Topics and Applications in Synthetic Data

Professor Donald B. Rubin **Bronwyn Wing Ling Loong**

Topics and Applications in Synthetic Data

Abstract

Releasing synthetic data in place of observed values is a method of statistical disclosure control for the public dissemination of survey data collected by national statistical agencies. The overall goal is to limit the risk of disclosure of survey respondents' identities or sensitive attributes, but simultaneously retain enough detail in the synthetic data to preserve the inferential conclusions drawn on the target population, in potential future legitimate statistical analyses. This thesis presents three new research contributions in the analysis and application of synthetic data. Firstly, to understand differences in types of input between the imputer, typically an agency, and the analyst, we present a definition of congeniality in the context of multiple imputation for synthetic data. Our definition is motivated by common examples of uncongeniality, specifically ignorance of the original survey design in analysis of fully synthetic data, and situations when the imputation model and analysis procedure condition upon different sets of records. We conclude that our definition provides a framework to assist the imputer to identify the source of a discrepancy between observed and synthetic data analytic results. Motivated by our definition, we derive an alternative approach to synthetic data inference, to recover the observed data set sampling distribution of sufficient statistics given the synthetic data. Secondly, we address the problem of negative method-of-moments variance estimates given fully synthetic data, which may be produced with the current inferential methods. We ap-

ply the adjustment for density maximization (ADM) method to variance estimation, and demonstrate using ADM as an alternative approach to produce positive variance estimates. Thirdly, we present a new application of synthetic data techniques to confidentialize survey data from a large-scale healthcare study. To date, application of synthetic data techniques to healthcare survey data is rare. We discuss identification of variables for synthesis, specification of imputation models, and working measures of disclosure risk assessment. Following comparison of observed and synthetic data analytic results based on published studies, we conclude that use of synthetic data for our healthcare survey is best suited for exploratory data analytic purposes.

Contents

Acknowledgments

Thank you to my advisor, Professor Donald Rubin. It has been a great learning opportunity to work with such a brilliant mind, and the support and freedom I have received to develop my own research interests and self-confidence has been invaluable.

Thank you also to Professor Carl Morris for guidance on Chapters 4 and 5, and constant encouragement to always question existing methods, and start with simple examples. My interest in synthetic data started in a survey design class taught by Professor Alan Zaslavsky. Again, it has been a privilege to learn from a very supportive and expert statistician.

I am indebted to Professor David Harrington for the opportunity to work on the CanCORS project in Chapter 6. The *real* data experience was a *real* turning point in my PhD.

Thank you to Jess, for our rock solid friendship, endless support and advice and good fun. Thank you to Ellen and her Mum, and Sachi, who brought me back from really difficult times, and to all the wonderful people I have met during my Harvard experience.

Lastly, thank you to my wonderful family for their never ending love, regardless of where I am in the world.

Chapter 1

Introduction

There is increasing demand from external researchers for access to individual record data (microdata) collected by national statistical agencies. In turn, the statistical agencies face a dilemma in the dissemination of microdata. On one hand, the privacy of survey respondents, and confidentiality of data collected[1] must be protected on legal and ethical grounds. On the other hand, there is a need to release enough detail in the microdata to preserve and maintain the validity of inference on the target population (as if given access to the original data set), from any potential statistical analysis, from any potential external analyst. To satisfy these dual objectives, one category of methods is to *restrict access* to the data to authorized individuals for approved analyses. A second category of methods is to *alter* the data before release, typically carried out by a *statistical disclosure control* (SDC) technique. Releasing

[1]In this thesis we adopt the definitions of privacy and confidentiality as used in Gates (2011), p. 3. 'Information privacy' is defined as the individual's desire (claim) to control the terms under which information about him/her is acquired, used or disclosed. 'Confidentiality' is closely related to privacy and refers to the agreement reached with the individual/business, when the information was collected, about who can see the identifiable information. Changes to this agreement can be made only with the explicit consent of the individual.

synthetic data is one such SDC method whereby observed data set values are replaced by synthetic data values generated to be representative of the same target population as the observed data set.

Traditional disclosure techniques (also referred to as masking techniques) include, but are not limited to, rounding, swapping or deleting values, and adding random noise (Little 1993). These methods are easy to implement and are widely used. However, a major drawback of traditional SDC methods is the potential distortion of the relationships among variables, such that results from standard likelihood methods are compromised. For example, treating rounded data as exact values will lead to an understatement of the posterior variance of the parameters (Sheppard 1898, Dempster and Rubin 1983). To analyze the masked data properly, users should apply the likelihood-based methods detailed in Little (1993), or the measurement error models described by Fuller (1993). These methods may be difficult to apply, especially for non-standard estimands, and may require analysts to learn new statistical methods and use specialized software packages.

Using synthetic data to replace observed values before public release was first proposed by Rubin (1993) based on the theory of multiple imputation (Rubin 1987). Synthetic data sets are created using samples drawn from the posterior predictive distribution of target population responses given the observed data set. Using an acceptable imputation model that captures correctly relationships among survey variables, and estimation methods based on the concepts of multiple imputation, analysts can make valid inferences on the target population of interest using standard likelihood methods, without accessing the original microdata. If all observed values are replaced

and no true values are released, this is known in the literature as fully synthetic data. A partially synthetic data set consists of a mix of multiply imputed and true values.

Some basic inferential methods for fully synthetic data were derived in Raghunathan, Reiter and Rubin (2003). Simulated and empirical data examples of fully synthetic data can be found in Reiter (2002), Raghunathan, Reiter and Rubin (2003) and Reiter (2005a). Since then, the basic fully synthetic data framework has been adapted to meet other disclosure control criteria. Some key developments requiring new inferential methods include inference for partially synthetic data (Reiter 2003), releasing multiply imputed synthetic data in two stages, which enables agencies to release different numbers of imputations for different variables (Reiter and Drechsler 2010), and sampling with synthesis, which combines the disclosure control benefits of partially synthetic data and random sampling, so that intruders are no longer guaranteed that their targets are in the released data (Drechsler and Reiter 2010). Other methodological developments in the synthetic data literature have focused on non-parametric approaches to synthetic data imputation, to reduce the reliance on the imputation model, in particular, using classification and regression trees (CART) (Reiter 2005b), and random forests (Caiola and Reiter 2010). There is also continuing research on measures for the assessment of disclosure risk and data utility of synthetic data. Statistical modeling approaches for assessment of identification risk have been proposed by a number of authors (e.g., Paass 1988, Duncan and Lambert 1989, and Fuller 1993) with extensions in subsequent papers (e.g., Fienberg et al. 1997, Reiter 2005c, Skinner and Shlomo 2008, and Reiter and Mitra 2009). Data utility measures attempt to characterize the quality of what can be learned about the target popu-

lation using the synthetic data, to what can be learned using the observed data set. Such comparisons can be tailored to specific estimands (Karr et al. 2006), or can be broadened to global differences in distributions (Woo et al. 2009).

Several US statistical agencies have been or are active in releasing or developing partially synthetic public data sets. Among these are the Survey of Consumer Finances (Kennickell 1997), the Longitudinal Employer Household Dynamics program (Abowd and Woodcock 2004), the Survey of Income and Program Participation (Abowd, Stinson and Benedetto 2006), the Longitudinal Business Database (Kinney and Reiter 2007), and the American Community Survey group quarters data (Hawala 2008). The German Institute for Employment Research (IAB) has also done extensive research to release a synthetic version of the IAB Establishment Panel, a business database on German firms' personnel structure, development and policy (Drechsler et al. 2008).

This thesis presents three new contributions to research in using synthetic data for statistical disclosure control.

1.1 Uncongeniality for synthetic data sets and recovery of the observed data set sampling distribution of sufficient statistics

In the context of multiple imputation for missing data, Meng (1994) coined the term 'uncongeniality' of the *analysis procedure* to the *imputation model*.

"Uncongeniality essentially means that the analysis procedure does not correspond to the imputation model. The uncongeniality arises when the analyst and the imputer have access to different amounts of information and have different assessments." (Meng 1994, p. 539).

Imputation input includes model assumptions, purpose of imputation, available information and data from the collection phase, as well as any other potentially useful resources (e.g., past similar surveys). Analysis input consists of the analyst's purpose of investigation, data made available for analysis, information on the imputation models if available, computational skills and so on.

During the imputation, resampling, and analysis of synthetic data sets, the potential for uncongeniality always exists because the imputer and analyst are separate bodies. Some examples of discrepancies between the types of information and assumptions used by imputers and analysts in a synthetic data setting include:

- Ignoring the original survey design when resampling and analyzing the synthetic data sets. For example, the original survey design may be stratified by income band, but resampling and analysis are carried out by simple random sample

methods ignoring the stratification. This may be the case if survey design variables are completely confidential and cannot be released to analysts.

- Deriving the imputation model from the entire observed data set when the analysis procedure utilizes a subset of records. The general imputation model may not capture some of the subset relationships of importance to analysts. For example, suppose an original sample of $n = 500,000$ units is drawn from the population of US adult males. The imputation models are conditional on the entire observed data set of $n = 500,000$ units, but the analyst only wants to study the 155,000 observed units sampled from the US southern states.

- Not using variables or structures of variables of interest to analysts in the imputation models. For example, the imputation model is conditional on income bands of size \$100,000. However, the analyst is interested to use income bands of size \$25,000.

Uncongeniality may lead to large discrepancies between inferential results from the observed data set and synthetic data. We cannot directly apply the congeniality definition in Meng (1994) to characterize any discrepancies, because the observed data set is not available to analysts in a synthetic data setting. Chapter 3 presents a definition of congeniality for multiple imputation of synthetic data. Uncongeniality is a very challenging theoretical topic, and so limited analytic results are available to justify the definition. Instead, we emphasize the practical use of the congeniality definition to explain discrepancies between analytic results from the observed data set and synthetic data, in two case studies used as motivating examples.

The discussion of uncongeniality is motivation for the development of new inferential methods to recover observed-data sufficient statistics given the synthetic data. This is the focus of Chapter 4 where we present alternative analysis equations and assess their performance. If the analyst cannot have access to the observed data set, the next best thing is to try to infer the observed-data sufficient statistics.

1.2 Application of adjustment for density maximization to sampling variance estimation in fully synthetic data inference

The second contribution provides an alternative approach to variance estimation for fully synthetic data to improve on a deficiency in the current inferential methods. A disadvantage of the method-of-moments variance estimate derived in Raghunathan, Reiter and Rubin (2003) is that it may be negative. Reiter (2002) and Reiter and Drechsler (2010) use slightly modified, more conservative estimates when negative variance estimates are calculated. Negative variance estimates may also be generally avoided by choosing a large synthetic sample size, or large number of imputations (Reiter 2002). The adjustment for density maximization (ADM) procedure, as first proposed by Morris (1988), provides an alternative approach to the variance estimation problem. ADM was originally proposed to produce shrinkage factor estimates in normal hierarchical models that lie inside the boundaries of the interval [0,1], which is not guaranteed by standard maximum likelihood methods. This procedure implies positive variance estimates. In Chapter 5 of this thesis, we investigate by simulation

and empirical data study, using ADM to produce positive variance estimates with fully synthetic data.

1.3 Partially synthetic data for a large-scale health-care study

In chapter 6 of this thesis, we demonstrate creating partially synthetic data for the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium. CanCORS is a multisite, multimode, multiwave study of the quality and patterns of care delivered to population-based and healthcare-system-based cohorts of newly diagnosed patients with lung and colorectal cancer. The Consortium is committed to sharing the data gathered to the widest possible audience to facilitate research by external analysts in healthcare, without comprising the confidentiality of respondents' identities and/or sensitive attributes. Compared to previous applications using socioeconomic data, the structure and potential analytic use of healthcare data are different. Hence, this applied contribution presents new challenges and innovations in using synthetic data for statistical disclosure control, for example imputing variable relationships that are clinically feasible. We review the methods for the creation and analysis of partially synthetic data, and some current measures for data utility and disclosure risk assessment. We discuss our approach to identification of variables to synthesize, and specification of the imputation models, followed by quantification of the data utility and disclosure risk of the synthetic data we generate.

1.4 Data set used in practical illustrations

To illustrate the new methods presented in this thesis, we synthesize data from the Joint Canada/United States Survey of Health (JCUSH), jointly produced by Statistics Canada and the United States National Center for Health Statistics. This data set contains data collected from November 4, 2002 to March 31, 2003. The public-use file was released in June 2004 after application of traditional statistical disclosure controls, including variable grouping and capping, and data suppression. Survey variables collected cover chronic health conditions, functional status, determinants of health, and healthcare utilization. The principal objectives of the study were to foster collaboration between the two national statistical offices, and to produce a single data source for comparability studies between the two health systems. The sample size consisted of $\approx 3,500$ respondents in Canada, and $\approx 5,000$ respondents in the United States. The sample was stratified by province in Canada (five strata), and by four geographic regions in the United States (northeast, midwest, west and south). The survey was administered by telephone using the random digit dialing sample selection method. The primary sampling unit was one adult aged 18 years or older per household from persons living in private occupied dwellings. The primary reasons for choosing this data set for empirical study of the new analysis methods in Chapters 4 and 5 are:

1. A healthcare data set has not been previously used for illustration purposes in the synthetic data literature. Previous data sets utilized have generally been surveys with a demographic or economic focus.

2. The original nine-stratum sampling mechanism provides a great data source to investigate different levels of design information to be included in the synthetic data.

For illustration purposes, we treated the publicly available data set as the population, and created observed and synthetic samples with 30% of the population units.

Central to Chapters 3-5 is an understanding of the creation and analysis of fully synthetic data sets. This material is reviewed next in Chapter 2.

Chapter 2

Multiple Imputation for Synthetic Data

Here we review the creation and analysis of fully synthetic data from the original paper by Raghunathan, Reiter and Rubin (2003).

2.1 Creation of synthetic data

Consider the following hypothetical example. Suppose a government statistical agency collects detailed information on expenditure, income and household characteristics, of a sample of $n = 10,000$ households resident in private dwellings throughout a legal district. The size of the target population is $N = 500,000$. The original sampling mechanism (denoted $\mathscr{I}_0$) is stratified by a geographical unit indicator (such as township), and is known for all units in the population, and this variable (and other background variables known for all N units prior to data collection) form the matrix

X. There are p survey variables (unknown prior to data collection) of interest including weekly expenditure on a defined list of goods and services, current weekly income, employment status, and socio-demographic information, which form the $(N \times p)$ matrix Y. Let Y_0 be the $(n \times p)$ matrix representing the portion of Y corresponding to sampled or included units. Let Y_{exc} be the $((N-n) \times p)$ matrix representing the portion of Y corresponding to excluded units. Define $Z_0 = \{X, Y_0\}$ to be the known and observed microdata available to the agency which includes original survey design variables.

The statistical agency wishes to release the survey data collected to meet the demand for public access, and to facilitate research by external analysts. However, for confidentiality reasons, the data set Z_0 cannot be released. To minimize disclosure risk, the agency chooses to synthesize all survey variables. The task is to create $m > 1$ synthetic data sets (denoted by $Z_{\text{syn}}^{(1)}, ..., Z_{\text{syn}}^{(\text{m})}$) for Z_0. Multiple data sets are required to capture accurately the sampling variance in population quantity estimates, given the synthetic data, due to additional uncertainty from imputation. The agency decides the sample size of each synthetic data set will be $n_{\text{syn}} = 5,000$. Each $Z_{\text{syn}}^{(\text{l})}$ $(l = 1, .., m)$ is created in two steps as follows:

- **Step 1**: Impute the excluded values Y_{exc} from the posterior predictive distribution $\pi(Y_{\text{exc}}|Z_0)$. (Note, the agency can choose to impute all N population records so that the imputed synthetic population data set contains no observed values Y_0). From Step 1 we obtain a complete-data population, $P_{\text{com}}^{(\text{l})} = (X, Y_{\text{com}}^{(\text{l})})$ $(l = 1, \ldots, m)$ where $Y_{\text{com}}^{(\text{l})} = (Y_0, Y_{\text{exc}}^{(\text{l})})$.

- **Step 2**: Randomly sample without replacement n_{syn} units by resampling mech-

anism $\mathscr{I}_{\text{syn}}$ from $P_{\text{com}}^{(l)}$, producing a synthetic data set $Z_{\text{syn}}^{(l)} = (X, Y_{\text{syn}}^{(l)})$.

Note, we have used the term *resampling* from the imputed complete-data population to distinguish from the *original sampling* mechanism $\mathscr{I}_0$, and **not** in the sense of drawing new samples from a given sample.

Define $Z_{\text{syn}} = \left\{ Z_{\text{syn}}^{(l)}, l = 1, 2, ...m \right\}$ to be the collection of synthetic samples that are released, and $P_{\text{com}} = \left\{ P_{\text{com}}^{(l)}, l = 1, 2, ...m \right\}$ to be the collection of imputed synthetic populations from which they were resampled. The definition of $Z_{\text{syn}}^{(l)}$ in Step 2 assumes there are no confidentiality constraints on releasing X. If X is completely confidential and cannot be released at all, one may use X to create the synthetic data but release only synthesized survey variables.

Conceptually for Step 1, we impute to replace observed values because we do not want to release the true values. Nor do we necessarily wish to release the same units as originally sampled, so we impute excluded values as well. Generally, it is not practical to release the entire imputed population, hence in Step 2 we resample from $P_{\text{com}}^{(l)}$. It was proposed in Raghunathan, Reiter and Rubin (2003), that $\mathscr{I}_{\text{syn}}$ and $\mathscr{I}_0$ can define different sampling mechanisms because each synthetic sample $Z_{\text{syn}}^{(l)}$ is redrawn from an imputed population $P_{\text{com}}^{(l)}$, which does not contain any information on $\mathscr{I}_0$. For our hypothetical example, this means the agency can resample $n_{\text{syn}} = 5,000$ households by simple random sampling, ignoring the stratum indicators. Step 2 can be merged into Step 1 by only imputing synthetic values for resampled units drawn by $\mathscr{I}_{\text{syn}}$. To generate draws of multiple survey variables from their posterior predictive distribution, joint modeling (Schafer 1997) and sequential regression multivariate imputation (SRMI) (Van Buuren and Oudshoorn 2000, Raghunathan et al. 2001) approaches

may be used. For technical guidelines on specification of the imputation models for common variable types, refer to Reiter (2005a).

2.2 Analysis of synthetic data sets

The synthetic data Z_{syn} generated by the agency are released to analysts. The analyst needs some rules to combine the data across the m synthetic data sets it receives, to draw inference on some scalar population quantity Q, such as a population mean or regression coefficient. Raghunathan, Reiter and Rubin (2003) derived approximations to the first and second moments of the posterior distribution $\pi(Q|Z_{\text{syn}})$. The authors assumed both $\mathscr{I}_0$ and $\mathscr{I}_{\text{syn}}$ defined simple random sampling mechanisms. The key assumptions and inferential equations from this paper are stated below. For theoretical justification, the reader is referred to Raghunathan, Reiter and Rubin (2003, Section 4, pp. 9-11).

Suppose that, given the observed data set, the analyst would use the point estimate q_0 and an associated measure of uncertainty v_0 for inference about Q. Let $(q^{(l)}, v^{(l)})$ be the values of q_0 and v_0 computed using synthetic data set $Z_{\text{syn}}^{(l)}$.

The key assumptions are:

(i) Sample sizes are (a) large enough to permit normal approximations to posterior distributions and thus only the first two moments are required for each distribution, which can be derived using standard large sample Bayesian arguments; and (b) non-informative priors are assumed for all parameters, such that the information in the likelihood function dominates any information in the analyst's prior distribution. Both are reasonable assumptions in large data sets.

(ii) The point estimate q_0 is unbiased for Q and asymptotically normal, with respect to repeated sampling from the finite population (X, Y). The variance of the unbiased estimator is V_0.

(iii) The sampling variance estimate v_0 is unbiased for V_0, and the repeated sampling variability in v_0 is negligible; that is, v_0 and V_0 are interchangeable.

(iv) Let $Q^{(l)}$ be the unbiased estimate of Q from $P^{(l)}_{\text{com}}$. Given synthetic data set $Z^{(l)}_{\text{syn}}$, the estimate $q^{(l)}$ is unbiased for $Q^{(l)}$, and asymptotically normal with sampling variance $V^{(l)}$, and $V^{(l)}$ is an unbiased estimate for V_0.

(v) The sampling variance estimate $v^{(l)}$ is unbiased for $V^{(l)}$, and the sampling variability in $v^{(l)}$ is negligible. That is, $v^{(l)}\big|P^{(l)}_{\text{com}} \approx V^{(l)}$. Thus, $v^{(l)}$ and $V^{(l)}$ are interchangeable.

(vi) The variation in $V^{(l)}$ across the m synthetic populations is negligible; that is, $V^{(l)} \approx V_0$. Then by (iv) and (v), $v^{(l)} \approx V_0$.

We assume the form of $(q^{(l)}, v^{(l)})$ reflects the resampling mechanism $\mathscr{I}_{\text{syn}}$. Combining information from all m synthetic data sets, the following quantities are needed for inference:

$$\bar{q}_{\text{m}} = \sum_{l=1}^{\text{m}} \frac{q^{(l)}}{m} \tag{2.1}$$

$$b_m = \sum_{l=1}^{\text{m}} \frac{(q^{(l)} - \bar{q}_{\text{m}})^2}{(m-1)} \ , \tag{2.2}$$

$$\bar{v}_{\text{m}} = \sum_{l=1}^{\text{m}} \frac{v^{(l)}}{m} \ . \tag{2.3}$$

The analyst uses $\bar{q}_{\mathrm{m}}$ as the estimate of Q. The sampling variance of $\bar{q}_{\mathrm{m}}$ is estimated by the method-of-moments estimate

$$V_{\mathrm{syn}} = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_{\mathrm{m}} \ . \tag{2.4}$$

Extensions for multivariate Q are presented in Reiter (2005b). Note that Raghunathan, Reiter and Rubin (2003) denoted the variance estimator as T_{m}. We have switched to the V_{syn} notation to avoid confusion with the variance estimator for multiple imputation for missing data $T_{\mathrm{m}} = \left(1 + \frac{1}{m}\right) b_{\mathrm{m}} + \bar{v}_{\mathrm{m}}$ (Rubin 1987); furthermore, V_{syn} is not a total of two components. Specifically, the mean within variance $\bar{v}_{\mathrm{m}}$ is subtracted in (2.4) because there is an additional level of sampling when creating the synthetic data already included in the estimate b_m. The randomization validity of the inferential methods for fully synthetic data was justified in Raghunathan, Reiter and Rubin (2003). It should also be noted that the method-of-moments estimator for V_{syn} may be negative. An alternative variance estimation method applying adjustment for density maximization (ADM) (Morris 1988) to produce positive variance estimates is investigated in Chapter 5.

For moderate m, inferences for scalar Q can be based on a t-distribution with degrees of freedom

$$df_{\mathrm{syn}} = (m-1)\left(1 - \frac{1}{r_{\mathrm{m}}}\right)^2 = (m-1)\left(1 - \frac{m}{m+1}\frac{\bar{v}_{\mathrm{m}}}{b_{\mathrm{m}}}\right)^2 \ , \tag{2.5}$$

where $r_{\mathrm{m}} = \frac{\left(1+\frac{1}{m}\right)b_{\mathrm{m}}}{\bar{v}_{\mathrm{m}}}$ such that a nominal $100(1-\alpha)\%$ confidence interval estimate

for Q is

$$\bar{q}_{\mathrm{m}} \pm t_{df_{\mathrm{syn}},\alpha/2}\sqrt{V_{\mathrm{syn}}}\ . \tag{2.6}$$

For large m, inference can be based on a standard normal distribution. For this thesis, we have assumed m to be large.

Chapter 3

Congeniality for Synthetic Data Sets

3.1 Case studies

3.1.1 Ignoring the original survey design

Example 3.1.1

Consider the following simulation setup. Let each unit $i = 1, .., N$ in the true population be a member of stratum j, where $j = 1, 2$. The stratum indicators are known for all units and are available for creating the synthetic data sets but not for public release. The population size of each stratum is $N_1 = N_2 = 10,000$. Survey values for stratum 1 are drawn from a $N(\mu_1 = 100, \sigma_1 = 1)$ distribution, and survey values for stratum 2 are drawn from a $N(\mu_2 = 10, \sigma_2 = 1)$ distribution. The stratum means have been chosen to define two very distinct strata so that the design effect

(Lohr 1999) for the observed data set is very small. It would be obvious if the analyst had access to the observed data set to use a stratified random sample (STRS) estimator to obtain an efficient estimate of the population mean (the quantity of interest).

