Published in final edited form as:

J Surv Stat Methodol. 2020 June ; 9(3): 598–625. doi:10.1093/jssam/smz047.

COMBINING INFORMATION FROM MULTIPLE DATA SOURCES TO ASSESS POPULATION HEALTH

TRIVELLORE RAGHUNATHAN[*] **[Professor of Biostatistics, Research Professor]**,
Department of Biostatistics, 1415 Washington Heights, University of Michigan, Ann Arbor, MI 48109; Survey Research Center, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48106.

KAUSHIK GHOSH [Research Specialists],
National Bureau of Economic Research (NBER), 1050 Massachusetts Ave, Cambridge, MA 02138.

ALLISON ROSEN [Associate Professor],
Department of Quantitative Health Sciences University of Massachusetts Medical School 368 Plantation Street, AS9-1083, Worcester, MA 01655; NBER.

PAUL IMBRIANO [former graduate student],
Department of Biostatistics, 1415 Washington Heights, University of Michigan, Ann Arbor, MI 48109.

SUSAN STEWART [Research Specialists],
National Bureau of Economic Research (NBER), 1050 Massachusetts Ave, Cambridge, MA 02138.

IRINA BONDARENKO [Statistician Lead],
Department of Biostatistics, 1415 Washington Heights, University of Michigan, Ann Arbor, MI 48109.

KASSANDRA MESSER [Restricted Data Architect],
Survey Research Center, Institute for Social Research, 426 Thompson Street, University of Michigan, Ann Arbor, MI 48106.

PATRICIA BERGLUND [Senior Research Associate],
Survey Research Center, Institute for Social Research, 426 Thompson Street, University of Michigan, Ann Arbor, MI 48106.

JAMES SHAFFER [Senior Statistician],
IQVIA, 4820 Emperor Blvd Durham, NC 27703.

DAVID CUTLER [Professor of Economics]
Department of Economics, Harvard University, 1805 Cambridge St, Cambridge, MA 02138; NBER

Abstract

[*]Address correspondence to Trivellore Raghunathan, Department of Biostatistics, 1415 Washington Heights, University of Michigan, Ann Arbor, MI 48109, USA; teraghu@umich.edu.

Information about an extensive set of health conditions on a well-defined sample of subjects is essential for assessing population health, gauging the impact of various policies, modeling costs, and studying health disparities. Unfortunately, there is no single data source that provides accurate information about health conditions. We combine information from several administrative and survey data sets to obtain model-based dummy variables for 107 health conditions (diseases, preventive measures, and screening for diseases) for elderly (age 65 and older) subjects in the Medicare Current Beneficiary Survey (MCBS) over the fourteen-year period, 1999–2012. The MCBS has prevalence of diseases assessed based on Medicare claims and provides detailed information on all health conditions but is prone to underestimation bias. The National Health and Nutrition Examination Survey (NHANES), on the other hand, collects self-reports and physical/ laboratory measures only for a subset of the 107 health conditions. Neither source provides complete information, but we use them together to derive model-based corrected dummy variables in MCBS for the full range of existing health conditions using a missing data and measurement error model framework. We create multiply imputed dummy variables and use them to construct the prevalence rate and trend estimates. The broader goal, however, is to use these corrected or modeled dummy variables for a multitude of policy analysis, cost modeling, and analysis of other relationships either using them as predictors or as outcome variables.

Keywords

Calibration; Measurement error; Multiple imputation; Propensity scores

1. INTRODUCTION

Information about the prevalence of an extensive set of diseases and preventive measures (such as well care, influenza vaccination, mammography screening, etc.) and a rich set of covariates on a well-defined sample of subjects from a target population of interest are essential for many kinds of analyses in health care policy and the science of health. Ideally, prevalence of diseases and preventive measures (generally, labelled as health conditions) will be represented by a collection of dummy variables indicating its presence (=1) or absence (=0). This type of statistical infrastructure can enable the society to continuously monitor population health at the macroscopic level. The goals of the analyses using such a statistical infrastructure, for example, may be to assess the health of the population through estimates of prevalence rates and the trend analysis, assess the impact of prevalence rates on the cost structure, study the impact of various policies, and investigate health disparities.

Unfortunately, there is no single data source that provides a rich collection of such dummy variables, and therefore, several sources must be brought together to construct the aforementioned statistical infrastructure. The national omnibus surveys such as the National Health and Nutrition Examination Survey (NHANES), the National Health Interview Survey (NHIS), and the Behavior Risk Factor Surveillance System (BRFSS) collect data on self-report for ever having a set of health conditions and clinical measures. For example, NHANES collects both self-report health conditions and laboratory measures that could be used to construct dummy variables, but the number of questions asked about diagnosis with specific health conditions is limited and may not be asked every year or administered to everyone in the sample. However, to the extent available, these surveys are the best source at the national level.

An alternative source is administrative data derived, for example, from providers of care. For the population age 65 and older, the Medicare Current Beneficiary Survey (MCBS), described in section 2, provides information on the prevalence of all health conditions of interest through linking with the administrative data on Medicare claims and using the International Classification of Diseases (ICD)-9 codes to define them. Also, this data source covers both institutionalized and noninstitutionalized populations.

Administrative data, however, have a number of issues. First, they provide information only on those individuals who sought care for the health condition in the billing year. Second, care is typically grouped during routine visits and may not always be identified separately in the claim data files. Third, information is lost through capitated billing practices where individual claims containing diagnosis and procedure codes are not filed for each encounter with the health care system. Finally, for acute conditions, such as acute myocardial infarction (AMI) or hip fracture, the claim data files may capture events that occurred during the recent past but not any distant past events (except for any potential follow-up). That is, the claims may fail to capture whether a subject “ever” had certain health conditions. Since our interest is in the prevalence of health conditions (defined as a subject ever having a particular health condition), just relying on the administrative data sources may not be an ideal strategy.

Several studies have investigated the differences between prevalence rate estimates from self-report and administrative data sources (example, Zuvekas and Olin 2009; Robinson, Young, Roos, and Gelskey 1997; Okura, Urban, Mahoney, Jacobsen, and Rodeheffer 2004; Muggah, Graves, Bennett, and Manuel 2013; O’Donnell, Vanderloo, McRae, Onysko, Patten et al. 2016; Yasaitis, Berkman, and Chandra 2015). These studies found disagreements between self-reported and claim-identified events, especially in chronic conditions and also report on validation studies where the self-reports have been shown to capture prevalence rates adequately.

To assess the extent of imperfectness of the claims-based prevalence rates, we compare three sets of estimates: self-reports from MCBS, claims-based (i.e., defined by ICD-9 codes) from MCBS, and self-reports from NHANES across the fourteen-year period, 1999–2012 (not all disease estimates are available from each of these resources). As an example, the claims-based prevalence rates from MCBS (1999, 2009) differ considerably from the self-report prevalence rates in NHANES (1999–2000, 2009–2010) or MCBS (1999, 2009). The claims-based prevalence rate of hyperlipidemia in 1999 is 27.9 percent, whereas the self-report rate from NHANES is 45.2 percent. The corresponding rates in 2009 are 51.4 percent and 62.4 percent, respectively. For acute myocardial infarction (AMI), the claims-based prevalence rate estimate in 2009 is 2.25 percent, whereas the self-report rate in NHANES (2009–2010) is 8.58 percent. A portion of this difference may be due to claims capturing the incidence of AMI during the year, as opposed to ever having AMI captured in the self-report. Another example is depression; the prevalence rate estimate based on claims alone is 7.6 percent, whereas the self-report rate in MCBS (2009) is 20.9 percent.

Let p_C be the design-weighted estimate of the prevalence rate of a health condition based on the claims-based dummy variable and p_S be the corresponding design-weighted estimate based on self-report. If $p_C < p_S$, then the claims-based estimate may suffer from underestimation bias. Define the effect size of underestimation bias as $\delta = (p_C - p_S) / \sqrt{p_S(1 - p_S)}$. There are 374 health conditions (out of a total of 1,498 [= 107 × 14], possible health conditions) for which both claims-based and self-report-based prevalence rate estimates could be constructed. Approximately 62 percent of the 374 effect sizes are negative, with an average effect size of −0.17 (for $\delta < 0$) and a maximum effect size of −0.5. As an alternate, consider the effect size using the transformed scale, $h = 2(sin^{-1}\sqrt{p_C} - sin^{-1}\sqrt{p_S})$ (Cohen 1988). The median value among $h < 0$ is −0.28, and 25 percent of the effect sizes h are less than −0.4 (note that the larger the magnitude of the negative numbers, the more potential for underestimation bias). Given the large sample size, almost all these differences are statistically significant. For the remaining 38 percent of the health conditions, where $p_C \geq p_S$, the effect size ranges from 0.02 to 0.09. Thus, the underestimation bias in the claims-based estimates may be a more important issue than the potential for overestimation.

Another limitation of the Medicare claim file is lack of information on subjects obtaining medical care through enrollment into a health maintenance organization (HMO, also known as the Medicare Advantage Program). The Medicare program in most cases reimburses the HMOs on a per-patient basis and does not require reporting of any specific health conditions or medicalspending information. The prevalence dummy variables, therefore, cannot be constructed for such individuals, and the number of individuals getting care through HMOs has been increasing over the fourteen-year period (1999–2012). Thus, the dummy variables purely based on Medicare claims, though available for all 107 health conditions, are subject to underestimation and potential bias due to ignoring HMO enrollees.

Using MCBS as the primary data source, we developed and implemented a two-pronged approach for correcting these biases. The first correction uses a propensity score model to weight the respondents in the survey who get care purely from Medicare Parts A and B to compensate for not using HMO enrollees in the study population. This is similar to weighting for unit nonresponse (or selection weighting) where the respondents (“pureMedicare”) are assigned a weight to compensate for the nonrespondents (“HMO enrollees”) based on the similarity of the covariates between the two groups.

