Published in final edited form as:

J Am Stat Assoc. 2013 January 1; 108(501): 34–47. doi:10.1080/01621459.2012.726889.

A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs

Roe Gutman [Assistant Professor],

Department of Biostatistics, Brown University, Providence, RI 02912

Christopher C. Afendulis [Lecturer], and

Department of Health Care Policy, Harvard Medical School, Boston, MA 02115

Alan M. Zaslavsky [Professor]

Department of Health Care Policy, Harvard Medical School, Boston, MA 02115

Roe Gutman: rgutman@stat.brown.edu; Christopher C. Afendulis: afendulis@hcp.med.harvard.edu; Alan M. Zaslavsky: zaslavsky@hcp.med.harvard.edu

Abstract

End-of-life medical expenses are a significant proportion of all health care expenditures. These costs were studied using costs of services from Medicare claims and cause of death (CoD) from death certificates. In the absence of a unique identifier linking the two datasets, common variables identified unique matches for only 33% of deaths. The remaining cases formed cells with multiple cases (32% in cells with an equal number of cases from each file and 35% in cells with an unequal number). We sampled from the joint posterior distribution of model parameters and the permutations that link cases from the two files within each cell. The linking models included the regression of location of death on CoD and other parameters, and the regression of cost measures with a monotone missing data pattern on CoD and other demographic characteristics. Permutations were sampled by enumerating the exact distribution for small cells and by the Metropolis algorithm for large cells. Sparse matrix data structures enabled efficient calculations despite the large dataset (≈ 1.7 million cases). The procedure generates m datasets in which the matches between the two files are imputed. The m datasets can be analyzed independently and results combined using Rubin's multiple imputation rules. Our approach can be applied in other file linking applications.

Keywords

Statistical Matching; Record Linkage; Administrative Data; Missing Data; Bayesian Analysis

1. Introduction

Estimating and forecasting health care costs of specific illnesses is crucially important for assessment of the long-term impact of programs designed to prevent or relieve specific diseases. Health care expenditures increase dramatically with time to death, but end-of-life expenditures have been shown to be a stable significant proportion of total expenditures over time (Lubitz and Riley, 1993; Hogan et al., 2001; Felder et al., 2000). These expenditures might be reduced by measures that shift patients from more to less expensive causes of death (CoD). For example, early interventions with diabetes patients might reduce deaths from complications of diabetes, which are relatively expensive. Study of this issue requires a dataset that includes cause of death, end-of-life expenditures and other demographic characteristics.

Medicare enrollment and claims data from the Centers for Medicare and Medicaid Services (CMS) include medical expenditures and demographic characteristics (e.g. age, sex, race, etc.), but not cause of death (CoD). The public-use Vital Statistics Mortality (VSM) records compiled from death certificates by the National Center for Health Statistics (NCHS) include CoD and demographic variables, but not medical expenditures. Although full linkage of these datasets would facilitate estimation of the mean and distribution of expenditures for each cause of death, it would require access to identifying information that is not publicly available, and so would be prohibitively expensive. Thus, linking records from these incomplete datasets would facilitate achievement of research aims.

File linkage has numerous administrative applications in marketing, customer relationship management, fraud detection, data warehousing, law enforcement and government administration. For these purposes, it is essential to link data on the same individual with regard to whom some decision or action will be taken. File linkage also has numerous applications in epidemiology, health, social science, and social policy. For such research applications, preservation of relationships among variables is crucial, but identification of specific individuals is not (Gu et al., 2003; D'Orazio et al., 2006).

The statistical literature addresses methods both for matching files and for analysis of linked datasets. Matching methods can be broadly classified into statistical matching and exact matching. In *statistical matching* (Rodgers, 1984; Rässler, 2002) there is no attempt to link records for the same individual; indeed the two files may represent disjoint samples. Thus the linking variables will typically be statistically (and perhaps scientifically) related variables, not identifying labels. Associations mediated through these variables can be estimated, but the partial associations (conditional on the matching variables) cannot (Rubin, 1974). Early applications ignored partial associations, essentially assuming conditional independence, when drawing inference from statistically matched files. Later procedures used multiple imputation (MI) (Rubin, 1987) and file concatenation to reflect prior uncertainty about partial associations, but did not estimate them (Rubin, 1986; Moriarity and Scheuren, 2003).

In exact matching or record linkage (Fellegi and Sunter, 1969; Scheuren and Winkler, 1993), the linked files represent overlapping or identical samples or populations, and matching of corresponding records is attempted in the two files that refer to the same individual. The matching variables might include identifiers such as names and addresses that identify entities or groups of entities but have no analytic significance in themselves. Even if potentially relevant variables like age or sex are used in matching, they are treated like identifiers: linkage models typically are concerned with the consistency with which the variables are recorded in the two files, and not with their scientific interpretation. Some exact matching algorithms calculate probabilities (or likelihoods under a probability model) that pairs of records from the two files are exact matches, which has been the subject of extensive work in statistics and computer science (Fellegi and Sunter, 1969; Winkler, 1988, 1993). These probabilities are then used in record matching algorithms, often of the “greedy” type proposed by Fellegi and Sunter (1969), which iteratively links and removes from the matching pool the pair with highest probability of a match. Possible matches with probabilities below a cutoff value are either clerically reviewed or declared nonmatches. The performance of this procedure is sensitive to the cutoff value (Belin and Rubin, 1995). An extension uses information from review of a few of the matched pairs to refit the model for match probabilities (Larsen and Rubin, 2001). These computationally simple algorithms may produce a globally suboptimal match because they do not consider interactions among different matched pairs. Linear sum optimization forces one-to-one matching after parameter estimation (Jaro, 1989), but uses estimated probabilities that consider pairs independently. Furthermore, none of those procedures offer any representation of uncertainty of the match.

To overcome these limitations, Bayesian approaches for record linkage were proposed (Fortini et al., 2001; Larsen, 2004). These procedures posit that a similarity measure comparing variables appearing in both files for possibly matched cases arises from a mixture of matches and non-matches. Bayesian calculations yield the probability of a match for each pair, and a one-to-one match is obtained using the posterior mode or minimizing a loss function. These procedures may rely heavily on the prior distributions of the parameters. Recently, another Bayesian approach was proposed that relies on a set of observable discrete matching variables rather than a similarity measure (Tancredi and Liseo, 2011). However, none of those Bayesian approaches use any information for matching contained in variables appearing in only one file.

Some methods for analysis of linked data use the probabilities of matches as weights in linear regressions that include non-linking variables from both files (Scheuren and Winkler, 1993; Lahiri and Larsen, 2005). An elaboration of this procedure (Scheuren and Winkler, 1997) iterates between regression analysis on the complete data set and record linking until no further improvement is obtained.

Wu (1995) proposed a Bayesian procedure that treats the unknown matches as missing data, and draws from the posterior distribution of the entire linkage, given identifying variables, while enforcing the restriction that each record can appear in at most one matched pair. Building on this key idea, we devised a general Bayesian procedure that jointly models the record linkage and associations between variables in the two files, thus improving matching and reducing bias in estimation of scientifically interesting relationships. In what follows, Section 2 defines notation and describes models, Section 3 presents a simulation, Section 4 describes the application of our algorithm to EoL medical expenditures, and Section 5 includes discussion and conclusions.

