Research Article

Statistics in Medicine

Received 25 February 2009, Accepted 23 April 2010 Published online 15 June 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3974

A multiple imputation approach to disclosure limitation for high-age individuals in longitudinal studies

Di An[a]*†, Roderick J. A. Little[b] and James W. McNally[c]

Disclosure limitation is an important consideration in the release of public use data sets. It is particularly challenging for longitudinal data sets, since information about an individual accumulates with repeated measures over time. Research on disclosure limitation methods for longitudinal data has been very limited. We consider here problems created by high ages in cohort studies. Because of the risk of disclosure, ages of very old respondents can often not be released; in particular, this is a specific stipulation of the Health Insurance Portability and Accountability Act (HIPAA) for the release of health data for individuals. Top-coding of individuals beyond a certain age is a standard way of dealing with this issue, and it may be adequate for cross-sectional data, when a modest number of cases are affected. However, this approach leads to serious loss of information in longitudinal studies when individuals have been followed for many years. We propose and evaluate an alternative to top-coding for this situation based on multiple imputation (MI). This MI method is applied to a survival analysis of simulated data, and data from the Charleston Heart Study (CHS), and is shown to work well in preserving the relationship between hazard and covariates. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: confidentiality; disclosure protection; longitudinal data; multiple imputation; survival analysis

1. Introduction

Statistical disclosure control (SDC) procedures deliberately alter data collected by statistical agencies before release to the public, to prevent the identity of survey respondents from being revealed. These methods have increased in importance, with the extensive use of computers and the internet. The goal of SDC methods is to reduce the risk of disclosure to acceptable levels, while releasing a data set that provides as much useful information as possible for researchers. One aspect of this is the ability to draw valid statistical inferences from the altered data.

Top-coding (TC) is a simple and common SDC method that seeks to prevent disclosure on the basis of extreme values of a variable, by censoring values above a pre-chosen 'top-code'. For example, in surveys that include income, extremely high income values are considered to be sensitive and to have the potential to reveal the identity of respondents. By recoding income values greater than a selected 'top-code' value to that value, the disclosure risk of respondents with very high income is reduced.

It is left to the analyst to decide how top-coded data are analyzed. One approach is to categorize the variable so that top-coded cases all fall in to one category—this is sensible, but does not work for analyses that treat the variable as continuous. Another approach is to ignore the TC and treat the top-coded values as the truth. This method is straightforward, but clearly the data distribution is distorted and biased estimates will be obtained. A better method is to treat the extreme values as censored. Under an assumed statistical model, maximum likelihood (ML) estimates can be obtained using algorithms such as the Expectation-Maximization (EM) algorithm [1]. This method is model-based, and should yield good inferences if the model is correctly specified. But we expect this method to be quite sensitive to

[a]*Merck Research Laboratories, Merck & Co., Inc., P.O. Box 1000, Upper Gwynedd, PA 19454, U.S.A.*
[b]*Department of Biostatistics, University of Michigan, 1420 Washington Heights M4045, Ann Arbor, MI 48109-2029, U.S.A.*
[c]*Institute for Social Research, University of Michigan, 330 Packard Street, Ann Arbor, MI 48109, U.S.A.*
**Correspondence to: Di An, Merck Research Laboratories, Merck & Co., Inc., P.O. Box 1000, Upper Gwynedd, PA 19454, U.S.A.*
†E-mail: di_an@merck.com

Contract/grant sponsor: National Institute of Child and Human Development; contract/grant number: P01 HD045753
Contract/grant sponsor: National Institute of Aging; contract/grant numbers: P30AG004590, R03AG021162

model misspecification, especially when the upper tail of the assumed distribution differs markedly from that of the true distribution. The data users can also apply an imputation method to the top-coded data set and fill in the censored values. A limitation is that the imputed data fail to reflect imputation uncertainty, and imputations are sensitive to assumptions about the right tail of the distribution. An and Little [2] propose an alternative to TC based on multiple imputation (MI), which allow valid inferences to be created based on applying multiple imputation combining rules described by Reiter [3], while preserving the SDC benefits of TC; for other discussions of MI in the disclosure control setting, see Little [4], Rubin [5], Raghunathan *et al*. [6], Little *et al*. [7], Reiter [8–10]. The methods in An and Little [2] are extended to handle covariate information in An and Little (2007, unpublished).