For the second correction, we use the omnibus survey, NHANES, as an external source for developing a model-based calibration and refinement of the claims-based dummy variables. We developed a combination of missing data and measurement error modeling frameworks to construct 107 multiply imputed corrected health condition dummy variables for each survey year (1999–2012) on to the primary data source—MCBS data sets. The work reported builds on and extends the earlier work (Raghunathan 2006; Schenker and Raghunathan 2007; Schenker, Raghunathan, and Bondarenko 2010).

We divide the task into several steps depending upon the information available for each of the 107 health conditions:

1. Setup: We carry out the correction or calibration separately for each year to avoid imposing any time structure in the imputation/measurement error models. Furthermore, this stratified process is computationally simpler when more years of data are included in the future (currently, this method is being extended for 2013–2017 data sets).

2. Data preparation: We use a total of seven NHANES (two-year cycles) and fourteen MCBS surveys in this project, and every survey had item missing values. We multiply impute these missing values using the sequential regression or chained equations approach (Kennickel 1991; van Buuren and Oudshoorn 1999; Raghunathan, Lepkowski, van Hoewyk, and Solenberger 2001) using the software IVEware (Raghunathan, Solenberger, and Van Hoewyk 2002). Section 2 provides details about this step.

3. Health maintenance organization weight construction: As indicated earlier, subjects enrolled in HMOs have no information from claims about various health conditions in the MCBS survey data and, hence, must be excluded. Given that MCBS is a nationally representative sample of Medicare beneficiaries age 65 and older, the goal is to maintain the representativeness to the extent possible. We constructed weights for the pure Medicare sample using a propensity score model and thus generalize the inferences to the full MCBS sample. Section 3 provides details about this step.

4. Calibration of health conditions with NHANES self-report for noninstitutionalized population: For about 25 percent of the health conditions, the self-report data are available in NHANES. We use a missing data framework that combines NHANES and MCBS to derive multiply imputed corrected/calibrated claims-based dummy variables in MCBS. The calibration refers to a constraint, where the corrected multiply imputed prevalence rates are to match the self-report NHANES prevalence rates. Section 4 provides details about this step.

5. Calibration of health conditions with no NHANES self-report for noninstitutionalized population: For the remaining health conditions (i.e., health conditions for which no self-report data is available in NHANES), we develop a regression model to relate the claims-based and calibrated/corrected claims-based definitions of the dummy variables obtained in step four. This can be viewed as a measurement error model, with calibrated claims-based dummy variables as an accurate measure and claims-based dummy variables as mismeasured. For those conditions not available in NHANES, we applied this measurement error model to obtain calibrated claims-based dummy variables. Section 5 provides details about this step.

6. Calibration of health conditions for institutionalized population: There are two different subpopulations among the elderly: noninstitutionalized (or community dwelling) and institutionalized (about 10 percent of the sample). The NHANES covers only the noninstitutionalized population, whereas MCBS covers both. Steps four and five, therefore, use only the noninstitutionalized portion of the

MCBS and the corresponding NHANES. Though the institutionalized sample is relatively small, it is an important component given the poorer health of those in institutions such as skilled nursing facilities. After obtaining calibrated dummy variables for the noninstitutionalized population as described in the previous steps, we developed imputation and regression models to extrapolate for the institutionalized population. For this extrapolation, we grouped the institutionalized and noninstitutionalized subjects based on a set of covariates using a propensity score model. We then developed regression models (similar to those described in step five) for the noninstitutionalized population in each matched group and then apply the estimated model to the institutionalized population in the same group. This approach may be viewed as an "hot-deck" approach where the respondents (non-institutionalized) and nonrespondents (institutionalized) are matched based on covariates and the models constructed for the respondents are then applied for the nonrespondents. This step is described in section 6.

7. Analysis of calibrated data: At the end of this process, we derived five imputed data sets with 107 calibrated-claim disease dummy variables for each subject in the MCBS survey data. We analyzed the claim, calibrated claim, and self-report prevalence rates to inspect trends. All analyses on each imputed data set incorporated complex design features (weighting, clustering, and stratification) and then combined across the five imputed data sets using the standard multiple imputation combining rules (Rubin 1987). All imputation and measurement error models include design variables as predictors.

8. Internal and external validation: Given all the complex modeling tasks, it is important to validate the results. For the internal validation, we routinely used model diagnostics and goodness of fit assessments as an integrated process with the model development. For external validation, we use all self-report health conditions in MCBS that were set aside and not used in the derivation of model-based dummy variables. That is, our pretense is that no self-report data are available in MCBS. This allows us to compare model-derived estimates (MCBS Claims + NHANES self-report) with the "direct" MCBS self-report estimates for each of the fourteen years of data used in this project. Analyses, including validation, are discussed in section 7.

9. Obviously, the task of building such an infrastructure using multiple data sources involves assumptions. Finally, we conclude in section 8, with the discussion of limitations and future work.

2. DATA SOURCES

2.1 MCBS and NHANES

We briefly describe the data sources used in this article. The MCBS is a survey of a nationally representative sample of Medicare beneficiaries including the aged, disabled, and institutionalized populations (approximately 6,500 noninstitutionalized and 700 institutionalized respondents per year). These survey data are linked to Medicare administrative claim data. The claim data includes information about medical care received for inpatient and outpatient hospital care, physician services, home health care, durable medical equipment, skilled nursing home services, and hospice care, including diagnosis (ICD-9-CM) and procedure codes (HCPCS, CPT). We created two separate data sets, one for the noninstitutionalized or community dwelling subjects and another for the institutionalized subjects in MCBS. All data sets are from the Center for Medicare Services (CMS) obtained under a data use agreement.

The data sets for correcting the claims-based dummy variables in MCBS are from the National Health and Nutrition Examination Surveys conducted during the same period, 1999–2012. The NHANES (CDC 1999–2012) data sets are publicly available and collected on a nationally representative survey of child and adult participants from the noninstitutionalized population in the United States. The current project, however, uses a subset of individuals who are 65 years of age or older (approximate sample size of 6,500 per two-year cycle). Both sets of surveys include information on demographic and socioeconomic characteristics, self-reported health status and functioning, health insurance status, and several other covariates. In addition, only MCBS has data on health care utilization and expenditures.

There are several common demographic variables and covariates in NHANES and MCBS. Tables 1 and 2 provide variables available across all fourteen years and the descriptive statistics for the common variables comparing NHANES (2009–2010) and MCBS (2009) as an illustration. The MCBS descriptive statistics include both noninstitutionalized and institutionalized populations. In addition, the imputation modeling includes a number of variables, including the expenditure variable from MCBS. That is, we use as many covariates as possible in the modeling process.

The covariate distributions between the two surveys are quite similar. Some notable differences include MCBS respondents being slightly older, more often reporting having a private insurance, having more difficulties with functioning and self-care, and more often being single or widowed. Some of these differences may be a result of including institutionalized subjects and excluding HMO enrollees (in spite of weighting adjustments) in MCBS, whereas NHANES has no such inclusions or exclusions. Nevertheless, we constructed a propensity score model analysis to assess the differences in the distribution of the covariates between the two surveys. Using a well-fitting logistic regression model with a dummy variable ($M = 1$ for MCBS and $M = 0$ for NHANES) as the dependent variable and the covariates listed in tables 1 and 2, as predictors, we obtain the predicted probability of being in the MCBS for each subject in the combined data set. Histograms of the propensity scores for the subjects in the two surveys show considerable overlap with no discernable differences across all fourteen years.

2.2 Definition of Health Conditions

We use the information provided by the Agency for Health Care Research and Quality (AHRQ), Health Care Cost and Utilization Project (HCUP), and Clinical Classification Software (CCS) for ICD-9-CM as the classification schema for chronic diseases and medical conditions of interest. The CCS collapses over 14,000 diagnosis codes and 3,900 procedure codes into a much smaller number of clinically meaningful CCS categories. While CCS codes are not provided on the MCBS claim files, ICD-9-CM codes can be mapped to the CCS categories via the CCS mapping software provided as a public use file by HCUP on the internet. The mapping uses the full five-digit ICD-9-CM diagnosis codes which are provided in the Medicare claim files.

Our project physicians identified a few conditions by ICD-9-CM code that split the larger CCS categories because they (1) should be stand-alone disease categories because of clinical significance or (2) should be grouped in a disease category different than assigned by the CCS (most commonly due to the significant changes to the mental health categories assigned by the CCS with the 2009 data release to more accurately reflect the Diagnostic and Statistical Manual of Mental Disorders disease classifications). In addition, relevant literature led to identification of the most exhaustive list of codes to capture the screening and preventative services.

The ICD-9-CM diagnosis and procedure codes, Health Care Procedural Coding System (HCPCS), the Current Procedural Terminology (CPT) codes and CCS mapping file along with our project physicians' clinical expertise and the extensive data management and analytic investigations of all Medicare Claim files for every MCBS subject lead to the mutually exclusive, collectively exhaustive categories of the 107 health conditions. Table 3 provides definitions of 107 health conditions ultimately used in the analysis.

2.3 Imputation of Covariates

Covariates listed in tables 1 and 2 and self-reported medical conditions are missing for some subjects in both NHANES and MCBS data files. Table 4 provides frequency distributions of missing values in the total of 1,400 variables in the two data files across the fourteen years.

We multiply impute the missing values using the sequential regression multiple imputation (SRMI) procedure as implemented in the software package IVEWARE. Sequential regression multiple imputation is an iterative procedure in which the missing values in each variable are imputed, conditional on all other variables, using appropriate regression models. Random draws from an approximate predictive distribution of the missing values under these models are then used as imputations. We routinize diagnostics checks described in Bondarenko and Raghunathan (2016) as the integral part of the SAS software code (SAS VERSION 9.4, SAS Institute, Cary, NC). The rates of missing values are generally highest in the MCBS institutionalized population. We impute the two variables in MCBS, with the missing value rates near 30 percent, by combining the noninstitutionalized and institutionalized MCBS populations to increase the stability of the regression coefficients in the imputation model.