2. Methods

2.1 Notation and Model Structure

Let A and B label two files that we wish to link for analysis. Let $\mathbf{Y}_A$ and $\mathbf{Y}_B$ respectively represent variables available exclusively in file A and file B , and $\mathbf{Z}$ the “blocking variables” that are assumed to be reported identically in both files. Note that if substantially the same variable appears in both files but variations in reporting are modeled rather than assumed to be identical, the versions in the two files would be regarded as distinct variables that are components of $\mathbf{Y}_A$ and $\mathbf{Y}_B$ respectively.

“Blocking” is a common file-linkage technique that reduces the number of possible matches by considering only pairs that agree on blocking variables (Newcombe and Kennedy, 1962, 1988). Let $j = 1, \dots, J$ index J cells (blocks) defined by values of $\mathbf{Z}$, each containing I_{Aj} observations from file A and I_{Bj} observations from file B . The blocking and hence the index j for each case are determined by the values $\mathbf{Z}$, which might be regarded as fixed in advance or as outcomes of a random process, such as the occurrence of deaths falling into various blocks in our application. Data for records in files A and B are $(\mathbf{y}_{Aji}, \mathbf{z}_j)$ and $(\mathbf{y}_{Bji}, \mathbf{z}_j)$ respectively; $\mathbf{z}_j$ has a single cell index j because the blocking variables are constant within each cell. Let C_j represent the (unknown) matching permutation indicating how cases in file A must be reordered to match corresponding cases in file B for cell j and C_{jk} , $k = 1, \dots, K_j$ be the possible values of this permutation, where in cell j , $C_{jk}(i)$ indexes the record in file B that is matched to the i th record in file A , $1 \leq i \leq I_{Aj}$. Hence for a given matching permutation the linked data for one case are $(\mathbf{y}_{Aji}, \mathbf{y}_{BjC_{jk}(i)}, \mathbf{z}_j)$. For example, $C_{j5}(2) = 4$ means that in the fifth possible linking permutation for cell j , the second case in file A is linked to the fourth case in file B , and the linked data are $(\mathbf{y}_{A2}, \mathbf{y}_{Bj4}, \mathbf{z}_j)$. Because linkages are known with certainty only to the level of the blocking cell, we refer to these datasets as

“partially linked.” In what follows we assume that the indices in the two files are uninformative so any permutation is equally likely *a priori*.

When $I_{Aj} = I_{Bj}$, all cases can be linked and $K_j = I_{Aj}!$. When $I_{Aj} \neq I_{Bj}$, only $\min(I_{Aj}, I_{Bj})$ observations can be linked. We assume for the moment that the remaining records in the larger file (for that cell) represent entities that were (non-informatively) omitted from the smaller file, but all records in the smaller file have a match in the larger. (In fact such mismatches might occur due to misreporting of blocking variables in one of the files, but for now we assume that such cases are un-linkable.) In cell j there are $K_j = \max(I_{Aj}, I_{Bj})! / (I_{Aj} - I_{Bj})!$ possible permutations linking records in the two files. Extending our notation to cover the case $I_{Aj} \neq I_{Bj}$, let $U_{Aj} = U_A(C_j)$ be the set of indices of unmatched records in file A and $U_{Bj} = U_B(C_j)$ the corresponding set for file B .

Given C_j the density for one case is

$$L_{ji}(\theta, C_j) = \begin{cases} f_{AB}(y_{Aji}, y_{BjC_j(i)} | \theta, z_j) f_z(z_j | \theta), & i \notin U_{Aj} \\ f_A(y_{Aji} | z_j, \theta) f_z(z_j | \theta), & i \in U_{Aj} \\ f_B(y_{Bji} | z_j, \theta) f_z(z_j | \theta), & i \in U_{Bj} \end{cases} \quad (1)$$

where θ is the parameter vector, f_A , f_B , f_{AB} are respectively the marginal densities of $\mathbf{y}_{Aji}$ and $\mathbf{y}_{Bji}$ and their joint density, all conditional on $\mathbf{z}_j$, and f_z is the marginal density of $\mathbf{z}_j$. The three cases in (1) represent respectively matched entities and unmatched ones appearing only in A and only in B . Multiplying over cases in a cell, the likelihood for θ and C_j for an entire cell j is

$$L_j(\theta, C_j) = f_z(z_j | \theta)^{\max(I_{Aj}, I_{Bj})} \times \left(\prod_{i \in U_{Aj}} f_A(y_{Aji} | \theta, z_j) \right) \times \left(\prod_{i \in U_{Bj}} f_B(y_{Bji} | \theta, z_j) \right) \times \left(\prod_{i \notin U_{Aj}} f_{AB}(y_{Aji}, y_{BjC_j(i)} | \theta, z_j) \right) \quad (2)$$

Assuming that indices (within cell) in the two files are noninformative, we postulate a uniform prior distribution over the possible permutations $\mathbf{C}$ which are independent across cells and distinct from θ . Integrating over possible permutations C_j , and combining across cells, the likelihood for θ is

$$L(\theta | \text{Data}) = \prod_{j=1}^J \sum_{k=1}^{K_j} L_j(\theta, C_{jk}). \quad (3)$$

While the form of f_{AB} is specific to the application, it will often be convenient to express it as a product of conditional distributions, for example

$$f_{AB}(y_A, y_B | z, \theta) = f_{A(1)}(y_{A(1)} | z, \theta) \cdot f_{B(1)}(y_{B(1)} | z, y_{A(1)}, \theta) \cdot f_{A(2)}(y_{A(2)} | z, y_{A(1)}, y_{B(1)}, \theta) \cdot f_{B(2)}(y_{B(2)} | z, y_{A(1)}, y_{B(1)}, y_{A(2)}, \theta) \quad (4)$$

where $\mathbf{y}_{A(1)}, \mathbf{y}_{A(2)}, \mathbf{y}_{B(1)}, \mathbf{y}_{B(2)}$ are components of $\mathbf{y}_A$ and $\mathbf{y}_B$, respectively. The sub-models represented by the factors might include both models of scientific interest and models for relationships that are only useful for better identifying the matches, such as those linking inconsistently recorded blocking variables.

2.2 Bayesian computations

We adopt a Bayesian approach to inference because it enables us to create complete linked data sets using samples from the posterior distribution of $\mathbf{C} = \{C_j\}$, reflecting posterior uncertainty about θ . These data sets can then be analyzed by researchers as multiple imputations of the complete (linked) data and summarized to provide posterior probabilities for pairwise matches and other summaries of the links. In this formulation, we treat the unknown matching permutation as missing data, and use a Data Augmentation (DA) (Tanner and Wong, 1987) scheme that iterates between sampling the unknown linking permutation and sampling the parameters θ . As noted above, the prior distribution for $\mathbf{C}$ is uniform and independent of θ , whose prior distributions are model-specific.

