORK.DXGR2 has 1976274 observations and 12 variables.
NOTE: PROCEDURE SUMMARY used:
      real time           12:35.78
      cpu time            2:49.31


97
98         ** Delete dataset no longer needed;
99         proc datasets library=work;
                            -----Directory-----

         Libref:            WORK
         Engine:            V8
         Physical Name:     /data/saswork/SAS_work20F7000049C2_genmed2
         File Name:         /data/saswork/SAS_work20F7000049C2_genmed2
         Inode Number:      4112064
         Access Permission: rwxrwx---
         Owner Name:        amresh
         File Size (bytes): 512


              #  Name     Memtype   File Size  Last Modified
              ---------------------------------------------------
              1  DXGR     DATA      774807552   01MAR2005:13:01:21
              2  DXGR2    DATA      223322112   01MAR2005:13:13:58
              3  FORMATS  CATALOG       28672   01MAR2005:12:12:34
100
100      !  delete dxgr;
101        run;

NOTE: Deleting WORK.DXGR (memtype=DATA).
102
```

```
103        ** Add diagnoses groups to dod3valid;

NOTE: PROCEDURE DATASETS used:
      real time           3.25 seconds
      cpu time            0.89 seconds


104        proc sort data=dod.dod3valid out=dod3valid; by pid; run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID.
NOTE: The data set WORK.DOD3VALID has 500000 observations and 30 variables.
NOTE: PROCEDURE SORT used:
      real time           2:23.02
      cpu time            19.68 seconds


105        proc sort data=dxgr2; by pid; run;

NOTE: There were 1976274 observations read from the data set WORK.DXGR2.
NOTE: The data set WORK.DXGR2 has 1976274 observations and 12 variables.
NOTE: PROCEDURE SORT used:
      real time           5:12.64
      cpu time            42.65 seconds


106
107        data dod3valid2;
108          merge dod3valid (in=one) dxgr2 (in=two);
109          by pid;
110          indod3val=one;
111          indx=two;
112        run;

NOTE: There were 500000 observations read from the data set WORK.DOD3VALID.
NOTE: There were 1976274 observations read from the data set WORK.DXGR2.
NOTE: The data set WORK.DOD3VALID2 has 2047716 observations and 43 variables.
NOTE: DATA statement used:
      real time           6:27.62
      cpu time            51.96 seconds


113
114        ** Delete dataset no longer needed;
115        proc datasets library=work;
                            -----Directory-----

         Libref:            WORK
         Engine:            V8
         Physical Name:     /data/saswork/SAS_work20F7000049C2_genmed2
         File Name:         /data/saswork/SAS_work20F7000049C2_genmed2
         Inode Number:      4112064
         Access Permission: rwxrwx---
         Owner Name:        amresh
```

-----Directory-----

File Size (bytes): 512

#	Name	Memtype	File Size	Last Modified
1	DOD3VALID	DATA	105046016	01MAR2005:13:16:25
2	DOD3VALID2	DATA	645210112	01MAR2005:13:28:05
3	DXGR2	DATA	223322112	01MAR2005:13:21:37
4	FORMATS	CATALOG	28672	01MAR2005:12:12:34

```
116
116      !  delete dod3valid dxgr2;
117        run;

NOTE: Deleting WORK.DOD3VALID (memtype=DATA).
NOTE: Deleting WORK.DXGR2 (memtype=DATA).
118
119        ** Check input of files--there are 500,000 people in dod3valid,
119      ! 1,976,274 people in dxgr2, and 428,558 in both;

NOTE: PROCEDURE DATASETS used:
      real time           1.39 seconds
      cpu time            0.28 seconds


120        proc freq data=dod3valid2;
121          title 'Check input of files dod3valid & dxgr2';
122          tables indod3val*indx;
123        run;

NOTE: There were 2047716 observations read from the data set WORK.DOD3VALID2.
NOTE: The PROCEDURE FREQ printed page 1.
NOTE: PROCEDURE FREQ used:
      real time           57.92 seconds
      cpu time            13.70 seconds


124
125        ** Create permanent dataset dod3valid2 which add diagnosis
125      ! information for validation sample;
126        data dod.dod3valid2; set dod3valid2;
127          where indod3val=1;
128          ** Convert missing values for diagnoses to zero;
129          array dx(9) diabetes cancer_fbrst mentalh ischhrt chf copd asthma
129      ! chrenfail osteoart;
130          do i=1 to 9;
131                 if dx(i)=. then dx(i)=0;
132          end;
133          drop i indod3val indx;
134        run;

NOTE: There were 500000 observations read from the data set WORK.DOD3VALID2.
```

```
      WHERE indod3val=1;
NOTE: The data set DOD.DOD3VALID2 has 500000 observations and 41 variables.
NOTE: DATA statement used:
      real time           2:07.76
      cpu time            18.86 seconds


135
136        proc freq data=dod.dod3valid2;
137          title 'Frequency for 9 diseases of interest';
138          tables diabetes cancer_fbrst mentalh ischhrt chf copd asthma
138       ! chrenfail osteoart;
139          run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID2.
NOTE: The PROCEDURE FREQ printed pages 2-3.
NOTE: PROCEDURE FREQ used:
      real time           21.49 seconds
      cpu time            5.24 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           1:19:01.47
      cpu time            12:59.66
```

Attachment A6_3
(task6f_agesexpred.log)

```
1   The SAS System                                18:06 Tuesday, March 8, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.


This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.


NOTE: SAS initialization used:
      real time           0.97 seconds
      cpu time            0.08 seconds

1          /*************************************/
2          /* task6f_agesexpred.sas             */
3          /* DoD1                              */
4          /* Jenn Fonda                        */
5          /* First created: March  8, 2005     */
6          /* Last modified: March  8, 2005     */
7          /*************************************/
8
9          /* This program regresses second-year expenditures with only age and
sex indicators and obtains predicted
10         expenditures for the validation sample. The predicted information is
later combined with dataset dod3valid2 to create
11         the new dataset dod3valid3.
12
13         Input file: dxcg, dod3valid2
14         Output file: dod3valid3
15         */
16
17         options ps=60 ls=80 nocenter compress=yes;
18         footnote '/data/dod1/saswork/fonda/task6_dcgpred.sas';
19         libname indcg '/data/dod1/saswork/fonda';
NOTE: Libref INDCG was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork/fonda
19        !
20         libname dod '/data/dod1/saswork/';
NOTE: Libref DOD was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
20        !
21         %include '/data/dod1/saswork/formats.sas';
NOTE: Format $URBANF has been output.
NOTE: Format $RACEF has been output.
NOTE: Format $BENCATF has been output.
NOTE: Format $RESREGF has been output.
NOTE: Format $RECSVCF has been output.
NOTE: Format TYPE has been output.
NOTE: Format COSTCAT has been output.