3. DEVELOPMENT OF WEIGHTS

As indicated earlier (see step three in section 1), a number of respondents in MCBS obtain care through a health care maintenance organization for part of or the full year. These subjects have incomplete information on the claims, so the claims-based dummy variables are not available for them. Of course, ignoring these individuals may introduce a selection bias. We developed a selection weight adjustment to compensate for not using these subjects in the analysis, assuming there is sufficient overlap in the covariate distribution of these two groups and there are no unmeasured confounding variables. This assumption is akin to the “missing at random” assumption in the missing data framework (i.e., conditional on the covariates listed in tables 1 and 2, the subjects enrolled in the HMOs are similar to those who receive care purely from Medicare).

We define a selection indicator or dummy variable taking the value one for inclusion in the study and zero otherwise. Two conditions have to be met for inclusion in the study: (1) subjects obtained care only through Medicare Parts A and B for the full twelve-month period (unless died); and (2) subjects did not participate in any HMO (Medicare Advantage Program) (i.e., the included subjects were pure Medicare participants). We develop a logistic regression model with the selection indicator as the dependent variable and the covariates listed in tables 1 and 2, in addition to some selected interaction effects as predictors. We use the Hosmer-Lemeshow goodness-of-fit test and other residual diagnostics to check for the goodness of fit of the model. We assess the balance of the covariates between pure Medicare and HMO groups using the methods described in Raghunathan (2015).

Figure 1 provides histograms of the propensity of obtaining care purely from Medicare for the two groups based on the MCBS data from the survey year 2009. The two groups have considerable overlap of the covariates but also show modest differences in the covariate distributions between them. Thus, the weighting adjustment may be important for removing the bias due to imbalance in the covariates that occurs with the exclusion of the HMO enrollees from the analysis.

We define the selection or HMO weights as the reciprocal of the predicted probabilities of selection obtained from the well-fitting logistic regression model. This approach is similar to making adjustments for unit nonresponse in surveys using a response propensity model. The “final weight” is the product of the existing MCBS weight and the calculated HMO weight. The final analytic data file consists of “pure” Medicare participants and the “final weight” for all subsequent analyses.

4. CALIBRATION FOR CONDITIONS IN NHANES

We now describe calibration/correction for the claims-based dummy variables using the self-report dummy variables available in NHANES for the noninstitutionalized population (step four in section 1). We append the NHANES data set to MCBS and assume, without any loss of generality, that the first n observations are from MCBS and the last m observations are from NHANES. Let X_i, i, = 1, 2, … , n, n + 1, … , $N = n + m$ be all the covariates for the N subjects in the combined data set and to be included in the model. Also, we include medical expenditure and some additional variables (say, U) in X, as they are important variables for future analysis and are available only in MCBS. It is important to include these variables in the imputation model to maintain the relationship between health condition dummy variables and these additional variables.

For subject $i = 1, 2, \dots, n$, let $C_{ik} = 1$ denote the presence of claim for the health condition $k = 1, 2, \dots, K (= 107)$ or set to zero otherwise, where K is the total number of health conditions. Note that the claims-based variables, C, are available only in MCBS. Without any loss of generality, suppose that NHANES self-report is available for the first $r (< K)$ of the K conditions. Let the corresponding self-report dummy variables in NHANES for the first r conditions be S_{ij}, $i = n + 1, n + 2, \dots, n + m$; $j = 1, 2, \dots, r$, where $S_{ij} = 1$ if the subject i self-reported having the health condition j and zero otherwise. Note that these self-report dummy variables are available only for subjects in NHANES. Henceforth, the subscript index $j = 1, 2, \dots, r$ will denote the health conditions for which NHANES self-report dummy variables are available, and the subscript index $k = r + 1, r + 2, \dots, K$ will denote health conditions for which self-report dummy variables are not available in NHANES.

We define a new composite indicator variable for the health condition, $j = 1, 2, \dots, r$, in the combined data set as follows:

$$\begin{gathered} D_{ij} = 1 \text{ if } S_{ij} = 1 \text{ or } C_{ij} = 1\,, \\ D_{ij} = 0 \text{ if } S_{ij} = 0, \end{gathered}$$

and

$$D_{ij} = \,.\, \text{ if } C_{ij} = 0\,.$$

That is, we assume the following: a subject has the health condition ($D = 1$) if the self-report (in the NHANES portion of the appended data set) or the claims (in the MCBS portion of the appended data set) indicate the presence of the disease; a subject does not have the health condition ($D = 0$) if the self-report indicates the subject does not have the health condition; and the actual disease status is missing if there is no claim for the health condition.

For now, setting aside (S_j, C_j, $j = 1, 2, \dots, r$) in the combined data set, the remainder set, (X, D_j, $j = 1, 2, \dots, r$; C_k, $k = r + 1, r + 2, \dots, K = 107$), represents the standard structure of data with missing values. Figure 2 provides a schematic display of combining the two data sets in preparation for multiply imputing the missing values.

The variables to be imputed in the combined data sets are U and $C = \{C_k, k = r + 1, r + 2, \dots, K\}$ for the NHANES subjects and $D = \{D_j, j = 1, 2, \dots r\}$ for the MCBS subjects. Recall that U is a portion of X which includes expenditure and other important variables available only in MCBS. When imputing the missing values in D (only in MCBS), we impose a constraint whereby the multiply imputed prevalence rates are calibrated to be equal to the observed prevalence rates in NHANES after adjusting for any differences in the covariate distribution between the two surveys. If the prevalence rate estimate using C in MCBS is greater than or equal to the prevalence rate estimate using S in NHANES, then no correction is made to D and all the missing values are set to zero. On the other hand, if the prevalence rate estimate using C is smaller than the one based on S, then the missing values in D are imputed such that prevalence rate estimate based on imputed D equal, in expectation, the prevalence rate estimate based on S. That is, this imputation procedure is designed to correct underestimation in the claims-based dummy variables and to match them, in expectation, with the NHANES rates.

We modify the sequential regression multivariate imputation methodology as follows:

1. Develop appropriate regression models for variables U. For example, we use a log-normal linear regression model to impute the missing values in the expenditure variable with rest of X and (D, C) as covariates. The logtransformation achieves approximate normality of the residuals in the regression model.

2. Impute the missing values in C_k, $k = r + 1, r + 2, \dots, K$ using a logistic regression model with $(X, D, C_{(-k)})$ as predictors, where $C_{(-\mathrm{k})}$ is all the claims-based dummy variables $(C_{r+1}, C_{r+2}, \dots, C_{k-1}, C_{k+1}, \dots, C_K)$.

3. For imputing the missing values in D_j, $j = 1, 2, \dots, r$, we use the following calibration procedure:

   a. Define a variable $M = 1$ for the subjects in MCBS and $M = 0$ for the subjects in NHANES. Let $D_{(-\mathrm{j})}$ denote the collection of health condition dummy variables for all conditions except j. Construct a propensity score based on fitting a logistic regression model predicting M with $(X, D_{(-\mathrm{j})}, C)$ as covariates and create strata based on the propensity scores. This step groups the subjects in the two surveys based on the similarity of the covariates and other health conditions (this is similar to creating hot-deck adjustment cells). For most health conditions, the number of strata is fixed at five—or four for very low prevalence conditions.

   b. For a particular propensity score class, let p_{jS} be the estimated prevalence rate (weighted) based on the self-report, S_j, and p_{jC} be corresponding estimate based on the claims, C_j. If $p_{jC} \geq p_{jS}$, then set all missing D_j to 0 because there is no underestimation in the concerned propensity score class.

   c. Suppose $p_{jC} < p_{jS}$ in a propensity score class. Let n_{1jD} be the number of subjects in MCBS with $D_j = 1$ in the class and n_{ojD} be number of subjects in MCBS with $D_j = .$ (missing). Let w_{1jD} and w_{ojD} be the sum of the weights of n_{1jD} and n_{ojD} subjects, respectively. The goal is to impute missing D_j such that after imputation $p_{jD} = p_{jS}$, in expectation, where p_{jD} is the prevalence estimated based on imputed D_j. That is, if θ is the imputation rate or probability of setting a missing D_j to be equal to one, then the calibration condition is $(w_{1jD} + \theta w_{ojD})/(w_{1jD} + W_{ojD}) = p_{jS}$ or $\theta = p_{jS} - w_{1jD}(1 - p_{jS})/w_{ojD}$.

Missing D_j are set to 1 or 0 by drawing n_{ojD} independent Bernoulli random variables with probability θ. This approach may be considered as a combination of the “hot-deck” approach (using the propensity score covariate matching of the respondents in the two surveys to create adjustment cells) and a binomial model for the missing values within each adjustment cell. The parameter in the binomial model is constrained to make the multiply imputed rate and the observed NHANES rate equal, in expectation.

We iterate these steps across all diseases several times until the multiply imputed prevalence rates stabilize.

5. CALIBRATION FOR CONDITIONS NOT AVAILABLE IN NHANES

The next step (step five in section 1) is the calibration of the claims-based dummy variables for the health conditions that are not available in NHANES. That is, derive D_k, $k = r + 1, r + 2, \dots, K$ using the measurement error model framework. The relationship between C_j and the corresponding imputed D_j (obtained in the previous section) where $j = 1, 2, \dots, r$ may be viewed as the measurement error (or “correction”) model by treating C_j as a mismeasured dummy variable and D_j as the variable a.” That is, the r measurement error models are given by estimates of $Pr(D_j | C_j, D_{(-j)}, X)$, $j = 1, 2, \dots, r$.

The following steps describe the imputation procedure for deriving D_k, $k = r + 1, r + 2, \dots, K$ using the measurement error models:

1. The first step is to build r measurement error models using logistic regression models,

$$\text{logit} Pr(D_j = 1 \mid C_j, D_{(-j)}, X) = Z_j^T \beta_j,$$

   where Z_j is a vector of predictors based on $(C_j, D_{(-j)}, X)$ and includes some interaction terms, as well. These are the r potential “donor” models for imputing the missing D_k, $k = r + 1, r + 2, \dots, K$.

