BMJ Open

Replicating Secondary Analyses of Clinical Trial Data Using Data Synthesis

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-043497
Article Type:	Original research
Date Submitted by the Author:	06-Aug-2020
Complete List of Authors:	Azizi, Zahra; McGill University Faculty of Medicine Zheng, Mina; Replica Analytics Ltd. Mosquera, Lucy; Replica Analytics Ltd. Pilote, Louise; McGill University, Medicine; Research Institute of the McGill University Health Centre, Centre for Outcomes Research and Evaluation Emam, Khaled; Children's Hospital of Eastern Ontario Research Institute,
Keywords:	EPIDEMIOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Information management < BIOTECHNOLOGY & BIOINFORMATICS, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, STATISTICS & RESEARCH METHODS

Replicating Secondary Analyses of Clinical Trial Data Using Data Synthesis

Zara Aziz[4], Mina Zheng[3], Lucy Mosquera[3], Louise Pilote[4], Khaled El Emam[1,2,3], and the GOING-FWD Collaborators

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[2]Childrens Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada

[3]Replica Analytics Ltd., Ottawa, Ontario, Canada

[4] Center for Outcomes research and Evaluation, McGill University, Montreal, Quebec, Canada

Corresponding Author:

Khaled El Emam
Children's Hospital of Eastern Ontario Research Institute
401 Smyth Road, Ottawa
Ontario K1J 8L1, Canada

E: kelemam@ehealthinformation.ca

Abstract

Objectives: There are increasing requirements to make research data, especially clinical trial data, more broadly available for secondary analyses. However, data availability remains a challenge due to complex privacy requirements. We propose to address this problem using synthetic data.

Setting: Secondary analysis of a published stage 3 colon cancer trial using synthetic data.

Participants: There were 2,686 patients recruited in the original trial.

Primary and Secondary Outcome measures: Analyses from a study published on the real dataset were replicated on synthetic data to investigate the relationship between bowel obstruction and event-free survival. Information theoretic metrics were used to compare the univariate distributions between real and synthetic data. Confidence interval overlap was used to assess the similarity in the size of the bivariate relationship by evaluating the percentage of confidence intervals which overlap for parameters computed from real and synthetic data, and similarly for the multivariate Cox models derived from the two datasets.

Results: Analysis results were similar between the real and synthetic datasets. The univariate distributions were within 1% of difference on an information theoretic metric. The confidence interval overlap for the effect size in bivariate relationships were all above 50%. The main conclusion from the published study, that bowel obstruction has a strong positive impact on survival, was replicated directionally and with the hazard ratio confidence interval overlap between the real and synthetic data varying from 61% to 86%.

Conclusions: The high concordance between the analytic results on synthetic and real data suggests that synthetic data can be used as a reasonable proxy for real clinical trial datasets. Synthetic data provides a good approach for making data more broadly available to meet journal, funder, and regulatory requirements.

Trial Registration (original study): NCT00079274

Strengths and Limitations

- The objective was to evaluate if synthetic data can be used instead of the real data
- A secondary analysis of a published oncology clinical trial was replicated
- The results and conclusions from the real and synthetic data were compared

1. Background

It is often difficult for researchers to get access to high quality individual-level data for secondary purposes (e.g. testing new hypotheses and building statistical and machine learning models). Specifically, for clinical trial data, secondary analysis of data from previous studies can provide new insights compared to the original publications [1], and has produced informative research results including on drug safety, evaluating bias, replication of studies, and meta-analysis [2].

However, data access remains a challenge [3]. An analysis of the success rates of getting individual-level data for research projects from authors found that the percentage of the time these efforts were successful varied significantly and was generally low at 58% [4], 46% [5], 14% [6], and 0% [7].

Therefore, there has been strong interest in making more clinical trial data available for secondary analysis by journals, funders, the pharmaceutical industry and regulators [8]–[13].

For example, the ICMJE's data sharing policy [14] indicates that articles reporting the results of clinical trials must include a data sharing statement when they are submitted to ICMJE journals for publication. Funders also have data sharing requirements. According to the Wellcome Trust's policy [15], researchers receiving funding are expected to share their data rapidly, an outputs management plan is a requirement for any funding proposal which anticipates the generation of significant outputs (e.g. data, software or other materials). These plans are factored into funding decisions. The NIH Statement on Sharing Research Data [16] indicates that applicants seeking $500,000 or more in funding per year are required to include a data sharing plan (or explain why it is not possible to share their research data). Data shared by researchers should be individual level data upon which the accepted publication was based.

One reason for this challenging data sharing environment is increasingly strict data protection regulations: a recent National Academy of Medicine/Government Accountability Office report highlights privacy as presenting a data access barrier for the application of Artificial Intelligence (AI) and machine learning in healthcare [17]. While patient (re-)consent is one legal basis for making data available for secondary purposes, it is often impractical to get retroactive consent under many circumstances and there is significant evidence of consent bias [18].

Anonymization is another approach to making clinical trial data available for secondary analysis. However, recently there have been repeated claims of successful re-identification attacks on anonymized data [19]–[25], eroding public and regulators' trust in this approach [25]–[35].

To solve this data access problem, we propose using synthetic data instead [36]. There are many use cases where synthetic data can provide a practical solution to the data access problem [37], and has been highlighted as a key privacy enhancing technology to enable data access for the coming decade [38].

To test the proposal that synthetic data can be a good proxy for real data, we compare the secondary analysis results from a synthetic version of a trial dataset with the results from a published clinical trial by replicating an analysis for a published oncology clinical trial study using synthetic data. We focus on replicating a secondary analysis rather than a primary analysis because by far the most common purposes for the re-analysis of clinical trial data are new analyses of the treatment effect and the disease state rather than replicating the primary analysis [39]. This will inform us whether the sharing of synthetic clinical trial data will still allow researchers performing new analyses on that data to draw similar conclusions as they would have had the original data been shared.

While the replication of clinical studies on synthetic data has been done before in the context of observational research [40], there has been a dearth of evaluations on clinical trial data. The small dataset size of clinical trials may affect the outcome of such replications. Furthermore, additional evidence on the utility of synthetic data across multiple contexts will inform the development and adoption of this approach to data sharing.

For an oncology trial, we obtained original datasets, synthesized these datasets, replicated a published secondary analysis, and compared the real and synthetic results and conclusions. There have thus far been no studies that examine the ability of synthetic to replicate secondary analyses of clinical trial data. This study is therefore contributing to the evidence base for enabling more access to clinical trial data through synthesis.

2. Methods

2.1 Data Sources

Some researchers note that getting access to datasets from authors can take from 4 months to 4 years [7]. Requests to access clinical trial data can be made to data sharing repositories such as clinicalstudydatarequest.com (CSDR) [41], and Project Data Sphere (PDS) [42]. Early experiences with CSDR noted that the process is lengthy [43]. This is consistent with recent reporting that it takes six months from proposal submission to data access on CSDR [44]. Accessing and downloading data from PDS only takes a few days [43].

We therefore identified a clinical trial from Project Data Sphere (PDS)[1]. The specific trial was selected because the PDS data were analyzed in a published study that we could successfully replicate (validating that we have the correct data and interpreted it the same way as the authors), and the description of the analyses performed was clear enough to allow replication.

[1] See <https://data.projectdatasphere.org/>

2.2 Summary of Original Trial Data

Trial N0147 was a randomized trial of 2,686 patients with stage 3 colon adenocarcinoma that were randomly assigned to adjuvant regimens with or without Cetuximab. After resection of colon cancer, Cetuximab was added to the modified sixth version of the FOLFOX regimen including oxaliplatin plus 5-fluorouracil and leucovorin (mFOLFOX6), fluorouracil, leucovorin, and irinotecan (FOLFIRI), or a hybrid regimen consisting of mFOLFOX6 followed up by FOLFIRI [45]. Our focus is on the secondary retrospective analysis of N0147 (the *published secondary analysis*) [46].

Participants in the control “chemotherapy-only” arm (FOLFOX, FOLFIRI or hybrid regimen without Cetuximab) was analyzed in the published secondary analysis. Presentation with acute obstruction of the bowel is a known risk factor for poor prognosis in patients with colon cancer [47], [48]. The main objective of this secondary analysis was to assess the role of obstruction as an independent risk factor for predicting outcomes in patients with stage 3 colon cancer. The primary endpoint of the study was defined as disease free survival (DFS) which was defined as time from random allocation to the first recurrence or death from any cause. The secondary endpoint was overall survival (OS) defined as time from random allocation to death from any cause.

The covariates in the published secondary analysis comprised of three types of variables: 1) Baseline demographics, including age, sex, and baseline BMI, 2) Baseline Eastern cooper- active oncology group (ECOG) performance score that describes patients’ level of functioning in terms of their ability to care for themselves, daily activity and physical ability, and 3) Baseline cancer characteristics, including clinical T stage, lymph node involvement, histologic status, and Kirsten rat sarcoma virus (KRAS) biomarker status.

2.3 Data Synthesis

We used sequential decision trees for data synthesis. Sequential decision trees are used quite extensively in the health and social sciences for the generation of synthetic data [49]–[57]. In these models, a variable is synthesized by using variables preceding it in the sequence as predictors. The method we used to generate synthetic data is called conditional trees [58], although other tree algorithms could also be used. A summary of the algorithm is provided in Figure 1. Other methods for data synthesis have been proposed in the literature for health data, such as deep learning [59][60]. However, compared to deep learning synthesis methods, sequential decision trees have the advantage of not requiring large training datasets. It is therefore suitable for synthesizing clinical trial data of this size.

A partial synthesis was performed on the trial dataset. The partial synthesis ensured that potentially identifying information in the dataset (the quasi-identifiers [61]) were synthesized. Quasi-identifiers are the variables that are potentially knowable by an adversary and can be used for re-identification attacks [61]–[63]. Such information is knowable because it is in the public domain (e.g., in obituaries or registries, such as voter registration lists), or is known by an adversary who is an acquaintance of someone in the dataset (e.g., a neighbor or relative).

The quasi-identifiers selected for the N0147 trial were age, gender, race, BMI, OS, DFS (since death status would be known by an adversary). All dates were converted to relative dates (consistent with a contemporary clinical trial de-identification standard [66]).

The synthesis of the quasi-identifiers used all the remaining information in the dataset to ensure that the relationships were maintained in the generated data. Only synthesizing the quasi-identifiers to protect against identification risks is consistent with the clinical trial data anonymization guidelines from the European Medicines Agency [64] and Health Canada [65].

2.4 Replication of Secondary Analysis on the Synthetic Data

We first replicated the published analysis on the original dataset. Once the results could be replicated, we re-ran the exact same analysis R code on the synthetic version of the data.

We first replicated the published analysis on the original dataset. Once the results could be replicated, we re-ran the exact same analysis R code on the synthetic version of the data.

The published secondary analysis [46] included descriptive statistics consisting of frequency (percentage) for categorical variables. The Pearson χ^2 test was used to investigate the statistical significance of the relationship between the baseline characteristics (clinical, pathological) and obstruction. Survival analysis was performed using the Kaplan-Meier curve. The log rank test and Cox proportional hazards model were used to plot OS and DFS at 5 years and to create a model adjusted for baseline clinical and pathological characteristics to assess the role of obstruction in predicting OS and DFS.

2.5 Evaluation of Results

Our objective is to evaluate the utility of the synthetic data. This means that we compare the analysis results using the original data with the analysis results using the synthetic data. Our utility evaluation method followed the recommendations to evaluate the utility of data that has been transformed to protect privacy, such as through data synthesis [67]. Specifically, we used two general approaches to compare real and synthetic analysis results: information theoretic methods based on the Kullback-Leibler (KL) divergence, and interval overlap for the confidence intervals of model parameters. Both are described further below.

To evaluate the utility of synthetic data we compared the published univariate and bivariate statistics on the original data and the synthetic data. We then also compared the multivariate model parameters for the models that were developed to explain survival and test the hypothesis that obstruction was an important predictor.

2.5.1 Univariate Analysis

The univariate results consisted of distributions on the categories of the variables (the relevant continuous variables were categorized in the published secondary analysis study). Relative entropy (KL-divergence) is often used in machine learning to compare two distributions and is given by [68]. However, KL-divergence is difficult to interpret because it has no fixed upper bound and is not compared to a yardstick to obtain a relative interpretation. We therefore convert it to a relative value so that it can be interpreted more easily.

By dividing KL-divergence by Shannon's entropy we get the *relative increase in entropy due to using synthetic data*, and we use it to compare the univariate distributions of the real and synthetic datasets. It is a form of normalization of the relative entropy to make it interpretable (in the same way that relative error is interpreted when computing model prediction accuracy). A value of zero means that there are no differences in the distributions. A value of one means that the entropy or uncertainty due to the use of synthetic data as opposed to the real data is twice that of using the real data.

2.5.2 Bivariate Analysis

In the published secondary analysis, the bivariate results were presented as contingency tables showing the cross-tabulations of the predictors with obstruction, OS after five years, and DFS after five years. The Pearson χ^2 test was used to evaluate all bivariate relationships. This type of testing when used in the current context has a number of disadvantages: (a) it does not give us an effect size and therefore we would not know if a bivariate relationship was strong or not (a test statistic can be significant with a very small effect size if there are many observations), (b) the tests did not account for multiple-testing, such as a Bonferroni adjustment, which means that there will be an elevated probability of finding significant results by chance, and (c) the chi-square tests considers independence whereas the relationship that is being tested is whether each of the covariates are predictive of the outcome. For these reasons we used a different statistic to compute the bivariate relationships on the original and synthetic datasets.

We use the Goodman and Kruskal tau statistic, which gives us a measure between zero and one of the extent to which the covariate is predictive of the outcome [69]. The effect size is computed for the real dataset, τ_r , and the synthetic dataset, τ_s . We compared the confidence interval overlap of the tau statistics between the two datasets. Confidence interval overlap has been proposed for evaluating the utility of privacy protective data transformations [67], which is defined as the percentage average of the real and synthetic confidence intervals that overlap:

$$\frac{1}{2}\times\left(\frac{\max\left(0,\min\left(u_r,u_s\right)-\max\left(l_r,l_s\right)\right)}{u_r-l_r}+\frac{\max\left(0,\min\left(u_r,u_s\right)-\max\left(l_r,l_s\right)\right)}{u_s-l_s}\right)\times 100 \qquad (1)$$

where u and l indicate the upper and lower limits of the confidenc einterval, and the r and s subscripts indicate real and symthetic data. This formulation gives an overlap of zero if the two intervals do not overlap at all. We express overlap as a percentage.

The published secondary analysis evaluated the bivariate relationship between each of the predictors and obstruction, and then evaluated each of the predictors and obstruction with event free survival. We repeated these analyses with tau statistic and confidence interval to provide a meaningful effect size.

2.5.3 Multivariate Analysis

For the multivariate models, we compared the confidence interval overlap of the Cox model parameters. Confidence interval comparisons using equation (1) was used for comparing the confidence intervals of the hazard ratios of the model.

3. Results for Trial N0147

We compare the results in the secondary analysis study that were published against the same analyses performed on the synthetic data.

4. Univariate Analyses

The first set of comparisons is shown in Table 1 with the univariate comparisons of the distributions on the I_1 metric. As can be seen, all of the values were less than 1%, therefore the relative increase in entropy is quite low due to data synthesis. The values that are zero in the table pertain to variables that were not synthesized in the partial synthesis process.

Variable	I_1
Age	0.147%
Sex	0.35%
BMI	0.06%
ECOG	0%
Race	0.049%
KRAS	0%
T Stage	0%
Histology	0%
Adjuvant Chemotherapy	0.095%
Positive LNs	0%
Adjuvant Regimen	0%
Overall survival	0.054%
Disease free survival	0.017%

Table 1: Comparing the real and synthetic univariate distributions.

4.1 Bivariate Analysis

The differences between real and synthetic data for the bivariate relationships of the covariates and obstruction are shown in Figure 2. When we look at the effect sizes (the tau metric) we see that the size of these bivariate relationships is very small. These covariates are not good predictors of obstruction. We also note that the effect sizes are similar between the real and synthetic datasets, and there are considerable confidence intervals overlap. One would draw the same conclusions from the real and synthetic datasets.

The next set of results are also the bivariate relationships between the covariates and the event free survival outcomes: overall survival and disease-free survival. The results in Figure 3 shows the effect sizes for the bivariate relationships with overall survival. There are two important observations. The first observation is that all the bivariate relationships are very weak – the covariate are not individually predictive of disease-free event outcomes. The second observation is that the effect sizes are very similar between the real and synthetic datasets. One would draw the same conclusions from the synthetic data as from the data in the published secondary analysis.