The observed sample consists of $n_1 = n_2 = 1,500$ units from each stratum. Given the observed sample, $m = 200$ synthetic populations are imputed using a normal imputation model conditional on the stratum indicator. We evaluate both simple random sample (SRS) and STRS resampling mechanisms for $n_{\text{syn}} = 3,000$. The true population value of $\bar{Y}$ is 55.00. We base our evaluation on the coverage of the nominal 95% confidence interval for the population mean across 1,000 replications. Thus the margin of error is given by $1.96\sqrt{0.05 \times 0.95/1000} = 0.014$, implying that we cannot distinguish interval estimate coverage from the 95% nominal level when the coverage rate of the true population mean is between 93.6% and 96.4%. Results are summarized in Table 3.1. The abbreviation **CR** stands for coverage rate, and **AW** for the average width of the interval estimate.

Table 3.1: *Simulation study results - ignoring the original survey design*

Data	**Samp. mech. & analysis proced.**	**Avg.**$\bar{\mathbf{q}}_{\text{m}}$	**Var.**$\bar{\mathbf{q}}_{\text{m}}$	**Avg.**$\mathbf{V}_{\text{syn}}$	**CR**	**AW**
Observed	STRS	55.00	4×10^{-4}	4×10^{-4}	95.0	0.083
Synthetic	SRS	55.00	9×10^{-4}	500×10^{-4}	99.3	2.46
	STRS	55.00	5×10^{-4}	5×10^{-4}	94.6	0.083

Avg.$\bar{\mathbf{q}}_{\text{m}}$: Average value for the posterior mean estimate of Q, across 1,000 replications
Var.$\bar{\mathbf{q}}_{\text{m}}$: Variance of the posterior mean estimates for Q, across 1,000 replications
Avg.$\mathbf{V}_{\text{syn}}$: Average value for the posterior variance estimate of Q, across 1,000 replications
CR: coverage rate
AW: average width

When $\mathscr{I}_0$ and $\mathscr{I}_{\text{syn}}$ both define STRS resampling and analysis procedures, syn-

thetic data results match the observed data set results, and we say data utility is preserved. But for the SRS resampling and analysis procedure, synthetic data interval estimates show gross overcoverage (CR=99%). We pose the question, what is driving the gross overcoverage when ignoring the original survey design in resampling and analyzing the synthetic data? What does this imply about the assumptions underlying the inferential methods for fully synthetic data?

3.1.2 Imputation model and analysis procedure condition on different sets of records

Example 3.1.2

This example comes from the investigation to create partially synthetic, public data for the Cancer Care Outcomes Research and Surveillance (CanCORS) lung cancer patient survey data set (Chapter 6). CanCORS is a large-scale study of the quality and patterns of care given to population-based and healthcare-system-based cohorts of newly diagnosed patients with lung and colorectal cancer, across 11 study sites in the US (Ayanian et al. 2004). More than 500 survey variables were collected. The demographic variables age, education, race, marital status and sex were identified as high disclosure risk and would be synthesized. All other variables (including clinical variables) were not synthesized. Predictors for the imputation models were identified by stepwise regression to minimize Akaike's Information Criterion (AIC). Refer to Chapter 6 for full justification of the selection of high disclosure risk variables and specification of the imputation models.

As part of our data utility assessment to compare analytic results using the syn-

thetic data relative to inference using the observed data set, we ran models based on the published analysis in Huskamp et al. (2009). In this paper, the authors seek to identify the factors associated with hospice discussion rates amongst stage IV lung cancer patients. The original analysis was based on the $1,517$ patients who had Stage IV lung cancer, which represents approximately 30% of the full set of records conditioned upon in the imputation models. Analytical results for the synthesized covariate 'race' as a predictor of hospice discussion, unadjusted for other covariates are presented in Table 3.2. Note that the hospice discussion response variable was not synthesized.

Table 3.2: *Descriptive characteristics and estimated probabilities of hospice discussion by race, unadjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Patients %**		**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Race/ethnicity						
White	73.7	72.0	55.2 (1.3)	54.4 (1.5)		
Black	10.7	11.4	42.6 (1.3)	50.0 (4.1)		
Hispanic	5.9	5.7	40.4 (1.3)	45.8 (5.3)	< 0.001	0.62
Asian	5.1	5.3	49.4 (1.3)	51.6 (6.0)		
Other	4.7	4.8	64.5 (1.2)	54.0 (6.9)		

The marginal sample counts using synthetic data agree with the observed data set statistics in Table 3.2. However, as a predictor of hospice discussion, 'race' was strongly significant using the observed data set, but insignificant using synthetic data ($p_{\text{obs}} < 0.001$ vs $p_{\text{syn}} = 0.62$). What explanation can be given for the change in conclusion of significance, beyond the additional uncertainty due to the multiple imputation procedure?

3.2 Congeniality for multiple imputation for synthetic data

We denote the different types of input in multiple imputation for synthetic data as follows:

- *Imputation input*: $Z_0 + A$
- *Analysis input*: Z_{syn}

where A denotes any additional information available to the imputer such as model assumptions and information from past similar surveys. Also as previously mentioned in Section 2.1, if X is completely confidential and cannot be released at all, then $Z_{\text{syn}} = Y_{\text{syn}}$. The analysis input also encompasses the analyst's purpose of investigation and assessment of any information provided by the agency.

In the synthetic data setting, Z_0 is not a source of input for the analyst. However, we shall define Z_0^{user} as input analysts would have used if they had access to the observed data set. We also define $\mathscr{P}_0^{\text{user}}$ to be the analysis procedure given data set Z_0^{user}, and $\mathscr{P}_{\text{syn}}$ to be the analysis procedure given data Z_{syn}. The analysis procedure encompasses a survey design assumption (for design-based estimation), or a distributional assumption (for model-based estimation). Let h denote a Bayesian model.

Definition 1

A Bayesian model h is said to be congenial to the analysis procedure $\mathscr{P} = \{\mathscr{P}_0^{\text{user}}, \mathscr{P}_{\text{syn}}\}$ if:

(i) The posterior mean and variance of Q under h given Z_0^{user} are asymptotically the same as the estimate and variance from the analysis procedure $\mathscr{P}_0^{\text{user}}$, that is

$$[\hat{Q}(Z_0^{\text{user}}), U(Z_0^{\text{user}})] \simeq [E_h[Q|Z_0^{\text{user}}], V_h[Q|Z_0^{\text{user}}]] \tag{3.1}$$

(ii) The posterior mean and variance of Q under h given Z_{syn} are asymptotically the same as the estimate and variance from the analysis procedure $\mathscr{P}_{\text{syn}}$, that is

$$[\hat{Q}(Z_{\text{syn}}), U(Z_{\text{syn}})] \simeq [E_h[Q|Z_{\text{syn}}], V_h[Q|Z_{\text{syn}}]] \tag{3.2}$$

Because multiple imputation is a Bayesian procedure, (3.1) - (3.2) are required to show inferential equivalence of the frequentist analysis procedure to some Bayesian model. Definition 1 is analogous to (2.3.1) and (2.3.2) in Meng (1994), p. 543.

Definition 2

The analysis procedure $\mathscr{P} = \{\mathscr{P}_0^{\text{user}}, \mathscr{P}_{\text{syn}}\}$ is said to be congenial to the imputation model $g(Y_{\text{syn}}|Z_0, A)$ if one can find an h such that asymptotically

(i) h is congenial to $\mathscr{P}$ under Definition 1.

(ii) The posterior predictive density for $Y_{\text{com}}^{(l)}$ $(l = 1, ..., m)$ derived under h is identical to the imputation model

$$h(Y_{\text{com}}^{(l)}|Z_0^{\text{user}}) = g(Y_{\text{com}}^{(l)}|Z_0, A) \tag{3.3}$$

for all possible $Y_{\text{com}}^{(l)}$.

(iii) The posterior predictive density for $Y_{\text{syn}}^{*(l)}$ $(l = 1, ..., m)$ derived under h is identical to the imputation model

$$h(Y_{\text{syn}}^{*(l)}|Z_{\text{syn}}) = g(Y_{\text{syn}}^{*(l)}|Z_0, A) \tag{3.4}$$

for all possible $Y_{\text{syn}}^{*(l)}$.

(note: we use the notation $Y_{\text{syn}}^{*(l)}$ $(l = 1, ..., m)$ to distinguish from $Y_{\text{syn}} \in Z_{\text{syn}}$).

The conditioning sets differ on each side of the equality in (3.4) and potentially in (3.3). For congeniality to hold, Definition 2 (ii) requires that if the user had access to some portion of the observed data set to conduct inference, any significant relationships in the analysis procedure must be included in the imputation model. Definition 2 (iii) requires that if we were to generate a new synthetic data set given Z_{syn}, the posterior predictive distribution we would use is identical to the imputation model used to generate Z_{syn}, because for congeniality to hold, both Z_0 and Z_{syn} must be drawn from the same target population. This means that Z_{syn} must be defined by the same sampling distribution as Z_0.

In the next section, we use the definition of congeniality for synthetic data to answer the questions posed in the motivating examples in Section 3.1.

3.3 Illustration of congeniality and uncongeniality for synthetic data

3.3.1 Ignoring the original survey design

Example 3.1.1 (*Continued*)

Refer back to the case study described in Section 3.1.1. We define $\bar{y}$ to be a sample mean and s to be a sample standard deviation. Let $\tilde{y}^{(\mathrm{l})}$ denote an imputed value from synthetic data set $Z_{\mathrm{syn}}^{(\mathrm{l})}$.

Table 3.3: *Imputation model and analysis procedure options*

Imputation model I (SRS)	**Imputation model II (STRS)**
$g_1(\tilde{y}_i^{(\mathrm{l})}\|Z_0, A)$ $\sim t_{\mathrm{n}-1}\left(\bar{y}_0, \left(1+\frac{1}{n}\right)^{1/2} s_0\right)$ $i = 1, ..., n_{\mathrm{syn}}$	$g_2(\tilde{y}_{\mathrm{ij}}^{(\mathrm{l})}\|Z_0, A)$ $\sim t_{\mathrm{n_j}-1}\left(\bar{y}_{0\mathrm{j}}, \left(1+\frac{1}{\mathrm{n_j}}\right)^{1/2} s_{0\mathrm{j}}\right)$ $i = 1, ..., n_{\mathrm{syn,j}}$; $j = 1, 2$
Analysis procedure I (SRS)	**Analysis procedure II (STRS)**
$\hat{Q}_1(Z_{\mathrm{syn}}) = \bar{y}_{\mathrm{syn}}$ $U_1(Z_{\mathrm{syn}}) = s_{\mathrm{syn}}^2/n_{\mathrm{syn}}$	$\hat{Q}_2(Z_{\mathrm{syn}}) = \frac{N_1}{N}\bar{y}_{\mathrm{syn1}} + \frac{N_2}{N}\bar{y}_{\mathrm{syn2}}$ $U_2(Z_{\mathrm{syn}}) = \left(\frac{N_1}{N}\right)^2 \frac{s_{\mathrm{syn1}}^2}{n_{\mathrm{syn1}}} + \left(\frac{N_2}{N}\right)^2 \frac{s_{\mathrm{syn2}}^2}{n_{\mathrm{syn2}}}$

Table 3.3 shows the imputation model and analysis procedure options available in this example (ignoring finite population correction factors). The question of interest is whether imputation model II (STRS) is congenial to analysis procedure I (SRS), as required by Definition 2 (iii) (3.4). That is, can we recover the two-stratum population structure from the synthetic data that is resampled and analyzed by simple random sampling. Now imputation model II is congenial to analysis procedure II (Definition

2 (ii) (3.3) - they both assume STRS estimators), and therefore we would like to show whether $\mathrm{E}[\hat{Q}_1(Z_{\text{syn}})] = \mathrm{E}[\hat{Q}_2(Z_{\text{syn}})]$ and $U_1(Z_{\text{syn}}) = U_2(Z_{\text{syn}})$.

Using simple random sampling, the sample counts by stratum are random variables defined by the distributions $n_{\text{syn1}} \sim \text{Bin}\left(n_{\text{syn}}, \frac{N_1}{N}\right)$ and $n_{\text{syn2}} \sim \text{Bin}\left(n_{\text{syn}}, \frac{N_2}{N}\right)$, whereas n_{syn1} and n_{syn2} are fixed and known under STRS. If we re-express $\hat{Q}_1(Z_{\text{syn}})$ as

$$\bar{y}_{\text{syn}} = \frac{\mathrm{n}_{\text{syn1}}}{n_{\text{syn}}}\bar{y}_{\text{syn1}} + \frac{n_{\text{syn2}}}{n_{\text{syn}}}\bar{y}_{\text{syn2}}$$

Then using conditional expectations we can show

$$\begin{aligned}
\mathrm{E}[\hat{Q}_1(Z_{\text{syn}})] &= \mathrm{E}\left[\mathrm{E}\left[\frac{n_{\text{syn1}}}{n_{\text{syn}}}\bar{y}_{\text{syn1}} + \frac{n_{\text{syn2}}}{n_{\text{syn}}}\bar{y}_{\text{syn2}}\middle| n_{\text{syn1}}\right]\right] \\
&= \mathrm{E}\left[\frac{n_{\text{syn1}}}{n_{\text{syn}}}\mu_1 + \frac{n_{\text{syn2}}}{n_{\text{syn}}}\mu_2\right] \\
&= \frac{N_1}{N}\mu_1 + \frac{N_2}{N}\mu_2 \\
&= \mathrm{E}[\hat{Q}_2(Z_{\text{syn}})] \ ,
\end{aligned}$$

and

$$
\begin{aligned}
\mathrm{Var}[\hat{Q}_1(Z_{\mathrm{syn}})] &= \mathrm{E}\left[\mathrm{Var}\left[\frac{n_{\mathrm{syn1}}}{n_{\mathrm{syn}}}\bar{y}_{\mathrm{syn1}} + \frac{n_{\mathrm{syn2}}}{n_{\mathrm{syn}}}\bar{y}_{\mathrm{syn2}}\middle| n_{\mathrm{syn1}}\right]\right] + \\
&\qquad \mathrm{Var}\left[\mathrm{E}\left[\frac{n_{\mathrm{syn1}}}{n_{\mathrm{syn}}}\bar{y}_{\mathrm{syn1}} + \frac{n_{\mathrm{syn2}}}{n_{\mathrm{syn}}}\bar{y}_{\mathrm{syn2}}\middle| n_{\mathrm{syn1}}\right]\right] \\
&= \mathrm{E}\left[\left(\frac{n_{\mathrm{syn1}}}{n_{\mathrm{syn}}}\right)^2 \sigma^2_{\bar{Y}_1} + \left(\frac{n_{\mathrm{syn2}}}{n_{\mathrm{syn}}}\right)^2 \sigma^2_{\bar{Y}_2}\right] + \mathrm{Var}\left[\frac{n_{\mathrm{syn1}}}{n_{\mathrm{syn}}}\mu_1 + \frac{n_{\mathrm{syn2}}}{n_{\mathrm{syn}}}\mu_2\right] \\
&= \frac{1}{n^2_{\mathrm{syn}}}\left[\mathrm{E}\left[n_{\mathrm{syn1}}\right]^2 \sigma^2_{\bar{Y}_1} + \mathrm{E}\left[n_{\mathrm{syn2}}\right]^2 \sigma^2_{\bar{Y}_2}\right] + \\
&\qquad \frac{1}{n^2_{\mathrm{syn}}}\left[\mathrm{Var}\left[n_{\mathrm{syn1}}\right](\sigma^2_{\bar{Y}_1} + \mu_1^2) + \mathrm{Var}\left[n_{\mathrm{syn2}}\right](\sigma^2_{\bar{Y}_2} + \mu_2^2)\right] \\
&\approx \left(\frac{N_1}{N}\right)^2 \sigma^2_{\bar{Y}_1} + \left(\frac{N_2}{N}\right)^2 \sigma^2_{\bar{Y}_2} \\
&= \mathrm{Var}[\hat{Q}_2(Z_{\mathrm{syn}})] \ .
\end{aligned}
$$

The approximation in the second to last line comes from ignoring the variability in n_{syn1} and n_{syn2}, which is justified when n_{syn} is large, and so asymptotically congeniality is satisfied. However, the size of n_{syn} required may reduce the disclosure risk benefits of using synthetic data. For the simulation study in Example 3.1.1, more than 80% of the population units are required to be released in each synthetic data set in order to eliminate the gross overcoverage in the interval estimate for the population mean. That is, analysis procedure I is not efficient relative to analysis procedure II, for fixed n_{syn} and/or m, because information is lost on the population structure at the analysis stage, and without the stratum indicators, it takes longer to recover the two-stratum structure of the target population.

3.3.2 Imputation model and analysis procedure condition on different sets of records

Example 3.1.2 (*Continued*)

Refer back to the case study described in Section 3.1.2. To understand better the statistical implications of different conditioning sets of records, we used a posterior predictive simulation given the observed data set to predict hospice discussion rates for black patients using (i) the full data set (Z_0); and (ii) stage IV patients only (Z_0^{user}). We chose black patients because they showed a large deviation in estimated probability of hospice discussion between observed and synthetic data analytic results (see Table 3.2). The mean posterior predictive probability of hospice discussion amongst black patients across the entire data set was $\hat{Q}(Z_0) = 0.29$. Let $Q(Z_0^{\text{user}})$ denote the posterior predictive probability for hospice discussion amongst black patients given Z_0^{user}, that is Stage IV patients only. We calculated the posterior predictive p-value to be $Pr(Q(Z_0^{\text{user}}) < 0.29) = 0.009$, and the difference in posterior predictive distributions is confirmed in Figure 3.1 by the contrast in location and spread of the two histograms. These results show Definition 2 (ii) (3.3) has been violated. These results make sense clinically because hospice discussion rates are lower if measured for cancer patients at all disease stages, whereas rates are elevated if measured for later Stage IV cancer patients only.

To confirm our reasoning, we ran the imputation model again but conditional on the same set of records as used in the analysis procedure. The revised analytical comparison results are in Table 3.4. Utilizing the same set of records has assisted in preserving the conclusion of significance ($p_{\text{obs}} < 0.001$ vs $p_{\text{syn}} = 0.003$) for the 'race'

covariate as a predictor of hospice discussion.

Table 3.4: *Descriptive characteristics and estimated probabilities of hospice discussion by race, unadjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Patients %**		**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Race/ethnicity						
White	73.7	74.6	55.2 (1.3)	55.6 (1.5)		
Black	10.7	9.9	42.6 (1.3)	42.0 (4.2)		
Hispanic	5.9	6.1	40.4 (1.3)	40.4 (5.3)	< 0.001	0.003
Asian	5.1	5.1	49.4 (1.3)	50.2 (5.8)		
Other	4.7	4.3	64.5 (1.2)	58.5 (6.5)		

3.4 Discussion

In this chapter we have proposed a definition of congeniality for multiple imputation for synthetic data. We have used the definition to understand better the role of the original survey design in the synthetic data resampling mechanism and analysis procedure, and establish in an example that congeniality holds asymptotically if the original survey design is ignored, and replaced by a simple random sample design. A simple random sample design just ignores the population structure from which a more complex design will produce more efficient estimators for a fixed synthetic data sample size or number of imputations. These results confirm previous simulation study results and comments in Reiter (2002), Drechsler et al. (2008), and Reiter and Drechsler (2010), on inclusion of survey design information in synthetic data sets. We have also used the definition to explain discrepancies between synthetic data and observed data set analytic results, when the imputation model does not condition

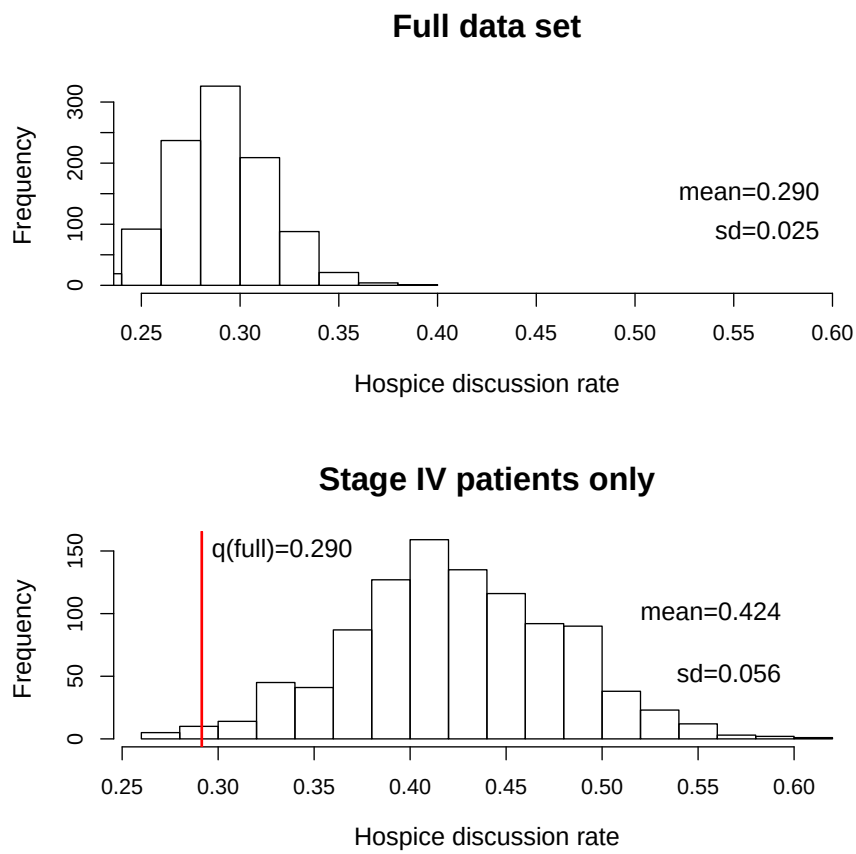


Figure 3.1: *Comparison of posterior predictive distributions for hospice discussion rate amongst black patients*

upon the same set of records as the analysis procedure. Both our case studies were common examples of uncongeniality in practice.

Definition 2 implies that in order for congeniality to be satisfied, there must exist asymptotically some true, unique imputation model g for given analysis procedures $\mathscr{P}(Z_0^{\text{user}})$ and $\mathscr{P}(Z_{\text{syn}})$. This raises two questions:

(i) How can imputers identify, and know they have identified, the 'true' model g?

(ii) Given the infinite number of potential future analyses, it is impossible for the im-

puter to satisfy congeniality for all potential future analyses. What constitutes a representative analysis, how many representative analyses should be run, and what results from the multiple analytical comparisons should be communicated to analysts?

Analytical solutions appear not to be available to answer the first question. The imputer must rely on the empirical data utility checks suggested in Karr et al. (2006) and Woo et al. (2009), such as confidence interval overlap and empirical distribution comparison, to check the quality of the synthetic data generated relative to inference using the observed data set. The theoretical model g represents the ideal imputation model, which can guide the imputer to identify the source of a discrepancy between observed data set and synthetic data analytic results, as illustrated by the case studies in Examples 3.1.1 and 3.1.2.

The second question is best answered on a case by case basis as it depends on the analysis procedures $\mathscr{P}(Z_0^{\text{user}})$ and $\mathscr{P}(Z_{\text{syn}})$ that are relevant to the observed and synthetic data. The most practical solution is for the imputer to endeavor to make the imputation models as complex as possible, without releasing confidential information, and provide assurance to the pool of respective analysts, that the synthetic data has been created by expert statisticians.

Chapter 4

Recovery of the Observed Data Set Sampling Distribution of Sufficient Statistics Using Synthetic Data

The definition of congeniality for synthetic data presented in Section 3.2 requires that in order for congeniality to be satisfied between the imputation model and the analysis procedure, we should be able to recover the observed data set sampling distribution given the synthetic data. This suggests an alternative approach to fully synthetic data inference:

(i) Infer the sufficient summaries (q_0, v_0) of the observed data set given Z_{syn}.

(ii) Proceed to draw inference on Q as if the analyst obtained q_0 and v_0 by direct access to the observed data set Z_0.

The objective is to recover select statistics from the observed data set that would be used in a frequentist analysis procedure for inference on Q, as opposed to the full posterior distribution of Q.

Because the alternative analysis equations need to be derived separately for every quantity of interest and survey design assumption, we cannot present a single set of combining rules that apply for any scalar quantity of interest. Instead, we will demonstrate the alternative approach by simulation and empirical data study. For all derivations in this chapter, we make the same assumptions as listed in Section 2.2.

