Multiple Imputation Methods for Disclosure Limitation in Longitudinal Data

Di An[1], Roderick J.A. Little[2], James W. McNally[3]

1 Merck Research Laboratories, Merck & Co., Inc., P.O. Box 1000, Upper Gwynedd, PA 19454

2 University of Michigan, 1420 Washington Heights M4045, Ann Arbor, Michigan 48109-2029

3 University of Michigan, 330 Packard Street, Ann Arbor, Michigan 48109

Abstract

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 over time. We consider problems created by high ages in cohort studies. Because of the risk of disclosure, ages of very old respondents can often not be released, as stipulated by the Health Insurance Portability and Accountability Act (HIPAA). Top-coding of individuals beyond a certain age is a standard way of dealing with this issue, but it has severe limitations in longitudinal studies. 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, and is shown to work well in preserving the relationship between hazard and covariates.

Key Words: confidentiality, disclosure protection, longitudinal data, multiple imputation, survival analysis

1. Introduction

Statistical disclosure control is a class of procedures that 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 dataset 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 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, respondents with very high income have reduced risk of disclosure.

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 one category – this is sensible, but precludes analyses that treat the variable as continuous. Another approach is to ignore the fact of top-coding 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 (Dempster, Laird and Rubin, 1977). 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 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 dataset 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 (2007a) propose an alternative to top-coding based on multiple imputation (MI), which allows valid inferences to be created based on applying multiple imputation combining rules described by Reiter (2003), while preserving the SDC benefits of top-coding; for other discussions of MI in the disclosure control setting, see Little (1993); Rubin (1993); Little, Liu and

Raghunathan (2004); Reiter (2005a, 2005b). The methods in An and Little (2007a) are extended to handle covariate information in An and Little (2007b).

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 for the release of health data for individuals. Top-coding 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 top-coding 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 (Nietert *et al.*, 2000), 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 top-coding 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 suffice to 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 procedure, where we create strata and draw deleted ages with replacement from each stratum. Our method concerns MI 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 datasets, calculate estimates of regression coefficients for putative risk factors, and compare these estimates, and 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 rest of this paper is organized as follows. Section 2 presents our SDC approaches for longitudinal data and describes 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 CHS data. Section 5 gives discussion and future work.

2. Methods

2.1 SDC methods for longitudinal data
An and Little (2007a) propose 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 top-coding.

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 time of survival. Individuals with $S \geq L$ are treated as censored ($C = 1$), and otherwise 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 approaches to SDC.

(a) Top-coding. 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 dataset is referred to as "top-coded" data.

(b) Hot-deck MI (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 (2007b), where we assign the deleted data into strata based on 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 predicted values of the logarithm of hazard computed from the proportional hazard model.
(ii) HD2. Strata are defined by predicted values of entry-age, from the regression of entry-age on other variables involved.
(iii) HD3. We develop a two-way stratification, where strata are defined by both predicted values of the logarithm of hazard, and predicted values of entry-age.
 (iv) HD4. Stratification depends on the value of C. For individuals that are censored, strata are defined by predicted values of entry-age; and for those not censored, strata are defined by both predicted values of the logarithm of hazard and predicted values of entry-age.
(v) HD5. We directly apply HDMI method without stratification, for comparison with the stratified methods.

Note that for methods HD1 – HD3, we delete values of Y_{end}, Y_{start}, and C of sensitive cases and jointly impute these values. HD4 retains values of C and imputes Y_{end} and Y_{start} only.

It is worth mentioning that for above stratified methods, we perform regression only on the deleted cases to obtain predicted values. We also consider an alternative way of stratification, where we perform regression on the complete data, and then stratify the sensitive cases for imputation. Results from these methods are briefly described in Section 3.

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 dataset 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) Top-coding (TC) – the estimates of regression coefficients calculated from top-coded dataset.

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

The five remaining methods HD1 – HD5 are as described in Section 2.1, yielding D MI datasets. The MI estimate is calculated as

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

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

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

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

$$B = \sum_{d=1}^{D} (\hat{\theta}^{(d)} - \hat{\theta}_{MI})^2 /(D-1)$$ is the between-imputation variance. The formula for variance 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, 2002, p86), for reasons discussed in Reiter (2003).