Our algorithm is a Gibbs sampler with two major steps. In one step the unknown parameter θ are sampled given the permutation $\mathbf{C}$, and in the other step the missing permutation $\mathbf{C}$ is sampled conditional on the parameters θ . The augmented-data likelihood is

$$L(\theta, \mathbf{C} | \mathbf{D}) = \prod_{j=1}^J L_j(\theta, \mathbf{C}_j). \quad (5)$$

Algorithms for sampling θ given $\mathbf{C}$ are model-specific. In some cases it is possible to improve the efficiency of the computations by only recalculating at each step the parts of the likelihood that are affected by the permutations and not those that depend only on $\mathbf{Y}_A$ or only on $\mathbf{Y}_B$.

We next consider sampling of the permutations $\mathbf{C}$. The posterior distribution of C_j given θ is a multinomial distribution with probabilities

$$p(C_j = C_{jk} | Y_A, Y_B, Z, \theta) = L_j(\theta, C_{jk}) / \sum_{k'=1}^{K_j} L_j(\theta, C_{jk'}). \quad (6)$$

Note that the factors involving $f_Z(\mathbf{z}_j | \theta)$ cancel out and hence modeling of Z is inessential unless it informs estimation of parameters of interest θ .

Direct sampling from this multinomial distribution requires enumerating all possible permutations $\{C_{jk}\}$ in each cell and calculating the corresponding likelihoods. For small K_j , this is relatively fast. Furthermore, the likelihood (2) is a function of likelihoods $f_{AB}(\mathbf{y}_{Aji}, \mathbf{y}_{Bj} | C_{ji} | \theta, \mathbf{z}_j)$ for pairs from A and B . Calculating the likelihood for all possible pairs is the most expensive part of the computation, and the number of pairs increases only as the square of the number of cases in the cell. However, K_j increases in factorial order, and computing all the likelihoods across the possible permutations requires K_j additions over I_j pairs, thus becoming computationally demanding in large cells. In such cases we use a version of the Metropolis-Hasting algorithm proposed by Wu (1995). Each iteration of this algorithm consists of two substeps:

- i. Randomly choose two observations (i_1, i_2) in cell j and propose a new permutation C_{jk*} by swapping the values of $C_{jk}(i_1)$ and $C_{jk}(i_2)$.

- ii. Accept the new permutation with probability $\min \left(1, \frac{L_j(\theta, C_{jk*} | Y_A, Y_B, Z)}{L_j(\theta, C_{jk} | Y_A, Y_B, Z)} \right)$.

The calculation only involves the likelihoods for four linked pairs since all others cancel out. This procedure is repeated one or more times in each cell, at each iteration of the major Gibbs steps for drawing **C**. The swapping algorithm might be made more efficient by using an adaptive proposal distribution, but this was not essential in our application.

In cells where $I_{Bj} = I_{Aj}$, factors of the likelihood that depend only on $\mathbf{Y}_A$ or only on $\mathbf{Y}_B$, possibly in a factorization like (4), cancel out of the posterior (6) and therefore modeling them is inessential to sampling of C_j . If $I_{Bj} \neq I_{Aj}$ for at least some j , then it might be necessary to evaluate f_A or f_B due to their appearance in (2); the precise requirements depend on the direction of the inequality and the form in which f_{AB} is factorized in expressing the model. This could be complicated if f_A and f_B are not expressed in closed form; in this case there are several approaches to the needed calculations. The first is to impute the missing records in file *A* or *B* as needed to make $I_{Bj} = I_{Aj}$, thus adding another step to the DA sampler. For example, if f_B can be evaluated directly but f_A cannot, it is only necessary to impute the missing y_B to create a monotone missing data pattern (Little and Rubin, 2002) in which the likelihood depends only on f_B and f_{AB} . We then impute $I_{Aj} - I_{Bj}$ observations to file *B* in all cells where $I_{Aj} > I_{Bj}$ and match I_{Aj} pairs. In cells where $I_{Aj} < I_{Bj}$, we choose only I_{Aj} observations from file *B* to be matched, exploiting the monotone missing data pattern. Convergence of the sampler may be slowed, however, by the additional augmented data. Two alternatives avoid imputation by linking only $\min(I_{Aj}, I_{Bj})$ observations in each cell, leaving the rest unlinked and evaluating the factors of likelihood (1) for unlinked cases. The marginal likelihood may be calculated by integrating out the data from the missing records, either analytically, by summing if the missing records' data is discrete-valued, or by numerical approximations. Alternatively, the parameters of f_A or f_B (as needed) may be drawn directly from their posterior distributions given the linked data; it might be difficult, however, to specify a marginal model f_A or f_B consistent with the joint model f_{AB} and properly relate the parameters of the joint and conditional models.

Lastly, if there is missing data in $\mathbf{Y}_A$ or $\mathbf{Y}_B$, we can incorporate a step to impute the missing observations or exploit specific missing data patterns (convenient forms of monotone missingness) that allow us to work directly with the observed-data likelihood (as illustrated in Section 4).

Our procedure samples from the joint distribution of the possible permutations and the parameters of the models describing the data-generating process. Some of these parameters might be of interest to the investigator. For more general analysis, not necessarily foreseen at the time of sampling, imputed matched datasets can be used for estimation of other scientifically interesting parameters and their standard errors (Sec. 4.5).

3. Bivariate Normal Simulation

We illustrate the potential benefits of full-likelihood modeling of partially linked data with a bivariate normal example. The blocking variables $\mathbf{Z}$ are assumed to be pure labels, playing no further role in models. The pairs (y_A, y_B) are independently and identically distributed as

$$\begin{pmatrix} y_{Aji} \\ y_{Bji} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right) \text{ where } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this specification, parameters of the marginal distributions of y_A and y_B are fixed leaving only the correlation parameter ρ that depends on the matching to be estimated. This simplifies assessment of the information loss due to inexact matching.

Substituting into (3) we obtain the loglikelihood

$$l(\rho|Y_A, Y_B) = -\frac{I}{2} \log(1 - \rho^2) - \frac{1}{2} \sum_{j=1}^J \log \sum_{k=1}^{K_j} \exp \sum_{i=1}^{I_j} \left(\frac{y_{Aji}}{y_{BCjk(i)}} \right)^T \sum^{-1} \left(\frac{y_{Aji}}{y_{BCjk(i)}} \right) \quad (7)$$

where $I = \sum I_j$. This can be decomposed as $l(\rho|Y_A, Y_B) = -\frac{I}{2} \log(1 - \rho^2) + S_B + S_w$, where

$$S_B = -\frac{1}{2} \sum_{j=1}^J I_j \left(\frac{\bar{y}_{Aj}}{\bar{y}_{Bj}} \right)^T \sum^{-1} \left(\frac{\bar{y}_{Aj}}{\bar{y}_{Bj}} \right)$$

represents the likelihood component for the cell means,

$$S_w = \sum_{j=1}^J \log \sum_{k=1}^{K_j} \exp \left(\sum_{i=1}^{I_j} -\frac{1}{2} \left(\frac{y_{Aij} - \bar{y}_{Aj}}{y_{BCjk(i)} - \bar{y}_{Bj}} \right)^T \sum^{-1} \left(\frac{y_{Aij} - \bar{y}_{Aj}}{y_{BCjk(i)} - \bar{y}_{Bj}} \right) \right)$$

and the component for the within-cell deviations (summed over possible matches).