4.1 Estimation of an observed sample mean using synthetic data

Example 4.1.1

Suppose the analyst wishes to estimate the population mean for a univariate survey variable Y. With no access to additional information other than the synthetic variable Y_{syn}, the analyst assumes $\mathscr{I}_{\text{syn}}$ to define simple random sampling. Thus the required observed data set statistics are $q_0 = \bar{y}_0$ and $v_0 = s_0^2/n$. Suppose the analyst correctly assumes the population data are normally distributed and without loss of generality, let $n_{\text{syn}} = n$.

Let $\tilde{y}_i^{(1)}$ be the imputed value for the ith record from synthetic data set $Z_{\text{syn}}^{(1)}$, which is drawn from the posterior predictive distribution

$$\tilde{y}_i^{(1)} | Z_0 \sim N\left(q_0, v_0(n+1)\right) \ ,$$

and so

$$q^{(l)}|Z_0 = \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i^{(l)}|Z_0 \sim N\left(\bar{y}_0, v_0\left(\frac{n+1}{n}\right)\right) \ ,$$

and

$$\bar{q}_{\mathrm{m}}|Z_0 = \frac{1}{m}\sum_{l=1}^{m} q^{(l)}\Big|Z_0 \sim N\left(\bar{y}_0, v_0\left(\frac{n+1}{m\times n}\right)\right) \ .$$

Given our assumptions and applying large sample Bayesian arguments to condition on Z_{syn}, we have

$$q_0|Z_{\mathrm{syn}} \sim N\left(\bar{q}_{\mathrm{m}}, \frac{\bar{v}_{\mathrm{m}}}{m}\left(\frac{n+1}{n}\right)\right) \ .$$

Therefore

$$\mathrm{E}[q_0|Z_{\mathrm{syn}}] \approx \bar{q}_{\mathrm{m}} \ , \tag{4.1}$$

and

$$V_{\mathrm{syn,alt}} = \mathrm{Var}[q_0|Z_{\mathrm{syn}}] \approx \frac{\bar{v}_{\mathrm{m}}}{m} + \frac{\bar{v}_{\mathrm{m}}}{mn} \ . \tag{4.2}$$

We now have an alternative variance estimator $V_{\mathrm{syn,alt}}$. The first term in (4.2) represents the within-imputation sampling variance, and the second term represents the variability from estimating the population mean $Q^{(l)}$ from the imputed population $P_{\mathrm{com}}^{(l)}$. We evaluate (4.1) and (4.2) by the same simulation study as in Example 3.1.1. Results are presented in Table 4.1, and are adjusted for the finite population correction factor.

Table 4.1: *Simulation study results - estimation of an observed sample mean given synthetic data*

Data	**Samp. mech. & analysis proced.**	**Variance estimator**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Avg.$\mathbf{V}_{\mathrm{syn}}$**	**CR**	**AW**
Observed	STRS		55.00	4×10^{-4}	4×10^{-4}	95.0	0.083
Synthetic	SRS	V_{syn}	55.00	9×10^{-4}	500×10^{-4}	99.3	2.46
		$V_{\mathrm{syn,alt}}$	55.00	9×10^{-4}	9×10^{-4}	94.0	0.374
	STRS	V_{syn}	55.00	5×10^{-4}	5×10^{-4}	94.6	0.083
		$V_{\mathrm{syn,alt}}$ †	55.00	5×10^{-4}	5×10^{-4}	94.4	0.083

Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1,000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1,000 replications
Avg.$\mathbf{V}_{\mathrm{syn}}$: Average value for the posterior variance estimate of Q, across 1,000 replications
CR: coverage rate
AW: average width
† analysis equations (4.1) - (4.2) applied within each strata

First examine the results using synthetic data and a SRS resampling mechanism and analysis procedure. The interval estimates using the V_{syn} estimator showed gross overcoverage as expected from previous results (see Section 3.1.1). The coverage rate using the alternative approach was approximately 95%, but the average width was 4.5 times greater than the observed data set interval width.

The optimal synthetic data results were obtained when using the STRS resampling mechanism and analysis procedure, and either the V_{syn} or $V_{\mathrm{syn,alt}}$ estimator (both confidence interval coverage and average width results matched the observed data set results). The approximate equivalence satisfies congeniality Definition 2 (iii) (3.4), which requires equivalence of the synthetic data sampling distribution to the observed data set sampling distribution for a fixed target population.

4.2 Theoretical results

Let synthetic data Z_{syn} be created as outlined in Section 2.1. We require the inferential quantities (2.1) - (2.3), and the following population quantities:

$$\bar{Q}_{\text{m}} = \sum_{l=1}^{\text{m}} \frac{Q^{(\text{l})}}{m} \ , \tag{4.3}$$

$$B_{\text{m}} = \frac{\sum_{l=1}^{\text{m}} \left(Q^{(\text{l})} - \bar{Q}_{\text{m}}\right)^2}{m-1} \ , \tag{4.4}$$

$$\bar{V}_{\text{m}} = \frac{1}{m} \sum_{l=1}^{m} V^{(\text{l})} \tag{4.5}$$

where $Q^{(\text{l})}$ denotes the computed value of the population quantity Q based on the imputed complete-data population $P_{\text{com}}^{(\text{l})}$. We require expressions for $\text{E}[q_0|Z_{\text{syn}}]$ and $\text{Var}[q_0|Z_{\text{syn}}]$.

We make use of the posterior distribution

$$Q^{(\text{l})}|Z_0 \sim N\left(q_0, v_0\right) \ , \tag{4.6}$$

and hence

$$\bar{Q}_{\text{m}}|Z_0 \sim N\left(q_0, v_0/m\right) \ . \tag{4.7}$$

Applying large sample Bayesian arguments, we have

$$q_0|P_{\text{com}}, v_0 \simeq N\left(\bar{Q}_{\text{m}}, \left(1+\frac{1}{m}\right)v_0\right) \ . \tag{4.8}$$

The additional v_0/m is to account for finite m (Rubin 1987, pp. 87-91).

In (4.8) we cannot replace v_0 with $\bar{v}_{\text{m}}$ because $\bar{v}_{\text{m}}$ is only available if we are conditioning on Z_{syn}. If we are conditioning on P_{com}, then $\bar{V}_{\text{m}} = 0$ because each $P_{\text{com}}^{(\text{l})}$ is a complete population. The quantity B_m is the estimate for the population quantity $\text{Var}(Q|P_{\text{com}})$. A naive approach would be to use the estimate $\hat{v}_0|P_{\text{com}} = B_m$ ignoring any adjustment for taking a sample from a population. Adopting this naive approach we have

$$q_0|P_{\text{com}} \simeq N\left(\bar{Q}_{\text{m}}, \left(1+\frac{1}{m}\right)B_m\right) \ . \tag{4.9}$$

Also note the posterior distribution

$$\bar{Q}_{\text{m}}|Z_{\text{syn}} \sim N(\bar{q}_{\text{m}}, \bar{v}_{\text{m}}/m) \ . \tag{4.10}$$

Combining (4.9) and (4.10) we have

$$\begin{aligned} \text{E}[q_0|Z_{\text{syn}}] &= \text{E}[\text{E}[q_0|P_{\text{com}}]\big|Z_{\text{syn}}] && (4.11)\\ &= E[\bar{Q}_{\text{m}}\big|Z_{\text{syn}}] && (4.12)\\ &= \bar{q}_{\text{m}} \ ; && (4.13) \end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[q_0|Z_{\mathrm{syn}}] &= \mathrm{E}[\mathrm{Var}[q_0|P_{\mathrm{com}}]\big|Z_{\mathrm{syn}}] + \mathrm{Var}[\mathrm{E}[q_0|P_{\mathrm{com}}]\big|Z_{\mathrm{syn}}] && (4.14)\\
&= \mathrm{E}[(1+1/m)B_m\big|Z_{\mathrm{syn}}] + \mathrm{Var}[\bar{Q}_{\mathrm{m}}\big|Z_{\mathrm{syn}}] && (4.15)\\
&= (1+1/m)E[B_m\big|Z_{\mathrm{syn}}] + \bar{v}_{\mathrm{m}}/m && (4.16)\\
&\approx (1+1/m)b_m - \bar{v}_{\mathrm{m}}\ , && (4.17)
\end{aligned}$$

where approximation (4.17) follows from equation [7] in Raghunathan, Reiter and Rubin (2003), p. 11.

Now compare the mean and variance in (4.13) and (4.17) to the posterior moments derived in Raghunathan, Reiter and Rubin (2003) (see equations (2.1) and (2.4) which are restated below).

$$\mathrm{E}[Q|Z_{\mathrm{syn}}] = \bar{q}_{\mathrm{m}}\ ;$$

$$V_{\mathrm{syn}} = \mathrm{Var}[Q|Z_{\mathrm{syn}}] = (1+1/m)b_m - \bar{v}_{\mathrm{m}}\ .$$

We have obtained the equivalent expressions for $\mathrm{E}[Q|Z_{\mathrm{syn}}]$ and $\mathrm{Var}[Q|Z_{\mathrm{syn}}]$. However, q_0 is a sample quantity, but Q is a population quantity. The error is due to the naive approximation $\mathrm{E}[v_0|P_{\mathrm{com}}] \approx B_m$, which ignores any sampling adjustment given by $\mathscr{I}_0$. That is, any efficiency gains from a more complex original survey design are ignored in the derivations in Raghunathan, Reiter and Rubin (2003), which assumed $\mathscr{I}_{\mathrm{syn}}$ was a simple random sample. This is the same conclusion drawn to explain the

gross overcoverage in interval estimates obtained in the case study in Section 3.3.1, where we ignored the original survey design in the resampling and analysis of fully synthetic data.

4.3 Estimation of an observed population proportion using synthetic data

Example 4.3.1

In this example, we wish to estimate the population proportion ($Q = p$) for a binary variable Y. The hypothetical population is composed of $N = 1,000$ units with two variables (X, Y). We draw X from a standard uniform distribution, and then draw $Y_i \sim \text{Bin}(1, p_i)$, where $\ln \frac{p_i}{1-p_i} = 0.5 + X_i$ for $i = 1, .., N$. We assume X is known for all units and is available for sampling the collected data, but not for public release. In the generated population, the proportion value is 0.626. We base our evaluation on the coverage of the nominal 95% confidence interval for the population proportion across 1,000 replications.

Observed data are collected by sampling $n = 100$ units with probability proportional to X, without replacement using the Midzuno sampling (Midzuno 1952) function for πPS sampling, in the *sampling* library package (Lumley 2011) of the R software environment for statistical computing and graphics. To create synthetic values for Y, we draw from the full Bayesian posterior predictive distribution given the logistic regression of Y_0 on X_0. That is, we first (i) draw values of the logistic model parameters (β) from their joint posterior distribution, or approximations to it given

the observed data set; and second (ii) generate $n_{\text{syn}} = 100$ synthetic values of p and Y given the posterior draw of β and known background variable X. A non-informative beta (conjugate) prior distribution is assumed for p. We generate $m = 200$ synthetic data sets.

With no access to the size variable X, the analyst assumes $\mathscr{I}_{\text{syn}}$ to define simple random sampling. The observed-data statistics of interest are $q_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_0}{n}$ and $v_0 = \frac{q_0(1-q_0)}{n}$. Given synthetic data set $Z_{\text{syn}}^{(l)}$, the sufficient summary statistics are $q^{(l)} = \frac{\sum_{i=1}^{n} \tilde{y}_i^{(l)}}{n}$ and $v^{(l)} = \frac{q^{(l)}(1-q^{(l)})}{n}$, assuming $n_{\text{syn}} = n$, and ignoring the finite population correction factor. We desire expressions for $E[q_0|Z_{\text{syn}}]$ and $Var[q_0|Z_{\text{syn}}]$.

First, from the posterior predictive distribution conditional on the observed data set, we derive expressions for the mean and variance of $q^{(l)}$:

$$
\begin{aligned}
\mathrm{E}[q^{(l)}\big|q_0] &= \frac{1}{n}\mathrm{E}\left[\mathrm{E}\left[q^{(l)}|p\right]\Big|y_0\right] \\
&= \frac{1}{n}\mathrm{E}\left[np\Big|y_0\right] \\
&= y_0 \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{Var}[q^{(l)}|q_0] &= \frac{1}{n^2}\left[\mathrm{E}\left[\mathrm{Var}\left[q^{(l)}|p\right]\Big|y_0\right] + \mathrm{Var}\left[\mathrm{E}\left[q^{(l)}|p\right]\Big|y_0\right]\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[np(1-p)\Big|y_0\right] + \frac{1}{n^2}\mathrm{Var}\left[np\Big|y_0\right] \\
&= \frac{1}{n}\mathrm{E}\left[p\Big|y_0\right]\left(1-\mathrm{E}\left[p\Big|y_0\right]\right) + \mathrm{Var}\left[p\Big|y_0\right]\left(1-\frac{1}{n}\right) \\
&= \frac{1}{n}\frac{y_0}{n}\left(1-\frac{y_0}{n}\right) + \frac{y_0(n-y_0)}{n^2(n+1)}\left(\frac{n-1}{n}\right) \\
&= v_0 \times \left(1+\frac{n-1}{n+1}\right) \ .
\end{aligned}
$$

Combining data across all m synthetic data sets, and applying large sample Bayesian arguments to condition on Z_{syn}, we have

$$
\mathrm{E}[q_0|Z_{\mathrm{syn}}] \approx \bar{q}_{\mathrm{m}} \ , \tag{4.18}
$$

and

$$
\mathrm{Var}[q_0|Z_{\mathrm{syn}}] \approx \frac{\bar{v}_{\mathrm{m}}}{m} \times \left(1+\frac{n-1}{n+1}\right) \ . \tag{4.19}
$$

Simulation study results to evaluate (4.18) and (4.19) are summarized in Table 4.2, and include the finite population correction factor. We have calculated observed data set estimates using both design-based and model-based approaches. The model-based approach is a regression-adjusted estimate given the set of predictors used in the imputation models.

Using synthetic data and a SRS resampling mechanism and analysis procedure, the V_{syn} estimator produced the same nominal coverage as the model-based, observed

Table 4.2: *Simulation study results - estimation of an observed population proportion given synthetic data*

Data	**Samp. mech. & analysis proced.**	**Variance estimator**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Avg.$\mathbf{V}_{\mathrm{syn}}$**	**CR**	**AW**
Observed	πPS	Design	0.626	3.2×10^{-3}	3.2×10^{-3}	96.6	0.258
		Model	0.626	3.2×10^{-3}	3.2×10^{-3}	97.5	0.223
Synthetic	SRS	V_{syn}	0.621	2.4×10^{-3}	3.5×10^{-3}	97.5	0.230
		$V_{\mathrm{syn,alt}}$	0.621	2.4×10^{-3}	3.9×10^{-3}	99.8	0.243
	πPS	V_{syn}	0.624	2.5×10^{-3}	3.5×10^{-3}	96.0	0.226
		$V^{\dagger}_{\mathrm{syn,alt}}$	-	-	-	-	

Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1,000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1,000 replications
Avg.$\mathbf{V}_{\mathrm{syn}}$: Average value for the posterior variance estimate of Q, across 1,000 replications
CR: coverage rate
AW: average width
† no analytic solution available

data set estimate. We did not obtain gross overcoverage in interval estimates from ignoring the original survey design, because the design effect for the observed data set was ≈ 1. The slightly larger average confidence interval width ($0.230 > 0.223$), reflects the additional variability from multiple imputation. However, the interval width is shorter than the width of the design-based observed data set interval estimate (AW=0.258), because the design-based estimate does not make use of the predictor information conditioned upon in the imputation model.

The alternative approach produced a coverage rate of $\approx 100\%$. The analyst needs to have access to the size variable X to reduce the overcoverage. However, there is no analytical solution using the alternative approach and a πPS synthetic data resampling mechanism and analysis procedure. On the other hand, for disclosure protection the overcoverage is advantageous, because there is more uncertainty in

the inferential estimates analysts compute if information on the observed data set is withheld.

4.4 Estimation under an incorrect distributional assumption

Example 4.4.1

In this example, we investigate the performance of the alternative approach when an incorrect distributional assumption is made by the analyst.

Assume we wish to estimate the population mean of some univariate survey variable Y, where $\ln Y \sim N(\mu = 3, \sigma^2 = 9)$. The population is of size $N = 10,000$ and an observed sample of size $n = 500$ is drawn by simple random sampling. The mean of the generated population is $\bar{Y} = 2.99$. The synthetic population data are imputed based on a normal model given observed data set statistics $q_0 = \bar{x}_0$ and $v_0 = s^2_{0,x}/n$, where $X = \ln Y$. A simple random sample of size $n_{\text{syn}} = 500$ is drawn from each synthetic population. This process is repeated to create $m = 50$ synthetic data sets, for each of 1,000 replications.

The analyst assumes Y follows a gamma distribution, so that $\bar{Y} \sim \frac{1}{n}\text{Gamma}(n, 1/\mu)$ and the required sufficient statistic is $\bar{y}_0$.

To derive the inferential equations, the Bayesian setup is as follows:

- Prior: $\mu|\alpha \sim \text{IGamma}[\mu_0, \frac{\mu_0^2}{(r-1)}]$ $(r > 1)$; for known hyperparameters $\alpha = (\mu_0, r)$;

- Likelihood: $\bar{y}_0|\mu \sim \text{Gamma}[\mu, \mu^2/n]$;
- Posterior: $\mu|y_0, \alpha \sim \text{IGamma}[\mu_y^*, \frac{(\mu_y^*)^2}{(n+r-1)}]$ where $\mu_y^* = (1-W)\bar{y}_0 + W\mu_0$ and $W = \frac{r}{r+n}$.

(The square brackets '[ ]' refer to definition of the distribution by the first and second moments).

Using the analyst's gamma distribution assumption, we have the following posterior predictive quantities

$$\text{E}[\tilde{y}_i^{(1)}|\bar{y}_0, \alpha] = \text{E}[\text{E}[\tilde{y}_i^{(1)}|\mu]\big|\bar{y}_0, \alpha] = \text{E}[\mu\big|\bar{y}_0, \alpha] = \mu_y^* \ ;$$

and

$$\begin{aligned}
\text{Var}[\tilde{y}_i^{(1)}|\bar{y}_0, \alpha] &= \text{E}[\text{Var}[\tilde{y}_i^{(1)}|\mu]\,|\bar{y}_0, \alpha] + \text{Var}[\text{E}[\tilde{y}_i^{(1)}|\mu]\big|\bar{y}_0, \alpha] \\
&= \text{E}[\mu^2/n\big|\bar{y}_0, \alpha] + \text{Var}[\mu\big|\bar{y}_0, \alpha] \\
&= \frac{1}{n}\left\{\text{E}[\mu\big|\bar{y}_0, \alpha]\right\}^2 + \frac{n+1}{n}\text{Var}[\mu\big|\bar{y}_0, \alpha] \\
&= \frac{(\mu_y^*)^2}{n}\left(1 + \frac{n+1}{n+r-1}\right) \ .
\end{aligned}$$

Given our assumptions and applying large sample Bayesian arguments to condition on Z_{syn}, we have

$$\text{E}[\bar{y}_0\big|Z_{\text{syn}}] \approx \bar{q}_{\text{m}} - W\mu_0 \ ; \tag{4.20}$$

and

$$\mathrm{Var}[\bar{y}_0 | Z_{\mathrm{syn}}] \approx \bar{v}_{\mathrm{m}} \left(1 + \frac{n+1}{n+r-1}\right) . \tag{4.21}$$

Table 4.3: *Simulation study results - incorrect distribution assumption - lognormal distributed population data*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Distrib. assum.**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Avg.$\mathbf{V}_{\mathrm{syn}}$**	**CR**	**AW**
Obs.	SRS		LN	2.99	8.0×10^{-3}	8.0×10^{-3}	95.3	0.362
Syn.	SRS	V_{syn}	LN	2.99	8.0×10^{-3}	8.0×10^{-3}	94.2	0.367
		$V_{\mathrm{syn,alt}}$	LN	3.04	8.0×10^{-3}	8.0×10^{-3}	94.1	0.353
			Gam	2.99	8.0×10^{-3}	16×10^{-3}	99.5	0.501

Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1,000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1,000 replications
Avg.$\mathbf{V}_{\mathrm{syn}}$: Average value for the posterior variance estimate of Q, across 1,000 replications
CR: coverage rate
AW: average width

Results in Table 4.3 (which include the finite population correction factor), show that using synthetic data, analysis under the incorrect gamma distribution assumption produces a sampling variance estimate twice the observed-data estimate, and overcoverage for the nominal 95% confidence interval.

Next, we reverse the roles of the gamma and lognormal distributions; we draw the population from a gamma distribution but posit that the analyst analyzes the synthetic data as lognormally distributed.

Results in Table 4.4 show that, using synthetic data, the incorrect lognormal distribution assumption results in severe undercoverage for the nominal 95% confidence interval estimate, and moreover, the point estimate is biased. This example demonstrates that the correct distributional assumption by the analyst is crucial for valid

Table 4.4: *Simulation study results - incorrect distribution assumption - gamma distributed population data*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Distrib. assum.**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Avg.$\mathbf{V}_{\mathrm{syn}}$**	**CR**	**AW**
Obs.	SRS		Gam	2.99	8.0×10^{-3}	8.0×10^{-3}	95.3	0.362
Syn.	SRS	V_{syn}	Gam	2.99	8.0×10^{-3}	10×10^{-3}	96.4	0.389
		$V_{\mathrm{syn,alt}}$	Gam	2.96	8.0×10^{-3}	8.0×10^{-3}	94.7	0.348
			LN	3.78	13×10^{-3}	54×10^{-3}	0.0	0.910

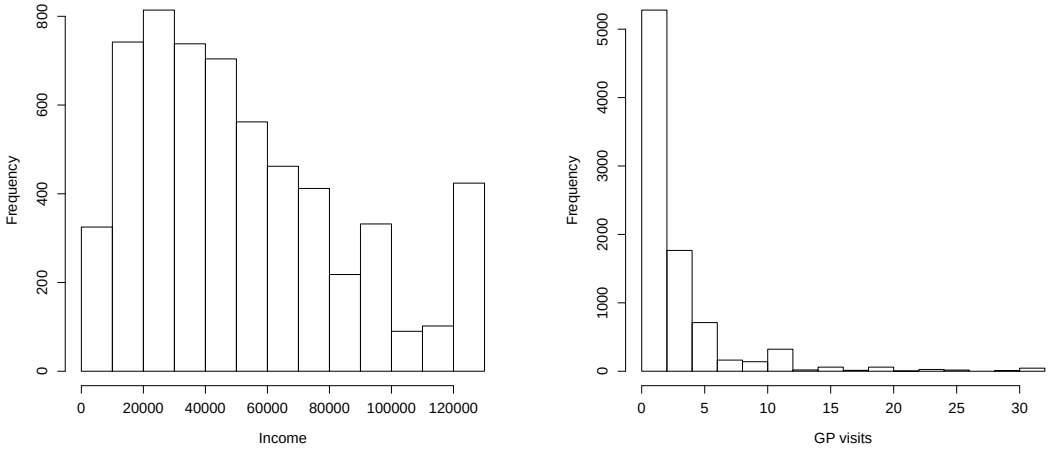
Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1,000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1,000 replications
Avg.$\mathbf{V}_{\mathrm{syn}}$: Average value for the posterior variance estimate of Q, across 1,000 replications
CR: coverage rate
AW: average width

inference when using the alternative approach. But if the correct distributional assumption is made, inferential results equivalent to the observed data set estimates are obtained using the $V_{\mathrm{syn,alt}}$ estimator.

4.5 Empirical study

For empirical evaluation of the alternative approach, we utilize data from the Joint Canada/United States Survey of Health (JCUSH) as described in Section 1.4. We investigate estimation of two quantities of interest: (i) mean annual household income; and (ii) mean number of general physician (GP) visits in a year per person. We chose these quantities because the distributions of the population data are skewed (skew(income)=0.78; skew(GP visits)= 3.27; see Figure 4.1) and the presence of outliers creates a disclosure risk. For the imputer, skewed data presents interesting modeling challenges because the popular linear regression model may not accurately describe the distribution of the observed variables. We investigated SRS and a two-

stratum resampling mechanism and analysis procedure, where the two strata are the US and Canada. Results are not shown for a nine-stratum synthetic data resampling method because we are interested in the effect of ignoring the original survey design, and the results would just be an extension of Example 4.1.1. We used two different imputation models: (A) conditional on region only; and (B) conditional on region, age, age^2, sex, race, marital status and education. Income was transformed by taking the cube root before imputation under a normal linear model to mitigate the skewness. The mean annual household population income is \$54,247. GP visits were modeled by a Poisson generalized linear model. The mean annual number of GP visits per person is 3.12. We used $m = 50$ imputations for each of 500 replications. We assessed the performance of the alternative approach by coverage of the nominal 95% confidence interval for each quantity of interest. Results are summarized in Tables 4.5 and 4.6.