Figure 4 shows the bivariate relationships with disease-free survival. The conclusions are like overall survival overall with one exception. The confidence intervals for the relationship between race and DFS do not overlap. However, the relationship is quite weak in both datasets and therefore the conclusions would be the same in both cases.

4.2 Multivariate Analysis

For the multivariate analyses, the real results are the same as those that were in the published secondary analysis. We first compare the survival curves for obstructed and non-obstructed patients on overall survival (Figure 5) and disease-free survival (Figure 6). We can see that the curves are very similar between the real and synthetic datasets.

The Cox models were intended to evaluate whether obstruction affects survival after accounting for the potential confounding effect of other covariates. The real and synthetic hazard ratio model parameters are generally in the same direction with relatively high overlap for the confidence intervals. This is the case for the overall survival model in Figure 7 and the disease free survival model in Figure 8.

The main hypothesis being tested in the published secondary analysis pertains to obstruction. For the OS model the obstruction overlap was quite high at 0.86 (1.56; 95% CI: 1.11-2.2 for real data, and 2.03; 95% CI: 1.44-2.87 for synthetic data) with both models showing a strong positive effect of obstruction on OS. Similarly for the DFS model, the real data parameter of 1.51 had a confidence interval of 1.18-1.95 and the synthetic data parameter of 1.63 had a confidence interval of 1.26-2.1, indicating that models show a positive association between obstruction and DFS. Therefore, one would draw the same conclusions about the negative impact of obstruction on event-free survival.

5. Discussion and Conclusions

5.1 Summary

The purpose of this study was to evaluate the extent to which a published secondary analysis of an oncology clinical trial could be replicated using a synthetic variant of the dataset. This replication is one of the first to test whether the same result sand conclusions would be drawn from the analysis of a synthetic version of a clinical trial dataset.

The published secondary analysis was investigating the relationship between bowel obstruction and event-free survival for colon cancer patients. We applied a commonly used synthesis approach that ensured the potentially identifying variables (the quasi-identifiers) in the dataset were appropriately synthesized.

We found that for the univariate and bivariate analyses in the published study, the synthetic data was quite similar in terms of distributions and effect sizes to the real data. With respect to the multivariate models that controlled for confounders, the published results were replicated in that there was a strong positive relationship between obstruction and overall survival and disease-free survival after five years in the both the real and synthetic datasets.

While this is a replication of a single clinical trial, it does provide evidence that synthesized datasets can be used as a reasonable proxy for real datasets. The data synthesis method is well established and has been applied extensively in the health social sciences. Further such replications should be performed to increase the weight evidence on the effectiveness of synthetic data as a proxy for real datasets. To the extent that synthetic data would allow drawing the same conclusions as real data, they can be more readily shared by researchers when publishing their studies and to meet funder requirements for data sharing.

In this particular case the dataset was available for us to use. However, in other situations where it is not easy to share the original data because of privacy concerns, a case can be made for sharing the synthetic dataset instead.

5.2 Limitations

This is an assessment of the ability to replicate a secondary analysis for a single trial. While this is a starting point, additional evaluations are necessary to increase the weight of evidence in support of using synthetic data as a proxy for the real dataset.

While we found that there were very little difference between the real and synthetic data on the bivariate comparisons, one may hypothesize that this was influenced by the fact that the effect sizes were small. However, was not the case for the multivariate models where the effect sizes were larger and the differences between the real and synthetic datasets remained small.

We did not explicitly evaluate the privacy risks in the synthetic data. However, multiple researchers have noted that synthetic data does not have an elevated identity disclosure (privacy) risk [59], [70]–[76]. In addition, we used a common data synthesis method that is widely applied in the health sciences.

Ethics

This project was approved by the CHEO Research Institute Research Ethics Board, protocol number CHEOREB# 20/75X.

Patient and Public Involvement

The comparative analysis of synthetic to real data did not have any patient or public involvement.

Acknowledgements

This article is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this article.

Data Statement

The data used in this study can be obtained from Project Data Sphere.

Funding Statement

This work was partially funded by a Discovery Grant RGPIN-2016-06781 from the Natural Sciences and Engineering Research Council of Canada, and by Replica Analytics Ltd.

Competing Interests Statement

This work was performed in collaboration with Replica Analytics Ltd. This company is a spin-off from the Children's Hospital of Eastern Ontario Research Institute. KEE is co-founder and has equity in this company. LM and MZ are data scientists employed by Replica Analytics Ltd.

Author Contributions

AZ contributed to designing the study, performing the analysis, and writing the paper. KEE contributed to designing the study, to performing the analysis, and to writing the paper. LM contributed to designing the study, implemented some of the code used to perform the analysis, and contributed to writing the paper. MZ contributed to designing the study, implemented some of the code used to perform the analysis, contributed to the data analysis, and contributed to writing the paper. LP contributed to designing the study and to writing the paper.

6. References

[1] Ebrahim S, Sohani ZN, Montoya L, and et al, "Reanalyses of randomized clinical trial data," *JAMA*, vol. 312, no. 10, pp. 1024–1032, Sep. 2014, doi: 10.1001/jama.2014.9646.

[2] Jean-Marc Ferran and Sarah Nevitt, "European Medicines Agency Policy 0070: an exploratory review of data utility in Clinical Study Reports for research," *BMC Medical Research Methodology*, vol. 19, 2019, [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0836-3.

[3] P. Doshi, "Data too important to share: do those who control the data control the message?," *BMJ*, vol. 352, Mar. 2016, doi: 10.1136/bmj.i1027.

[4] J. R. Polanin, "Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing," *Journal of Clinical Epidemiology*, vol. 98, pp. 157–159, Jun. 2018, doi: 10.1016/j.jclinepi.2017.12.014.

[5] F. Naudet *et al.*, "Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine," *BMJ*, vol. 360, Feb. 2018, doi: 10.1136/bmj.k400.

[6] B. Villain, A. Dechartres, P. Boyer, and P. Ravaud, "Feasibility of individual patient data meta-analyses in orthopaedic surgery," *BMC Med*, vol. 13, no. 1, p. 131, Jun. 2015, doi: 10.1186/s12916-015-0376-6.

[7] M. Ventresca *et al.*, “Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide,” *BMC Medical Research Methodology*, vol. 20, no. 1, p. 113, May 2020, doi: 10.1186/s12874-020-00964-6.

[8] Phrma & EFPIA, “Principles for Responsible Clinical Trial Data Sharing,” Jul. 2013. [Online]. Available: http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

[9] TransCelerate Biopharma, “DE-IDENTIFICATION AND ANONYMIZATION OF INDIVIDUAL PATIENT DATA IN CLINICAL STUDIES: A Model Approach,” 2017.

[10] TransCelerate Biopharma, “PROTECTION OF PERSONAL DATA IN CLINICAL DOCUMENTS – A MODEL APPROACH,” Jan. 2017.

[11] European Medicines Agency, “European Medicines Agency policy on publication of data for medicinal products for human use: Policy 0070.” Oct. 02, 2014, [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

[12] D. B. Taichman *et al.*, “Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal EditorsSharing Clinical Trial Data,” *Ann Intern Med*, vol. 164, no. 7, pp. 505–506, Apr. 2016, doi: 10.7326/M15-2928.

[13] Institute of Medicine, “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk,” Washington, D.C., 2015.

[14] International Committee of Medical Journal Editors, “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals,” 2019. http://www.icmje.org/icmje-recommendations.pdf (accessed Jun. 29, 2020).

[15] The Wellcome Trust, “Policy on data, software and materials management and sharing,” *Wellcome*, 2017. https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing (accessed Sep. 12, 2017).

[16] National Institutes of Health, “Final NIH Statement on Sharing Research Data,” 2003. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed Jun. 29, 2020).

[17] “Artificial Intelligence in Health Care,” National Academy of Medicine and the General Accountability Office, Dec. 2019.

[18] K. El Emam, E. Jonker, E. Moher, and L. Arbuckle, “A Review of Evidence on Consent Bias in Research,” *American Journal of Bioethics*, vol. 13, no. 4, pp. 42–44, 2013.

[19] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific Reports*, vol. 3, Mar. 2013, doi: 10.1038/srep01376.

[20] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. “Sandy” Pentland, “Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata,” *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297.

[21] L. Sweeney, J. Su Yoo, L. Perovich, K. E. Boronow, P. Brown, and J. Green Brody, “Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study,” *Journal of Technology Science*, no. 2017082801, Aug. 2017, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2017082801/.

[22] J. Su Yoo, A. Thaler, L. Sweeney, and J. Zang, “Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data,” *Journal of Technology Science*, no. 2018100901, Oct. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018100901/.

[23] L. Sweeney, “Matching Known Patients to Health Records in Washington State Data,” Harvard University. Data Privacy Lab, 2013.

[24] L. Sweeney, M. von Loewenfeldt, and M. Perry, “Saying it’s Anonymous Doesn’t Make It So: Re-identifications of ‘anonymized’ law school data,” *Journal of Technology Science*, no. 2018111301, Nov. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018111301/.

[25] A. Zewe, “Imperiled information: Students find website data leaks pose greater risks than most people realize,” *Harvard John A. Paulson School of Engineering and Applied Sciences*, Jan. 17, 2020. https://www.seas.harvard.edu/news/2020/01/imperiled-information (accessed Mar. 23, 2020).

[26] K. Bode, “Researchers Find ‘Anonymized’ Data Is Even Less Anonymous Than We Thought,” *Motherboard: Tech by Vice*, Feb. 03, 2020. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought (accessed May 11, 2020).

[27] E. Clemons, “Online Profiling and Invasion of Privacy: The Myth of Anonymization,” *HuffPost*, Feb. 20, 2013.

[28] C. Jee, “You’re very easy to track down, even when your data has been anonymized,” *MIT Technology Review*, Jul. 23, 2019. https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/ (accessed May 11, 2020).

[29] G. Kolata, “Your Data Were ‘Anonymized’? These Scientists Can Still Identify You,” *The New York Times*, Jul. 23, 2019.

[30] N. Lomas, “Researchers spotlight the lie of ‘anonymous’ data,” *TechCrunch*, Jul. 24, 2019. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/ (accessed May 11, 2020).

[31] S. Mitchell, “Study finds HIPAA protected data still at risks,” *Harvard Gazette*, Mar. 08, 2019. https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/ (accessed May 11, 2020).

[32] S. A. Thompson and C. Warzel, “Twelve Million Phones, One Dataset, Zero Privacy,” *The New York Times*, Dec. 19, 2019.

[33] “‘Anonymised’ data can never be totally anonymous, says study,” *the Guardian*, Jul. 23, 2019.

[34] Alex van der Wolk, “The (Im)Possibilities of Scientific Research Under the GDPR,” *Cybersecurity Law Report*, Jun. 17, 2020.

[35] S. Ghafur, J. V. Dael, M. Leis, A. Darzi, and A. Sheikh, “Public perceptions on data sharing: key insights from the UK and the USA,” *The Lancet Digital Health*, vol. 0, no. 0, Jul. 2020, doi: 10.1016/S2589-7500(20)30161-8.

[36] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O’Reilly, 2020.

[37] Khaled E Emam and Richard Hoptroff, “The synthetic data paradigm for using and sharing data,” *Cutter Executive Update*, vol. 19, no. 6, 2019.

[38] Jules Polonetsky and Elizabeth Renieris, “10 Privacy Risks and 10 Privacy Technologies to Watch in the Next Decade,” Future of Privacy Forum, Jan. 2020.

[39] A. M. Navar, M. J. Pencina, J. A. Rymer, D. M. Louzao, and E. D. Peterson, “Use of Open Access Platforms for Clinical Trial Data,” *JAMA*, vol. 315, no. 12, p. 1283, Mar. 2016, doi: 10.1001/jama.2016.2374.

[40] A. R. Benaim *et al.*, “Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies,” *JMIR Medical Informatics*, vol. 8, no. 2, p. e16492, 2020, doi: 10.2196/16492.

[41] CSDR: Clinical Study Data Request", 2015. https://www.clinicalstudydatarequest.com/ (accessed Nov. 26, 2015).

[42] CEO Life Sciences Consortium, “Share, Integrate & Analyze Cancer Research Data | Project Data Sphere.” https://projectdatasphere.org/projectdatasphere/html/home.

[43] N. Geifman, J. Bollyky, S. Bhattacharya, and A. J. Butte, “Opening clinical trial data: are the voluntary data-sharing portals enough?,” *BMC Medicine*, vol. 13, no. 1, p. 280, Nov. 2015, doi: 10.1186/s12916-015-0525-y.

[44] National Academies of Sciences, Engineering, and Medicine, *Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop*. 2020.

[45] S. R. Alberts *et al.*, “Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial,” *JAMA*, vol. 307, no. 13, pp. 1383–1393, Apr. 2012, doi: 10.1001/jama.2012.385.

[46] F. S. Dahdaleh *et al.*, “Obstruction predicts worse long-term outcomes in stage III colon cancer: A secondary analysis of the N0147 trial,” *Surgery*, vol. 164, no. 6, pp. 1223–1229, 2018, doi: 10.1016/j.surg.2018.06.044.

[47] P. G. Carraro, M. Segala, B. M. Cesana, and G. Tiberio, “Obstructing colonic cancer: failure and survival patterns over a ten-year follow-up after one-stage curative surgery,” *Dis. Colon Rectum*, vol. 44, no. 2, pp. 243–250, Feb. 2001, doi: 10.1007/BF02234300.

[48] J. Mella, A. Biffin, A. G. Radcliffe, J. D. Stamatakis, and R. J. Steele, “Population-based audit of colorectal cancer management in two UK health regions. Colorectal Cancer Working Group, Royal College of Surgeons of England Clinical Epidemiology and Audit Unit,” *Br J Surg*, vol. 84, no. 12, pp. 1731–1736, Dec. 1997.

[49] J. Drechsler and J. P. Reiter, “An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets,” *Computational Statistics & Data Analysis*, vol. 55, no. 12, pp. 3232–3243, Dec. 2011, doi: 10.1016/j.csda.2011.06.006.

[50] R. C. Arslan, K. M. Schilling, T. M. Gerlach, and L. Penke, “Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior,” *J Pers Soc Psychol*, Aug. 2018, doi: 10.1037/pspp0000208.

[51] D. Bonnéry *et al.*, “The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-Agency Longitudinal Data,” *Journal of Research on Educational Effectiveness*, vol. 12, no. 4, pp. 616–647, Oct. 2019, doi: 10.1080/19345747.2019.1631421.

[52] A. Sabay, L. Harris, V. Bejugama, and K. Jaceldo-Siegl, “Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data,” *SMU Data Science Review*, vol. 1, no. 3, Aug. 2018, [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/12.

[53] Michael Freiman, Amy Lauger, and Jerome Reiter, “Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau,” US Census Bureau, Working paper, 2017.

[54] B. Nowok, “Utility of synthetic microdata generated using tree-based methods,” 2015.

[55] G. M. Raab, B. Nowok, and C. Dibben, “Practical Data Synthesis for Large Samples,” *1*, vol. 7, no. 3, pp. 67–97, 2016, doi: 10.29012/jpc.v7i3.407.

[56] B. Nowok, G. M. Raab, and C. Dibben, “Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1,” *Statistical Journal of the IAOS*, vol. 33, no. 3, pp. 785–796, Jan. 2017, doi: 10.3233/SJI-150153.

[57] D. S. Quintana, “A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation,” *eLife*, vol. 9, 2020, doi: 10.7554/eLife.53275.

[58] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased Recursive Partitioning: A Conditional Inference Framework,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Sep. 2006, doi: 10.1198/106186006X133933.

[59] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, “Data synthesis based on generative adversarial networks,” *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018, doi: 10.14778/3231751.3231757.

[60] K. Chin-Cheong, T. Sutter, and J. E. Vogt, “Generation of Heterogeneous Synthetic Electronic Health Records using GANs,” presented at the Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Dec. 2019, doi: 10.3929/ethz-b-000392473.

[61] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[62] K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O’Reilly, 2013.

[63] K. El Emam, S. Rodgers, and B. Malin, “Anonymising and Sharing Individual Patient Data,” *BMJ*, vol. 350, p. h1139, Mar. 2015, doi: 10.1136/bmj.h1139.

[64] European Medicines Agency, “External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use,” Sep. 2017.

[65] Health Canada, “Guidance document on Public Release of Clinical Information,” Apr. 01, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html (accessed Jun. 04, 2019).

[66] PhUSE De-Identification Working Group, “De-Identification Standards for CDISC SDTM 3.2,” 2015.

[67] A. Karr, C. Koonen, A. Oganian, J. Reiter, and A. Sanil, “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality: The American Statistician: Vol 60, No 3,” *The American Statistician*, vol. 60, no. 3, pp. 224–232, 2006.

[68] Thomas Cover and Joy Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[69] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, New Jersey: Wiley, 2002.

