Expanding the National Health Expenditure Accounts (NHEA) Technical Documentation

Chapter 7. Calibration processes: community dwelling population.

The goal of the calibration process was to create a set of disease group indicators that reflects the prevalence of disease groups both treated (resulted in a claim), as well as latent (not medically addressed within a given year). This task was accomplished in two steps: first calibrating 25 disease groups with self-report (SR) available in NHANES; then calibrating the remaining 80 disease groups. This chapter provides details of the methods used in each of the 2 steps.

Calibration of community dwelling population for disease groups available in NHANES

For each year the data from MCBS and NHANES were appended. Without loss of any generality, assume that in the combined data set, the first n observations are from MCBS and the last m observations are from NHANES. Let $X_i, i = 1, 2, \cdots, n, n+1, \cdots, n+m$ be the common covariates for $N = n + m$ subjects in the combined data set. Let $C_{ij} = 1$ denote the presence of claim for disease group $j = 1, 2, \cdots, p$ or set to 0 otherwise for subject $i = 1, 2, \cdots, n$ where p is the number of disease groups for which we have self-report data in NHANES. Note that this variable is defined only for subjects in MCBS. Let $S_{ij} = 1, i = n+1, \cdots n+m$ indicate if the subject self-reported having disease group j or set to 0 otherwise. Note that this variable is available only for subjects in NHANES.

Define a new indicator variable for the disease groups in a combined data set as follows:
$D_{ij} = 1$ if $S_{ij} = 1$ or $C_{ij} = 1,$
$D_{ij} = 0$ if $S_{ij} = 0$ and
$D_{ij} = .$ if $C_{ij} = 0$

That is, the subject is assumed to have presence of the disease group, if the self-report (in NHANES portion of the appended data set) or the claim (in the MCBS portion of the appended data set) indicates the presence of disease; and not having the disease, if the self-report indicates the subject does not have disease; and, if there was no claim for the disease the actual disease group status is missing. We now have a standard missing data problem by setting aside variables (S, C) and only working with (X, D) in the combined data set.

When imputing the missing values in D (only in MCBS) one needs to make sure that prevalence rates after imputation match with prevalence rates in NHANES. That is, the multiply imputed prevalence rates are calibrated to the observed prevalence rates in NHANES after adjusting for any differences in the covariate distribution. The sequential regression methodology was modified to result in calibrated prevalence indicator variables. Specifically, we defined a variable $R = 1$ for the subjects in MCBS and $R = 0$ for the subjects NHANES. The following steps describe an iteration in the sequential regression approach for calibrated multiple imputation using the variables (X, D, R):

1. Let $D_{(-j)}$ denote the collection of disease group indicators for all disease groups except disease group j. Construct a propensity score based on fitting a logistic regression model predicting R with $(X, D_{(-j)})$ as covariates and create strata based on the propensity scores. This step groups the subjects in the two surveys based on similarity of the covariates and other disease groups.
2. Within each propensity score class, estimate the prevalence rate based on the self-report, S_j, and the claims C_j. If the prevalence rate based on the claims is greater than or equal to that based on the self-report then set all missing D_j to 0. That is, no additional imputation is necessary and all those without a claim for that disease group are considered as not having that disease.
3. If the prevalence rate based on the claims is smaller than the self-report prevalence rate then randomly some missing D_j were set to 1 so that the prevalence rates after the imputation will match the self-report prevalence rates. We used several Bernoulli draws within each propensity score class to achieve this calibration.
4. Note that medical expenditure and disease groups without self-report are missing in the NHANES portion of the appended data. To be fully conditional, we imputed these missing values in the NHANES.
5. These steps were iterated across all diseases several times until the multiply imputed prevalence rates stabilized.

Calibration of community dwelling population for disease groups not available in NHANES

For calibrating disease groups that are not available in NHANES, relationships between the multiply imputed D_j and claims based $C_j, j = 1, 2, \cdots, p$ were developed. This can be viewed as an measurement error model and this relationship is then used to calibrate the disease groups, $k = p+1, p+2, \cdots, 105$. The following steps describe the imputation procedure for these disease groups:

1. Fit a measurement error regression model,
2. Fit two propensity models for the claims based disease groups, $\Pr(C_j = 1 | X)$ and $\Pr(C_k = 1 | X)$.
3. Match the disease groups j and k based on the propensity score to find the closest match of the measurement error model developed in step 1.
4. Use the closest match measurement error model to impute $D_k, k = p+1, p+2, \cdots, 105$.

This process was repeated for each year. Resulting claim-based and calibrated disease prevalence estimates from MCBS and NHANES SR estimates for each year are provided in Appendices 7_8a-k.