(a) Income (b) GP visits

Figure 4.1: *Histograms of selected observed data in JCUSH, section 4.5*

For the estimation of mean income using synthetic data, interval estimates using

Table 4.5: *Empirical study results - estimation of mean(income)*

Data	**Samp. mech. & analysis proced.**	**Variance estimator**	**Avg.$\bar{q}_m$**	**Var.$\bar{q}_m$**	**Avg.V_{syn}**	**CR**	**AW**
Observed	9-STRS		54,207	4.61×10^5	4.61×10^5	95.8	2,661
Imputation model A							
Synthetic	SRS	V_{syn}	54,853	5.01×10^5	1.06×10^6	97.6	3,996
		$V_{syn,alt}$	54,856	5.01×10^5	5.54×10^5	95.0	2,917
	2-STRS	V_{syn}	54,490	4.55×10^5	1.02×10^6	98.8	3,908
		$V_{syn,alt}$	54,490	4.55×10^5	5.51×10^5	95.2	2,907
Imputation model B							
Synthetic	SRS	V_{syn}	54,833	3.96×10^5	8.16×10^5	94.2	3,479
		$V_{syn,alt}$	54,833	3.96×10^5	5.63×10^5	96.2	2,940
	2-STRS	V_{syn}	54,816	4.21×10^5	8.06×10^5	97.4	3,462
		$V_{syn,alt}$	54,816	4.21×10^5	5.63×10^5	94.4	2,940

Avg.$\bar{q}_m$: Average value for the posterior mean estimate of Q, across 500 replications
Var.$\bar{q}_m$: Variance of the posterior mean estimates for Q, across 500 replications
Avg.V_{syn}: Average value for the posterior variance estimate of Q, across 500 replications
CR: coverage rate
AW: average width

the $V_{syn,alt}$ estimator were closer to the observed data set results. There were no striking differences between results for the two imputation models studied, nor between results for the SRS and the 2-STRS resampling mechanism and analysis procedure.

For estimation of the mean number of GP visits, the synthetic data results appear unbiased, but the sampling variance (by both V_{syn} and $V_{syn,alt}$ estimators) was underestimated, producing coverage rates well below the 95% nominal level. Figure 4.2 compares histograms of GP visits imputed for one population and of the true population of values. We see the scale of the imputed population is reduced by about one-third relative to the true population. Although an essentially unbiased estimate of the mean was obtained, for estimation of scale-variant quantities data utility is compromised. The results in Table 4.6 indicate the true population distribution has

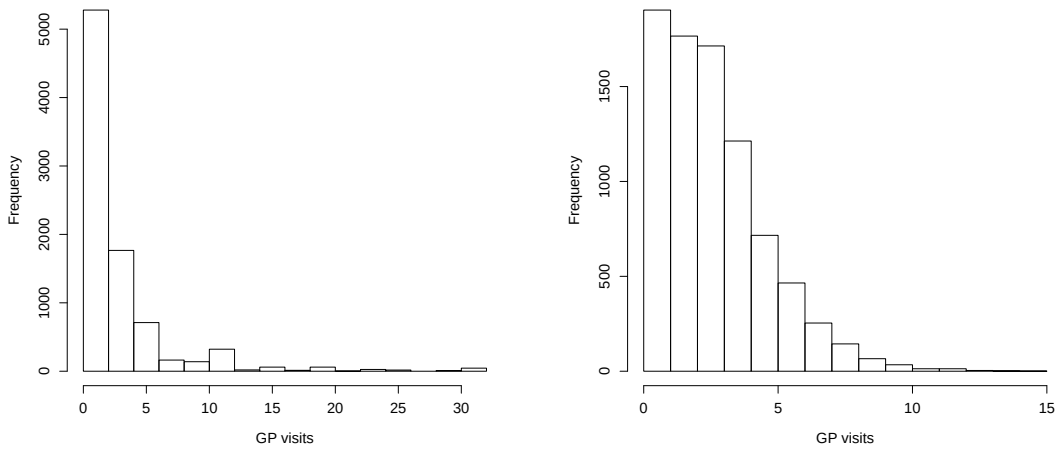
Table 4.6: *Empirical study results - estimation of mean(GP visits)*

Data	**Samp. mech. & analysis proced.**	**Variance estimator**	**Avg.$\bar{q}_m$**	**Var.$\bar{q}_m$**	**Avg.V_{syn}**	**CR**	**AW**
Observed	9-STRS		3.12	5.0×10^{-3}	5.0×10^{-3}	95.0	0.277
Imputation model A							
Synthetic	SRS	V_{syn}	3.13	5.0×10^{-3}	1.6×10^{-3}	69.6	0.156
		$V_{syn,alt}$	3.13	5.0×10^{-3}	1.0×10^{-3}	56.2	0.119
	2-STRS	V_{syn}	3.13	5.0×10^{-3}	1.6×10^{-3}	72.2	0.157
		$V_{syn,alt}$	3.13	5.0×10^{-3}	1.0×10^{-3}	61.4	0.118
Imputation model B							
Synthetic	SRS	V_{syn}	3.13	5.0×10^{-3}	1.8×10^{-3}	76.2	0.162
		$V_{syn,alt}$	3.13	5.0×10^{-3}	1.3×10^{-3}	66.4	0.141
	2-STRS	V_{syn}	3.13	5.0×10^{-3}	1.8×10^{-3}	75.2	0.163
		$V_{syn,alt}$	3.13	5.0×10^{-3}	1.3×10^{-3}	72.2	0.140

Avg.$\bar{q}_m$: Average value for the posterior mean estimate of Q, across 500 replications
Var.$\bar{q}_m$: Variance of the posterior mean estimates for Q, across 500 replications
Avg.V_{syn}: Average value for the posterior variance estimate of Q, across 500 replications
CR: coverage rate
AW: average width

not been correctly modeled, and that the imputation model requires improvement either by a different response distribution assumption, and/or conditioning on more predictors in the imputation model. On the other hand, for the protection of disclosure risk, this may be of benefit because true values of strong outliers are not released.

In this chapter we illustrated an alternative approach to fully synthetic data inference to recover the observed data set sampling distribution of sufficient statistics using synthetic data. The alternative approach requires equations for inference to be derived separately for every quantity of interest. The empirical data example also demonstrated the importance of an acceptable imputation model, that gives a representative indication of the distribution of the observed variables without releasing



(a) True population

(b) Imputed population

Figure 4.2: *Histogram comparison for GP visits - true and imputed population values*

confidential information, as the first priority before selection of an inference approach. The alternative approach is not presented as a replacement to the existing combining rules, but as a complementary tool. Practitioners are still advised to apply the existing inferential equations, but the alternative approach has shed some light on the role of the original survey design in the analysis of fully synthetic data, to satisfy congeniality.

Chapter 5

Application of Adjustment for Density Maximization to Sampling Variance Estimation in Fully Synthetic Data Inference

In chapter 2, Section 2.2, we reviewed the inferential combining rules for fully synthetic data (equations (2.1) - (2.4)) to estimate a scalar population quantity Q. It was noted that a disadvantage of the method-of-moments sampling variance estimator $V_{\text{syn}} = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_{\text{m}}$, is that it can be negative, (although asymptotically it is an unbiased estimator of the posterior quantity $\text{Var}(Q|Z_{\text{syn}})$). In order to obtain confidence intervals for Q, a positive variance estimate is required. Reiter (2002) and Reiter and Drechsler (2010) used slightly modified, more conservative estimates when negative variance estimates were calculated. Negative variance estimates may also

be avoided by choosing a large synthetic sample size (n_{syn}), or a large number of imputations (m) (Reiter 2002).

The adjustment for density maximization (ADM) procedure, proposed by Morris (1988), provides an alternative approach to sampling variance estimation. The ADM procedure multiples the posterior density of interest by an adjustment factor determined by a matching Pearson distribution, such that the posterior mean is estimated and not the posterior mode, and positive variance estimates are produced. ADM has been used to estimate shrinkage factors in a normal hierarchical model (Morris and Tang 2011), and to improve inferences of random effects in other multi-level models, as in Christiansen and Morris (1997) for a Poisson multi-level model. In this chapter, we propose using ADM for sampling variance estimation with fully synthetic data.

5.1 Background on ADM

This section provides a brief outline of ADM from Morris (1988), which justified using ADM for estimation of shrinkage factors in hierarchical models, so that there is zero probability of a shrinkage estimate outside the boundaries of the interval $[0, 1]$ (which implies non-zero and non-negative variance estimates), which is not guaranteed by standard maximum likelihood methods. Using ADM to produce positive variance estimates has not been theoretically justified, hence the investigation in this chapter is exploratory and requires substantial theoretical development to justify a principled method for variance estimation with synthetic data.

Let Y be a random variable which follows a Pearson distribution with mean parameter μ_0. The density of Y is

$$\pi(y) = K_W(\gamma, \mu_0)\exp\left(-\gamma \int \frac{y-\mu_0}{W(y)} dy\right) \frac{1}{W(y)} \ , \tag{5.1}$$

with respect to dy, y varying over an interval with $0 < W(y) < \infty$, where $W(y) = w_2 y^2 + w_1 y + w_0$, with w_0, w_1, and w_2 all known, and $K_W(\gamma, \mu_0)$ is the normalizing constant. The variance $\mathrm{Var}(y) = \frac{W(\mu_0)}{\gamma - w_2}$ is finite if $\gamma > w_2$, and is a quadratic function of the mean parameter μ_0. For fixed W, we can think of (5.1) as a two parameter distribution, denoted by

$$\mathrm{Pearson}(\gamma, \mu_0; W) = \mathrm{Pearson}\left[\mu_0, \frac{W(\mu_0)}{\gamma - w_2}\right] \ , \tag{5.2}$$

for unknown parameters μ_0 and γ. We define "Pearson measure" to be the density with respect to $dy/W(y)$.

For a unimodal density $f(y) > 0$, which is to be approximated by a $\mathrm{Pearson}(\gamma, \mu_0; W)$ density for specified W, let $l(y) = \ln(f(y)W(y))$. Then, with respect to Pearson measure $dy/W(y)$, $f(y)W(y)$ is a density and

$$f(y)W(y) = \exp(l(y)) \ . \tag{5.3}$$

We also express $f(y)W(y)$ as

$$\exp\left(-\gamma \int \frac{y-\mu_0}{W(y)} dy\right) \ , \tag{5.4}$$

by matching two derivatives of the logarithms at the modal value. Letting $\frac{dl(y)}{dy} = 0$, with y_0 the root of this derivative, then $\mu_0 = y_0$ and $\frac{d^2 l(y)}{dy^2} = \frac{\gamma}{W(y_0)}$, because the logarithm of (5.4) has first and second derivatives $-\gamma(y-\mu_0)/W(y)$ and $-\gamma W(y) +$

$\gamma(y-\mu_0)W'(y)/W^3(y)$. Given $f(y)$ and $W(y)$, one then chooses the Pearson$(\gamma, \mu_0; W)$ distribution with $\mu_0 = y_0$ and $\gamma = -\frac{d^2 l(y)}{dy^2} W(y_0)$, where $\frac{dl(y)}{dy} = 0$. We can then easily obtain the first and second moments of Y, which are

$$\mathrm{E}[Y] = \mu_0 \ , \qquad \mathrm{Var}[Y] = \frac{W(\mu_0)}{\gamma - w_2}. \tag{5.5}$$

That is, by maximizing the adjusted likelihood $f(y)W(y)$ with respect to Pearson measure, we obtain an estimate of the posterior mean of Y, and not the posterior mode, assuming a non-informative prior on μ_0. Hence, if the approximation Pearson$(\gamma, \mu_0; W)$ is chosen because the density of its range agrees with $f(y)$, we can produce parameter estimates in the domain of values for y where the density of $f(y)$ is positive, and not outside or on the boundary of this range.

5.2 Hierarchical framework for synthetic data inference

We now express the inferential methods for synthetic data derived in Raghunathan, Reiter and Rubin (2003) (and reviewed in Chapter 2), in a two-level hierarchical model framework. We apply the same assumptions as listed in Section 2.2. In particular, sample sizes (n and n_{syn}) are (a) large enough to permit normal approximations to posterior distributions and thus only the first two moments are required for each distribution, which can be derived using standard large sample Bayesian arguments; and (b) non-informative priors are assumed for all parameters, such that the information in the likelihood function dominates any information in the analyst's

prior distribution. Both are reasonable assumptions in large data sets, and distributions stated in the hierarchical framework are asymptotic distributions (denoted by the symbol $\sim$).

Our scalar population quantity of interest is Q. We assume a non-informative prior distribution for Q. The synthetic data sets are created by the agency as described in Section 2.1, and released to external analysts. Each synthetic data set is drawn from an imputed complete-data population $P_{\text{com}}^{(\text{l})}$. We do not draw repeated samples from the same $P_{\text{com}}^{(\text{l})}$. We define $Q^{(\text{l})}$ to be an unbiased point estimate of Q given $P_{\text{com}}^{(\text{l})}$, which the agency can calculate, but the analyst cannot because they do not have access to $P_{\text{com}}^{(\text{l})}$. From the collection of imputed populations, $P_{\text{com}} = (P_{\text{com}}^{(1)}, ..., P_{\text{com}}^{(\text{m})})$, the agency could calculate the following quantities:

$$\bar{Q}_{\text{m}} \equiv \sum_{l=1}^{\text{m}} \frac{Q^{(\text{l})}}{m} \ , \tag{5.6}$$

and

$$B_{\text{m}} \equiv \sum_{l=1}^{\text{m}} \frac{(Q^{(\text{l})} - \bar{Q}_{\text{m}})^2}{(m-1)} \ . \tag{5.7}$$

Using each synthetic data set $Z_{\text{syn}}^{(\text{l})}$, for $l = 1, \ldots, m$ and $m < \infty$, the analyst calculates the complete data statistic $q^{(\text{l})}$ as an unbiased point estimate of $Q^{(\text{l})}$, and $v^{(\text{l})}$ as an unbiased estimate of the sampling variance of $q^{(\text{l})}$. We assume negligible sampling variability in the estimates $v^{(\text{l})}$ across the collection of synthetic data sets Z_{syn}=$(Z_{\text{syn}}^{(1)}, ..., Z_{\text{syn}}^{(\text{m})})$. The analyst calculates the following inferential quantities:

$$\bar{q}_{\mathrm{m}} \equiv \sum_{l=1}^{\mathrm{m}} \frac{q^{(\mathrm{l})}}{m} \ , \tag{5.8}$$

$$b_m \equiv \sum_{l=1}^{\mathrm{m}} \frac{(q^{(\mathrm{l})} - \bar{q}_{\mathrm{m}})^2}{(m-1)} \ , \tag{5.9}$$

and

$$\bar{v}_{\mathrm{m}} \equiv \sum_{l=1}^{\mathrm{m}} \frac{v^{(\mathrm{l})}}{m} \ . \tag{5.10}$$

We assume independence between any pair of complete data statistics $(q^{(\mathrm{l})}, v^{(\mathrm{l})})$ and $(q^{(\mathrm{k})}, v^{(\mathrm{k})})$ for $l = 1, \ldots, m$, $k = 1, \ldots, m$, and $l \neq k$. This is justified because each synthetic data set $Z^{(\mathrm{l})}_{\mathrm{syn}}$ is sampled from an imputed population $P^{(\mathrm{l})}_{\mathrm{com}}$, which conditional on the observed data, is independent of any other imputed population $P^{(\mathrm{k})}_{\mathrm{com}}$. Combining the data across the m synthetic data sets, the hierarchical model is

$$\mathrm{Level-1:}\ \ \bar{q}_{\mathrm{m}} | \bar{Q}_{\mathrm{m}}, \bar{v}_{\mathrm{m}}, B_{\mathrm{m}} \sim N\left(\bar{Q}_{\mathrm{m}}, \frac{\bar{v}_{\mathrm{m}}}{m}\right) \ ; \tag{5.11}$$

$$\mathrm{Level-2:}\ \ \bar{Q}_{\mathrm{m}} | Q, \bar{v}_{\mathrm{m}}, B_{\mathrm{m}} \sim N\left(Q, \left(1 + \frac{1}{m}\right) B_{\mathrm{m}}\right) \ . \tag{5.12}$$

There is no b_m term in the Level-1 model (5.11) because each $q^{(\mathrm{l})}$ is an estimate for a different $Q^{(\mathrm{l})}$, $l = 1, .., m$; that is, we do not draw repeated samples from the same population $P^{(\mathrm{l})}_{\mathrm{com}}$, to calculate multiple $q^{(\mathrm{l})}$ estimates for a single $Q^{(\mathrm{l})}$, for fixed l. Also, because we assume there is negligible sampling variability in the estimates $v^{(\mathrm{l})}$, then $v^{(\mathrm{l})}$ and $\bar{v}_{\mathrm{m}}$ are interchangeable. The extra B_m/m in (5.12) is an adjustment required

for finite m because each imputed complete-data population $P_{\text{com}}^{(\text{l})}$ is generated from the infinite number of possible complete-data populations (Rubin 1987, pp.87-91).

It follows that the approximate marginal distribution of $\bar{q}_{\text{m}}$ is

$$\bar{q}_{\text{m}}|Q, \bar{v}_{\text{m}}, B_m \dot{\sim} N\left(Q, \frac{\bar{v}_{\text{m}}}{m} + \left(1 + \frac{1}{m}\right) B_m\right) \ , \tag{5.13}$$

and the approximate posterior distribution of Q is

$$Q|\bar{q}_{\text{m}}, \bar{v}_{\text{m}}, B_m \dot{\sim} N\left(\bar{q}_{\text{m}}, \frac{\bar{v}_{\text{m}}}{m} + \left(1 + \frac{1}{m}\right) B_m\right) \ , \tag{5.14}$$

assuming a non-informative prior on Q; that is, uniform measure on the positive real line $(-\infty, \infty)$. For the analyst, the quantity B_m is an unknown variance parameter and requires estimation given the synthetic data Z_{syn}. One approach is to use method-of-moments to derive the approximation $E[B_m|\bar{q}_{\text{m}}, \bar{v}_{\text{m}}, b_m] \approx b_m - \bar{v}_{\text{m}}$, as shown in Raghunathan, Reiter and Rubin (2003). Using this approach, we obtain the approximate posterior distribution of Q derived in Raghunathan, Reiter and Rubin (2003) (see (2.1) and (2.4)); that is,

$$Q|\bar{q}_{\text{m}}, \bar{v}_{\text{m}}, b_m \dot{\sim} N(\bar{q}_{\text{m}}, \left(1 + \frac{1}{m}\right) b_m - \bar{v}_{\text{m}}) \ , \tag{5.15}$$

and we can see that the variance estimate may be negative.

5.3 Using ADM for variance estimation with fully synthetic data

In this section, we derive an alternative variance estimator for fully synthetic data using ADM to produce a positive estimate. Specifically, we desire an estimate $\widehat{B_{\mathrm{m}}}$ using synthetic data, which we will obtain by maximizing a posterior distribution of B_m, approximated by the density of a Pearson distribution.

Although $\bar{Q}_{\mathrm{m}}$ and B_m can be calculated by the agency, for the analyst, $\bar{Q}_{\mathrm{m}}$ and B_{m} are unknown parameters and require estimation using statistics calculated from the synthetic data Z_{syn}. Using the standard one-way analysis of variance setup, conditional on B_m, the distribution of $\frac{(m-1)b_m}{(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})}$ is χ^2_{m-1}. Thus, the likelihood function of B_{m} is

$$L(B_{\mathrm{m}}) \propto (\bar{v}_{\mathrm{m}} + B_{\mathrm{m}})^{-(m-1)/2} \times \exp\left(-\frac{(m-1)b_{\mathrm{m}}}{2(\bar{v}_{\mathrm{m}} + B_{\mathrm{m}})}\right) . \tag{5.16}$$

To use ADM for variance estimation, firstly we need a prior density $\pi(B_{\mathrm{m}})$. Because we are assuming non-informative prior distributions for all parameters, we select the scale-invariant prior $\pi(B_m) \propto B_{\mathrm{m}}^{(c-1)}$, for known $c > 0$, and set $c = 1$ so that $B_{\mathrm{m}} \sim \mathrm{Unif}(0, \infty)$, and the posterior density of B_{m} is the same as the likelihood function for B_{m}.

There are three choices available for the Pearson distribution, namely the scaled Gamma, Inverse Gamma and F-distributions. To be consistent with Morris (1988), we choose the scaled Gamma distribution for our approximation. The adjustment factor with respect to Pearson measure is B_m (the variance function of the Gamma

distribution as a conjugate prior is linear in B_m). The adjusted posterior density for B_{m} using ADM is

$$\pi(B_{\mathrm{m}}\big|b_{\mathrm{m}},\bar{v}_{\mathrm{m}}) \propto B_{\mathrm{m}}(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})^{-(m-1)/2}\exp\left(-\frac{(m-1)b_{\mathrm{m}}}{2(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})}\right) \quad .$$

The adjusted log-posterior density is

$$l(B_{\mathrm{m}}) \propto \log B_{\mathrm{m}} - \frac{(m-1)}{2}\log(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}}) - \frac{(m-1)b_{\mathrm{m}}}{2(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})} \quad .$$

We wish to solve for $\widehat{B_m}$ that satisfies the equation

$$\frac{\partial l}{\partial B_{\mathrm{m}}} = \frac{1}{B_{\mathrm{m}}} - \frac{m-1}{2(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})} + \frac{(m-1)b_{\mathrm{m}}}{2(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})^2} = 0 \ , \tag{5.17}$$

which requires finding the roots of a quadratic equation. Rewriting (5.17) as

$$\frac{\partial l}{\partial B_{\mathrm{m}}} = \frac{-\left((m-3)B_{\mathrm{m}}^2 - \left[(m-1)(b_{\mathrm{m}}-\bar{v}_{\mathrm{m}})+4\bar{v}_{\mathrm{m}}\right]B_{\mathrm{m}} - 2\bar{v}_{\mathrm{m}}^2\right)}{2B_{\mathrm{m}}(\bar{v}_{\mathrm{m}}+B_{\mathrm{m}})^2} = 0 \ , \tag{5.18}$$

we see that for $m > 3$, the numerator of (5.18) is a convex quadratic function of B_{m}, which is negative at $B_{\mathrm{m}} = 0$, and therefore has two real roots. The positive root is the ADM estimator $\widehat{B}_{\mathrm{m,ADM}}$ (Morris and Tang 2011), which is

$$\widehat{B}_{\mathrm{m,ADM}} = \frac{((m-1)(b_{\mathrm{m}}-\bar{v}_{\mathrm{m}})+4\bar{v}_{\mathrm{m}}) + \sqrt{((m-1)(b_{\mathrm{m}}-\bar{v}_{\mathrm{m}})+4\bar{v}_{\mathrm{m}})^2 - 8(m-3)\bar{v}_{\mathrm{m}}^2}}{2(m-3)} \ , \tag{5.19}$$

where $\widehat{B}_{\mathrm{m,ADM}} > 0$, which we input into the variance equation from (5.14)

$$\mathrm{Var}(Q\big|Z_{\mathrm{syn}}) = V_{\mathrm{ADM}} \approx \frac{\bar{v}_{\mathrm{m}}}{\mathrm{m}} + \left(1 + \frac{1}{m}\right) \widehat{B}_{\mathrm{m,ADM}} \ ,$$

and hence we have a positive estimate for $\mathrm{Var}(Q\big|Z_{\mathrm{syn}})$, as opposed to the negative estimates which may be calculated with the variance estimate $\mathrm{Var}(Q\big|Z_{\mathrm{syn}}) = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_{\mathrm{m}}$ (5.15). However, V_{ADM} is not an unbiased estimate of B_m.

5.4 Evaluation of the ADM variance estimator

For continuity, our evaluation utilizes the same examples as in Chapter 4. If negative variance estimates are calculated using the method-of-moments estimator (2.4), we use the modified variance estimate $V^* = \max(0, V) + \delta \times (\frac{n_{\mathrm{syn}}}{n} \bar{v}_{\mathrm{m}})$, where $\delta = 1$ if $V = 0$ (Reiter 2002). In this section, we denote the method-of-moments estimator (2.4) as V_{MOM}, to clarify that we are evaluating the method-of-moments estimation approach with the proposed ADM-based estimator.