[70] J. P. Reiter, “New Approaches to Data Dissemination: A Glimpse into the Future (?),” *CHANCE*, vol. 17, no. 3, pp. 11–15, Jun. 2004, doi: 10.1080/09332480.2004.10554907.

[71] J. Hu, “Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data,” *arXiv:1804.02784 [stat]*, Apr. 2018, Accessed: Mar. 15, 2019. [Online]. Available: http://arxiv.org/abs/1804.02784.

[72] J. Taub, M. Elliot, M. Pampaka, and D. Smith, “Differential Correct Attribution Probability for Synthetic Data: An Exploration,” in *Privacy in Statistical Databases*, 2018, pp. 122–137.

[73] J. Hu, J. P. Reiter, and Q. Wang, “Disclosure Risk Evaluation for Fully Synthetic Categorical Data,” in *Privacy in Statistical Databases*, 2014, pp. 185–199.

[74] L. Wei and J. P. Reiter, “Releasing synthetic magnitude microdata constrained to fixed marginal totals,” *Statistical Journal of the IAOS*, vol. 32, no. 1, pp. 93–108, Jan. 2016, doi: 10.3233/SJI-160959.

[75] N. Ruiz, K. Muralidhar, and J. Domingo-Ferrer, “On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective,” in *Privacy in Statistical Databases*, 2018, pp. 59–74.

[76] J. P. Reiter, “Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 168, no. 1, pp. 185–205, 2005, doi: 10.1111/j.1467-985X.2004.00343.x.

Figures

Figure 1: A description of the sequential data synthesis process using classification and regression trees. Although any set of classification and regression methods can be used in principle.

Figure 2: The tau coefficient on the real and synthetic data, and the confidence interval overlap for the bivariate relationship with obstruction.

Figure 3: The effect size for the real and synthetic variables against overall survival.

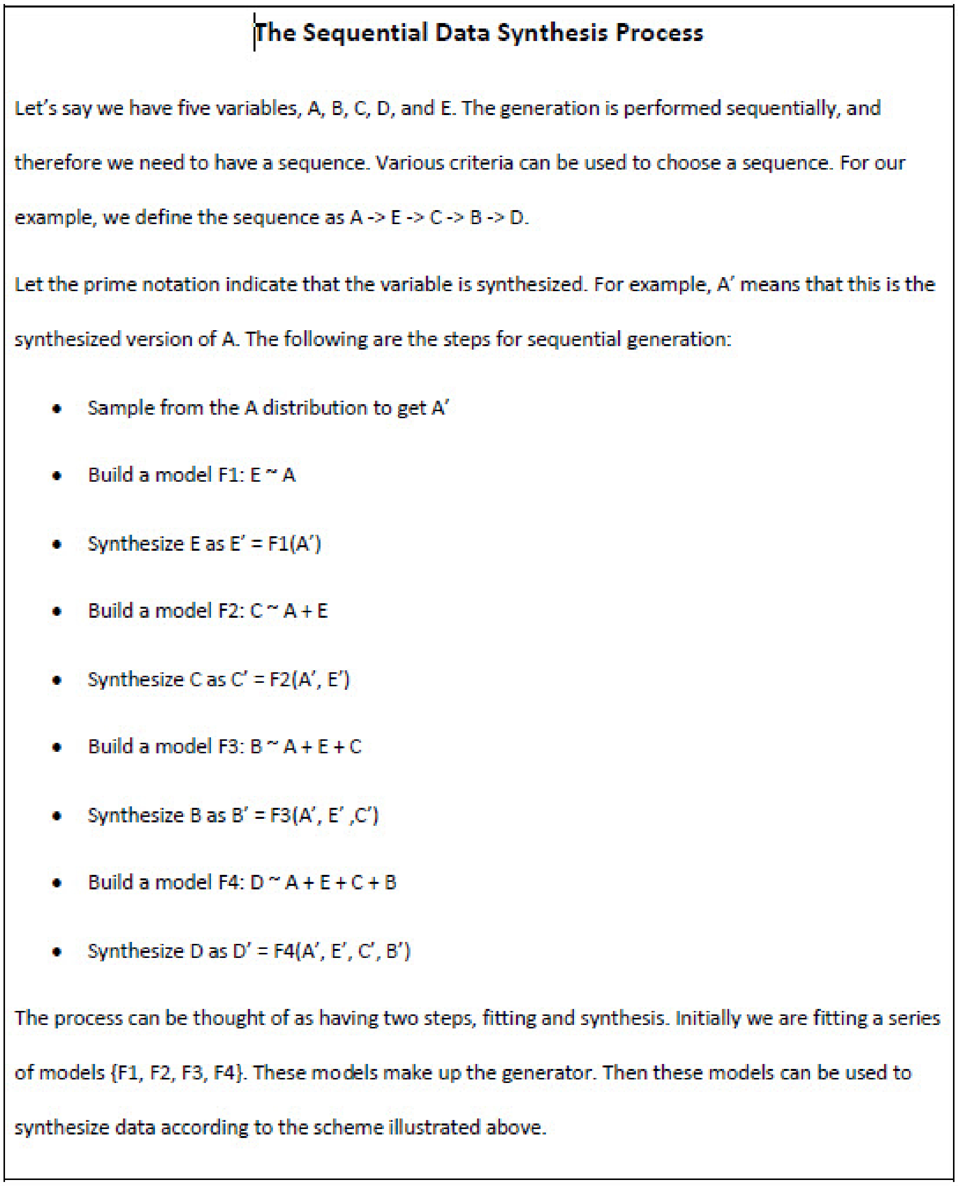
Figure 4: The effect size for the real and synthetic variables against disease-free survival.

Figure 5: Survival curve comparing overall survival (OS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B) .

Figure 6: Survival curve comparing disease free survival (DFS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B) .

Figure 7: Comparison of real and synthetic Cox model parameters (hazard ratios) with the overall survival outcome variable for the real and synthetic datasets.

Figure 8: Comparison of real and synthetic Cox model parameters (hazard ratios) with the disease free survival outcome variable for the real and synthetic datasets.



The Sequential Data Synthesis Process

Let’s say we have five variables, A, B, C, D, and E. The generation is performed sequentially, and therefore we need to have a sequence. Various criteria can be used to choose a sequence. For our example, we define the sequence as A -> E -> C -> B -> D.

Let the prime notation indicate that the variable is synthesized. For example, A’ means that this is the synthesized version of A. The following are the steps for sequential generation:

- Sample from the A distribution to get A’
- Build a model F1: E ~ A
- Synthesize E as E’ = F1(A’)
- Build a model F2: C ~ A + E
- Synthesize C as C’ = F2(A’, E’)
- Build a model F3: B ~ A + E + C
- Synthesize B as B’ = F3(A’, E’ ,C’)
- Build a model F4: D ~ A + E + C + B
- Synthesize D as D’ = F4(A’, E’, C’, B’)

The process can be thought of as having two steps, fitting and synthesis. Initially we are fitting a series of models {F1, F2, F3, F4}. These models make up the generator. Then these models can be used to synthesize data according to the scheme illustrated above.

Figure 1: A description of the sequential data synthesis process using classification and regression trees. Although any set of classification and regression methods can be used in principle.

117x144mm (120 x 120 DPI)

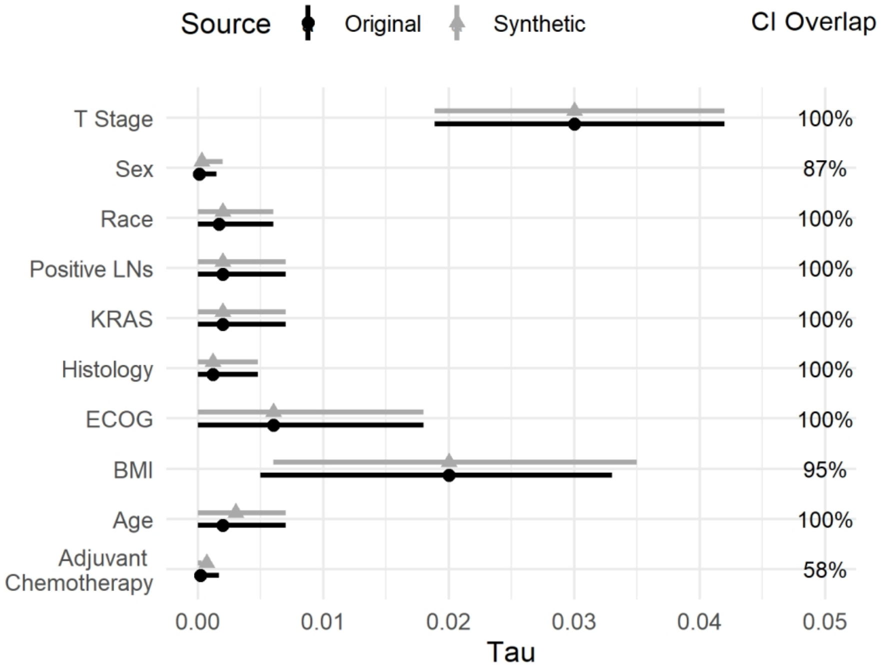


Figure 2: The tau coefficient on the real and synthetic data, and the confidence interval overlap for the bivariate relationship with obstruction.

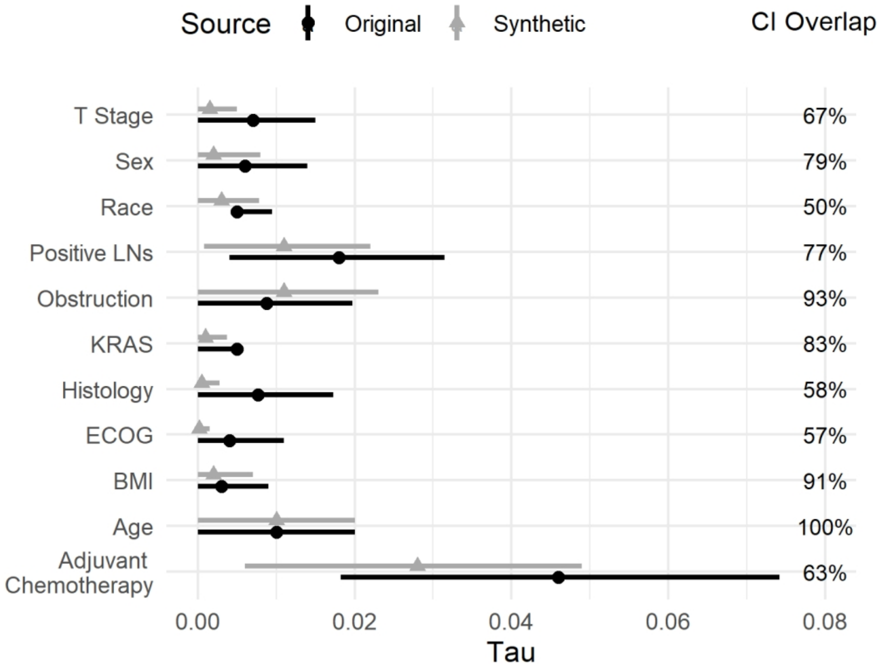


Figure 3: The effect size for the real and synthetic variables against overall survival.

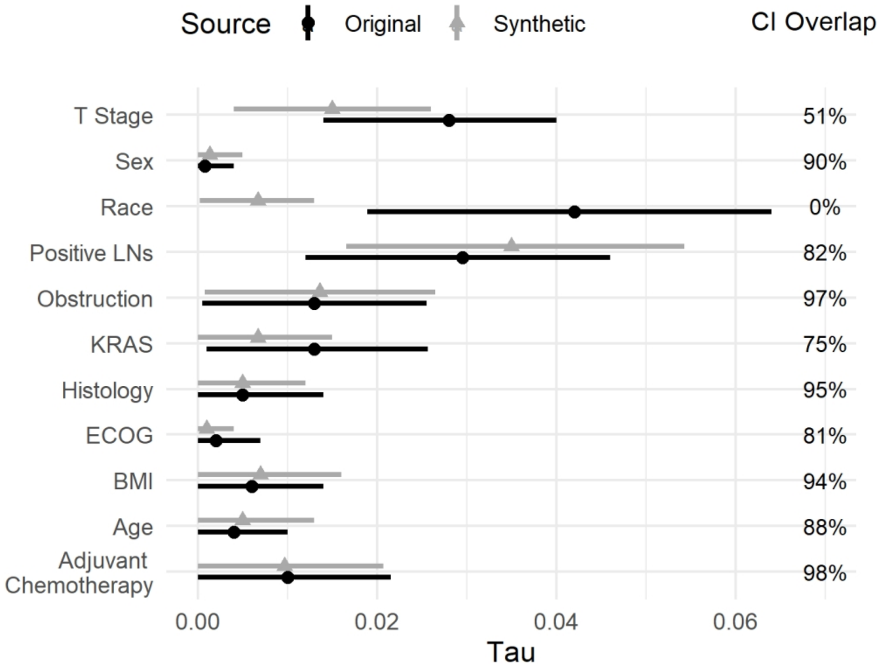


Figure 4: The effect size for the real and synthetic variables against disease-free survival.

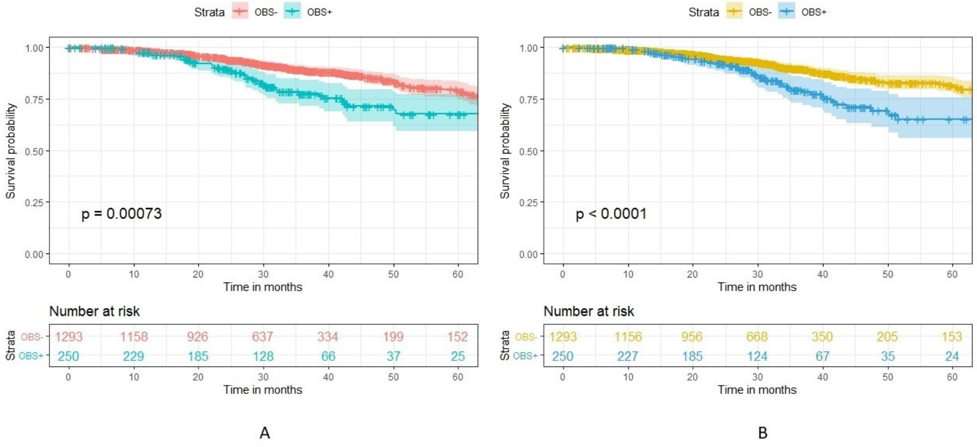


Figure 5: Survival curve comparing overall survival (OS)   in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B) .

337x165mm (96 x 96 DPI)

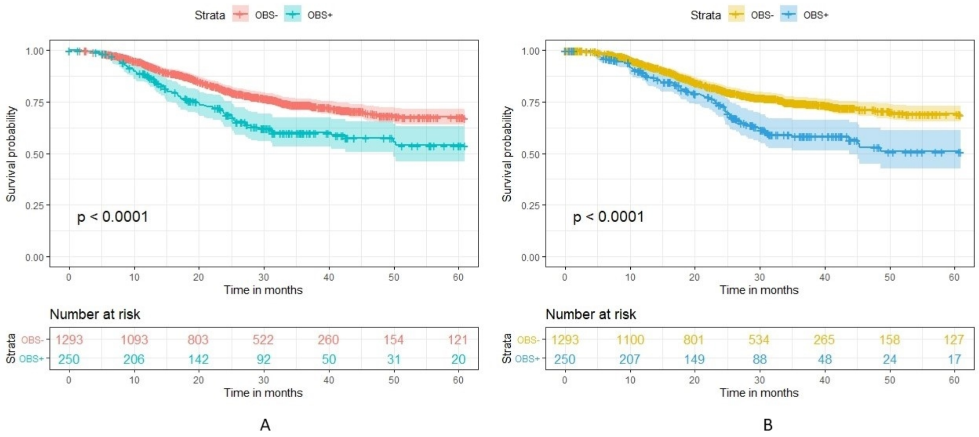


Figure 6: Survival curve comparing disease free survival (DFS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B) .