NOTE: PROCEDURE FORMAT used:
      real time           0.42 seconds
```

```
      cpu time            0.03 seconds


54
55          /* Create dataset that contains only idno, expenditures for
55        ! 2002(EXPEND2),
56          and the values for agesex and HCC001-HCC184 */
57          data agesex; set indcg.dxcg (keep=IDNO AGE SEX EXPEND2 datatype
57        ! rename=(EXPEND2=expyr));
58             **  Create topcoded expenditures;
59                   expyrt25 = expyr;
60                   if expyr GT 25000 then expyrt25 = 25000;
61                   expyrt50 = expyr;
62                   if expyr GT 50000 then expyrt50 = 50000;
62        !
63             ** Create age and gender indicators;
64             if SEX=1 & AGE LE 1 then m0_1=1; else m0_1=0;
65             if SEX=1 & (AGE GE 2) & (AGE LE 10) then m2_10=1; else m2_10=0;
66             if SEX=1 & (AGE GE 11) & (AGE LE 17) then m11_17=1; else m11_17=0;
67             if SEX=1 & (AGE GE 18) & (AGE LE 25) then m18_25=1; else m18_25=0;
68             if SEX=1 & (AGE GE 26) & (AGE LE 30) then m26_30=1; else m26_30=0;
69             if SEX=1 & (AGE GE 31) & (AGE LE 35) then m31_35=1; else m31_35=0;
70             if SEX=1 & (AGE GE 36) & (AGE LE 40) then m36_40=1; else m36_40=0;
71             if SEX=1 & (AGE GE 41) & (AGE LE 45) then m41_45=1; else m41_45=0;
72             if SEX=1 & (AGE GE 46) & (AGE LE 50) then m46_50=1; else m46_50=0;
73             if SEX=1 & (AGE GE 51) & (AGE LE 55) then m51_55=1; else m51_55=0;
74             if SEX=1 & (AGE GE 56) & (AGE LE 60) then m56_60=1; else m56_60=0;
75             if SEX=1 & (age GE 61) & (AGE LE 64) then m61_64=1; else m61_64=0;
76             if SEX=2 & AGE LE 1 then f0_1=1; else f0_1=0;
77             if SEX=2 & (AGE GE 2) & (AGE LE 10) then f2_10=1; else f2_10=0;
78             if SEX=2 & (AGE GE 11) & (AGE LE 17) then f11_17=1; else f11_17=0;
79             if SEX=2 & (AGE GE 18) & (AGE LE 25) then f18_25=1; else f18_25=0;
80             if SEX=2 & (AGE GE 26) & (AGE LE 30) then f26_30=1; else f26_30=0;
81             if SEX=2 & (AGE GE 31) & (AGE LE 35) then f31_35=1; else f31_35=0;
82             if SEX=2 & (AGE GE 36) & (AGE LE 40) then f36_40=1; else f36_40=0;
83             if SEX=2 & (AGE GE 41) & (AGE LE 45) then f41_45=1; else f41_45=0;
84             if SEX=2 & (AGE GE 46) & (AGE LE 50) then f46_50=1; else f46_50=0;
85             if SEX=2 & (AGE GE 51) & (AGE LE 55) then f51_55=1; else f51_55=0;
86             if SEX=2 & (AGE GE 56) & (AGE LE 60) then f56_60=1; else f56_60=0;
87             if SEX=2 & (AGE GE 61) & (AGE LE 64) then f61_64=1; else f61_64=0;
88          run;

NOTE: Character values have been converted to numeric
      values at the places given by: (Line):(Column).
      64:5   65:5   66:5   67:5   68:5   69:5   70:5   71:5   72:5   73:5
      74:5   75:5   76:5   77:5   78:5   79:5   80:5   81:5   82:5   83:5
      84:5   85:5   86:5   87:5
NOTE: There were 2304926 observations read from the data set INDCG.DXCG.
NOTE: The data set WORK.AGESEX has 2304926 observations and 31 variables.
NOTE: Compressing data set WORK.AGESEX decreased size by 73.08 percent.
      Compressed is 9402 pages; un-compressed would require 34924 pages.
NOTE: DATA statement used:
      real time           23:23.01
      cpu time            4:11.53


88        !
89
```

```
90         ** Create a revised version of expenditure variables (expyr,
expyrt25
90       !  and expyrt50) that have missing entries
91         for validation sample;
92         data agesex2; set agesex;
93           if datatype=1 then expyrs = expyr;
94           if datatype=1 then expyrt25s = expyrt25;
95           if datatype=1 then expyrt50s = expyrt50;
96           label datatype='1=fitting sample, 2=Validation sample';
97           label expyr = 'Total health care expenditures, FY2002';
98           label expyrt25 = 'expyr topped at $25K';
99           label expyrt50 = 'expyr topped at $50K';
100        run;

NOTE: There were 2304926 observations read from the data set WORK.AGESEX.
NOTE: The data set WORK.AGESEX2 has 2304926 observations and 34 variables.
NOTE: Compressing data set WORK.AGESEX2 decreased size by 68.45 percent.
      Compressed is 12120 pages; un-compressed would require 38416 pages.
NOTE: DATA statement used:
      real time           11:58.72
      cpu time            1:09.69


100      !
101
102        /* since some regressors (age groups or diagnoses) may take only one
102      ! value for all data,
103        first identify these (note that regression is only on the fitting
103      ! sample) */
104        ** List of all regressors--
105        Note: we will exclude f61_64 since it will be our reference group
(it
105      !  is also a linear combination
106        of all other age-sex groups).We will also exclude HCC129 and HCC173
106      ! since they have a zero frequency count and will interfere with
107        the matrix operation for regression (it will give the note that the
107      ! model is not full rank);
108        %let xlist = m0_1 m2_10 m11_17 m18_25 m26_30 m31_35 m36_40 m41_45
108      ! m46_50
109        m51_55 m56_60 m61_64 f0_1 f2_10 f11_17 f18_25 f26_30 f31_35 f36_40
109      ! f41_45
110        f46_50 f51_55 f56_60;
111
112        ** Run regression;
113        proc reg data=agesex2;
114          model expyrs expyrt25s expyrt50s = &xlist;
115          output out=agesex3 (keep=IDNO pexp_agesex pexpt25_agesex
115      ! pexpt50_agesex)
116            p=pexp_agesex pexpt25_agesex pexpt50_agesex;
117        run;

NOTE: 2304926 observations read.
NOTE: 500000 observations have missing values.
NOTE: 1804926 observations used in computations.
117      !
118
119        ** Delete unnecessary files;

NOTE: There were 2304926 observations read from the data set WORK.AGESEX2.
NOTE: The data set WORK.AGESEX3 has 2304926 observations and 4 variables.
```

```
NOTE: Compressing data set WORK.AGESEX3 increased size by 27.98 percent.
      Compressed is 15285 pages; un-compressed would require 11943 pages.
NOTE: The PROCEDURE REG printed pages 1-3.
NOTE: PROCEDURE REG used:
      real time           14:33.07
      cpu time            1:43.31


120        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work3A2600000E5D_genmed2
File Name:         /data/saswork/SAS_work3A2600000E5D_genmed2
Inode Number:      4323520
Access Permission: rwxrwx---
Owner Name:        jfonda
File Size (bytes): 512


#  Name     Memtype    File Size  Last Modified
-----------------------------------------------------
1  AGESEX   DATA       154050560   08MAR2005:18:29:44
2  AGESEX2  DATA       198582272   08MAR2005:18:41:43
3  AGESEX3  DATA       125222912   08MAR2005:18:56:16
4  FORMATS  CATALOG        28672   08MAR2005:18:06:22
5  PROFILE  CATALOG        12288   08MAR2005:18:06:22
6  REGSTRY  ITEMSTOR       32768   08MAR2005:18:06:21
120      !
121
121      !  delete agesex agesex2;
122        run;

NOTE: Deleting WORK.AGESEX (memtype=DATA).
NOTE: Deleting WORK.AGESEX2 (memtype=DATA).
122      !
123
124        ** Save predicted expenditures from dcg in new file saswork.dod3dcg;
125        ** Merge with dod2 (dod2 has individual characteristics);

NOTE: PROCEDURE DATASETS used:
      real time           1.51 seconds
      cpu time            0.32 seconds


126        proc sort data=agesex3 (rename=(IDNO=pid)); by pid; run;

NOTE: There were 2304926 observations read from the data set WORK.AGESEX3.
NOTE: The data set WORK.AGESEX3 has 2304926 observations and 4 variables.
NOTE: Compressing data set WORK.AGESEX3 increased size by 27.98 percent.
      Compressed is 15285 pages; un-compressed would require 11943 pages.
NOTE: PROCEDURE SORT used:
      real time           7:49.08
      cpu time            53.87 seconds


