HHS Public Access

Author manuscript

Health Serv Outcomes Res Methodol. Author manuscript; available in PMC 2022 February 03.

Published in final edited form as:

Health Serv Outcomes Res Methodol. 2021 February 03; 21: 389–406. doi:10.1007/s10742-021-00241-z.

Using Synthetic Data to Replace Linkage Derived Elements: A Case Study

Dean M. Resnick, **Christine S. Cox**

NORC at the University of Chicago, Bethesda, Maryland

Lisa B. Mirel

Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland

Abstract

While record linkage can expand analyses performable from survey microdata, it also incurs greater risk of privacy-encroaching disclosure. One way to mitigate this risk is to replace some of the information added through linkage with synthetic data elements. This paper describes a case study using the National Hospital Care Survey (NHCS), which collects patient records under a pledge of protecting patient privacy from a sample of U.S. hospitals for statistical analysis purposes. The NHCS data were linked to the National Death Index (NDI) to enhance the survey with mortality information. The added information from NDI linkage enables survival analyses related to hospitalization, but as the death information includes dates of death and detailed causes of death, having it joined with the patient records increases the risk of patient re-identification (albeit only for deceased persons). For this reason, an approach was tested to develop synthetic data that uses models from survival analysis to replace vital status and actual dates-of-death with synthetic values and uses classification tree analysis to replace actual causes of death with synthesized causes of death. The degree to which analyses performed on the synthetic data replicate results from analysis on the actual data is measured by comparing survival analysis parameter estimates from both data files. Because synthetic data only have value to the degree that they can be used to produce statistical estimates that are like those based on the actual data, this evaluation is an essential first step in assessing the potential utility of synthetic mortality data.

Keywords

National Center for Health Statistics; National Death Index; National Hospital Care Survey Linked Mortality Data; Survival Analysis; Synthetic Data

Resnick-Dean@norc.org .

Conflicts of interest/Competing interests (include appropriate disclosures)
The authors have no conflicts of interest to report.

Availability of data and material (data transparency)
Researchers who wish to obtain access to the linked 2016 NHCS to 2016/2017 NDI file must submit and have an approved research proposal to the NCHS Research Data Center (RDC): https://www.cdc.gov/rdc/index.htm.

Code availability (software application or custom code)
Code for this analysis is available upon request.

1. Introduction

The National Center for Health Statistics (NCHS) has established a Data Linkage Program designed to maximize the scientific value of the Center's health surveys by linking its health survey data with health-related administrative data resources. These linked files create new longitudinal data resources that expand the analytic potential beyond the individual data sources and create new research opportunities to understand the factors that influence disability, health care utilization, morbidity, and mortality among different U.S. subpopulations. NCHS has previously linked several of its large population health surveys, including the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Surveys (NHANES) to mortality data from the National Death Index (NDI) (NCHS NDI). In an effort to maximize access to linked mortality data, NCHS has created public-use linked mortality files for NHIS and NHANES data that contain partially, not fully, synthetic mortality information, including date and cause of death data (https://www.cdc.gov/nchs/data-linkage/mortality-public.htm).

NCHS recently completed a mortality linkage with data on patients that received inpatient (IP) or emergency department (ED) services at sampled hospitals participating in the National Hospital Care Survey (NHCS). The NHCS collects complete sets of encounter records for participating hospitals, which also contain patient identification information, such as name, address, birth date, and Social Security Number (SSN), to facilitate linkage to other sources of health-related data. Patient data collected in the 2016 NHCS were linked to the NDI to obtain information on mortality status and cause of death for deaths occurring after the hospitalization through the subsequent calendar year (2017) (NCHS 2019).

While the linked patient records and NDI records are putatively for patients who have died, NCHS data confidentiality standards require patient information (including that for persons presumed to be deceased) to be unidentifiable for data users. However, if data were made available that include a patient's date-of-birth, date-of-death, and other demographics, such as sex and state-of-residence, re-identification of the patient may be enabled, even when withholding direct identifiers such as name and SSN. Additionally, the linked NDI records add detailed cause of death information to patient records which further increases the risk of re-identification. Due to requirements to protect the confidentiality of the NHCS data, restricted-use versions of the Linked Mortality Files (LMFs) were made available only through the NCHS and Federal Statistical Research Data Centers (RDCs).

NCHS is exploring making more detailed mortality information, including age of death and cause of death, more widely available to researchers. One way of achieving this is by creating a synthetic linked NHCS-NDI file that maintains certain associations between variables and making the file publicly available via the NCHS website. This paper will serve as a case study to explore the methodology, using the linked NHCS-NDI data as an example, of how synthetic data could be generated and evaluated to determine its concordance with the actual data. This paper does not address whether the generated synthetic data file adequately reduces the risk of re-identification once the file is produced, but rather takes an essential first step in evaluating whether the proposed synthetic data can produce statistically valid results.

The data elements that will be synthesized include vital status (alive or dead), date of death, the underlying or primary cause of death from the top nine leading causes of death and all residual deaths, and indicators of the presence of multiple or contributing causes of death for diabetes, hypertension, or both (NHCS NDI). The presence of diabetes and hypertension in the multiple cause-of-death codes were included, as these conditions are frequently reported as contributing causes of death.

2. Methods

2.1 Data sources

2.1.1 NHCS description—The NHCS is an establishment survey that collects IP, ED, and outpatient department (OPD) encounter-level data from sampled hospitals. NHCS is one of the National Healthcare Surveys, a family of surveys covering a wide spectrum of healthcare delivery settings from ambulatory and OPD to hospital and long-term care providers. The goal of NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance-involved ED visits (NCHS NHCS).

From participating hospitals, NHCS collects data on all IP and ambulatory (ED and OPD) care visits occurring during the calendar year. The 2016 NCHS data collection procedures provided hospitals with the option to submit data in the form of electronic health records (EHR) or as UB-04 claims records. NHCS collects patient personally identifiable information (PII) such as name, date of birth, and SSN, which allows for the linkage of each patient's health care encounters within a surveyed hospital as well as to other external data sources, such as the NDI. The analysis described in this paper includes only IP and ED visits - other, non-ED OPD visits have been excluded because OPD visits were not included in the 2016 NHCS- 2016/2017 NDI linkage. NHCS is not currently nationally representative due to low response rates, 158/581=27%[1]. Still, linking NHCS with the NDI does allow for new analyses, such as studying mortality post hospital discharge, along with specific causes of death. (NCHS 2019)

2.1.2 NDI description—The NDI is a centralized database of United States death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. The NDI became operational in 1981 and includes death record information for all persons officially known to have died in the U.S. or a U.S. territory from 1979 onward. The records, which are compiled annually, include detailed information on the underlying and multiple or contributing causes of death (NHCS NDI).

2.1.3 Linked data—The linkage of the 2016 NHCS to the 2016/2017 NDI has been described elsewhere (NCHS 2019). Briefly, patients with sufficient PII[2] were linked in

[1]Responding hospitals were those providing records for at least 50 encounters covering at least six months of the year.
[2]Sufficient PII is defined as having two of the following three items: valid date of birth (month, day, and year), name (first, middle, and last), and/or a valid format 9-digit SSN. See (NCHS 2019, p. 5).

two steps, using deterministic and probabilistic linkage techniques. About five percent of the 2016 NHCS linkage eligible patients linked to the 2016/2017 NDI with the largest percentage of links in the 65 or older age category (NCHS 2019). The total number of 2016 NHCS patients identified as deceased through NDI linkage was 212,155.