337x159mm (96 x 96 DPI)

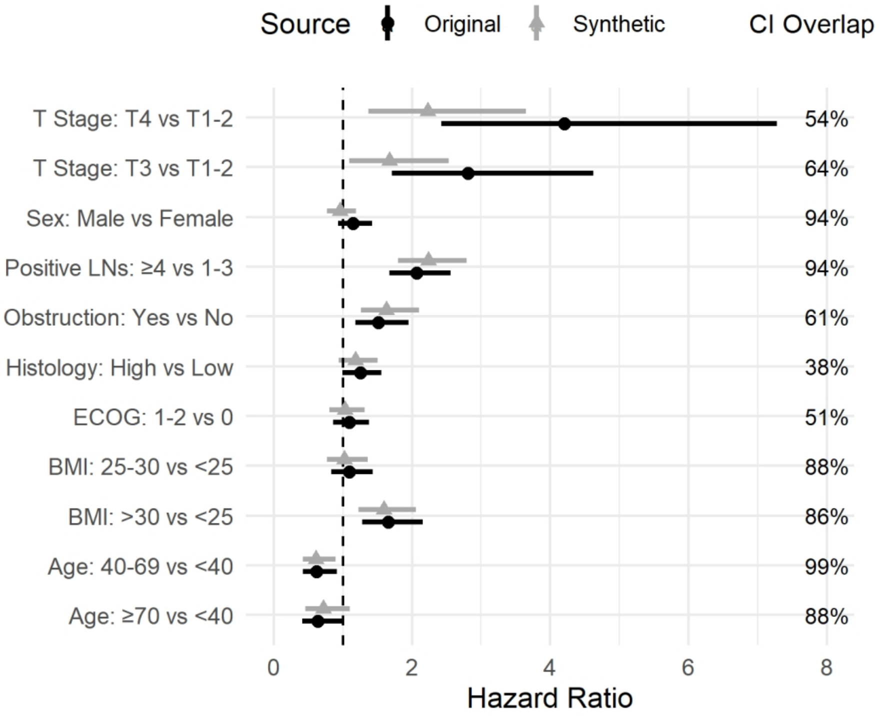


Figure 7: Comparison of real and synthetic Cox model parameters (hazard ratios) with the overall survival outcome variable for the real and synthetic datasets.

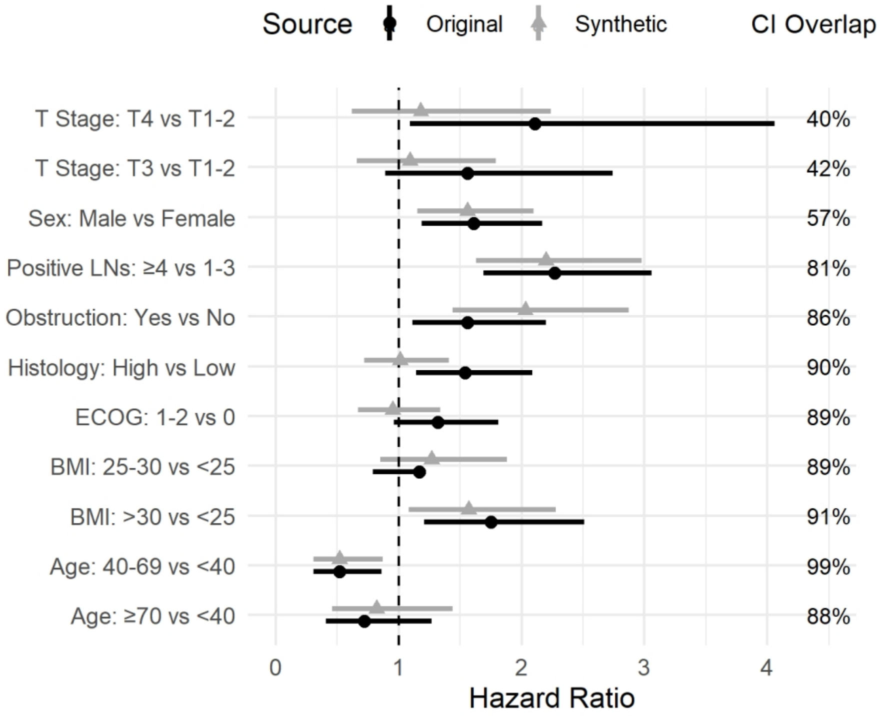


Figure 8: Comparison of real and synthetic Cox model parameters (hazard ratios) with the disease free survival outcome variable for the real and synthetic datasets.

BMJ Open

Can Synthetic Data Be A Proxy for Real Clinical Trial Data: A Validation Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-043497.R1
Article Type:	Original research
Date Submitted by the Author:	14-Jan-2021
Complete List of Authors:	Azizi, Zahra; McGill University Faculty of Medicine, Center for Outcomes Research and Evaluation Zheng, Mina; Replica Analytics Ltd., Data Science Mosquera, Lucy; Replica Analytics Ltd., Data Science Pilote, Louise; McGill University, Medicine; Research Institute of the McGill University Health Centre, Centre for Outcomes Research and Evaluation El Emam, Khaled; Children's Hospital of Eastern Ontario Research Institute, Electronic Health Information Laboratory
<b>Primary Subject Heading</b>:	Health informatics
Secondary Subject Heading:	Research methods
Keywords:	EPIDEMIOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Information management < BIOTECHNOLOGY & BIOINFORMATICS, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, STATISTICS & RESEARCH METHODS

Can Synthetic Data Be A Proxy for Real Clinical Trial Data: A Validation Study

Zahra Azizi[4], Mina Zheng[3], Lucy Mosquera[3], Louise Pilote[4,5], and Khaled El Emam[1,2,3]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[2]Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada

[3]Data Science, Replica Analytics Ltd., Ottawa, Ontario, Canada

[4] Center for Outcomes Research and Evaluation, Faculty of Medicine, McGill University, Montreal, Quebec, Canada

[5] Research Institute of the McGill University Health Centre, Faculty of Medicine, McGill University, Montreal, Quebec, Canada

Corresponding Author:

Khaled El Emam
Children's Hospital of Eastern Ontario Research Institute
401 Smyth Road, Ottawa
Ontario K1J 8L1, Canada

E: kelemam@ehealthinformation.ca

Abstract

Objectives: There are increasing requirements to make research data, especially clinical trial data, more broadly available for secondary analyses. However, data availability remains a challenge due to complex privacy requirements. This challenge can potentially be addressed using synthetic data.

Setting: Replication of a published stage 3 colon cancer trial secondary analysis using synthetic data generated by a machine learning method.

Participants: There were 1,543 patients in the control arm that were included in our analysis.

Primary and Secondary Outcome measures: Analyses from a study published on the real dataset were replicated on synthetic data to investigate the relationship between bowel obstruction and event-free survival. Information theoretic metrics were used to compare the univariate distributions between real and synthetic data. Percentage confidence interval overlap was used to assess the similarity in the size of the bivariate relationships, and similarly for the multivariate Cox models derived from the two datasets.

Results: Analysis results were similar between the real and synthetic datasets. The univariate distributions were within 1% of difference on an information theoretic metric. All of the bivariate relationships had confidence interval overlap on the tau statistic above 50%. The main conclusion from the published study, that lack of bowel obstruction has a strong impact on survival, was replicated directionally and the hazard ratio confidence interval overlap between the real and synthetic data was 61% for overall survival (HR of 1.56; 95% CI: 1.11-2.2 for real data, and HR of 2.03; 95% CI: 1.44-2.87 for synthetic data), and 86% for disease-free survival (HR of 1.51; 95% CI: 1.18-1.95 for real data, and 1.63; 95% CI: 1.26-2.1 for synthetic data).

Conclusions: The high concordance between the analytic results and conclusions from synthetic and real data suggests that synthetic data can be used as a reasonable proxy for real clinical trial datasets.

Trial Registration (original study): NCT00079274

Strengths and Limitations

- The study evaluated whether a synthetic clinical trial dataset gives similar analysis results and the same conclusions as does analysis of the real dataset
- A machine learning method was used to generate the synthetic data
- A published analysis evaluating the effect of bowel obstruction on survival of colon cancer patients was replicated
- The results and conclusions from real and synthetic data were compared in univariate, bivariate and multivariate analyses
- The identity disclosure (privacy) risks of the synthetic data were not explicitly evaluated although existing evidence in the literature suggests that it is low

1. Background

It is often difficult for researchers to get access to high quality individual-level data for secondary purposes (e.g. testing new hypotheses and building statistical and machine learning models). Specifically, for clinical trial data, secondary analysis of data from previous studies can provide new insights compared to the original publications [1], and has produced informative research results including on drug safety, evaluating bias, replication of studies, and meta-analysis [2]. Therefore, there has been strong interest in making more clinical trial data available for secondary analysis by journals, funders, the pharmaceutical industry and regulators [3]–[8].

For example, the ICMJE's data sharing policy [9] indicates that articles reporting the results of clinical trials must include a data sharing statement when they are submitted to ICMJE journals for publication. Funders also have data sharing requirements. According to the Wellcome Trust's policy [10], researchers receiving funding are expected to share their data rapidly, an outputs management plan is a requirement for any funding proposal which anticipates the generation of significant outputs (e.g., data, software or other materials). These plans are factored into funding decisions. The NIH Statement on Sharing Research Data [11] indicates that applicants seeking $500,000 or more in funding per year are required to include a data sharing plan (or explain why it is not possible to share their research data). Data shared by researchers should be individual-level data upon which the accepted publication was based.

However, data access for secondary analysis remains a challenge [12]. To highlight this challenge, an examination of the success rates of getting individual-level data for research projects from authors found that the percentage of the time these efforts were successful varied significantly and was generally low at 58% [13], 46% [14], 25% [15], 14% [16], and 0% [17].

One reason for this challenging data sharing environment is increasingly strict data protection regulations. A recent National Academy of Medicine/Government Accountability Office report highlights privacy as presenting a data access barrier for the application of Artificial Intelligence (AI) and machine learning in healthcare [18]. While patient (re-)consent is one legal basis for making data available for secondary purposes, it is often impractical to get retroactive consent under many circumstances and there is significant evidence of consent bias [19].

Anonymization is one approach to making data available for secondary analysis. However, recently there have been repeated claims of successful re-identification attacks on anonymized data [20]–[26], eroding public and regulators' trust in this approach [26]–[36]. Although, it should be noted that there are no known successful re-identification attacks on anonymized clinical trial data at the time of writing.

To provide additional options and methods for sharing the information from clinical trials, in this paper we propose using synthetic data [37]. To create synthetic data, a machine learning generative model is constructed from the real individual-level data, capturing its patterns and statistical properties. Then new data is generated from that model. This step is performed by the data controller / custodian who has access to that real data. The synthetic version of the data would then be provided to analysts to conduct their studies.

There are many use cases where synthetic data can provide a practical solution to the data access problem [38], and has been highlighted as a key privacy enhancing technology to enable data access for the coming decade [39]. Furthermore, there are recent examples of research studies using synthetic data not requiring ethics review because they are considered to contain no patient information [40]. To the extent that this becomes a common practice, it would accelerate data access.

An important question with the analysis of synthetic data is whether similar results and the same conclusions would be obtained as with the real data. To answer this question, we compared the analysis results and conclusions using real and synthetic data for a published oncology trial. Given that by far the most common purposes for the re-analysis of clinical trial data are new analyses of the treatment effect and the disease state rather than replicating the primary analysis [41], we focused on replicating a published secondary analysis rather than a primary analysis. This approach will inform us about the extent to which synthetic data can be useful for the secondary analysis of clinical trials.

There have been limited replications of clinical studies using synthetic data, with only a handful of examples in the context of observational research [42], [43] and larger clinical trial data [44]. The current study adds to this body of work and contributes to the evidence base for enabling more access to clinical trial data through synthesis.

2. Methods

2.1 Data Sources

We obtained the dataset for an oncology trial, N0147, from Project Data Sphere (PDS)[1] [45]. The specific trial was selected because the PDS data were analyzed in a published study that we could successfully replicate (validating that we have the correct data and interpreted it the same way as the authors), and the description of the analyses performed was clear enough to allow replication. In the current paper, we will refer to this PDS dataset as the "real" data since that is our source dataset for synthesis.

PDS data is already perturbed to anonymize it. The level of perturbation is dependent on the sponsor. Therefore, the use of term "real" should be interpreted to mean "real and anonymized" data.

[1] See <https://data.projectdatasphere.org/>

2.2 Summary of Trial Data

Trial N0147 was a randomized trial of 2,686 patients with stage 3 colon adenocarcinoma that were randomly assigned to adjuvant regimens with or without Cetuximab. After resection of colon cancer, Cetuximab was added to the modified sixth version of the FOLFOX regimen including oxaliplatin plus 5-fluorouracil and leucovorin (mFOLFOX6), fluorouracil, leucovorin, and irinotecan (FOLFIRI), or a hybrid regimen consisting of mFOLFOX6 followed up by FOLFIRI [46]. Our focus is on the secondary retrospective analysis of N0147 (the *published secondary analysis*) [47].

The primary endpoint in the original trial was disease-free survival (DFS), defined as time from random allocation to the first of either tumor recurrence or death from any cause. Secondary trial endpoints were time to recurrence (TTR) and overall survival (OS). TTR was measured from random allocation to tumor recurrence, whereas OS was from random allocation to death from any cause. OS was censored at 8 years, whereas DFS and TTR were censored at 5 years. Patients who died without recurrence were censored for TTR at the time of death. Patients who were lost to follow-up were censored at the date of their most recent disease assessment or contact.

Participants in the control “chemotherapy-only” arm (FOLFOX, FOLFIRI or hybrid regimen without Cetuximab) were analyzed in the published secondary analysis, which consisted of 1,543 patients. Presentation with acute obstruction of the bowel is a known risk factor for poor prognosis in patients with colon cancer [48], [49]. The main objective of this secondary analysis was to assess the role of obstruction presentation as an independent risk factor for predicting outcomes in patients with stage III colon cancer. The primary endpoint of the in the published secondary analysis was disease free survival (DFS), and the secondary endpoint was overall survival (OS), and both DFS and OS were censored at five years.

The covariates in the published secondary analysis comprised of three types of variables: 1) Baseline demographics, including age, sex, and baseline BMI, 2) Baseline Eastern cooper- active oncology group (ECOG) performance score that describes patients' level of functioning in terms of their ability to care for themselves, daily activity and physical ability, and 3) Baseline cancer characteristics, including clinical T stage, lymph node involvement, histologic status, and Kirsten rat sarcoma virus (KRAS) biomarker status.

2.3 Data Synthesis Method

The data synthesis process takes a real dataset as input, trains a generative model from it, then generates synthetic data using the model. Multiple statistical or machine learning methods can be used to create a generative model.

We used sequential decision trees for data synthesis to fit a generative model. Sequential decision trees are used quite extensively in the health and social sciences for the generation of synthetic data [50]–[58]. In these models, a variable is synthesized by using variables preceding it in the sequence as predictors. The method we used to generate synthetic data is called conditional trees [59], although other parametric or tree algorithms could also be used. Methods such as deep learning have been proposed for the synthesis of health data [60][61]. However, compared to deep learning synthesis methods, sequential decision trees have the advantage of not requiring a large input dataset that is used for training. It is therefore suitable for creating synthetic variants of clinical trial data that typically have a relatively small number of participants. More details about how sequential synthesis was applied are included in the supplementary materials.

2.4 Replication of Secondary Analysis on the Synthetic Data

We first replicated the published analysis on the real dataset. Once the results could be replicated, we re-ran the exact same analysis R code on the synthetic version of the data.

The published secondary analysis [47] included descriptive statistics consisting of frequency (percentage) for categorical variables. The Pearson χ^2 test was used to investigate the statistical significance of the relationship between the baseline characteristics (clinical, pathological) and obstruction. Survival analysis was performed using the Kaplan-Meier curve. The log rank test and Cox proportional hazards model were used to plot OS and DFS at 5 years and to create a model adjusted for baseline clinical and pathological characteristics to assess the role of obstruction in predicting OS and DFS.

2.5 Evaluation of Results

Our objective was to evaluate the utility of the synthetic data. Thus, we compared the results using the real data with those using the synthetic data. Our utility evaluation method followed the recommendations to evaluate the utility of data that has been transformed to protect privacy, such as through data synthesis [62]. Specifically, we used two general approaches to compare real and synthetic analysis results: information theoretic methods based on the Kullback-Leibler (KL) divergence, and interval overlap for the confidence intervals of model parameters. Both are described further below.

To evaluate the utility of synthetic data we compared the published univariate and the bivariate statistics on the real data and the synthetic data. The methods for the univariate comparisons are in the appendix. We then compared the multivariate model parameters for the models that were developed to explain survival and test the hypothesis that obstruction was an important predictor.

2.5.1 Bivariate Analysis

In the published secondary analysis, the bivariate results were presented as contingency tables showing the cross-tabulations of the predictors with obstruction, OS after five years, and DFS after five years. The Pearson χ^2 test was used to evaluate all bivariate relationships. This type of testing when used in the current context has a number of disadvantages: (a) it does not give us an interpretable effect size and therefore we would not know if a bivariate relationship was strong or not (a test statistic can be significant with a very small effect size if there are many observations), (b) the tests did not account for multiple-testing, such as a Bonferroni adjustment, which means that there will be an elevated probability of finding significant results by chance, and (c) the chi-square tests considers independence whereas the relationship that is being tested is whether each of the covariates are predictive of the outcome. For these reasons we used a different statistic to compute the bivariate relationships on the real and synthetic datasets.