We propose here MI for disclosure control in the context of the treatment of age in longitudinal data sets. Because of the risk of disclosure, ages of very old respondents can often not be released; in particular this is a specific stipulation of HIPAA regulations [11, 12] for the release of health data for individuals. TC of individuals beyond a certain age (say 80) is a standard way of dealing with this issue, and it may be adequate for cross-sectional data, since the number of cases affected may be modest. However, this approach has severe limitations in longitudinal studies, when individuals have been in the study for many years; for example, consider an individual in a 40-year longitudinal study, who enters the study at age 42 at time t and is still in the study at age 82 at time $t+40$. The age at time $t+40$ cannot simply be replaced by a top code of 80, since age at time $t+40$ can be inferred by simply adding 40 to the age at time t. A strict application of TC would replace all individuals aged 40 or older at time t by a top code of 40, but this strategy seriously limits the ability to do longitudinal analysis, particularly survival analyses where chronological age is a key variable of interest. In particular, since age at entry is a marker for cohorts, differences in outcomes between cohorts aged 40 or greater at entry can no longer be estimated, since these cohorts are all top-coded to the same value.

This problem arises in the Charleston Heart Study [13], a longitudinal study that collects data over 40 years (1960–2000). The study was originally conducted to understand the natural aging process in a community-based cohort. The data include baseline characteristics such as age, race, gender, occupation, education, as well as death information for respondents. For longitudinal data from this study to be included in the National Archive of Computerized Data on Aging (NACDA)—the gerontological data archive at the University of Michigan, individual ages beyond age 80 cannot be disclosed because of HIPAA regulation, given the geographic specificity of the respondents. Also, given the longitudinal nature of the data, a TC approach would need to be applied to all individuals aged 40 or older in 1960, which has the limitation discussed above.

The goal of this research is to develop MI methods that adequately limit disclosure risk and preserve the relationship between hazard and covariates in survival analysis. We propose a non-parametric MI method, specifically a stratified hot-deck (HD) procedure, where we create strata and draw deleted ages with replacement from each stratum. Our method multiplies imputed values of two age variables—entry age and final age (age at death or age at last contact).

To assess the proposed method, we apply a proportional hazard (PH) model to the multiply imputed data sets, calculate estimates of regression coefficients for putative risk factors, and compare these estimates, and the corresponding estimates from top-coded data, with estimates from the PH model applied to the original data prior to SDC. We also present simulation studies where data are simulated according to a known survival model, and inferences for parameters of this model are compared with the true values.

The remainder of this paper is organized as follows. Section 2 presents our SDC approaches for longitudinal data and describes the corresponding methods of inference for regression coefficients. Section 3 describes a simulation study to evaluate the approaches in Section 2, and Section 4 applies the methods to the Charleston heart study (CHS) data. Section 5 gives the discussion and future work.

2. Methods

2.1. SDC methods for longitudinal data

An and Little [2] proposed SDC methods for a single variable with extreme values. In this paper, we investigate a more complicated situation with longitudinal data, where two age variables are subject to TC.

Let Y_{end} denote participants' age at the end of study (referred to as final age) and Y_{start} denote their entry age. Let C be the censoring indicator. Let L represent the length of study and S denote survival time. Individuals with $S \geqslant L$ are treated as censored ($C=1$), and otherwise as died ($C=0$). We consider individuals with values of Y_{end} greater than a particular value y_0 to be at risk of disclosure, and refer to these individuals as sensitive cases. Thus, values of Y_{end} and Y_{start} of the sensitive cases are treated as sensitive values. We consider the following SDC approaches:

(a) *TC*: Replace values of Y_{end} greater than y_0 by y_0 and replace values of Y_{start} greater than $y_0 - L$ by $y_0 - L$. The resulting data set is referred to as 'top-coded' data.

(b) *HDMI*: Classify sensitive and non-sensitive values into strata, to be defined below. Then delete the values of Y_{end}, Y_{start}, and C for sensitive cases and replace them with random draws from the set of deleted values in the same stratum. Our stratified HDMI method is similar to the approach described in An and Little (2007, unpublished), where we assign the deleted data into strata based on the predicted values of either age variables from regression on other variables, and apply HDMI within each stratum to impute deleted values. The following choices of strata are considered here:

(i) HD1: Strata are defined by the predicted values of the logarithm of hazard computed from a PHs model for survival. This choice is motivated by the idea that if survival analysis is a primary analysis involving the imputed age data, imputing within strata defined by predicted hazard will minimize the distortions of survival analysis applied to the imputed data.
(ii) HD2: We develop a two-way stratification, where strata are defined by both the predicted values of the logarithm of hazard as in (i), and the predicted values of entry-age from the regression of entry-age on other variables involved.
(iii) HD3: Stratification depends on the value of C. For individuals that are censored, strata are defined by the predicted values of entry-age; and for those not censored, strata are defined by both the predicted values of the logarithm of hazard and the predicted values of entry-age as in (ii).
(iv) HDU: Unstratified HDMI, which we include as a baseline for comparison with the stratified methods.