2.2 Generation of synthetic data

The generation of the synthetic data relied on two main steps and then an assessment. These steps are outlined below.

2.2.1 Step 1. Modeling occurrence and date of death—

The first stage of the synthetic data generation relates to assigning occurrence and date of death to certain patient records. The patients who are assigned a status of assumed deceased during the follow-up period do not always represent the same patients who were linked to NDI death records. Similarly, patients who are assigned a vital status of dead in the synthetic data may in fact be presumed alive based on a non-match status to an NDI record. However, it is important that the synthesized death status reflects the propensity of death as it depends on sex, age, health conditions, and other factors related to mortality. For this reason, a Cox proportional hazard model (Cox 1972) (estimated using SAS's PHREG procedure) (SAS PHREG) was used to generate death status.

In this context it was decided to use the known diagnoses from the latest survey-collected patient encounter record, whether this is an IP or ED visit. The Charlson Comorbidity Index scoring system was designed as a weighted composite index for predicting mortality risk within 1 year of hospitalization for patients with specific co-morbidities (Charlson 1987). The Charlson index is meant to reflect near-term mortality experience based on the presence of one or more of 17 major diagnostic condition categories. Each of these categories has a weight associated with it and the index value is equal to the sum of these weights (Table 1). Code developed by the University of Calgary (Sundararajan 2004) was used both to create the 17 indicator variables, apply the weights, and compute the composite index value.

Using these recoded condition categories rather than specific diagnoses seemed more suitable for regression analyses, which would have otherwise required estimating a regression parameter for thousands of diagnosis code levels. Additionally, the risk associated with these conditions was not expected to be necessarily linearly additive. That is, a computed risk level

$$\text{Charlson Index} = \sum C_i \cdot W_i \qquad \text{(Eq. 1)}$$

C_i – 0/1 indicator for presence of condition i

W_i –Weight for condition C_i

associated with having both congestive heart failure and pulmonary disease is likely more than the sum of having these conditions individually. For this reason, both the 17 condition indicators (Variables # 21–37, as shown in the Appendix) and a Charlson index value (Variables # 38–43), which was top coded at a score of six so as to address collinearity

of these two variable sets (i.e., so the index value is not a linear function of the present conditions) were included in the synthetic data generation model to represent the level of comorbidities:

$$\text{Charlson Index Value} = \min\{\text{Charlson Index}, 6\}$$

Incorporating these Charlson values, a multivariable survival analysis regression procedure, SAS's PHREG (proportional hazard regression), was used to estimate the risk of dying based on the known characteristics of the patient, hospital utilization, and hospital characteristics (SAS PHREG). Patient characteristics included age at discharge, sex, Census division of residence[3], as noted on the claim or EHR, and imputed race and ethnicity. Modeling for race and ethnicity was conducted based on Census distribution of last names (reported on claim or EHR) combined with tract-level race/ethnicity distributions from 2010 Decennial Census in a probabilistic model (Fiscella 2006). Categories included Hispanic, non-Hispanic white, non-Hispanic black or African American, non-Hispanic American Indian or Alaskan Native, and non-Hispanic Asian or Pacific Islander. Hospital utilization was defined as the patient's length of inpatient stay in the calendar year (top coded at 50) and the number of ED visits in the calendar year (top coded to five) for the patient. Hospital characteristics were defined by, ownership status (for profit, government, and non-profit), type of hospital (general acute, children's, psychiatric, and Long-Term Acute Care and Rehabilitation), urban-rural classification (large central metropolitan, large fringe metropolitan, medium metropolitan, small metropolitan, micropolitan, and non-core) and categories of bed size (1 – 25, 26 – 100, 101 – 500, and more than 500).

For each patient, the regression procedure estimates a linear predictor of death risk. Additionally, the procedure has been set to compute baseline survival rates for each day subsequent to the latest known discharge. Thus, for any patient i, the probability of surviving to day t is equal to:

$$\lambda(t|X_i) = \lambda_0(t)\exp(X_i \cdot \beta) \tag{2}$$

$X_i \cdot \beta = \sum_{k=1}^{n_k} X_{i,k} \cdot \beta_k$, k represents index for each of 69 covariates shown in Appendix. where $\lambda_0(t)$ is the baseline survival rate for time t, β is the vector of regression parameters, and X_i is the set of predictors for patient i. In this implementation of the Cox proportional hazard estimation model, the Breslow estimator is used to estimate the baseline hazard function. (Breslow 1972) (SAS PHREG).

To impute the death status we used the survival probability function. To start, a date of death or a date of censoring for each patient was imputed by drawing a pseudo-random value from a standard uniform distribution ($u \sim uniform[0,1]$) and then comparing that value to the patient's survival probabilities $\hat{S}\left(t|x\hat{\beta}\right)$ The first date where the survival probability was

[3]Division is coded based on patient home state based on U.S. Census Bureau Schema. See https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

less than or equal to the drawn random value, was assigned as the imputed date of death: $t^* \ni \hat{S}\left(t|\boldsymbol{x}\hat{\boldsymbol{\beta}}\right) \leq u$ (Lipkovich 2016).

Since duplicating the level of death reporting to within the follow-up period was sought, if the random value was less that the survival probability for the last day of the follow-up period, then the patient was modeled as having survived through the follow-up period and assigned a date of censoring. Those assigned a date of death were imputed as died, and those assigned a date of censoring were imputed as alive. Note then that a completely different set of synthesized data can be generated by using a different sequence, initiated by using a different seed to the pseudo-random number generator, of random values to be compared to survival probabilities.

Exhibit 1 demonstrates the imputation of date of death or date of censoring. It shows a plot of one patient's model-estimated survival probabilities (computed from Eq. 2, from date of discharge to end of follow-up) compared to a uniform random variable value (for this example, we use RV=0.6732); note that in an actual synthetization run, a distinct random variable value is generated independently for each patient.

To estimate the date of death or date of censoring, one would move across the plot at a level equal to RV, until reaching the survival probability plot (actually, the first day with survival probability just under the RV). There is a specific date associated with this RV (for this example, January 23, 2017) and this is the value used in the imputation. If the RV selected is less than the survival probability on December 31, 2017 (the end of the follow-up period) the patient is not assigned a death status (they are imputed to have survived follow-up).

2.2.2 Step 2. Modeling the contributing and underlying causes of death—

Synthetization of cause of death is contingent upon synthetization of occurrence of death (i.e. a synthetic cause of death code is generated only for patients assigned a synthetic mortality status of assumed deceased). There are two sets of information about cause of death to be synthesized:

1. The presence of diabetes or hypertension, as contributing causes of death (similar to the two multiple causes of death available on the NHANES and NHIS partially synthetic public use linked mortality files (https://www.cdc.gov/nchs/data-linkage/mortality-public.htm))
2. The underlying cause of death was limited to include only the top nine leading underlying causes of death (similar to the NHIS and NHANES public use linked mortality files) based on the National Vital Statistics System 2017 annual report (https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_06-508.pdf). The top nine causes of death included diseases of the heart, malignant neoplasms (cancer), accidents (unintentional injuries), chronic lower respiratory diseases, cerebrovascular diseases, Alzheimer disease, diabetes mellitus, influenza and pneumonia, and deaths due to nephritis, nephrotic syndrome and nephrosis. All other underlying causes of death were grouped together and placed in a residual category. The modeling for underlying cause of death was dependent on whether

synthesized diabetes, hypertension, both, or neither as contributing causes of the death were present.

