MEPS HC-092:
1996-2004 Risk Adjustment Scores
Public Use File

April 2008

Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
(301) 427-1406

TABLE OF CONTENTS

A. Data Use Agreement

Individual identifiers have been removed from the micro-data contained in these files. Nevertheless, under sections 308 (d) and 903 (c) of the Public Health Service Act (42 U.S.C. 242m and 42 U.S.C. 299 a-1), data collected by the Agency for Healthcare Research and Quality (AHRQ) and/or the National Center for Health Statistics (NCHS) may not be used for any purpose other than for the purpose for which they were supplied; any effort to determine the identity of any reported cases is prohibited by law.

Therefore in accordance with the above referenced Federal Statute, it is understood that:

1. No one is to use the data in this data set in any way except for statistical reporting and analysis; and

2. If the identity of any person or establishment should be discovered inadvertently, then (a) no use will be made of this knowledge, (b) the Director Office of Management AHRQ will be advised of this incident, (c) the information that would identify any individual or establishment will be safeguarded or destroyed, as requested by AHRQ, and (d) no one else will be informed of the discovered identity; and

3. No one will attempt to link this data set with individually identifiable records from any data sets other than the Medical Expenditure Panel Survey or the National Health Interview Survey.

By using these data you signify your agreement to comply with the above stated statutorily based requirements with the knowledge that deliberately making a false statement in any matter within the jurisdiction of any department or agency of the Federal Government violates Title 18 part 1 Chapter 47 Section 1001 and is punishable by a fine of up to $10,000 or up to 5 years in prison.

The Agency for Healthcare Research and Quality requests that users cite AHRQ and the Medical Expenditure Panel Survey as the data source in any publications or research based upon these data.

B. Background

1.0 Household Component

The Medical Expenditure Panel Survey (MEPS) provides nationally representative estimates of health care use, expenditures, sources of payment, and health insurance coverage for the U.S. civilian non-institutionalized population. The MEPS Household Component (HC) also provides estimates of respondents' health status, demographic and socio-economic characteristics, employment, access to care, and satisfaction with health care. Estimates can be produced for individuals, families, and selected population subgroups. The panel design of the survey, which includes 5 Rounds of interviews covering 2 full calendar years, provides data for examining person level changes in selected variables such as expenditures, health insurance coverage, and health status. Using computer assisted personal interviewing (CAPI) technology, information about each household member is collected, and the survey builds on this information from interview to interview. All data for a sampled household are reported by a single household respondent.

The MEPS-HC was initiated in 1996. Each year a new panel of sample households is selected. Because the data collected are comparable to those from earlier medical expenditure surveys conducted in 1977 and 1987, it is possible to analyze long-term trends. Each annual MEPS-HC sample size is about 15,000 households. Data can be analyzed at either the person or event level. Data must be weighted to produce national estimates.

The set of households selected for each panel of the MEPS HC is a subsample of households participating in the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. The NHIS sampling frame provides a nationally representative sample of the U.S. civilian non-institutionalized population and reflects an oversample of blacks and Hispanics. MEPS oversamples additional policy relevant sub-groups such as Asians and low income households. The linkage of the MEPS to the previous year's NHIS provides additional data for longitudinal analytic purposes.

2.0 Medical Provider Component

Upon completion of the household CAPI interview and obtaining permission from the household survey respondents, a sample of medical providers are contacted by telephone to obtain information that household respondents can not accurately provide. This part of the MEPS is called the Medical Provider Component (MPC) and information is collected on dates of visit, diagnosis and procedure codes, charges and payments. The Pharmacy Component (PC), a subcomponent of the MPC, does not collect charges or diagnosis and procedure codes but does collect drug detail information, including National Drug Code (NDC) and medicine name, as well as date filled and sources and amounts of payment. The MPC is not designed to yield national estimates. It is primarily used as an imputation source to supplement/replace household reported expenditure information.