5.4.1 Simulation study I

Example 5.4.1

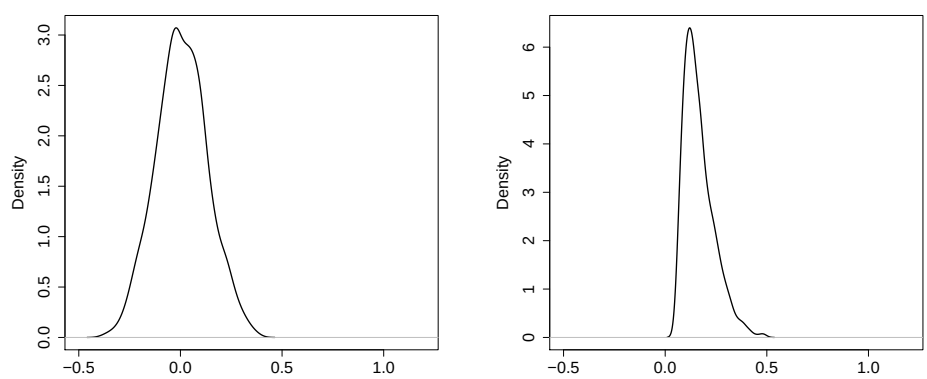
Refer to Example 4.1.1 for a description of the simulation setup for estimation of the population mean in a two-stratum normal population. Survey variables from stratum 1 are drawn from a $N(100, 1)$ distribution, and survey variables from stratum 2 are drawn from a $N(10, 1)$ distribution. Also note that the choice of the stratum mean parameters ensures a sizable proportion of negative variance estimates to assess

the ADM approach to sampling variance estimation. The observed stratum sample sizes are 1,500 units each, and $m = 200$ synthetic data sets are created.

Table 5.1: *Simulation study results - example 5.4.1*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**% V < 0**	**Avg.V**	**CR**	**AW**
Obs.	STRS		55.00	4×10^{-4}	0	4×10^{-4}	95.0	0.083
Syn.	SRS	V_{MOM}	55.00	9×10^{-4}	46	500×10^{-4}	99.3	2.46
		V_{ADM}	55.00	9×10^{-4}	0	161×10^{-4}	100	1.54
Syn.	STRS	V_{MOM}	55.00	5×10^{-4}	0	5×10^{-4}	94.6	0.083
		V_{ADM}	55.00	5×10^{-4}	0	5×10^{-4}	94.2	0.085

Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1000 replications
% V < 0: Percentage of posterior variance estimates that were negative
Avg.V: Average value for the posterior variance estimate of Q, across 1000 replications
CR: coverage rate
AW: average width



(a) V_{MOM} (b) V_{ADM}

Figure 5.1: *Example 5.4.1 - synthetic data variance estimate density plots (SRS resampling mechansim and analysis procedure).*

Simulation results are summarized in Table 5.1. The top half of the table is for results using synthetic data and a SRS resampling mechanism and analysis procedure.

We see that 46% of variance estimates were negative using the V_{MOM} estimator. In contrast, using ADM to obtain a variance estimate has to produce positive variance estimates. For both variance estimation approaches however, interval estimates for the population mean showed gross overcoverage. The 100% coverage rate arises from ignoring the more complex original survey design when resampling and analyzing the synthetic data. This research problem was addressed in Chapter 3. We conclude based on efficiency criteria, that the V_{MOM} estimator does no better or worse than the ADM approach to variance estimation.

We can see in Figure 5.1 the benefits of using an ADM approach to variance estimation when there is a large proportion of negative variance estimates using the V_{MOM} estimator. Approximately half of the area under the density plot for V_{MOM} (Figure 5.1 (a)) lies in the domain of values $V_{\text{MOM}} < 0$. In contrast, the density plot for V_{ADM} (Figure 5.1 (b)) has positive density in the domain of values $V_{\text{ADM}} > 0$, with zero density for values $V_{\text{ADM}} < 0$.

For completeness, we also show in Table 5.1 the simulation results when the synthetic data are resampled and analyzed according to the original survey design (STRS). Negative variances are not a problem for the V_{MOM} estimator. The ADM-based confidence interval is just slightly larger, consistent with the overestimation in B_m required to produce positive variance estimates. We conclude all variance estimation approaches do equally well to approximate the observed data set results.

5.4.2 Simulation study II

Example 5.4.2

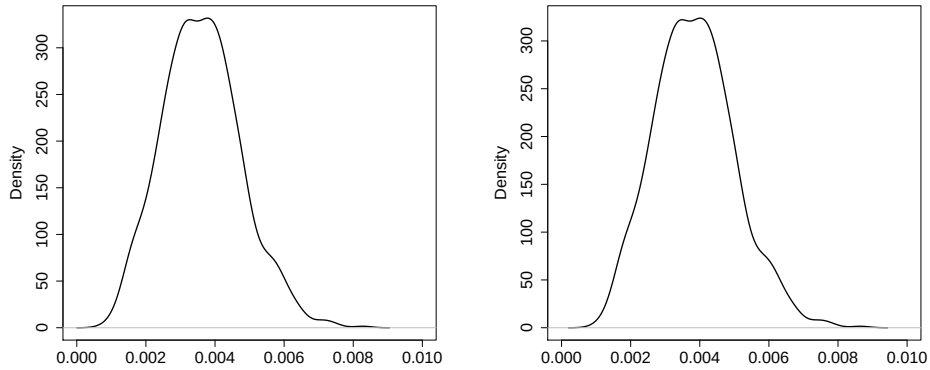
Refer to Example 4.3.1 for a description of the simulation setup for estimation of a population proportion where the observed data set is sampled by πPS without replacement. The observed data set sample size is $n = 100$, and $m = 200$ synthetic data sets are created. Results are summarized in Table 5.2.

Table 5.2: *Simulation study results - example 5.4.2*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**Var.$\bar{\mathbf{q}}_{\mathrm{m}}$**	**% V < 0**	**Avg.V**	**CR**	**AW**
Obs.	πPS	design	0.626	3.2×10^{-3}	0	3.2×10^{-3}	96.6	0.258
		model	0.626	3.2×10^{-3}	0	3.2×10^{-3}	97.5	0.223
Syn.	SRS	V_{MOM}	0.621	2.4×10^{-3}	0	3.5×10^{-3}	97.5	0.230
		V_{ADM}	0.621	2.4×10^{-3}	0	3.8×10^{-3}	97.9	0.238
Syn.	πPS	V_{MOM}	0.624	2.5×10^{-3}	0	3.5×10^{-3}	96.0	0.226
		V_{ADM}	0.624	2.5×10^{-3}	0	3.9×10^{-3}	96.9	0.237

Avg.$\bar{\mathbf{q}}_{\mathrm{m}}$: Average value for the posterior mean estimate of Q, across 1000 replications
Var.$\bar{\mathbf{q}}_{\mathrm{m}}$: Variance of the posterior mean estimates for Q, across 1000 replications
% V < 0: Percentage of posterior variance estimates that were negative
Avg.V: Average value for the posterior variance estimate of Q, across 1000 replications
CR: coverage rate
AW: average width

The V_{MOM} estimator did not produce negative variance estimates in this simulation example. Both V_{MOM} and V_{ADM} estimators did equally well to approximate the observed data set results. The variance estimate density plots in Figure 5.2 appear identical to each other, and all plots have positive density in the domain of values $V_{\mathrm{MOM}} > 0$ (or $V_{\mathrm{ADM}} > 0$), with zero density for values $V_{\mathrm{MOM}} < 0$ (or $V_{\mathrm{ADM}} < 0$).



(a) V_{MOM} (b) V_{ADM}

Figure 5.2: *Example 5.4.2 - synthetic data variance estimate density plots (SRS resampling mechansim and analysis procedure).*

5.4.3 Empirical study

For our empirical evaluation, we utilize the Joint Canada/United States Survey of Health (JCUSH) as described in Section 1.4, and as used in the empirical evaluation in Section 4.5. To recap, we investigate estimation of two quantities of interest (i) mean annual household income; and (ii) mean number of general physician (GP) visits in a year per person. We chose these variables because the distributions of the assumed population data are skewed (skew(income)=0.78; skew(GP visits)= 3.27; see Figure 4.1), and hence, the ADM approach is of potential use to produce positive variance estimates. The mean annual household population income is $54,247. The mean annual number of GP visits is 3.12 per person. The original survey is a nine-stratum survey design. The imputation models are conditional on the predictors strata indicator, age, age^2, gender, race, marital status and education. Income is transformed by taking the cube root before imputation under a normal linear model

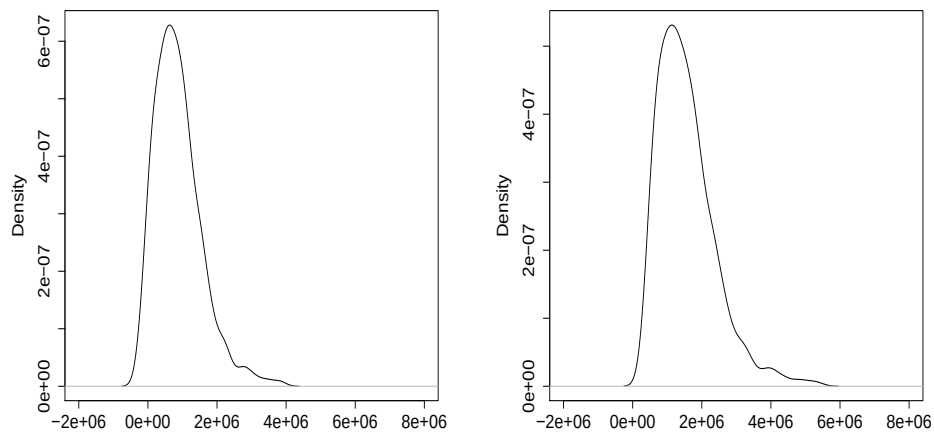
to mitigate the skewness. GP visits are modeled by a poisson generalized linear model. We use simple random sampling to resample and analyze the synthetic data. Results are not shown for a nine-stratum synthetic data sampling design because negative variance estimates are not a problem in this case. We used $m = 10$ and $m = 50$ imputations, across 500 replications each. We assessed the performance of the ADM variance estimates by coverage of their nominal 95% confidence intervals for each quantity of interest.

Table 5.3: *Empirical study results - estimation of mean(income)*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Avg.$\bar{q}_m$**	**Var.$\bar{q}_m$**	**% V < 0**	**Avg.V**	**CR**	**AW**
Obs.	9-STRS		54,206	4.61×10^5	0	4.61×10^5	95.8	2,661
$m = 10$								
Syn.	SRS	V_{MOM}	54,800	5.04×10^5	7.4	9.24×10^5	99.8	7,165
		V_{ADM}	54,800	5.04×10^5	0	1.56×10^6	100	10,304
$m = 50$								
Syn.	SRS	V_{MOM}	54,853	3.96×10^5	0	8.16×10^5	94.2	3,479
		V_{ADM}	54,853	3.96×10^5	0	8.96×10^5	96.8	3,665

Avg.$\bar{q}_m$: Average value for the posterior mean estimate of Q, across 500 replications
Var.$\bar{q}_m$: Variance of the posterior mean estimates for Q, across 500 replications
% V < 0: Percentage of posterior variance estimates that were negative
Avg.V: Average value for the posterior variance estimate of Q, across 500 replications
CR: coverage rate
AW: average width

In this empirical example, negative variance estimates were a problem using the V_{MOM} estimator when the number of imputations was $m = 10$. Specifically, 7.4% of variance estimates were negative for estimation of mean income, and 4.6% of variance estimates were negative for estimation of mean GP visits. The negative variance problem was resolved using $m = 50$, or using the ADM approach to variance estima-



(a) V_{MOM} (b) V_{ADM}

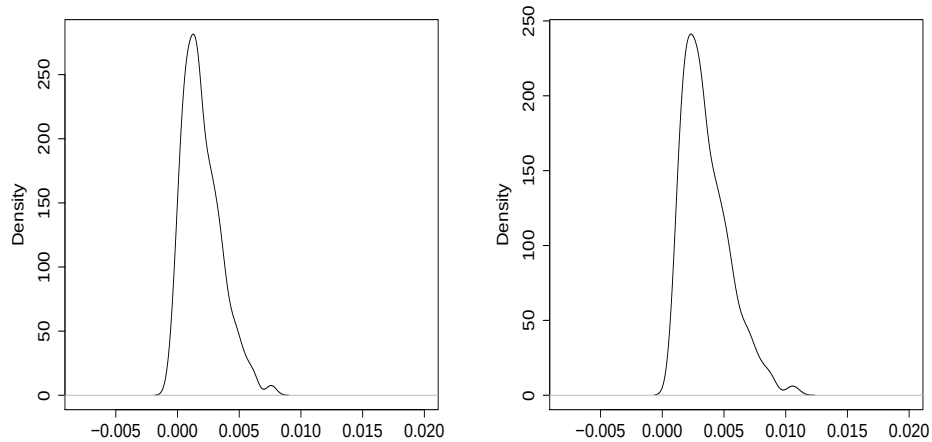
Figure 5.3: *Variance estimate density plots for* ***mean(income)***

Table 5.4: *Empirical study results - estimation of mean(GP visits)*

Data	**Samp. mech. & analysis proced.**	**Var. estim.**	**Avg.$\bar{\mathbf{q}}_{\text{m}}$**	**Var.$\bar{\mathbf{q}}_{\text{m}}$**	**% V < 0**	**Avg.V**	**CR**	**AW**
Obs.	9-STRS		3.12	5.0×10^{-3}	0	5.0×10^{-3}	95.0	0.277
$m = 10$								
Syn.	SRS	V_{MOM}	3.13	5.0×10^{-3}	4.6	2.1×10^{-3}	97.6	0.349
		V_{ADM}	3.13	5.0×10^{-3}	0	3.6×10^{-3}	99.6	0.503
$m = 50$								
Syn.	SRS	V_{MOM}	3.13	5.0×10^{-3}	0	1.8×10^{-3}	76.2	0.162
		V_{ADM}	3.13	5.0×10^{-3}	0	2.0×10^{-3}	75.0	0.172

Avg.$\bar{\mathbf{q}}_{\text{m}}$: Average value for the posterior mean estimate of Q, across 500 replications
Var.$\bar{\mathbf{q}}_{\text{m}}$: Variance of the posterior mean estimates for Q, across 500 replications
% V < 0: Percentage of posterior variance estimates that were negative
Avg.V: Average value for the posterior variance estimate of Q, across 500 replications
CR: coverage rate
AW: average width

tion. The overcoverage in interval estimates for $m = 10$ is due to ignoring the original survey design and a small number of imputations. For the estimation of mean in-



(a) V_{MOM} (b) V_{ADM}

Figure 5.4: *Variance estimate density plots for* ***mean(GP visits)***

come, coverage rates are reduced to the nominal 95% level using $m = 50$ imputations, and both V_{MOM} and V_{ADM} interval estimates were approximately equivalent to the observed data set results. For estimation of mean GP visits, the interval estimates show severe undercoverage using $m = 50$ imputations, for all variance estimators. This is likely due to an inadequate imputation model, as discussed in Section 4.5. The effects of an inadequate imputation model to capture accurately the true target population distribution are hidden by the small number of imputations when $m = 10$.

This chapter has demonstrated an application of ADM to achieve positive variance estimates when analyzing fully synthetic data, which is not guaranteed by the method-of-moments estimator to approximate $\text{Var}[Q|Z_{\text{syn}}]$. A key feature of ADM is to overestimate the sampling variance to produce a positive (but biased) estimate. Hence, even if applied when not required (and assuming an acceptable imputation model), underestimation of variance will not be a problem, but interval estimates will

be wider than those produced by the V_{MOM} estimator. As mentioned at the beginning of Section 5.1, the work in this chapter is preliminary, because the theoretical justification to use ADM for variance estimation has not been addressed. In addition, further work is required to investigate the sensitivity of results to informative prior distributions for $\pi(B_m)$. These steps are required before we can promote usage of ADM for variance estimation with synthetic data as a principled approach.

Chapter 6

Partial Synthesis of a Large-Scale Healthcare Study

The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium is a large-scale healthcare study of services and outcomes of care delivered to population and healthcare-system-based cohorts of newly diagnosed patients with lung and colorectal cancer (Ayanian et al. 2004). The study is administered across 11 study sites in the US, in multiple waves. Patients are surveyed by telephone and information is gathered on the care received during different stages of illness, clinical and patient-reported outcomes, and patient preferences and behaviors. Additional data are obtained from physician surveys and medical records. The wide scope of data collection and numerous measurements provide opportunities for research on multiple topics, both by members of the Consortium and by external investigators.

As discussed in Chapter 1, several statistical agencies have begun to use partially synthetic approaches to create public-use data for major surveys, but application of

synthetic data techniques to large-scale healthcare survey data is rare. Disclosure control for CanCORS data entails some interesting challenges and innovations. Firstly, the presence of clinical variables requires that any modification to the observed data set must produce variable relationships that are clinically feasible. Secondly, the external analysts of CanCORS data are likely to be clinicians and health services researchers. It is important to demonstrate that these analysts can obtain valid statistical inferences by running typical healthcare analyses on synthetic data. Thirdly, CanCORS is a large sample from a small, well-defined target population. The target population size in the first wave of the study was approximately 15,000 patients for lung cancer, and 12,000 patients for colorectal cancer. After identification of appropriate samples, approximately 5,000 patients were surveyed for each cancer type. This is a sampling fraction of between 30% and 40%, and thus disclosure risk protection by random sampling is limited in this study, because there is more certainty that a unit in the target population will be sampled for inclusion in the observed data set.

Partially synthetic data sets (Reiter 2003) consist of a mix of synthetic values for sensitive variables or key identifiers, and the originally observed values for all other data points. In contrast to fully synthetic data, not all values are replaced by imputations; hence we maintain the benefits of fully synthetic data to protect confidentiality, but with decreased sensitivity to the specification of the imputation models. The inferential methods for partially synthetic data derived in Reiter (2003) are reviewed in the next section.

6.1 Background on partially synthetic data

6.1.1 Inference with partially synthetic data

Let $S_i = 1$ if unit i is selected to have any of its observed values replaced, and let $S_i = 0$ otherwise. Let $S_0 = (S_1, \ldots, S_n)$, where n is the number of records in the observed data set. Let $Y_0 = (Y_{\text{rep}}, Y_{\text{nrep}})$ be the data collected in the original survey, where Y_{rep} includes all values to be replaced with multiple imputations, and Y_{nrep} includes all values not replaced with imputations. Let $Y_{\text{rep}}^{(\text{l})}$ be the replacement values for Y_{rep} in synthetic data set l, for $l = 1, .., m$. Each $Y_{\text{rep}}^{(\text{l})}$ is generated by simulating values from the posterior predictive distribution $\pi(Y_{\text{rep}}^{(\text{l})}|Y_0, S_0)$, or some close approximation to the distribution such as those of Raghunathan et al. (2001). The agency repeats the process m times creating synthetic data sets $Z_{\text{syn}}^{(\text{l})} = (Y_{\text{nrep}}, Y_{\text{rep}}^{(\text{l})})$, for $l = 1, ..., m$, and releases the collection of synthetic data sets $Z_{\text{syn}} = (Z_{\text{syn}}^{(1)}, \ldots, Z_{\text{syn}}^{(\text{m})})$ to the public. We refer to the agency as the 'imputer', and in the context of CanCORS, this refers to the Statistical Consulting Center (SCC). Investigators outside the Consortium take the role of the 'analysts' or 'public-users'.

To obtain valid inference for a scalar estimand Q, analysts can use the combining rules presented by Reiter (2003). Suppose that given the original data set, the analyst would estimate Q with some point estimate q_0, and the sampling variance of q_0 with some estimate v_0. Let $q^{(\text{l})}$ and $v^{(\text{l})}$ be the complete data estimates from synthetic data set $Z_{\text{syn}}^{(\text{l})}$. The analyst computes $q^{(\text{l})}$ and $v^{(\text{l})}$ by acting as if each $Z_{\text{syn}}^{(\text{l})}$ is the observed data set.

The point estimate of Q given the synthetic data is $\bar{q}_{\text{m}} = \sum_{l=1}^{m} q^{(\text{l})}/m$. The

estimated variance of $\bar{q}_{\mathrm{m}}$ is $T_p = b_m/m + \bar{v}_{\mathrm{m}}$, where $b_m = \sum_{l=1}^{m}(q^{(l)} - \bar{q}_{\mathrm{m}})^2/(m-1)$ and $\bar{v}_{\mathrm{m}} = \sum_{l=1}^{m} v^{(l)}/m$. To understand how the variance estimator T_{p}, differs from the variance estimator for fully synthetic data V_{syn} (2.4), consider the case when $m = \infty$. Because the same, original units are released in each synthetic data set, the quantity $\bar{v}_\infty$ is by itself an estimate of $\mathrm{Var}(Q|Z_{\mathrm{syn}})$. For $m < \infty$, we replace $\bar{v}_\infty$ with $\bar{v}_{\mathrm{m}}$, and we add b_m/m for the additional variance due to the use of a finite number of imputations. Inferences for scalar Q, when n is large, can be based on a t-distribution with degrees of freedom $\nu_{\mathrm{m}} = (m-1)(1 + r_{\mathrm{m}}^{-1})^2$, where $r_{\mathrm{m}} = (m^{-1}b_{\mathrm{m}}/\bar{v}_{\mathrm{m}})$. Methods for multivariate inferences are derived in Reiter (2005d).

6.1.2 Disclosure risk

Disclosure risk can be defined as the risk of identification of sampled units in the released data. To compute probabilities of identification, we use the Duncan-Lambert risk framework (Duncan and Lambert (1989)), with related approaches in Fienberg, Makov and Sanil (1997), Reiter (2005c), and Reiter and Mitra (2009). Under this framework, we mimic the behavior of an ill-intentioned public user (hereon referred to as the intruder), who possesses the true values of unique or quasi-identifiers for select target units, and seeks to identify the records in the synthetic data that have matching identifier values.

Our key modeling assumptions are:

(a) The intruder knows the target is in the survey and the identifiers of all units in the population.

(b) We investigate 3 sets of identifying variables to illustrate the variation in dis-

closure risk from varying assumed levels of intruder information:

(i) *Set 1*: Age, sex, marital status, race

(ii) *Set 2*: Set 1 + education+ income level

(iii) *Set 3*: Set 2 + disease stage + study site

Assumption (a) is justified because CanCORS is a large survey from a small, well-defined target population. That is, there is limited disclosure control benefit from random sampling (Duncan and Lambert 1989, Fienberg, Makov and Sanil 1997, and Reiter 2005c). The data vectors listed in (b) do not constitute unique identifiers (as would name, address, date of birth, case ID), but are the best substitutes for unique identifiers from the variables available for public release, hence the term quasi-identifiers. A potential identification risk for target record i occurs when its true quasi-identifier values match the corresponding values for a synthetic data set record k ($k = 1, .., n$). The risk is potential because there will be other records in the synthetic data with the same set of identifying variables, either true or synthesized, such that the intruder does not know if a match is the correct match or not. Furthermore, the set of synthetic data records which match the identifying information for a given target unit will vary across the released synthetic data sets. We assume the intruder assigns equal probability of being the correct match to each synthetic data record identified as a potential match for a given target unit.

Let $R_{\text{ilk}} = 1$ (for $i = 1, .., n; l = 1, .., m; k = 1, .., n$) if the quasi-identifying information for original unit i, matches the quasi-identifying information of unit k, in synthetic data set l. Define $F_{\text{il}} = \sum_{k=1}^{n} R_{\text{ilk}}$, to be the number of records in synthetic

data set l, that match the quasi-identifying information of original unit i. Define the following risk measures:

1. The *maximum number of matches* across the collection of synthetic data sets is

$$MXM = \sum_{i=1}^{n} \frac{1}{m} \sum_{l=1}^{m} C_{\text{il}} \ , \tag{6.1}$$

   where $C_{\text{il}} = 1$ when record i is among the F_{il} matches from synthetic data set $Z_{\text{syn}}^{(\text{l})}$, and $C_{\text{il}} = 0$ otherwise. The maximum number of matches is reached if the intruder always correctly selects the target from the set of potential candidates.

