BMJ Open

Can synthetic data be a proxy for real clinical trial data? A validation study

Zahra Azizi,[1] Chaoyi Zheng,[2] Lucy Mosquera,[2] Louise Pilote [ID],[3,4] Khaled El Emam [ID],[5,6] GOING-FWD Collaborators

[1]Center for Outcomes Research and Evaluation, Faculty of Medicine, McGill University, Montreal, Québec, Canada
[2]Data Science, Replica Analytics Ltd, Ottawa, Ontario, Canada
[3]Medicine, McGill University, Montreal, Québec, Canada
[4]Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, Québec, Canada
[5]Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada
[6]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

Correspondence to
Dr Khaled El Emam;
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 III colon cancer trial secondary analysis using synthetic data generated by a machine learning method.

Participants There were 1543 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 CI 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 CI 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 HR CI overlap between the real and synthetic data was 61% for overall survival (real data: HR 1.56, 95% CI 1.11 to 2.2; synthetic data: HR 2.03, 95% CI 1.44 to 2.87) and 86% for disease-free survival (real data: HR 1.51, 95% CI 1.18 to 1.95; synthetic data: HR 1.63, 95% CI 1.26 to 2.1).

Conclusions The high concordance between the analytical 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 number NCT00079274.

Strengths and limitations of this study

► 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 patients with colon cancer 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.

including those 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 International Committee of Medical Journal Editors (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 (eg, 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 on which the accepted publication was based.

However, data access for secondary analysis remains a challenge.[12] To highlight this

BACKGROUND

It is often difficult for researchers to get access to high-quality individual-level data for secondary purposes (eg, 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 with the original publications[1] and has produced informative research results

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 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]

Anonymisation is one approach to making data available for secondary analysis. However, recently there have been repeated claims of successful reidentification attacks on anonymised data,[20–26] eroding public and regulators' trust in this approach.[26–36] However, it should be noted that there are no known successful reidentification attacks on anonymised 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 are 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 have 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 reanalysis 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.

METHODS
Data sources
We obtained the dataset for an oncology trial, N0147, from Project Data Sphere (PDS) (see https://data.projectdatasphere.org/).[45] The specific trial was selected because the PDS data were analysed 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 are already perturbed to anonymise it. The level of perturbation is dependent on the sponsor. Therefore, the use of the term 'real' should be interpreted to mean 'real and anonymised' data.

Summary of trial data
Trial N0147 was a randomised trial of 2686 patients with stage III 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 tumour 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 tumour 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 analysed in the published secondary analysis, which consisted of 1543 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 in the published secondary analysis was DFS, and the secondary endpoint was OS, and both DFS and OS were censored at 5 years.

The covariates in the published secondary analysis comprised three types of variables: (1) baseline

demographics, including age, sex and baseline BMI; (2) baseline Eastern Cooperative 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, histological status and Kirsten rat sarcoma virus biomarker status.

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 synthesised 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 with 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 online supplemental materials.

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 reran 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 and pathological) and obstruction. Survival analysis was performed using the Kaplan-Meier curve. The log rank test and the 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.

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 have 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 divergence and interval overlap for the CIs of model parameters. Both are described further.

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 to test the hypothesis that obstruction was an important predictor.

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 5 years and DFS after 5 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: (1) 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); (2) 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 (3) the χ^2 tests considers independence, whereas the relationship that is being tested is whether each of the covariates is 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 0 and 1 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 CIs were compared. CI 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 CIs that overlap. Our formulation gives an overlap value of 0 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 CIs.

Multivariate analysis

For the multivariate models, we compared the Cox model HR estimates between the real and synthetic data. We also computed the CI overlap of the HRs from the Cox models.