3. Simulation study

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

3.1 Study design

For simplicity we simulated survival data with just two binary covariates, representing gender (male and female) and entry age (say 30 - 40 and 40 - 50). Datasets were simulated from multinomial distribution in four categories defined by these variables. Values of entry-age were generated from uniform distribution. Survival times (in years) were generated from piece-wise exponential distributions. 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: Distributions of entry age do not depend on gender; both male and female have same entry-age distributions.
Scenario II: Distributions of entry age are different for males and females.
Scenario III: Distributions of entry age are the same for males and females, and there is interaction between entry age and gender.

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. For each simulated dataset, 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 top-coding method, with 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 variances of the estimates, as well as 95% confidence intervals (CI's) based on normal approximation, and the confidence coverage of these intervals.

3.2 Results

Simulation results are based on 500 datasets 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 datasets. For stratified HDMI methods, we create strata with stratum size around 25.

	Entry-age (40~50)				Gender (female)			
	Bias $(*10^4)$	RMSE $(*10^4)$	Rel-wid	Cover (%)	Bias $(*10^4)$	RMSE $(*10^4)$	Rel-wid	Cover (%)
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	25	571	1.01	95.4	257	622	1.01	91.8
HD3	7	569	1.01	95.2	276	645	1.01	91.2
HD4	36	573	1.01	94.8	-17	585	1	93.6
HD5	7	581	1.03	94.2	325	648	1.01	91

Table 1: Simulation study scenario I: inferences of regression coefficients from PH model

 "RMSE" refers to root mean squared error. "Rel-wid" refers to "relative width", which is fraction of 95 CI % width comparing to estimate 1. "Cover" refers to the 95% CI coverage.

Table 1 presents results from scenario I, where distributions of entry-age are the same for male and female. TC yields estimate of regression coefficient with serious bias and RMSE, and zero confidence coverage for the entry-age variable. As for gender, TC estimate has relatively better properties, yet it still has sizable bias and low coverage. All stratified HDMI methods produce quite satisfactory results for the entry-age variable, with negligible bias and confidence coverage close to before deletion. HD5 also work well in terms of bias and coverage, but it is somewhat less efficient than the stratified HD methods. HD4 method works best for gender variable, yielding estimate of regression coefficient with minimal bias and good confidence coverage. Estimates from other HD methods are also acceptable, though they are in general more biased and have less coverage. When male and female have different entry-age distributions as in scenario II (results not shown), most methods behave similarly as in the first scenario, except that HD3 yields larger bias, RMSE and less coverage for estimate of the regression coefficient of gender. In fact, it has even worse results than TC method.

	Entry-age (40~50)				Gender (female)				Interaction			
	Bias (*10^4)	RMSE (*10^4)	Rel-wid	Cover (%)	Bias (*10^4)	RMSE (*10^4)	Rel-wid	Cover (%)	Bias (*10^4)	RMSE (*10^4)	Rel-wid	Cover (%)
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	-244	803	1.02	94.8	-53	803	1	93.8	568	1166	1.01	93.6
HD3	-241	823	1.01	94	-123	850	1.01	92.8	550	1298	1.01	89.4
HD4	-20	760	1.01	96.4	-67	798	1	94.6	104	1070	1.01	95.4
HD5	-706	985	1.04	88.8	-437	854	1.01	91	1452	1646	1.03	81.4

Table 2: Simulation study scenario III: inferences of regression coefficients from PH model

"RMSE" refers to root mean squared error. "Rel-wid" refers to "relative width", which is fraction of 95 CI % width comparing to estimate 1. "Cover" refers to the 95% CI coverage.

Table 2 displays results from scenario III, where there is interaction between the age and gender variables. TC yields estimates with considerable bias and poor coverage for regression coefficients of age, gender and the interaction between these two variables. Among stratified HD methods, HD4 has the best performance and yields estimates with good inferences for both variables and the age-gender interaction. HD2 also has satisfactory results for all three terms, though it is more biased than HD4. Estimates from HD1 and HD3 methods have similar properties as from HD2, except that they have less sufficient coverage for the interaction term. Estimates from HD5 have larger bias and less confidence coverage than those from the stratified HD methods.

We also applied the alternative stratified method described in Section 2.1, where we obtained predicted values from regression on the complete data, and then stratified the sensitive cases for imputation. Estimates from these methods (not shown) are more biased and have less confidence coverage compared to the methods above. This suggests that when a regression model is fitted to the data that are being deleted, it makes the method more robust to model mis-specification and yield better result (see Section 5 for more discussion).