2. For unit i, the *expected match risk* is

$$EMR_i = \frac{1}{m} \sum_{l=1}^{m} \frac{1}{F_{\text{il}}} \times C_{\text{il}} \ , \tag{6.2}$$

   The contribution of unit i to the expected match risk reflects the intruder randomly guessing at the correct match from the F_{il} candidates.

   The overall expected match risk rate is

$$EMR = \frac{1}{n} \sum_{i=1}^{n} EMR_i \ . \tag{6.3}$$

3. For unit i, the *presumed true match risk* is

$$TMR_i = \frac{1}{m} \sum_{l=1}^{m} K_{\text{il}} \ , \tag{6.4}$$

   where $K_{\text{il}} = 1$ when $F_{\text{il}} = 1$ and $C_{\text{il}} = 1$, and $K_{\text{il}} = 0$ otherwise.

The overall true match risk is

$$TMR = \frac{1}{n}\sum_{i=1}^{n} TMR_i \ , \tag{6.5}$$

and reflects the intruder correctly and uniquely identifying records, averaging over the collection of synthetic data sets.

6.1.3 Data utility

In a broad sense, the utility of a particular data release is the benefit to society of the released information (Woo et al. 2009). A more quantitative definition might characterize what can be learned from the synthetic data, relative to what can be learned from the observed data set. Such comparisons can be tailored to specific analyses, or can be broadened to global differences in distributions. Karr et al. (2006) discuss measures of data utility for specific estimands using confidence interval overlap. Woo et al. (2009) introduce some global measures of data utility using propensity scores, cluster analysis, and empirical distribution estimation. The confidence interval overlap measure is commonly cited in the literature, and is defined in Karr et al. (2006) as follows.

For an estimate q, the average overlap is calculated by:

$$J_q = \frac{1}{2}\left(\frac{U_{\text{over,q}} - L_{\text{over,q}}}{U_{\text{0,q}} - L_{\text{0,q}}} + \frac{U_{\text{over,q}} - L_{\text{over,q}}}{U_{\text{syn,q}} - L_{\text{syn,q}}}\right) \ , \tag{6.6}$$

where ($L_{\text{0,q}}$, $U_{\text{0,q}}$) denote the lower and the upper bound of the confidence interval for the estimate q using the observed data set, and similarly ($L_{\text{syn,q}}$, $U_{\text{syn,q}}$) using the

synthetic data, and ($L_{\text{over,q}}$, $U_{\text{over,q}}$) denote the intersection of these intervals. High values of overlap ($0.9 \leq J_q \leq 1$) are favored over low values. A low overlap could be the result of wide or poorly centered synthetic data confidence intervals.

Although the synthetic data estimate $\bar{q}_{\text{m}}$ is an asymptotically unbiased estimate of q_0, the actual value computed from the synthetic data generated will deviate from q_0. Define $Bias_q = \bar{q}_{\text{m}} - q_0$ to be the bias of the synthetic data estimate from the observed data set estimate, and V_{q} is the sampling variance estimate of $\bar{q}_{\text{m}}$ given the synthetic data. Define $Bias_q/\sqrt{V_{\text{q}}}$ to be the standardized bias. Following Cochran (1977) Section 1.8, p. 13, we can compute the effect of bias on the coverage of a nominal 95% confidence interval.

The upper tail error probability is given by

$$P_{\text{U,err}} = \frac{1}{\sqrt{2\pi}} \int_{1.96-(Bias_{\text{q}}/\sqrt{V_{\text{q}}})}^{\infty} e^{-t^2/2} dt \ , \tag{6.7}$$

and the lower tail error probability is given by

$$P_{\text{L,err}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1.96-(Bias_{\text{q}}/\sqrt{V_{\text{q}}})} e^{-t^2/2} dt \ . \tag{6.8}$$

The total error probability is $P_{\text{Tot,err}} = P_{\text{L,err}} + P_{\text{U,err}}$, which should be 0.05 for a nominal 95% confidence interval estimate.

In this thesis, we compare descriptive statistics of the observed and synthetic CanCORS data, and compare estimates for models based on two published analyses of the observed CanCORS data set. Because we expect the data utility results for non-synthesized variables to be better than those for synthesized variables, we also

summarize results by this grouping.

6.2 Application to the CanCORS patient survey data set

6.2.1 Identification of high disclosure risk variables

For the protection of data confidentiality, high disclosure risk and sensitive variables should potentially be synthesized. Identifying all the variables that pose disclosure risk is an important and labor intensive task, and the research literature on strategies to identify high disclosure risk variables is limited. The general strategy adopted for our study is discussed below and is based on the strategy in Hawala (2008), with additional consideration for the clinical variables in our healthcare study.

We define high disclosure risk variables to be those (either individually or in combination) with two characteristics: (i) an intruder is likely to have an external data source containing them; and (ii) they can uniquely identify an individual when matched against the intruders' external information source. The observed CanCORS data set for synthesis already excluded the variables: patient name, address, and exact age. From the remaining variables, we identified the demographic variables age (grouped into 5 year age bands), sex, education, marital status, and race as posing high disclosure risk.

Clinical variables in combination with demographic variables also pose high disclosure risk. However, we chose to retain their original values to preserve clinical relationships, and eliminate the need to impute with skip patterns determined by

the structure of the patient survey. We included clinical variables in our disclosure risk assessment, however, as components of the quasi-identifying sets of information (Section 6.1.2).

We defined sensitive variables as those whose disclosure, if attributed to the correct individual, would be considered a breach of data confidentiality for the respondent. All information collected can be considered sensitive from the perspective of the respondent, in particular, health insurance details, grouped income levels and medical history.

To summarize the above qualitative definitions, we have four possible types of variables:

(i) High disclosure risk and sensitive

(ii) High disclosure risk and non-sensitive

(iii) Low disclosure risk and sensitive

(iv) Low disclosure risk and non-sensitive

Our focus is on minimizing the disclosure risk from variables in groups (i) and (ii). If a sensitive variable in group (iii) cannot be attributed to an identifiable target, data confidentiality is not threatened.

6.2.2 Imputation models

Imputation models were run separately for lung and colorectal cancer because there might be different relationships among clinical and demographic variables depending on the cancer type. The observed data sets each had > 500 variables. The number of variables to be included in the partially synthetic, public data was reduced to ≈ 300 based on the following exclusions:

- Categorical variables whose responses are highly concentrated ($> 95\%$) in one category.
- Variables containing names or addresses of people or places.
- Variables on consumption of specific types of alternative therapy, vitamins and herbal supplements, but retain the general indicator for usage of these services/products.
- Variables not associated with, or a consequence of the active patient treatment plan, such as recollection of symptoms at notification of cancer.

The first two criteria protect data confidentiality prior to application of any statistical disclosure control method. The third and fourth criteria were applied to avoid issues with multicollinearity, such as including indicators for both symptoms experienced at diagnosis, and in the last 4 weeks.

Apart from sex, which is binary, all variables were defined as unordered categorical variables for imputation. We simplified the structure of the categorical variables to be synthesized by reducing the number of levels (see Appendix A, Table A.1) to avoid sparse cell counts, and thereby ensure stable parameter estimates in the model fit

required for the imputation models. The recoded structure is also akin to that used commonly in analyses.

We chose a parametric approach to imputation. Imputation for the binary variable, sex, was based on a logistic regression model. Imputations for all other variables were based on multinomial logit models. Our imputation models are consistent with those for similar demographic variables synthesized in previous studies (see Reiter 2005a, Kinney and Reiter 2007, Drechsler and Reiter 2008, and Reiter and Mitra 2009). Non-parametric approaches to imputation of synthetic data have been studied, including classification and regression trees (CART) (Reiter 2005b), and random forests (Caiola and Reiter 2010). There is no published literature to favor one approach over the other, and evaluation of a non-parametric approach to create partially synthetic data for CanCORS is an area for future work.

Using a parametric approach, we need to specify the set of predictors to impute each variable. To try to capture all variable relationships of importance to the public user, a large number of predictors are desired in each imputation model. However, fitting ≈ 300 predictor variables in each imputation model is impractical. To select the set of predictors for each imputation model, a stepwise regression was conducted within each of the 12 sections of the administered survey questionnaire. Within each section, variables were added and dropped until Akaike's Information Criterion (AIC) was minimized. We ran the stepwise regression by section to include the important predictor variables from each section. This variable reduction procedure resulted in an average of ≈ 50 predictor variables for each imputation model. To avoid the potential for over fitting which increases disclosure risk, no interactions were included

in the imputation models, only main effects. The baseline model for each stepwise regression always included the variables survey version code, cancer stage, cancer histology, vital status at time of interview (alive/dead/unknown) and PDCRID site, because these variables are generally used as control variables in analyses. We did not consider variable selection approaches for high-dimensional data, such as the LASSO (Tibshirani 1996) or SCAD (Fan and Li 2001), because we do not anticipate that external analysts of CanCORS data would use high-dimensional data analysis methods.

Synthetic data sets are drawn from the posterior predictive distributions for each variable using the sequence of conditional distributions specified below. The notation $Y_{\text{nrep}}^{(*)}$ denotes the predictor variables selected by stepwise regression for the variable to be imputed. The notation X denotes the variables conditioned upon in all models.

1. Impute sex using a logistic model to draw from

$$\pi(Y_{\text{rep}}^{(\text{sex})} \big| Y_{\text{nrep}}^{(\text{sex})}, X) \ .$$

2. Impute race using a multinomial model to draw from

$$\pi(Y_{\text{rep}}^{(\text{race})} \big| Y_{\text{rep}}^{(\text{sex})}, Y_{\text{nrep}}^{(\text{race})}, X) \ .$$

3. Impute marital status using a multinomial model to draw from

$$\pi(Y_{\text{rep}}^{(\text{marstat})} \big| Y_{\text{rep}}^{(\text{race})}, Y_{\text{rep}}^{(\text{sex})}, Y_{\text{nrep}}^{(\text{marstat})}, X) \ .$$

4. Impute education using a multinomial model to draw from

$$\pi(Y_{\text{rep}}^{(\text{educ})} \big| Y_{\text{rep}}^{(\text{marstat})}, Y_{\text{rep}}^{(\text{race})}, Y_{\text{rep}}^{(\text{sex})}, Y_{\text{nrep}}^{(\text{educ})}, X) \ .$$

5. Impute age using a multinomial model to draw from

$$\pi(Y_{\text{rep}}^{(\text{age})} \big| Y_{\text{rep}}^{(\text{educ})}, Y_{\text{rep}}^{(\text{marstat})}, Y_{\text{rep}}^{(\text{race})}, Y_{\text{rep}}^{(\text{sex})}, Y_{\text{nrep}}^{(\text{age})}, X) \ .$$

Other possible orderings of the conditional distributions are possible.

For each posterior predictive distribution, we first (i) draw values of the model parameters from their posterior distribution, or an approximation to the distribution given the observed data set; and second (ii) generate synthetic values given the drawn values of the parameters and the selected predictor variables given the results of the stepwise regression. Non-informative prior distributions were assumed for all parameters. For full technical details on methods used to draw parameters and synthetic values, refer to Reiter (2005a), Appendix B pp. 203-204. We used the software package for multiple imputation 'mi' (Su et al. 2011), in the *R* software and computing environment to run our imputation models. We ran into some computational constraints when imputing the multinomial variables due to dimensionality or scarcity. To avoid this issue, we used the Gaussian based routines for categorical variable imputation in health surveys developed in Yucel, He and Zaslavsky (2011), to make use of the existing functionality in the 'mi' package for multivariate imputation for continuous data.

6.3 Data utility for the partially synthesized data

For all synthetic data results in this section, we apply the inferential methods for partially synthetic data as described in Section 6.1.1.

6.3.1 Analytic comparison I - Logistic regression model for probability of hospice discussion

We used a logistic regression based on the statistical model in Huskamp et al. (2009), applied to the 1,517 patients diagnosed as having stage IV lung cancer, to identify factors associated with whether or not patients have discussed hospice care with their physicians. The authors argue that discussing hospice care with a health-care provider could increase awareness of hospice, and possibly result in earlier use.

The analysis in Huskamp et al. (2009) was based on five complete, observed data sets imputed for missing values (He et al. 2009). The results in this section are based on the partial synthesis of one of these five data sets. Hence, the observed data set results reported here closely, but do not exactly match the analytic results reported in Huskamp et al. (2009). The combining rules for tests of multivariate hypotheses when using multiple imputation simultaneously for missing data and partial synthesis are detailed in Kinney and Reiter (2010). Given that the item non-response rates were relatively low ($< 3\%$, He et al. 2009), it is unlikely that the synthetic data results will be sensitive to ignoring the imputation of missing data.

Table 6.1 displays descriptive characteristics for synthesized variables and estimated probabilities of hospice discussion, not adjusting for other covariates. Adjusted

estimates are reported in Table 6.2. We also quantify data utility in Table 6.3 with respect to the parameters of the hospice logistic model. Note that we do not focus our discussion on the practical interpretation of the synthetic data results to draw clinical conclusions. Our aim is to illustrate interpretation of the data utility measures described in Section 6.1.3.

Table 6.1: *Descriptive characteristics and estimated probabilities of hospice discussion by synthesized variables, unadjusted results. (Standard errors in parentheses)*

Characteristic	**Patients (%)**		**Discussed Hospice (%)**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Overall			53.2	53.2		
Sex						
Male	61.3	57.1	53.3 (1.3)	54.7 (2.4)		
Female	38.7	42.9	53.0 (1.3)	55.6 (2.6)	0.89	0.43
Race/ethnicity						
White	73.7	72.0	55.2 (1.3)	54.4 (1.5)		
Black	10.7	11.4	42.6 (1.3)	50.0 (4.1)		
Hispanic	5.9	5.7	40.4 (1.3)	45.8 (5.3)	< 0.001	0.62
Asian	5.1	5.3	49.4 (1.3)	51.6 (6.0)		
Other	4.7	4.8	64.5 (1.2)	54.0 (6.9)		
Married/live with partner						
Yes	61.0	60.2	50.4 (1.3)	55.3 (2.4)	0.006	0.06
No	39.0	39.8	57.6 (1.3)	56.3 (2.6)		
Age (yrs)						
21-54	12.5	12.5	45.5 (1.3)	44.5 (4.1)		
55-59	12.4	12.5	52.1 (1.3)	49.1 (4.0)		
60-64	13.2	13.1	51.0 (1.3)	49.3 (4.0)		
65-69	16.0	16.4	50.8 (1.3)	50.3 (3.6)	0.002	0.007
70-74	17.8	16.7	50.4 (1.3)	55.6 (3.5)		
75-79	13.8	14.3	57.1 (1.3)	57.3 (3.5)		
80+	14.4	14.0	65.1 (1.2)	64.7 (3.3)		
Education						
$<$ High school	22.7	21.6	54.8 (1.3)	55.9 (1.8)		
High school or some college	60.6	62.2	53.5 (1.3)	53.8 (3.2)	0.46	0.22
$\geq$ College degree	16.7	16.3	49.8 (1.3)	47.0 (3.7)		

Table 6.2: *Estimated probabilities for hospice discussion, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Sex				
Male	54.5 (9.4)	54.4 (9.9)	0.72	0.93
Female	53.2 (9.9)	54.7 (10.4)		
Race/ethnicity				
White	54.5 (9.4)	54.4 (9.9)		
Black	46.6 (10.6)	53.3 (11.3)		
Hispanic	38.1 (11.2)	39.2 (12.2)	< 0.001	0.22
Asian	60.0 (11.4)	59.2 (11.7)		
Other	78.2 (7.7)	65.3 (11.6)		
Married/live with partner				
Yes	54.5 (9.4)	54.4 (9.9)	0.031	0.047
No	62.6 (9.6)	62.1 (10.1)		
Age (yrs)				
21-54	54.5 (9.4)	54.4 (9.9)		
55-59	57.5 (9.5)	52.0 (10.0)		
60-64	53.7 (9.2)	55.3 (9.8)		
65-69	50.5 (8.5)	53.9 (8.8)	0.63	0.94
70-74	52.2 (8.5)	60.5 (8.4)		
75-79	60.2 (8.4)	56.3 (8.7)		
80+	59.1 (8.6)	55.2 (9.1)		
Speaks english in home				
Yes	54.5 (9.4)	54.4 (9.9)	0.045	0.061
No	35.7 (12.3)	37.7 (12.7)		
Education				
$<$ High school	56.3 (9.9)	60.3 (10.2)		
High school/some college	54.5 (9.4)	54.4 (9.9)	0.87	0.48
$\geq$ College degree	56.3 (10.1)	53.1 (10.5)		
Income ($)				
$<$ 20 000	66.2 (8.6)	63.1 (9.7)		
20 000-39 999	64.7 (8.5)	62.3 (9.3)		
40 000 - 59 999	64.4 (8.9)	63.0 (9.6)	0.19	0.42
$\geq$ 60 000	54.5 (9.4)	54.4 (9.9)		

Table 6.2: *(Continued) Estimated probabilities for hospice discussion, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Insurance				
Medicare	54.5 (9.4)	54.4 (9.9)		
Medicaid	48.3 (11.1)	49.5 (11.6)	0.018	0.004
Private	57.3 (7.9)	50.3 (8.6)		
Other	77.1 (8.1)	79.9 (7.8)		
Treated in VA facility				
Yes	46.6 (14.4)	51.1 (14.7)	0.43	0.74
No	54.5 (9.4)	54.4 (9.6)		
HMO member				
Yes	60.4 (9.3)	60.0 (9.9)	0.25	0.42
No	54.5 (9.4)	54.4 (9.9)		
Region				
South	60.7 (8.9)	59.3 (9.6)		
West	54.5 (9.4)	54.4 (9.9)	0.12	0.055
Other	47.9 (10.0)	44.7 (10.5)		
Days from diagnosis to interview				
Quartile 1	54.5 (9.4)	54.4 (9.9)		
Quartile 2	67.1 (8.2)	66.1 (8.9)	0.055	0.079
Quartile 3	62.0 (8.6)	61.1 (9.4)		
Quartile 4	38.7 (7.6)	24.8 (8.1)		
Days from interview until death				
Deceased prior to interview	54.5 (9.4)	54.4 (9.9)		
1-59 d	11.0 (4.2)	12.2 (4.7)		
60-119 d	7.6 (3.1)	8.7 (3.6)		
120-179 d	5.7 (2.6)	6.0 (2.8)	< 0.001	< 0.001
180-239 d	5.9 (2.2)	6.1 (2.4)		
$\geq$ 240 d	64.2 (8.9)	65.0 (9.2)		

Table 6.2: *(Continued) Estimated probabilities for hospice discussion, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Received chemo. before interview				
Yes	54.5 (9.4)	54.4 (9.9)	0.006	0.002
No	59.1 (9.3)	58.4 (9.6)		
Comorbidity				
None	54.5 (9.4)	54.4 (9.9)		
Mild	57.1 (9.6)	57.0 (9.8)	0.23	0.15
Moderate	64.7 (9.1)	65.4 (9.2)		
Severe	70.3 (9.4)	68.8 (10.2)		
Alive but surrogate completed interview				
Yes	61.7 (9.4)	59.2 (11.2)	0.004	0.008
No	54.5 (9.4)	54.4 (9.9)		

Table 6.1 shows that the marginal sample counts for each synthesized variable have been preserved. Differences between observed data set and synthetic data sample proportions were generally $< 1\%$, with the exception of sex. The synthetic data sample proportions by sex are consistent with the observed data set sample proportions using the entire data set of 5,000 records, but there is a higher observed proportion of males in the subset of Stage IV lung cancer patients, hence the larger deviation in marginal sample counts by sex. The synthetic data estimated probabilities of hospice discussion were generally within 3% of the observed data set estimates. However, there is a 7% discrepancy in estimated probability for the 'black' race factor level, and the conclusion of significance for race changes from strongly significant to strongly insignificant. The synthetic data standard errors were approximately 3 times greater than the observed data set standard errors because of the additional between-imputation variability, which was not offset by the presence of non-synthesized covariates.

Table 6.3: *Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β		$\left\lvert \frac{Bias_q}{\sqrt{V_q}} \right\rvert$	**CI. overlap**	**Coverage error**
	Obs.	Syn.			
Intercept	0.182 (0.381)	0.179 (0.409)	0.010	0.832	0.050
Sex					
Female	-0.054 (0.148)	0.015 (0.188)	0.364	0.745	0.065
Race/ethnicity					
Black	-0.323 (0.226)	-0.043 (0.254)	1.101	0.767	0.196
Hispanic	-0.670 (0.315)	-0.632 (0.371)	0.101	0.831	0.051
Asian	0.213 (0.335)	0.205 (0.351)	0.024	0.803	0.050
Other	1.092 (0.323)	0.495 (0.404)	1.479	0.707	0.316
Married/live with partner					
No	0.336 (0.155)	0.322 (0.163)	0.085	0.707	0.051
Age (yrs)					
55-59	0.127 (0.266)	-0.100 (0.333)	0.684	0.823	0.105
60-64	-0.042 (0.266)	0.034 (0.301)	0.253	0.791	0.057
65-69	-0.166 (0.297)	-0.022 (0.321)	0.447	0.794	0.073
70-74	-0.097 (0.304)	0.249 (0.312)	1.109	0.783	0.199
75-79	0.226 (0.322)	0.073 (0.322)	0.474	0.784	0.076
80+	0.183 (0.320)	0.031 (0.349)	0.436	0.809	0.072
Speaks english in home					
No	-0.769 (0.382)	-0.700 (0.371)	0.193	0.800	0.054
Education					
$<$ High school	0.081 (0.190)	0.247 (0.213)	0.778	0.745	0.122
$\geq$ College degree	0.078 (0.175)	-0.054 (0.238)	0.554	0.785	0.088
Income ($)					
$<$ 20 000	0.474 (0.241)	0.367 (0.254)	0.422	0.758	0.071
20 000-39 999	0.413 (0.218)	0.333 (0.226)	0.355	0.742	0.065
40 000 - 59 999	0.416 (0.236)	0.359 (0.238)	0.241	0.745	0.057

$\left\lvert \frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}} \right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table 6.3: *(Continued) Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β		$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. overlap**	**Coverage error**
	Obs.	Syn.			
Insurance					
Medicaid	-0.223 (0.343)	-0.202 (0.346)	0.059	0.895	0.050
Private	0.109 (0.222)	0.202 (0.213)	0.437	0.726	0.072
Other	1.021 (0.360)	1.215 (0.358)	0.542	0.798	0.084
Treated in VA facility					
Yes	-0.298 (0.400)	-0.135 (0.401)	0.406	0.817	0.069
HMO member					
Yes	0.276 (0.245)	0.200 (0.246)	0.312	0.748	0.061
Region					
South	-0.266 (0.241)	-0.399 (0.247)	0.540	0.751	0.084
Other	0.242 (0.179)	0.230 (0.183)	0.062	0.716	0.050
Days from diagnosis to interview					
Quartile 2	0.263 (0.190)	0.200 (0.190)	0.329	0.718	0.063
Quartile 3	0.543 (0.193)	0.498 (0.193)	0.231	0.720	0.056
Quartile 4	0.329 (0.201)	0.278 (0.202)	0.252	0.725	0.057
Days from interview until death					
1-59 d	-1.36 (0.255)	-1.31 (0.261)	0.196	0.758	0.054
60-119 d	-2.26 (0.252)	-2.18 (0.255)	0.331	0.754	0.063
120-179 d	-2.67 (0.276)	-2.56 (0.274)	0.377	0.761	0.066
180-239 d	-2.98 (0.335)	-2.97 (0.338)	0.042	0.792	0.050
$\geq$ 240 d	-2.94 (0.191)	-2.94 (0.196)	0.025	0.724	0.050

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table 6.3: *(Continued) Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β Obs.	Syn.	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. overlap**	**Coverage error**
Received chemo. before interview					
Yes	0.409 (0.147)	0.447 (0.148)	0.258	0.692	0.058
Comorbidity					
Mild	0.181 (0.185)	0.163 (0.184)	0.100	0.713	0.051
Moderate	0.098 (0.206)	0.106 (0.204)	0.039	0.725	0.050
Severe	0.417 (0.215)	0.469 (0.214)	0.244	0.731	0.057
Alive but surrogate completed interview					
Yes	0.677 (0.235)	0.622 (0.236)	0.234	0.744	0.056
Summary avg.					
Synthesized			0.56	0.78	0.11
Non synthesized			0.26	0.75	0.06
All			0.36	0.76	0.08

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

For the model results adjusted for other covariates in Table 6.2, estimated probabilities using synthetic data differed on average by 1.5% from the observed data set estimates for non-synthesized variables. There were non-zero discrepancies because the model fit also depends on values of synthesized variables. The average deviation for synthesized variables was 4.1%, with race and age estimated probabilities showing the largest deviations. Again there was a change in significance at conventional levels for race. For all other factors, conclusions of significance were preserved.