126      !
```

```
127        proc sort data=dod.dod3valid2; by pid; run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID2.
NOTE: The data set DOD.DOD3VALID2 has 500000 observations and 41 variables.
NOTE: PROCEDURE SORT used:
      real time           2:54.28
      cpu time            22.02 seconds


127      !
128
129        data dod.dod3valid3;
130          merge agesex3 (in=infirst) dod.dod3valid2 (in=insecond);
131          by pid;
132          if infirst=1 & insecond=1;
133        run;

NOTE: There were 2304926 observations read from the data set WORK.AGESEX3.
NOTE: There were 500000 observations read from the data set DOD.DOD3VALID2.
NOTE: The data set DOD.DOD3VALID3 has 500000 observations and 44 variables.
NOTE: Compressing data set DOD.DOD3VALID3 decreased size by 29.87 percent.
      Compressed is 6876 pages; un-compressed would require 9805 pages.
NOTE: DATA statement used:
      real time           5:46.67
      cpu time            51.46 seconds


133      !
134
135        * produces mean of predicted 2002 expenditures for the validation
135      ! sample overall and for each disease category;
136        proc means data=dod.dod3valid3 mean;
137          var pexp_agesex;
138        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 4.
NOTE: PROCEDURE MEANS used:
      real time           20.11 seconds
      cpu time            5.67 seconds


138      !
139
140        proc means data=dod.dod3valid3 mean;
141          class diabetes;
142          var pexp_agesex;
143        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 5.
NOTE: PROCEDURE MEANS used:
      real time           26.37 seconds
      cpu time            6.56 seconds


143      !
144
```

```
145        proc means data=dod.dod3valid3 mean;
146          class cancer_fbrst;
147          var pexp_agesex;
148        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 6.
NOTE: PROCEDURE MEANS used:
      real time           31.42 seconds
      cpu time            6.40 seconds


148      !
149
150        proc means data=dod.dod3valid3 mean;
151          class mentalh;
152          var pexp_agesex;
153        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 7.
NOTE: PROCEDURE MEANS used:
      real time           27.96 seconds
      cpu time            6.60 seconds


153      !
154
155        proc means data=dod.dod3valid3 mean;
156          class ischhrt;
157          var pexp_agesex;
158        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 8.
NOTE: PROCEDURE MEANS used:
      real time           28.98 seconds
      cpu time            6.40 seconds


158      !
159
160        proc means data=dod.dod3valid3 mean;
161          class chf;
162          var pexp_agesex;
163        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 9.
NOTE: PROCEDURE MEANS used:
      real time           25.76 seconds
      cpu time            6.55 seconds


163      !
164
165        proc means data=dod.dod3valid3 mean;
166          class copd;
```

```
167          var pexp_agesex;
168        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 10.
NOTE: PROCEDURE MEANS used:
      real time           24.38 seconds
      cpu time            6.38 seconds


168      !
169
170        proc means data=dod.dod3valid3 mean;
171          class asthma;
172          var pexp_agesex;
173        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 11.
NOTE: PROCEDURE MEANS used:
      real time           24.09 seconds
      cpu time            6.58 seconds


173      !
174
175        proc means data=dod.dod3valid3 mean;
176          class chrenfail;
177          var pexp_agesex;
178        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 12.
NOTE: PROCEDURE MEANS used:
      real time           30.01 seconds
      cpu time            6.38 seconds


178      !
179
180        proc means data=dod.dod3valid3 mean;
181          class osteoart;
182          var pexp_agesex;
183        run;

NOTE: There were 500000 observations read from the data set DOD.DOD3VALID3.
NOTE: The PROCEDURE MEANS printed page 13.
NOTE: PROCEDURE MEANS used:
      real time           23.67 seconds
      cpu time            6.55 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           1:10:53.86
      cpu time            10:16.46
```

Attachment A6_4
(task6g_pcmtype.log)

```
proc printto print='/data/dod1/saswork/task6g_pcmtype.log'; run;
/*
task6g_pcmtype.sas
Amresh Hanchate
DOD1 Project
created: April 5, 2005
last revision: April 25, 2005 (by Jenn Fonda)

This job adds following fields to dod2 and dod3valid3
i) adds pcmtype field (whether primary care manager in September 2001 is
military (0) or civilian(1)
ii) adds datatype field (only to dod2 file -- already present in dod3valid3) --
note that datatype field identifies whether observation is in validation or
estimation sample

To create pcmtype I used the dmidid file (dmis_table_200408.v8x) sent by Nancy
(likely downloaded from www.dmisid.com). This file has the list of all dmisids
along with names of the military facilities, their location and their service
affiliation. It is




the service affiliation field (servcode) that is used here to determine
pcmtype. Civilian has service code M and all others are military.

Input: dod2, dod2charb, dod3valid3
Output: dod2v2.v8x (new comprehensive version of dod2)
          dod3valid4.v8x (new comprehensive version of dod3valid)

*/

options ps=55 ls=80 nocenter formdlim=' ';
libname saswork '/data/dod1/saswork/';

*sort dmis data file;
proc cimport data=dmis file='/data/dod1/saswork/dmis_table_200408.v8x';
run;
data dmis;
       set dmis(keep=dmisid servcode);
       temp = length(left(right(dmisid)));
run;

** Check to see if there are any missing dmisid's--there are 460 missing
dmisids;
proc freq data=dmis;
       title 'Check if there are any missing dmisids';
       where dmisid = ' ';
       tables dmisid;
run;

** Check the length of the temp variable--87% of the dmisid's have length of 4.
The remaining dmisids are missing
13% of the dmisids are missing and have a length of 1;
proc freq data=dmis;
       title 'Check length of dmisid in dmis';
       tables temp;
```

```
run;

** We found that in the Excel Table (dmis_table_200408) there were only 2,968
dmisids.  When transfering to an export
file we included an additional 460 lines that didn't have any information.
Thus, we will delete the 460 extra lines
that are missing the dmisid & all other information;
data dmis; set dmis;
       where dmisid ne '';
       ** Change character length of dmisid from 14 to 4;
       length dmisid2 $ 4.;
       ** Create dmisid2 by left-justifying dmisid & take the 4-digit dmisid;
       dmisid2 = substr(left(dmisid),1,4);
       drop dmisid temp;
run;

proc sort data=dmis; by dmisid2; run;

*add dmisid field to dod2 -- dmisid field for FY2001 is in datafile lenr01;
*first sort dod2 by pid;
proc sort data=saswork.dod2; by pid; run;
*obtain dmisid field -- called enr0109, but renamed as dmisid, from lenr01;
libname xpt1 xport '/data/dod1/rawdata/lenr01.xpt';
proc sort data=xpt1.lenr01 (keep=pid enr0109 rename=(enr0109=dmisid))
out=enr01; by pid; run;
*merge dmisid to dod2;
data revdod2;
       merge saswork.dod2 (in=one) enr01 (in=two);
       by pid;
       infirst = one;
       insecond = two;
       label dmisid='Enrollment dmisid in September 2001';
       temp = length(left(right(dmisid)));
run;

** Check to see if there are any missing dmisid's--there are 584,453
records that are missing dmisid's;
proc freq data=revdod2;
       title 'Check if there are any missing dmisids';
       where dmisid = ' ';
       tables dmisid;
run;

** Check the length of the temp variable--87% of the dmisid's have length of 4.
The remaining dmisids are missing
13% of the dmisids are missing and have a length of 1;
proc freq data=revdod2;
       title 'Check length of dmisid in revdod2';
       tables temp;
run;

data revdod2; set revdod2;
       ** Change character length of dmisid from 14 to 4;
       length dmisid2 $ 4.;
       ** Create dmisid2 by left-justifying dmisid & take the 4-digit dmisid;
       dmisid2 = substr(left(dmisid),1,4);
       drop dmisid temp;
run;
proc datasets library=work;
       delete enr01;
run;
proc freq data=revdod2; tables infirst*insecond; run;
data revdod2;
```