The model for underlying cause of death is contingent on the contributing cause of death, thus the model for underlying cause of death is nested within that for contributing cause of death, and both of these are nested within the model for occurrence and date of death.

Each of these sets is synthesized using a classification tree model (Breiman 1984) and the factors tested for the model are generally similar to those used in the survival analysis. This approach is expected to result in a joint distribution that will align with the patterns that are present in the underlying data.

To build the classification trees, SAS's HPSPLIT was used (SAS HPSLIT). The data used to build the classification tree are the actual set of NHCS records linked to NDI records (i.e., the records for patients shown to have died based on record linkage procedures).

To assign the multiple or contributing causes of death, there were four possible statuses to be selected from:

- Neither diabetes nor hypertension is a contributing cause of death
- Diabetes but not hypertension is a contributing cause of death
- Hypertension but not diabetes is a contributing cause of death
- Both diabetes and hypertension are contributing causes of death

The variables used for classification are:

- Days of inpatient care
- Number of emergency department visits
- Age at time of discharge
- Sex
- Imputed race and ethnicity
- Charlson Comorbidity Index groups (among 17 possible)
- Charlson Index Value Summary (range 0–6)
- Census division of residence
- Length of interval (days) from discharge to death

and they were defined identically to those used in the original survival model.

The developed classification tree schema places each patient in a leaf based on the above characteristics using logic that is developed to minimize the entropy of resulting classifications. For each leaf, there is an associated probability of each of the four contributing cause statuses, which sum to one. By partitioning the interval from zero to one into segments having lengths equal to their corresponding probabilities, a pseudo random value drawn from uniform distribution specifies the assigned contributing cause status.

Once the contributing cause status has been synthesized, it is used along with the other predictive variables listed above to model the underlying cause of death from among these ten exclusive categories:

• Malignant neoplasms (cancer)	• Diabetes mellitus
• Diseases of the heart	• Alzheimer disease
• Accidents (unintentional injuries)	• Influenza and pneumonia
• Chronic lower respiratory diseases	• Nephritis, nephrotic syndrome, and nephrosis
• Cerebrovascular diseases	• All other

The classification variables for this sub-model are identical to those used in the contributing causes of death sub-model except for the addition of the synthesized value for contributing cause of death. The assignment of the underlying cause of death is performed in a parallel manner to the major causes. It should be noted that because each of the assignments is dependent on a drawn random variable a different sequence of drawn random variables will produce different synthesized causes of death for the same patients.

2.2.3 Step 3. Assessing the estimates from the actual and synthesized data

—Since the synthesized data would probably most frequently be used is in survival analysis, the statistical utility of the synthetic data is evaluated by comparing the similarity of survival parameter estimates between the synthetic data and the actual data. Survival analysis, using SAS's PHREG, was conducted (SAS PHREG). The variables in this model (there are a total of 69 and these are listed in the Appendix) were defined identically to those used in the original survival model except that age was binned in five-year increments.

For the synthetic data, the at-risk period for each patient is the number of days from the last hospital discharge (either from IP or ED setting) until synthesized death or the end of the follow-up period. To assess the results of the synthetic data versus the actual, the methods below were used:

- Histograms of the ratios of the parameter estimates of the two approaches with a peak occurring at a ratio of 1, indicating similarity.
- Scatter plots of actual versus synthetic parameter estimates: with conformity to the y = x (45° line), indicating similarity.
- Regression analysis of synthetic parameter estimates to actual parameter estimates: with estimated R-Square and β nearness to 1, indicating similarity.
- Cross tabulations of actual versus synthetic parameter statistical significance evaluated at alpha=0.01: a higher percent agreement indicates greater similarity. Cohen's Kappa statistic was used to measure agreement of the statistically significant parameters from the synthetic and actual data. Kappa statistics were generated for those at alpha=0.01. The standard range of the Kappa statistic is 0 for no agreement and 1 for complete agreement, albeit values from −1 to 0 are possible and would indicate negative correlation. Landis and Koch (Landis 1977) suggest the following interpretation for the Kappa statistic: < 0.00: Poor; 0.00–0.20: Slight; 0.21–0.40: Fair; 0.41–0.60: Moderate; 0.61–0.80: Substantial; 0.81–1.00: Almost Perfect. The Kappa statistic was used as a way to account for

agreement by chance. Note, for example, that if 80% of both actual and synthetic parameter estimates are significant, but this status is randomly assigned, then just by chance, 68% of statuses would agree. Thus, high levels of percent agreement may be less confirmatory of general concordance than might be expected, and Kappa is a way to get a better assessment than the raw agreement rate.

For all of these analyses, the results are presented as unweighted estimates. At this time the NHCS cannot be used to make nationally representative estimates due to the low response rate.

3. Results

An initial assessment looked at the distributions of all-cause mortality and cause specific mortality of the actual data and the synthetic data. The results from this assessment were similar (Table 2). Exhibit 2 shows the distribution of the ratio of the survival parameter estimates for the synthetic data to the survival parameter estimates for the actual data. There is a high peak right around the value 1 (69.1% are ±5% of this)), and for most of the estimates the ratio is between 0.8 and 1.2 (about + or − 20%, where a value of 1 indicates that the parameter estimated from the actual survey data exactly equals the value of the parameter estimated from the synthetic data).

Exhibit 3 shows the plot of these estimates (Synthetic vs. Actual). The plotted points fall very close to the 45° line suggesting similarity between the two sets of values.

Similarly, proportional hazard regressions (both using actual death data and synthesized death data) for each of the nine specific causes of death were conducted. For each of the nine causes of death a survival analysis was conducted that considered death by cause as the event and the at-risk period as time from discharge to death or to the end of the follow-up period, whichever came first. Patients dying of other causes were included in the model with the at-risk period running from date of discharge to the date of death (synthetic or actual), but not considered as having a death outcome. They were censored at their time of death. Exhibits 4 and 5 shows plots for cancer and heart disease.

When comparing the resulting parameter estimates displayed in Exhibit 4 and 5, it is seen they follow less on the 45° line than in the analysis for all causes of death but still generally approach it.

However, there is a strong relationship between the actual and synthetically derived parameter estimates as can be demonstrated by conducting linear regression between them. Here the response variable is the parameter estimates for underlying cause of death generated from the proportional hazard model using the actual data and the predictor is the parameter estimates for underlying cause of death generated for the same predictor from the proportional hazard model using the synthetic data. In the appendix table, this would be the regression of ***Estimate (Syn.)*** *(3rd column)* on ***Estimate*** *(2nd column)*. Thus, as R-square approaches 1, the estimated parameter value from actual data correlates with the estimated parameter value from the synthetic data (i.e., the line connecting them is straight). In addition, a β of 1 indicates that the degree of change in the actual parameter estimates is

of the same scale the degree of change in the synthetic data parameter estimates (i.e., the fit line is on 45°). Table 3 summarizes the results of these regressions.

