Published in final edited form as:

Stat Med. 2013 October 30; 32(24): 4139–4161. doi:10.1002/sim.5841.

Disclosure Control using Partially Synthetic Data for Large-Scale Health Surveys, with Applications to CanCORS

Bronwyn Loong^{a,*†}, Alan M. Zaslavsky^b, Yulei He^b, and David P. Harrington^c

^aResearch School of Finance, Actuarial Studies and Applied Statistics, The Australian National University, Canberra, ACT 0200, Australia

^bDepartment of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

^cDepartment of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 3 Blackfan Circle, Boston, MA 02215, U.S.A.

Abstract

Statistical agencies have begun to partially synthesize public-use data for major surveys to protect the confidentiality of respondents' identities and sensitive attributes, by replacing high disclosure risk and sensitive variables with multiple imputations. To date, there are few applications of synthetic data techniques to large-scale healthcare survey data. Here, we describe partial synthesis of survey data collected by CanCORS, a comprehensive observational study of the experiences, treatments, and outcomes of patients with lung or colorectal cancer in the United States. We review inferential methods for partially synthetic data, and discuss selection of high disclosure risk variables for synthesis, specification of imputation models, and identification disclosure risk assessment. We evaluate data utility by replicating published analyses and comparing results using original and synthetic data, and discuss practical issues in preserving inferential conclusions. We found that important subgroup relationships must be included in the synthetic data imputation model, to preserve the data utility of the observed data for a given analysis procedure. We conclude that synthetic CanCORS data are suited best for preliminary data analyses purposes. These methods address the requirement to share data in clinical research without compromising confidentiality.

Keywords

data confidentiality; data utility; disclosure risk; multiple imputation; synthetic data

1. Introduction

The Cancer Care Outcomes Research and Surveillance (CanCORS) project [1] is a large-scale study of services delivered to and outcomes of care for newly diagnosed patients with

Copyright © 2012 John Wiley & Sons, Ltd.

*Correspondence to: Bronwyn Loong, Research School of Finance, Actuarial Studies and Applied Statistics, College of Business and Economics Building 26C, The Australian National University, Canberra, ACT 0200, Australia.. [†]bronwyn.loong@anu.edu.au.. Bronwyn Loong was a doctoral student in the Department of Statistics at Harvard University when this research was conducted.

lung and colorectal cancer. Population and healthcare system based cohorts were recruited at eleven study sites in the United States (US). Patients were surveyed by telephone gathering information on the care received, patient-reported outcomes, and patient preferences and behaviors. Additional data were obtained from physician surveys and medical records. The wide scope of data collection and numerous measurements provide opportunities for research on multiple topics by the CanCORS Consortium and external investigators.

Under requirements established by the funding agency, CanCORS must provide for access by researchers outside the Consortium. Because of the high risk of re-identification of the distinctive CanCORS population, access to the original data must be limited by protective restriction and constraints on the information content, that impose administrative burdens on potential users or conflict with research needs. Data sharing may pose similar challenges to access to other health and healthcare data sets.

Replacing observed values with synthetic data before public release was first proposed by Rubin [2] based on the theory of multiple imputation [3]. 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 the relationships among observed survey variables, and estimation methods based on the concepts of multiple imputation, analysts can draw valid inferences about the target population of interest using standard methods, without accessing the original microdata. Partially synthetic data sets combine synthetic values for sensitive variables or key identifiers with the originally observed values for all other data points [4].

Several US statistical agencies have been or are active in releasing or developing partially synthetic public data sets, predominantly for socioeconomic microdata. Among these are the Survey of Consumer Finances [5], the American Communities Survey [6], Survey of Income and Program Participation [7], US Longitudinal Business Database [8], and the Longitudinal Employer-Household Dynamics survey [9]. 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 [10]. To date, application of partially synthetic data techniques to disclosure control for healthcare survey data has been limited.

Disclosure control for healthcare survey data entails some interesting challenges and innovations. Firstly, any modification to the observed data set must maintain variable relationships that are clinically plausible. Secondly, external users of healthcare data may have had limited exposure to synthetic data for disclosure control, and must be persuaded of its validity.

In what follows, Section 2 reviews methods for the creation and analysis of partially synthetic data, and working measures for data utility and identification disclosure risk assessment. Section 3 discusses our selection of variables to synthesize, and specifies our imputation models. In Section 4 we compare analytic results using observed and synthetic data based on two illustrative published studies conducted by CanCORS investigators. In

Section 5 we quantify and discuss the identification disclosure risk of the synthetic data we created. Section 6 provides some concluding remarks.

2. Background on partially synthetic data

2.1. Inference with partially synthetic data

Inferential methods for partially synthetic data were derived in Reiter [4] and are reviewed here. Let $S_i = 1$ if unit i is selected to have any of its observed values replaced, and let $S_i = 0$ otherwise. Let $S_0 = (S_1, \dots, S_n)$, 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 other values. Let $Y_{\text{rep}}^{(j)}$ be the replacement values for Y_{rep} in synthetic data set j . Each $Y_{\text{rep}}^{(j)}$ is generated by simulating values from the posterior predictive distribution $f(Y_{\text{rep}}^{(j)} | Y_0, S_0)$, or some close approximation to the distribution as in Raghunathan et al. [11]. The data producer or ‘imputer’ repeats the process m times creating synthetic data sets $D^{(j)} = (Y_{\text{nrep}}, Y_{\text{rep}}^{(j)})$ for $j = 1, \dots, m$, and releases the collection of synthetic data sets $D = (D^{(1)}, \dots, D^{(m)})$. Multiply-imputed data sets are required to capture the uncertainty in relationships among survey variables when drawing from the posterior predictive distribution. Here the ‘imputer’ is the CanCORS Statistical Consulting Center, and investigators outside the Consortium are ‘analysts’ or ‘public-users’.

To obtain valid inference for a scalar estimand Q , analysts can use the combining rules presented by Reiter [4]. Suppose that using 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 $q^{(j)}$ and $v^{(j)}$ be the complete-data estimates for Q from synthetic data set $D^{(j)}$. The analyst computes $q^{(j)}$ and $v^{(j)}$ by treating $D^{(j)}$ like observed data.

The point estimate of Q using the synthetic data is

$$\bar{q}_m = \sum_{j=1}^m q^{(j)} / m. \quad (1)$$

The estimated variance of $\bar{q}_m$ is

$$T_p = b_m / m + \bar{v}_m, \quad (2)$$

where $b_m = \sum_{j=1}^m (q^{(j)} - \bar{q}_m)^2 / (m - 1)$ and $\bar{v}_m = \sum_{j=1}^m v^{(j)} / m$. Inferences for scalar Q , when n is large, can be based on a t -distribution with degrees of freedom $\nu_m = (m - 1) (1 + r_m^{-1})^2$, where $r_m = (m^{-1} b_m / \bar{v}_m)$. Methods for multivariate inferences appear in Reiter [12].

2.2. Disclosure risk

Disclosure risk arises from the potential identification of sampled units in the released data (*identification disclosure*, [13], [14]), and inference of new information about a known participant in the survey (*inferential disclosure*, [15]). Inferential disclosure can occur even if no released record is correctly associated with the respondent, and the new information is inexact. For example, suppose there are ten potential matches in the synthetic data for unit i in the original data set, and all ten have similar values for income. The intruder can infer income level for unit i without knowing the true match. Inferential disclosure risk assessment methods were considered in Duncan and Lambert [16] for tabular data masked by traditional disclosure control techniques such as aggregation and cell suppression. Methods to evaluate the risk of inferential disclosure for partially synthetic microdata are still being developed, and thus, we have evaluated only the identification disclosure risk (hereon referred to as identification risk) of the synthetic data we generated. We recognize that evaluation of inferential disclosure risk is an important area of future research.