Patient and public involvement

The comparative analysis of synthetic to real data did not have any patient or public involvement.
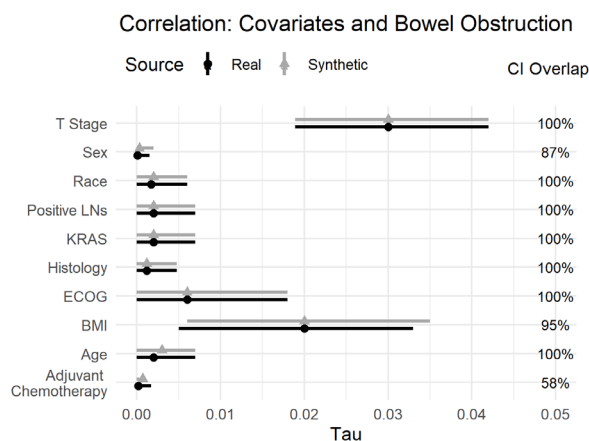
Figure 1 Tau coefficient for the real and synthetic data, and the CI overlap for the bivariate relationship with obstruction. BMI, Body Mass Index; ECOG, Eastern Cooperative Oncology Group; KRAS, Kirsten rat sarcoma virus; LNs, Lymph Nodes.

RESULTS FOR TRIAL N0147

We compared 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 online supplemental materials.

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 CI overlaps. 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: OS and DFS. The results in figure 2 show the effect sizes for the bivariate relationships with OS. 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 DFS. The conclusions are like OS with one exception: the CIs for the relationship between race and DFS do not overlap. Given the weak relationship between race and DFS, this lack of CI 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.
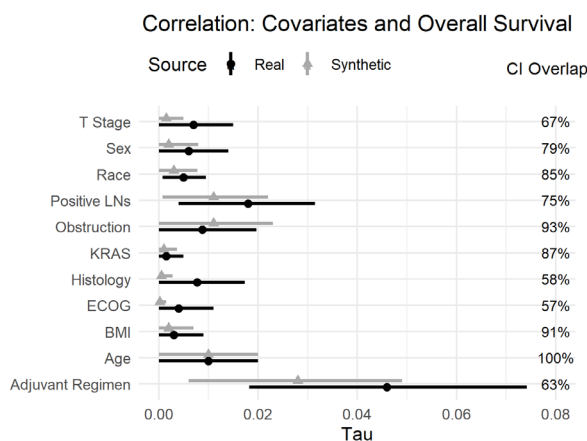

Figure 2 Tau coefficient and CI overlap for the real and synthetic variables against overall survival. BMI, Body Mass Index; ECOG, Eastern Cooperative Oncology Group; KRAS, Kirsten rat sarcoma virus; LNs, Lymph Nodes.

Multivariate analysis

For the multivariate analyses, the real data results were like those that were in the published secondary analysis. We first compared the survival curves for obstructed and non-obstructed patients on OS (figure 4) and DFS (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 HR model parameters were generally in the same direction with relatively high overlap for the CIs. This is the case for the OS model in figure 6 and the DFS model in figure 7.

The main hypothesis being tested in the published secondary analysis pertains to obstruction. For the OS model the HR for obstruction overlap was high at 61%
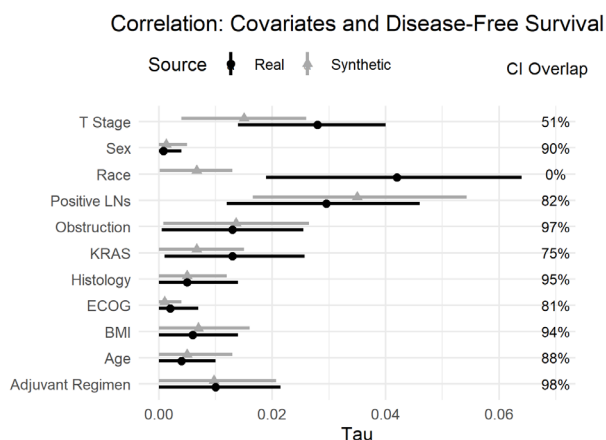