The R-square values presented in Table 3 are generally close to 1, all are above 0.90 except for Alzheimer disease, diabetes melliltus, and influenza and pneumonia. Also, the slope of regression line is generally close to 1 except for Alzheimer disease, diabetes melliltus, and influenza and pneumonia.

For an analyst who is using synthetic data to evaluate relationships between various factors and survival rates, it may be of interest whether specific relationships are shown to be significant rather than the precise size of estimates. Ideally, factors which are statistically significant, using actual linked data, would remain statistically significant when using the synthesized data, and any factor not statistically significant using actual linked data would remain so when using the synthesized data. To evaluate this, a cross-classification, actual to synthetic, of statistical significance status for survival analysis parameter estimates was assessed. For example, if of the 69 estimated parameter estimates (this is the sum of the levels in each categorical variable less one for the reference level, and they are listed in the Appendix) in the chronic lower respiratory disease cause-of-death survival analysis, 58 agreed (actual vs. synthetic) on the statistical significance status at the alpha=0.01 level, the percent agreement would be 58/69=84.1%, with a Kappa of 0.67, indicating moderate agreement.

Table 4 shows the concordance of statistically significant survival parameter estimates between the actual and synthetic data. The percent agreement for all-cause mortality is 97.1% and for cause specific mortality the percent agreement is generally between 70 and 85%. The Kappa statistic, assessing the agreement of the number of statistically significant parameters, for all-cause mortality is 0.84, suggest almost perfect agreement, while for cause specific mortality the Kappa statistic ranges from 0.27 to 0.67, suggesting slight to moderate agreement. Thus for a researcher using the synthetic data to determine which variables are statistically significant predictors of survival for specific causes of death with the actual data they would usually, but not always, get a correct indication of statistical significance. Still, to confirm the results using the synthetic data researchers would need to gain access to the actual restricted-use linked NHCS-NDI data files.

4. Conclusion

A new methodology was employed to create synthetic data for the NHCS linked mortality data. The occurrence of death was synthesized using a proportional hazard model that incorporated demographics and health status information collected from patient encounter records. Classification trees to assign underlying and selected multiple or contributing causes of death for modeled deaths were used.

The analysis shows that the synthetic data yield similar parameter estimates for occurrence and time until death such that this approach to creating publicly available synthetic data could be a reliable substitute for access to the restricted-use linked NHCS-NDI data. With regard to the causes of death, the comparison of regression parameters shows fairly high

R-square estimates for most causes of death, with some level of non-alignment (e.g., Alzheimer's disease, diabetes mellitus, and influenza and pneumonia). This approach to creating synthetic data would allow data users to form preliminary analyses of condition-specific death rates; however, subsequent access to the actual data may in some cases yield different conclusions about the statistical significance of factors relating to that survival experience. By limiting underlying cause categories to those which present the most similar results to the actual data, it would be possible to minimize the instances where the estimates made from synthetic data differ significantly from the estimates made from the actual data. Should NCHS decide to proceed with the production of NHCS-NDI synthetic linked data set based on the data generation models presented in this paper, the next steps would entail conducting appropriate disclosure protections analyses to determine whether the synthetic data generated by these processes provide acceptable levels of privacy protection. In addition, it may be worth considering creating multiple replicates for the synthetic data to assess the uncertainty of the statistical models.

The most substantial limitation to the analysis presented in this paper, of which we are aware, is that the survival models used to evaluate the synthetic data are very similar in structure to the models used to develop the synthetic data, particularly with regard to the predictors used in them. Thus, this analysis does not demonstrate that unmodelled variables will be synthesized in a way that generate results that would be obtained with actual data and users of the synthetic data must be made aware of this limitation. Another limitation of this study is lack of an explanation of why certain causes of death can be modeled with synthetic data more similarly to actual data than others. Additionally, it would be beneficial to data users if they were provided a means to establish a confidence interval for actual parameter estimates based on results obtained from synthetic data.

Still, the methods presented in this paper suggest strategies that could be used effectively when linked or even non-linked data needs to be protected from disclosure. In particular it presents a viable strategy for incorporating survival models into a synthetic database generation model. It also shows how synthetic data can be evaluated using parameter estimates from explanatory data.

Funding (information that explains whether and by whom the research was supported)

This work was supported in part with funding from the Department of Health and Human Services' Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF).

Appendix

NHCS 2016 Survival Estimates Comparison of Actual to Synthetic Data Comparison for All Deaths

#	Variable	Estimate	Estimate (Synth.)	StdErr	StdErr (Synth.)	ProbChiSq	ProbchiSq (Synth.)
	Age Group (rounded to nearest five-year, reference category: 70 years old)						

#	Variable	Estimate	Estimate (Synth.)	StdErr	StdErr (Synth.)	ProbChiSq	ProbchiSq (Synth.)
1	0	−3.37	−3.34	0.036	0.035	<.0001	<.0001
2	5	−4.12	−4.07	0.059	0.057	<.0001	<.0001
3	10	−3.97	−3.95	0.063	0.062	<.0001	<.0001
4	15	−3.47	−3.41	0.048	0.047	<.0001	<.0001
5	20	−2.75	−2.78	0.032	0.032	<.0001	<.0001
6	25	−2.34	−2.33	0.025	0.025	<.0001	<.0001
7	30	−2.13	−2.15	0.023	0.023	<.0001	<.0001
8	35	−1.79	−1.80	0.021	0.021	<.0001	<.0001
9	40	−1.48	−1.50	0.019	0.019	<.0001	<.0001
10	45	−1.19	−1.16	0.017	0.016	<.0001	<.0001
11	50	−0.90	−0.87	0.014	0.014	<.0001	<.0001
12	55	−0.62	−0.62	0.012	0.012	<.0001	<.0001
13	60	−0.38	−0.39	0.011	0.011	<.0001	<.0001
14	65	−0.21	−0.22	0.011	0.011	<.0001	<.0001
15	75	0.21	0.21	0.010	0.010	<.0001	<.0001
16	80	0.48	0.47	0.010	0.010	<.0001	<.0001
17	85	0.79	0.78	0.010	0.010	<.0001	<.0001
18	90	1.12	1.11	0.011	0.011	<.0001	<.0001
19	95	1.52	1.50	0.012	0.013	<.0001	<.0001
20	Age missing	0.26	0.15	0.041	0.043	<.0001	0.0004
	Conditions not present (reference category: condition not present)						
21	Myocardial Infarction	0.03	0.03	0.016	0.016	0.0804	0.0309
22	Diabetes without complications	−0.31	−0.33	0.023	0.023	<.0001	<.0001
23	Diabetes with complications	−0.03	−0.06	0.036	0.037	0.3462	0.1008
24	Paraplegia and Hemiplegia	0.33	0.31	0.021	0.022	<.0001	<.0001
25	Renal Disease	0.06	0.06	0.014	0.015	<.0001	<.0001
26	Cancer	0.56	0.54	0.015	0.016	<.0001	<.0001
27	Moderate or Severe Liver Disease	0.59	0.58	0.076	0.076	<.0001	<.0001
28	Metastatic Carcinoma	1.30	1.30	0.029	0.031	<.0001	<.0001
29	AIDS/HIV	−0.08	−0.12	0.045	0.047	0.0598	0.0081
30	Congestive Heart Failure	0.39	0.38	0.009	0.010	<.0001	<.0001
31	Peripheral Vascular Disease	−0.15	−0.15	0.016	0.016	<.0001	<.0001
32	Cerebrovascular Disease	−0.05	−0.06	0.012	0.012	0.0001	<.0001
33	Dementia	0.50	0.48	0.010	0.011	<.0001	<.0001
34	Chronic Pulmonary Disease	0.03	0.03	0.009	0.009	0.0002	0.0013
35	Connective Tissue Disease-Rheumatic Disease	−0.05	−0.03	0.020	0.020	0.0203	0.1004