To compute probabilities of identification disclosure, we use the Duncan-Lambert risk framework [17], with related approaches in Fienberg, Makov and Sanil [18], Reiter [19] and Reiter and Mitra [20]. 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.
- b. The intruder knows one of 3 sets of identifying variables that illustrate identification risk at 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 comprises a large fraction of 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 determination of appropriate samples, approximately 5,000 patients were surveyed for each cancer type; that is, between 30% and 40% of the target population. Identification risk is not well protected by random sampling, because of the high probability that a unit in the target population was sampled for inclusion in the observed data set [17], [18], [19].

The variable sets listed in assumption (b) do not constitute unique identifiers (like name, address, date of birth, case ID), but best approach unique identification among the variables available for public release; hence they are called quasi-identifiers. A potential identification risk for target record i occurs when its quasi-identifying values match the corresponding values for a record k ($k = 1, \dots, n$), in synthetic data set j ($j = 1, \dots, m$). The risk is potential because there will be other synthetic data records with matching values for the identifying variables, either true or synthesized, so the intruder does not know whether or not a match is

correct. Furthermore, for a given target unit, the set of matching records will vary across the multiply-imputed data sets.

Let $R_{ijk} = 1$ (for $i = 1, \dots, n; j = 1, \dots, m; k = 1, \dots, n$) if the quasi-identifying information for original unit i matches the quasi-identifying information of unit k in synthetic data set j , and $R_{ijk} = 0$ otherwise. For $k = i$, let $C_{ij} = R_{iji}$, which indicates whether original unit i matches its synthesized version in synthetic data set j . Define $F_{ij} = \sum_{k=1}^n R_{ijk}$ to be the number of records in synthetic data set j , that match the quasi-identifying information of original unit i .

Define the following identification disclosure risk measures:

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

$$MXM = \sum_{i=1}^n \sum_{j=1}^m C_{ij}, \quad (3)$$

The maximum number of matches is reached if the intruder always correctly selects the target from the F_{ij} candidates.

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

$$EMR_i = \sum_{j=1}^m \frac{1}{F_{ij}} \times C_{ij}, \quad (4)$$

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

The overall expected match risk rate is

$$EMR = \sum_{i=1}^n EMR_i. \quad (5)$$

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

$$TMR_i = \sum_{j=1}^m K_{ij}, \quad (6)$$

where $K_{ij} = 1$ when $F_{ij} = 1$ and $C_{ij} = 1$, and $K_{ij} = 0$ otherwise.

The overall true match risk is

$$TMR = \sum_{i=1}^n TMR_i, \quad (7)$$

and reflects the intruder correctly and uniquely identifying records.

2.3. Data utility

In a broad sense, the utility of a particular data release is the benefit to society of the released information [21]. A more quantitative definition might characterize the quality of 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 [22], or can be broadened to global differences in distributions [21]. The confidence interval overlap measure is commonly cited, and is defined in Karr et al. [22] as follows:

For a point estimate q of a scalar estimand Q , the confidence interval overlap is calculated by:

$$J_q = \frac{1}{2} \left(\frac{U_{\text{over},q} - L_{\text{over},q}}{U_{0,q} - L_{0,q}} + \frac{U_{\text{over},q} - L_{\text{over},q}}{U_{\text{syn},q} - L_{\text{syn},q}} \right), \quad (8)$$

where $(L_{0,q}, U_{0,q})$ 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, which could be the result of a wide or poorly centered synthetic data confidence interval.

To better identify the effect of bias in synthetic data estimates relative to observed data set results, we introduce a new measure for data utility assessment. Define

$$\frac{\text{Bias}}{\sqrt{T_p}} = \frac{\bar{q}_m - q_0}{\sqrt{T_p}}, \quad (9)$$

as the standardized bias of the synthetic data estimate, where T_p is the sampling variance estimate of $\bar{q}_m$ (2). Following Cochran [23] Section 1.8 p. 13, we can compute the effect of bias on the nominal coverage of a 95% confidence interval; that is, the probability of an error greater than $1.96 \times \sqrt{T_p}$ where error is measured from the observed data set estimate q_0 . The upper tail error probability is given by

$$P_{\text{Uq,err}} = \frac{1}{\sqrt{2\pi}} \int_{1.96 - (\text{Bias}/\sqrt{T_p})}^{\infty} e^{-t^2/2dt}, \quad (10)$$

and the lower tail error probability is given by

$$P_{\text{Lq,err}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1.96 - (\text{Bias}/\sqrt{T_p})} e^{-t^2/2dt}, \quad (11)$$

The total error probability is $P_{\text{Totq,err}} = P_{\text{Lq,err}} + P_{\text{Uq,err}}$, which should be 0.05 for an unbiased nominal 95% confidence interval estimate.

3. Application to the CanCORS patient survey data set

3.1. Identification of variables to synthesize

To protect confidentiality, high identification risk and sensitive variables might be synthesized. Determining the variables that pose identification risk is labor intensive with limited guidance in the research literature. The general strategy adopted for our study is based on the strategy in Hawala [6], with additional consideration for the clinical variables in our healthcare study.

There were fifteen direct identifiers (including the patient's name, postal address and social security number) defined by the CanCORS Institutional Review Board that are unavailable to both internal and external investigators. For the remaining survey variables, we defined high identification risk variables (either individually or in combination) as those 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 intruder's external information source. We identified the demographic variables age (grouped into 5 year age bands), marital status, race, sex, and education as quasi-identifiers to synthesize.

Clinical variables in combination with demographic variables might pose high identification risk. We chose to retain their original values for the following reasons: Firstly, clinical variables have a very complex joint structure with many categorical variables, logical relationships, and structural skip (missingness) patterns; this is much harder to synthesize in a plausible way than the limited set of demographic variables, which furthermore have no structural missingness. Secondly, the clinical variables (unlike the demographic variables) are likely to be known only to an intruder who already has much of the patient's private information, thereby limiting exposure to re-identification by outside parties, and particularly limiting exposure to massive identification disclosure. Thirdly, some further protection is afforded to the extent that clinical variables are subject to judgmental variations in record abstraction and incomplete medical record acquisition, limiting their value for re-identification.

Our decision not to synthesize clinical variables means we have assumed that clinical variables would not be available in the intruder's external information source; that is, the intruder comes from the general population who do not have access to confidential clinical data. Extending our identification risk assessment, to incorporate explicitly the external databases available to clinical and other health professionals for re-identification purposes, is an item of future research. We did include the clinical variable 'disease stage' in our identification risk assessment, however, as a component of quasi-identifying Set 3.

From the respondent's perspective, all information gathered can be considered sensitive, in particular, health insurance details, grouped income levels and medical history. If sensitive information for a target unit is disclosed, this constitutes an inferential disclosure. This suggests we should synthesize all variables, but full synthesis generally reduces data utility and increases the reliance on the imputation model to produce plausible relationships. Guidelines to define sensitive variables and methods for inferential disclosure risk

assessment are still in development, as mentioned in Section 2.2. We have synthesized only high identification risk variables in this paper.

3.2. Imputation models

Imputation models were fitted separately for lung and colorectal cancer because relationships among variables might differ by cancer type. To create synthetic data sets, we want to simulate values from the joint posterior predictive distribution

$$f\left(Y_{\text{age}}^{(j)}, Y_{\text{educ}}^{(j)}, Y_{\text{marstat}}^{(j)}, Y_{\text{race}}^{(j)}, Y_{\text{sex}}^{(j)} | Y_0, S_0\right), \quad (12)$$

where Y_0 includes observed values ($Y_{\text{age}}, Y_{\text{educ}}, Y_{\text{marstat}}, Y_{\text{race}}, Y_{\text{sex}}$). All variables were defined as unordered categorical variables for imputation, collapsing adjacent factor levels (see Appendix A, Table 7) to avoid sparse cell counts, stabilizing parameter estimates, and for consistency with recodings commonly used in analyses.