To compare the information content in fully- and partially-matched data, we calculated the Fisher (expected) information $-E^{-2} l / \rho^2$ for both likelihood functions as well as for the likelihood function for data reduced to cell means, representing the simplest consistent estimation strategy with partial matches. We divided the information from the latter likelihood functions (7) by that from an exactly matched sample of the same size to calculate relative efficiency (RE). Figure 1 summarizes RE for $0 < \rho < 1$ and for $I_j \equiv I_0 = 2, 3, 4, 5$. The horizontal line at 1 represent fully matched data and those at $1/I_0$ the RE when data for each cell of size I_0 are reduced to the cell mean, with $1/I_0$ times the original number of observations. The curved lines represent the relative information from (7) for the various cell sizes. For each cell size $RE \rightarrow 1$ as $\rho \rightarrow 1$, because we are able to distinguish with increasing certainty which observations should be matched. When $\rho \rightarrow 0$, $RE \rightarrow 1/I_0$; almost all the information is carried in the cell because the observations are nearly independent and each match is almost equally likely. RE decreases with increasing I_0 since there are more possible matches. These results coincide with those of DeGroot and Goel (1980) for the RE when $\rho = 0$.

Applying Bayesian estimation methods as in Section 2.2 yields very similar results, and additionally provides draws of the exact matches (see on-line supplement).

In this example, statistical matching by hot-decking (random matching) within cell would yield attenuated estimates of correlation. Because the probability that each observation is matched to a correlated observation is only $1/I_0$, the sample correlation of Y_A and Y_B would tend to ρ/I_0 instead of ρ .

4. Application: End-Of-Life Expenditures

In this analysis, we apply our procedure to Medicare enrollment and claims data and the Vital Statistics Mortality records to facilitate estimation of the mean and distribution of expenditures for each cause of death. Our file linking procedure provides researchers with multiple imputations of linked datasets sampled from the posterior distribution of the missing matches under models that include relationships of interest among variables in the two sets of records. Analyses of the resulting datasets can be combined using standard procedures for multiply imputed data.

4.1 Data

Almost all (95%) elderly (age ≥ 65 years) residents of the United States are covered by Medicare health insurance. The Medicare enrollment database contains information on all beneficiaries such as demographic characteristics (date of birth, female/male sex, black/non-black race), date of death, and state and county of residence. The file also identifies beneficiaries enrolled in a Medicare managed care plan ("Medicare Advantage" or HMO).

Beneficiaries fall into three groups with respect to availability of data on medical expenditures in the last 6 months of life from Medicare claims files. For each beneficiary enrolled in traditional, fee-for-service Medicare during the last six months of life (87% of Medicare-enrolled decedents in 2004), we calculated end-of-life "Medicare Part A" expenditures incurred during this time period from inpatient and outpatient hospitals, skilled nursing facilities, home health agencies, hospices and vendors of durable medical equipment. We also derived a claims-based measure of place of death (hospital inpatient, hospital outpatient or emergency room, or other). For a 20% random sample of these beneficiaries we also had "Medicare Part B" data, from which we calculated end-of-life expenditures for non-institutional providers, including physicians and laboratories. Reliable claims data is not available for beneficiaries enrolled in the Medicare Advantage program (13%). For these beneficiaries, the two expenditure variables are coded as missing and place of death is coded as "HMO".

The NCHS works with vital statistics registrars in each state to collect all death certificates for each calendar year, which are then collated into the VSM file. The VSM file included a similar set of demographic, residence, and time and place of death variables, but age at death is provided only in whole years, and the timing of death is specific only to the day of week and month of death. Cause of death (CoD) for each decedent was coded into one of 28 categories, based on the 39-category scheme developed by NCHS for recoding cause of death codes from the International Classification of Diseases (ICD), 10th Revision, Clinical Modification (ICD-10-CM). Table 1 summarizes all the variables in both files.

We restricted our study population to those aged 66 and older at the time of death to guarantee six months of Medicare coverage prior to death and accommodate the recording of age in completed years in the VSM file.

We blocked records by age, sex, race, month and day of week of death, and state and county of residence. Because of differences in assessing place of death in the two data systems we did not treat this as a blocking variable but rather as two distinct variables whose associations help with matching. After blocking, only 33.2% of the 1,724,368 decedents in the VSM file were exactly (one-to-one) matched to decedents in the EoL file. The rest were matched to cells with either the same number of decedents from each file (balanced cells, 31.5% of cases in 16.4% of cells) or with unequal number of decedents (unbalanced cells, 37.3% of cases in 36.5% of cells) as detailed in Table 2. An analysis using only the exactly matched cases would be inefficient, using only about a third of the data. Moreover, it might be biased since exactly matched cases more commonly occur in smaller counties while larger cells occur in the larger counties. We discarded data (9% of all data) from cells that had no observations from one of the files, since there was no possible match that respected the blocking.

4.2 Models

Following the notation of section 2 and the variable labels from Table 1, and for legibility suppressing indices for cells and individuals, we have $\mathbf{Z} = (1, F, B, \text{Age})$, $\mathbf{Y}_A = (Y_1, Y_2, \text{PoDe})$ and $\mathbf{Y}_B = (\text{CoD}, \text{PoDv})$ where $\text{CoD} = (\text{CoD}_1, \dots, \text{CoD}_p)$ and $\text{CoD}_p = 1$ if the

decedent's CoD was p and 0 otherwise. Define $g_u(x, \beta) = \exp(x^T \beta_u) / \sum_{u'=1}^U \exp(x^T \beta_{u'})$, the probability function of a multinomial logistic regression (with $\beta_1 \equiv 0$). Define $X = (Z, Y_B, \text{Age} \times \text{CoD})$, where $\text{Age} \times \text{CoD}$ is the interaction between Age and CoD. Also define the combined parameter vector $\Theta = (\beta_{B1}, \beta_{B2}, \beta_{A0}, \beta_{A1}, \beta_{A2}, \sigma_1, \sigma_2)$. As noted in Section 2.2, modeling of the blocking variables Z is inessential. Hence we define our model using a series of conditional models as in (4), summarized in Table 3. We specify $Y_B|Z$ through multinomial logistic regressions for $\text{CoD}|Z$ and $\text{PoDv}|Z, \text{CoD}$:

$$\begin{aligned} f_{B1}(u; z, \beta_{B1}) &= P(\text{CoD} = u | Z, \Theta) = g_u(Z, \beta_{B1}) \\ f_{B2}(u; z, \beta_{B2}) &= P(\text{PoDv} = u | Z, \text{CoD}, \Theta) = g_u(Z, \beta_{B2}) \end{aligned}$$

and similarly $\text{PoDe}|X$ (where X includes PoDv) as

$$f_{A0}(u, X) = P(\text{PoDe} = u | X, \Theta) = g_u(X, \beta_{A0})$$

Decedents who were enrolled in managed care ($\text{PoDe} = \text{HMO}$) are missing expenditure values. We have logged Part A expenditures Y_1 for all of the remaining decedents, but logged total expenditures Y_2 for only a 20% sample. This constitutes a monotone missing data pattern, under which (Y_1, Y_2) can be jointly modeled with two linear regressions, one for $Y_1|X, \text{PoDe}$ (f_{A1}) and the other (f_{A2}) for $Y_2|X, \text{PoDe}, Y_1$, both conditional on $\text{PoDe} = \text{HMO}$:

$$\begin{aligned} Y_1 &\sim N(X, \text{PoDe}^T \beta_{A1}, \sigma_1^2) \\ Y_2 &\sim N(X, \text{PoDe}, Y_1)^T \beta_{A2}, \sigma_2^2) \end{aligned}$$