#	Variable	Estimate	Estimate (Synth.)	StdErr	StdErr (Synth.)	ProbChiSq	ProbchiSq (Synth.)
36	Peptic Ulcer Disease	−0.06	−0.03	0.025	0.025	0.0189	0.2764
37	Mild Liver Disease	0.82	0.82	0.016	0.017	<.0001	<.0001
	Charlson Index Summary (reference category: 1)						
38	0	−0.64	−0.65	0.010	0.010	<.0001	<.0001
39	2	0.38	0.37	0.011	0.011	<.0001	<.0001
40	3	0.59	0.58	0.016	0.017	<.0001	<.0001
41	4	0.78	0.77	0.022	0.024	<.0001	<.0001
42	5	0.85	0.84	0.030	0.032	<.0001	<.0001
43	6	0.62	0.59	0.040	0.043	<.0001	<.0001
	Imputed Race/Ethnicity (reference category: White)						
44	Asian	−0.21	−0.20	0.017	0.017	<.0001	<.0001
45	Black	−0.04	−0.04	0.007	0.007	<.0001	<.0001
46	Hispanic	−0.24	−0.23	0.009	0.009	<.0001	<.0001
47	Amer. Ind	0.06	0.14	0.070	0.070	0.3613	0.0505
48	**Sex: Female** ***(reference category: Male)***	−0.27	−0.26	0.005	0.005	<.0001	<.0001
	Division (reference category: Mid-Atlantic)						
49	Northeast	−0.05	−0.03	0.018	0.018	0.0128	0.1139
50	East North Central	0.12	0.12	0.009	0.009	<.0001	<.0001
51	West North Central	0.10	0.10	0.013	0.013	<.0001	<.0001
52	South Atlantic	0.21	0.22	0.009	0.009	<.0001	<.0001
53	East South Central	0.17	0.19	0.011	0.011	<.0001	<.0001
54	West South Central	0.89	0.90	0.012	0.012	<.0001	<.0001
55	Mountain	0.18	0.19	0.013	0.013	<.0001	<.0001
56	Pacific	0.40	0.41	0.010	0.010	<.0001	<.0001
	Hospital Bed Size (reference category: > 500)						
57	0 – 25	−0.37	−0.38	0.038	0.038	<.0001	<.0001
58	26 – 100	−0.19	−0.19	0.017	0.017	<.0001	<.0001
59	101 – 500	−0.08	−0.09	0.006	0.006	<.0001	<.0001
	Hospital Ownership (reference category: Non-profit)						
60	For Profit	−0.55	−0.54	0.019	0.019	<.0001	<.0001
61	Government	−0.03	−0.03	0.009	0.009	0.0005	0.0043
	Hospital Type (reference category: General Acute)						
62	Children’s	−0.14	−0.16	0.046	0.045	0.0027	0.0004
63	Psychiatric	0.37	0.35	0.024	0.024	<.0001	<.0001
64	Long-Term Acute Care, Rehab., etc.	0.17	0.20	0.026	0.027	<.0001	<.0001
	Hospital Urban Rural Classification (reference category: Large Central Metropolitan)						
65	Large Fringe Metropolitan	−0.10	−0.09	0.008	0.008	<.0001	<.0001
66	Medium Metropolitan	−0.05	−0.05	0.007	0.007	<.0001	<.0001
67	Small Metropolitan	0.25	0.22	0.011	0.011	<.0001	<.0001

#	Variable	Estimate	Estimate (Synth.)	StdErr	StdErr (Synth.)	ProbChiSq	ProbchiSq (Synth.)
68	Micropolitan	0.19	0.20	0.013	0.013	<.0001	<.0001
69	Non-Core	0.10	0.12	0.045	0.045	0.0218	0.0078

References

1. Breiman Leo, et al. Classification and regression trees. CRC press, 1984
2. Breslow NE (1972) Discussion of the paper by D. R. Cox. J R Statist Soc B 34:216–217
3. Charlson Mary, et al. “A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation.” Journal of Chronic Disease Vol 40, No 5 (1987): 373–383.
4. David Cox R. Regression models and life tables (with discussion). Journal of the Royal Statistical Society 34.2 (1972): 187–220.
5. Fiscella Kevin, and Fremont Allen M.. “Use of geocoding and surname analysis to estimate race and ethnicity.” Health services research 41.4p1 (2006): 1482–1500. [PubMed: 16899020]
6. Landis J. Richard, and Koch Gary G.. “The measurement of observer agreement for categorical data.” Biometrics (1977): 159–174. [PubMed: 843571]
7. Lipkovich Ilya, Ratitch Bohdana, and Michael O’Kelly. “Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints.” Pharmaceutical statistics 15.3 (2016): 216–229. [PubMed: 26997353]
8. National Center for Health Statistics. National Death Index (NDI). http://www.cdc.gov/nchs/ndi/index.htm
9. National Center for Health Statistics. Division of Analysis and Epidemiology. The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations, 8 2019. Hyattsville, Maryland. https://www.cdc.gov/nchs/data/datalinkage/NHCS16_NDI16_17_Methodology_Analytic_Consider.pdf
10. National Center for Health Statistics. National Hospital Care Survey (NHCS). http://www.cdc.gov/nchs/dhcs/index.htm
11. SAS HPSPLIT documentation. https://documentation.sas.com/?docsetId=stathpug&docsetTarget=stathpug_hpsplit_syntax01.htm&docsetVersion=15.1&locale=en
12. SAS PHREG documentation. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#phreg_toc.htm
13. Sundararajan Vijaya, et al. “New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality.” Journal of Clinical Epidemiology 57.12 (2004): 1288–1294.). [PubMed: 15617955]