3.0 Survey Management and Data Collection

MEPS HC and MPC data are collected under the authority of the Public Health Service Act. Data are collected under contract with Westat, Inc. Data sets and summary statistics are edited and published in accordance with the confidentiality provisions of the Public Health Service Act and the Privacy Act. The National Center for Health statistics (NCHS) provides consultation and technical assistance.

As soon as data collection and editing are completed, the MEPS survey data are released to the public in staged releases of summary reports, micro data files, and tables via the MEPS web site: www.meps.ahrq.gov. Selected data can be analyzed through MEPSnet, an on-line interactive tool designed to give data users the capability to statistically analyze MEPS data in a menu-driven environment.

Additional information on MEPS is available from the MEPS project manager or the MEPS public use data manager at the Center for Financing Access and Cost Trends, Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850 (301-427-1406).

C. Technical Information

1.0 Data File Content

This documentation describes the 1996-2004 Relative Risk Scores Public use File derived from the respondents to the Medical Expenditures Panel Survey (MEPS) sample for Panels 1 through 9. To obtain analytic variables, the records on this file must be linked to the corresponding MEPS public use data sets by the sample person identifier (DUPERSID).

This file contains a total of 152,589 persons. Each record contains a PANEL indicator which identifies the time period the respondent was in the survey. For example if PANEL=1, the respondent was in the MEPS survey for 1996 and 1997.

2.0 Relative Risk Scores based on the DCG Model in MEPS

A large literature describes methods for estimating the relative propensity to consume health services. These methods are used to adjust for the risk of future utilization when predicting or explaining health care utilization and costs. These "risk adjustment" methods are typically based on diagnostic information from claims data. One well known risk adjustment model, the DCG model, has been developed by researchers at DxCG Inc. DCG prospective relative risk scores (RRSs) are good "generic" measures of disease burden. Studies have shown that people with higher RRS scores go on to use more hospitalizations, ER services and home care, and to experience higher mortality. These scores are widely employed in health policy studies, budgeting, payment, pricing, negotiation, provider profiling, disease management reconciliation, and resource planning.

To add value for health services researchers, AHRQ, in collaboration with DxCG Inc., used diagnosis codes in MEPS to generate a relative risk score for each individual respondent, to enable risk-adjustment for examining future health care spending, and as a general proxy for morbidity due to disease burden. RiskSmart,^{TM} Version 2.2 software was used to calculate relative risk scores.

This Public Use File contains **Relative Risk Scores** for most respondents in Panels 1 through 9 (1996–2004) in the Medical Expenditure Panel Survey. A previous Public Use File (PUF HC-81) contained relative risk scores for respondents in MEPS Panels 1–5. It was subsequently discovered that age had been calculated in error for some respondents in the process of using the risk-adjustment software to produce the data in that PUF. In addition, further refinements have been made in DCG risk adjustment models since this prior PUF was released. Thus, the risk scores for respondents in Panels 1–5 have been recalculated, using updated DCG software and the correct age. The relative risk scores for respondents in Panels 1–5 in the current file supercede and replace the risk scores in PUF81.

The relative risk scores in this PUF have been calculated using DCG models. The DCG model takes diagnostic information, based on claims data, and aggregates specific diagnoses into broader clinically meaningful categories. DxCG's Hierarchical Condition Categories (HCCs) are based on the 5-digit level ICD-9-CM diagnosis codes. Each code is classified into one of 184

condition categories, and hierarchies are further imposed to make predictions more robust to variations in how disease codes are captured, to reward specific coding, and to increase model stability. Thus, to avoid double counting, only the most severe condition in a hierarchy is considered when developing a risk score for an individual. However, the risk models do consider multiple conditions from different hierarchies. Regression models have been developed using large national samples to predict various outcomes. Age, sex, HCCs and interaction terms are included in the models. The individual-level prediction is a relative risk score (RRS). The relative risk score is a summary of disease burden and expected annual health care resource use at the individual level. The RRS can be converted into a dollar prediction by multiplying by an appropriate sample mean. For example, if a reference population has $2000 mean costs, then multiply RRS by $2000. HCC/DCG models are described in several articles referenced at the end of this note.