Note that for methods HD1 and HD2, we delete values of Y_{end}, Y_{start}, and C of sensitive cases and jointly impute these values. HD3 retains values of C and imputes Y_{end} and Y_{start} only.

It is worth noting that for the above stratified methods, we perform regression only on the set of sensitive cases with values deleted to obtain the predicted values. We also considered an alternative way of creating strata, where we perform regression on the complete data, and then stratify the sensitive cases for imputation. These methods did not perform as well in terms of empirical bias, root mean squared error (RMSE), and confidence coverage in the simulation study reported in Section 3, hence we do not consider them further.

2.2. *Methods of inference*

We consider the properties of the SDC methods for inferences about the regression coefficient, where a PH model is fitted to the data set before and after imputation. The following estimates and associated standard errors are considered:

(1) *Before deletion* (*BD*): The estimates of regression coefficients calculated from original data prior to SDC, used as a benchmark for comparing SDC methods.
(2) *TC*: The estimates of regression coefficients calculated from the top-coded data set.

The standard errors for methods (1) and (2) are computed by the bootstrap.

The four remaining methods HD1–HD3 and HDU are as described in Section 2.1, yielding D MI data sets. The MI estimate is calculated as

$$\hat{\theta}_{\text{MI}} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}^{(d)}, \tag{1}$$

where $\hat{\theta}^{(d)}$ is the parameter estimate from the dth data set. The MI estimate of variance is

$$T_{\text{MI}} = \text{Var}(\hat{\theta}_{\text{MI}}) = \bar{W} + B/D, \tag{2}$$

where $\bar{W} = \sum_{d=1}^{D} W^{(d)}/D$ is the average of the within-imputation variances $W^{(d)}$ for imputed data set d, and $B = \sum_{d=1}^{D}(\hat{\theta}^{(d)} - \hat{\theta}_{\text{MI}})^2/(D-1)$ is the between-imputation variance. Formula (2) differs from the original MI formula for missing data (where B is multiplied by a factor $(D+1)/D$, see e.g. Little and Rubin [14], p. 86), for reasons discussed in Reiter [3].

3. Simulation study

A simulation study was carried out to evaluate the SDC methods in Section 2. We computed estimates of regression coefficients, their corresponding variances and confidence intervals from the imputed and top-coded data sets, and compared them with those calculated from the original data set prior to SDC.

Table I. Hazard rate for simulation study, scenario I and II.

	Age at death					
	30–40	40–50	50–60	60–70	70–80	80+
Category 1 Male Entry-age 30–40	0.003	0.005	0.011	0.04	0.06	0.1
Category 2 Female Entry-age 30–40	0.024	0.004	0.0088	0.032	0.048	0.08
Category 3 Male Entry-age 40–50		0.0075	0.0165	0.06	0.09	0.15
Category 4 Female Entry-age 40–50		0.006	0.0132	0.048	0.072	0.12

Table II. Hazard rate for simulation study, scenario III.

	Age at death					
	30–40	40–50	50–60	60–70	70–80	80+
Category 1 (0,0) Male Entry-age 30–40	0.003	0.005	0.011	0.04	0.06	0.1
Category 2 (1,0) Female Entry-age 30–40	0.003	0.005	0.011	0.04	0.06	0.1
Category 3 (0,1) Male Entry-age 40–50		0.0075	0.0165	0.06	0.09	0.15
Category 4 (1,1) Female Entry-age 40–50		0.006	0.0132	0.048	0.072	0.12

3.1. Study design

For simplicity we simulated survival data with only two binary covariates, representing gender (male and female) and entry-age (say 30–40 and 40–50). The values of these variables were simulated from a multinomial distribution for the four categories. Values of entry-age were generated from a uniform distribution. Survival times (in years) were generated from piece-wise exponential distributions with hazard rates specified in Tables I and II. An individual was treated as censored if (s)he survived more than 40 years from age at entry. We investigated the following three scenarios:

Scenario I: Males and females have same entry-age distributions. Entry-age values are generated from the Uniform distribution with the ranges 30–40 and 40–50. For both males and females, values from the former distribution are 1.5 times those from the latter distribution. Gender and entry-age have additive effects on the log-hazard of survival; hazards are shown in Table I.