Figure 3 Tau coefficient and CI overlap for the real and synthetic variables against disease-free survival. BMI, Body Mass Index; ECOG, Eastern Cooperative Oncology Group; KRAS, Kirsten rat sarcoma virus; LNs, Lymph Nodes.
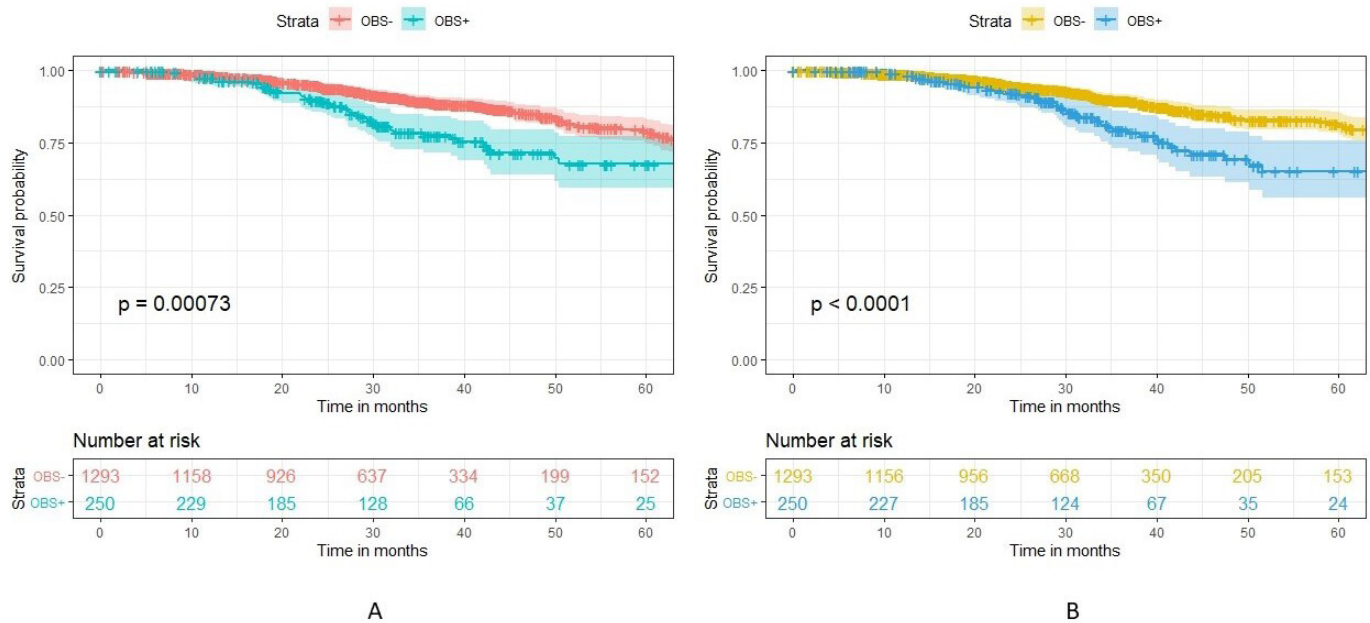
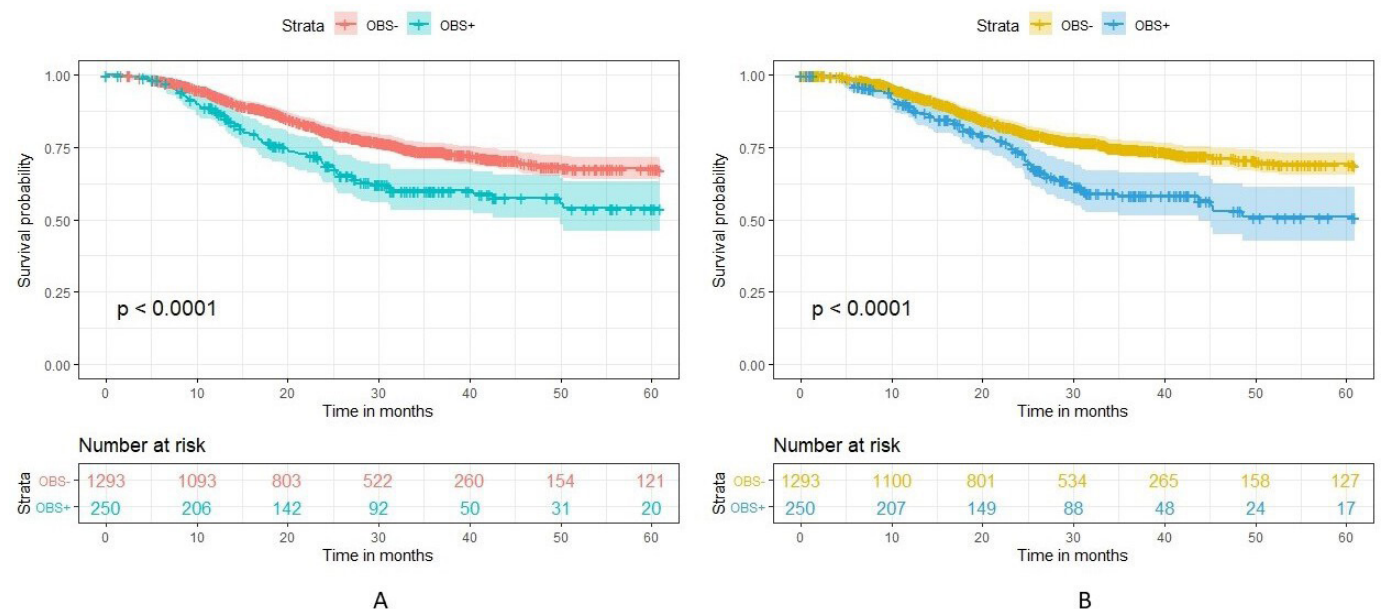**Figure 4** Survival curve comparing overall survival in OBS+ and OBS– patients in the real (A) versus synthetic (B) datasets. OBS+, obstructed; OBS–, non-obstructed.