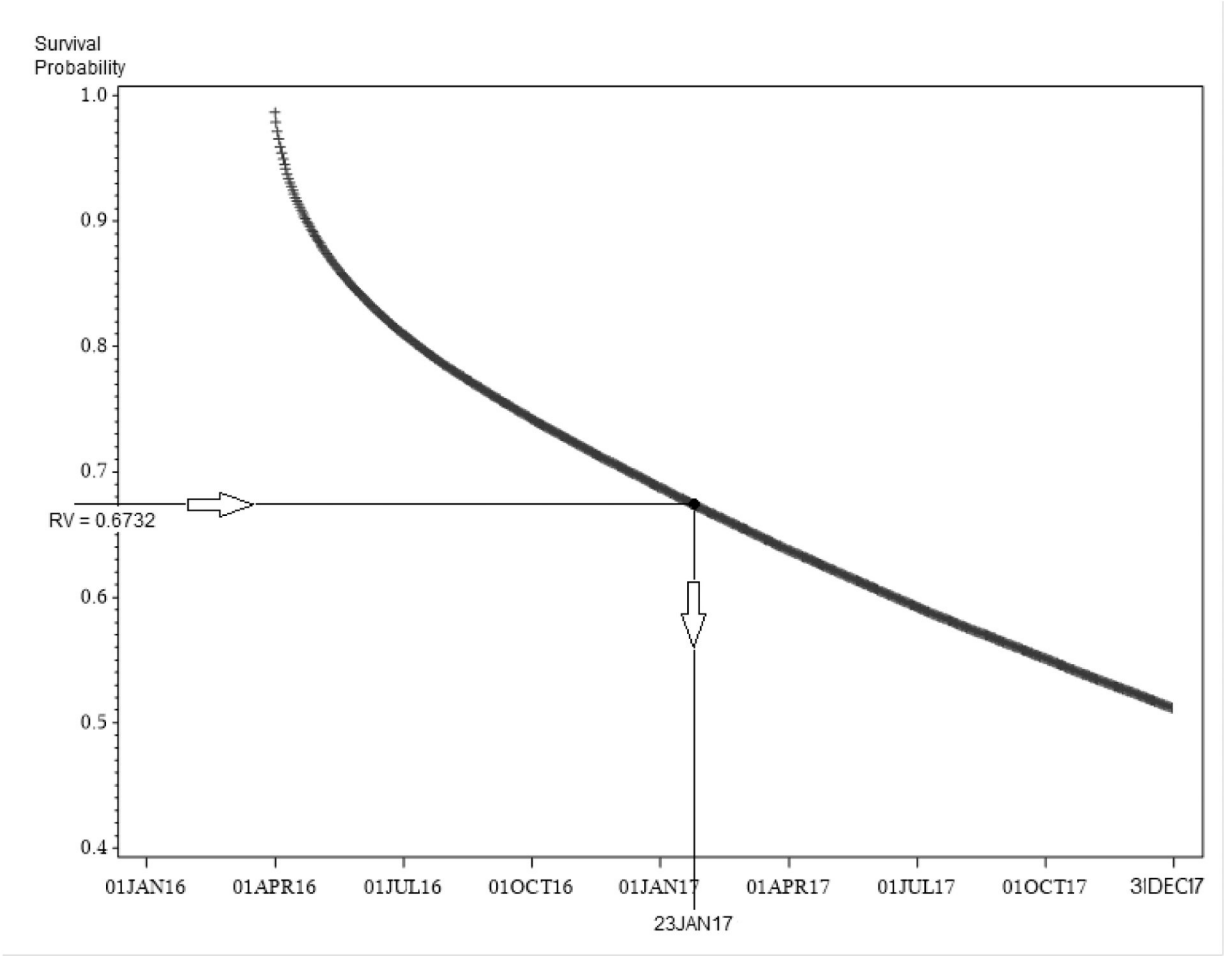

Exhibit 1.
Imputation of Date-of-Death from Estimated Survival Probability

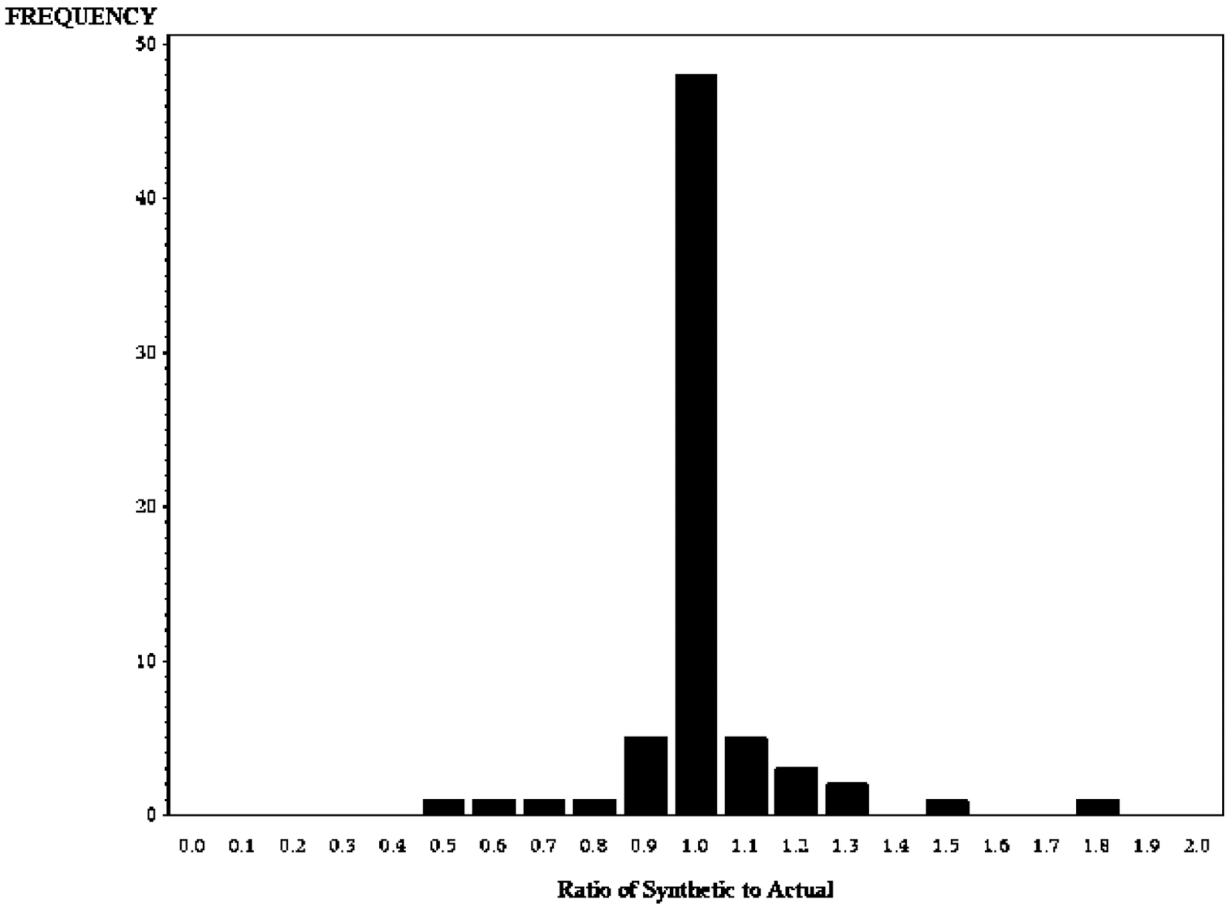

Exhibit 2.