To capture variable relationships of importance to analysts, a large number of predictors are desired in the imputation model to approximate (12). The observed data Y_0 is composed of approximately 500 variables, and including all of these as predictors is impractical. The number of variables to be included in the partially synthetic data was reduced to approximately 350 based on the following exclusions:

- Variables containing names or addresses of healthcare providers.
- Variables on consumption of specific types of alternative therapy, vitamins and herbal supplements, although we retained the general indicator for usage of these services and products.
- Variables neither associated with, nor a consequence of the active patient treatment plan, such as recollection of symptoms at notification of cancer.

The first criterion protects data confidentiality prior to application of any statistical disclosure control method. The second and third criteria were applied to avoid issues with multicollinearity [10, 20], such as including indicators for both symptoms experienced at diagnosis and in the last 4 months. We assumed that excluded variables are not of interest to analysts.

The joint distribution in (12) defines a multivariate distribution for categorical variables. However, there are twenty-five factor levels to be imputed at one time, and it is not possible to draw directly from the joint distribution in (12) without encountering limitations due to the analytic effort required to consider all necessary relationships, and the adequacy of the data to define all relationships in the joint model. The alternative is to generate imputations variable by variable to approximate the joint distribution using sequential regression multiple imputation (SRMI [24], [11]), iteratively imputing values until approximate convergence. We chose a parametric approach to imputation and used the software package for multiple imputation ‘mi’ by Su et al. [25] in the R [26] software and computing environment. Imputation for sex was based on a logistic regression model. Imputations for all other variables were based on multinomial logit models. Sample code is available upon request.

Our imputation models are consistent with those for similar demographic variables synthesized in previous studies [8, 20, 27, 28]. However, we did not consider the ordering of age-education to be important in our imputation models because the patients in the CanCORS survey are predominantly aged 55 or older, and the association that older age implies more education does not apply. However, we did check the consistency of age-education cross-tabulation counts between the original and synthetic data, and found there to be no significant differences beyond random sampling error. Five partially synthetic data sets were created ($m = 5$), following the minimum number of imputations used in previous applied studies to create partially synthetic data. Increasing the number of imputations to $m = 10$ gave us no improvement in data utility.

To select the predictors for each imputation model, a stepwise regression (considering both forward and backward selection at each stage) was conducted within each of the 12 sections of the survey questionnaire, identifying the subspace of predictors most correlated with the response variable of interest in the observed data set by minimizing Akaike's Information Criterion (AIC). We ran the stepwise regression by section to include important predictors from each section. The selected models included on average about 50 predictor variables. The stepwise regression was conducted only once, prior to any imputations, and not repeatedly at each iteration of the imputation algorithm. We did not consider variable selection approaches for high-dimensional data, such as the LASSO [29] or SCAD [30], because we do not anticipate that external analysts of CanCORS data will use high-dimensional data analysis methods. To avoid the potential for overfitting, which increases disclosure risk, no interactions were included in the imputation models, only main effects. This also helped to control the computational burden, and avoid multicollinearity and separation issues if linear combinations of our predominantly discrete predictor variables are perfectly aligned with the categorical response.

To draw from each variable's posterior predictive distribution, we first draw values of the observed model parameters from their approximate posterior distribution, and then generate synthetic values given the drawn posterior values of the parameters and the selected observed predictor data from the results of the stepwise regression. Non-informative prior distributions were assumed for all parameters, to illustrate the utility of generic methods that do not rely on topic-specific prior information. Moreover, non-informative prior distributions are required for valid application of the inferential methods proposed by Reiter [4], described in Section 2.1 of this paper. The following variables were always included in the imputation models, however: version of the survey (minor modifications in readability led to small changes), cancer stage, cancer histology, vital status at time of interview (alive/dead/unknown) and study site, because these variables are generally used as control variables in analyses.

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 [31], to make use of the existing functionality in the 'mi' package for multivariate imputation for continuous data. This calibration-based rounding strategy preserves observed distributions of categorical variables when imputing from a multivariate normal distribution.

4. Data utility for the partially synthesized data

We compare analytic results based on two published analyses of the observed CanCORS data set by internal CanCORS investigators: Huskamp et al. [32] and Keating et al. [33]. These two papers were chosen because they are important in themselves and as examples of the types of analyses of interest to external researchers. Our aim is to illustrate interpretation of the data utility measures described in Section 2.3; for the subject-matter impact of the analyses, see the original articles.

Results in this section rely on the inferential methods for partially synthetic data described in Section 2.1. Because we expect the data utility results for non-synthesized variables to be better than those for synthesized variables, we also report summary results by this grouping.

4.1. A logistic regression model for probability of hospice discussion

Huskamp *et al.* [32] estimated a logistic regression model from 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 healthcare provider could increase awareness of hospice, and possibly result in earlier use.

The analysis in Huskamp *et al.* [32] was based on five multiply-imputed data sets for missing values [34]. Our imputations and analysis are based on the partial synthesis of one of these five data sets. Hence, the observed data set results reported here closely but not exactly match those previously reported. (Combining rules for hypotheses tests when using multiple imputation simultaneously for missing data and partial synthesis are detailed in Kinney and Reiter [35]). Given the low item non-response rates ($< 3\%$ [34]), it is unlikely that the synthetic data results will be sensitive to ignoring the imputation of missing data.

Table 1 describes characteristics of synthesized variables and estimated probabilities of hospice discussion for Stage IV lung cancer patients by patient subgroup, unadjusted for other covariates. Differences between observed and synthetic data marginal subgroup proportions were generally $< 1\%$, with the exception of sex. The synthetic data sample proportions by sex are consistent, however, with the observed data set sample proportions using the entire data set of 5,000 records, conditioned on in our imputations. For most subgroups, synthetic data estimated probabilities of hospice discussion deviated by less than 3% from observed data set estimates, but deviations were greater than 5% for some race and age subgroups. Synthetic data standard errors were marginally greater than observed data set standard errors due to the additional between-imputation variability, $b_m(2)$. The conclusion of significance for ‘race’ changes from strongly significant ($p\text{-value} < 0.001$) to insignificant ($p\text{-value} = 0.62$). We attributed this discrepancy to different sets of records being used between the imputation model and analysis procedure, as discussed further below.

Subgroup probabilities of hospice discussion adjusted for other covariates are reported in Table 2. For non-synthesized variables, estimated probabilities using synthetic data differed on average by 1.5% from the observed data set estimates. There were non-zero discrepancies because the model fit also depends on synthesized data. The average deviation for

synthesized variables was 4.1%, with race and age subgroup estimated probabilities showing the largest deviations. Again there was a change in significance at conventional levels for race. For all other covariates, conclusions of significance were preserved.

Data utility measures with respect to the parameters of the adjusted hospice logistic model are reported in Table 3. The average absolute standardized bias of non-synthesized variable estimates 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 subgroup 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. That is, we have not synthesized the clinical variable hospice discussion, but we have preserved its association with some of the synthesized demographic variables. Data utility was not preserved however, for age and race covariates. Meng [36] coined the term *uncongeniality* to describe situations in multiple imputation where the analyst and the imputer have access to different types of information and data, and assess and use the information and data in different ways. In other words, there are systematic differences between the imputation model and analysis procedure inputs. In our example, the general imputation model was derived from the entire data set, but the analysis procedure used the subset of Stage IV lung cancer records. If the imputation model does not capture all the important subgroup relationships, results from the synthetic data may be biased. In our application of the inferential equations described in Section 2.1, we assumed that the imputation model and analysis procedure inputs were congenial.