2. Next, for each condition, $k = r + 1, r + 2, \dots, K$, find health conditions from $j = 1, 2, \dots, r$ with similar prevalence rates, conditional on the covariates, X. Specifically, let $\pi_k(X) = Pr(C_k = 1 | X)$ be the prevalence rate function for the claims-based condition C_k as a function of the covariates. We consider a collection of all $j = 1, 2, \dots, r$ with similar values of $\pi_j(X) = Pr(C_j = 1 | X)$ as a set of donor candidates of the measurement error models developed in step one. To accomplish this task, we create groups by stratifying the subjects based on their predicted values using $\pi_k(X)$. For each group, all the conditions with the values of $\pi_j(X)$, $j = 1, 2, \dots, r$ in the same group constitutes a donor pool. The heuristic reasoning is that if the estimated probability of having health conditions k and j, given X, match for individuals, then the imputation/measurement error models for j can be used for k.

   The rationale that a match on the disease etiology based on X implies one can borrow strength from the corresponding measurement error process involves an unverifiable assumption. However, comparisons of the regression coefficients in the measurement error model described in step one show that if $|\pi_j(X) - \pi_{j'}(X)|$ is small then $|\hat{\beta}_j - \hat{\beta}_{j'}|$ is also small, with a strong positive correlation for $j, j' =$

$1, 2, \ldots, r$, $j \neq j'$. That is, this assumption seems to be reasonable for the conditions in $\{1, 2, \ldots, r\}$.

3. Let $\bar{\hat{\beta}}$ be the average value of the measurement error model coefficients (derived in step one) for the donor candidates identified in step two. Imputation model for condition $k = r + 1, r + 2, \ldots, K$ is the logit model,

$$\text{logit}\ Pr(D_k = 1 \mid C_k = 0, D_1, D_2, \ldots, D_r, X) = Z_k^T \bar{\hat{\beta}}\,.$$

Note that $D_k = 1$, if $C_k = 1$ and imputation is needed only if $C_k = 0$. The heuristic reasoning for the imputation model is the hot-deck analogy where the adjustment cells are formed based on the similarity of $\pi_k(X)$ and $\pi_j(X)$, and the imputation model for the non-NHANES condition, k, is the average of the measurement error models for the NHANES conditions.

6. INSTITUTIONALIZED POPULATION

The final step is the calibration of the health conditions for the institutionalized population (step six in section 1). The number of institutionalized subjects is relatively small (about 10 percent) but an important segment from the scientific perspective. The NHANES does not include any institutionalized subjects. Thus inherently, calibration for the institutionalized sample is an extrapolation. Furthermore, the covariate distributions (including cost) might be different for institutionalized and noninstitutionalized subjects and, therefore, should be accounted for in developing the calibration. However, the analysis shows considerable overlap in the distribution of covariates (tables 1 and 2) between these two groups.

Using the same hot-deck analogy, the institutionalized and noninstitutionalized subjects are matched on common set of covariates to create adjustment cells; we use the calibration/correction for the claims-based dummy variables for all 107 health conditions just completed for the noninstitutionalized population (as described in sections 4 and 5) as the donor pool to calibrate for the institutionalized. The following procedure carries out calibration for the institutionalized population.

1. Match the subjects in the institutionalized and noninstitutionalized (community dwelling) populations for each health condition. Note that by definition, $D = 1$ when $C = 1$. Hence, this matching needs to performed only for the subset with $C = 0$. Let $I = 1$ for institutionalized subjects and $I = 0$ for noninstitutionalized subjects. For health condition, $k = 1, 2, \ldots, K$, define the covariates as $(X, C_{(-k)}, E)$ and fit a propensity score model with I as the dependent variable. Let $v_k =$ logit $Pr(I = 1 | X, C_{(-k)}, E, C_k = 0)$ be the estimated logit (or the linear predictor) of the probability of being institutionalized.

2. Conditional on the match score, v_k, assume that

$$Pr(D_k = 1 \mid v_k, I = 1) = Pr(D_k = 1 \mid v_k, I = 0)\,.$$

This is akin to missing at random assumption in the missing data analysis, where, conditional on the propensity score v_k, the probability of misclassification (that is, $D = 1$, given $C = 0$) is the same for institutionalized and noninstitutionalized samples. Paucity of self-report data from the institutionalized populations is a severe limitation, and therefore, this assumption of being able to predict using the model based on the matched noninstitutionalized subjects to impute for the institutionalized subjects is inevitable. However, the considerable overlap in the distributions of a rich set of covariates between the two populations, despite differences, may make this assumption more palatable.

3. Consider now the estimation of the predictive distribution for the imputation based on the assumption in step two. Assume that $v_k \mid D_k = l \sim N(a_l, b_l^2)$, $l = 0, 1$, where a_l and b_l are computed from the propensity scores defined in step one. Bayes theorem yields

$$q_k = Pr(D_k = 1 \mid v_k, I = 0, C_k = 0) = \frac{Pr(v_k \mid I = 0, D_k = 1, C_k = 0)Pr(D_k = 1 \mid I = 0, C_k = 0)}{\sum_{l=0}^{1} Pr(v_k \mid I = 0, D_k = l, C_k = 0)Pr(D_k = l \mid I = 0, C_k = 0)},$$

   where $Pr(v_k|I = 0, D_k = l, C_k = 0)$ is the normal density evaluated at v_k with mean a_l and variance b_l^2.

4. For each subject in the institutionalized sample and with $C_k = 0$, compute their value of q_k and draw an uniform random number. If it is less than q_k, then set $D_k = 1$ and zero otherwise.

7. EVALUATION OF THE PROCEDURE

We use a series of complex steps to create model-based health condition dummy variables by borrowing strength from the nationally representative sample. This section describes evaluation of the procedure, the effect of calibration, and the comparison of resulting estimates with other internal and external sources.

7.1 Model Diagnostics

Derivation of the calibrated health condition dummy variables relies on various models for matching subjects between the two surveys, between two subpopulations (institutionalized versus noninstitutionalized) and a number of imputation and measurement error models. We routinize the model diagnostics for every model used in the matching, imputation, and measurement error process as an integral part of the macros. For the propensity score models, we assess the overall goodness of fit using the Hosmer-Lemeshow test but also check for the balance of covariates using the methods described in Raghunathan (2015). We modify the models by adding interaction and nonlinear terms to improve the fit, if necessary. The modeling tasks are tailored for each health condition and the year of the survey and thus involve several hundred models.

Imputation diagnostics for each imputed variable compare the distribution of the observed and imputed values as described in Bondarenko and Raghunathan (2016). The checking of measurement error models involves cross-validation by setting a portion of actual values aside and then checking the draws from the predictive distribution against the actual values. We provide the details about the procedures and steps taken to check the model assumptions in a technical report available on the website at NBER (https://www.nber.org/aging/nha/techandresults_spending.html, last accessed November 1, 2019).

The measurement error models involve matching the claims-based predictive density for the health conditions available in NHANES with the conditions not available in NHANES. We use different choices for the cut points to create groups (see step three in section 5) to explore the sensitivity of this choice on prevalence estimates. We find the estimates to be quite similar for five to ten groups and subsets of health conditions chosen among the matching health conditions. All these diagnostics indicate that the models used in the imputation or predictions are well fitting for all 1,498 dummy variables (107 health conditions for each of the fourteen years).

7.2 External Validation

As indicated earlier, MCBS collects self-report data on several conditions that were intentionally set aside in all the analyses. We reserved them for the evaluation of the calibration process. The first analysis, therefore, is to use these self-report prevalence rate estimates and compare them to claims-based and model- or calibration-based prevalence estimates. Let p_M denote the design-weighted prevalence rate estimate based on the self-report question in MCBS. As described earlier, we define $\delta^* = (p_D - p_M) / \sqrt{p_M(1 - p_M)}$ as the effect size, where p_D is the design-weighted estimate of the prevalence rate using the calibrated claim dummy variables (i.e., using D). This discrepancy measure could be computed for the 352 (out of 1,498) health conditions. Only twentysix of these 352 health conditions have negative values of δ^*, with the average of −0.02 (for $\delta^* < 0$) and a maximum of −0.05. In contrast, $\delta = (p_C - p_M) / \sqrt{p_M(1 - p_M)}$, where p_C is the design-weighted claims-based prevalence estimate, are considerably larger, as expected.

The calibration is useful if $|p_D - p_M| < |p_C - p_M|$. Of the 352 health conditions, the strict inequality is satisfied for 254 estimates (72 percent). That is, the calibration produces estimates closer to the self-report data in MCBS (which were not included in the calibration process) than the claims-based estimates.

We evaluate trend estimates over the fourteen-year period using p_M and corresponding p_D (i.e., for the conditions which could be externally validated). We use a regression model with time (with reference year 1999 as zero and 2012 as fourteen) as a continuous predictor and the prevalence rate estimate as the outcome variable to estimate the trend. The goal of this analysis is to assess whether the slopes using the calibrated claims–based prevalence rate estimates are similar to the slope estimates from self-report prevalence rate estimates and then compare and contrast them to the slope estimates using the claims-based prevalence rate estimates.

Figure 3 compares the slope estimates using the calibrated claims-based estimates to slopes estimated using MCBS self-reports. The scatter around the 45-degree line suggests that the slope estimates from the calibrated claims–based prevalence rate estimates are generally similar to the ones obtained from self-report prevalence rate estimates. In general, the trend is not strong in any of these conditions. Note that all trend estimates are subject to regression model residual or error (deviation from the model with time as a linear predictor), sampling error, and imputation uncertainty. Thus, one would expect some scatter around the 45-degree line.