In summary, HD4 performs best under all circumstances. Other stratified HD methods yield estimates of regression coefficient with good inferential properties for the entry-age variable. These methods also provide satisfactory results for gender, except for HD3 in scenario II. With presence of interaction between age and gender, estimates for the interaction term from HD1 and HD3 methods do not have sufficient coverage. HD5 tends to be slightly less efficient than the stratified HD methods, but it works surprisingly well in the first two scenarios, indicating stratification may not be necessary in such data setting. For more complicated situation (scenario III), it yields biased estimates with low confidence coverage.

4. Application in Charleston Heart Study data

We chose a subset of the CHS data and studied the relationship between hazard rate and certain risk factors. Since 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 infraction (MI), history of diabetes, history of hypertension, electro-cardiographic interpretation (EKG), living place between age 20 to 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 effects of our chosen risk factors, we applied the PH model to the dataset prior to SDC. All factors have 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% increase in risk of death. White females tend to have 34% less of risk than white males. Achieving education after high school reduces hazard by

30% comparing to non-high school education. Smoking cigarette increases death risk by 76%. Participants with definite history of myocardial infraction 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%, while history of hypertension increases risk of death by 17%. Rural residents have 25 % less of 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 intuitively leads to top-code values of 40 for entry-age and 80 for final-age. For this dataset, top-coding 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 to the data and computed estimates of regression coefficient from a PH model. Results are based on 500 replications (not shown). 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. Of the stratified HDMI methods, HD3 and HD4 yield estimates of coefficients of entry-age close to those from BD. HD1 provides better estimates of regression coefficients than other methods for the gender variable. For the rest of covariates, none of the stratified HD methods seems to have an obvious advantage, with HD2 being slightly inferior. HD5 has less satisfactory results, though it still yields better estimates than TC for some covariates. Overall, the stratified HD methods all work better than top-coding in preserving the relationship between risk of death and the covariates on this dataset.

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 Charleston Heart Study. 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 top-coding age has severe limitations in this longitudinal setting, especially for survival analyses with age being a key variable of interest. HIPPA restrictions make full public releases 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 (2007b), 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 dataset in this study, we have focused on inference about regression coefficients from Cox's proportional hazard model. As expected, top-coding method yields seriously biased estimate especially for the entry-age variable. Among our stratified HDMI methods, HD4 has the best performance and yields results close to before deletion in simulation studies. The other stratified methods also work well overall, except that sometimes they do not quite attain 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), HD4 does not have obvious advantage over other methods, though it still yields satisfactory results. The no-stratification method HD5 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 low confidence coverage.

An and Little (2007a) 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.

Similarly, we develop two alternatives in this study. The first method calculates predicted values from regression on the deleted data; and the second one utilizes the complete data for regression. Results show the first method yield estimates with better inferential properties. This finding supports the justification in An and Little (2007a), as regression on deleted data tends to be more robust to model mis-specification.

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

Moreover, we have confined attention to individuals with high age values. The whole field of SDC methods raised by other variables (e.g. geographic) in longitudinal health data like the CHS data remains rather unexplored. We also plan to consider other possible confidential concerns and develop suitable SDC methods for these problems.

Acknowledgements

References

An, D. and Little, R.J. (2007a). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170, pp. 923-940.

An, D. and Little, R.J. (2007b). Extensions of multiple imputation methods as disclosure control procedure for multivariate data. In preparation.

Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, *Series B*, 39, pp. 1-37.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 2, pp. 383-406.

Little, R.J.A. and Rubin, DB (2002). *Statistical Analysis with Missing Data.* Wiley: New York.

Little, R.J., Liu, F. and Raghunathan, T. (2004). Statistical Disclosure Techniques Based on Multiple Imputation. In "*Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*", A. Gelman and X.-L. Meng, eds., pp. 141-152. Wiley: New York.

Nietert P.J., Sutherland S.E., Bachman D.L., Keil J.E., Gazes P., and Boyle E. (2000). CHARLESTON HEART STUDY [Computer file]. ICPSR version. Charleston, SC: Medical University of South Carolina [producer], 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

R Project (2007). The R project for statistical computing. See http://www.r-project.org/.

Raghunathan, T.E., Reiter J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, pp. 1-16.

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics,* 18, pp. 531-544.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology,* 29, pp. 181-188.

Reiter, J.P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, pp. 185 - 205.

Reiter, J.P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference,* 131 (2), pp. 365 - 377.

U.S. Department of Health and Human Services. The Health Insurance Portability and Accountability Act (HIPAA) of 1996.

U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information (the Privacy Rule).