Scenario II: The distribution of entry-age differs for males and females. Males have the same distribution of entry-age as in Scenario I. Entry-age values for females are generated in a similar manner, except that 70 per cent of the values lies within the range of 35–45. Gender and entry-age effects on survival are as for Scenario I.

Scenario III: Males and females have the same entry-age distribution as specified in Scenario I, and there is interaction between entry-age and gender on the log-hazard of survival; hazards are shown in Table II.

In this study, we considered individuals with final age greater than or equal to 75 years to be at risk of disclosure, and refer to these individuals as sensitive cases. About 25 per cent of the cases have sensitive values, and about one-third of the cases are censored. For each simulated data set, we applied the stratified HDMI methods to both final-age and entry-age variables for sensitive cases as described in Section 2. We also applied the TC method, with the top-code being 75 for final-age and 35 for entry-age (as the length of study is 40 years). We then calculated estimates of regression coefficients from the PH model, the corresponding empirical bias and RMSE of the estimates, average width of the

Table III. Simulation study scenario I: inferences for regression coefficients in the PH model.

	Entry-age (40–50)				Gender (female)			
Method	Empirical bias ($*10^4$)	RMSE ($*10^4$)	Rel-wid	Cover (per cent)	Empirical Bias ($*10^4$)	RMSE ($*10^4$)	Rel-wid	Cover (per cent)
BD	38	570	1	95.2	−38	582	1	92.6
TC	11501	11513	0.94	0	486	746	0.99	84.8
HD1	8	574	1.01	94.6	183	623	1.01	93
HD2	7	569	1.01	95.2	276	645	1.01	91.2
HD3	36	573	1.01	94.8	−17	585	1	93.6
HDU	7	581	1.03	94.2	325	648	1.01	91

Table IV. Simulation study scenario II: inferences of regression coefficients from the PH model.

	Entry-age (40–50)				Gender (female)			
Method	Empirical bias ($*10^4$)	RMSE ($*10^4$)	Rel-wid	Cover (per cent)	Empirical Bias ($*10^4$)	RMSE ($*10^4$)	Rel-wid	Cover (per cent)
BD	36	583	1	93.6	−15	580	1	93.6
TC	11463	11475	0.94	0	486	737	0.99	83.8
HD1	6	578	1.01	93.8	204	609	1.01	93.2
HD2	13	582	1.01	93.4	560	884	1.01	78.6
HD3	30	581	1.01	93.6	−7	577	1.01	94.2
HDU	96	599	1.03	93.6	225	588	1.02	94.2

*Here 'RMSE' refers to root mean squared error. 'Rel-wid' refers to 'relative width', which is fraction of 95 per cent CI width comparing with estimate 1. 'Cover' refers to the 95 per cent CI coverage.

95 per cent confidence intervals (CIs) based on a normal approximation relative to the CI from the data BD, and the confidence coverage of these intervals.

3.2. Results

The simulation results are based on 500 data sets of sample size 2000. We set the number of bootstraps B to be 100 for calculating standard errors of BD and TC estimates; and create $D=5$ imputed data sets. For stratified HDMI methods, we create strata with stratum size around 25.

Table III presents the results from scenario I, where entry-age and gender are independent and their log hazards are additive. TC yields estimate of regression coefficients with serious empirical bias and high RMSE, and zero confidence coverage for the entry-age variable. The TC estimate of the gender coefficient is less biased, but it still has sizeable empirical bias, and the CI has below nominal coverage. All stratified HDMI methods produce quite satisfactory results for the coefficient of entry-age, with negligible empirical bias and close to nominal confidence coverage. The unstratified method, HDU, also works well in terms of empirical bias and coverage, but it is somewhat less efficient than the stratified HD methods. HD3 works best for the gender coefficient, yielding an estimate with minimal empirical bias and good confidence coverage. Estimates from the other HD methods are also acceptable, although they have slightly higher empirical bias and below nominal confidence coverage. When males and females have different entry-age distributions as in scenario II (Table IV), most methods perform as in the first scenario, except that HD2 yields larger empirical bias, RMSE and less coverage for estimate of the regression coefficient of gender. In fact, it has even worse results than the TC method.