```
       set revdod2;
       if infirst=1;
       drop infirst insecond;
run;
*still need to add servcode field from dmis to create pcmtype, but first do
following (since already sorted by pid);

*add datatype field also to dod2v2 -- this field is in dod2charb;
proc sort data=revdod2; by pid; run;
proc sort data=saswork.dod2charb (keep=recipno datatype rename=(recipno=pid))
out=dod2charb; by pid; run;
data revdod2b;
       merge revdod2 (in=one) dod2charb (in=two);
       by pid;
       infirst=one;
       insecond=two;
run;
proc datasets library=work;
       delete dod2charb revdod2;
run;
proc freq data=revdod2b; tables infirst*insecond; run;
data revdod2b;
       set revdod2b;
       if infirst=1;
       drop infirst insecond;
run;
proc freq data=revdod2b; tables datatype / missing; run;

*now add servcode field from dmis and create pcmtype field;
proc sort data=revdod2b; by dmisid2; run;
data revdod2c;
       merge revdod2b (in=one) dmis (in=two);
       by dmisid2;
       infirst = one;
       insecond = two;
       pcmtype=(servcode='M');
       label pcmtype='If PCM in Sept 2001 is civilian (1=Yes, 0=No)';
run;
proc datasets library=work;
       delete revdod2b dmis;
run;
proc freq data=revdod2c; tables infirst*insecond; run;
data revdod2c;
       set revdod2c;
       if infirst=1;
       rename dmisid2=dmisid;
       drop infirst insecond SERVCODE;
run;
proc freq data=revdod2c; tables pcmtype / missing; run;


*save file as dod2v2 -- sort by pid first;
proc sort data=revdod2c; by pid; run;
proc cport data=revdod2c file='/data/dod1/saswork/dod2v2.v8x'; run;
proc contents data=revdod2c varnum; run;

*attach dmisid and pcmtype variables to dod3valid3;
proc sort data=saswork.dod3valid3; by pid; run;
proc sort data=revdod2c(keep=pid dmisid pcmtype) out=temp; by pid; run;
proc datasets library=work; delete revdod2c; run;
data dod3valid4;
       merge saswork.dod3valid3 (in=one) temp (in=two);
       by pid;
```

```
       infirst=one;
       insecond=two;
run;
proc freq data=dod3valid4; tables infirst*insecond; run;
data dod3valid4;
       set dod3valid4;
       if infirst=1;
       drop infirst insecond obsnum _TYPE_ _FREQ_;
       label costcat2='Previous year cost -- 2% categories';
run;
proc cport data=dod3valid4 file='/data/dod1/saswork/dod3valid4.v8x'; run;
proc contents data=dod3valid4 varnum; run;
```

Attachment A6_5
(task6h_pcmtype_exclude.log)

```
1   The SAS System                                19:00 Monday, April 25, 2005

NOTE: Copyright (c) 1999-2000 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 8.1 (TS1M0)
      Licensed to BOSTON UNIV. INFO. TECHNOLOGY, Site 0001506004.
NOTE: This session is executing on the SunOS 5.8 platform.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           4.93 seconds
      cpu time            0.09 seconds

1          /*********************************/
2          /* task6h_pcmtype_exclude.sas    */
3          /* DOD1                          */
4          /* Jenn Fonda                    */
5          /* First created: April 22, 2005 */
6          /* Last modified: April 25, 2005 */
7          /*********************************/
8
9          /* This program adds the pcmtype field (whether primary care manager
in September 2001 is military (0) or civilian(1)
10         for the 1.1 million excluded from our sample.
11
12         Input files: dmis_table_200408.v8x, lenr01.xpt, excluded
13         Output file: excluded2
14         */
15
16         proc printto print='/data/dod1/saswork/task6h_pcmtype_exclude.log';
run;

NOTE: PROCEDURE PRINTTO used:
      real time           0.00 seconds
      cpu time            0.00 seconds


16       !
17         options ps=55 ls=80 nocenter formdlim=' ';
18         libname saswork '/data/dod1/saswork';
NOTE: Libref SASWORK was successfully assigned as follows:
      Engine:        V8
      Physical Name: /data/dod1/saswork
18       !
19
20
21         *sort dmis data file;
22         proc cimport data=dmis
22       ! file='/data/dod1/saswork/dmis_table_200408.v8x';
```

```
23         run;

NOTE: Proc CIMPORT begins to create/update data set WORK.DMIS
NOTE: Data set contains 20 variables and 3428 observations.
      Logical record length is 274

NOTE: PROCEDURE CIMPORT used:
      real time           5.26 seconds
      cpu time            0.19 seconds


23       !
24         data dmis;
25           set dmis(keep=dmisid servcode);
26           temp = length(left(right(dmisid)));
27         run;

NOTE: There were 3428 observations read from the data set WORK.DMIS.
NOTE: The data set WORK.DMIS has 3428 observations and 3 variables.
NOTE: DATA statement used:
      real time           3.60 seconds
      cpu time            0.07 seconds


27       !
28
29         ** Check to see if there are any missing dmisid's--there are 460
29       ! missing dmisids;
30         proc freq data=dmis;
31           title 'Check if there are any missing dmisids';
32           where dmisid = ' ';
33           tables dmisid;
34         run;
```


The FREQ Procedure

DMISID	Frequency	Percent	Cumulative Frequency	Cumulative Percent

```
Frequency Missing = 460
NOTE: There were 460 observations read from the data set WORK.DMIS.
      WHERE dmisid=' ';
NOTE: The PROCEDURE FREQ printed page 1.
NOTE: PROCEDURE FREQ used:
      real time           1.14 seconds
      cpu time            0.06 seconds


34       !
35
36         ** Check the length of the temp variable--87% of the dmisid's have
36       ! length of 4.  The remaining dmisids are missing
37         13% of the dmisids are missing and have a length of 1;
```

```
38         proc freq data=dmis;
39           title 'Check length of dmisid in dmis';
40           tables temp;
41         run;
```

```
Check length of dmisid in dmis                  19:00 Monday, April 25, 2005
2

The FREQ Procedure

                                    Cumulative    Cumulative
temp    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
   1         460       13.42           460        13.42
   4        2968       86.58          3428       100.00
NOTE: There were 3428 observations read from the data set WORK.DMIS.
```

```
NOTE: The PROCEDURE FREQ printed page 2.
NOTE: PROCEDURE FREQ used:
      real time           1.98 seconds
      cpu time            0.03 seconds


41       !
42
43         ** We found that in the Excel Table (dmis_table_200408) there were
43       ! only 2,968 dmisids.  When transfering to an export
44         file we included an additional 460 lines that didn't have any
44       ! information.  Thus, we will delete the 460 extra lines
45         that are missing the dmisid & all other information;
46         data dmis; set dmis;
47           where dmisid ne '';
48           ** Change character length of dmisid from 14 to 4;
49           length dmisid2 $ 4.;
50           ** Create dmisid2 by left-justifying dmisid & take the 4-digit
50       ! dmisid;
51           dmisid2 = substr(left(dmisid),1,4);
52           drop dmisid temp;
53         run;

NOTE: There were 2968 observations read from the data set WORK.DMIS.
      WHERE dmisid not = '';
NOTE: The data set WORK.DMIS has 2968 observations and 2 variables.
NOTE: DATA statement used:
      real time           2.80 seconds
      cpu time            0.03 seconds


53       !
54
55         proc sort data=dmis; by dmisid2; run;

NOTE: There were 2968 observations read from the data set WORK.DMIS.
NOTE: The data set WORK.DMIS has 2968 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           3.12 seconds
      cpu time            0.02 seconds


55       !
56
57         *add dmisid field to excluded sample dataset -- dmisid field for
57       ! FY2001 is in datafile lenr01;
58         *first sort dod2 by pid;
59         proc sort data=saswork.excluded; by pid; run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used:
      real time           0.18 seconds
      cpu time            0.00 seconds
```