To investigate further, we ran the imputation models again but only on the 1,517 Stage IV lung cancer units in the analysis procedure (and not the entire data set). Appendix B, Table 8 reports the revised estimates for synthesized variables, unadjusted for other covariates. Conclusions of significance were identical to those from the observed data set, and deviations of synthetic from observed data estimated probabilities were generally less than 1.5%.

Revised hospice discussion model estimates adjusted for other covariates, and data utility measures are in Appendix B, Tables 9 - 10. Synthetic data estimates of hospice discussion probabilities differed by 4% on average (equivalent to 0.5 standard errors) from observed data estimates, for both non-synthesized and synthesized variables. Deviations for age and race subgroups were less than 5%. The largest deviation (10%) was for the non-synthesized 'non-English speaking' subgroup, but non-English speakers represented < 5% of the analyzed sample, and larger discrepancies are more likely for smaller observed subgroups. There were no differences in conclusions of significance. The average absolute standardized bias of 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 increased for age subgroups '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 bias for the 'black' race subgroup estimate was removed, as well as smaller bias for non-synthesized variable estimates. Furthermore, any remaining discrepancies noted for the age subgroup estimates in Tables 8 - 10, were found to be removed after conditioning on hospice discussion in the imputation model for age.

4.2. A multinomial logistic regression model for cancer patients' roles in treatment decisions

A multinomial logistic regression based on the study by Keating *et al.* [33] assessed the effects of several predictors on patient role in treatment decision. The authors argue that patients with more active roles in decisions are more satisfied and may have better health outcomes. Predictors included evidence about the treatment's benefits, whether the decision was likely preference-sensitive, and treatment modality. Table 4 shows how the response options for role in treatment decision-making were categorized for analysis.

The analysis in Keating *et al.* [33] used 10,939 decisions from 5,383 lung and colorectal cancer patients who self-completed the full patient survey, and who had discussed at least one treatment (surgery, radiation or chemotherapy) with a clinician. The unit of analysis was the decision about one of these treatments. Standard errors were adjusted to account for correlation among repeated decisions by a single patient, by using a robust variance estimator [37] in the *Stata* software package. We adopted the same statistical methods for analysis of each synthetic data set. As in the previous subsection, our results are based on partial synthesis of one of the five sets multiply imputed for missing values [34], and approximate analytic results reported in Keating *et al.* [33].

Appendix B, Table 11 reports synthetic and observed data estimated proportions reporting each decision role category by various patient and tumor characteristics, adjusted for all other covariates fit in the same model. Estimates are relative to the reference group for each characteristic. Ethnicity, marital status and education were not significant predictors at conventional levels using synthetic data, but were significant in the observed data set. Conclusions of significance were preserved for all non-synthesized variables, including all clinical variables, with minimal deviations in the magnitude of the p-values. Data utility results for coefficients of the decision role model are reported in Appendix B, Table 12. 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.

To explain the synthetic data deviations from original data results, first note that we selected predictors for our imputation models from a larger set of variables than modeled in [33]. Hence, some variables relevant to the analysis may not be included in the imputation

models, and any associated significant relationship is attenuated. For example, decision role for any type of treatment was not selected as a predictor in any imputation model except for education. Nevertheless, we still observed a change in conclusion of significance for education. A second reason is that the unit of analysis in the imputation model was patient, not decision as assumed in the analysis procedure. Without accounting for the within-patient correlation in the imputation model, variable relationships imputed at the patient level may be attenuated from those inferred at the decision role level. Hence we observed decreased significance for some synthesized covariates.

Both our analytical comparisons used the same imputed synthetic data. Imputers generally run one synthesis model that aims to capture relationships of importance to all external analysts, because the imputer cannot anticipate all potential future analyses on the synthetic data by external investigators, and it is impractical to run separate synthesis models for each analysis procedure. Hence, there may be some variables or variable structures, or subgroup relationships of importance to analysts that are not included in the imputation models. As shown in our two analytic comparisons, this may lead to large discrepancies between observed and synthetic data analytic results.

5. Identification disclosure risk assessment

Tables 5 and 6 quantify the identification disclosure risk for the partially synthetic data we generated, using the measures described in Section 2.2. For comparison, we also calculated identification risk measures on the observed data set.

The maximum number of matches (MXM) is highest for quasi-identifier Set 1, because the fewest number of synthesized variables define a potential match, increasing the number of cases where the target is amongst the potential match candidates F_{ij} . The maximum number of matches does not change between quasi-identifier Sets 2 and 3, because the additional variables in Set 3 were not synthesized.

For the lung cancer cohort, as the number of variables in the quasi-identifier set increases, the expected match risk (EMR) also increases. Requiring more criteria to identify a match decreases F_{ij} , and if $C_{ij} = 1$, there are fewer candidates the intruder needs to randomly choose from to select the correct match. For colorectal cancer, the EMR for quasi-identifier Set 2 is 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 matches. This is evidence of strong identification risk protection from creating synthetic values for education or marital status.

The true match risk (TMR) is zero using quasi-identifier Set 1 for both lung and colorectal cancer synthetic data, meaning that no original record was correctly and uniquely identified as a match with any record, in any synthetic data set. The TMR is greatest for quasi-identifier Set 3 after including disease stage and study site. If intruders exist who do have access to such clinical information, then clinical variables need to be synthesized. Note that our quasi-identifier set assumptions are for illustration purposes, and further research is required to identify explicitly the external databases available to potential intruders for re-identification purposes.

To illustrate interpretation of the identification risk measures, consider the results for lung cancer, quasi-identifier Set 2. The true match risk is 88, that is, 88 cases where an original record was correctly and uniquely matched to a synthetic data set record. The 88 true match counts correspond to 76 records, because a record may be correctly and uniquely identified more than once across the five synthetic data sets. Of these, 68 records were correctly and uniquely identified in only 1 out of the 5 synthetic data sets, 6 were identified twice, 1 record was identified three times, and only 1 record was identified correctly and uniquely in all 5 synthetic data sets. The true match risk is ‘presumed’ because the intruder will not know which unique matches are correct. The overall expected match risk is 217 out of 5,000 patients: allowing for uncertainty from randomly guessing among the potential matches, approximately 4% of patients could be correctly identified. The maximum number of matches is 1158, which corresponds to 925 unique records (because the true record may be a match candidate more than once across the five synthetic data sets). That is, the worst case scenario is 925 units are identified if the intruder always correctly selects the target from the potential matches.

Comparing observed and synthetic data results, we conclude that identification risk is greatly reduced when releasing synthetic data instead of the observed values. The target unit is always in the set of potential matches for observed data, and therefore the expected and presumed true match risk values are larger.

There are no universal standards for acceptable levels of identification risk. Risk tolerance depends on many factors, including the risk attitude of the database owners, size of the target population and sampling fraction, the features of the data set which are vulnerable to identification disclosure, and realistic assessment and modeling of intruders’ strategies to identify targets. The results in Tables 5 and 6 illustrate one statistically based approach to quantify the identification risk of partially synthetic databases.

6. Concluding Remarks

We applied synthetic data techniques for data confidentiality protection in the public-data sharing plan of a large-scale healthcare survey. We synthesized high disclosure risk variables and employed stepwise variable selection to select predictors for our imputation models. We quantified the data utility and identification disclosure risk of the synthetic data we generated, using measures selected from current research. These are the key steps that need to be addressed when creating synthetic data for public use.

Our application highlights a number of areas where further research is required. Firstly, we used stepwise regression to select from about 350 potential covariates in the parametric imputation models. Non-parametric approaches such as classification and regression trees (CART) [38], and random forests [39] provide more flexibility to model relationships among a variety of variable types (categorical, numerical, and mixed), which may not be easily modeled by standard parametric approaches [40]. CART synthesis models were evaluated in Reiter [41], and used in empirical studies by Drechsler and Reiter [28] and Drechsler and Reiter [42]. Caiola and Reiter [43] described and illustrated using random forests to generate partially synthetic data. Further research is required to compare the

performance between parametric and non-parametric imputation approaches in meeting disclosure control and data utility criteria.