Table V displays the results from scenario III, where there is interaction between the age and gender variables. TC yields estimates with considerable empirical bias and poor coverage for regression coefficients of age, gender, and the interaction between these two variables. Among stratified HD methods, HD3 has the best performance and yields estimates with good inferences for both variables and the age–gender interaction. Estimates from HD1 and HD2 methods have satisfactory results for all three terms, although they have more empirical bias than HD3. Estimates from HDU have larger empirical bias and smaller confidence coverage than the stratified HD methods.

In summary, HD3 performs best under all circumstances. Other stratified HD methods yield estimates of regression coefficients with good inferential properties for the entry-age variable. These methods also provide satisfactory results for gender, except for HD2 in scenario II. With the presence of interaction between age and gender, estimates for the interaction term from HD1 and HD2 methods do not have sufficient coverage. HDU tends to be slightly less efficient than the stratified HD methods, but it works surprisingly well in the first two scenarios, indicating that stratification may

Table V. Simulation study scenario III: inferences of regression coefficients from the PH model.

	Entry-age (40–50)				Gender (female)				Interaction			
Method	Empirical bias ($*10^4$)	RMSE ($*10^4$)	Rel-width	Cover (per cent)	Empirical Bias ($*10^4$)	$RMSE$ ($*10^4$)	Rel-width	Cover (per cent)	Empirical Bias ($*10^4$)	RMSE ($*10^4$)	Rel-width	Cover (per cent)
BD	28	781	1	94.2	−39	810	1	94.4	13	1094	1	95
TC	10383	10411	0.95	0	−710	1129	1.07	84.6	2423	2646	0.97	38.6
HD1	−217	836	1.01	92.8	−128	839	1.01	93	501	1277	1.01	90.2
HD2	−241	823	1.01	94	−123	850	1.01	92.8	550	1298	1.01	89.4
HD3	−20	760	1.01	96.4	−67	798	1	94.6	104	1070	1.01	95.4
HDU	−706	985	1.04	88.8	−437	854	1.01	91	1452	1646	1.03	81.4

not be necessary in these settings. For the more complicated situation (scenario III), it yields biased estimates with low confidence coverage.

4. Application to the Charleston Heart Study data

We chose a subset of the CHS data and studied the relationship between hazard rate and certain risk factors. As an intact data file prior to disclosure control was available to us, the effectiveness of our SDC methods can be readily assessed.

4.1. Primary data analysis

After deletion of missing values and recoding on some variables, our sample included 1344 individuals, of which 303 survived the study. The variables involved were entry-age, final-age, censoring indicator, race/gender, education level, current cigarette smoking status, history of myocardial infarction (MI), history of diabetes, history of hypertension, electro-cardiographic interpretation (EKG), living place between age 20 and 65 and body mass index (BMI). For the PH regression model, final-age instead of survival time was treated as the time-scale variable.

To examine the effects of our chosen risk factors, we applied the PH model to the data set prior to SDC. Table VI displays the results from the regression. All the factors have a significant effect on participant's hazard ratio except BMI and entry-age (overall). Comparing to individuals that enter the study between 35 and 40 years old, those with entry-age greater than 50 have about a 30 per cent increase in risk of death. White females tend to have 34 per cent less risk than white males. Achieving education after high school reduces hazard by 30 per cent comparing to non-high school education. Smoking cigarettes increases death risk by 76 per cent. Participants with a definite history of MI have twice the risk of death as those without a history. History of diabetes as well as EKG problems increases the hazard by over 50 per cent, whereas history of hypertension increases risk of death by 17 per cent. Rural residents have 25 per cent less hazard than urban residents. Most of these coefficients are in the expected direction.

4.2. Results from SDC methods

As described earlier, variables subject to disclosure limitation are entry-age and final-age variables. Respondents with final-age greater than or equal to 80 years are considered to be sensitive cases, which leads to top-code values of 40 for entry-age and 80 for final-age. For this data set, TC the age variables has great impact on the analysis, since the entry-age variable is recoded into only two categories (40 or below 40), in contrast to the five categories for entry-age in the original data. We applied HDMI methods with $D=5$ imputed data sets to the data and computed estimates of regression coefficient from a PH model.