We use the Goodman and Kruskal tau statistic, which gives us a measure between zero and one of the extent to which the covariate is predictive of the outcome [63]. The tau coefficient was computed for the real dataset and the synthetic dataset and the confidence intervals compared. Confidence interval overlap has been proposed for evaluating the utility of privacy protective data transformations [62], which is defined as the percentage average of the real and synthetic confidence intervals that overlap. Our formulation gives an overlap value of zero if the two intervals do not overlap at all. We express overlap as a percentage.

The published secondary analysis evaluated the bivariate relationship between each of the predictors and obstruction, and then evaluated each of the predictors and obstruction with event free survival. We repeated these analyses with the tau statistic and confidence intervals.

2.5.2 Multivariate Analysis

For the multivariate models, we compared the Cox model hazard ratio estimates between the real and synthetic data. We also computed the confidence interval overlap of the hazard ratios from the Cox models.

2.6 Patient and Public Involvement

The comparative analysis of synthetic to real data did not have any patient or public involvement.

3. Results for Trial N0147

We compare the results in the secondary analysis study that were published against the same analyses performed on the synthetic data. The results for the univariate analysis show little difference in distributions and are in the supplementary materials.

3.1 Bivariate Analysis

The differences between real and synthetic data for the bivariate relationships of the covariates and obstruction are shown in Figure 1. When we look at the effect sizes (the tau metric), we see that the size of these bivariate relationships is very small. These covariates individually are not good predictors of obstruction. We also note that the effect sizes are similar between the real and synthetic datasets, and there are considerable confidence intervals overlap. One would draw the same conclusions from the real and synthetic datasets.

The next set of results are also the bivariate relationships between the covariates and the event free survival outcomes: overall survival and disease-free survival. The results in Figure 2 show the effect sizes for the bivariate relationships with overall survival. There are two noteworthy observations. The first observation is that all the bivariate relationships are very weak – the covariates are not individually predictive of OS. The second observation is that the effect sizes are very similar between the real and synthetic datasets. One would draw the same conclusions from the synthetic data as from the data in the published secondary analysis.

Figure 3 shows the bivariate relationships with disease-free survival. The conclusions are like overall survival with one exception: the confidence intervals for the relationship between race and DFS do not overlap. Given the weak relationship between race and DFS, this lack of confidence interval overlap is likely due to the stochastic nature of synthesis. In addition, the relationship is quite weak in both datasets and of very similar magnitude, therefore the conclusions would still be the same in both cases.

3.2 Multivariate Analysis

For the multivariate analyses, the real data results are like those that were in the published secondary analysis. We first compare the survival curves for obstructed and non-obstructed patients on overall survival (Figure 4) and disease-free survival (Figure 5). We can see that the curves are very similar between the real and synthetic datasets.

The Cox models were intended to evaluate whether obstruction affects survival after accounting for the potential confounding effect of other covariates. The real and synthetic hazard ratio model parameters are generally in the same direction with relatively high overlap for the confidence intervals. This is the case for the overall survival model in Figure 6 and the disease-free survival model in Figure 7.

The main hypothesis being tested in the published secondary analysis pertains to obstruction. For the OS model the hazard ratio for obstruction overlap was high at 61% (HR of 1.56; 95% CI: 1.11-2.2 for real data, and HR of 2.03; 95% CI: 1.44-2.87 for synthetic data) with both models showing a strong effect of obstruction on OS (No obstruction related to higher OS). Similarly, for the DFS model, the overlap was 86% (real data HR of 1.51; 95% CI: 1.18-1.95, and the synthetic data HR of 1.63; 95% CI: 1.26-2.1), indicating that the model shows an association between obstruction and DFS. Therefore, one would draw the same conclusion about the impact of obstruction on event-free survival.

The point estimates for the T stage covariates differ the most in Figure 6 for overall survival and Figure 7 for the disease-free survival model, with lower confidence interval overlap than many of the other covariates. The same is true for histology in Figure 6. While some variation in the numeric values is expected in the synthetic data, the parameters were directionally the same, and the inclusion of these covariates did allow us to control for their effect in the assessment of obstruction, which was the main objective of the analysis.

One other observation from the overall survival model in Figure 6 and the disease-free survival model in Figure 7 is that the confidence intervals from the synthetic data are narrower than the real data. A generative model captures the patterns in the data. A plausible explanation is that the machine learning methods used during synthesis capture the signal or patterns in the data well and these are produced more clearly (or with less noise) in the synthetic data.

4. Discussion

4.1 Summary

The purpose of this study was to evaluate the extent to which a published secondary analysis of an oncology clinical trial could be replicated using a synthetic variant of the dataset. This replication is one of the first to test whether similar results and the same conclusions would be drawn from the re-analysis of a published clinical trial analysis using a synthetic version of the dataset.

The published secondary analysis was investigating the relationship between bowel obstruction and event-free survival for colon cancer patients. We applied a commonly used synthesis approach that ensured the potentially identifying variables (the quasi-identifiers) in the dataset were appropriately synthesized.

We found that for the univariate and bivariate analyses in the published study, the synthetic data was quite similar in terms of distributions and effect sizes to the real data. With respect to the multivariate models that controlled for confounders, the published results were replicated in that there was a strong positive relationship between obstruction and overall survival and disease-free survival after five years in the both the real and synthetic datasets.

4.2 Relevance and Application of Results

In addition to offering more options for addressing privacy concerns, sharing synthetic versions of clinical trial datasets can potentially alleviate the need for obtaining ethics board reviews for such analysis projects [40]; simplifying and accelerating research studies.

If the objective of a secondary analysis of a clinical trial dataset is the replication / validation of a published study, then working with a synthetic variant of the dataset will not give the exact numeric results but would be expected to produce the same conclusions as was demonstrated in our study. Another type of secondary analysis is to assess bias in trial design, misreporting or selective outcome reporting where "keeping the same conclusions and comparable numerical results of all primary, secondary and safety endpoints […] is of utmost importance." [2]. The data synthesis approach we presented here achieves these objectives by including the primary and secondary endpoints in the generative model to ensure that relationships with other covariates are maintained, and it does not synthesize adverse event data to maintain the accuracy of safety data. More generally, a review of protocols found that most secondary analysis of clinical trial datasets focused on novel analyses rather than replication or validation of results [64]. In such cases, the conclusions from using synthetic data would be expected to be the same as using the real data. However, it is more difficult to make the case for using synthetic data for the primary analysis of a clinical trial dataset since the investigators and sponsors would have ready access to the real data.

While we are already starting to see published (observational) health research using synthetic data only [40], there will be situations where there is a requirement for additional verification that the model parameters produced from synthetic data are numerically similar to the real data, and that the conclusions are the same. This step can be achieved by implementing a verification server. With such a setup synthetic data is shared, and the analysts build their models on the synthetic data. Then their analysis code (say an R or SAS program) is sent to a verification server which is operated by the data controller / custodian. The analysis code is executed on the real data, and the results returned to the analysts. The returned results would either be the model parameters on the real data or the difference in parameter values between the real data model and the synthetic data model. That way the analysts can get feedback as to the accuracy of the synthetic data model parameters without having direct access to the real data themselves. The deployment of a verification server balances the need for rapid access to data with minimal constraints with the need for ensuring model accuracy from the synthetic data. On the other hand, it does introduce an additional process step.

The need for a verification server can arise, for example, when results are going to be submitted to a regulator. Generally, in the early days of adoption of data synthesis there will likely be a greater need for verification, and one would expect that need would dissipate as successful applications of data synthesis increase over time.

This study is a replication of a single clinical trial. However, it does provide evidence that synthesized datasets can be used as a reasonable proxy for real datasets. The data synthesis method is well established and has been applied extensively in the health social sciences. Further such replications should be performed to increase the weight evidence on the effectiveness of synthetic data as a proxy for real datasets. To the extent that synthetic data would allow drawing the same conclusions as real data, they can be more readily shared by researchers when publishing their studies and to meet funding agency requirements for data sharing, and by sponsors to meet their data transparency commitments.

4.3 Limitations

The data we used in our analysis came from Project Data Sphere, which shares datasets that have already gone through a perturbation to anonymize the data. This would not affect our results or conclusions because the published study that we replicated used the same (perturbed) dataset from Project Data Sphere. More generally, synthetic data can be generated from pseudonymous data rather than from fully anonymized data. Multiple researchers have noted that synthetic data does not have an elevated identity disclosure (privacy) risk [60], [65]–[72], and therefore anonymization before synthesis is not necessary.

This study was an assessment of the ability to replicate a secondary analysis for a clinical trial dataset. It is a reasonable expectation that as more similar replications using synthetic data demonstrate equivalent results and conclusions as real data, there will be greater acceptance of synthetic derivatives as a reliable way to share clinical trial datasets. In fact, we are already starting to see published (observational) health research using synthetic derivatives only [40].

While we found that there were very little differences between the real and synthetic data on the bivariate comparisons, one may hypothesize that this was influenced by the fact that the effect sizes were small. However, that was not the case for the multivariate models where the effect sizes were larger and the differences between the real and synthetic datasets remained small.

5. Conclusions

As interest in the potential of synthetic data has been growing, an important question that remains is the extent to which similar results and the same conclusions would be obtained from the synthetic datasets compared to the real datasets. In this study we have provided one answer to that question. Our re-analysis of a published oncology clinical trial analysis demonstrated that the same conclusions can be drawn from the synthetic data. These results suggest that synthetic data can serve as a proxy for real data and would therefore make useful clinical trial data more broadly available for researchers.

Ethics

This project was approved by the CHEO Research Institute Research Ethics Board, protocol number CHEOREB# 20/75X.

Acknowledgements

The work in this paper was performed in collaboration with the GOING FWD consortium.

This article is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this article.

Data Statement

The dataset can be obtained by registering at Project Data Sphere: https://www.projectdatasphere.org/

Funding Statement

This work was partially funded by a Discovery Grant RGPIN-2016-06781 from the Natural Sciences and Engineering Research Council of Canada, by Replica Analytics Ltd., and by the GOING FWD consortium. The GOING-FWD Consortium is funded by the GENDER-NET Plus ERA-NET Initiative (Project Ref. Number: GNP-78). ZA is funded by the Canadian Institutes of Health Research (GNP-161904).

Competing Interests Statement

This work was performed in collaboration with Replica Analytics Ltd. This company is a spin-off from the Children's Hospital of Eastern Ontario Research Institute. KEE is co-founder and has equity in this company. LM and MZ are data scientists employed by Replica Analytics Ltd.

Author Contributions

AZ contributed to designing the study, performing the analysis, and writing the paper. KEE contributed to designing the study, to performing the analysis, and to writing the paper. LM contributed to designing the study, implemented some of the code used to perform the analysis, and contributed to writing the paper. MZ contributed to designing the study, implemented some of the code used to perform the analysis, contributed to the data analysis, and contributed to writing the paper. LP contributed to designing the study and to writing the paper.

6. References

[1] Ebrahim S, Sohani ZN, Montoya L, and et al, "Reanalyses of randomized clinical trial data," *JAMA*, vol. 312, no. 10, pp. 1024–1032, Sep. 2014, doi: 10.1001/jama.2014.9646.

[2] Jean-Marc Ferran and Sarah Nevitt, "European Medicines Agency Policy 0070: an exploratory review of data utility in Clinical Study Reports for research," *BMC Medical Research Methodology*, vol. 19, 2019, [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0836-3.

[3] Phrma & EFPIA, "Principles for Responsible Clinical Trial Data Sharing," Jul. 2013. [Online]. Available: http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

[4] TransCelerate Biopharma, "DE-IDENTIFICATION AND ANONYMIZATION OF INDIVIDUAL PATIENT DATA IN CLINICAL STUDIES: A Model Approach," 2017.

[5] TransCelerate Biopharma, "PROTECTION OF PERSONAL DATA IN CLINICAL DOCUMENTS – A MODEL APPROACH," Jan. 2017.

[6] European Medicines Agency, "European Medicines Agency policy on publication of data for medicinal products for human use: Policy 0070." Oct. 02, 2014, [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

[7] D. B. Taichman *et al.*, “Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal EditorsSharing Clinical Trial Data,” *Ann Intern Med*, vol. 164, no. 7, pp. 505–506, Apr. 2016, doi: 10.7326/M15-2928.

[8] Institute of Medicine, “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk,” Washington, D.C., 2015.

[9] International Committee of Medical Journal Editors, “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals,” 2019. http://www.icmje.org/icmje-recommendations.pdf (accessed Jun. 29, 2020).

[10] The Wellcome Trust, “Policy on data, software and materials management and sharing,” *Wellcome*, 2017. https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing (accessed Sep. 12, 2017).

[11] National Institutes of Health, “Final NIH Statement on Sharing Research Data,” 2003. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed Jun. 29, 2020).

[12] P. Doshi, “Data too important to share: do those who control the data control the message?,” *BMJ*, vol. 352, Mar. 2016, doi: 10.1136/bmj.i1027.

[13] J. R. Polanin, “Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing,” *Journal of Clinical Epidemiology*, vol. 98, pp. 157–159, Jun. 2018, doi: 10.1016/j.jclinepi.2017.12.014.

[14] F. Naudet *et al.*, “Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine,” *BMJ*, vol. 360, Feb. 2018, doi: 10.1136/bmj.k400.

[15] S. J. Nevitt, A. G. Marson, B. Davie, S. Reynolds, L. Williams, and C. T. Smith, “Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review,” *BMJ*, vol. 357, p. j1390, Apr. 2017, doi: 10.1136/bmj.j1390.

[16] B. Villain, A. Dechartres, P. Boyer, and P. Ravaud, “Feasibility of individual patient data meta-analyses in orthopaedic surgery,” *BMC Med*, vol. 13, no. 1, p. 131, Jun. 2015, doi: 10.1186/s12916-015-0376-6.

[17] M. Ventresca *et al.*, “Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide,” *BMC Medical Research Methodology*, vol. 20, no. 1, p. 113, May 2020, doi: 10.1186/s12874-020-00964-6.

[18] “Artificial Intelligence in Health Care,” National Academy of Medicine and the General Accountability Office, Dec. 2019.

[19] K. El Emam, E. Jonker, E. Moher, and L. Arbuckle, “A Review of Evidence on Consent Bias in Research,” *American Journal of Bioethics*, vol. 13, no. 4, pp. 42–44, 2013.

[20] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific Reports*, vol. 3, Mar. 2013, doi: 10.1038/srep01376.

[21] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. “Sandy” Pentland, “Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata,” *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297.

[22] L. Sweeney, J. Su Yoo, L. Perovich, K. E. Boronow, P. Brown, and J. Green Brody, “Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study,” *Journal of*

Technology Science, no. 2017082801, Aug. 2017, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2017082801/.

[23] J. Su Yoo, A. Thaler, L. Sweeney, and J. Zang, "Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data," *Journal of Technology Science*, no. 2018100901, Oct. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018100901/.

[24] L. Sweeney, "Matching Known Patients to Health Records in Washington State Data," Harvard University. Data Privacy Lab, 2013.

[25] L. Sweeney, M. von Loewenfeldt, and M. Perry, "Saying it's Anonymous Doesn't Make It So: Re-identifications of 'anonymized' law school data," *Journal of Technology Science*, no. 2018111301, Nov. 2018, Accessed: Mar. 23, 2020. [Online]. Available: https://techscience.org/a/2018111301/.

[26] A. Zewe, "Imperiled information: Students find website data leaks pose greater risks than most people realize," *Harvard John A. Paulson School of Engineering and Applied Sciences*, Jan. 17, 2020. https://www.seas.harvard.edu/news/2020/01/imperiled-information (accessed Mar. 23, 2020).

[27] K. Bode, "Researchers Find 'Anonymized' Data Is Even Less Anonymous Than We Thought," *Motherboard: Tech by Vice*, Feb. 03, 2020. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought (accessed May 11, 2020).

[28] E. Clemons, "Online Profiling and Invasion of Privacy: The Myth of Anonymization," *HuffPost*, Feb. 20, 2013.

[29] C. Jee, "You're very easy to track down, even when your data has been anonymized," *MIT Technology Review*, Jul. 23, 2019. https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/ (accessed May 11, 2020).

[30] G. Kolata, "Your Data Were 'Anonymized'? These Scientists Can Still Identify You," *The New York Times*, Jul. 23, 2019.

[31] N. Lomas, "Researchers spotlight the lie of 'anonymous' data," *TechCrunch*, Jul. 24, 2019. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/ (accessed May 11, 2020).

[32] S. Mitchell, "Study finds HIPAA protected data still at risks," *Harvard Gazette*, Mar. 08, 2019. https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/ (accessed May 11, 2020).