Secondly, we faced some challenges in our data utility assessment because of uncongeniality. Both our analytic comparisons suggested a variety of analyst-defined variables, statistical models and data subsets, which may be used in the analysis procedure. Considering the large number of potential public-users whose research questions concern subgroups or relationships not included in the imputation models, some large discrepancies between observed and synthetic data analytic results are likely. The imputer can try to fit complex imputation models that capture relationships of importance to analysts, but also must be careful not to release all information on the original data distribution. Some uncongeniality between the imputation model and analysis procedure may help to protect data confidentiality, because it makes it less likely that external users will be able to recover observed-data values. External users of synthetic data should be aware, that the objective of using synthetic data techniques is not to reproduce exactly observed data set results, but rather a balance is required between preserving inferential conclusions and minimizing disclosure risk. With this in mind, analysts may best use synthetic data for exploratory analyses purposes. Even with imperfect synthesis, analysts can gain familiarity with data structures, and formulate and test preliminary hypotheses, without the burdens involved in gaining full access to the observed data. With more experience with synthetic data techniques for statistical disclosure control, practical guidelines may be developed to deal with uncongeniality in applications.

Thirdly, our data utility assessment was focused on two analytic comparisons based on two published studies out of many possible analyses of CanCORS data. It remains to be determined which and how many representative published studies should be selected for analytical comparison, and how results of data utility assessment should be communicated to public-users. These questions are best answered on a case-by-case basis, depending on the analysis procedures of the pool of potential analysts of the synthetic data.

Finally, we used a basic intruder strategy and a statistical framework to quantify identification disclosure risk. Further work is required to quantify inferential disclosure risk for partially synthetic data, specifically, defining the events that characterize inferential disclosure, and modeling the behavior and thought processes of an intruder who attempts to make an inferential disclosure. Our re-identification assessment was based on exact matching between original records in CanCORS, and synthetic data records within each synthetic data set. Further work is required to investigate more sophisticated matching algorithms, which estimate the probability that a target record belongs to the sample space defined by the distribution of quasi-identifiers of a synthetic data set record. In addition, empirical research is required to identify explicitly the pool of potential intruders (for example, the general public or clinical professionals), and external data sources available to the intruder for disclosure attack purposes (for example, medical claims data or cancer registry data). Such research can improve the practical relevance of statistical approaches to disclosure risk assessment. There are many assumptions to make and modeling the uncertainty in these assumptions is also important. Limited research has been done on the

use of Bayesian prior distributions to capture prior intruder information, and this is another area for future investigation.

Acknowledgments

This research was supported by grants from the National Cancer Institute (NCI) to the Statistical Coordinating Center (U01 CA093344) and the NCI supported Primary Data Collection and Research Centers (Dana-Farber Cancer Institute/Cancer Research Network (U01 CA093332).

Appendix A: Recoded variable structure - CanCORS data

Table 7

Recoded variable structure for synthesis

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
Race	White	White
	African American	African American
	Hispanic or Latino	Hispanic or Latino
	Asian	Asian
	American Indian	Other
	Native Hawaiian	
	Alaskan Native	
	Other Pacific Islander	
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	

Variable	Original structure	Revised structure
	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/prof	
	Doctorate degree (J.D., M.D., Ph.D, etc) or > 2 years grad/prof	
Marital Status	Married	Married
	Divorced	Divorced
	Living with partner	Living with partner
	Never married	Never married
	Separated	Separated
	Widowed	Widowed

Appendix B: Supplementary analytic comparison results

Table 8

Descriptive characteristics and estimated probabilities of hospice discussion using synthesized variables, unadjusted results. Imputation model and analysis procedure use same subset of records. (Standard errors in parentheses)

Characteristic	Patients (%)		Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic
Sex					0.89	0.63
Male	61.3	62.4	53.3 (1.6)	53.6 (1.7)		
Female	38.7	37.6	53.0 (2.1)	52.3 (2.2)		
Race/ethnicity					< 0.001	0.003
White	73.7	74.6	55.2 (1.5)	55.6 (1.5)		
Black	10.7	9.9	42.6 (3.9)	42.0 (4.2)		
Hispanic	5.9	6.1	40.4 (5.2)	40.4 (5.3)		
Asian	5.1	5.1	49.4 (5.9)	50.2 (5.9)		
Other	4.7	4.3	64.5 (5.7)	58.5 (6.5)		
Married/live with partner					0.006	0.038
Yes	61.0	60.8	50.4 (1.6)	50.0 (1.7)		
No	39.0	39.2	57.6 (2.0)	58.1 (2.0)		
Age (yrs)					0.002	0.011
21-54	12.5	12.3	45.5 (3.6)	46.7 (3.8)		
55-59	12.4	12.1	52.1 (3.6)	50.0 (3.7)		
60-64	13.2	13.5	51.0 (3.5)	52.7 (3.8)		

Characteristic	Patients (%)		Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic
65-69	16.0	15.5	50.8 (3.2)	52.0 (3.6)		
70-74	17.8	18.6	50.4 (3.0)	52.0 (3.3)		
75-79	13.8	14.0	57.1 (3.4)	51.7 (3.9)		
80+	14.4	14.0	65.1 (3.2)	66.6 (3.3)		
Education					0.46	0.37
< High school	22.7	22.9	54.8 (2.7)	54.5 (2.9)		
High school/some college	60.6	62.2	53.5 (1.6)	53.8 (1.7)		
College degree	16.7	14.9	49.8 (3.1)	48.7 (3.4)		

Table 9

Estimated probabilities for hospice discussion, adjusted for other covariates. Imputation model and analysis procedure use same subset of records. (Standard errors in parentheses)

Characteristic	Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic
Sex			0.72	0.95
Male	54.5 (9.4)	60.0 (9.7)		
Female	53.2 (9.9)	59.7 (10.2)		
Race/ethnicity			< 0.001	0.018
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)		
Asian	60.0 (11.4)	63.3 (11.6)		
Other	78.2 (7.7)	74.8 (9.6)		
Married/live with partner			0.031	0.011
Yes	54.5 (9.4)	60.0 (9.7)		
No	62.6 (9.6)	69.2 (9.1)		
Age (yrs)			0.63	0.93
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)		
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			0.045	0.136
Yes	54.5 (9.4)	60.0 (9.7)		
No	35.7 (12.3)	45.9 (13.9)		
Education			0.87	0.86
< High school	56.3 (9.9)	58.0 (11.0)		
High school/some college	54.5 (9.4)	60.0 (9.7)		

Characteristic	Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic
College degree	56.3 (10.1)	61.6 (10.0)		
Income \$			0.19	0.23
< 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)		
60 000	54.5 (9.4)	60.0 (9.7)		
Insurance			0.018	0.015
Medicare	54.5 (9.4)	60.0 (9.7)		
Medicaid	48.3 (11.1)	54.3 (11.5)		
Private	57.3 (7.9)	61.8 (8.2)		
Other	77.1 (8.1)	81.3 (7.3)		
Treated in VA facility			0.43	0.44
Yes	46.6 (14.4)	52.3 (14.8)		
No	54.5 (9.4)	60.0 (9.7)		
HMO member			0.25	0.27
Yes	60.4 (9.3)	64.9 (9.5)		
No	54.5 (9.4)	60.0 (9.7)		
Region			0.12	0.12
South	60.7 (8.9)	64.8 (9.1)		
West	54.5 (9.4)	60.0 (9.7)		
Other	47.9 (10.0)	52.3 (10.6)		
Days from diagnosis to interview			0.055	0.100
Quartile 1	54.5 (9.4)	60.0 (9.7)		
Quartile 2	67.1 (8.2)	70.8 (8.3)		
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			< 0.001	< 0.001
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)		
180-239 d	5.9 (2.2)	7.3 (2.9)		
240 d	64.2 (8.9)	69.5 (8.7)		
Received chemotherapy before interview			0.006	0.005
Yes	54.5 (9.4)	60.0 (9.7)		
No	59.1 (9.3)	64.5 (9.1)		
Comorbidity			0.23	0.13
None	54.5 (9.4)	60.0 (9.7)		
Mild	57.1 (9.6)	62.6 (9.5)		
Moderate	64.7 (9.1)	70.8 (8.5)		
Severe	70.3 (9.4)	74.8 (9.1)		