Table VII shows the results from original, top-coded, and imputed data sets based on 500 replications. Figure 1 summarizes these results with box plots of the percentage deviations of the TC and HD estimates of the regression coefficients from the BD estimates. Estimates of coefficients of entry-age variable have not been plotted as the TC method cannot differentiate between the age categories. Predictably, TC considerably alters the relationship between hazard and covariates and yields estimates of the regression coefficients with serious bias, especially for the entry-age variable. The unstratified HD method, HDU, yields better estimates than TC for some covariates, but one coefficient is seriously underestimated. The stratified methods all do considerably better, yielding box plots with narrower inter-quartile ranges and less extreme outliers than TC and HDU. There is not much to choose between the stratified HD methods—HD2 and HD3 yield better estimates of the entry-age coefficients than HD1, but HD1 provides better estimates of the regression

Table VI. Estimates of regression coefficients from PH model, original CHS data.

	Parameter estimate ($*10^4$)	Standard error ($*10^4$)	Pr $> \chi^2$	Hazard ratio
Entry-age 1 (40–44)	1977	1128	0.08	1.22
Entry-age 2 (45–49)	1814	1151	0.1	1.2
Entry-age 3 (50–59)	2786	1072	0.009	1.32
Entry-age 4 (60+)	2878	1242	0.02	1.33
Race/Gender 2 (white woman)	−4171	955	<0.0001	0.66
Race/Gender 3 (black man)	−241	949	0.8	0.98
Race/Gender 4 (black woman)	−1870	1031	0.07	0.83
Education 1 (some high school)	−1100	832	0.2	0.9
Education 2 (after high school)	−3761	1000	0.0002	0.69
Current cigarette smoking 1 (yes)	5677	701	<0.0001	1.76
History of MI 1 (possible)	3741	3416	0.3	1.45
History of MI 2 (definite)	6949	1889	0.0002	2
History of diabetes 1 (yes)	4330	1602	0.007	1.54
History of hypertension 1 (yes)	1547	750	0.04	1.17
EKG 1 (with problem)	4644	947	<0.0001	1.59
Living place 20–65 2 (rural)	−2947	1028	0.004	0.75
Living place 20–65 3 (mix of rural and urban)	−1361	1467	0.4	0.87
BMI	28	74	0.7	1

coefficient for gender than HD2 and HD3. Overall, the stratified HD methods all work better than TC in preserving the relationship between risk of death and the covariates on this data set.

5. Discussion

Longitudinal data raise particular confidential concerns with potentially extensive longitudinal information gathered over time. We consider a specific application concerning disclosure risk caused by some participants attaining high ages because of prolonged participation in a longitudinal study, as in the CHS. One of the authors (McNally) has the responsibility to prepare a public use version of this data set through NACDA that meets HIPAA regulations. As discussed earlier, the standard approach of TC age has severe limitations in this longitudinal setting, especially for survival analyses with age being a key variable of interest. HIPPA restrictions make a full public release impossible and require a formal Limit Use Agreement which imposed significant barriers to accessing the data. We develop MI-based SDC methods for this particular data setting. Similar to the methods in An and Little (2007, unpublished), our proposed MI methods are based on stratification, with strata defined by the predicted values of the age variables from a regression model.

Regarding the longitudinal nature of the data set in this study, we have focused on inference about regression coefficients from Cox's PH model for survival. As expected, the TC method yields seriously biased estimates, especially for the entry-age variable. In principle, it is possible to improve the statistical performance of the TC method by treating top-coded values as censored, but this yields a non-standard problem of survival analysis with censored covariates, and does not address the problem of severe loss of information from TC in this setting.

Among our stratified HDMI methods, HD3 has the best performance and yields results close to those before deletion in simulation studies. The other stratified methods also work well overall, except that sometimes they do not quite attain

Table VII. Estimates of regression coefficients from the PH model, CHS data after SDC.