```
59       !
60         *obtain dmisid field -- called enr0109, but renamed as dmisid, from
60       ! lenr01;
61         libname xpt1 xport '/data/dod1/rawdata/lenr01.xpt';
NOTE: Libref XPT1 was successfully assigned as follows:
      Engine:        XPORT
      Physical Name: /data/dod1/rawdata/lenr01.xpt
61       !
62         proc sort data=xpt1.lenr01 (keep=pid enr0109
rename=(enr0109=dmisid))
62       !  out=enr01; by pid; run;

NOTE: There were 4486060 observations read from the data set XPT1.LENR01.
NOTE: The data set WORK.ENR01 has 4486060 observations and 2 variables.
NOTE: PROCEDURE SORT used:
      real time           10:58.40
      cpu time            2:55.18


63         *merge dmisid to dod2;
64         data revexclude;
65           merge saswork.excluded (in=one) enr01 (in=two);
66           by pid;
67           infirst = one;
68           insecond = two;
69           label dmisid='Enrollment dmisid in September 2001';
70           temp = length(left(right(dmisid)));
71         run;

NOTE: There were 1079613 observations read from the data set SASWORK.EXCLUDED.
NOTE: There were 4486060 observations read from the data set WORK.ENR01.
NOTE: The data set WORK.REVEXCLUDE has 4486060 observations and 19 variables.
NOTE: DATA statement used:
      real time           5:48.35
      cpu time            1:15.59


71       !
72
73         ** Check to see if there are any missing dmisid's;
74         proc freq data=revexclude;
75           title 'Check if there are any missing dmisids';
76           where dmisid = ' ';
77           tables dmisid;
78         run;
```


The FREQ Procedure

Enrollment dmisid in September 2001

DMISID	Frequency	Percent	Cumulative Frequency	Cumulative Percent

```
Frequency Missing = 584453
NOTE: There were 584453 observations read from the data set WORK.REVEXCLUDE.
      WHERE dmisid=' ';
NOTE: The PROCEDURE FREQ printed page 3.
NOTE: PROCEDURE FREQ used:
      real time           15.71 seconds
      cpu time            13.08 seconds
```

```
78       !
79
80         ** Check the length of the temp variable--should be a length of 4
for
80       !  all dmisid's;
81         proc freq data=revexclude;
82           title 'Check length of dmisid in revexclude';
83           tables temp;
84         run;
```


The FREQ Procedure

temp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	584453	13.03	584453	13.03
4	3901607	86.97	4486060	100.00

```
NOTE: There were 4486060 observations read from the data set WORK.REVEXCLUDE.
NOTE: The PROCEDURE FREQ printed page 4.
NOTE: PROCEDURE FREQ used:
      real time           5:45.85
      cpu time            15.54 seconds


84       !
85
86         data revexclude; set revexclude;
87           ** Change character length of dmisid from 14 to 4;
88           length dmisid2 $ 4.;
89           ** Create dmisid2 by left-justifying dmisid & take the 4-digit
89       ! dmisid;
90           dmisid2 = substr(left(dmisid),1,4);
91           drop dmisid temp;
92         run;

NOTE: There were 4486060 observations read from the data set WORK.REVEXCLUDE.
NOTE: The data set WORK.REVEXCLUDE has 4486060 observations and 18 variables.
NOTE: DATA statement used:
      real time           14:31.54
      cpu time            53.54 seconds


92       !
93
94         proc datasets library=work;
                    -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work108F0000441A_genmed2
File Name:         /data/saswork/SAS_work108F0000441A_genmed2
Inode Number:      5490304
Access Permission: rwxrwx---
Owner Name:        jfonda
File Size (bytes): 512
```

```
6 The SAS System                                    19:00 Monday, April 25, 2005

#  Name        Memtype    File Size  Last Modified
-----------------------------------------------------------
1  DMIS        DATA           40960   25APR2005:19:00:55
2  ENR01       DATA        99606528   25APR2005:19:11:56
3  REGSTRY     ITEMSTOR       32768   25APR2005:19:00:35
4  REVEXCLUDE  DATA       506912768   25APR2005:19:38:18
94       !
95
95       !  delete enr01;
96         run;

NOTE: Deleting WORK.ENR01 (memtype=DATA).
96       !

NOTE: PROCEDURE DATASETS used:
      real time           0.73 seconds
      cpu time            0.08 seconds


97         proc freq data=revexclude; tables infirst*insecond; run;



Check length of dmisid in revexclude            19:00 Monday, April 25, 2005
5

The FREQ Procedure

Table of infirst by insecond

infirst
          insecond
Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|  Total
---------+--------+
       0 |3406447 |3406447
         |  75.93 |  75.93
         | 100.00 |
         |  75.93 |
---------+--------+
       1 |1079613 |1079613
         |  24.07 |  24.07
         | 100.00 |
         |  24.07 |
---------+--------+
Total     4486060  4486060
           100.00   100.00
NOTE: There were 4486060 observations read from the data set WORK.REVEXCLUDE.
NOTE: The PROCEDURE FREQ printed page 5.
NOTE: PROCEDURE FREQ used:
      real time           53.37 seconds
      cpu time            19.27 seconds


97       !
98         data revexclude;
99           set revexclude;
100          if infirst=1;
```

```
101          drop infirst insecond;
102        run;

NOTE: There were 4486060 observations read from the data set WORK.REVEXCLUDE.
NOTE: The data set WORK.REVEXCLUDE has 1079613 observations and 16 variables.
NOTE: DATA statement used:
      real time           7:19.46
      cpu time            23.52 seconds


102      !
103
104        *now add servcode field from dmis and create pcmtype field;
105        proc sort data=revexclude; by dmisid2; run;

NOTE: There were 1079613 observations read from the data set WORK.REVEXCLUDE.
NOTE: The data set WORK.REVEXCLUDE has 1079613 observations and 16 variables.
NOTE: PROCEDURE SORT used:
      real time           2:24.94
      cpu time            30.68 seconds
```

```
105      !
106        data revexclude2;
107          merge revexclude (in=one) dmis (in=two);
108          by dmisid2;
109          infirst = one;
110          insecond = two;
111          pcmtype=(servcode='M');
112          label pcmtype='If PCM in Sept 2001 is civilian (1=Yes, 0=No)';
113        run;

NOTE: There were 1079613 observations read from the data set WORK.REVEXCLUDE.
NOTE: There were 2968 observations read from the data set WORK.DMIS.
NOTE: The data set WORK.REVEXCLUDE2 has 1082097 observations and 20 variables.
NOTE: DATA statement used:
      real time           1:07.26
      cpu time            13.58 seconds


113      !
114
115        proc freq data=revexclude2; tables infirst*insecond; run;
```


The FREQ Procedure

Table of infirst by insecond

infirst	insecond		
Frequency Percent Row Pct Col Pct	0	1	Total
0	0 0.00 0.00 0.00	2484 0.23 100.00 0.25	2484 0.23
1	76633 7.08 7.10 100.00	1002980 92.69 92.90 99.75	1079613 99.77
Total	76633 7.08	1005464 92.92	1082097 100.00

```
NOTE: There were 1082097 observations read from the data set WORK.REVEXCLUDE2.
NOTE: The PROCEDURE FREQ printed page 6.
NOTE: PROCEDURE FREQ used:
      real time           5.18 seconds
      cpu time            4.87 seconds


115      !
116        data saswork.excluded2;
```

```
117          set revexclude2;
118          if infirst=1;
119          rename dmisid2=dmisid;
120          drop infirst insecond SERVCODE;
121        run;

NOTE: There were 1082097 observations read from the data set WORK.REVEXCLUDE2.
NOTE: The data set SASWORK.EXCLUDED2 has 1079613 observations and 17 variables.
NOTE: DATA statement used:
      real time           57.56 seconds
      cpu time            10.46 seconds