Data utility measures with respect to the parameters of the adjusted hospice logistic model are reported in Table 6.3. The average standardized bias for non-synthesized variables was 0.26 versus 0.56 for synthesized variables, and the average error in coverage probability due to bias was 0.06 for non-synthesized variables, versus

0.11 for synthesized variables. The largest estimated errors were in excess of 0.20 for some race and age coefficient estimates. Average confidence interval overlap results were similar between non-synthesized and synthesized variables, at 0.75 and 0.78 respectively.

Overall, we conclude that data utility has been preserved for non-synthesized variables and synthesized sex, education and marital status, within reasonable bounds. However, data utility was not preserved for age and race covariates. We ran the imputation models again but only on the $n = 1,517$ units in the analysis procedure (and not the entire data set). The revised bivariate association estimates for synthesized variables are in Appendix B, Table B.1. There were no changes in conclusions of significance relative to the observed data set, and deviations in synthetic data estimated probabilities from observed data set estimates were generally less than one standard error.

Revised logistic regression model results adjusted for other covariates are in Appendix B, Tables B.2 - B.3. The synthetic data estimated probabilities differed by 4% on average (equivalent to 0.5 standard errors) from the observed data set estimates for both non-synthesized and synthesized variables. Deviations for age and race factors were less than 5%. The largest deviation (10%) was actually for the 'English speaking' factor, which was not synthesized. We attribute this deviation to variability between repeated draws from the posterior predictive distributions to impute the synthetic data. Furthermore, there were no changes in conclusions of significance. The average standardized bias for coefficient estimates was 0.14 for non-synthesized variables, and 0.44 for synthesized variables. The overall average overlap was unchanged at 0.76.

The overall average error in coverage probability was slightly lower at 0.069 (0.054 for non-synthesized variables; 0.093 for synthesized variables). However, the coverage error estimates were greater for age groups '55-59' and '75-79'.

To summarize, utilizing the same subset of records in the imputation model as for the analysis procedure assisted in preserving data utility. In particular, the large deviations from observed data set estimates for the 'black' race factor level were removed, as well as producing smaller bias for non-synthesized variables. Any remaining discrepancies noted for the age factor level estimates were found to be removed after conditioning on hospice discussion in the imputation model for age. Meng (1994) coined the term *uncongeniality* to refer to the case where the analyst and the imputer have access to different types of information. In our example, the general imputation model was derived from the entire data set, but the analysis procedure used a subset of records. If the imputation model does not capture all the important subgroup relationships, results from the synthetic data may be biased. Refer to Chapter 3 for a statistical definition of uncongeniality for synthetic data.

6.3.2 Analytic comparison II - Multinomial logistic regression model for cancer patients' roles in treatment decisions

The second analysis was a multinomial logistic regression based on the analytic study by Keating et al. (2010). The objective of the study was to assess whether the characteristics of the decision, including evidence about the treatment's benefits, whether the decision was likely preference-sensitive, and treatment modality, influ-

enced patients' roles in that decision. The authors argue that patients with more active roles in decisions are more satisfied and may have better health outcomes. The analysis used $10,939$ decisions from $5,383$ lung and colorectal cancer patients who self-completed the full patient survey (that is, a surrogate did not complete the survey), and who had discussed at least one treatment (surgery, radiation or chemotherapy) with a clinician. The unit of analysis was the decision, and it included up to three observations per patient (one for each of three treatment modalities that a patient may have discussed). Standard errors were adjusted to account for correlation among repeated decisions within patients by using a robust variance estimator in the *Stata* software package. We adopt the same statistical analysis methods for each synthetic data set. Table 6.4 explains the different categories for role in treatment decision making. Shared control was the base level for decision role.

The results in this section are based on partial synthesis of one of the five complete, observed data sets imputed for missing values (He et al. 2009). Hence, the observed data set results reported here closely, but do not exactly match the analytic results reported in Keating et al. (2010).

Table 6.4: *Categories for role in cancer treatment decision making*

Level	**Description**	**Corresponding survey response**
1	Patient control	You made the decision with little or no input from your doctors
		You made the decision after considering your doctors' opinions
2	Shared control *	You and your doctors made the decision together
3	Physician control	Your doctors made the decision after considering your opinion
		Your doctors made the decision with little or no input from you

* - reference group

Appendix B, Table B.4 reports the adjusted differences in proportion reporting by cancer treatment decision role relative to the reference group for each characteristic. Marital status was not significant at conventional levels using the synthetic data, but was significant using the observed data set. Conclusions of significance were preserved for all non-synthesized variables, including all clinical variables, with minimal deviation in the magnitude of the p-values. Data utility results for coefficients of the decision role model are reported in Appendix B, Table B.5. For non-synthesized characteristics, aggregate data utility was better than utility from our first analytic comparison on hospice discussion. One explanation is the larger subset of patients analyzed in the second analytic comparison. For synthesized characteristics, data utility was lowest for physician control coefficient estimates. For example, the bias in the marital status coefficient estimate has an estimated error in coverage probability of 0.75. That is, only 25% of the time would we expect the confidence interval for marital status to cover the true value in repeated sampling. The reasons for the poor data utility results may include the following:

(i) Decision role for any type of treatment was not a predictor in any imputation model except for education.

(ii) The treatment indicator variable was not always an explicit predictor in the imputation models, but other related variables were included. For example, a response to 'rate the quality of surgery' implies the patient had undergone surgery.

(iii) Differing variable structures between the imputation models and analysis procedure. For example, marital status was imputed as a 6-level categorical variable,

but analyzed as a binary variable. Decision role was included in the stepwise regressions with 5 factor levels, but analyzed with 3 factor levels.

(iv) The decision role analysis model is not a standard multinomial logistic model because of the multiple observations per patient. Existing inferential methods for obtaining interval estimates for scalar quantities, and for performing large sample tests of multicomponent hypotheses, have not been shown to extend to valid inference for more complex analyses on synthetic data, for example, cluster and factor analysis, or hierarchical models (Reiter 2009).

The mixed results in our data utility assessment demonstrate that it is difficult to generate synthetic data that would preserve the inferential conclusions from the observed data set, for all potential future analyses. Our analytic comparisons suggest a variety of analyst-defined variables, statistical models and data subsets, that may be used in the analysis procedure, and it is impossible for the imputer to foresee and capture all such analyses in the imputation models. Given this, it is no surprise to observe poor data utility results for some quantities of interest. For CanCORS, we recommend use of synthetic data primarily for exploratory data-analytic purposes, as a screening device for preliminary research hypotheses prior to requesting full access to the original data. The benefit is the ability of analysts to explore the data without incurring the time and monetary costs to gain authorized access. The CanCORS Consortium can meet this demand at lower risk of respondent identification (see Section 6.4).

6.4 Disclosure risk assessment

Tables 6.5 and 6.6 quantify the disclosure risk for the partially synthetic data we generated, using the measures described in Section 6.1.2.

Table 6.5: *Disclosure risk - CanCORS lung cancer* ***partially synthetic*** *data*

Quasi-identifier set	**MXM**	**EMR**	**TMR**	Max (F_{il})	Mean(F_{il})
Set 1	771	8	0	485	210
Set 2	232	43	18	97	9
Set 3	232	178	143	14	0.5

Table 6.6: *Disclosure risk - CanCORS colorectal cancer* ***partially synthetic*** *data*

Quasi-identifier set	**MXM**	**EMR**	**TMR**	Max (F_{il})	Mean(F_{il})
Set 1	601	7	0	325	153
Set 2	20	4	1	42	4
Set 3	20	18	17	5	0.1

$\mathbf{F}_{\text{il}}$: the number of units in synthetic data set l, that match the quasi-identifying information of original unit i.
MXM: Maximum number of matches $= \sum_{i=1}^{n} \frac{1}{m} \sum_{l=1}^{m} C_{\text{il}}$, where $C_{\text{il}} = 1$ when record i is among the F_{il} matches from synthetic data set $Z_{\text{syn}}^{(\text{l})}$, and $C_{\text{il}} = 0$ otherwise
EMR: Expected match risk for unit $i = \frac{1}{m} \sum_{l=1}^{m} \frac{1}{F_{\text{il}}} \times C_{\text{il}}$
TMR: Presumed true match risk for unit $i = \frac{1}{m} \sum_{l=1}^{m} K_{\text{il}}$, where $K_{\text{il}} = 1$ when $F_{\text{il}} = 1$ and $C_{\text{il}} = 1$, and $K_{\text{il}} = 0$ otherwise.

The true match risk rate is zero for quasi-identifying Set 1 variables for both lung and colorectal cancer synthetic data, zero meaning that the target record was never correctly and uniquely identified as a match. For lung cancer disclosure risk results in Table 6.5, as the number of variables in the quasi-identifying set increases, the expected (EMR) and true match risk (TMR) increase, because more criteria are required to identify a match, decreasing the value of F_{il} (the number of potential candidates), and increasing the number of cases where $F_{\text{il}} = 1$ (a unique match). For colorectal cancer, the EMR for quasi-identifying Set 2 is actually less than the EMR

for Set 1, because of a corresponding reduction in the number of cases where the target was in the set of potential candidates; that is, more cases where C_{il} equals 0. This is evidence of strong disclosure risk protection from creating synthetic values for education or marital status, for the colorectal cancer data set.

To illustrate the interpretation of the disclosure risk results, consider quasi-identifying Set 2 for lung cancer. We anticipate the intruder could uniquely identify 18 patients out of $\approx 5,000$. The intruder will not know how many of the unique matches are actually true matches. Furthermore, across the five partially synthetic data sets, there were 76 unique matches. Of these, 68 records were uniquely identified in only 1 out of the 5 synthetic data sets, 6 were identified twice, and only 1 record was identified uniquely in all 5 data sets. This is the reason why the true match risk is a 'presumed' risk because the intruder will not know if a unique match is the correct one, and the set of potential matches will vary across the released synthetic data sets. The overall expected match risk is 43 out of 5,000 patients. That is, allowing for uncertainty from randomly guessing from the potential candidates, it is expected less than 1% of patients could be correctly identified.

Using quasi-identifiers, it is possible that partial synthesis could reduce the number of potential match candidates. Thus, we have also quantified disclosure risk measures on the observed data set. Results are shown in Tables 6.7 and 6.8. The true match risk of the observed data set is the number of unique matches, and ranges in value from 0 to 4000. Because we are dealing with original values, $C_{\text{il}} = 1$ for all $i = 1, .., n$, and therefore the expected and true match risk values are larger for the observed data set. Comparing the disclosure risk values in Tables 6.7 and 6.8 to the values in

Tables 6.5 and 6.6, we conclude that, as expected, disclosure risk is greatly reduced when releasing synthetic data instead of the observed values.

Table 6.7: *Disclosure risk - CanCORS lung cancer* ***original*** *data set*

Quasi-identifier set	**EMR**	**TMR**	Max (F_{il})	Mean(F_{il})
Set 1	70	0	380	214
Set 2	1770	989	87	11
Set 3	4495	4000	11	2

Table 6.8: *Disclosure risk - CanCORS colorectal cancer* ***original*** *data set*

Quasi-identifier set	**EMR**	**TMR**	Max (F_{il})	Mean(F_{il})
Set 1	70	0	345	156
Set 2	1824	1070	42	7
Set 3	4105	3810	8	1

$\mathbf{F}_{\text{il}}$: the number of units in synthetic data set l, that match the quasi-identifying information of original unit i.
EMR: Expected match risk for unit $i = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{F_{\text{il}}} \times C_{\text{il}}$
TMR: Presumed true match risk for unit $i = \frac{1}{m}\sum_{l=1}^{m} K_{\text{il}}$, where $K_{\text{il}} = 1$ when $F_{\text{il}} = 1$ and $C_{\text{il}} = 1$, and $K_{\text{il}} = 0$ otherwise.

It should be noted that there are no universal rules on acceptable levels of disclosure risk. The level of disclosure risk tolerated depends on many factors, including the risk attitude of the database owners, size of the target population and the sampling fraction, realistic assessment of assumed levels of external information available to the intruder, and the intruders' strategies to identify targets. The results in Tables 6.5 and 6.6 illustrate one, statistically based approach to quantify the disclosure risk of partially synthetic databases.

In this chapter we have demonstrated partial synthesis of the CanCORS cancer patient survey data set. Our practical application has highlighted a number of areas where further research is required. There remain open questions on how to select representative published studies for analytical comparison, how many to select, and

whether to combine the multiple data utility results into a single index. The content of, and mode to communicate results of data utility assessment to public- users also needs to be addressed. These issues were also raised in Chapter 3; hence this chapter provides practical motivation to address these questions. In terms of disclosure risk assessment, we used a basic intruder strategy and a statistical framework for calculation. Disclosure risk assessment can be improved by more circumstantiated identification of the pool of potential intruders, available information sources, strategies for record identification, and modeling the uncertainty in these assumptions. Limited research has been done on use of Bayesian prior distributions to capture uncertainty in intruder information, and this is another area for future investigation.

Chapter 7

Conclusion

This thesis has presented three new research contributions for using synthetic data techniques for statistical disclosure control.

The first contribution is a definition of congeniality for multiple imputation for synthetic data. For imputers, the definition provides a theoretical framework to identify the sources of actual and potential differences between observed data set and synthetic data inferential results, the onus being on the imputer to justify data utility has been adequately preserved in the synthetic data for public release. For analysts, a conceptual understanding of sources of uncongeniality and its statistical implications are important. Analysts need to understand that by virtue of the synthetic data creation process, synthetic data results will generally not be exactly the same as observed data set results, and there may be large deviations because of uncongeniality, either unforeseen, or intentional to protect data confidentiality. Given this, analysts may wish to adjust their purpose of investigation to more exploratory data analytic purposes.

Motivated by the definition of congeniality, we presented an alternative approach to fully synthetic data inference to recover the observed data set sampling distribution of sufficient statistics from synthetic data. The alternative approach assisted to understand better the role of the original survey design in the existing inferential methods for analysis of fully synthetic data. In our simulation and empirical data studies, when the observed survey design, and the synthetic data resampling mechanism and analysis procedure were the same, fully Bayesian-derived and alternative approach inferential estimates were equivalent. Our recommendation is to set the synthetic data resampling and analysis procedure to be the same as the original design, and that original design information be made available to the analyst where possible.

The second contribution demonstrated that application of Adjustment for Density Maximization (ADM) can achieve positive variance estimates when analyzing fully synthetic data, which is not guaranteed by the existing method-of-moments variance estimator. This new approach required specification of synthetic data inference in a hierarchical model framework. The ADM approach is offered as an alternative to, (not a replacement for) the existing combining rules when the analyst is concerned the existing combining rules will produce non-positive variance estimates, but further theoretical justification is required to establish ADM as a principled approach to variance estimation with fully synthetic data.

Finally, the third contribution demonstrated application of synthetic data techniques for disclosure control in the CanCORS Consortium. In our data utility assessment, we found data utility was not preserved if the set of records for the imputation

model was not the same as the set of records used in the analysis procedure, an example of *uncongeniality*. We discussed how it would not be uncommon to encounter such *uncongeniality* considering the large number of potential public users with specific research questions, relevant to only a subset of the data, and using variable structures not included in the imputation models. It is recommended that partially synthetic, public-data for CanCORS be used primarily to assess preliminary hypotheses, but special access to the original data set should be requested to answer more specific questions with confirmatory results. For disclosure risk, we found the risk of identification to be greatly reduced by replacing the original data set with synthetic values. Our research presents a building block for addressing the increasing demand of sharing data, yet protecting data confidentiality in clinical outcomes research.

Research in synthetic data is a growing field and attracting increasing interest from statistical agencies as a method of statistical disclosure control. Reiter (2009) summarized some future research challenges in multiple imputation for disclosure limitation. These are '*flexible synthesis models*, *synthesis design strategies*, *confidence in synthetic data*, and *expansion of analysis methods*'. We encountered all these challenges in our applied work in Chapter 6, specifically (i) identifying strategies to deal with hundreds of variables when building imputation models; (ii) developing less subjective approaches to identify high disclosure risk variables and to specify the intruder attack method; (iii) finding a balance between the search for the 'perfect' imputation model and inevitable uncongeniality; and (iv) adapting synthetic data inferential methods to more complex analytic procedures. Our work in Chapter 3 also highlighted the limitations, due to uncongeniality, for gaining public confidence in

synthetic data. Continued research in these areas will help to increase the acceptance of using synthetic data for disclosure control, by both statistical agencies and analysts.

Appendix A

Recoded variable structure - CanCORS data

Table A.1: *Recoded variable structure - CanCORS data (Section 6.2.2)*

Variable	**Original structure**	**Revised structure**
Age	0-52	
	53-54	0-54
	55-59	55-59
	60-64	60-64
	65-69	65-69
	70-74	70-74
	75-79	75-79
	80-81	80+
	82+	
Gender	Male	Male
	Female	Female
Marital Status	Married	Married
	Divorced	Divorced
	Living with partner	Living with partner
	Never married	Never married
	Separated	Separated
	Widowed	Widowed
Race	White	White
	African American	African American
	Hispanic or Latino	Hispanic or Latino
	Asian	Asian
	American Indian	Other
	Pacific Islander	

Table A.1: *(Continued) Recoded variable structure - CanCORS data (Section 6.2.2)*

Variable	**Original structure**	**Revised structure**
Education	1st grade	< High school
	3rd grade	High school diploma
	4th grade	Some college
	5th grade	College degree
	6th grade	> College
	7th grade	
	8th grade	
	9th grade	
	10th grade	
	11th grade	
	High School Diploma or GED or completed 12th grade	
	Vocational Diploma	
	More than 2 years	
	More than 4 years	
	1st year (freshman)	
	2nd year (sophomore)	
	3rd year (junior)	
	College Degree (BA/BS) or 4th year (senior)	
	1st year grad or prof school	
	Masters degree (MA/MS/MPH/MBA etc.) or 2nd year grad or prof school	
	Doctorate degree (J.D., M.D., PhD, etc) or more than 2 years grad or prof school	

Appendix B

Supplementary analytic comparison results

Table B.1: *Descriptive characteristics and estimated probabilities of hospice discussion by synthesized variables, unadjusted results. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Patients (%)**		**Discussed Hospice (%)**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Overall			53.2	53.2		
Sex						
Male	61.3	52.8	53.3 (1.3)	54.1 (2.4)		
Female	38.7	47.2	53.0 (1.3)	54.9 (2.4)	0.89	0.63
Race/ethnicity						
White	73.7	74.6	55.2 (1.3)	55.6 (1.5)		
Black	10.7	9.9	42.6 (1.3)	42.0 (4.2)		
Hispanic	5.9	6.1	40.4 (1.3)	40.4 (5.3)	< 0.001	0.003
Asian	5.1	5.1	49.4 (1.3)	50.2 (5.8)		
Other	4.7	4.3	64.5 (1.2)	58.5 (6.5)		
Married/live with partner						
Yes	61.0	60.8	50.4 (1.3)	54.5 (2.1)	0.006	0.038
No	39.0	39.2	57.6 (1.3)	54.9 (2.3)		
Age (yrs)						
21-54	12.5	12.3	45.5 (1.3)	46.7 (3.8)		
55-59	12.4	12.1	52.1 (1.3)	50.0 (3.7)		
60-64	13.2	13.5	51.0 (1.3)	52.7 (3.8)		
65-69	16.0	15.5	50.8 (1.3)	52.0 (3.6)	0.002	0.011
70-74	17.8	18.6	50.4 (1.3)	52.0 (3.3)		
75-79	13.8	14.0	57.1 (1.3)	51.7 (3.9)		
80+	14.4	14.0	65.1 (1.2)	66.6 (3.3)		
Education						
$<$ High school	22.7	22.9	54.8 (1.3)	54.5 (2.9)		
High school/some college	60.6	62.2	53.5 (1.3)	53.8 (1.7)	0.46	0.37
$\geq$ College degree	16.7	14.9	49.8 (1.3)	48.7 (3.4)		

Table B.2: *Estimated probabilities for hospice discussion, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Sex				
Male	54.5 (9.4)	60.0 (9.7)	0.72	0.95
Female	53.2 (9.9)	59.7 (10.2)		
Race/ethnicity				
White	54.5 (9.4)	60.0 (9.7)		
Black	46.6 (10.6)	52.1 (11.7)		
Hispanic	38.1 (11.2)	42.0 (12.1)	< 0.001	0.018
Asian	60.0 (11.4)	63.3 (11.6)		
Other	78.2 (7.7)	74.8 (9.6)		
Married/live with partner				
Yes	54.5 (9.4)	60.0 (9.7)	0.031	0.011
No	62.6 (9.6)	69.2 (9.1)		
Age (yrs)				
21-54	54.5 (9.4)	60.0 (9.7)		
55-59	57.5 (9.5)	56.1 (10.2)		
60-64	53.7 (9.2)	57.6 (10.0)		
65-69	50.5 (8.5)	54.2 (9.1)	0.63	0.93
70-74	52.2 (8.5)	56.6 (8.8)		
75-79	60.2 (8.4)	54.5 (9.3)		
80+	59.1 (8.6)	61.7 (8.8)		
Speaks english in home				
Yes	54.5 (9.4)	60.0 (9.7)	0.045	0.136
No	35.7 (12.3)	45.9 (13.9)		
Education				
$<$ High school	56.3 (9.9)	58.0 (11.0)		
High school/some college	54.5 (9.4)	60.0 (9.7)	0.87	0.86
$\geq$ College degree	56.3 (10.1)	61.6 (10.0)		
Income $				
$<$ 20 000	66.2 (8.6)	70.6 (8.5)		
20 000-39 999	64.7 (8.5)	69.1 (8.5)		
40 000 - 59 999	64.4 (8.9)	69.0 (8.9)	0.19	0.23
$\geq$ 60 000	54.5 (9.4)	60.0 (9.7)		

Table B.2: *(Continued) Estimated probabilities for hospice discussion, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Insurance				
Medicare	54.5 (9.4)	60.0 (9.7)		
Medicaid	48.3 (11.1)	54.3 (11.5)	0.018	0.015
Private	57.3 (7.9)	61.8 (8.2)		
Other	77.1 (8.1)	81.3 (7.3)		
Treated in VA facility				
Yes	46.6 (14.4)	52.3 (14.8)	0.43	0.44
No	54.5 (9.4)	60.0 (9.7)		
HMO member				
Yes	60.4 (9.3)	64.9 (9.5)	0.25	0.27
No	54.5 (9.4)	60.0 (9.7)		
Region				
South	60.7 (8.9)	64.8 (9.1)		
West	54.5 (9.4)	60.0 (9.7)	0.12	0.117
Other	47.9 (10.0)	52.3 (10.6)		
Days from diagnosis to interview				
Quartile 1	54.5 (9.4)	60.0 (9.7)		
Quartile 2	67.1 (8.2)	70.8 (8.3)	0.055	0.100
Quartile 3	62.0 (8.6)	66.2 (8.8)		
Quartile 4	38.7 (7.6)	27.5 (8.9)		
Days from interview until death				
Deceased prior to interview	54.5 (9.4)	60.0 (9.7)		
1-59 d	11.0 (4.2)	13.3 (5.2)		
60-119 d	7.6 (3.1)	9.8 (4.1)		
120-179 d	5.7 (2.6)	6.9 (3.2)	< 0.001	< 0.001
180-239 d	5.9 (2.2)	7.3 (2.9)		
$\geq$ 240 d	64.2 (8.9)	69.5 (8.7)		

Table B.2: *(Continued) Estimated probabilities for hospice discussion, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Received chemo. before interview				
Yes	54.5 (9.4)	60.0 (9.7)	0.006	0.005
No	59.1 (9.3)	64.5 (9.1)		
Comorbidity				
None	54.5 (9.4)	60.0 (9.7)		
Mild	57.1 (9.6)	62.6 (9.5)	0.23	0.13
Moderate	64.7 (9.1)	70.8 (8.5)		
Severe	70.3 (9.4)	74.8 (9.1)		
Alive but surrogate completed interview				
Yes	61.7 (9.4)	66.2 (10.5)	0.004	0.004
No	54.5 (9.4)	60.0 (9.7)		