Users should be aware of several factors that affect the calculation and interpretation of the relative risk scores:

Prospective Models. DCG software provides several different types of models. While all use the same basic DCG framework, the models differ in their details. Some models predict concurrent costs, while prospective models predict future costs. **For developing relative risk scores in MEPS, prospective models were used.** Recall that each Panel in MEPS provides information for a two-year observation period. For a prospective model, information from Year 1 is used to predict costs in Year 2.

The implication is that the relative risk scores are based on diagnostic information reported in Year 1 of a panel. **Any diagnoses that were reported for the first time in Year 2 are not included in the calculation of risk scores.** In the models, age was coded as age at the beginning of the second year of a panel.

Ineligible Cases. Respondents who were not eligible for MEPS in Year 1 (e.g., entered the MEPS during Year 2, such as newborns, people returning from the military) have no diagnostic information from Year 1 and thus have no risk score calculated. There are 5,076 such respondents in the data file; these respondents have a code of "2" for the variable YEARONE. Risk scores for these individuals have been assigned the missing value code of -1.

In addition, relative risk scores were set to the missing value code (-1) for those respondents who had a longitudinal MEPS weight of zero. In panels 1–9, there were 15,784 such cases among respondents younger than 65, and 975 cases among respondents aged 65 or older, for a total of 16,759 cases with a zero weight. Users of the relative risks cores should ensure that cases with codes of -1 are not included in substantive analyses.

ICD-9 Coding Level. The MEPS public use data contain 3-digit level ICD-9-CM diagnosis codes. However, the 5-digit level ICD-9 codes were used when calculating the risk scores.

DxCG Inc. staff have examined how using 3-digit diagnoses (rather than 5-digit codes) would affect the prospective DCG/HCC model's performance. They concluded that, although using 3-digit codes would reduce the model's specificity in clinical classification and its predictive accuracy, the loss in specificity and predictive power was small.

Insurance Coverage. Insurance coverage presents a complication in applying DCG models to the MEPS data. DCG models have been developed using linear regression on large national claims datasets from particular insurers. Different models have been developed for different datasets: One risk adjustment model was derived for Medicare claims, another for claims for privately insured individuals, and a third for Medicaid claims data. While the majority of MEPS respondents have one source of insurance coverage during a calendar year, people can be uninsured, and they can change insurance coverage during a year. To accommodate this complexity, we developed a variable that represents the **predominant** form of coverage for each respondent during Year 1 of the Panel. This variable, INSCAT1, has four categories:

1 Medicare
2 Private
3 Medicaid
4 Uninsured

Respondents were assigned to a category based on the number of months of each type of coverage (or no coverage) during the first panel year. Thus, if someone had seven months of private coverage and five months of Medicare, the person was coded as private (INSCAT1 = 2). If someone had equal months of coverage for two or more different sources, their classification was based on the following hierarchy: Medicare, private, Medicaid, uninsured.

The DCG models were developed to predict health care costs. **Note that costs refer to the kinds of costs covered within an insurance system.** Thus, for example, a person with high long term care costs may look less expensive to a Medicare model (since Medicare does not pay long term care costs) than he or she would to a Medicaid model (which does pay such costs).

For those familiar with DxCG Inc. RiskSmart ™ software (Stand-Alone Version 2.2), we used the following specific model options:

Commercial: Model 26
Medicare: Model 3
Medicaid: Model 64

Additional investigation by staff at DxCG Inc. showed that, for respondents who were uninsured, the commercial model provided the best prospective prediction of costs, compared with the Medicare or Medicaid models.