[33] S. A. Thompson and C. Warzel, "Twelve Million Phones, One Dataset, Zero Privacy," *The New York Times*, Dec. 19, 2019.

[34] "'Anonymised' data can never be totally anonymous, says study," *the Guardian*, Jul. 23, 2019.

[35] Alex van der Wolk, "The (Im)Possibilities of Scientific Research Under the GDPR," *Cybersecurity Law Report*, Jun. 17, 2020.

[36] S. Ghafur, J. V. Dael, M. Leis, A. Darzi, and A. Sheikh, "Public perceptions on data sharing: key insights from the UK and the USA," *The Lancet Digital Health*, vol. 0, no. 0, Jul. 2020, doi: 10.1016/S2589-7500(20)30161-8.

[37] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly, 2020.

[38] Khaled E Emam and Richard Hoptroff, “The synthetic data paradigm for using and sharing data,” *Cutter Executive Update*, vol. 19, no. 6, 2019.

[39] Jules Polonetsky and Elizabeth Renieris, “10 Privacy Risks and 10 Privacy Technologies to Watch in the Next Decade,” Future of Privacy Forum, Jan. 2020.

[40] A. Guo, R. E. Foraker, R. M. MacGregor, F. M. Masood, B. P. Cupps, and M. K. Pasque, “The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation,” *Front. Digit. Health*, vol. 2, 2020, doi: 10.3389/fdgth.2020.576945.

[41] A. M. Navar, M. J. Pencina, J. A. Rymer, D. M. Louzao, and E. D. Peterson, “Use of Open Access Platforms for Clinical Trial Data,” *JAMA*, vol. 315, no. 12, p. 1283, Mar. 2016, doi: 10.1001/jama.2016.2374.

[42] A. R. Benaim *et al.*, “Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies,” *JMIR Medical Informatics*, vol. 8, no. 2, p. e16492, 2020, doi: 10.2196/16492.

[43] R. E. Foraker *et al.*, “Spot the difference: comparing results of analyses from real patient data and synthetic derivatives,” *JAMIA Open*, no. ooaa060, Dec. 2020, doi: 10.1093/jamiaopen/ooaa060.

[44] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, and C. S. Greene, “Privacy-preserving generative deep neural networks support clinical data sharing,” *bioRxiv*, p. 159756, Jul. 2017, doi: 10.1101/159756.

[45] CEO Life Sciences Consortium, “Share, Integrate & Analyze Cancer Research Data | Project Data Sphere.” https://projectdatasphere.org/projectdatasphere/html/home.

[46] S. R. Alberts *et al.*, “Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial,” *JAMA*, vol. 307, no. 13, pp. 1383–1393, Apr. 2012, doi: 10.1001/jama.2012.385.

[47] F. S. Dahdaleh *et al.*, “Obstruction predicts worse long-term outcomes in stage III colon cancer: A secondary analysis of the N0147 trial,” *Surgery*, vol. 164, no. 6, pp. 1223–1229, 2018, doi: 10.1016/j.surg.2018.06.044.

[48] P. G. Carraro, M. Segala, B. M. Cesana, and G. Tiberio, “Obstructing colonic cancer: failure and survival patterns over a ten-year follow-up after one-stage curative surgery,” *Dis. Colon Rectum*, vol. 44, no. 2, pp. 243–250, Feb. 2001, doi: 10.1007/BF02234300.

[49] J. Mella, A. Biffin, A. G. Radcliffe, J. D. Stamatakis, and R. J. Steele, “Population-based audit of colorectal cancer management in two UK health regions. Colorectal Cancer Working Group, Royal College of Surgeons of England Clinical Epidemiology and Audit Unit,” *Br J Surg*, vol. 84, no. 12, pp. 1731–1736, Dec. 1997.

[50] J. Drechsler and J. P. Reiter, “An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets,” *Computational Statistics & Data Analysis*, vol. 55, no. 12, pp. 3232–3243, Dec. 2011, doi: 10.1016/j.csda.2011.06.006.

[51] R. C. Arslan, K. M. Schilling, T. M. Gerlach, and L. Penke, “Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior,” *J Pers Soc Psychol*, Aug. 2018, doi: 10.1037/pspp0000208.

[52] D. Bonnéry *et al.*, “The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-Agency Longitudinal Data,” *Journal of Research on Educational Effectiveness*, vol. 12, no. 4, pp. 616–647, Oct. 2019, doi: 10.1080/19345747.2019.1631421.

[53] A. Sabay, L. Harris, V. Bejugama, and K. Jaceldo-Siegl, “Overcoming Small Data Limitations in Heart Disease Prediction by Using Surrogate Data,” *SMU Data Science Review*, vol. 1, no. 3, Aug. 2018, [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/12.

[54] Michael Freiman, Amy Lauger, and Jerome Reiter, “Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau,” US Census Bureau, Working paper, 2017.

[55] B. Nowok, “Utility of synthetic microdata generated using tree-based methods,” 2015.

[56] G. M. Raab, B. Nowok, and C. Dibben, “Practical Data Synthesis for Large Samples,” *1*, vol. 7, no. 3, pp. 67–97, 2016, doi: 10.29012/jpc.v7i3.407.

[57] B. Nowok, G. M. Raab, and C. Dibben, “Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1,” *Statistical Journal of the IAOS*, vol. 33, no. 3, pp. 785–796, Jan. 2017, doi: 10.3233/SJI-150153.

[58] D. S. Quintana, “A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation,” *eLife*, vol. 9, 2020, doi: 10.7554/eLife.53275.

[59] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased Recursive Partitioning: A Conditional Inference Framework,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Sep. 2006, doi: 10.1198/106186006X133933.

[60] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, “Data synthesis based on generative adversarial networks,” *Proc. VLDB Endow.*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018, doi: 10.14778/3231751.3231757.

[61] K. Chin-Cheong, T. Sutter, and J. E. Vogt, “Generation of Heterogeneous Synthetic Electronic Health Records using GANs,” presented at the Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Dec. 2019, doi: 10.3929/ethz-b-000392473.

[62] A. Karr, C. Koonen, A. Oganian, J. Reiter, and A. Sanil, “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality: The American Statistician: Vol 60, No 3,” *The American Statistician*, vol. 60, no. 3, pp. 224–232, 2006.

[63] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, New Jersey: Wiley, 2002.

[64] A. M. Navar, M. J. Pencina, J. A. Rymer, D. M. Louzao, and E. D. Peterson, “Use of Open Access Platforms for Clinical Trial Data,” *JAMA*, vol. 315, no. 12, p. 1283, Mar. 2016, doi: 10.1001/jama.2016.2374.

[65] J. P. Reiter, “New Approaches to Data Dissemination: A Glimpse into the Future (?),” *CHANCE*, vol. 17, no. 3, pp. 11–15, Jun. 2004, doi: 10.1080/09332480.2004.10554907.

[66] J. Hu, “Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data,” *arXiv:1804.02784 [stat]*, Apr. 2018, Accessed: Mar. 15, 2019. [Online]. Available: http://arxiv.org/abs/1804.02784.

[67] J. Taub, M. Elliot, M. Pampaka, and D. Smith, “Differential Correct Attribution Probability for Synthetic Data: An Exploration,” in *Privacy in Statistical Databases*, 2018, pp. 122–137.

[68] J. Hu, J. P. Reiter, and Q. Wang, “Disclosure Risk Evaluation for Fully Synthetic Categorical Data,” in *Privacy in Statistical Databases*, 2014, pp. 185–199.

[69] L. Wei and J. P. Reiter, “Releasing synthetic magnitude microdata constrained to fixed marginal totals,” *Statistical Journal of the IAOS*, vol. 32, no. 1, pp. 93–108, Jan. 2016, doi: 10.3233/SJI-160959.

[70] N. Ruiz, K. Muralidhar, and J. Domingo-Ferrer, “On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective,” in *Privacy in Statistical Databases*, 2018, pp. 59–74.

[71] J. P. Reiter, “Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 168, no. 1, pp. 185–205, 2005, doi: 10.1111/j.1467-985X.2004.00343.x.

[72] Khaled E Emam, Lucy Mosquera, and Jason Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” *JMIR*, vol. 22, no. 11, Nov. 2020, Accessed: Oct. 13, 2020. [Online]. Available: https://www.jmir.org/2020/11/e23139.

Figures

Figure 1: The tau coefficient for the real and synthetic data, and the confidence interval overlap for the bivariate relationship with obstruction.

Figure 2: The tau coefficient and confidence interval overlap for the real and synthetic variables against overall survival.

Figure 3: The tau coefficient and confidence interval overlap for the real and synthetic variables against disease-free survival.

Figure 4: Survival curve comparing overall survival (OS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B).

Figure 5: Survival curve comparing disease free survival (DFS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B).

Figure 6: Comparison of real and synthetic Cox model parameters (hazard ratios) with the overall survival outcome variable.

Figure 7: Comparison of real and synthetic Cox model parameters (hazard ratios) with the disease-free survival outcome variable.

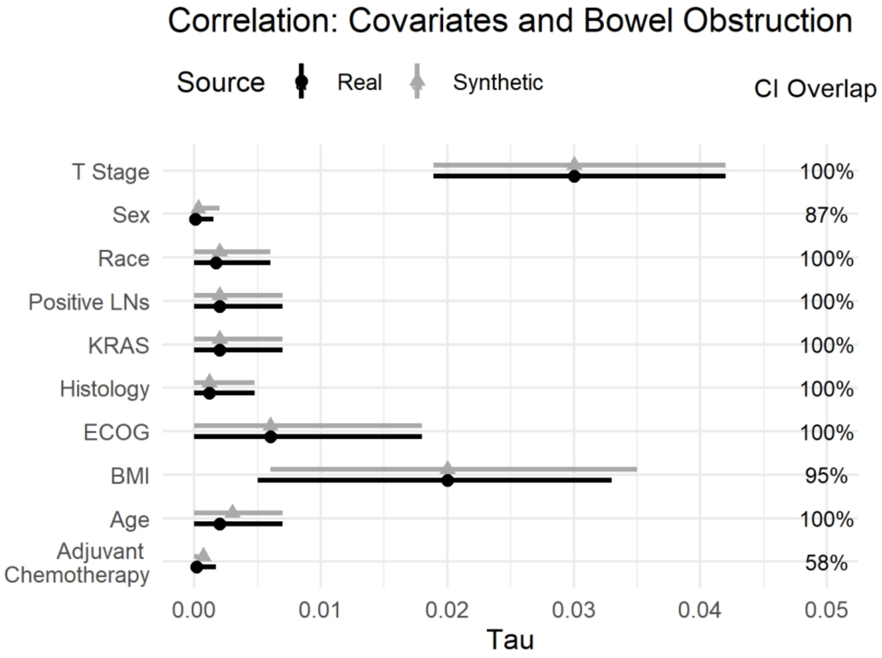


Figure 1: The tau coefficient for the real and synthetic data, and the confidence interval overlap for the bivariate relationship with obstruction.

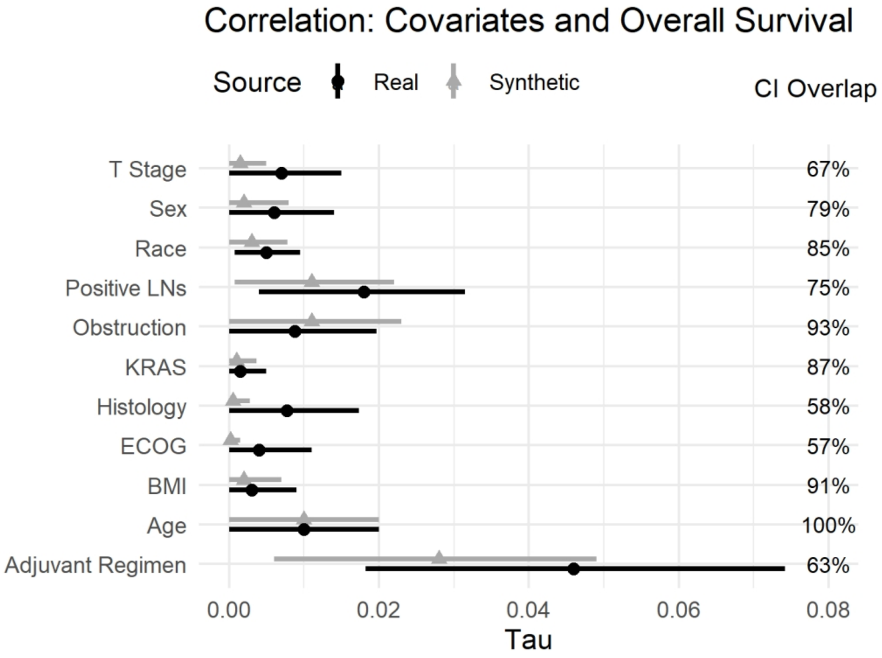


Figure 2: The tau coefficient and confidence interval overlap for the real and synthetic variables against overall survival.

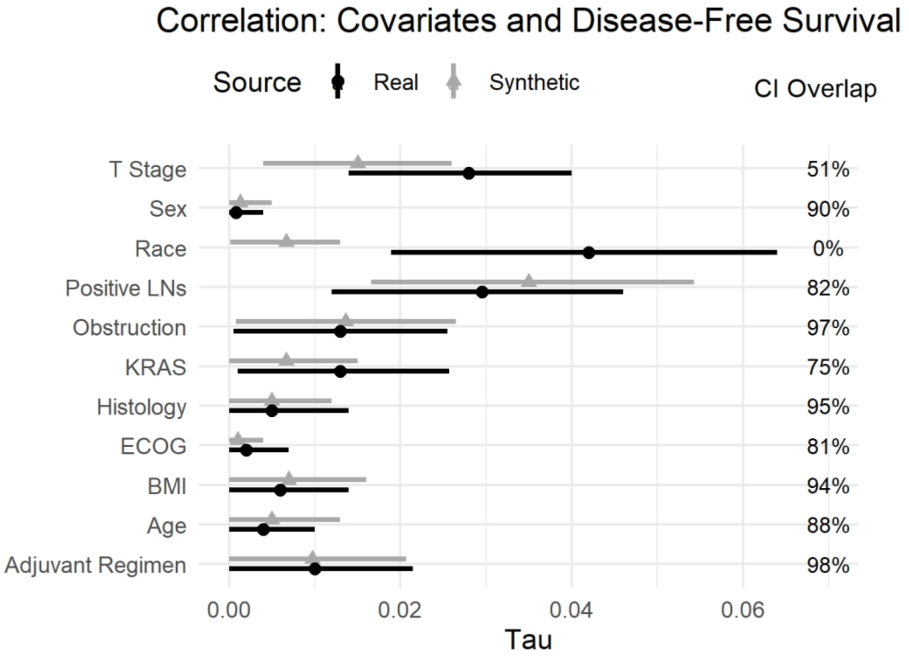


Figure 3: The tau coefficient and confidence interval overlap for the real and synthetic variables against disease-free survival.

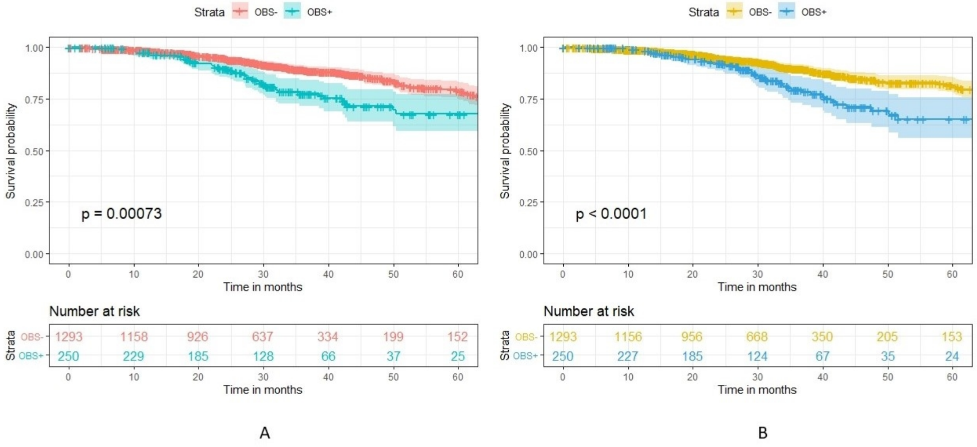


Figure 4: Survival curve comparing overall survival (OS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B).