To further investigate the usefulness of the calibration process, we define a dummy variable $A = 1$ if $|S_D - S_M| < |S_C - S_M|$ (success in the sense that the calibrated-claim slope estimate is closer to the self-report slope estimate than the claims-based slope estimate) and zero otherwise, where S_D, S_C, and S_M are the slope estimates using the calibrated claim, claim, and MCBS self-report prevalence rate estimates, respectively. Seventy-three percent of the health conditions satisfy the strict inequality. Thus, the analysis of the point estimates and trend estimates demonstrates that the calibration process is achieving the goal of correcting for the underestimation bias in the claims-based dummy variables and results in estimates that mimic the properties of prevalence rate estimates based on the self-reports.

Calibrated estimates also compare favorably with rates in other published literature (using National Health Interview Survey, Behavior Risk Factor Surveillance System, Health and Retirement Study and National Comorbidity Survey), and rates provided by disease-specific interest groups and associated websites and clinical experts. Obviously, there is no known true prevalence rate estimates to compare against except for a few diseases. The modeling task is, to some extent, to extrapolate all 107 conditions.

7.3 Internal Validation

The purpose for constructing model-based dummy variables for each of the 107 health conditions across the fourteen years is to treat them as a statistical structure where these dummy variables could be used in a variety of analyses. As an example, Cutler, Ghosh, Messer, Raghunathan, Stewart et al. (2019) use these dummy variables to explain the slowdown of Medicare spending and to attribute health conditions potentially contributing to this slowdown. This analysis collapsed the 107 health conditions into thirty-two categories.

The three panels (Panel (c) is on the next page) in figure 4 present the effect of calibration on the prevalence estimates in the years 2000, 2005, and 2012, for the thirty-two major health condition categories used in Cutler et al. (2019). The darker color bar is the claims-based estimates, and the lighter color bar is the increase in prevalence rate due to calibration. The calibration makes large corrections for cardiovascular conditions, arthritis, and a number of other conditions, modest for screening and small for cancers.

The two panels in figure 5 provide a scatter plot of calibrated claims–based and claims–based prevalence rate estimates, along with a 45-degree line for the noninstitutionalized and institutionalized populations, respectively, for all 1,498 conditions. The prevalence rate estimates on the 45-degree line are not calibrated because p_C was larger than p_S. In general, the scatter plots show that impact of calibration depends on the value of p_C. We fit a regression line, $p_D = \alpha_o + \alpha_1 (p_C - 0.5)$, to assess the extent to which the calibration moves the prevalence rates away from a 45-degree line. The least squares estimates are $\hat{\alpha}_o = 0.03$ and $\hat{\alpha}_1 = 1.04$ for the noninstitutionalized population. That is, on average, the calibrated claims–based prevalence estimates are about 3 percentage points higher than the claims-based prevalence rate estimates. For the institutionalized population, the estimated intercept and slopes are 0.03 and 1.01, respectively. Thus, the corrections made (movement away from a 45-degree line) to the claims-based dummy variables are more for the noninstitutionalized population than for the institutionalized population.

All these analyses indicate that the calibration process produces meaningful changes to the prevalence dummy variables.

8. DISCUSSION AND LIMITATIONS

An ideal data set would have been generated by asking self-report questions on all 107 health conditions or, better yet, clinically or biologically assessed in a nationally representative survey for all fourteen years. For the elderly population, the primary data source, MCBS, provides information on all 107 health conditions using data on Medicare claims but has potential problems. However, nationally representative survey data sources do not collect information on all 107 health conditions. Both data sources are in a sense imperfect. We developed model-based dummy variables by combining information from both these sources of data. The modeling approach used the measurement error and missing data framework to create a “modeled” ideal data set which then can be used for a range of subsequent analyses.

Though the focus of this article in on the elderly, the same approach is being applied for other age groups (younger than 18 years of age, 18 to 44, and 45 to 64 years of age) over the same time period but using different data sources. The data sources for these age groups are scarce. The Medical Expenditure Panel Survey (MEPS), National Comorbidity Survey (NCS), and several other national surveys are being assembled for these age groups. Most—if not all—of these analyses have to be performed in the Federal Statistical Research Data Center (FSRDC).

The self-reports in surveys may also be subject to bias as discussed in Schenker et al. (2010), where clinical measures from NHANES were used to correct the self-reports in the National Health Interview Survey. The current project of this article incorporates physical/laboratory measures from NHANES (the same method described in the aforementioned reference) to estimate the prevalence rates of diabetes, hypertension, and hyperlipidemia. It is possible that further refinements can be made if clinical measures for some or all other measures become available through targeted clinical studies or surveys.

The claims-based prevalence estimates, though assumed to suffer only from “false negatives” (that is having no claims for a given health condition were considered as unknown health condition status), may also suffer from “false positives.” For example, a screening or diagnostic visit with a negative outcome may have been coded with the corresponding disease ICD-9 codes. For modeling such “false positives” some additional data will be needed.

The accuracy of calibration depends on the imputation model assumptions, quality of available covariates, and caliber of match between the sample of interest and the external source. Given the reliance on numerous model assumptions, further exploration of sensitivity of the estimates to the underlying assumptions is needed. Several analysts cross-checked the model diagnostics and different models, but a more systematic treatment may be needed.

This study was based on two fairly rich data sources. There are several other possible sources, such as the National Health Interview Survey, Behavior Risk Factor Surveillance System, Medical Expenditure Panel Survey, National Comorbidity Survey, Health and Retirement Study, etc. These sources are used in a limited fashion (e.g., for assessing the quality of calibration and model checking through comparisons of the calibrated estimates to the self-reports from these surveys). A more thorough approach will be to combine all these data sources and incorporate them in the calibration/correction process. The goal of such an endeavor (by no means an easy task) would be to reduce the number conditions without self-reports and increase the precision of the imputation process.

The calibration can be made more statistically efficient through temporal modeling of the prevalence rates in the calibration process rather than stratified by year. This would complicate the computational task considerably and might also over-smooth the trends in the prevalence rate estimates. Further investigation is necessary to explore the possibility of borrowing strength across time and additional surveys.

If the prevalence rate estimates and trends are the only inferential quantities of interest (instead of the creation of dummy variables as a part of the statistical infrastructure), a variety of other methods can be used as discussed in Lohr and Raghunathan (2017) and Dong, Elliott, and Raghunathan (2014). One option is using a fully Bayesian model for all prevalence rate estimates, self-reports, and claims across all years. The bias correction model similar to the one used in Raghunathan, Xie, Schenker, Parsons, Davis, et al. (2007) could be developed assuming that the self-reports are unbiased estimates of the population prevalence rates and the claims-based rates are biased where the bias term is explicitly modeled. This approach might be suitable for handling both underand overestimation while estimating the model based prevalence rates. Investigation along this line is currently in progress to correct for both underand overestimation in the claims-based prevalence rate estimates.

Acknowledgments

Authors thank the two referees and associate editor for many insightful comments. The research was funded by National Institute of Health grants P01-AG031098 and R37-AG047312.

References

Bondarenko I, and Raghunathan TE (2016), “Graphical and Numerical Diagnostic Tools to Assess Suitability of Multiple Imputations and Imputation Models,” Statistics in Medicine, 35, 3007–3020. [PubMed: 26952693]

Cutler D, Ghosh K, Messer K, Raghunathan T, Stewart S, and Rosen A (2019), "Explaining the Slowdown in Medical Spending Growth among the Elderly, 1999–2012," Health Affairs, 38, 222–229. [PubMed: 30715965]

Centers for Disease Control and Prevention (CDC) (1999–2012), "National Center for Health Statistics (NCHS)." National Health and Nutrition Examination Survey Data. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention. Available at www.cdc.gov/nchs/nhanes (accessed November 1, 2019).

Center for Medicare and Medicare Services (CMS) (1999–2012), "Medicare Current Beneficiary Survey." US Department of Health and Human Services, Center for Medicare and Medicaid Services. Available at www.cms.gov/MCBS (accessed November 1, 2019).

Cohen J (1988), Statistical Power Analysis for the Behavioral Sciences, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dong Q, Elliott M, and Raghunathan T (2014), "Combining Information from Multiple Complex Surveys," Survey Methodology, 40, 347–354. [PubMed: 29200609]

Kennickell AB (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," Proceedings of the Survey Research Methods Section of the American Statistical Association, 1,40.

Lohr S, and Raghunathan T (2017), "Combining Survey Data with Other Data Sources," Statistical Science, 32, 293–312.

Muggah E, Graves E, Bennett C, and Manuel DG (2013), "Ascertainment of Chronic Diseases Using Population Health Data: A Comparison of Health Administrative Data and Patient Self-Report," BMC Public Health, 13, 16. [PubMed: 23302258]

NBER. "NBER Program Project on Satellite National Health Accounts." Available at https://www.nber.org/aging/nha/techandresults_spending.html (accessed November 1, 2019).

O'Donnell S, Vanderloo S, McRae L, Onysko J, Patten SB, and Pelletier L (2016), "Comparison of the Estimated Prevalence of Mood and/or Anxiety Disorders in Canada between Self-Report and Administrative Data," Epidemiology and Psychiatric Sciences, 25, 360–369. [PubMed: 26081585]

Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, and Rodeheffer RJ (2004), "Agreement between Self-Report Questionnaires and Medical Record Data Was Substantial for Diabetes, Hypertension, Myocardial Infarction and Stroke but Not for Heart Failure," J Clin Epidemiol, 57, 1096–1103. [PubMed: 15528061]

Raghunathan TE, Lepkowski JM, van Hoewyk JV, and Solenberger PW (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," Survey Methodology, 27, 85–95.

Raghunathan TE, Solenberger PW, and Van Hoewyk JH (2002), "IVEware: Imputation and Variance Estimation Software." Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan. Available at www.iveware.org.

Raghunathan TE (2006), "Combining Information from Multiple Surveys for Assessing Health Disparities," Allgemeines Statistisches Archiv, 90, 515–526.

Raghunathan TE, Xie D, Schenker N, Parsons VL, Davis WW, Dodd KW, and Feuer EJ (2007), "Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening," Journal of American Statistical Association, 102, 474–486.