Users should note that the Medicaid model excludes individuals who are aged 65 or older. There are 17,196 such respondents in Panels 1–9. These respondents have been assigned the missing value code of -1 for risk scores based on the Medicaid model. **Users should deal with these missing values appropriately in their analyses.**

Age/Sex and HCC Specifications. Within each type of DCG model (Medicare, private, and Medicaid) there are two model specifications: A basic model includes only information on the person's age and sex ("age/sex" or "A/S" model), and a more elaborate model also includes information on the HCCs (in addition to age and sex), based on medical conditions reported for

each respondent in MEPS. This file includes relative risk scores from both the A/S specification and the HCC specification.

To provide maximum flexibility and information for users of MEPS data, each of the three established DCG prediction models (Medicare, private, and Medicaid) was applied to each MEPS respondent, regardless of the person's insurance status. Thus, **six** relative risk scores, based on a combination of model type (Medicare, private, and Medicaid) and model specification ("A/S" only or age/sex and HCCs), have been produced for each person. (Respondents aged 65 and older have been assigned a missing value code of -1 for age/sex and HCC Medicaid model scores.)

Table 1 shows the variable names, corresponding to the models used to implement the DCG prediction, and the inputs used in each model.

Table 1 – Prospective DCG Relative Risk Scores in MEPS

DCG Risk Score Name (in DxCG, Inc. software)	Model Type*	YEARONE Model Inputs
RRSASMC	A/S_Medicare	Age, Sex
RRSHCCMC	HCC_Medicare	Age, Sex, Diagnoses
RRSASPV	A/S_Private	Age, Sex
RRSHCCPV	HCC_Private	Age, Sex, Diagnoses
RRSASMD	A/S_Medicaid	Age, Sex
RRSHCCMD	HCC_Medicaid	Age, Sex, Diagnoses, Eligibility Categories

* "A/S" refers to models based on age and sex alone. "HCC" stands for the Hierarchical Condition Category modeling framework that organizes diagnostic information into profiles, which, in conjunction with demographic data, are used (in these prospective models) to predict next year's health care cost. The second part of each type name refers to the population on which the model was originally derived: Medicare, commercially (privately) insured, or Medicaid.

Normalization

Risk scores are "made relative" by multiplying by a normalizing constant, chosen so that the scores average to 1.00 within specified MEPS subpopulations. Thus, relative risk scores (RRSs) are normalized, positive predictions of future (prospective) total health care spending, where a score of 1 refers to a person whose expected costs next year are "average" in a specified population. Regardless of how they are normalized, relative risk scores convey relative expected costliness, so that, when applying the same model to any group of people under a given type of health care benefit, RRS = 1.5 indicates expected costs 50% higher than RRS = 1.0.

For the MEPS data, a separate normalization was performed for each combination of panel and INSCAT1. Table 2 shows the standard RRSs produced by the DCG modeling software, for each combination of panel and INSCAT1, prior to normalization. The entries are the mean RRS for each cell; in calculating the mean, data were weighted by the analytic weight derived for longitudinal analyses of each panel (LONGWT). For example, we applied DxCG's HCC Medicare model to all (n = 147,512) members of MEPS panels 1 through 9 who were eligible in Year 1, producing the "standard" (i.e., not normalized) Medicare relative risk scores. People without Medicare coverage received a risk score. The mean of these scores, among only the (n =

2,566) people in the MEPS panel 1 subgroup with INSCAT1 = Medicare and a positive longitudinal weight, was calculated as 0.5840127 (see Table 2). Similarly, the mean standard (not normalized) Medicare relative risk score among only the (n = 1,722) people in the MEPS panel 2 subgroup with INSCAT1 = Medicare and a positive longitudinal weight was 0.6034610.