	Estimate (SE) ($*10^4$)					
	BD	TC	HD1	HD2	HD3	HDU
Entry-age 1 (40–44)	1992 (1154)	Entry-age 1 (<40)	2597 (1164)	1962 (1157)	1977 (1152)	1801 (1173)
Entry-age 2 (45–49)	1815 (1153)	−792 (975)	1817 (1181)	1872 (1180)	1999 (1178)	2269 (1187)
Entry-age 3 (50–59)	2711 (1056)		1640 (1097)	2371 (1098)	2706 (1090)	2638 (1095)
Entry-age 4 (60+)	2799 (1254)		2393 (1240)	2922 (1262)	3099 (1262)	3716 (1230)
Race/Gender 2 (white woman)	−4200 (913)	−3813 (1189)	−4724 (1002)	−3798 (979)	−3971 (965)	−2177 (960)
Race/Gender 3 (black man)	−205 (1004)	982 (1142)	−248 (966)	54 (975)	16 (963)	845 (971)
Race/Gender 4 (black woman)	−1876 (1073)	−1734 (1346)	−1984 (1036)	−1771 (1054)	−1660 (1054)	−1267 (1043)
Education 1 (some high school)	−1127 (829)	−1347 (1029)	−996 (841)	−1224 (843)	−1257 (846)	−924 (847)
Education 2 (after high school)	−3806 (963)	−4958 (1257)	−3559 (1024)	−3721 (1027)	−3793 (1013)	−3290 (1025)
Current cigarette smoking 1 (Yes)	5785 (718)	7328 (891)	5763 (714)	5874 (724)	5596 (711)	4875 (706)
History of MI 1 (possible)	4211 (4548)	5360 (6113)	3397 (3515)	2467 (3516)	2946 (3599)	3863 (3552)
History of MI 2 (definite)	7080 (1936)	5622 (2766)	4678 (1980)	5029 (1979)	5280 (1954)	4716 (2017)
History of diabetes 1 (yes)	4616 (2158)	6234 (2189)	4013 (1681)	3695 (1676)	4414 (1674)	4744 (1677)
History of hypertension 1 (yes)	1637 (840)	2581 (977)	2006 (775)	1877 (778)	1678 (777)	1823 (778)
EKG 1 (with problem)	4754 (1091)	4717 (1197)	4129 (982)	3936 (982)	3754 (974)	3327 (992)
Living place 20–65 2 (rural)	−3042 (1029)	−3719 (1299)	−3297 (1054)	−3189 (1058)	−3162 (1047)	−2522 (1039)
Living place 20–65 3 (mix of rural and urban)	−1296 (1887)	−594 (1969)	−1375 (1545)	−1239 (1500)	−559 (1480)	−410 (1519)
BMI	28 (81)	20 (98)	57 (76)	61 (76)	13 (75)	10 (76)

Note: The BD and TC estimates are bootstrap estimates based on 100 bootstraps.

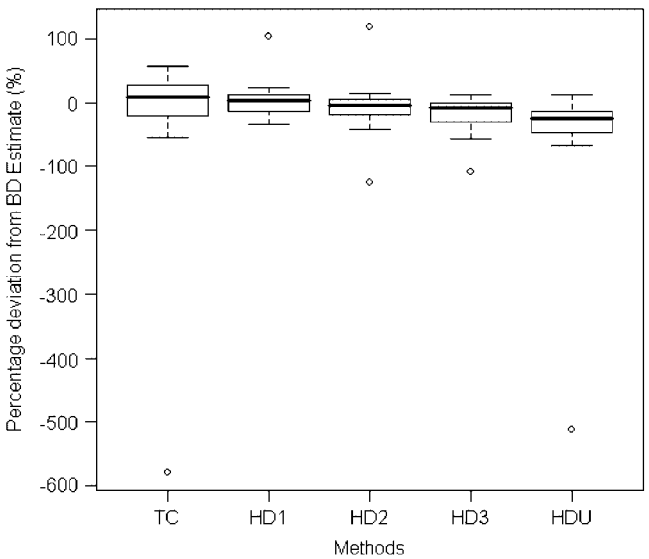

Figure 1. Percentage deviation from BD estimates for TC and HD estimates of the regression coefficients from the PH model, CHS data after SDC.

the nominal confidence coverage. When there are fewer censored cases, as with the CHS data (number of censored cases is one-fourth the total sample size), HD3 does not have the obvious advantage over other methods, although it still yields satisfactory results. The unstratified method, HDU, works almost as well as stratified HD methods in simple data settings. In situations with more covariates and a larger number of sensitive cases, it yields biased estimates with below-nominal confidence coverage.

An and Little [2] present two versions of MI methods, the 'C' method, which is based on a model fitted to the complete data; and the 'D' method based on a model fitted to the deleted values alone. The 'D' method is somewhat less efficient than the 'C' method, but it is more robust to model misspecification, since the model is fitted to the data that are being deleted. As mentioned in Section 2, we present the results for the 'D' method here, since the 'C' method was inferior in simulations.