Characteristic	Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic
Alive at survey but surrogate completed interview			0.004	0.004
Yes	61.7 (9.4)	66.2 (10.5)		
No	54.5 (9.4)	60.0 (9.7)		

Table 10

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

Characteristic	Coef. Est		Std. Bias	CI. overlap	Coverage error
	Observed	Synthetic			Total
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
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

Characteristic	Coef. Est		Std. Bias	CI. overlap	Coverage error
	Observed	Synthetic			Total
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
240 d	--2.94 (0.191)	--2.97 (0.197)	0.150	0.755	0.052
Received chemotherapy 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 at survey but surrogate completed interview					
Yes	0.677 (0.235)	0.684 (0.239)	0.029	0.746	0.050
Summary averages					
Synthesized			0.44	0.78	0.09
Non synthesized			0.14	0.75	0.05
Overall			0.25	0.76	0.07

$$\text{Std. Bias} = \left| \frac{\text{Bias}}{\sqrt{T_p}} \right| \text{ (standardized bias), (9)}$$

CI. overlap: confidence interval overlap, (8)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (10) & (11)

Table 11

Adjusted associations of patient and tumor characteristics with roles in decisions. (Standard errors in parentheses)

Characteristic	Estimated difference in proportion reporting						p-value	
	Patient Control		Shared Control		Physician Control		Observed	Synthetic
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic		
Level of evidence for treatment							< 0.001	< 0.001
Evidence for	*	*	*	*	*	*		
Uncertain	5.7 (1.4)	5.7 (1.4)	-4.2 (1.4)	-4.0 (1.4)	-1.6 (1.0)	-1.6 (1.0)		
No evidence for	-1.8 (1.8)	-1.7 (1.8)	-2.6 (1.7)	-2.6 (1.7)	4.4 (1.2)	4.3 (1.2)		
Missing	5.2 (2.7)	5.0 (2.7)	-3.8 (2.6)	-3.6 (2.6)	-1.4 (1.8)	-1.5 (1.8)		
Preference sensitive							< 0.001	< 0.001
No	*	*	*	*	*	*		
Yes	-6.5 (1.5)	-6.2 (1.5)	1.5 (1.6)	1.4 (1.6)	4.8 (1.2)	4.8 (1.2)		
Treatment modality							< 0.001	< 0.001
Surgery	*	*	*	*	*	*		
Radiation	-2.7 (1.2)	-2.6 (1.2)	2.2 (1.2)	2.2 (1.2)	0.5 (0.9)	0.4 (0.9)		
Chemotherapy	4.3 (0.9)	4.3 (0.9)	-1.4 (0.9)	-1.3 (0.9)	-2.9 (0.7)	-3.0 (0.7)		
Received treatment							< 0.001	< 0.001
No	*	*	*	*	*	*		
Yes	3.9 (1.4)	3.9 (1.4)	12.3 (1.3)	12.4 (1.4)	-16.2 (1.2)	-16.3 (1.3)		
Cancer site							0.54	0.59
Lung	*	*	*	*	*	*		
Colorectal	-0.7 (1.5)	-0.4 (1.6)	-0.5 (1.5)	-0.6 (1.5)	1.2 (1.0)	1.0 (1.1)		
Age at diagnosis (yrs)							0.19	0.92
21-55	*	*	*	*	*	*		
56-70	1.6 (2.1)	2.6 (3.5)	1.2 (2.1)	-2.7 (2.4)	-2.8 (1.4)	0.1 (2.0)		
71-80	1.4 (2.3)	1.9 (2.0)	0.9 (2.2)	-1.7 (2.4)	-2.2 (1.5)	-0.1 (1.6)		
81	1.4 (2.3)	0.6 (4.1)	-1.4 (2.3)	-2.5 (2.5)	0.1 (1.6)	1.8 (2.8)		
Sex							0.28	0.85
Male *	*	*	*	*	*	*		
Female	-2.0 (1.3)	0.3 (1.9)	1.1 (1.3)	-0.8 (1.8)	1.0 (1.0)	0.5 (1.0)		
Ethnicity							0.05	0.64
White	*	*	*	*	*	*		
Black	1.4 (1.9)	-1.4 (2.8)	1.6 (1.9)	0.7 (2.5)	-3.0 (1.3)	0.7 (2.5)		
Hispanic	-5.1 (2.5)	-4.4 (3.7)	4.4 (2.6)	0.8 (3.3)	0.7 (1.9)	3.7 (2.4)		
Asian	-2.2 (3.0)	0.6 (3.7)	-2.3 (3.0)	-2.8 (3.4)	4.4 (2.5)	2.1 (2.7)		

Characteristic	Estimated difference in proportion reporting						p-value	
	Patient Control		Shared Control		Physician Control		Observed	Synthetic
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic		
Other	0.7 (2.8)	1.2 (3.3)	1.7 (2.9)	-0.8 (3.1)	-2.4 (1.9)	-0.4 (2.8)		
Marital status							0.005	0.42
Married	*	*	*	*	*	*		
Not married	-0.2 (1.4)	1.3 (1.4)	-3.0 (1.4)	-2.1 (1.6)	3.2 (1.1)	0.7 (1.1)		
Education							0.07	0.58
< High school (HS)	-2.1 (1.8)	-0.5 (1.8)	0.6 (1.8)	0.1 (2.2)	1.5 (1.3)	0.6 (1.6)		
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)	-1.5 (1.0)	-0.3 (1.6)		
Income \$							0.08	0.08
< 20,000	2.3 (2.1)	1.2 (2.1)	-1.3 (2.1)	-2.3 (2.1)	1.0 (1.5)	1.0 (1.5)		
20,000 to < 40,000	-2.5 (1.8)	-3.1 (1.8)	2.4 (1.9)	1.7 (1.9)	0.1 (1.3)	1.5 (1.4)		
40,000 to < 60,000	-0.9 (1.9)	-1.2 (1.9)	-0.7 (1.9)	-1.2 (1.9)	1.7 (1.5)	2.4 (1.5)		
60,000	*	*	*	*	*	*		
No. self-reported comorbid conditions							0.79	0.77
0	*	*	*	*	*	*		
1	1.8 (1.4)	1.8 (1.4)	-1.8 (1.4)	-1.9 (1.4)	-0.1 (1.0)	0.0 (1.0)		
2	1.3 (1.9)	1.2 (1.9)	-2.3 (1.9)	-2.4 (1.9)	1.0 (1.4)	1.1 (1.4)		
3	0.8 (2.6)	0.9 (2.6)	-0.2 (2.7)	-0.3 (2.6)	-0.6 (1.9)	-0.6 (1.9)		
Prediagnosis health status							0.02	0.02
Quartile 1	*	*	*	*	*	*		
Quartile 2	0.4 (1.7)	0.4 (1.7)	1.0 (1.7)	0.8 (1.7)	-1.3 (1.2)	-1.1 (1.2)		
Quartile 3	4.9 (1.8)	4.9 (1.8)	-2.8 (1.8)	-3.0 (1.8)	-2.1 (1.3)	-1.9 (1.3)		
Quartile 4	3.7 (1.8)	3.8 (1.8)	-0.4 (1.8)	-0.5 (1.8)	-3.3 (1.2)	-3.3 (1.2)		
CES-D short form							0.26	0.28
5	*	*	*	*	*	*		
6	2.6 (1.7)	2.2 (1.7)	-2.7 (1.7)	-2.4 (1.7)	0.1 (1.3)	0.2 (1.3)		
Study site							< 0.001	< 0.001
Los Angeles county	*	*	*	*	*	*		
Alabama	-1.7 (2.3)	-0.9 (2.3)	5.7 (2.3)	5.6 (2.3)	-4.1 (1.5)	-4.7 (1.5)		
8 counties in North California	2.5 (1.9)	2.0 (1.9)	-2.2 (1.9)	-2.2 (1.9)	-0.3 (1.3)	0.1 (1.3)		
22 counties in eastern North Carolina	-8.4 (2.1)	-7.9 (2.2)	11.2 (2.4)	10.9 (2.4)	-2.8 (1.5)	-3.0 (1.5)		
Iowa	1.1 (2.6)	1.3 (2.6)	4.0 (2.7)	3.4 (2.7)	-5.1 (1.5)	-4.7 (1.5)		
5 HMOs	-1.1 (2.1)	-0.9 (2.1)	2.1 (2.1)	1.8 (2.1)	-1.1 (1.4)	-0.9 (1.4)		