337x165mm (96 x 96 DPI)

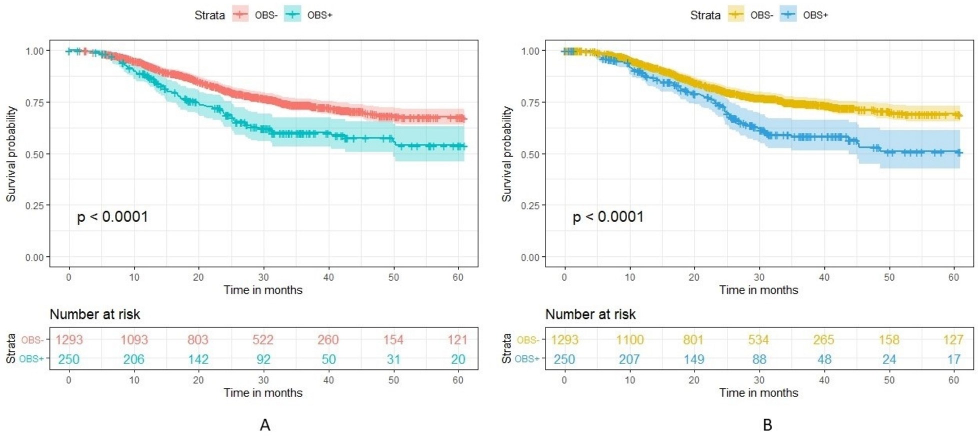


Figure 5: Survival curve comparing disease free survival (DFS) in obstructed (OBS+) and non obstructed (OBS-) patients in the real (A) versus synthetic dataset (B).

337x159mm (96 x 96 DPI)

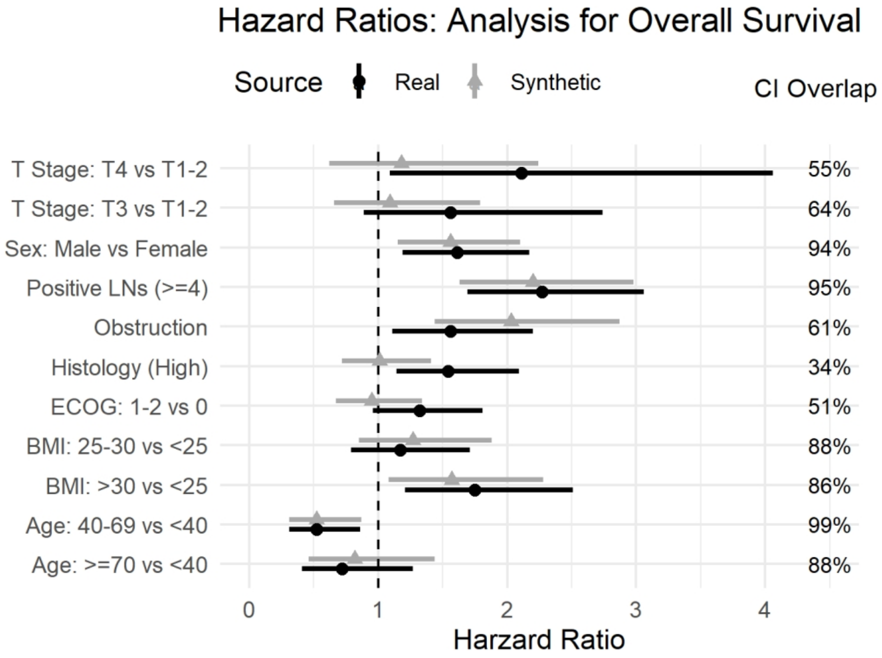


Figure 6: Comparison of real and synthetic Cox model parameters (hazard ratios) with the overall survival outcome variable.

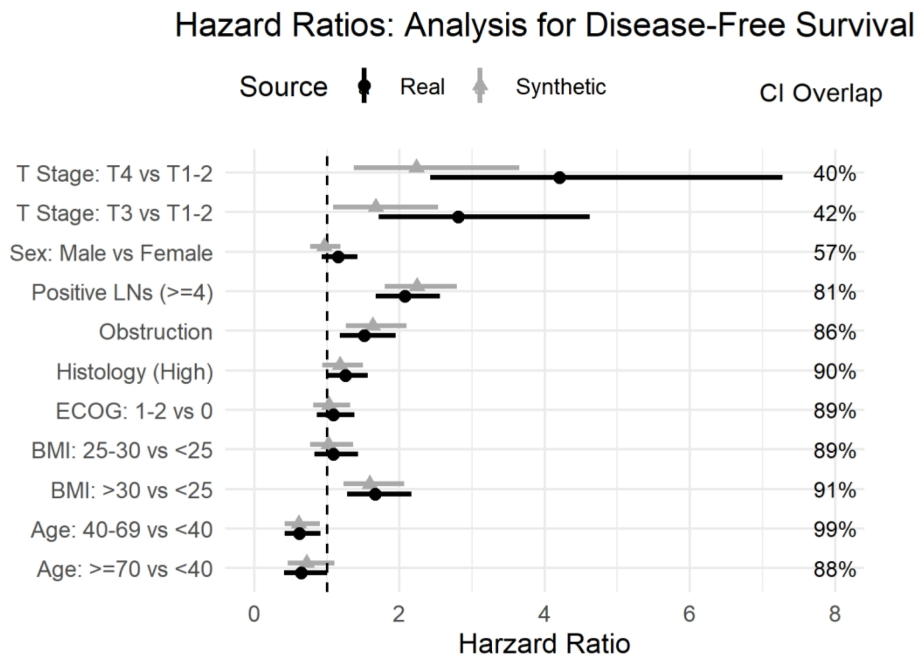
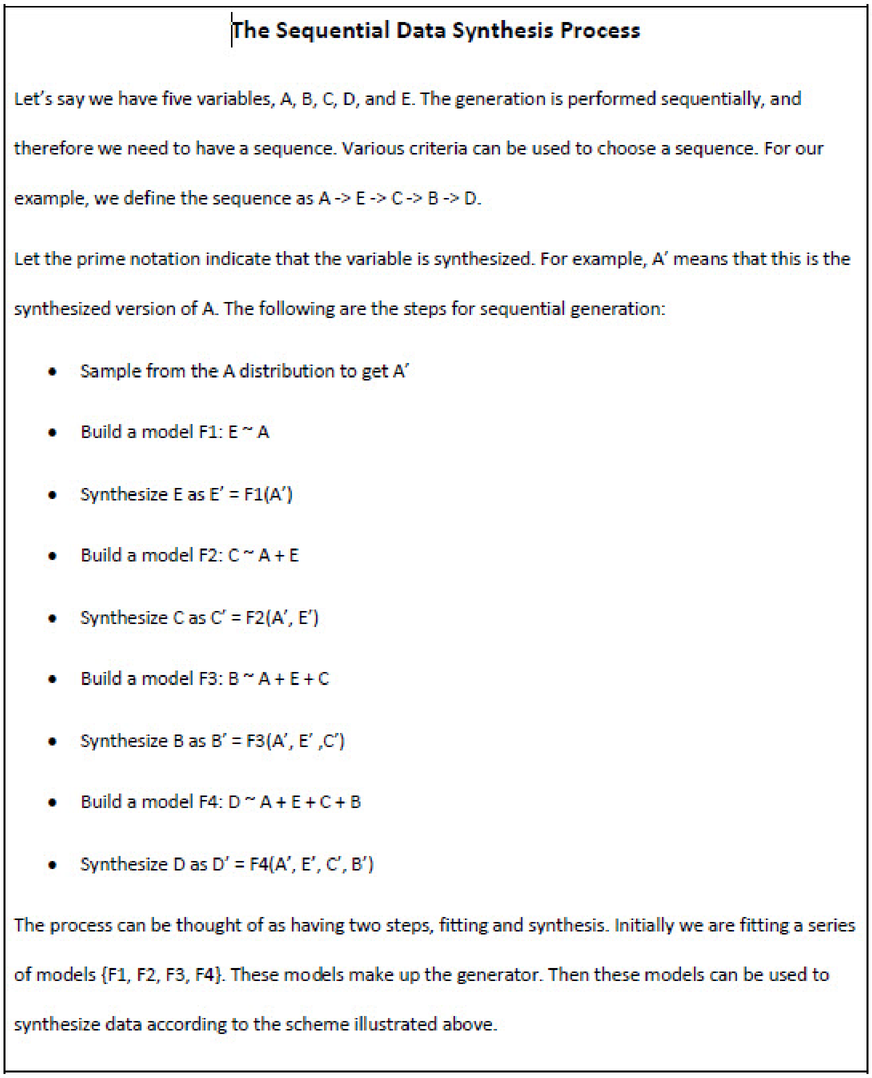


Figure 7: Comparison of real and synthetic Cox model parameters (hazard ratios) with the disease-free survival outcome variable.

Appendix A: Additional Details on Synthesis Method

A. Sequential Synthesis Method

A description of the general sequential synthesis algorithm is provided in Figure 1 [1].



The Sequential Data Synthesis Process

Let's say we have five variables, A, B, C, D, and E. The generation is performed sequentially, and therefore we need to have a sequence. Various criteria can be used to choose a sequence. For our example, we define the sequence as A -> E -> C -> B -> D.

Let the prime notation indicate that the variable is synthesized. For example, A' means that this is the synthesized version of A. The following are the steps for sequential generation:

- Sample from the A distribution to get A'
- Build a model F1: E ~ A
- Synthesize E as E' = F1(A')
- Build a model F2: C ~ A + E
- Synthesize C as C' = F2(A', E')
- Build a model F3: B ~ A + E + C
- Synthesize B as B' = F3(A', E' ,C')
- Build a model F4: D ~ A + E + C + B
- Synthesize D as D' = F4(A', E', C', B')

The process can be thought of as having two steps, fitting and synthesis. Initially we are fitting a series of models {F1, F2, F3, F4}. These models make up the generator. Then these models can be used to synthesize data according to the scheme illustrated above.

Figure 1: A description of the sequential data synthesis process using classification and regression trees. Although any set of classification and regression methods can be used.

In this study partial synthesis was performed on the trial dataset. The partial synthesis ensured that potentially identifying information in the dataset (called the quasi-identifiers [2]) were synthesized. Quasi-identifiers are the variables that are knowable by an adversary and can be used for re-identification attacks [2]–[4]. Such information is knowable because it is in the public domain (e.g., in obituaries or registries, such as voter registration lists), or is known by an adversary who is an acquaintance of someone in the dataset (e.g., a neighbor or relative). Only synthesizing the quasi-identifiers to protect against identification risks is consistent with the clinical trial data anonymization guidelines from the European Medicines Agency [5] and Health Canada [6].

The process for partial synthesis is illustrated in Figure 2. In this example we have four variables overall and two of them are quasi-identifiers (Q1 and Q2). During the fitting step two models are built in sequence for each of the quasi-identifiers. Then the non-quasi-identifiers are used as inputs during the synthesis process to generate the synthetic quasi-identifiers. This process ensures that relationships between the synthetic quasi-identifiers and the other non-synthesized (input) variables are maintained. The input data into this process is tabular, and so is the synthetic version that is produced.

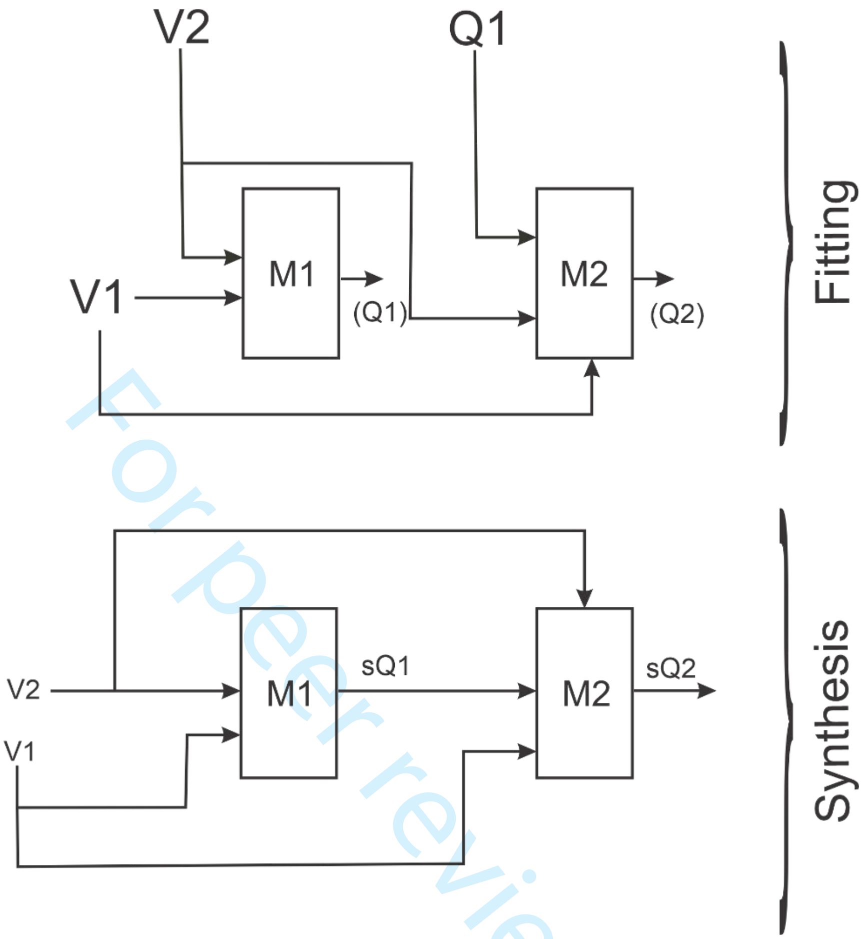


Figure 2: A schematic diagram illustrating the partial synthesis process where only the quasi-identifiers are synthesized. The Q1 and Q2 variables are the original quasi-identifier values, and sQ1 and sQ2 are the synthesized quasi-identifier values.

Appendix B: Definition of Quasi-identifiers for Clinical Trial Data

B. Introduction

The conclusion of this analysis is that a certain type of quasi-identifier contributes the most to the risk of identifying individuals: the public quasi-identifiers. Therefore, the public quasi-identifiers should be synthesized.

Public quasi-identifiers are the information that would be publicly known by an adversary about individuals. This information includes the demographics and socio-economic status variables. One exception to that rule is death information due to a serious adverse event (SAE). Data protection methods when applied to clinical trial datasets should not perturb SAE counts as that can affect the interpretation of the results.

In the following we will use the term “depersonalized” to apply to a dataset that has been transformed to protect patient privacy using any of a number of privacy enhancing techniques, such as data synthesis.

B.1 Definitions

To start off with we will provide some definitions.

B.1.1 Population, Real and Depersonalized Samples

To make this analysis more general, we will refer to the original dataset as the “Original Sample” and the depersonalized version of that dataset as the “Depersonalized Sample”. The depersonalized sample can be created using data synthesis, for example.

There is also the concept of the population. The Original Sample is assumed to be drawn from a population. We will discuss how the population is defined further below. The relationships between these datasets are illustrated in Figure 3.

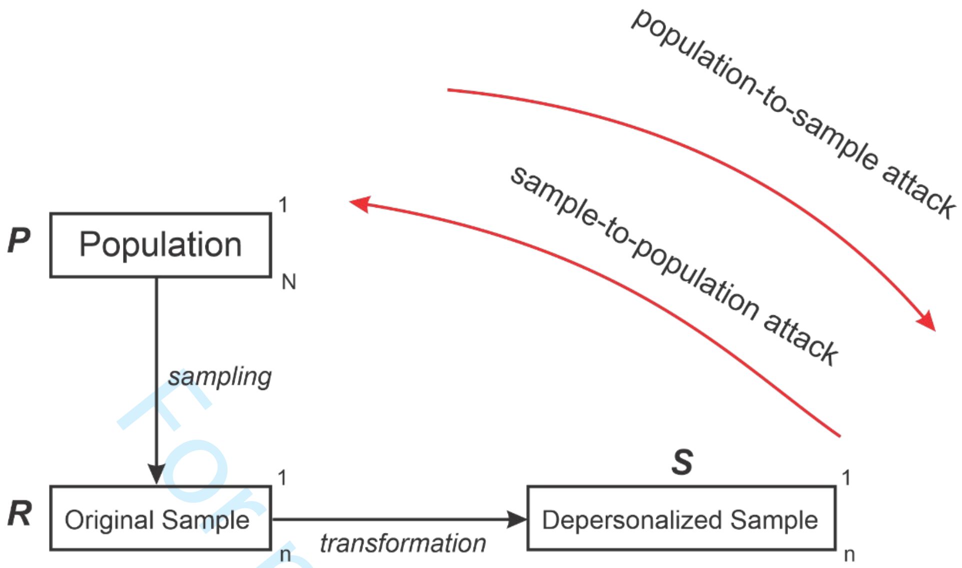


Figure 3: Directions of attack. The population has N records and the samples have n records.

B.1.2 Directions of Attack

When an adversary attacks a depersonalized sample, she can do so in one of two directions [7], [8]. The direction for these types of attacks is illustrated in Figure 3.