(HR of 1.56; 95% CI: 1.11 to 2.2 for real data, and HR of 2.03; 95% CI: 1.44 to 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 to 1.95, and the synthetic data HR of 1.63; 95% CI: 1.26 to 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 OS and figure 7 for the DFS model,

with lower CI 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 OS model in figure 6 and the DFS model in figure 7 is that the CIs 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

Figure 5 Survival curve comparing disease-free survival in OBS+ and OBS– patients in the real (A) versus synthetic (B) datasets. OBS+, obstructed; OBS–, non-obstructed.
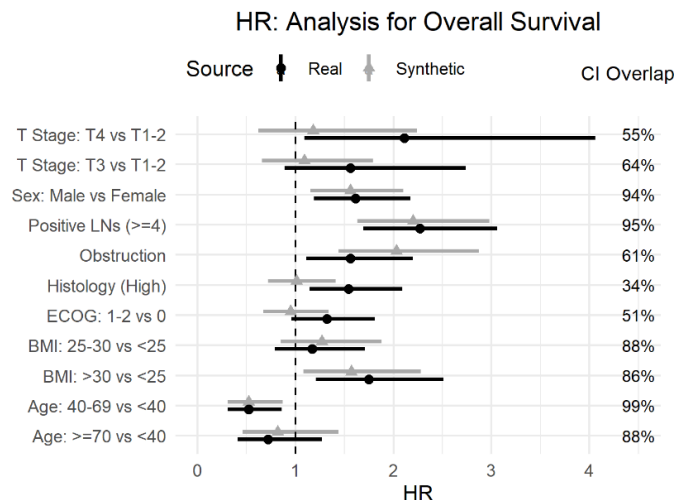
HR: Analysis for Overall Survival