Raghunathan TE (2015), Missing Data Analysis in Practice, Boca Raton, Florida: CRC Press.

Robinson JR, Young TK, Roos LL, and Gelskey DE (1997), "Estimating the Burden of Disease. Comparing Administrative Data and Self-Reports," Medical Care, 35, 932–947. [PubMed: 9298082]

Rubin DB (1987), Multiple Imputation for Nonresponse in Surveys, New York: Wiley.

Schenker N, and Raghunathan TE (2007), "Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health," Statistics in Medicine, 26, 1802–1811. [PubMed: 17278184]

Schenker N, Raghunathan TE, and Bondarenko I (2010), "Improving on Analyses of Self-Reported Data in a Large-Scale Health Survey by Using Information from an Examination-Based Survey," Statistics in Medicine, 29, 533–545. [PubMed: 20029804]

van Buuren S, and Oudshoorn K (1999), "Flexible Multivariate Imputation by MICE." Leiden: TNO Preventie En Gezondheid Technical Report TNO/VGZ/PG 99.054.

Yasaitis LC, Berkman LF, and Chandra A (2015), “Comparison of Self-Reported and Medicare Claims-Identified Acute Myocardial Infarction,” Circulation, 131, 1477–1485. [PubMed: 25747935]

Zuvekas S, and Olin G (2009), “Validating Household Reports of Health Care Use in the Medical Expenditure Panel Survey,” Health Services Research, 44, 1679–1700. [PubMed: 19619249]

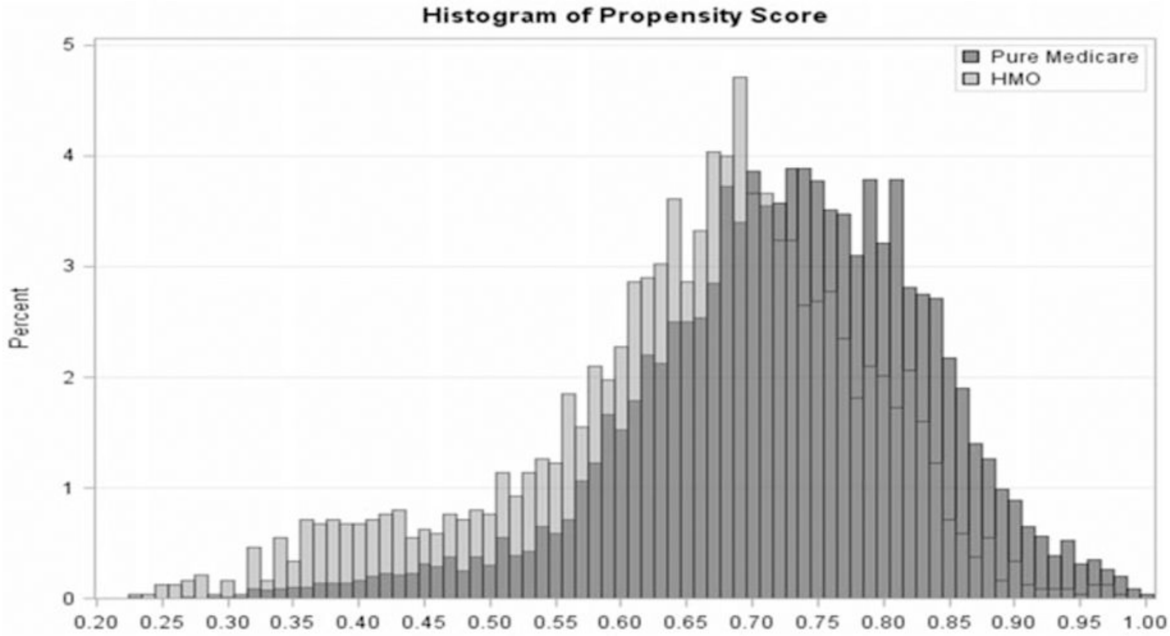


Figure 1.

Histograms of the Propensity of Obtaining Care Purely from Medicare (Not Enrolled in HMO) for Pure Medicare and HMO Respondents for the Survey Year 2009.

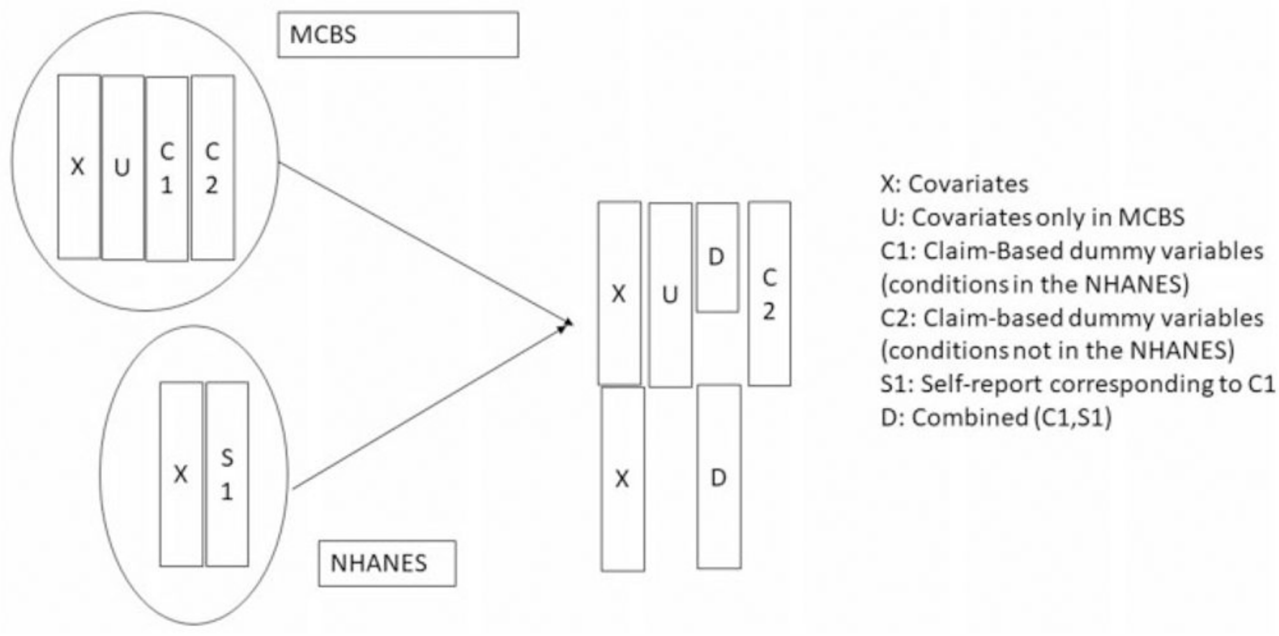

MCBS
X
U
C 1
C 2
X
S 1
NHANES
X
U
D
C 2
X
D
X: Covariates
U: Covariates only in MCBS
C1: Claim-Based dummy variables (conditions in the NHANES)
C2: Claim-based dummy variables (conditions not in the NHANES)
S1: Self-report corresponding to C1
D: Combined (C1,S1)

Figure 2.

Schematic Display of Combining the Two Data Sets to Multiply Impute the Derived Health Conditions for the Subjects Missing in in MCBS.

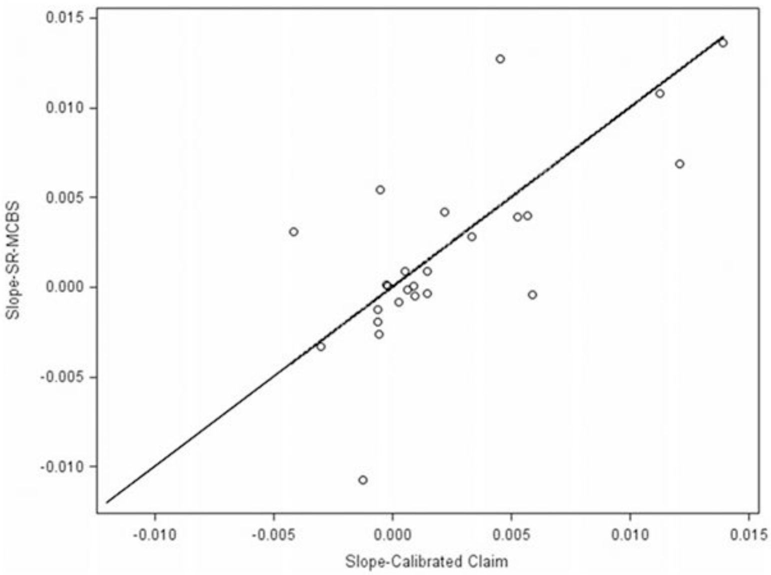
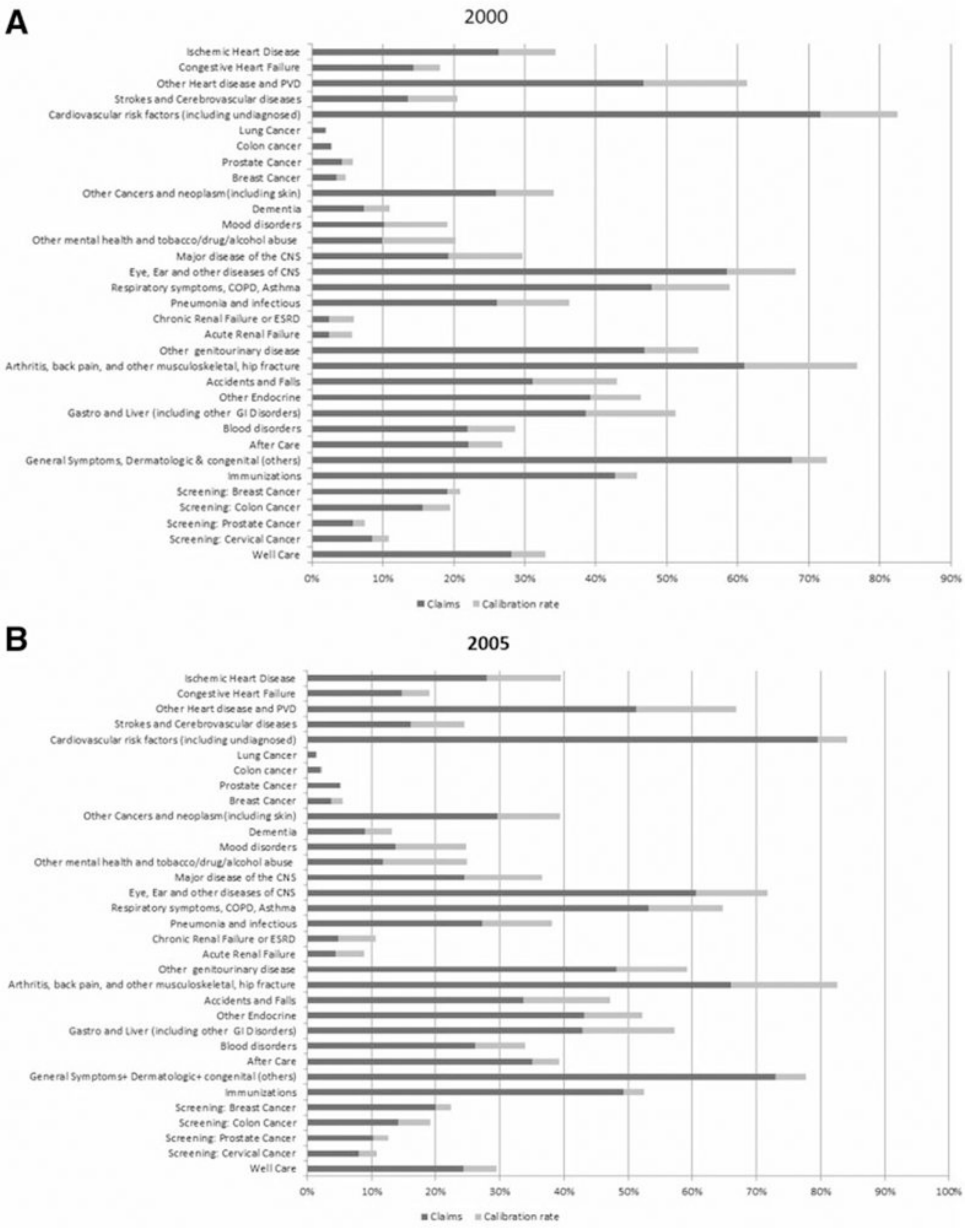