121      !
122        proc freq data=saswork.excluded2; tables pcmtype / missing; run;



Check length of dmisid in revexclude            19:00 Monday, April 25, 2005
7

The FREQ Procedure

        If PCM in Sept 2001 is civilian (1=Yes, 0=No)

                                     Cumulative    Cumulative
pcmtype    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
      0      908793       84.18        908793        84.18
      1      170820       15.82       1079613       100.00
NOTE: There were 1079613 observations read from the data set SASWORK.EXCLUDED2.
NOTE: The PROCEDURE FREQ printed page 7.
NOTE: PROCEDURE FREQ used:
      real time           3.58 seconds
      cpu time            3.03 seconds
```

```
122      !
123
124        proc datasets library=work;
                     -----Directory-----

Libref:            WORK
Engine:            V8
Physical Name:     /data/saswork/SAS_work108F0000441A_genmed2
File Name:         /data/saswork/SAS_work108F0000441A_genmed2
Inode Number:      5490304
Access Permission: rwxrwx---
Owner Name:        jfonda
File Size (bytes): 512


#  Name         Memtype    File Size  Last Modified
-----------------------------------------------------------
1  DMIS         DATA           40960   25APR2005:19:00:55
2  REGSTRY      ITEMSTOR       32768   25APR2005:19:00:35
3  REVEXCLUDE   DATA       105299968   25APR2005:19:48:58
4  REVEXCLUDE2  DATA       130375680   25APR2005:19:50:06
124      !
125
125      !  delete revexclude revexclude2 dmis;
126        run;

NOTE: Deleting WORK.REVEXCLUDE (memtype=DATA).
NOTE: Deleting WORK.REVEXCLUDE2 (memtype=DATA).
NOTE: Deleting WORK.DMIS (memtype=DATA).
NOTE: PROCEDURE DATASETS used:
      real time           0.70 seconds
      cpu time            0.23 seconds


NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           50:43.13
      cpu time            7:19.33
```

Appendix E: Selection of Medical Conditions for Grouped Prediction

Appendix E
Selection of Medical Conditions for Grouped Prediction

One of the comparison criteria for assessing risk adjustment model performance is by groups of persons with specified medical conditions in the first year. We selected these conditions as follows. We used the 3-digit ICD-9 diagnostic codes found in the inpatient (SIDR, HCSR-I) and outpatient (SADR, HCSR-N) files for FY2001, excluding diagnoses from telephone consults. Three variables were created for each of the 3-digit ICD-9 codes; the average total cost per person with the condition, the total cost of all individuals with the condition and the prevalence of the condition in the half million validation sample. The ICD-9 codes were then sorted by each of these three variables and then a subset of 3-digit codes was chosen. The subset included all codes found in at least one of the following categories:

1. Top 25 most prevalent
2. Top 25 with highest average cost of treatment
3. Top 25 with highest total treatment cost

Furthermore, we included the ICD-9 code only if it had a prevalence of at least 400 and a total cost of at least 10 million dollars. Appendix Table E1 shows this subset of codes.

Transient clinical conditions were then excluded because we are interested in exploring the predictive capabilities of the risk adjustment models. Since different manifestations of the same disease could be classified across more than one 3-digit ICD-9 code, we used published medical literature to group related codes. From this list of grouped ICD-9 codes, or chronic medical conditions, in consultation with the Department of Defense sponsor we chose nine for analysis. These conditions are: diabetes mellitus, female breast cancer, major mental health disorders, ischemic heart disease, congestive heart failure, chronic obstructive pulmonary disease (COPD), asthma, chronic renal failure, and osteoarthritis. Appendix Table E2 shows the 3-digit ICD-9 codes used to classify these disease conditions.

Table E1. 25 Most Prevalent, Highest Total Cost or Highest Average Cost 3-Digit ICD-9 Codes in TRICARE Prime Validation Data (N=500,000)*
Sorted by ICD-9 Code

ICD-9 Code & Description	Frequency (n)	Prevalence (%)	Average Cost ($)	Total Cost (millions $)
041 Bacterial infection in CCE and of unspec sites	1,700	0.3	**9,827**	16.7
079 Viral and chlamydial infection in CCE and of unspec sites	**27,758**	**5.6**	2,127	**59.0**
174 Malignant neoplasm of the female breast	1,282	0.3	**7,880**	10.1
250 Diabetes Mellitus	12,112	2.4	**6,513**	**78.9**
272 Dis of lipoid metabolism	**29,594**	**5.9**	4,214	**124.7**
276 Dis of fluid, electrolyte, and acid-base balance	5,887	1.2	**6,529**	38.4
278 Obesity & othr hyperalimentation	11,745	2.3	4,154	**48.8**
280 Iron deficiency anemias	2,777	0.6	**6,129**	17.0
285 Othr and unspec anemias	5,068	1.0	**7,254**	36.8
296 Affective dis	8,175	1.6	5,542	**45.3**
300 Neurotic dis	**13,183**	**2.6**	4,793	**63.2**
311 Depressive disorder, not elsewhere classified	12,516	2.5	4,883	**61.1**
345 Epilepsy	1,584	0.3	**6,957**	11.0
355 Mononeuritis of lower limb and unspec sites	1,879	0.4	**6,199**	11.6
362 Othr retinal dis	2,964	0.6	**6,298**	18.7
366 Cataract	3,555	0.7	**6,280**	22.3
367 Dis of refraction and accommodation	**63,457**	**12.7**	2,227	**141.3**
372 Dis of conjunctiva	**21,785**	**4.4**	2,073	45.2
380 Dis of external ear	**13,042**	**2.6**	2,432	31.7
381 Nonsuppurative otis media & eustachian tube dis	**17,701**	**3.5**	2,315	41.0
382 Suppurative and unspec otitis media	**34,700**	**6.9**	1,725	**59.9**
401 Essential hypertension	**36,396**	**7.3**	4,699	**171.0**
413 Angina Pectoris	1,465	0.3	**8,542**	12.5
414 Othr forms of chronic ischemic heart disease	4,537	0.9	**8,494**	38.5
424 Othr diseases of endocardium	2,298	0.5	**7,617**	17.5
428 Heart failure	1,182	0.2	**13,897**	16.4
429 Ill-defined descriptions and complications of heart disease	1,467	0.3	**10,910**	16.0
443 Othr peripheral vascular disease	1,102	0.2	**9,676**	10.7
461 Acute sinusitis	**33,737**	**6.7**	2,951	**99.6**
462 Acute pharyngitis	**43,089**	**8.6**	1,970	**84.9**
465 Acute upper respiratory infections of multiple or unspec sites	**80,503**	**16.1**	2,171	**174.8**
466 Acute bronchitis & bronchiolitis	**20,884**	**4.2**	3,077	**64.3**
477 Allergic rhinitis	**44,863**	**9.0**	2,586	**116.0**
491 Chronic bronchitis	1,618	0.3	**6,803**	11.0
493 Asthma	**21,752**	**4.4**	3,184	**69.3**
518 Othr diseases of lung	1,959	0.4	**11,795**	23.1
530 Diseases of esophagus	**15,864**	**3.2**	5,156	**81.8**
558 Othr noninfectious gastroenteritis & colitis	**15,637**	**3.1**	2,697	42.2
574 Cholelithiasis	1,682	0.3	**6,091**	10.2
585 Chronic renal failure	467	0.1	**25,038**	11.7
593 Othr dis of kidney and ureter	1,492	0.3	**9,306**	13.9
599 Othr dis of urethra & urinary tract	**21,253**	**4.3**	3,844	**81.7**
626 Dis of menstruation & othr abn bleeding fr female genital tract	**12,710**	**2.5**	3,620	**46.0**
692 Contact dermatitis & othr eczema	**22,139**	**4.4**	2,406	**53.3**
706 Diseases of sebaceous glands	**15,684**	**3.1**	2,222	34.8
714 Rheumatoid arthritis & othr inflammatory polyarthropathies	1,496	0.3	**7,067**	10.6
715 Osteoarthrosis & allied dis	9,369	1.9	5,547	**52.0**

ICD-9 Code & Description	Frequency (n)	Prevalence (%)	Average Cost ($)	Total Cost (millions $)
719 Othr and unspec dis of joint	**39,116**	**7.8**	3,484	**136.3**
721 Spondylosis and allied dis	2,159	0.4	**6,570**	14.2
724 Othr and unspec dis of back	**33,312**	**6.7**	3,788	**126.2**
726 Peripheral enthesopathies & allied syndromes	**18,425**	**3.7**	3,595	**66.2**
729 Othr dis of soft tissues	**27,913**	**5.6**	4,133	**115.4**

*Boldface type indicates the 25 most prevalent, highest total cost or highest average cost cases. Only ICD-9s with at least 400 occurrences among the beneficiary sample and a total cost of at least $10 million were eligible to be included. *Total cost* includes all costs in FY2002 for the subset of people who have at least one such code in 2001. Determination of the highest *average cost* (i.e. per person) ICD-9 codes is subject to a restriction that Total Cost > $10,000,000 and that there were at least 400 occurrences.
CCE =conditions classified elsewhere; Dis = Disorder(s)

Table E2. Important Disease Conditions Using 3-digit ICD-9 Codes in TRICARE Prime Validation Data (N=500,000)