Characteristic	Estimated difference in proportion reporting						p-value	
	Patient Control		Shared Control		Physician Control			
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic
15 Veterans Affairs hospitals	0.2 (2.4)	1.8 (2.5)	1.9 (2.5)	1.1 (2.5)	-2.1 (1.7)	-2.9 (1.7)		

*
reference group

Table 12

Data utility for parameters of the multinomial logit model for patient decision role.

Characteristic	Patient Control				Physician Control			
	Std. Bias	CI. overlap	Cov. (Total)	Err.	Std. Bias	CI. overlap	Cov. (Total)	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 (yrs)								
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	
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	
Ethnicity								

Characteristic	Patient Control				Physician Control			
	Std. Bias	CI. overlap	Cov. (Total)	Err.	Std. Bias	CI. overlap	Cov. (Total)	Err.
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	
No. self-reported comorbid conditions								
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	*	*	*		*	*	*	
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	
Prediagnosis 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	0.24	0.94	0.06		0.03	0.99	0.05	

Characteristic	Patient Control				Physician Control			
	Std. Bias	CI. overlap	Cov. (Total)	Err.	Std. Bias	CI. overlap	Cov. (Total)	Err.
Carolina								
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 Affairs 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	

$$\text{Std. Bias} = \left| \frac{\text{Bias}}{\sqrt{T_p}} \right| \text{ (standardized bias), (9)}$$

CI. overlap: confidence interval overlap, (8)

Cov. Err.: estimated error in coverage of nominal 95% confidence interval, (10) & (11)

* reference group

References

1. Ayanian JZ, Chrischilles EA, Wallace RB, Fletcher RH, et al. Understanding cancer treatment and outcomes: The Cancer Care and Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*. 2004; 22:2992–2996. DOI: 10.1200/JCO.2004.06.020. [PubMed: 15284250]
2. Rubin DB. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*. 1993; 9:461–468.
3. Rubin, DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc; New York: 1987.
4. Reiter JP. Inference for partially synthetic, public use microdata sets. *Survey Methodology*. 2003; 29:181–188.
5. Kennickell, AB. Multiple imputation and disclosure protection: The case of the 1995 Consumer Finances. In: Alvey, W.; Jamerson, B., editors. *Record Linkage Techniques*. National Academy Press; Washington DC: 1997. p. 248-267.
6. Hawala, S. Producing partially synthetic data to avoid disclosure; Proceedings of the Joint Statistical Meetings; American Statistical Association; Alexandria VA. 2008;
7. Abowd, J.; Stinson, M.; Benedetto, G. Longitudinal Employer Household Dynamics Program. US Census Bureau; Washington DC: 2006. Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical Report.
8. Kinney, SK.; Reiter, JP. Making public use synthetic files of the longitudinal business database; Proceedings of the Joint Statistical Meetings; American Statistical Association; Alexandria VA. 2007;
9. Abowd, J.; Woodcock, SD. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In: Domingo-Ferrer, J.; Torra, V., editors. *Privacy in Statistical Databases*. Springer-Verlag; New York: 2004. p. 290-297.
10. Drechsler J, Dundler A, Bender S, Rässler S, Zwick T. A new approach for disclosure control in the IAB Establishment Panel-multiple imputation for a better data access. *Advances in Statistical Analysis*. 2008; 92:439–458. DOI:10.1007/s10182-008-0090-1.

11. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*. 2001; 27:85–95.
12. Reiter JP. Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*. 2005; 131:365–377. DOI: 10.1016/j.jspi.2004.02.003.
13. Spruill, NL. The confidentiality and analytic usefulness of masked business microdata; Proceedings of the section on survey research methods; American Statistical Association: Alexandria VA. 1983; p. 602–607.
14. Paass, G. Disclosure risk and disclosure avoidance for microdata; Presented to the Conference on Access to Public Data; Social Science Research Council: Washington DC. 1985;
15. Dalenius T. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*. 1977; 5:429–444.
16. Duncan GT, Lambert D. Disclosure-limited data dissemination (with comments). *Journal of the American Statistical Association*. 1986; 81:10–28. DOI: 10.1080/01621459.1986.10478229.
17. Duncan GT, Lambert D. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*. 1989; 7:207–217. DOI: 10.1080/07350015.1989.10509729.
18. Fienberg SE, Makov UE, Sanil AP. A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*. 1997; 13:75–89.
19. Reiter JP. Estimating identification risks in microdata. *Journal of the American Statistical Association*. 2005; 100:1103–1113. DOI: 10.1198/016214505000000619.
20. Reiter JP, Mitra R. Estimating risks of identification and disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*. 2009; 1:99–110.
21. Woo MJ, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*. 2009; 1:111–124.
22. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil A. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*. 2006; 60:224–232. DOI: 10.1198/000313006X124640.
23. Cochran, WG. Sampling techniques. John Wiley and Sons; New York: 1977.
24. Van Buuren, S.; Oudshoorn, CGM. Multivariate imputation by chained equations: MICE v1.0 user's manual. TNO; Leiden: 2000.
25. Su YS, Gelman A, Hill H, Yajima M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. 2011; 45(2)
26. The R project for statistical computing. <http://www.r-project.org>
27. Reiter JP. Releasing multiply imputed, synthetic public use-microdata: an illustration and empirical study. *Journal of the Royal Statistical Society - Series A*. 2005; 168:185–205. DOI: 10.1111/j.1467-985X.2004.00343.x.
28. Drechsler J, Reiter JP. Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. *Journal of Official Statistics*. 2008; 25:589–603.
29. Tibshirani RJ. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
30. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360. DOI: 10.1198/016214501753382273.
31. Yucel RM, He Y, Zaslavsky AM. Gaussian-based routines to impute categorical variables in health surveys. *Statistics in Medicine*. 2011; 30:3447–3460. DOI: 10.1002/sim.4355. [PubMed: 21976366]
32. Huskamp HA, Keating NL, Malin JM, Zaslavsky AM, Weeks JC, Earle CC, Teno JM, Virnig BA, Kahn KL, He Y, Ayanian JZ. Discussions with physicians about hospice among patients with metastatic lung cancer. *Arch Intern Med*. 2009; 169:954–962. DOI: 10.1001/archinternmed.2009.127. [PubMed: 19468089]