Table B.3: *Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β		$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. overlap**	**Coverage Error**
	Obs.	Syn.			
Intercept	0.182 (0.381)	0.179 (0.409)	0.556	0.832	0.068
Sex					
Female	-0.054 (0.148)	-0.014 (0.218)	0.183	0.783	0.054
Race/ethnicity					
Black	-0.323 (0.226)	-0.326 (0.267)	0.013	0.781	0.050
Hispanic	-0.670 (0.315)	-0.736 (0.339)	0.194	0.802	0.055
Asian	0.213 (0.335)	0.143 (0.341)	0.208	0.794	0.055
Other	1.092 (0.323)	0.705 (0.361)	1.073	0.818	0.188
Married/live with partner					
No	0.336 (0.155)	0.408 (0.162)	0.438	0.705	0.072
Age (yrs)					
55-59	0.127 (0.266)	-0.163 (0.292)	0.995	0.783	0.172
60-64	-0.042 (0.266)	-0.096 (0.309)	0.174	0.800	0.053
65-69	-0.166 (0.297)	-0.240 (0.314)	0.238	0.788	0.057
70-74	-0.097 (0.304)	-0.142 (0.315)	0.142	0.786	0.052
75-79	0.226 (0.322)	-0.229 (0.343)	1.327	0.793	0.264
80+	0.183 (0.320)	0.072 (0.328)	0.338	0.790	0.063
Speaks english in home					
No	-0.769 (0.382)	-0.574 (0.386)	0.503	0.812	0.080
Education					
< High school	0.081 (0.190)	-0.073 (0.204)	0.758	0.734	0.118
≥ College degree	0.078 (0.175)	0.070 (0.203)	0.037	0.742	0.050
Income $					
< 20 000	0.474 (0.241)	0.473 (0.244)	0.006	0.748	0.050
20 000-39 999	0.413 (0.218)	0.401 (0.220)	0.052	0.735	0.050
40 000 - 59 999	0.416 (0.236)	0.396 (0.237)	0.084	0.744	0.050

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table B.3: *(Continued) Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β		$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. overlap**	**Coverage Error**
	Obs.	Syn.			
Insurance					
Medicaid	-0.223 (0.343)	-0.236 (0.357)	0.037	0.805	0.050
Private	0.109 (0.222)	0.076 (0.229)	0.147	0.743	0.054
Other	1.021 (0.360)	1.073 (0.371)	0.141	0.809	0.052
Treated in VA facility					
Yes	-0.298 (0.400)	-0.317 (0.410)	0.046	0.824	0.050
HMO member					
Yes	0.276 (0.245)	0.271 (0.247)	0.022	0.750	0.050
Region					
South	-0.266 (0.241)	-0.315 (0.241)	0.204	0.714	0.051
Other	0.242 (0.179)	0.212 (0.181)	0.160	0.745	0.057
Days from diagnosis to interview					
Quartile 2	0.263 (0.190)	0.208 (0.190)	0.286	0.718	0.059
Quartile 3	0.543 (0.193)	0.482 (0.193)	0.311	0.720	0.062
Quartile 4	0.329 (0.201)	0.268 (0.202)	0.302	0.725	0.051
Days from interview until death					
1-59 d	-1.36 (0.255)	-1.38 (0.259)	0.089	0.756	0.051
60-119 d	-2.26 (0.252)	-2.30 (0.256)	0.129	0.755	0.051
120-179 d	-2.67 (0.276)	-2.64 (0.277)	0.088	0.764	0.050
180-239 d	-2.98 (0.335)	-3.03 (0.339)	0.160	0.793	0.053
$\geq$ 240 d	-2.94 (0.191)	-2.97 (0.197)	0.150	0.755	0.052

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table B.3: *(Continued) Data utility for parameters of the hospice discussion logistic model, adjusted for other covariates. Imputation model and analysis procedure use same set of records. (Standard errors in parentheses)*

Characteristic	**Coef. Est** β		$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. overlap**	**Coverage Error**
	Obs.	Syn.			
Received chemo. before interview					
Yes	0.409 (0.147)	0.420 (0.148)	0.073	0.693	0.051
Comorbidity					
Mild	0.181 (0.185)	0.193 (0.185)	0.065	0.715	0.051
Moderate	0.098 (0.206)	0.111 (0.205)	0.063	0.725	0.050
Severe	0.417 (0.215)	0.483 (0.215)	0.306	0.732	0.062
Alive but surrogate completed interview					
Yes	0.677 (0.235)	0.684 (0.239)	0.029	0.746	0.050
Summary avg.					
Synthesized			0.44	0.78	0.09
Non synthesized			0.44	0.75	0.05
Overall			0.25	0.76	0.07

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table B.4: *Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting**			
	Patient		**Shared**	
	Obs.	Syn.	Obs.	Syn.
Level of Evidence for treatment				
Evidence for	*	*	*	*
Uncertain	5.7 (1.4)	5.7 (1.4)	-4.2 (1.4)	-4.0 (1.4)
No evidence for	-1.8 (1.8)	-1.7 (1.8)	-2.6 (1.7)	-2.6 (1.7)
Missing	5.2 (2.7)	5.0 (2.7)	-3.8 (2.6)	-3.6 (2.6)
Preference sensitive				
No	*	*	*	*
Yes	-6.5 (1.5)	-6.2 (1.5)	1.5 (1.6)	1.4 (1.6)
Treatment modality				
Surgery	*	*	*	*
Radiation	-2.7 (1.2)	-2.6 (1.2)	2.2 (1.2)	2.2 (1.2)
Chemotherapy	4.3 (0.9)	4.3 (0.9)	-1.4 (0.9)	-1.3 (0.9)
Received treatment				
No	*	*	*	*
Yes	3.9 (1.4)	3.9 (1.4)	12.3 (1.3)	12.4 (1.4)
Cancer site				
Lung	*	*	*	*
Colorectal	-0.7 (1.5)	-0.4 (1.6)	-0.5 (1.5)	-0.6 (1.5)
Age at diagnosis, (years)				
21-55	*	*	*	*
56-70	1.6 (2.1)	2.6 (3.5)	1.2 (2.1)	-2.7 (2.4)
71-80	1.4 (2.3)	1.9 (2.0)	0.9 (2.2)	-1.7 (2.4)
$\geq$ 81	1.4 (2.3)	0.6 (4.1)	-1.4 (2.3)	-2.5 (2.5)
Sex				
Male	*	*	*	*
Female	-2.0 (1.3)	0.3 (1.9)	1.1 (1.3)	-0.8 (1.8)

* - reference group

Table B.4: *(Continued) Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting**			
	Patient		**Shared**	
	Obs.	Syn.	Obs.	Syn.
Ethnicity				
White	*	*	*	*
Black	1.4 (1.9)	-1.4 (2.8)	1.6 (1.9)	0.7 (2.5)
Hispanic	-5.1 (2.5)	-4.4 (3.7)	4.4 (2.6)	0.8 (3.3)
Asian	-2.2 (3.0)	0.6 (3.7)	-2.3 (3.0)	-2.8 (3.4)
Other	0.7 (2.8)	1.2 (3.3)	1.7 (2.9)	-0.8 (3.1)
Marital status				
Married	*	*	*	*
Not married	-0.2 (1.4)	1.3 (1.4)	-3.0 (1.4)	-2.1 (1.6)
Education				
< High school (HS)	-2.1 (1.8)	-0.5 (1.8)	0.6 (1.8)	0.1 (2.2)
HS graduate or some college	*	*	*	*
College degree or higher	3.3 (1.5)	3.2 (2.0)	-1.8 (1.5)	-2.9 (1.8)
Income $				
< 20,000	2.3 (2.1)	1.2 (2.1)	-1.3 (2.1)	-2.3 (2.1)
20,000 to < 40,000	-2.5 (1.8)	-3.1 (1.8)	2.4 (1.9)	1.7 (1.9)
40,000 to < 60,000	-0.9 (1.9)	-1.2 (1.9)	-0.7 (1.9)	-1.2 (1.9)
≥ 60,000	*	*	*	*
No. self-reported comorbid conditions				
0	*	*	*	*
1	1.8 (1.4)	1.8 (1.4)	-1.8 (1.4)	-1.9 (1.4)
2	1.3 (1.9)	1.2 (1.9)	-2.3 (1.9)	-2.4 (1.9)
≥ 3	0.8 (2.6)	0.9 (2.6)	-0.2 (2.7)	-0.3 (2.6)

* - reference group

Table B.4: *(Continued) Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting**			
	Patient		**Shared**	
	Obs.	Syn.	Obs.	Syn.
Prediag.health status				
Quartile 1	*	*	*	*
Quartile 2	0.4 (1.7)	0.4 (1.7)	1.0 (1.7)	0.8 (1.7)
Quartile 3	4.9 (1.8)	4.9 (1.8)	-2.8 (1.8)	-3.0 (1.8)
Quartile 4	3.7 (1.8)	3.8 (1.8)	-0.4 (1.8)	-0.5 (1.8)
CES-D short form				
≤ 5	*	*	*	*
≥ 6	2.6 (1.7)	2.2 (1.7)	-2.7 (1.7)	-2.4 (1.7)
Study site				
Los Angeles county	*	*	*	*
Alabama	-1.7 (2.3)	-0.9 (2.3)	5.7 (2.3)	5.6 (2.3)
8 counties in North California	2.5 (1.9)	2.0 (1.9)	-2.2 (1.9)	-2.2 (1.9)
22 counties in eastern North Carolina	-8.4 (2.1)	-7.9 (2.2)	11.2 (2.4)	10.9 (2.4)
Iowa	1.1 (2.6)	1.3 (2.6)	4.0 (2.7)	3.4 (2.7)
5 HMOs	-1.1 (2.1)	-0.9 (2.1)	2.1 (2.1)	1.8 (2.1)
15 Veteran Aff. hospitals	0.2 (2.4)	1.8 (2.5)	1.9 (2.5)	1.1 (2.5)

* - reference group

Table B.4: *(Continued) Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting Physician**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Level of Evidence for treatment			< 0.001	< 0.001
Evidence for	*	*		
Uncertain	-1.6 (1.0)	-1.6 (1.0)		
No evidence for	4.4 (1.2)	4.3 (1.2)		
Missing	-1.4 (1.8)	-1.5 (1.8)		
Preference sensitive			< 0.001	< 0.001
No	*	*		
Yes	4.8 (1.2)	4.8 (1.2)		
Treatment modality			< 0.001	< 0.001
Surgery	*	*		
Radiation	0.5 (0.9)	0.4 (0.9)		
Chemotherapy	-2.9 (0.7)	-3.0 (0.7)		
Received treatment			< 0.001	< 0.001
No	*	*		
Yes	-16.2 (1.2)	-16.3 (1.3)		
Cancer site			0.54	0.59
Lung	*	*		
Colorectal	1.2 (1.0)	1.0 (1.1)		
Age at diagnosis, (years)			0.19	0.92
21-55	*	*		
56-70	-2.8 (1.4)	0.1 (2.0)		
71-80	-2.2 (1.5)	-0.1 (1.6)		
≥ 81	0.1 (1.6)	1.8 (2.8)		
Sex			0.28	0.85
Male	*	*		
Female	1.0 (1.0)	0.5 (1.0)		

* - reference group

Table B.4: *(Continued) Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting Physician**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Ethnicity			0.05	0.64
White	*	*		
Black	-3.0 (1.3)	0.7 (2.5)		
Hispanic	0.7 (1.9)	3.7 (2.4)		
Asian	4.4 (2.5)	2.1 (2.7)		
Other	-2.4 (1.9)	-0.4 (2.8)		
Marital status			0.005	0.42
Married	*	*		
Not married	3.2 (1.1)	0.7 (1.1)		
Education			0.07	0.58
< High school (HS)	1.5 (1.3)	0.6 (1.6)		
HS graduate or some college	*	*		
College degree or higher	-1.5 (1.0)	-0.3 (1.6)		
Income $			0.08	0.08
< 20,000	1.0 (1.5)	1.0 (1.5)		
20,000 to < 40,000	0.1 (1.3)	1.5 (1.4)		
40,000 to < 60,000	1.7 (1.5)	2.4 (1.5)		
≥ 60,000	*	*		
No. self-reported comorbid conditions			0.79	0.77
0	*	*		
1	-0.1 (1.0)	0.0 (1.0)		
2	1.0 (1.4)	1.1 (1.4)		
≥ 3	-0.6 (1.9)	-0.6 (1.9)		

* - reference group

Table B.4: *(Continued) Adjusted differences in patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)*

Characteristic	**Adjusted difference in proportion reporting Physician**		**p-value**	
	Obs.	Syn.	Obs.	Syn.
Prediag. health status			0.02	0.02
Quartile 1	*	*		
Quartile 2	-1.3 (1.2)	-1.1 (1.2)		
Quartile 3	-2.1 (1.3)	-1.9 (1.3)		
Quartile 4	-3.3 (1.2)	-3.3 (1.2)		
CES-D short form			0.26	0.28
≤ 5	*	*		
≥ 6	0.1 (1.3)	0.2 (1.3)		
Study site			< 0.001	< 0.001
Los Angeles county	*	*		
Alabama	-4.1 (1.5)	-4.7 (1.5)		
8 counties in North California	-0.3 (1.3)	0.1 (1.3)		
22 counties in eastern North Carolina	-2.8 (1.5)	-3.0 (1.5)		
Iowa	-5.1 (1.5)	-4.7 (1.5)		
5 HMOs	-1.1 (1.4)	-0.9 (1.4)		
15 Veterans Aff. hospitals	-2.1 (1.7)	-2.9 (1.7)		

* - reference group

Table B.5: *Data utility for parameters of the multinomial logit model for patient decision role.*

Characteristic	**Patient**			**Physician**		
	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**
Level of Evidence for treatment						
Evidence for	*	*	*	*	*	*
Uncertain	0.03	0.99	0.05	0.09	0.98	0.05
No evidence for	0.01	1.00	0.05	0.12	0.97	0.05
Missing	0.07	0.98	0.05	0.03	0.99	0.05
Preference sensitive						
No	*	*	*	*	*	*
Yes	0.08	0.98	0.05	0.01	1.00	0.05
Treatment modality						
Surgery	*	*	*	*	*	*
Radiation	0.06	0.98	0.05	0.09	0.98	0.05
Chemotherapy	0.05	0.99	0.05	0.06	0.98	0.05
Received treatment						
No	*	*	*	*	*	*
Yes	0.01	1.00	0.05	0.04	0.99	0.05
Cancer site						
Lung	*	*	*	*	*	*
Colorectal	0.06	0.98	0.05	0.05	0.99	0.05
Age at diagnosis, (years)						
21-55	*	*	*	*	*	*
56-70	0.47	0.90	0.08	2.44	0.44	0.68
71-80	0.80	0.81	0.13	2.18	0.48	0.59
$\geq$ 81	0.61	0.84	0.09	0.69	0.82	0.11
Sex						
Male	*	*	*	*	*	*
Female	1.34	0.65	0.27	0.12	0.97	0.05

* - reference group

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Cov. Err.: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table B.5: *(Continued) Data utility for parameters of the multinomial logit model for patient decision role.*

Characteristic	**Patient**			**Physician**		
	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**
Ethnicity						
White	*	*	*	*	*	*
Black	0.15	0.96	0.05	2.11	0.44	0.56
Hispanic	0.89	0.77	0.14	1.62	0.59	0.37
Asian	0.66	0.83	0.10	0.25	0.94	0.05
Other	0.77	0.80	0.12	1.24	0.68	0.24
Marital status						
Married	*	*	*	*	*	*
Not married	0.12	0.95	0.05	2.63	0.36	0.75
Education						
< High school (HS)	0.75	0.81	0.12	0.41	0.89	0.07
HS graduate or some college	*	*	*	*	*	*
College degree or higher	0.10	0.95	0.05	1.11	0.71	0.20
Income $						
< 20,000	0.30	0.93	0.06	1.28	0.69	0.25
20,000 to < 40,000	0.29	0.93	0.06	0.92	0.77	0.15
40,000 to < 60,000	0.08	0.98	0.05	0.44	0.89	0.07
≥ 60,000	*	*	*	*	*	*
No. self-reported co-morbid conditions						
0	*	*	*	*	*	*
1	0.04	0.99	0.05	0.05	0.99	0.05
2	0.03	0.99	0.05	0.03	0.99	0.05
≥ 3	0.01	1.00	0.05	0.01	1.00	0.05

* - reference group

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Cov. Err.: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Table B.5: *(Continued) Data utility for parameters of the multinomial logit model for patient decision role.*

Characteristic	**Patient**			**Physician**		
	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**	$\left\lvert\frac{Bias_q}{\sqrt{V_q}}\right\rvert$	**CI. over-lap**	**Cov. Err.**
Prediag. health status						
Quartile 1	*	*	*	*	*	*
Quartile 2	0.02	0.99	0.05	0.14	0.96	0.05
Quartile 3	0.05	0.99	0.05	0.17	0.96	0.05
Quartile 4	0.02	0.99	0.05	0.04	0.99	0.05
CES-D short form						
≤ 5	*	*	*	*	*	*
≥ 6	0.07	0.98	0.05	0.09	0.98	0.05
Study site						
Los Angeles county	*	*	*	*	*	*
Alabama	0.24	0.94	0.06	0.28	0.93	0.06
8 counties in North California	0.13	0.97	0.05	0.23	0.94	0.06
22 counties in eastern North Carolina	0.24	0.94	0.06	0.03	0.99	0.05
Iowa	0.16	0.96	0.05	0.27	0.93	0.06
5 HMOs	0.13	0.97	0.05	0.11	0.97	0.05
15 Veterans Aff. hospitals	0.52	0.87	0.08	0.24	0.94	0.06
Summary averages						
Synthesized	0.60	0.84	0.11	1.35	0.66	0.33
Non synthesized	0.11	0.97	0.05	0.20	0.95	0.07
All	0.27	0.93	0.07	0.56	0.86	0.15

* - reference group

$\left\lvert\frac{\mathbf{Bias_q}}{\sqrt{\mathbf{V_q}}}\right\rvert$: standardized bias, Section 6.1.3

CI. overlap: confidence interval overlap, (6.6)

Cov. Err.: estimated error in coverage of nominal 95% confidence interval, (6.7) & (6.8)

Bibliography

[1] Abowd, J. and Woodcock, S.D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer and V. Torra, New York: Springer-Verlag, 290-297.

[2] Abowd, J. Stinson, M. and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical Report, *US Census Bureau Longitudinal Employer Household Dynamics Program*,

[3] Ayanian J.Z., Chrischilles E.A., Wallace R.B., Fletcher R.H. et al. (2004) . Understanding cancer treatment and outcomes: The Cancer Care and Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 22 (15), 2992-2996.

[4] Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2006). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26 (6), 1368-1382.

[5] Caiola G. and Reiter J.P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3 (1), 27-42.

[6] Christiansen, C. L. and Morris, C. N. (1997). Hierarchical poisson regression modeling. *Journal of the American Statistical Association*, 92 (438), 618-632.

[7] Cochran, W. G. (1977). *Sampling techniques*, Wiley: New York.

[8] Dempster, A.P. and Rubin, D.B. (1983). Rounding error in regression: The appropriateness of sheppard's corrections. *Journal Royal Statistical Society, Series B*, 45 (1), 51-59.

[9] Deville, J.C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85 (1), 89-101.

[10] Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008). A new approach for disclosure control in the IAB Establishment Panel - Multiple imputation for a better data access. *Advances in Statistical Analysis*, 92 (4), 439-458.

[11] Drechsler, J. and Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer and Y. Saygin, New York: Springer-Verlag, 227-238.

[12] Drechsler, J. and Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use Census microdata. *Journal of the American Statistical Association*, 105 (492), 1347-1357.

[13] Duncan, G.T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7 (2), 207-217.

[14] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 (456), 1348-1360.

[15] Fienberg, S.E. Makov, U.E. and Sanil, A.P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*, 13 (1), 75-89.

[16] Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.

[17] Gates, G.W. (2011). How uncertainty about privacy and confidentiality is hampering efforts to more effectively use administrative records in producing US national statistics. *Journal of Privacy and Confidentiality*, 3 (2), 3-40.

[18] Graham, P., Young, J. and Penny, R. (2008). Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. *Journal of Official Statistics*, 25 (2), 245-268.

[19] Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.

[20] He, Y., Zaslavsky, A. Landrum, M., Harrington, D. and Catalano, P. (2010). Multiple imputation in a large-scale complex survey: a practical guide. *Statistical Methods in Medical Research*, 19 (6), 663-670.

[21] Huskamp et al. (2009). Discussions with physicians about hospice among patients with metastatic lung cancer. *Arch Intern Med*, 169 (10), 954-962.

[22] Karr, A.F. Kohnen, C.N. Oganian, A., Reiter, J.P. and Sanil, A.P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60 (3), 224-232.

[23] Keating et al. (2010). Cancer patients' roles in treatment decisions: do characteristics of the decision influence roles?. *Journal of Clinical Oncology*, 28 (28), 4364-4370.

[24] Kennickell, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Consumer Finances. In *Record Linkage Techniques*, eds. W. Alvey and B. Jamerson, Washington, DC: National Academy Press, 248-267.

[25] Kinney S.K and Reiter J.P. (2007). Making public use synthetic files of the Longitudinal Business Database. In *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.

[26] Kinney S.K and Reiter J.P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and partial synthesis. *Journal of Official Statistics*, 26 (2), 301-315.

[27] The Joint Canada/United States Survey of Health. URL=<http://www.cdc.gov/nchs/nhis/jcush.htm>.

[28] Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9 (2), 407-426.

[29] Lohr, S. (1999). *Sampling: Design and analysis*. Brooks/Cole Publishing Company: Pacific Grove, California.

[30] Lumley, T. (2011). Package 'survey'. URL=<http://faculty.washington.edu/tlumley/survey/>.

[31] Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9 (4), 538-558.

[32] Midzuno, H. (1952). On the sampling system with probability proportional to sum of size. *Annals of the Institute of Statistical Mathematics*, 3 (1), 99-107.

[33] Morris, C. N. (1983). Parametric empirical bayes confidence intervals. In *Scientific Inference Data Analysis, and Robustness*, eds. G. E. Box, T. Leonard and C.F. Wu, vol. 48. Mathematics Research Center, University of Wisconsin, Madison, Wisconsin, 25-50.

[34] Morris, C.N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian Statistics 3*. Oxford University Press, Oxford, 327-344.

[35] Morris, C.N. and Lock, K.F. (2009). Unifying the named natural exponential families and their relatives. *The American Statistician*, 63 (3), 247-253.

[36] Morris, C.N. and Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science*, 26 (2), 271-287.

[37] Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6 (4), 487-500.

[38] R Project for Statistical Computing, n.d. $http://www.r-project.org/$. *Survey Sampling* package (2011).

[39] Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27 (1), 85-96.

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

[41] Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic datasets. *Journal of Official Statistics*, 18 (4), 531-543.

[42] Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29 (2), 181-189.

[43] 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 (1), 185-205.

[44] Reiter, J.P. (2005b). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21 (3), 441-462.

[45] Reiter, J.P. (2005c). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100 (472), 1103-1113.

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

[47] Reiter, J.P. (2009). Multiple imputation for disclosure limitation: future research challenges. *Journal of Privacy and Confidentiality*, 1 (2), 223-233.

[48] Reiter, J.P. and Mitra, R. (2009). Estimating risks of identification and disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1 (1), 99-110.

[49] Reiter, J.P. and Drechsler, J. (2010). Releasing multiply imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 20 (1), 405-421.

[50] Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, New York: John Wiley & Sons, Inc.

[51] Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9 (2), 461-468.

[52] Särndal, C., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag: New York.

[53] Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

[54] Sheppard, W. F. (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proceedings London Math. Soc*, 29 (1), 353-380.

[55] Skinner, C.J. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal American Statistical Association*, 103 (483), 989-1001.

[56] Su, Y.S., Gelman, A., Hill, H. and Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45 (2), 1-31.

[57] Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58 (1), 267-288.

[58] Van Buuren S., and Oudshoorn, C.G.M. (2000). *Multivariate imputation by chained equations: MICE v1.0 user's manual*, TNO: Leiden.

[59] Wilson E.B. and Hilferty M.M. (1931). The distribution of chi-squares. In *Proceedings of National Academy of Sciences USA*, 17, 684-688.

[60] Woo, M.J., Reiter, J.P., Oganian, A., and Karr, A.F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1 (1), 111-124.

[61] Yucel R.M., He Y. and Zaslavsky A.M. (2008). Using calibration to improve rounding in multiple imputation. *The American Statistician*, 62 (2), 125-129.

[62] Yucel, R.M, He, Y. and Zaslavsky A.M. (2011). Imputation of categorical variables using gaussian based routines. *Statistics in Medicine*, 30 (29), 3447-3460.