Disease Cohort	ICD-9 Codes & Descriptions
Female breast cancer	Malig neo female breast (174)
Diabetes Mellitus	Diabetes Mellitus (250)
Major Mental Health Disorders	Schizophrenic disorders (295), Affective disorders (296), Neurotic Disorders (300), Acute reaction to stress (308), Adjustment reaction (309)
Ischemic Heart Disease	Acute myocardial infarction (410), Other acute and subacute forms of ischemic heart disease (411),Old myocardial infarction (412), Angina pectoris (413), Other forms of chronic ischemic heart disease (414)
Congestive Heart Failure	Congestive Heart Failure (428)
Chronic Obstructive Pulmonary Disease (COPD)	Chronic Bronchitis (491), Emphysema (492), Bronchiectasis (494), and Chronic Airway Obstruction, NEC (496)
Asthma	Asthma (493)
Chronic Renal Failure	Chronic renal failure (585)
Osteoarthritis	Osteoarthrosis and Allied Disorders (715)

Appendix F: Extended Results

Appendix F
Extended Results

Included are tables and figures which give more detail about the results included in the main text of the report. These include:

- Tables F1–F5, showing the intermediate calculations of the grouped R-square for service type, beneficiary category, rank, catchment area and primary care manager type (tables F1-F5)
- Figures F1–F2, showing the predicted versus actual FY2002 cost when the total costs are top coded at $25 thousand and $50 thousand (figures F1 and F2)
- Table F6, showing the comparison of predictive accuracy of the CDPS model when run with and without exclusion of lab-only diagnoses from the civilian data (HCSR-N)

Table F1. Grouped R-square for service type

	Model	Army	Air Force	Navy or Marines	Other	All	All (except Other)
N		176,613	160,609	152,389	10,389	500,000	489,611
Total cost, $							
Mean, $		1,769	1,902	1,745	1,347	1,796	1,805
Std. Deviation, $		5,525	5,295	6,384	3,975	5,705	
Mean, predicted cost, $	ACG	1,793	1,858	1,739	1,603		
	CDPS	1,780	1,845	1,759	1,626		
	CRG	1,819	1,849	1,726	1,486		
	DCG	1,809	1,844	1,733	1,530		
Prediction error= (Actual – Predicted), $	ACG	–24	44	7	–256		
	CDPS	–11	57	–13	–279		
	CRG	–50	53	19	–138		
	DCG	–40	58	12	–182		
Contribution to SSE (sum of squared errors) (million $)	ACG	109	319	7	682		
	CDPS	20	525	28	810		
	CRG	437	451	58	199		
	DCG	286	541	24	346		
Grouped R-squared	ACG					0.75	0.81
	CDPS					0.69	0.75
	CRG					0.74	0.59
	DCG					0.73	0.64

Table F2. Grouped R-square for beneficiary category

	Model	Dependent Active Duty/Guard	Retired	Dep. of Retired or Survivor	Active Duty and Guard	All
N		226,327	46,042	88,801	138,830	500,000
Total cost, $						
Mean, $		1,515	2,897	2,538	1,413	1,796
Std. Deviation, $		5,303	8,365	7,696	3,042	5,705
Mean, predicted cost, $	ACG	1,514	2,992	2,595	1,339	
	CDPS	1,529	2,991	2,571	1,322	
	CRG	1,524	2,910	2,517	1,400	
	DCG	1,516	2,993	2,570	1,343	
Prediction error= (Actual – Predicted), $	ACG	2	–95	–57	73	
	CDPS	–14	–94	–33	91	
	CRG	–8	–13	22	13	
	DCG	–1	–95	–32	70	
Contribution to SSE (sum of squared errors) (million $)	ACG	1	414	285	749	
	CDPS	42	404	99	1,159	
	CRG	15	8	41	24	
	DCG	0	419	92	675	
Grouped R-squared	ACG					0.99
	CDPS					0.99
	CRG					0.99
	DCG					0.99

Table F3. Grouped R-square for rank

	Model	Jr Enlist	Sr Enlist	Jr Officer	Sr Officer	Warrant	Other	All	All (except Other)
N		56,408	328,919	35,851	65,638	12,249	935	500,000	499,065
Total cost, $									
Mean, $		1,696	1,826	1,577	1,838	1,819	2,254	1,796	1,795
Std. Deviation, $		4,254	5,785	6,256	6,056	5,424	8,750	5,705	
Mean, predicted cost, $	ACG	1,554	1,849	1,496	1,858	1,962	1,463		
	CDPS	1,570	1,837	1,524	1,878	1,935	1,437		
	CRG	1,567	1,565	1,516	1,831	1,904	1,927		
	DCG	1,622	1,836	1,500	1,850	1,938	1,528		
Prediction error= (Actual – Predicted), $	ACG	142	–22	81	–20	–143	791		
	CDPS	126	–10	52	–39	–115	817		
	CRG	128	12	738	–4	–66	–108		
	DCG	74	–10	77	–11	–119	726		
Contribution to SSE (sum of squared errors) (million $)	ACG	1,141	166	235	26	250	585		
	CDPS	892	35	98	102	163	624		
	CRG	930	5	509	7	284	142		
	DCG	308	30	210	8	173	493		
Grouped R-squared	ACG							0.18	0.33
	CDPS							0.34	0.53
	CRG							0.36	0.50
	DCG							0.58	0.73

Table F4. Grouped R-square for catchment area

	Model	Not in catchment area	In catchment area	All
N		162,569	337,431	500,000
Total cost, $				
Mean, $		1,732	1,826	1,796
Std. Deviation, $		4,989	6,019	5,705
Mean, predicted cost, $	ACG	1,881	1,751	
	CDPS	1,845	1,765	
	CRG	1,742	1,818	
	DCG	1,843	1,767	
Prediction error= (Actual – Predicted), $	ACG	–149	75	
	CDPS	–112	61	
	CRG	–10	9	
	DCG	–111	60	
Contribution to SSE (sum of squared errors) (million $)	ACG	3,598	1,911	
	CDPS	2,044	1,254	
	CRG	16	25	
	DCG	1,987	1,209	
Grouped R-squared	ACG			–4.7
	CDPS			–2.4
	CRG			0.96
	DCG			–2.3

Table F5. Grouped R-square for primary care manager type (in Sept. 2001)

	Model	Military PCM	Civilian PCM	All
N		423,826	76,174	500,000
Total cost, $				
Mean, $		1,784	1,864	1,796
Std. Deviation, $		5,663	5,933	5,705
Mean, predicted cost, $	ACG	1,716	2,224	
	CDPS	1,723	2,168	
	CRG	1,742	1,818	
	DCG	1,724	2,168	
Prediction error= (Actual – Predicted), $	ACG	68	–361	
	CDPS	60	–305	
	CRG	–10	9	
	DCG	60	–304	
Contribution to SSE (sum of squared errors) (million $)	ACG	1,940	9,901	
	CDPS	1,540	7,070	
	CRG	16	25	
	DCG	1,521	7,049	
Grouped R-squared	ACG			–27.7
	CDPS			–19.8
	CRG			0.96
	DCG			–19.7

Figure F1. Actual vs. predicted costs by 2 percent groups with a $50,000 limit imposed

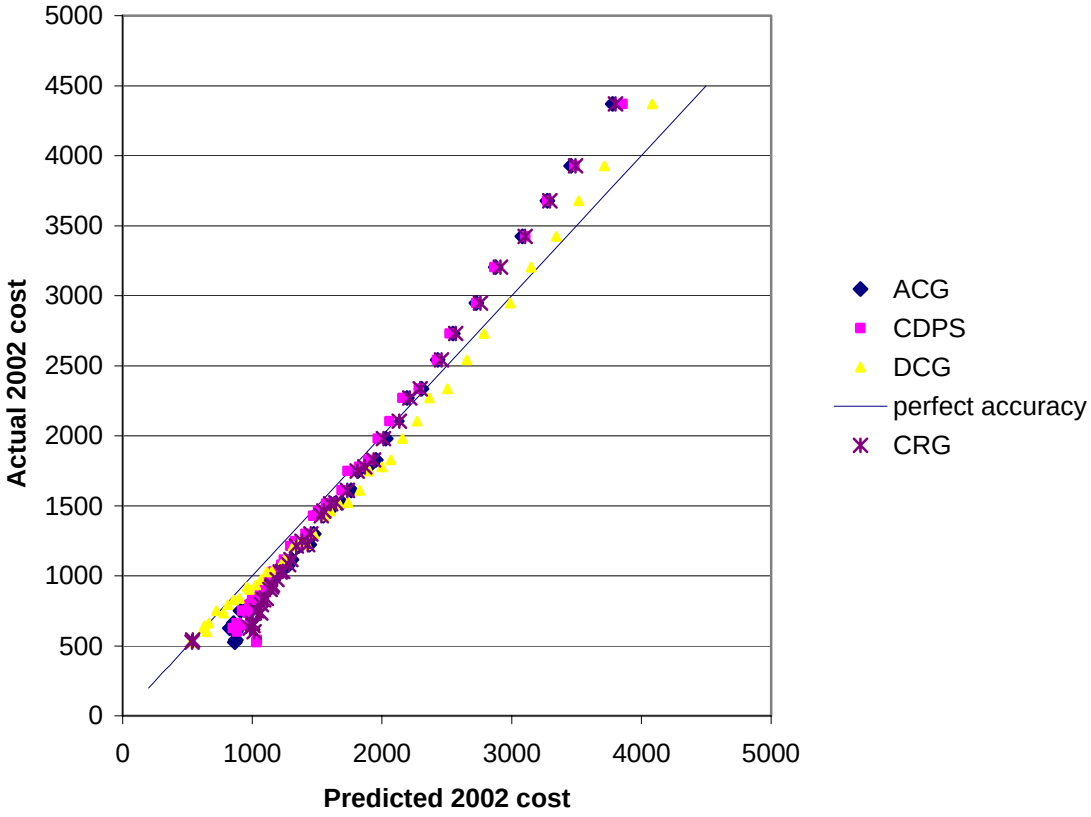


Figure F2. Actual vs. predicted costs by 2 percent groups with a $25,000 limit imposed

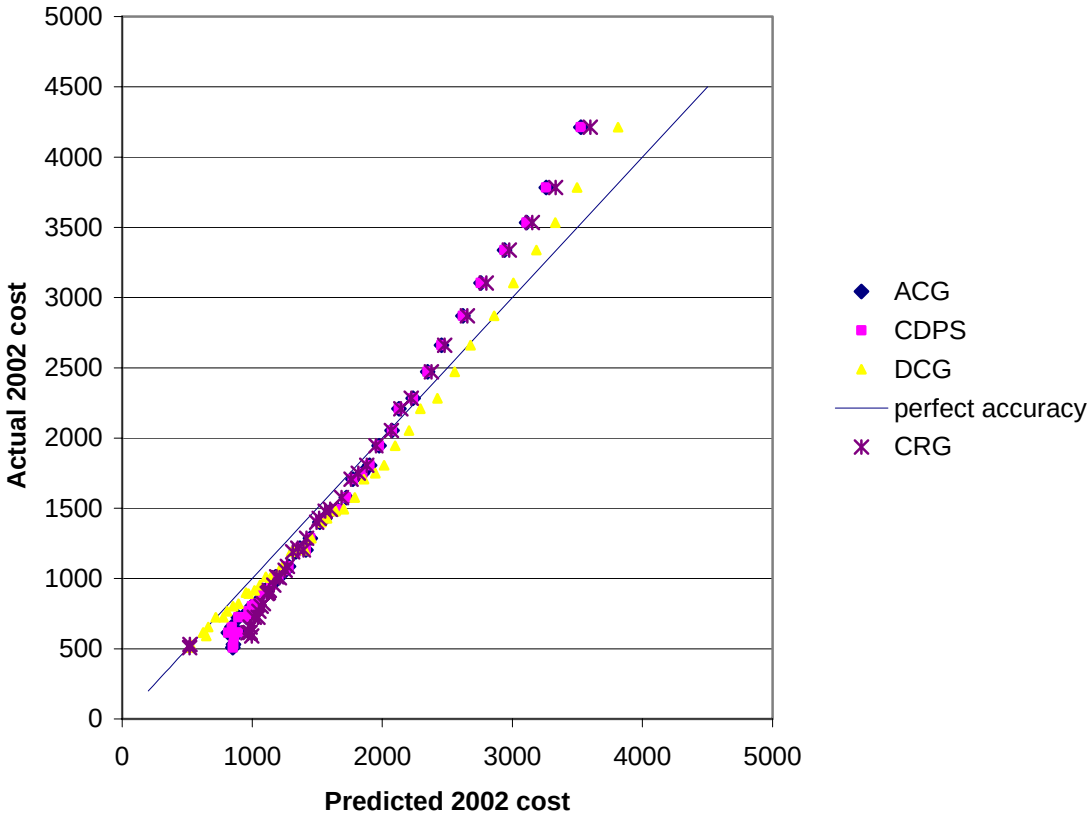


Table F6. Comparing predictive accuracy with lab-only diagnoses dropped (CDPS model only)

	Unaltered expenditures		**Expenditures top-coded at $50K**		**Expenditures top-coded at $25K**	
	Lab-only diagnoses included	Lab-only diagnoses excluded	Lab-only diagnoses included	Lab-only diagnoses excluded	Lab-only diagnoses included	Lab-only diagnoses excluded
R^2	0.1471	0.1475	0.2141	0.2131	0.2371	0.2359
Cumming's Prediction Measure (CPM)	0.1621	0.1615	0.1710	0.1703	0.1765	0.1757
Mean absolute prediction error (MAPE)	$1,632.0	$1,632.8	$1,533.0	$1,534.7	$1,433.0	$1,433.9

Comparisons of actual to predicted FY2002 cost in the validation subset (N=500,000) using risk weights developed on the validation subset (N=1.8 million).

* Better models have higher R^2s and CPMs. In each row-block the model that performs best is bolded. Since CPM = 1 – constant × MAPE, the higher the CPM the lower the MAPE. Thus, CPM and MAPE are different ways of looking at the same thing, not independent measures of model performance