33. Keating NL, Landrum MB, Arora NK, Malin JL, Ganz PA, van Ryn M, Weeks JC. Cancer patients' roles in treatment decisions: do characteristics of the decision influence roles? *Journal of Clinical Oncology*. 2010; 28:4364–4370. DOI: 10/1200/JCO.2009.26.8870. [PubMed: 20713872]
34. He Y, Zaslavsky AM, Landrum M, Harrington DP, Catalano P. Multiple imputation in a large-scale complex survey: a practical guide. *Statistical Methods in Medical Research*. 2009; 19:663–670. DOI: 10.1177/0962280208101273.
35. Kinney SK, Reiter JP. Tests of multivariate hypotheses when using multiple imputation for missing data and partial synthesis. *Journal of Official Statistics*. 2010; 26:301–315.
36. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 1994; 9:538–558.
37. White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50:1–25.
38. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. *Classification and Regression Trees*. Wadsworth Inc; Belmont California: 1984.
39. Breiman L. Random forests. *Machine Learning*. 2001; 45:5–32.
40. Reiter JP. Multiple imputation for disclosure limitation: future research challenges. *Journal of Privacy and Confidentiality*. 2009; 1:223–233.
41. Reiter JP. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*. 2005; 21:441–462.
42. Drechsler J, Reiter JP. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*. 2010; 105:1347–1357. DOI: 10.1198/jasa.2010.ap09480.
43. Caiola G, Reiter JP. Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*. 2010; 3:27–42.

Table 1

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

Characteristic	Patients (%)		Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic	Observed	Synthetic
Sex					0.89	0.43
Male	61.3	57.1	53.3 (1.6)	50.9 (2.8)		
Female	38.7	42.9	53.0 (2.1)	55.5 (3.0)		
Race/ethnicity					< 0.001	0.62
White	73.7	72.0	55.2 (1.5)	54.4 (1.5)		
Black	10.7	11.4	42.6 (3.9)	50.0 (4.1)		
Hispanic	5.9	5.7	40.4 (5.2)	45.8 (5.3)		
Asian	5.1	5.3	49.4 (5.9)	51.6 (6.0)		
Other	4.7	4.8	64.5 (5.7)	54.0 (6.9)		
Married/live with partner					0.006	0.06
Yes	61.0	60.2	50.4 (1.6)	50.4 (1.7)		
No	39.0	39.8	57.6 (2.0)	57.4 (2.0)		
Age (yrs)					0.002	0.007
21-54	12.5	12.5	45.5 (3.6)	44.5 (4.1)		
55-59	12.4	12.5	52.1 (3.6)	49.1 (4.0)		
60-64	13.2	13.1	51.0 (3.5)	49.3 (4.0)		
65-69	16.0	16.4	50.8 (3.2)	50.3 (3.6)		
70-74	17.8	16.7	50.4 (3.0)	55.6 (3.5)		
75-79	13.8	14.3	57.1 (3.4)	57.3 (3.5)		
80+	14.4	14.0	65.1 (3.2)	64.7 (3.3)		
Education					0.46	0.22
< High school	22.7	21.6	54.8 (2.7)	55.9 (3.2)		
High school/some college	60.6	62.2	53.5 (1.6)	53.8 (1.8)		
College degree	16.7	16.3	49.8 (3.1)	47.0 (3.7)		

Table 2

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

Characteristic	Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic
Sex			0.72	0.93
Male	54.5 (9.4)	54.4 (9.9)		
Female	53.2 (9.9)	54.7 (10.4)		
Race/ethnicity			< 0.001	0.22
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)		
Asian	60.0 (11.4)	59.2 (11.7)		
Other	78.2 (7.7)	65.3 (11.6)		
Married/live with partner			0.031	0.047
Yes	54.5 (9.4)	54.4 (9.9)		
No	62.6 (9.6)	62.1 (10.1)		
Age (yrs)			0.63	0.94
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)		
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			0.045	0.061
Yes	54.5 (9.4)	54.4 (9.9)		
No	35.7 (12.3)	37.7 (12.7)		
Education			0.87	0.48
< High school	56.3 (9.9)	60.3 (10.2)		
High school/some college	54.5 (9.4)	54.4 (9.9)		
College degree	56.3 (10.1)	53.1 (10.5)		
Income (\$)			0.19	0.42
< 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)		
60 000	54.5 (9.4)	54.4 (9.9)		
Insurance			0.018	0.004
Medicare	54.5 (9.4)	54.4 (9.9)		
Medicaid	48.3 (11.1)	49.5 (11.6)		
Private	57.3 (7.9)	50.3 (8.6)		
Other	77.1 (8.1)	79.9 (7.8)		
Treated in VA facility			0.43	0.74

Characteristic	Discussed Hospice %		p-value	
	Observed	Synthetic	Observed	Synthetic
Yes	46.6 (14.4)	51.1 (14.7)		
No	54.5 (9.4)	54.4 (9.6)		
HMO member			0.25	0.42
Yes	60.4 (9.3)	60.0 (9.9)		
No	54.5 (9.4)	54.4 (9.9)		
Region			0.12	0.055
South	60.7 (8.9)	59.3 (9.6)		
West	54.5 (9.4)	54.4 (9.9)		
Other	47.9 (10.0)	44.7 (10.5)		
Days from diagnosis to interview			0.055	0.079
Quartile 1	54.5 (9.4)	54.4 (9.9)		
Quartile 2	67.1 (8.2)	66.1 (8.9)		
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			< 0.001	< 0.001
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)		
180-239 d	5.9 (2.2)	6.1 (2.4)		
240 d	64.2 (8.9)	65.0 (9.2)		
Received chemotherapy before interview			0.006	0.002
Yes	54.5 (9.4)	54.4 (9.9)		
No	59.1 (9.3)	58.4 (9.6)		
Comorbidity			0.23	0.15
None	54.5 (9.4)	54.4 (9.9)		
Mild	57.1 (9.6)	57.0 (9.8)		
Moderate	64.7 (9.1)	65.4 (9.2)		
Severe	70.3 (9.4)	68.8 (10.2)		
Alive at survey but surrogate completed interview			0.004	0.008
Yes	61.7 (9.4)	59.2 (11.2)		
No	54.5 (9.4)	54.4 (9.9)		

Table 3

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

Characteristic	Coef. Est		Std. Bias	CI. overlap	Coverage error
	Observed	Synthetic			Total
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
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
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

Characteristic	Coef. Est		Std. Bias	CI. overlap	Coverage error
	Observed	Synthetic			Total
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
240 d	-2.94 (0.191)	-2.94 (0.196)	0.025	0.724	0.050
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

$$\text{Std. Bias} = \left| \frac{\text{Bias}}{\sqrt{T_p}} \right| \text{ (standardized bias), (9)}$$

CI. overlap: confidence interval overlap, (8)

Coverage Error: estimated error in coverage of nominal 95% confidence interval, (10), (11)

Table 4

Categories for role in 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

Table 5
Disclosure risk - CanCORS lung cancer partially synthetic data

Quasi-identifier set	MXM	EMR		TMR	
		Observed	Synthetic	Observed	Synthetic
1	3856	70	38	0	0
2	1158	1770	217	989	88
3	1158	4495	890	4000	717

MXM: maximum number of matches

EMR: expected match risk

TMR: presumed true match risk

Table 6
Disclosure risk - CanCORS colorectal cancer partially synthetic data

Quasi-identifier set	MXM	EMR		TMR	
		Observed	Synthetic	Observed	Synthetic
1	3007	70	32	0	0
2	100	1824	22	1070	3
3	100	4495	90	4105	83

MXM: maximum number of matches

EMR: expected match risk

TMR: presumed true match risk