Table 2 – Conversion Factors (numbers needed to multiply by to recover the original DxCG-model risk score values)

DCG/HCC Model

Panel	Private	Medicare	Medicaid	Uninsured
1	0.8667726	0.5840127	0.3895308	0.7180602
2	0.8679859	0.6034610	0.3899222	0.7176932
3	0.8365996	0.5778647	0.3729121	0.6658390
4	0.8264264	0.5832314	0.3669336	0.6896421
5	0.8585728	0.5626942	0.3489640	0.7303295
6	0.9129664	0.6050517	0.3566948	0.7133616
7	0.9258927	0.6105160	0.3736620	0.7540057
8	0.9425828	0.6274948	0.3505671	0.7520604
9	0.9738929	0.6191015	0.3430512	0.8053602

Age/Sex Model

Panel	Private	Medicare	Medicaid	Uninsured
1	0.9078039	0.9945072	0.4246647	0.8031818
2	0.9166051	0.9910251	0.4251920	0.7916005
3	0.9207479	0.9900677	0.3995558	0.7017388
4	0.9326035	0.9925572	0.4012468	0.8195732
5	0.9315218	0.9878520	0.3816353	0.8330917
6	0.9470494	0.9980819	0.3927610	0.8422467
7	0.9574624	1.0033320	0.4009821	0.8523306
8	0.9703112	1.0033437	0.3975427	0.8747223
9	0.9689793	0.9925472	0.4095152	0.8845804

The mean standard (not normalized) RRSs were then used to normalize the individual relative risk scores, by panel and INSCAT1. Thus, all Panel 1 relative risk scores based on the DCG Medicare model (n = 19,529, including everyone in panel 1, regardless of insurance, if LONGWT was >0 and YEARONE = 1) were divided by 0.5840127 to produce the variable labeled RRSHCCMC for panel 1. Similarly, all Medicare relative risk scores in panel 2 were divided by 0.6034610, to create the RRSHCCMC score for panel 2. Thus, the average RRSHCCMC score for panel 1 people in Medicare (INSCAT1=1) is 1, and the average RRSHCCMC score for panel 2 people in Medicare is also 1. This process was repeated for each of the other panels. The overall process was then repeated for the DCG private model, yielding the variable RRSHCCPV, and for the Medicaid model, yielding RRSHCCMD.

In other words, within each combination of panel and INSCAT1, the average risk score is normalized to 1.000. This allows researchers to conduct analysis by panel or by insurance coverage type across panels or both.

The following normalized risk scores are thus included in the file:

RRSHCCPV –Normalized RRS, HCC model, private
RRSASPV – Normalized RRS, age-sex model, private

RRSHCCMC – Normalized RRS, HCC model, Medicare
RRSASMC – Normalized RRS, age-sex model, Medicare
RRSHCCMD – Normalized RRS, HCC model, Medicaid
RRSASPMD – Normalized RRS, age-sex model, Medicaid
RRSHCCUN – Normalized RRS, HCC model, uninsured
RRSASUN – Normalized RRS, age-sex model, uninsured

Because persons with a longitudinal weight of zero and those with no data from the first year of a panel are excluded from calculations of normalized relative risk scores, there are 18,931 cases with a missing value code (-1) for the normalized RRSs for private, Medicare, and uninsured. The number of cases with missing values rises to 35,278 for Medicaid normalized RRS, due to exclusion of people 65 and older.

If a researcher wants to convert the relative risk scores to dollar predictions, he/she needs to multiply the average expenditure for a combination of panel and INSCAT1 by the relative risk score for that combination. To move from a relative prediction to a dollar prediction for a person in any of these three insured populations, multiply the risk scores by the average expenditure for the corresponding panel*INSCAT1, as given in Table 3.

The HCC private insurance model predicts subsequent costs best (in terms of R-squared) for the uninsured. To create dollar predictions (that match the observed costs) for an uninsured respondent in a panel, you can multiply the RRSHCCPV relative risk score for an uninsured respondent in a panel by the mean observed cost for uninsured respondents in that panel.