Histograms of the Ratio of Synthetic Survival Analysis Parameter Estimates to Actual Estimates for All Causes of Death

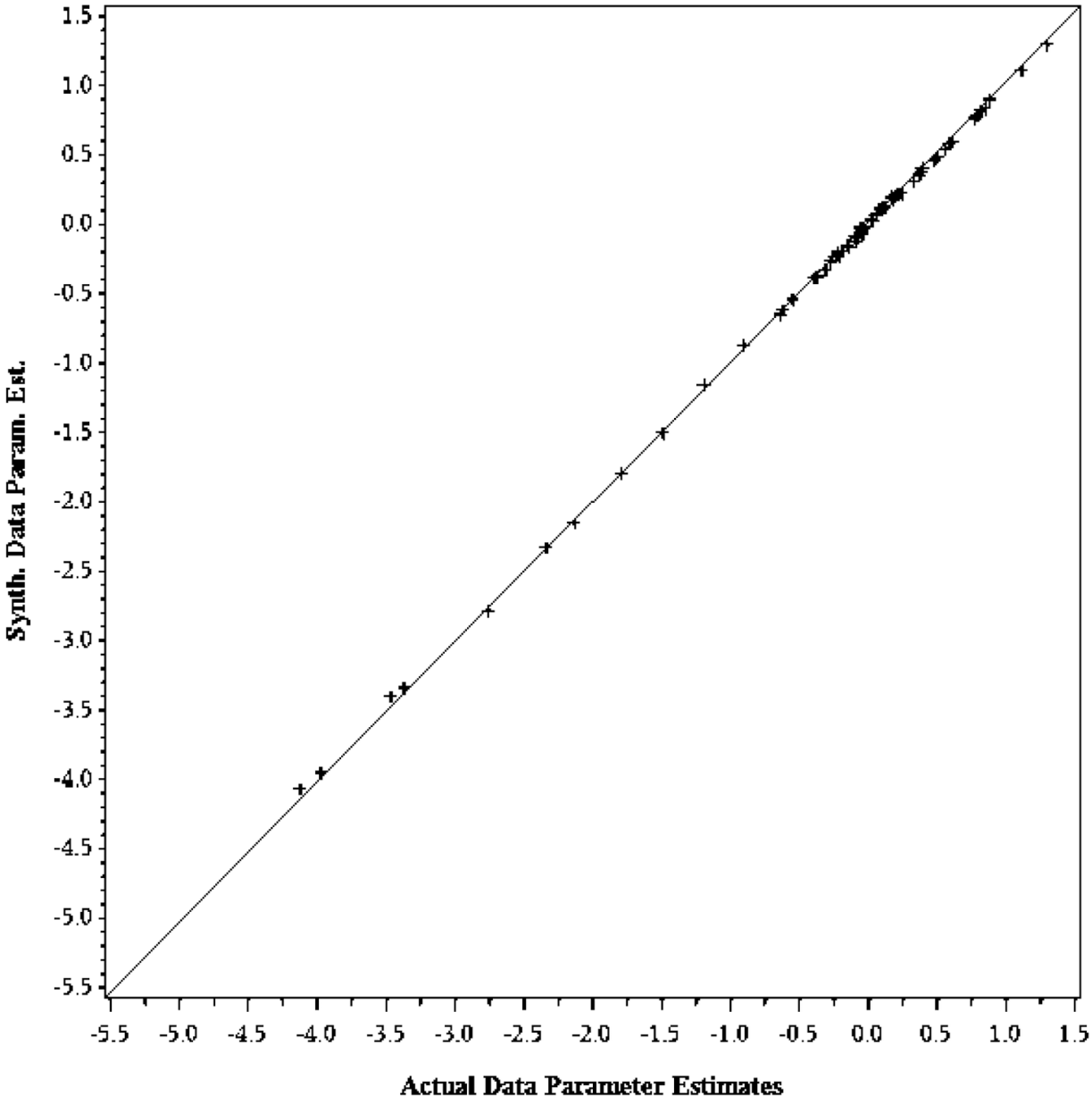

Exhibit 3.
Plot of Synthetic to Actual Survival Parameter Estimates for All Causes of Death

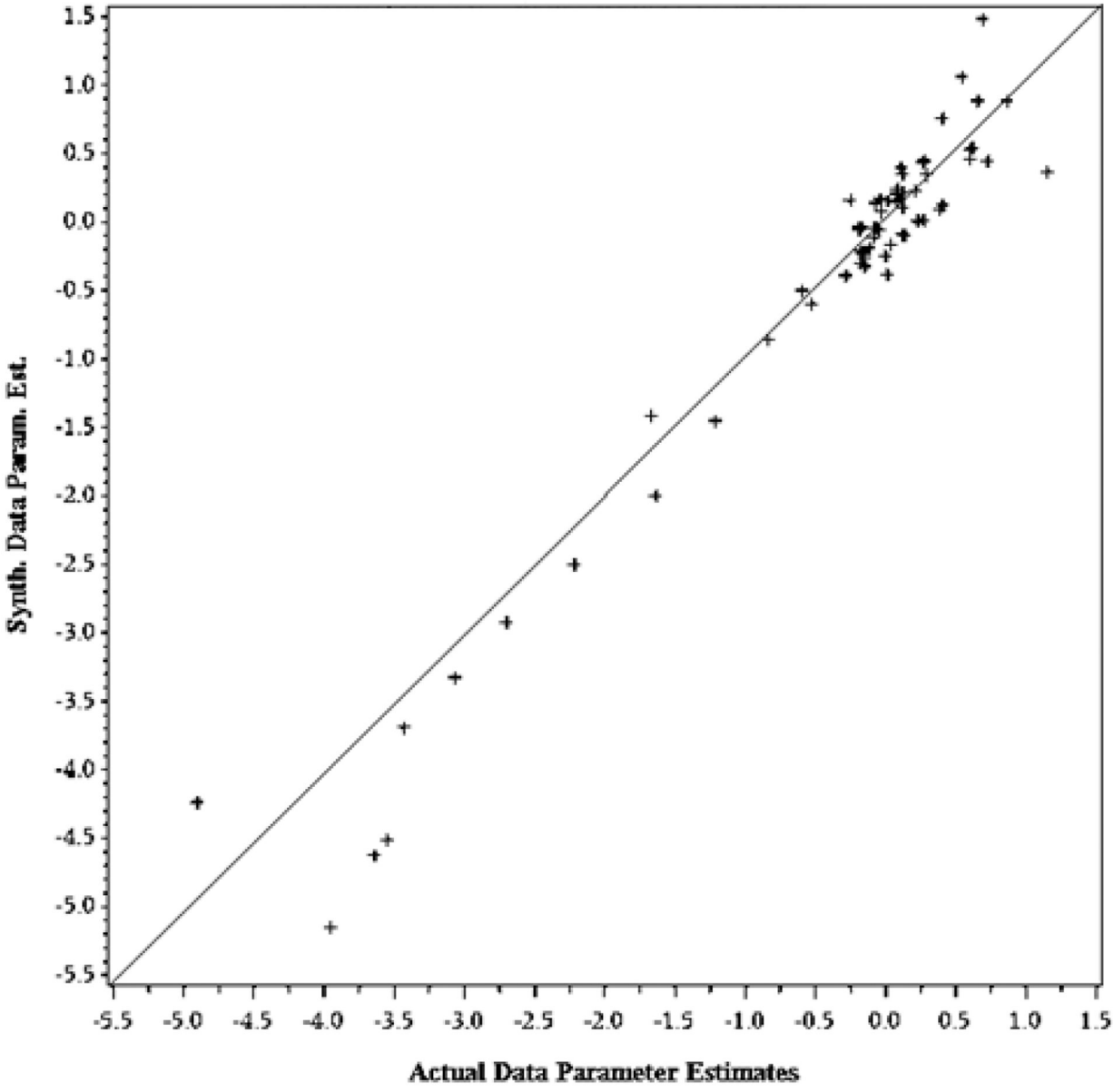

Exhibit 4.
Plot of Synthetic to Actual Survival Parameter Estimates for Cancer Mortality