Note that in this study, the predicted logarithm of the hazard is considered an appropriate factor for stratification, as the primary focus is survival analysis, and preserving the original relationship between the hazard and covariates. The current size of strata is selected based on the empirical experience, by trying to maintain a good balance between limiting disclosure risk, and best retaining the utility of the data. Therefore, it is not a universal recommendation. The statistical agencies/data producers are encouraged to make reasonable choice of the stratification factor and the size of strata, based on the interest of the specific data.

Our stratified HDMI methods produce excellent inferences, but they arguably have the limitation as SDC methods that original values in the data set are retained, although not attached to the right records. As multiply imputed data sets protect an individual with extremely high-age value from being linked to a specific record, a potential data snooper may still recognize the fact that this individual is included in the data set, especially for data with geographic specificity. To address this concern, we will develop parametric MI methods in our future work.

Reiter [9] proposes the use of classification and regression trees (CART) to generate partially synthetic data. For the CART approach, subpopulations with relatively homogeneous outcome (imputation classes) are created by partitioning the predictor space. The imputation model is fit on the cases with sensitive values only, and sensitive values of the outcome can be replaced by random draws from the same class according to the predictor values by Bayesian bootstrap. This is in spirit quite similar to the stratified HD methods in this paper, where we create strata (imputation classes) based on some predicted values from a regression model fitted to the cases with sensitive values, and replace sensitive values with random draws from a set of sensitive values in the same stratum. Both methods are non-parametric, and have been shown to have good repeated sampling properties. Reiter also suggested an alternative method of drawing samples from a kernel density estimator based on the random draws from the first step, which yields added protection since it avoids releasing real data values.

An important issue that is not addressed in this article is quantifying the reduction in disclosure risk from multiple imputations of the ages of high-age individuals, compared with alternatives such as TC. This is a complex question that depends on the set of 'key' variables available to the intruder from external databases that include the target individuals, the probability that target individuals are in the sample, and the joint distribution of the key variables for the high-age individuals. Reiter and Mitra [15] and Drechsler and Reiter [16] describe approaches for addressing this issue with partially synthesized data, and the future research should address how these methods translate into the longitudinal data setting.

We have confined attention here to imputing age-related variables for individuals with high-age values, and SDC methods for other types of variables (such as geography) in longitudinal health data like the CHS data remain a topic for future research.

Acknowledgements

This work was supported by the National Institute of Child and Human Development grant (P01 HD045753). The Charleston Heart Study is supported by the National Institute of Aging grants (P30AG004590 and R03AG021162). The authors thank Trivellore Raghunathan, Michael Elliott, and Myron Gutmann, for useful comments.

References

1. Dempster AP, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**:1–37.
2. An D, Little RJ. Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A* 2007; **170**:923–940.
3. Reiter JP. Inference for partially synthetic, public use microdata sets. *Survey Methodology* 2003; **29**:181–188.
4. Little RJ. Statistical analysis of masked data. *Journal of Official Statistics* 1993; **9**:407–426.
5. Rubin DB. Discussion: statistical disclosure limitation. *Journal of Official Statistics* 1993; **9**:461–468.
6. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 2003; **19**:1–16.

7. Little RJ, Liu F, Raghunathan T. Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Gelman A, Meng X-L (eds). Wiley: New York, 2004; 141–152.
8. Reiter JP. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* 2002; **18**:531–544.
9. Reiter JP. Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society*, *Series A* 2005; **168**:185–205.
10. Reiter JP. Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* 2005; **131**(2):365–377.
11. U.S. Department of Health and Human Services. The Health Insurance Portability and Accountability Act (HIPAA) of 1996.
12. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information (the Privacy Rule).
13. Nietert PJ, Sutherland SE, Bachman DL, Keil JE, Gazes P, Boyle E. Charleston Heart Study (Computer file). CPSR version. Medical University of South Carolina (producer): Charleston, SC, 2000. Inter-university Consortium for Political and Social Research (distributor): Ann Arbor, MI, 2004.
14. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
15. Reiter JP, Mitra R. Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* 2009; **1**(1):99–110.
16. Drechsler J, Reiter JP. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, Domingo-Ferrer J, Saygin Y (eds). Springer-Verlag: New York, 2008; 227–238.