To complete the Bayesian model we specify prior distributions for the unknown parameters θ . We assume independent normal priors with mean 0 for families of related coefficients corresponding to multilevel categorical variables, specifically those for age ranges, place of death (PoDv), causes of death (CoD), and age by CoD interactions. Thus, $\beta_{op} \sim N(0, \gamma_o^2)$, where o indexes a family of parameters for one of these predictors in a specific model (e.g., all the CoD coefficients in the model for Y_1) and p indexes a specific parameter in that family. This specification improves precision through shrinkage and facilitates a symmetrical prior specification of the categorical variables without singling out a baseline category (Gelman and Hill, 2007, Ch. 11). All other regression coefficients are given Uniform $(-\infty, \infty)$ prior distributions. Priors for residual variances σ_1^2, σ_2^2 and variances of coefficient families γ_o^2 are Uniform $(0, \infty)$. These prior distributions are improper. Because the posterior distribution is a finite mixture of the posterior distributions conditional on each of the possible matches, propriety of the unconditional posterior is equivalent to propriety of all of those conditionals. In our application, with a large sample of cases that are unambiguously matched, we are confident that the default priors we used yield proper posteriors. In an application with small samples, more attention should be given to verifying posterior propriety and if necessary the priors should be modified.

4.3 Model Fitting

Sampling from this model requires customized data structures and Markov Chain Monte Carlo (MCMC) algorithms. Each of the two files includes approximately 1.7 million decedents. Models f_{A0} , f_{A1} and f_{A2} each involve 178 or 179 explanatory variables. Most of these, however, are indicator variables and can be stored in an efficient sparse matrix data structure (Stoer and Bulirsch, 2002, Ch. 4), reducing the memory and computational requirements for regressions. We modified the customary algorithm to facilitate speedy calculation of the $\mathbf{X}^T\mathbf{X}$ matrix when rows are switched between iterations due to sampling of permutations (on-line supplement), exploiting the fact that matrix blocks $\mathbf{Y}_B^T\mathbf{Y}_B$, $\mathbf{Y}_B^T\mathbf{Z}$ and $\mathbf{Z}^T\mathbf{Z}$ that involve only variables within the same dataset or blocking variables do not depend on the permutations and hence need be calculated only once.

As proposed in Section 2.2, our Gibbs sampler alternately draws from the conditional distributions of the matching permutation and the model parameters. We sampled from (6) for small cells by calculating the exact distribution, and for large cells by the MCMC approach described in section 2 applying 30 “pair-switching” iterations per cell, per Gibbs iteration. Gibbs substeps iterated drawing from the conditional distributions of the model parameters. We sampled linear regression parameters β_{A1} , β_{A2} , σ_1^2 , σ_2^2 using standard methods (Gelman et al., 2003, Ch. 14), once for every draw of the permutations. We sampled multinomial logistic regression coefficients β_{A0} , β_{B1} and β_{B2} by an auxiliary variable method proposed by Holmes and Held (2006), which proved to be fast and to yield low autocorrelations. β_{B1} was sampled once for every draw of the matching permutations while β_{B2} was sampled twice and β_{A0} three times. The conditional posterior distributions of each γ_o are scaled Inverse- χ^2 with the appropriate degrees of freedom.

We dealt with unbalanced cells ($I_{Aj} - I_{Bj}$) using a procedure based on the first approach discussed in Section 2.2, adding a step to our Gibbs sampling procedure that imputed cases in file B for cells in which $I_{Bj} < I_{Aj}$. Since all the observations in each cell had the same age, sex and race, only PoDv and CoD had to be imputed. This can be done by sampling from two multinomial predictive distributions, expressed (suppressing arguments, for legibility) as:

$$\begin{aligned} P(\text{CoD}=u | \text{PoDv}, \mathbf{Z}, \mathbf{Y}_1, \mathbf{Y}_2, \boldsymbol{\Theta}) &\propto f_{B2} \times f_{B1} \times f_{A0} \times f_{A1} \times f_{A2} \\ P(\text{PoDv}=u | \mathbf{Z}, \text{CoD}, \mathbf{Y}_1, \mathbf{Y}_2, \boldsymbol{\Theta}) &\propto f_{B1} \times f_{A0} \times f_{A1} \times f_{A2} \end{aligned}$$

After these imputations, $I_{Bj} = I_{Aj}$ in each cell. The resulting monotone missing data pattern made it unnecessary to calculate f_A .

Initial analysis using the Raftery and Lewis (1995) statistic suggested that we could estimate the medians of the coefficients in f_{A0} , f_{A1} and f_{A2} to within an accuracy of ± 0.05 with probability 95% with a mean 456 iterations with only 8 coefficients requiring more than 1000 iterations. We applied the full sampling algorithm using three MCMC chains starting from different positions, with 1,000 Gibbs sample iterations and 30 “pair-switching” steps per iteration in cells of 5 or more observations, resulting in 1,000 samples from each chain. The Gelman and Rubin (1992) potential scale reduction statistics projected little potential improvement in the estimates by increasing the number of iterations ($\hat{R} < 1.1$ for all scalar parameters). Autocorrelations for most of the coefficients were modest. In the models of primary interest (f_{A0} , f_{A1} and f_{A2}), only 3 out of 536 coefficients had absolute autocorrelation exceeding 0.15 at lag 15, and only 16% of the 324 coefficients of f_{B1} and f_{B2}

did so. These results indicate that the MCMC chains converged to a common distribution. (Selected convergence and autocorrelation plots appear in the on-line supplements.)

4.4 Evaluations with simulated partial linkage

For the simulations described in this section we used data on the 555,227 decedents for whom exact matches were known. We ignored county, month and day of week of death and created 325,658 random cells with an equal number of observations from the two files based on the remaining variables. Cell sizes ranged from one to nine with the proportions (0.65, 0.194, 0.074, 0.035, 0.018, 0.01, 0.006, 0.007, 0.006), which resembles the size distribution of the complete dataset.

The first simulation evaluated the performance of our algorithm with balanced cells, comparing coefficient estimates from the original exactly matched datasets to those from our algorithms and from a dataset that was randomly matched within cells, a so-called “hot deck” procedure. We assessed the algorithm in terms of bias, standard error, and information loss. The estimands of interest are coefficients of CoD in the regression of total costs Y_2 on CoD and demographic characteristics (age, sex, race); for consistency with our other evaluations we included PoD as a predictor, although the final policy analysis will exclude PoD, which might be affected by a counterfactual change in CoD. Combining f_{A1} and f_{A2} , the coefficients of CoD are $\beta_{A^*, \text{CoD}(p)} = \beta_{A2, Y_1} \cdot \beta_{A1, \text{CoD}(p)} + \beta_{A2, \text{CoD}(p)}$. Using our procedure and the hot-deck procedure one can obtain Bayesian estimates for the coefficients of interest by imputing the missing matches. Figure 2A compares the posterior medians of these coefficients obtained from our procedure to those from the exactly matched data. The average absolute difference between the estimates is 0.06 (range 0.0005 to 0.18). The largest bias is observed for car accidents, the CoD with the smallest sample size, the lowest mean EoL expenditures and relatively high variance, compared to other CoD.