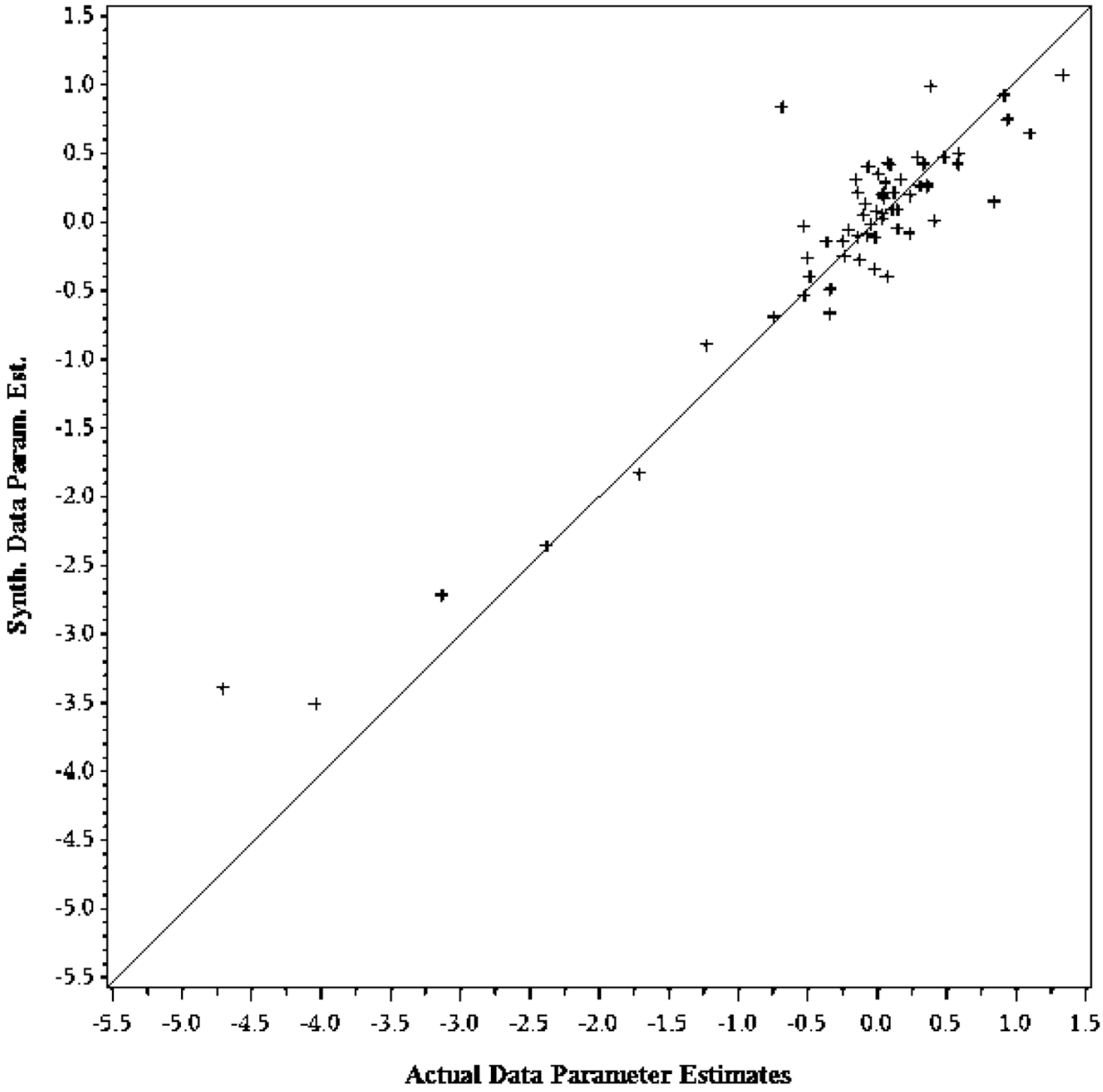

Exhibit 5.
Plot of Synthetic to Actual Survival Parameter Estimates for Heart Disease Mortality

Table 1.

Charlson Index, Conditions and Weights

Condition	Weight
Acute myocardial infarction	1
Congestive heart failure	1
Peripheral vascular disease	1
Cerebrovascular accident	1
Dementia	1
Pulmonary disease	1
Connective tissue disorder	1
Peptic ulcer	1
(Non-Severe) Liver disease	1
Diabetes	1
Diabetes complications	2
Paraplegia	2
Renal disease	2
Cancer	2
Metastatic cancer	3
Severe liver disease	3
HIV	6

Table 2.

Distribution of Underlying Cause of Death from the Synthetic and Actual Linked 2016 NHCS-2016/2017 NDI Data

Underlying cause of death[&]	Synthetic %	Actual %
Malignant neoplasms (cancer)	27.4	26.7
Diseases of the heart	12.6	12.7
Accidents (unintentional injuries)	5.4	5.7
Chronic lower respiratory diseases	5.2	5.4
Cerebrovascular diseases	4.9	4.6
Diabetes mellitus	3.0	3.0
Alzheimer disease	2.8	2.7
Influenza and pneumonia	2.4	2.3
Nephritis, nephrotic syndrome and nephrosis	1.7	1.7

[&]Underlying cause of death codes are based upon International Statistical Classification of Diseases, Injuries and Causes of Death, Tenth Revision, recode into 113 selected causes.

Source: NCHS, 2016 NHCS linked 2016/2017 NDI file

Table 3.

R-Square and Beta Estimates from OLS Model: $Param_{ACT}(Predictor) = \alpha + \beta \cdot Param_{SYNTH}(Predictor) + \varepsilon$ (n=69)

Cause of Death	R-Square	β	Standard error of β
Malignant neoplasms (cancer)	0.98	0.94	0.02
Diseases of the heart	0.94	1.15	0.04
Accidents (unintentional injuries)	0.92	0.93	0.03
Chronic lower respiratory diseases	0.97	1.06	0.02
Cerebrovascular diseases	0.97	1.00	0.02
Diabetes mellitus	0.72	1.31	0.10
Alzheimer disease	0.58	1.44	0.15
Influenza and pneumonia	0.75	0.64	0.05
Nephritis, nephrotic syndrome and nephrosis	0.98	1.00	0.02

Table 4.

Percent Agreement and Concordance Actual vs. Synthetic for the Number of Statistically Significant Survival Parameter Estimates

Cause of Death	Percent agreement at α = 0.01	Kappa	Interpretation of Kappa[4]
All-cause mortality	97.1%	0.84	Almost perfect
Malignant neoplasms (cancer)	78.3%	0.27	slight
Diseases of the heart	73.9%	0.32	slight
Accidents (Unintentional injuries)	72.5%	0.45	fair
Chronic lower respiratory diseases	84.1%	0.67	moderate
Cerebrovascular diseases	78.3%	0.54	fair
Diabetes mellitus	68.1%	0.34	slight
Alzheimer disease	73.9%	0.45	fair
Influenza and pneumonia	76.8%	0.54	fair
Nephritis, nephrotic syndrome and nephrosis	76.8%	0.53	fair

[4](Landis 1977)