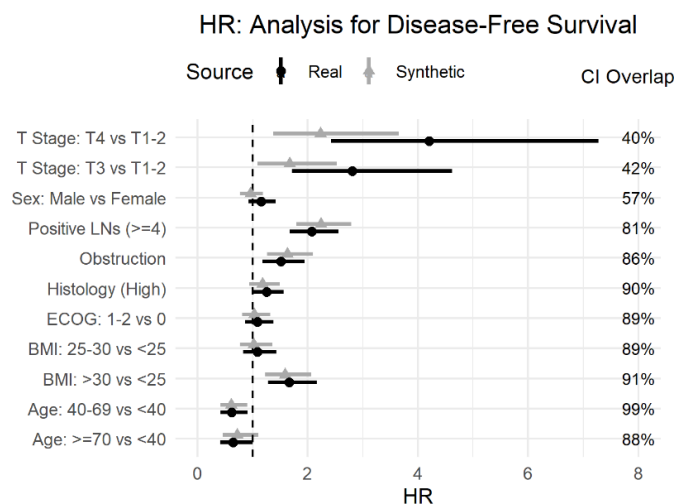**Figure 6** Comparison of real and synthetic Cox model parameters (HR) with the overall survival outcome variable. BMI, Body Mass Index; ECOG, Eastern Cooperative Oncology Group; LNs, Lymph Nodes.

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.

DISCUSSION
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 reanalysis 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 patients with colon cancer. We applied

HR: Analysis for Disease-Free Survival

Figure 7 Comparison of real and synthetic Cox model parameters (HR) with the disease-free survival outcome variable. BMI, Body Mass Index; ECOG, Eastern Cooperative Oncology Group; LNs, Lymph Nodes.

a commonly used synthesis approach that ensured the potentially identifying variables (the quasi-identifiers) in the dataset were appropriately synthesised.

We found that for the univariate and bivariate analyses in the published study, the synthetic data were 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 OS and DFS after 5 years in the both the real and synthetic datasets.

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 synthesise 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.[41] 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 those from 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 are shared, and the analysts build their models on the synthetic data. Then their analysis code (say an R or SAS programme) 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 are 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 provides evidence that synthesised 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.

Limitations

The data we used in our analysis came from PDS, which shares datasets that have already gone through a perturbation to anonymise the data. This would not affect our results or conclusions because the published study that we replicated used the same (perturbed) dataset from PDS. More generally, synthetic data can be generated from pseudonymous data rather than from fully anonymised data. Multiple researchers have noted that synthetic data do not have an elevated identity disclosure (privacy) risk,[60 64–71] and therefore anonymisation 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 hypothesise 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.

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 with the real datasets. In this study, we have provided one answer to that question. Our reanalysis 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.

ORCID iDs
Louise Pilote http://orcid.org/0000-0002-6159-0628
Khaled El Emam http://orcid.org/0000-0003-3325-4149