Some users might prefer to use a different normalization procedure than the one used here. To accommodate this possibility, the file also includes 6 (insurance type by model specification) "standard" risk scores prior to normalization. These are

HCCMC – Not normalized risk score, HCC model, Medicare
HCCMD – Not normalized risk score, HCC model, Medicaid
HCCPV – Not normalized risk score, HCC model, commercial
ASMC – Not normalized risk score, age/sex model, Medicare
ASMD – Not normalized risk score, age/sex model, Medicaid
ASPV – Not normalized risk score, age/sex model, commercial

(The means of these scores, by INSCAT1 and Panel, appear in Table 2.)

Table 3 – Average Expenditure by Panel and Insurance Category (INSCAT1)

Panel	Private	Medicare	Medicaid	Uninsured
1	$1,727.24	$5,568.79	$1,641.08	$682.56
2	$1,726.42	$6,044.14	$1,600.56	$661.21
3	$1,680.42	$6,232.56	$1,583.85	$629.04
4	$1,569.08	$6,905.62	$1,684.28	$745.46
5	$1,776.17	$5,789.48	$1,816.00	$869.41
6	$2,117.51	$7,233.54	$2,131.74	$844.78
7	$2,329.69	$8,213.69	$2,443.24	$1,007.16
8	$2,575.74	$8,806.04	$2,289.97	$965.64
9	$2,928.18	$9,826.39	$2,125.23	$1,055.05

D. DxCG Bibliography

Publications by DxCG Senior Scientists

Zhao Y, A Ash, RP Ellis et al	"Predicting Pharmacy Costs and Other Medical Costs Using Diagnosis and Drug Claims." Medical Care 43 (1): 34–43, January 2005.
Pope G, J Kautter, RP Ellis, A Ash, J Ayanian, et al	"Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC model." Health Care Financing Review. Summer 2004.
Ellis RP, MS Kramer, JF Romano, R Yi	"Applying Diagnosis-based Predictive Models to Group Underwriting." Health Section News; August 2003. 1, 4–8.
Zhao Y, A Ash, J Haughton, B McMillan	"Identifying Future High-Cost Cases Through Predictive Modeling." Disease Management and Health Outcomes 2003; 11(6): 389–397.
Zhao Y, A Ash, RP Ellis, et al	"Disease burden profiles An Emerging Tool for Managing Managed Care." Health Care Management Science (2002); 5(3) 211–219.
Shen, Y, RP Ellis	"Cost-Minimizing Risk Adjustment." Journal of Health Economics. (2002) 21(3) pp 515-530.
Shen, Y, RP Ellis	"How Profitable is Risk Selection? A Comparison of Four Risk Adjustment Models." Health Economics. (2002) 11(2) 165–174
Ash, A, Y Zhao, RP Ellis, MS Kramer	"Finding Future High-cost Cases Comparing Prior Cost Versus Diagnosis-based Methods." Health Services Research 36(6) Part II December 194–206. (2001), http//www.hsr.org/AliceHersh/12-Ash.pdf
Zhao Y, Ellis RP, A Ash, et al	"Measuring Population Health Risks Using Inpatient Diagnosis and Outpatient Pharmacy Data". Health Services Research. 36(6) Part II December 180–193. (2001) http//www.hsr.org/AliceHersh/11-Zhao.pdf
Ash, A, F Porell, L Gruenberg, et al	"Adjusting Medicare Capitation Payments Using Prior Hospitalization." Health Care Financing Review 10(4) 17–29, 1989.
Ellis, RP, A Ash	Refinements to the Diagnostic Cost Group Model. Inquiry 32 1–12, Winter 1995.
Ellis, RP, G Pope, et al	"Diagnosis-Based Risk Adjustment for Medicare Capitation Payments." Health Care Financing Review, Spring 1996.
Pope, G, CF Liu, RP Ellis, A Ash et al	"Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment." Health Care Financing Review, (2000) Spring 21 (3) 93–118.
Pope, G et al	"Evaluating Alternative Risk Adjusters for Medicare," Health Care Financing Review, 1998.