With the hot-deck procedure all of the coefficients are systematically attenuated toward 0 (Figure 2B), due to the assumption of conditional independence of $\mathbf{Y}_A$ and $\mathbf{Y}_B$ given $\mathbf{Z}$. The mean absolute value of the posterior medians was 0.44 with exact matches and 0.47 with our Bayesian procedure, but only 0.25 with the hot deck. The average absolute bias of posterior medians from the hot-deck procedure is 0.19 (range 0.02 to 0.6). This illustrates the bias of the hot deck due to the conditional independence assumption $\mathbf{Y}_A \perp \mathbf{Y}_B | \mathbf{Z}$.

Comparison of posterior standard error (Figure 2C) illustrates the information loss due to partial rather than exact matching. Direct calculation of the information in complex multi-parameter models is more difficult than in the univariate example in Section 3. Instead, we estimate the relative information content of the exactly and partially matched cases using the reciprocals of the posterior variances of the coefficients. Define I_{total} as the total amount of information in the data, I_1 and I_2 the information per case for the exactly and partially linked cases respectively, and n_1 and n_2 the corresponding sample sizes. Then

$$1/\text{Var}(\theta|\text{data}) \approx I_{\text{total}} \approx n_1 I_1 + n_2 I_2. \quad (8)$$

We calculated for each of the CoD coefficients the reciprocal of its posterior variance under our procedure to estimate I_{total} for each. Similarly, we estimate I_1 as the reciprocal posterior variance from the analysis of exactly matched cases analysis divided by the total number observations, and finally estimated I_2 by substitution into (8). The ratio I_2/I_1 represent the information content of partially missing matches cases relative to exactly matched cases. Its mean over all CoD coefficients was 0.69 (range 0.40-0.95). This quantifies how the correlation between the two datasets is used by our method to extract information from the partial matches.

To investigate the quality of the matches we examined the predicted probabilities of match. Figures 3A and 3B summarize the estimated probabilities of the most likely match of a specific decedent in file *A* to one in file *B*. For cells with 2 decedents, the proportions of the most likely match has a peak for nearly random permutations (50%) and a higher peak for near-certain matching (near 100%). The former corresponds to cells in which the CoD for the two decedents are either the same or have similar medical expenditures, and PoD is also the same or unobserved. Thus, there is little information from which to infer which cause of death is related to which expenditure, but this is not a problem for inference since either match gives similar information about expenditures. Similar peaks are observed around 1/3, 1/2 and 1 for cells with 3 cases. The dark shading in each bar represents the portion of most likely imputed matches in the probability range that are correct. These suggest that the probabilities are well calibrated (confirmed for larger cells by additional figures in the on-line supplement); for example about half of the most likely matches with estimated probability around 50% are actually correct, as are almost all of those with probability close to 100%. Columns 2 and 3 of Table 4 summarize these probabilities up to cell size 7. The average probability of the most likely match decreases with cell size. However, even in cells of size 7 decedents, correct matches are predicted 24% of the time, compared to 14% predicted by chance.

Table 4 (columns 4–5) summarizes the probability of identifying the correct *permutation* as most likely, by cell size. Although this probability falls off sharply with size, it is much greater than would be expected by chance. Figure 3C displays the distribution of this posterior probability for cells of 3 decedents (with 2 decedents, correct permutation is equivalent to correct individual match). As before, the dark shading indicates correct identification and suggests well-calibrated probabilities (again, confirmed by additional figures in the on-line supplement).

Two other simulations assessed the performance of our procedure when some cases are missing at random from one of the files, resulting in imbalanced cells. For each simulation we randomly deleted observations from cells with at least two observations with probability 17%, with the constraints that $I_{Aj} > 0$, $I_{Bj} > 0$ after deletion. We simulated deletions from the EoL and VSM files separately, to assess the distinct procedures used in cells with excess cases from one or the other of the files. This missingness pattern is otherwise similar to the one observed in the data.

Figures 4A and 4C compare point estimates (posterior medians) and standard errors of the coefficients of CoD in f_{A2} when records are removed from the EoL file to those based on the file prior to deletions. The point estimates are very similar, while the standard errors are generally slightly larger (mean percentage increase 5.4%), as expected due to the loss of observations.

Deleting records in the VSM file has slightly larger effects on point estimates (Figure 4B) and standard errors (Figure 4D, mean increase 15.9%). In this scenario our algorithm imputes the missing records in the VSM file. Larger differences in median values are observed for causes of death associated with lower mean EOL expenditure, which generally have larger relative variances and an excess of zero expenditures which are not captured well by our model. Despite the imperfect modeling in this case, MI estimates of the mean expenditure for each cause of death, using imputed linkages of the datasets, are similar for the three scenarios (no missing records, missing VSM records, and missing EoL records), with mean relative discrepancy of 1.7% for missing VSM vs. no missing and 0.6% for missing EoL vs. no missing. Standard errors are slightly larger with missing records (Figure 5). The relatively large discrepancies in coefficients for some low-cost CoD have little effect on the MI estimates. Our use of imputed matches is in this regard similar to predictive mean

matching for missing data: even if the predicted means are incorrect due to limitations of the model, the method still chooses reasonable donors for imputation as long as the model is able to approximate the ordering of the data (Andridge and Little, 2010). Furthermore, model misspecification will tend to increase estimated variance, yielding valid coverage intervals (Little and Rubin, 2002, Sec. 10.2.4). Similar results are observed in comparisons to estimates based on exact matches (on-line supplement).

4.5 Analysis of the full dataset

We describe the results in three main parts: (1) the multinomial logistic regression of the place of death (PoD) in the Medicare file on variables in the VSM file, (2) an MI analysis of mean costs by CoD, and (3) the linear regression of logged medical expenditures on CoD and demographic characteristics.

The multinomial logit model for place of death is not of scientific interest, but it may contribute to better matching. Despite discrepancies in assessment of this variable in the two files, we found them to be strongly associated, as expected. The posterior mean of Cohen's κ across the imputations of matches (excluding the cases with PoDe=HMO, that is, missing) was 0.63 (SD = 0.005), indicating substantial although not perfect agreement. Odds ratios in the multinomial also strongly favored agreement.