REFERENCES

1 Ebrahim S, Sohani ZN, Montoya L, *et al*. Reanalyses of randomized clinical trial data. *JAMA* 2014;312:1024–32.
2 Ferran J-M, Nevitt SJ. European medicines Agency policy 0070: an exploratory review of data utility in clinical study reports for academic research. *BMC Med Res Methodol* 2019;19:204.
3 Phrma & EFPIA. Principles for responsible clinical trial data sharing, 2013. 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*, 2017.
6 European Medicines Agency. European medicines Agency policy on publication of data for medicinal products for human use: policy, 2014. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf
7 Taichman DB, Backus J, Baethge C, *et al*. Sharing clinical trial data: a proposal from the International Committee of medical Journal editors. *Ann Intern Med* 2016;164:505–6.
8 Institute of Medicine,. *Sharing clinical trial data: maximizing benefits, minimizing risk*. Washington, DC, 2015.
9 International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals, 2019. Available: http://www.icmje.org/icmje-recommendations.pdf [Accessed 29 Jun 2020].
10 The Wellcome Trust. Policy on data, software and materials management and sharing, 2017. Available: https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing [Accessed 12 Sep 2017].
11 National Institutes of Health. Final NIH statement on sharing research data, 2003. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [Accessed 29 Jun 2020].
12 Doshi P. Data too important to share: do those who control the data control the message? *BMJ* 2016;352:i1027.
13 Polanin JR. Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing. *J Clin Epidemiol* 2018;98:157–9.
14 Naudet F, Sakarovitch C, Janiaud P, *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* 2018;360:k400.
15 Nevitt SJ, Marson AG, Davie B, *et al*. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. *BMJ* 2017;357:j1390.
16 Villain B, Dechartres A, Boyer P, *et al*. Feasibility of individual patient data meta-analyses in orthopaedic surgery. *BMC Med* 2015;13:131.
17 Ventresca M, Schünemann HJ, Macbeth F, *et al*. Obtaining and managing data sets for individual participant data meta-analysis: Scoping review and practical guide. *BMC Med Res Methodol* 2020;20:113.
18 Artificial Intelligence in Health Care. *National Academy of medicine and the general accountability office*, 2019.
19 El Emam K, Jonker E, Moher E, *et al*. A review of evidence on consent bias in research. *Am J Bioeth* 2013;13:42–4.
20 de Montjoye Y-A, Hidalgo CA, Verleysen M, *et al*. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 2013;3:1376.
21 de Montjoye Y-A, Radaelli L, Singh VK, *et al*. Identity and privacy. unique in the Shopping mall: on the reidentifiability of credit card metadata. *Science* 2015;347:536–9.
22 Sweeney L, Yoo JS, Perovich L, *et al*. Re-Identification risks in HIPAA safe harbor data: a study of data from one environmental health study. *Technol Sci* 2017;2017:2017082801.
23 Su Yoo J, Thaler A, Sweeney L. Risks to patient privacy: a re-identification of patients in Maine and Vermont statewide hospital data. *J Technol Sci* 2018:2018100901.
24 Sweeney L. *Matching known patients to health records in Washington State Data, Harvard University. data privacy lab*, 2013.
25 Sweeney L, von Loewenfeldt M, Perry M. Saying it's anonymous doesn't make it so: re-identifications of 'anonymized' law school data. *J Technol Sci* 2018:2018111301.
26 Zewe A. Imperiled information: Students find website data leaks pose greater risks than most people realize, *Harvard John A. Paulson School of Engineering and Applied Sciences*, 2020. Available: https://www.seas.harvard.edu/news/2020/01/imperiled-information [Accessed 23 Mar 2020].
27 Bode K. *Researchers find 'anonymized' data is even less anonymous than we thought. Motherboard: Tech by Vice*, 2020. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought
28 Clemons E. Online profiling and invasion of privacy: the myth of anonymization. *HuffPost* 2013.
29 Jee C. You're very easy to track down, even when your data has been anonymized, MIT Technology Review, 2019. Available: https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/ [Accessed 11 May 2020].
30 Kolata G. Your data were 'anonymized'? These scientists can still identify you. *The New York Times* 2019.
31 Lomas N. Researchers spotlight the lie of 'anonymous' data, 2019. Available: https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/ [Accessed 11 May 2020].
32 Mitchell S. Study finds HIPAA protected data still at risks, 2019. Available: https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/ [Accessed 11 May 2020].
33 Thompson SA, Warzel C. Twelve million phones, one dataset, zero privacy. *The New York Times* 2019.
34 Hern A. 'Anonymised' data can never be totally anonymous, says study. *The Guardian* 2019 https://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds#:~:text=%E2%80%9CAnonymised%E2%80%9D%20data%20lies%20at%20the,impossible%20for%20any%20complex%20dataset.
35 Wolk A. The (Im)Possibilities of Scientific Research Under the GDPR. Cybersecurity Law Report 2020.
36 Ghafur S, Van Dael J, Leis M, *et al*. Public perceptions on data sharing: key insights from the UK and the USA. *Lancet Digit Health* 2020;2:e444–6.
37 El Emam K, Mosquera L, Hoptroff R. *Practical synthetic data generation: balancing privacy and the broad availability of data*, 2020.
38 El Emam K, Hoptroff R. The synthetic data paradigm for using and sharing data. *Cutter Executive Update* 2019;19.
39 Polonetsky J, Renieris E. 10 privacy risks and 10 privacy technologies to watch in the next decade. *Future of Privacy Forum* 2020.
40 Guo A, Foraker RE, MacGregor RM, *et al*. 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* 2020;2.
41 Navar AM, Pencina MJ, Rymer JA, *et al*. Use of open access platforms for clinical trial data. *JAMA* 2016;315:1283.
42 Reiner Benaim A, Almog R, Gorelik Y, *et al*. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR Med Inform* 2020;8:e16492.
43 Foraker RE, Yu SC, Gupta A, *et al*. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 2020;3:ooaa060.
44 Beaulieu-Jones BK, Wu ZS, Williams C. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 2017;12:159756.
45 CEO life sciences consortium. Project Data Sphere; Share, integrate & analyze cancer research data. Available: https://projectdatasphere.org/projectdatasphere/html/home
46 Alberts SR, Sargent DJ, Nair S, *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* 2012;307:1383–93.
47 Dahdaleh FS, Sherman SK, Poli EC, *et al*. Obstruction predicts worse long-term outcomes in stage III colon cancer: a secondary analysis of the N0147 trial. *Surgery* 2018;164:1223–9.
48 Carraro PG, Segala M, Cesana BM, *et al*. Obstructing colonic cancer: failure and survival patterns over a ten-year follow-up after one-stage curative surgery, *Dis*. *Colon Rectum* 2001;44:243–50.
49 Mella J, Biffin A, Radcliffe AG, *et al*. 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* 1997;84:1731–6.