The adversary can start from an acquaintance. This may be, for example, a relative, a co-worker, or a neighbor. The adversary will have some background information about that acquaintance and then use that background information to find the record that matches. This is called a *population-to-sample attack*.

An adversary can also start from the depersonalized sample and select a record to match against someone in the population. The adversary will need a population registry to match against.[1] This would be an identified database[2] of people in the population that the same quasi-identifiers as in the depersonalized sample. Typically these registries exist, such as voter registration lists [9]. Or the adversary can try to construct one from, say, social media [10]. This is called a *sample-to-population attack*.

B.1.3 Quasi-identifiers

The adversary matches an acquaintance with depersonalized sample records or matches a depersonalized sample record with population registry. In both cases the matches are performed using the quasi-identifiers. The quasi-identifiers are variables that are present in the depersonalized sample record and also either known to the adversary or that are in the population registry.

There are two types of quasi-identifiers: (a) acquaintance quasi-identifiers, and (b) public quasi-identifiers. Public quasi-identifiers are a subset of acquaintance quasi-identifiers.

[1] In practice, population registries may be incomplete (such as the voter registration list). However, for the purposes of our analysis we will make the conservative assumption that a complete population registry exists.

[2] An identified database is one that has identities of individuals in it, such as a name, address, and possibly SIN.

Acquaintance quasi-identifiers are those known about an acquaintance. An adversary can know many things about an acquaintance, including their medical history. Public quasi-identifiers are those typically included in a population registry. This is the kind of information that exists in voter registration lists and that *many* people share about themselves on public social media. It also includes what can be inferred from public information. For example, income can be inferred via a person's ZIP code, which is easily obtainable for most people. Public quasi-identifiers are typically demographic and socio-economic information, as well as major events (e.g., births and deaths).

A population-to-sample attack is potentially more potent because the adversary will have more quasi-identifiers to work with. However, sampling can be protective in this case because the attacker's acquaintance may or may not be in the depersonalized sample.

B.2 Risk Model for the Identification of Clinical Trial Participants

In this section we will formulate the basic risk model specific to clinical trials and illustrate it under different assumptions. Our main conclusion is that the baseline risk of identification of participants in clinical trials under the population-to-sample attack is lower than generally accepted thresholds. Therefore, the focus should be on managing sample-to-population attacks, where the relevant quasi-identifiers are public quasi-identifiers. These are the quasi-identifiers that we synthesized in our study.

Without loss of generality, in this analysis we focus on identification risk to participants in the US. One main reason is that there is much more data that can be used by us to conduct meaningful risk analyses from the US. The US population that we use is 330 million. We also assume the attacker performs exact matching when performing attacks in either direction.

B.2.1 Evaluation of Population-to-sample Risk

We show that the population-to-sample risk is low under different assumptions about the adversary knowledge. We examine three assumptions: (a) the adversary knows that a target individual participated in an industry sponsored trial, but not knowing which one, (b) the adversary knows which specific trial the target individual participated in, and (c) the adversary knows the disease being studied for the trial that the target individual participated in, but not the specific trial or sponsor of the trial.

B.2.1.1 Background

An adversary needs to know the population from which the target individual was selected from. For example, if all that the adversary knows is that the target individual participated in a clinical trial then the relevant population is all individuals who participated in a trial. We will make a series of different assumptions about what the adversary knows and evaluate the risk of a successful population-to-sample attack. For our purposes we will always assume that the adversary knows that the target individual has participated in a clinical trial.

Based on a commonly used methodology for evaluating risk, there are three attacks that can occur on a dataset [7], [8]. The first is a deliberate attack on the clinical trial dataset by the adversary. The overall risk of matching the acquaintance with a depersonalized sample record under a deliberate attack model can be expressed as:

$$pr\left(b \mid attempt, a\right) \times pr\left(attempt \mid a\right) \times pr\left(a\right) \qquad (1)$$

where $pr(a)$ is the probability that the adversary knows someone in the population (has an acquaintance), $pr(attempt \mid a)$ is the probability of the adversary attempting to identify a record in the dataset given that the adversary knows someone in the population, and $pr(b \mid attempt, a)$ is the probability of identification given that the adversary will attempt to identify a record and knows someone in the population. For our purposes we make the conservative assumption that the probability of an adversary attempting an attack on the data is one: $pr(attempt \mid a) = 1$. Therefore, we can simplify the model to:

$$pr(b \mid a) \times pr(a) \tag{2}$$

The second attack is when an adversary inadvertently identifies a record while working with a dataset. Under worst case assumptions, that risk is the same as in equation (2).

The third attack is when a breach occurs. This is modeled as:

$$pr(b \mid breach, a) \times pr(breach \mid a) \times pr(a) \tag{3}$$

The probability of a breach occurring will by definition be some number smaller than one.

The maximum of the probabilities from these three attacks is taken as the overall risk of identification [7], [8], which will be the same as equation (2). This equation gives us the probability of a successful population-to-sample attack.

This number needs to be smaller than the commonly used risk threshold of 0.09 by the European Medicines Agency (EMA) [5] and Health Canada [6] for a dataset to be considered to have a low risk of identification.

Furthermore, the value for $pr(a)$ can be expressed as [7]:

$$pr(a) = 1 - (1 - v)^{150} \tag{4}$$

where v is the prevalence in the population that we are looking at, and 150 is the Dunbar number (see the literature review in [7]), which is the average number of “friends” or acquaintances that a person has. For example, if the population is all individuals who participated in clinical trials in the US, then v is the proportion of individuals in the US who have participated in clinical trials.

An additional parameter that we will need is the probability of identification under a population-to-sample attack [7] (also see the appendix of that reference for the derivation):

$$pr(b \mid a) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{f_i} \tag{5}$$

where f_i is the size of the equivalence class in the depersonalized sample that record i is in. An equivalence class is the group of records in the depersonalized sample that have the same values on the quasi-identifiers. For example, if the quasi-identifiers are age and sex, then an equivalence class would be “50-year-old males”, and for a depersonalized sample record i that has these values on age and sex, f_i is the number of 50 year old males in the depersonalized sample.

We now need to define the population, which will depend on the assumption we make about what the adversary knows. We will consider examine three different assumptions.

B.2.1.2 Assumption 1: Industry Sponsored Trial

Here we focus on clinical trials that lead to approvals of FDA-regulated products, and therefore we limit our analysis to clinical trials sponsored by the life sciences industry. We do not consider investigator-initiated trials or those funded by governments and foundations under this assumption.

If we start off by asking what is the probability that an adversary would identify a target individual who has participated in an FDA-regulated trial, then we are interested in calculating the probability that an adversary would know someone who participated in an FDA-regulated trial which resulted in an approved product.

The 2019 snapshot from the FDA noted that there 46,391 individuals who participated in clinical trials submitted as part of approved New Molecular Entities (NMEs) and original biologics.[3] NMEs and original biologics are medications made of new molecular structures that have not been approved by the FDA before.

Only 40% of these patients were recruited in the US (see the snapshot report). These trials would have been ongoing for multiple years before approval.

Therefore, for this population we have $v = \frac{(46,391 \times 0.4)}{330,000,000} = 0.00005623$. If we use this in equation (4) , then the probability of an adversary knowing someone who has participated in one of these trials in the US is 0.008.

However, only considering NMEs ignores clinical trials for line extensions. It has been estimated that industry’s R&D direct costs for these post-NME approvals is one fourth than that of an NME [12]. If we extrapolate that to the number of participants, then we can recompute the risk value above as $v = \frac{(46,391 \times 0.4 \times 1.25)}{330,000,000} = 0.00007028$. If we use this in equation (4), then the probability of an adversary knowing someone who has participated in one of these trials in the US is 0.01.

If we sum the FDA snapshot numbers from 2015 to 2019 inclusive, then the overall baseline risk is $pr(a) = 0.0553$. This assumes that a dataset was for a trial which was part of an FDA approval over that five year period.

This means that if a life sciences company is sharing a clinical trial dataset for a trial that was submitted as part of an FDA approval from 2015-2019, the estimated probability of an adversary knowing a participant is 0.0553.

[3] See <https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots>

This number is already lower than the commonly used risk threshold of 0.09. From equation (1), irrespective of the value of the $pr(b|a)$ in equation (5), the overall population-to-sample risk will already be below the threshold.

B.2.1.3 Assumption 2: Specific Trial

The above is an industry-wide estimate. We can also compute the probability of successful identification for a specific trial as well since risk assessments are performed on a per trial basis. In this case we are assuming that the adversary knows that the target individual has participated in a specific trial.

If we have a study with 5,500 participants, the estimated probability of an adversary knowing someone who has participated in that trial (in the US) from equation (4) is 0.0025. This estimate is shown in Figure 4 for different trial sizes.

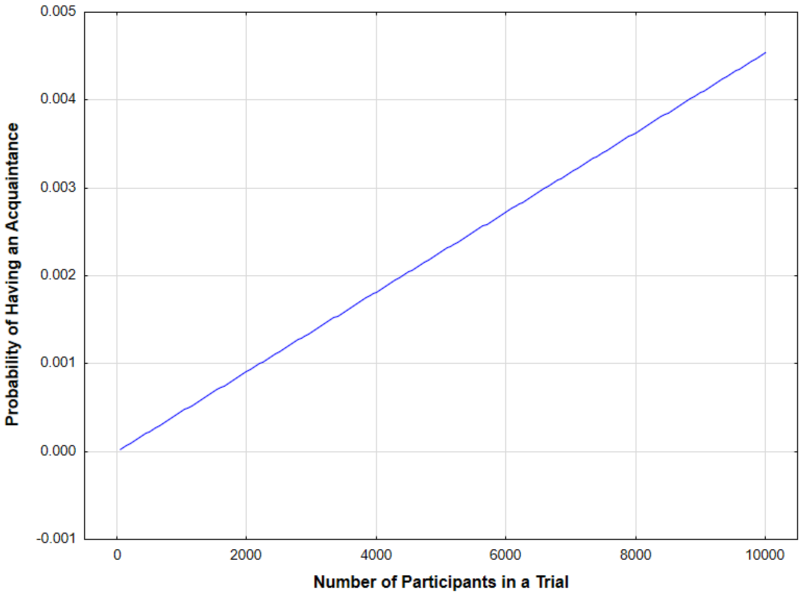


Figure 4: The probability of an adversary having an acquaintance in a specific trial (y-axis) against the size of the trial (x-axis).

For the trial sizes shown in Figure 4, the baseline population-to-sample risk will be below the commonly used threshold of 0.09.

B.2.1.4 Assumption 3: Cancer Trial

We now consider a specific disease: cancer and assume that the adversary knows that the target is participating in an oncology trial. The prevalence of cancer in the US in January 2017 is 15,760,939 individuals[4], which is 4.7% of the population. It is estimated that less than 5% of cancer patients participate in cancer trials [13], which represents 0.2% or less of the US population.

Based on that prevalence, we then have $pr(a)=0.259$, which means that a randomly selected adversary has approximately a 25% chance of knowing someone who has participated in a cancer trial.

Let us consider $pr(b|a)$. Under the worst case assumption (i.e., highest identification risk) we have $f_i=1$, which gives us $pr(b|a)={}^{n}\!/_{N}$, which is the sampling fraction. If we assume a 10,000 participant trial, then the sampling fraction is: 0.012. The overall risk of identification is $0.259\times 0.012=0.0033$, which is quite a low population-to-sample risk. If the $f_i=1$ is not correct (i.e., that any $f_i>1$) then the population-to-sample risk would be even smaller.

Therefore, the baseline probability of an adversary knowing someone who has participated in an oncology trial and successfully identifying their record is quite small.

B.2.2 Evaluation of Sample-to-Population Risk

In this type of risk assessment, the adversary is matching a record in the clinical trial dataset with a population registry. The kinds of quasi-identifiers in population registries are demographic and socio-economic indicators. Therefore, identification risk management should focus on reducing the identification risk on these public quasi-identifiers (rather than the full set of acquaintance quasi-identifiers).

B.2.3 Quasi-identifiers Synthesized in the Current Study

This analysis has made the case that under some common assumptions about adversary knowledge and types of clinical trial datasets, the baseline population-to-sample risk for clinical trial data is below the commonly used 0.09 threshold. Therefore, the focus for data synthesis should be on the public quasi-identifiers to protect against sample-to-population attacks. These are the types of quasi-identifiers that were synthesized in our study.

Therefore, the public quasi-identifiers selected for the N0147 trial were (in that order) age, sex, BMI, race, DFS event status, OS event status, time in days to DFS, and time in days to OS. In the case of OS, the event is death (which here is an outcome rather than an SAE), and for DFS the event was death or disease progression. All dates were converted to relative dates (consistent with a contemporary clinical trial de-identification standard [14]).

B.3 Limitations

The model we use makes assumptions, such as using the Dunbar number. To the extent that these assumptions are reasonable, the estimates here can be relied upon.

Our analysis assumes that the adversary will perform exact matching on the quasi-identifiers. This is a common assumption in the disclosure control literature.

[4] See https://seer.cancer.gov/explorer/application.html?site=1&data_type=5&graph_type=11&compareBy=sex&chk_sex_1=1&series=9&age_range=1&advopt_compprev_y_axis_var=0

Appendix C: Univariate Comparisons

C. Introduction

In this appendix we present the methods and results for comparing the univariate distributions in the real and synthetic data.

C.1 Methods

The univariate results consisted of distributions on the categories of the variables (the relevant continuous variables were categorized in the published secondary analysis study). Relative entropy (KL-divergence [15]) is often used in machine learning to compare two distributions. However, KL-divergence is difficult to interpret because it has no fixed upper bound and is not compared to a yardstick to obtain a relative interpretation. We therefore convert it to a relative value so that it can be interpreted more easily.

Dividing KL-divergence by Shannon's entropy we get the *relative increase in entropy due to using synthetic data*, and we use it to compare the univariate distributions of the real and synthetic datasets. It is a form of normalization of the relative entropy to make it interpretable (in the same way that relative error is interpreted when computing model prediction accuracy). A value of zero means that there are no differences in the distributions. A value of one means that the entropy or uncertainty due to the use of synthetic data as opposed to the real data is twice that of using the real data.

C.2 Results

The univariate comparisons of the distributions on the KL-divergence metric are shown in Table 1. As can be seen, all the values were less than 1%, therefore the relative increase in entropy is quite low due to data synthesis. The values that are zero in the table pertain to variables that were not synthesized in the partial synthesis process.

Variable	Normalized KL-divergence metric
Age	0.147%
Sex	0.35%
BMI	0.06%
ECOG	0%
Race	0.049%
KRAS	0%
T Stage	0%
Histology	0%
Adjuvant Chemotherapy	0.095%
Positive LNs	0%
Adjuvant Regimen	0%
Overall survival	0.054%
Disease free survival	0.017%

Table 1: Comparing the real and synthetic univariate distributions on the normalized KL-divergence metric.

References

[1] K. El Emam, L. Mosquera, and C. Zheng, "Optimizing the synthesis of clinical trial data using sequential trees," *J Am Med Inform Assoc*, Nov. 2020, doi: 10.1093/jamia/ocaa249.

[2] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[3] K. El Emam and L. Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly, 2013.

[4] K. El Emam, S. Rodgers, and B. Malin, "Anonymising and Sharing Individual Patient Data," *BMJ*, vol. 350, p. h1139, Mar. 2015, doi: 10.1136/bmj.h1139.

[5] European Medicines Agency, "External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use," Sep. 2017.

[6] Health Canada, "Guidance document on Public Release of Clinical Information," Apr. 01, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.

[7] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

[8] Institute of Medicine, "Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk," Washington, D.C., 2015.

[9] K. Benitez and B. Malin, "Evaluating Re-Identification Risks with Respect to the HIPAA Privacy Rule," *J Am Med Inform Assoc*, vol. 17, no. 2, pp. 169–177, Mar. 2010, doi: 10.1136/jamia.2009.000026.

[10] Janice Branson, Nathan Good, Jung-Wei Chen, Guillermo Monge, Christian Probst, and Khaled El Emam, "Evaluating the Re-identification Risk of a Clinical Study Report Anonymized under EMA Policy 0070 and Health Canada Regulations," *Trials*, vol. 21, p. 200, 2020.

[11] European Medicines Agency, "European Medicines Agency policy on publication of data for medicinal products for human use: Policy 0070." Oct. 02, 2014, [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

[12] CBO, "Research and Development in the Pharmaceutical Industry," Congressional Budget Office, 2006.

[13] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, "Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies," *Am Soc Clin Oncol Educ Book*, vol. 35, pp. 185–198, 2016, doi: 10.14694/EDBK_156686.

[14] PhUSE De-Identification Working Group, "De-Identification Standards for CDISC SDTM 3.2," 2015.

[15] Thomas Cover and Joy Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.