Figure 3.

Trend Estimates from MCBS Self-Reports Compared with Calibrated Claims.


A
2000
Ischemic Heart Disease
Congestive Heart Failure
Other Heart disease and PVD
Strokes and Cerebrovascular diseases
Cardiovascular risk factors (including undiagnosed)
Lung Cancer
Colon cancer
Prostate Cancer
Breast Cancer
Other Cancers and neoplasm (including skin)
Dementia
Mood disorders
Other mental health and tobacco/drug/alcohol abuse
Major disease of the CNS
Eye, Ear and other diseases of CNS
Respiratory symptoms, COPD, Asthma
Pneumonia and infectious
Chronic Renal Failure or ESRD
Acute Renal Failure
Other genitourinary disease
Arthritis, back pain, and other musculoskeletal, hip fracture
Accidents and Falls
Other Endocrine
Gastro and Liver (including other GI Disorders)
Blood disorders
After Care
General Symptoms, Dermatologic & congenital (others)
Immunizations
Screening: Breast Cancer
Screening: Colon Cancer
Screening: Prostate Cancer
Screening: Cervical Cancer
Well Care
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
■ Claims ■ Calibration rate
B
2005
Ischemic Heart Disease
Congestive Heart Failure
Other Heart disease and PVD
Strokes and Cerebrovascular diseases
Cardiovascular risk factors (including undiagnosed)
Lung Cancer
Colon cancer
Prostate Cancer
Breast Cancer
Other Cancers and neoplasm (including skin)
Dementia
Mood disorders
Other mental health and tobacco/drug/alcohol abuse
Major disease of the CNS
Eye, Ear and other diseases of CNS
Respiratory symptoms, COPD, Asthma
Pneumonia and infectious
Chronic Renal Failure or ESRD
Acute Renal Failure
Other genitourinary disease
Arthritis, back pain, and other musculoskeletal, hip fracture
Accidents and Falls
Other Endocrine
Gastro and Liver (including other GI Disorders)
Blood disorders
After Care
General Symptoms+ Dermatologic+ congenital (others)
Immunizations
Screening: Breast Cancer
Screening: Colon Cancer
Screening: Prostate Cancer
Screening: Cervical Cancer
Well Care
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
■ Claims ■ Calibration rate

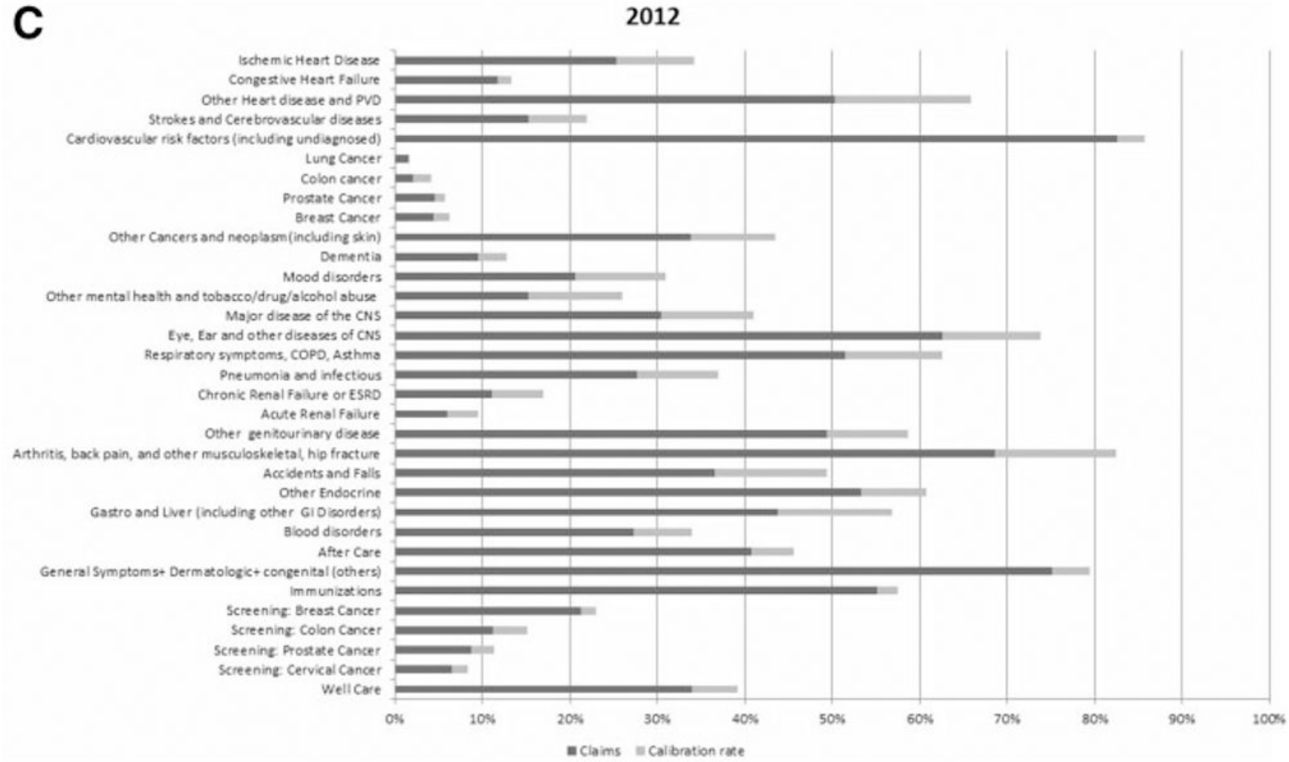


Figure 4.

Bar Graph Comparing the Calibrated Claim and Claims-Based Prevalence Rate Estimates of the 32 Health Condition Categories for Years 2000, 2005, and 2012. (a) Year 2000. (b) Year 2005. (c) Year 2012.