One of the primary goals of our analysis is to estimate mean EoL expenditure by cause of death using imputed matches. We sampled $m = 30$ imputations of the completed datasets from the posterior distribution of the missing matches. From the l -th imputed set we calculated the sample mean of Medicare part A and B expenditure for each CoD ($\hat{Q}_l$) and its estimated variance U_l . The combined multiple-imputation point estimate is

$\bar{Q}_m = m^{-1} \sum_{l=1}^m \hat{Q}_l$, with estimated variance $T_m = m + ((m+1)/m) B_m$, where $\bar{U}_m = m^{-1} \sum_{l=1}^m U_l$ and $B_m = (m-1)^{-1} \sum_{l=1}^m (\hat{Q}_l - \bar{Q}_m)^2$ (Rubin, 1987, ch. 3). Table 5 presents the causes of death with the three highest and three lowest EoL Medicare part A and B expenditure. Death in a car accident is associated with the lowest mean medical expenditures, and death from non-Hodgkin's lymphoma or leukemia with the highest. Figure 6 compares the means and posterior standard deviations of EoL expenditure as obtained from the exactly matched cases to those from the multiply imputed datasets. Point estimates show similar patterns but some substantial differences: the mean absolute difference across all causes of death was \$1,620, and in 18 out of the 28 causes of death there was an absolute difference of over a \$1,000. Such differences are not surprising since the exactly-matched cases systematically underrepresent the most populous counties, where multiple deaths are more likely to occur in the same blocking cell. The posterior standard deviations of these

means are fairly similar. The quantity $\hat{\eta}_m = [(m+1)/m] \frac{B_m}{T_m}$ is approximately the fraction of information about Q that is missing due to inexact matching (Rubin, 1987, Sec. 3.3). The mean fraction of missing information for all causes of death was 23% (standard deviation = 8%), although only one third of the decedents were exactly matched.

We also compared the coefficient estimates of the logged linear regression (summarized as posterior median and standard error) from our procedure using the entire dataset to those computed from only the exactly matched decedents. (We omit detailed reporting of hot-deck estimates, which suffered from the same attenuation reported in Section 4.4.) Point estimates are fairly similar, with some notable differences that might seem small on the log expenditure scale, but could represent substantial differences in expenditures (as observed in the MI analysis). Out of 168 CoD and CoD×Age coefficients, posterior medians for 72 differed by more than 0.05, corresponding to about 5% difference in expenditures. The

standard error plot (on-line supplement) shows more posterior variability in estimates from the complete dataset than the exactly matched one. We conjecture two reasons for this phenomenon. The first is the biased population of the exact-matched cells, as discussed for the MI analysis. Second, the unbalanced cells with an excess EOL cases have a concentration of zero expenditures, which are poorly fit by our log-transformed model. However, despite the imperfect linking model, the matches and the MI analysis results appear plausible, similar to the results observed at the end of Sec. 4.4.

It is noteworthy that the estimator in our multiple-imputation analysis is different from that in the log-linear linking model, although it addresses the same scientific issue of expenditures by CoD. While the log-linear model is relatively well adapted to identify likely matches reflecting the expenditure-CoD relationship, using it to estimate mean expenditures would require retransformation of logged expenditures, which is sensitive to distributional assumptions (Manning and Mullahy, 2001), unlike direct calculation of means in the imputed matched datasets.

5. Discussion

Analysis of partially linked datasets is likely to become increasingly important as researchers and policy analysts seek to integrate administrative and registry datasets while adapting to privacy regulations that limit access to unique identifiers. In our application, separate datasets, containing either cause of death or medical expenditures, were readily available, but exact linkage through unique personal identifiers would have been prohibitively expensive. Using only the exactly matched cases may be inefficient and probably biased, while the hot deck procedure (random permutations) is almost certainly biased toward null relationships.

Our procedure shows promising results by modeling both scientifically important relationships and additional variables that can help in the matching process. Our procedure provides draws of the parameters of the linking models. It also generates completely linked data sets that can be analyzed by any model and combined using standard multiple-imputation rules. This approach makes fewer demands on the time and expertise of researchers while giving them more flexibility in the analysis of the data, as illustrated in Section 4.5. It also provides probabilities for any potential match that can be used to assess the strength of the matching information and the potential benefits of using more matching variables and models.

Blocking restricts the potential matches, making sampling more computationally feasible by breaking the overall matching task into numerous much smaller tasks. In our application, the large number of unbalanced blocking cells (constituting about a third of our data) was almost certainly due not only to file undercoverage but also to inconsistent recording of blocking variables (especially county of residence of decedents with multiple homes), which placed the correct matches in different blocking cells. We were able to deal with this by treating the shortfall of cases in a cell as missing data. A better algorithm might allow matching across cells, recovering information from these misclassified cases. Development of such an algorithm requires further research, perhaps using more sophisticated network data structures that take into account the degree of similarity between different cells to generate proposal matches.

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grant P01AG031098 from the National Institute on Aging. The authors thank Thomas Herzog for helpful comments.

References

- Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*. 2010; 78:40–64. [PubMed: 21743766]
- Belin TR, Rubin DB. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*. 1995; 90:694–707.
- DeGroot MH, Goel PK. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*. 1980; 8:264–278.
- D'Orazio, M.; Di Zio, M.; Scanu, M. *Statistical Matching Theory and Practice*. Hoboken: John Wiley & Sons; 2006.
- Felder S, Meier M, Schmitt H. Health care expenditure in the last months of life. *Journal of Health Economics*. 2000; 19:679–695. [PubMed: 11184799]
- Fellegi IP, Sunter AB. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969; 64:1183–1210.
- Fortini M, Liseo B, Nuccitelli A, Scanu M. On Bayesian record linkage. *Research in Official Statistics*. 2001; 4:185–198.
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis*. 2nd. New York: Chapman & Hall; 2003.
- Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press; 2007.
- Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. 1992; 7:457–472.
- Gu, L.; Baxter, R.; Vickers, D.; Rainsford, C. Tech rep. CSIRO Mathematical and Information Sciences; 2003. Record Linkage: Current Practice and Future Directions.
- Hogan C, Lunney J, Gabel J, Lynn J. Medicare beneficiaries' costs of care in the last year of life. *Health Affairs*. 2001; 20:188–195. [PubMed: 11463076]
- Holmes CC, Held L. Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*. 2006; 1:145–168.
- Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*. 1989; 84:414–420.
- Lahiri P, Larsen MD. Regression analysis with linked data. *Journal of the American Statistical Association*. 2005; 469:222–230.
- Larsen, MD. Record linkage using finite mixture models. In: Gelman, A.; Meng, XL., editors. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. Wiley; 2004. p. 309–318.
- Larsen MD, Rubin DB. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*. 2001; 96:32–41.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. Second. Hoboken: Wiley-Interscience; 2002.
- Lubitz J, Riley G. Trends in Medicare Payments in the Last Year of Life. *New England Journal of Medicine*. 1993; 328:1092–1096. [PubMed: 8455667]
- Manning WG, Mullahy J. Estimating Log Models: To Transform or Not to Transform? *Journal of Health Economics*. 2001; 20:461–494. [PubMed: 11469231]
- Moriarity C, Scheuren F. A Note on Rubins Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*. 2003; 21:65–73.
- Newcombe HB, Kennedy JM. Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. *Journal of the American Statistical Association*. 1962; 82:528–540.