50 Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput Stat Data Anal* 2011;55:3232–43.

51 Arslan RC, Schilling KM, Gerlach TM, *et al*. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *J Pers Soc Psychol* 2018 doi:10.1037/pspp0000208

52 Bonnéry D, Feng Y, Henneberger AK, *et al*. The promise and limitations of synthetic data as a strategy to expand access to State-Level Multi-Agency longitudinal data. *J Res Educ Eff* 2019;12:616–47. doi:10.1080/19345747.2019.1631421

53 Sabay A, Harris L, Bejugama V. Overcoming small data limitations in heart disease prediction by using surrogate data. *SMU Data Science Review* 2018.

54 Freiman M, Lauger A, Reiter J. Data synthesis and perturbation for the American community survey at the US. census bureau, us census bureau, working paper 2017.

55 Nowok B. *Utility of synthetic microdata generated using tree-based methods*, 2015.

56 Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality* 2016;7:67–97.

57 Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R1. *Statistical Journal of the IAOS* 2017;33:785–96. doi:10.3233/SJI-150153

58 Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife* 2020;9 doi:10.7554/eLife.53275

59 Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006;15:651–74.

60 Park N, Mohammadi M, Gorde K, *et al*. Data synthesis based on generative adversarial networks. *Proc VLDB Endow* 2018;11:1071–83. doi:10.14778/3231751.3231757

61 Chin-Cheong K, Sutter T, Vogt JE. 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) 2019.

62 Karr AF, Kohnen CN, Oganian A, *et al*. A framework for evaluating the utility of data altered to protect confidentiality. *Am Stat* 2006;60:224–32. doi:10.1198/000313006X124640

63 Agresti A. *Categorical data analysis*. 2nd edn. Hoboken: Wiley, 2002.

64 Reiter JP. New approaches to data dissemination: a glimpse into the future (?). *CHANCE* 2004;17:11–15.

65 Hu J. Bayesian estimation of attribute and identification disclosure risks in synthetic data, 2018. Available: http://arxiv.org/abs/1804.02784 [Accessed 15 Mar 2019].

66 Taub J, Elliot M, Pampaka M. Differential correct attribution probability for synthetic data: an exploration. *Privacy in Statistical Databases* 2018:122–37.

67 Hu J, Reiter JP, Wang Q. Disclosure risk evaluation for fully synthetic categorical data. *Privacy in Statistical Databases* 2014:185–99.

68 Wei L, Reiter JP. Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Stat J IAOS* 2016;32:93–108.

69 Ruiz N, Muralidhar K, Domingo-Ferrer J. On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. *Privacy in Statistical Databases* 2018:59–74.

70 Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J Royal Statistical Soc A* 2005;168:185–205.

71 El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J Med Internet Res* 2020;22:e23139.