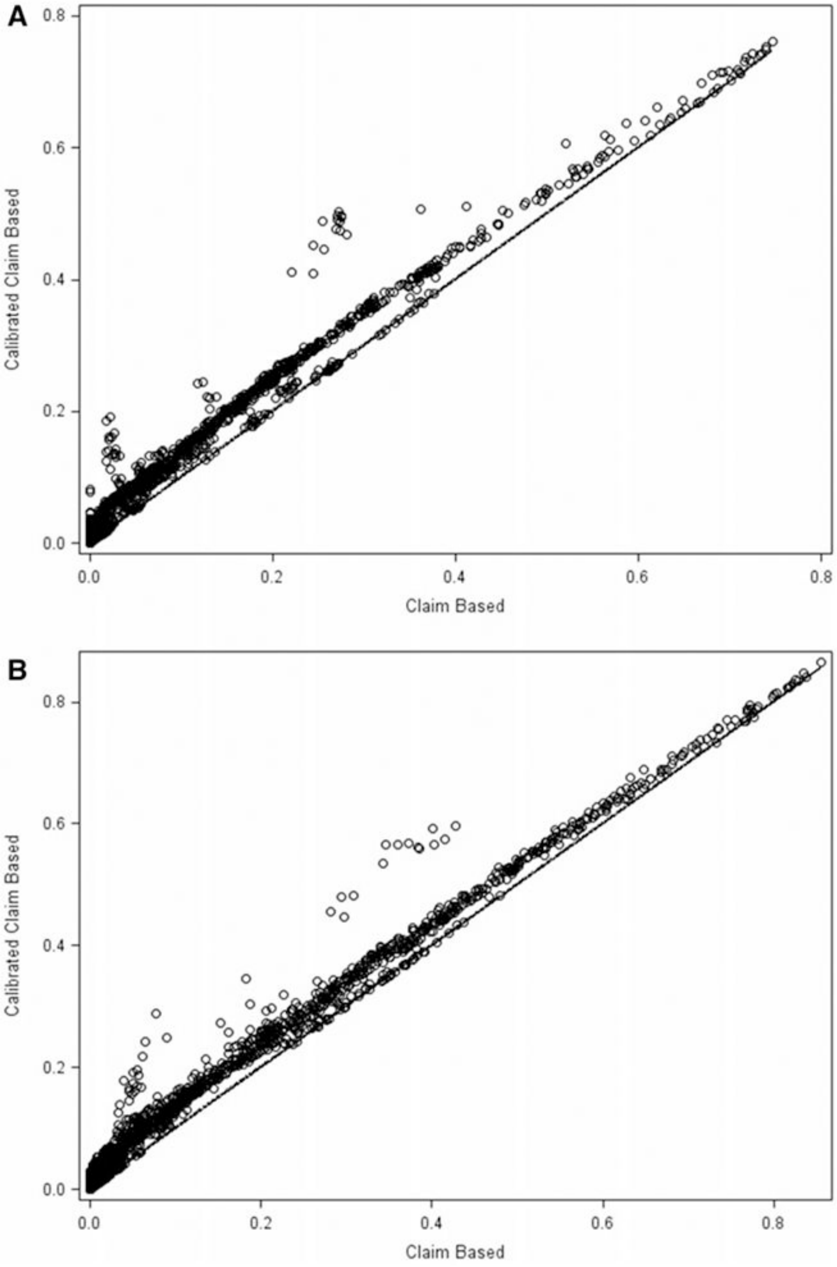


Figure 5.

Scatter Plot Comparing the Calibrated Claim and Claims-Based Prevalence Rate Estimates of the 107 Health Conditions Over a Fourteen-Year Period (1999–2012). (a) Noninstitutionalized Population. (b) Institutionalized Population.

Table 1.

Descriptive Statistics of Demographic Variables for 2009–2010 NHANES and 2009 MCBS Survey Respondents

Variable label and coded values	**NHANES (2009–10) % (SE)**	**MCBS (2009) % (SE)**
Age (Mean/SE)	73.17 (0.22)	75.61 (0.11)
Race/ethnicity		
White	80.27 (2.52)	80.00 (1.06)
Black	8.31 (1.18)	8.16 (0.87)
Hispanic	6.99 (2.01)	7.41 (0.80)
Other	4.43 (1.04)	4.44 (0.43)
Male	44.49 (1.13)	43.44 (0.92)
Education		
Less than 9th grade	10.54 (1.29)	10.38 (0.47)
9-11th grade	15.26 (1.78)	13.32 (0.55)
High school	24.60 (1.63)	30.18 (0.72)
Some college or AA degree	26.87 (1.73)	26.18 (0.80)
College grad or more	22.73 (2.09)	19.93 (0.89)
Marital status		
Married	62.18 (1.96)	53.24 (0.81)
Widowed	26.42 (1.55)	32.37 (0.79)
Divorced or separated	8.56 (0.78)	11.00 (0.52)
Never married	2.84 (0.38)	3.39 (0.24)
Poverty category		
Poor/negative	9.86 (1.20)	12.85 (0.63)
Near poor	8.06 (0.73)	6.76 (0.42)
Low income	17.89 (1.57)	17.78 (0.64)
Middle income	33.14 (2.13)	34.12 (0.76)
High income	31.06 (2.22)	28.49 (0.93)
Ever served in the armed forces	26.64 (1.17)	25.10 (0.79)
Description of home		
One-family house detached		73.07 (1.06)
One-family house attached to other house(s)		5.46 (0.43)
Apartment		13.65 (0.73)
Mobile home or trailer		7.42 (0.57)
Other		0.41 (0.10)
Total number of people in household (Mean/SE)	2.11 (0.05)	1.95 (0.02)
Number of rooms in home (Mean/SE)	6.30 (0.09)	5.90 (0.04)

Table 2.

Descriptive Statistics of Common Variables for 2009–2010 NHANES and 2009 MCBS Survey Respondents

Variable label and coded values	NHANES (2009–10) % (SE)	MCBS (2009) % (SE)
Private health insurance coverage	63.14 (1.93)	72.81 (0.85)
Have a routine place to go for health care	97.60 (0.60)	96.16 (0.32)
General health condition		
Excellent	12.46 (1.09)	16.40 (0.76)
Very good	26.64 (1.42)	29.84 (0.63)
Good	35.15 (1.46)	32.36 (0.58)
Fair	20.27 (1.55)	16.19 (0.62)
Poor	5.48 (0.63)	5.21 (0.27)
Health compared to one year ago		
Better	15.12 (0.97)	14.43 (0.65)
Same	70.38 (1.38)	63.68 (0.83)
Worse	14.51 (1.02)	21.89 (0.60)
Height in centimeters (Mean/SE)	167.7 (0.29)	167.4 (0.17)
Weight in kilograms (Mean/SE)	78.64 (0.50)	76.71 (0.26)
Inpatient stays (Mean/SE)	0.33 (0.03)	0.31 (0.01)
Difficulty walking 1/4 mile or 2–3 blocks	38.73 (1.45)	37.01 (0.71)
Difficulty lifting/carrying 10 pounds		
No/little difficulty	77.09 (1.69)	72.70 (0.72)
Some difficulty	10.82 (0.72)	8.75 (0.42)
Much difficulty	4.59 (0.89)	6.82 (0.34)
Unable to do	7.50 (0.98)	11.73 (0.42)
Difficulty stooping/crouching/kneeling		
No/little difficulty	49.80 (1.75)	48.42 (0.81)
Some difficulty	29.43 (1.40)	19.51 (0.60)
Much difficulty	11.59 (1.04)	16.88 (0.53)
Unable to do	9.18 (0.98)	15.19 (0.55)
Difficulty eating	5.21 (0.58)	4.37 (0.26)
Difficulty dressing	10.49 (1.20)	10.07 (0.42)
Ever smoked cigarettes/cigars/tobacco	49.52 (1.18)	56.39 (0.87)
Currently smoke cigarettes/cigars/tobacco	7.83 (0.71)	9.07 (0.53)
Had a hysterectomy	46.01 (1.98)	34.77 (1.00)

Table 3.

107 Health Conditions for Which Model Based Calibrated Claim Dummy Variables were Constructed Using Multiple Imputations

Tuberculosis	STD, non-HIV
HIV	Immunizations and screening for infection disease
Other Infectious Disease	Colon cancer
Lung Cancer	Skin Cancer
Breast Cancer	Cervical Cancer
Prostate Cancer	Hematologic Cancers
Benign Neoplasm	Other cancer
Thyroid Disorders	Diabetes Mellitus
Undiagnosed Diabetes Mellitus	Hyperlipidemia
Undiagnosed Hyperlipidemia	Gout and other crystal arthropathies
Other Endocrine Diseases	Anemias
Other Hematologic Disease	ETOH Abuse
Illicit Drug Use	Tobacco Use
Dementia	Depression
Bipolar Disorder	Schizophrenia
Anxiety	Posttraumatic Stress Disorder (PTSD)
Attention Deficit Hyperactivity Disorder (ADD-ADHD)	Mental Retardation (HCC term)
Other Mental Health Disorders	Otitis Media
Parkinson Disease Other Mental Health Disorders	Multiple Sclerosis Otitis Media
Paralysis Parkinson Disease	Seizure Disorders Multiple Sclerosis
Headaches Paralysis	Migraine Seizure Disorders
Cataract Headaches	Glaucoma Migraine
Eye Disorders Cataract	Vestibular Disorders Glaucoma
Other Ear Disorders Eye Disorders	Other Disease of the Vestibular Disorders
Central Nervous System (CNS) Other Ear Disorders	Hypertension Other Disease of the
Undiagnosed Hypertension Central Nervous System (CNS)	Generic Illness Hypertension
Well care A	Well Care B
Undiagnosed Hypertension	Generic Illness
Accidents and E-codes Well care A	Well Care B
Acute myocardial infarction (AMI) Accidents and E-codes	Atrial fibrillation and flutter
Acute myocardial infarction (AMI)	Atrial fibrillation and flutter
Coronary atherosclerosis and other heart diseases	
Other arrhythmias	Cardiac arrest (includes VF)
Congestive heart failure	Acute hemorrhagic stroke
Ischemic stroke	Cerebrovascular disease
Peripheral vascular disease	Other cardiovascular diseases
Other vascular diseases	Pulmonary embolism

DVT	Pneumonia (non-TB, non-STD)
Influenza	Chronic obstructive pulmonary Disease (aka Emphysema)
Asthma	Acute respiratory infection
Respiratory symptoms	Other respiratory diseases
Reflux/ulcer disease	Biliary tract disease
Liver disease	Gastrointestinal bleeding
Other gastrointestinal disorders	Acute renal failure
Chronic renal failure	Endstage renal disease (ESRD)
UTI	Urinary incontinence
Hyperplasia of the prostate	Other genitourinary diseases
Pregnancy and childbirth	Menopause
Contraception and procreation	Dermatologic diseases
Rheumatoid arthritis	Osteoarthritis
Back pain	Osteoporosis
Other rheumatic diseases	Congenital disorders
Newborn conditions	Trauma
Hip fracture	Fractures
Poisoning and other injury	Motor vehicle accident
Screening: breast cancer	Screening: colon cancer
Screening: prostate cancer	Screening: cervical cancer

Table 4.

Item Missing Data on Covariates and Self-Report Health Conditions Multiply Imputed

Missing data percentage	Number	Percent
0	325	23.1%
<2%	1, 012	72.3%
≥2% to <5%	36	2.6%
≥5% to <10%	23	1.6%
≥10% to <15%	2	0.1%
≥15%	4	0.3%