- Newcombe, HB.; Kennedy, JM. Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford: Oxford University Press; 1988.
- Raftery, AE.; Lewis, SM. The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: Gilks, WR.; Spiegelhalter, DJ.; Richardson, S., editors. Practical Markov Chain Monte Carlo. Chapman and Hall; 1995. p. 115-130.
- Rässler, S. Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches. Springer Verlag; 2002.
- Rodgers WL. An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics*. 1984; 2:91–102.
- Rubin DB. Characterizing the Estimation of Parameters in Incomplete-Data Problems. *Journal of the American Statistical Association*. 1974; 69:467–474.
- Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*. 1986; 4:87–94.
- Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc; 1987.
- Scheuren F, Winkler WE. Regression Analysis of Data Files That are Computer Matched. *Survey Methodology*. 1993; 19:39–58.
- Scheuren F, Winkler WE. Regression Analysis of Data Files that are Computer Matched II. *Survey Methodology*. 1997; 23:157–165.
- Stoer, J.; Bulirsch, R. Introduction to Numerical Analysis. 3rd. Berlin, New York: Springer-Verlag; 2002.
- Tancredi A, Liseo B. A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*. 2011; 5:1553–1585.
- Tanner MA, Wong WH. The calculation of posterior densities by data augmentation. *Communications of the Association for Computing Machinery*. 1987; 5:563–567.
- Winkler WE. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 1988:667–671.
- Winkler WE. Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 1993:274–279.
- Wu Y. Random Shuffling: A New Approach to Matching Problem. *ASA Proceedings of the Statistical Computing Section, American Statistical Association*. 1995:69–74.

Figure 1. Expected information for ρ in bivariate normal model, as a function of ρ , for cells of size 2 (top curve) to 5 (lowest curve)

(A) Posterior medians: Permutation sampling vs. exact matches

(B) Posterior medians: Random sampling vs. exact matches

(C) Standard errors: Permutation sampling vs. exact matches

Figure 2. Simulated partial linkage: Comparison of estimating procedures to exactly known matches for Medicare Part A & B coefficients

Figure 3.

Probability of most likely match / permutation in simulated partial linkage, for cells of size 2 (A) and 3 (B, C). Each bar represents a range of predicted probabilities that the most likely match or permutation is the correct one. The combined light/dark bars form a histogram of these predicted probabilities while the dark part of each bar represents the instances in which the correct match or permutation was correctly identified. For example, for cells of size 2, the first bar indicates that among pairs with predicted probability around 0.53 of being matches, slightly over half were correct. The last bar indicates that among those with predicted probabilities near 0.97, almost all were correct.

(A) Posterior medians: without and with simulated missing EoL observations

(B) Posterior medians: without and with simulated missing VSM observations

(C) Standard errors: without and with simulated missing EoL observations

(D) Standard errors: without and with simulated missing VSM observations

Figure 4. Estimation of Part A & B coefficients under simulated partial linkage, comparing estimates with balanced cells to estimates with simulated missing cases in either EoL or VSM file

(A) Mean expenditures: without and with simulated missing EoL observations

(B) Mean expenditures: without and with simulated missing VSM observations

(C) Standard errors: without and with simulated missing EoL observations

(D) Standard errors: without and with simulated missing VSM observations

Figure 5. Estimates of mean expenditures by cause of death under simulated partial linkage, comparing estimates with balanced cells to estimates with simulated missing cases in either EoL or VSM file

(A) Mean expenditures

(B) Posterior Standard Deviations of Mean

Figure 6. Estimates of mean expenditures by cause of death, comparing estimates using permutation sampling to to those using exactly matched cases

Table 1
Variables in data sets

	Code	Variable name	Values
Blocking variables, in both files	F	Female	male, female
	B	Black	Non-Black, Black 5 categories, 5-year ranges, age 66 and older
	Age	Age	
	Time	Time of Death	Month and day of the week
	State	State of residence	Federal Information Processing Standard (FIPS) State
	County	County of Residence	FIPS County
Predictor variables EoL	PoDe	Place of Death	In, Out, other, missing (HMO)
Predictor variables VSM	PoDv	Place of Death	In, Out, other, NA
	CoD	Cause of Death	28 Causes
	Age $\times$ CoD	Interaction	1:5 $\times$ 1:28
Outcome variable EoL	Y ₁	Medicare Part A log(expenditure + 50\$)	Non-negative continuous
	Y ₂	Medicare parts A & B log(expenditure + 50\$)	Non-negative continuous

EoL - End of Life dataset (Medicare)

VSM - Vital Statistics Mortality dataset (NCHS)

Table 2
Distribution of decedents by blocking status

Cell Type	Number (%) of Cells	Number (%) of Decedents VSM	Number (%) of Decedents EoL
Exact matches (1 EoL, 1 VSM)	555,227 (47.1)	555,227 (32.2)	555,227 (33.2)
Inexact, equal number cases	193,007 (16.4)	526,149 (30.5)	526,149 (31.5)
Inexact, unequal number case	430,050 (36.5)	642,992 (37.3)	588,982 (35.3)
Total	1,178,284	1,724,418	1,670,408

*The numbers in parenthesis are percentages

Table 3
Summary of models

Model	Outcome	Predictors	Model form	Parameters
f_Z	Demographics (age, sex, race), day/month of death, county	—	Unspecified	—
f_{B1}	CoD (VSM)	Age, sex, race	Multinomial logistic regression	β_{B1}
f_{B2}	PoDv = Place of death (VSM)	Age, sex, race, CoD	Multinomial logistic regression	β_{B2}
f_{A0}	PoDe = Place of death, HMO (EoL)	Age, sex, race, Place of Death (VSM), CoD, CoD×Age	Multinomial logistic regression	β_{A0}
f_{A1}	Y_1 = Medicare Part A expenditures (EoL, non-HMO only)	Age, sex, race, Place of Death (EoL), CoD, CoD×Age	Linear regression	β_{A1}, σ_1
f_{A2}	Y_2 = Medicare part A & B expenditures (EoL, non-HMO only)	Age, sex, race, Place of Death (EoL), Medicare part A expenditure, CoD, CoD×Age	Linear regression	β_{A2}, σ_2

EoL - End of Life dataset (Medicare)

VSM - Vital Statistics Mortality dataset (NCHS)

CoD - Cause of Death

HMO - Decedents who were enrolled in managed care

Table 4
Matching Accuracy in Simulation

Cell Size	<i>P</i> (Correct Match)		<i>P</i> (Correct Permutation)	
	Hot deck	Permutation sampling	Hot deck	Permutation sampling
2	0.50	0.72	0.5000	0.7199
3	0.33	0.53	0.1667	0.4021
4	0.25	0.40	0.0417	0.1860
5	0.20	0.32	0.0083	0.0679
6	0.17	0.27	0.0014	0.0234
7	0.14	0.23	0.0002	0.0080

Table 5
Average Part A & B Expenditure Obtained by Multiple Imputation Procedure

	Mean	2.50%	97.50%
Car Accidents	14,025	11,681	16,370
Alzheimer's disease	15,343	14,822	15,865
Hypertensive heart disease	17,683	16,595	18,771
Nephritis	39,541	37,529	41,554
Leukemia	40,785	38,874	42,697
Non Hodgkin's lymphoma	41,533	39,802	43,264