JAMIA Open, 00(0), 2021, 1–10
doi: 10.1093/jamiaopen/ooab012
Research and Applications

Research and Applications

Evaluating the utility of synthetic COVID-19 case data

Khaled El Emam,[1,2,3] Lucy Mosquera,[3] Elizabeth Jonker,[2] and Harpreet Sood[4,5]

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada, [2]Electronic Health Information Laboratory, Childrens Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada, [3]Data Science, Replica Analytics Ltd, Ottawa, Ontario, Canada, [4]London School of Economics, London, UK and [5]National Health Service, London, UK

Corresponding Author: Khaled El Emam, PhD, Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada; kelemam@ehealthinformation.ca

Received 13 December 2020; Revised 1 February 2021; Editorial Decision 5 February 2021; Accepted 10 February 2021

ABSTRACT

Background: Concerns about patient privacy have limited access to COVID-19 datasets. Data synthesis is one approach for making such data broadly available to the research community in a privacy protective manner.
Objectives: Evaluate the utility of synthetic data by comparing analysis results between real and synthetic data.
Methods: A gradient boosted classification tree was built to predict death using Ontario's 90 514 COVID-19 case records linked with community comorbidity, demographic, and socioeconomic characteristics. Model accuracy and relationships were evaluated, as well as privacy risks. The same model was developed on a synthesized dataset and compared to one from the original data.
Results: The AUROC and AUPRC for the real data model were 0.945 [95% confidence interval (CI), 0.941–0.948] and 0.34 (95% CI, 0.313–0.368), respectively. The synthetic data model had AUROC and AUPRC of 0.94 (95% CI, 0.936–0.944) and 0.313 (95% CI, 0.286–0.342) with confidence interval overlap of 45.05% and 52.02% when compared with the real data. The most important predictors of death for the real and synthetic models were in descending order: age, days since January 1, 2020, type of exposure, and gender. The functional relationships were similar between the two data sets. Attribute disclosure risks were 0.0585, and membership disclosure risk was low.
Conclusions: This synthetic dataset could be used as a proxy for the real dataset.

Key words: data sharing, data synthesis, synthetic data, data access

LAY SUMMARY

There remains a strong need for sharing COVID-19 data with the research community. This study evaluates whether data synthesis can address that need. We synthesized the Ontario case database of 90 514 individuals testing positive for SARS-CoV-2 and created a synthetic version of that. The synthesis method used sequential decision trees. A machine learning (gradient boosted trees) mortality prediction model was constructed using the synthetic data and its accuracy and the relationships it detected were compared to the real data. The results of the real and synthetic data models were similar and the conclusions were the same. A privacy risk assessment on the synthetic data showed that the attribute and membership disclosure risks were low. We conclude that the synthetic version of the COVID-19 testing dataset can be shared more broadly as it has high utility and privacy characteristics.

INTRODUCTION

COVID-19 has created demand for an unprecedented level of data sharing with researchers, health care providers, and public health organizations.[1–3] Even before the current pandemic, global health and funding agencies have been calling for greater sharing of public health data.[4,5] To address the needs of global health research, data sharing must include clinical data collected from routine care as well as clinical trials.[6] There is also significant potential value in using Artificial Intelligence methods to analyze COVID-19 data,[7] but such analytic methods require large volumes of data,[8] further amplifying the need for efficient and scalable data sharing mechanisms.

Some organizations have already set up large scale COVID-19 data sharing programs. For example, South Korea is providing access to 5 years of health insurance benefit claims for COVID-19 patients for research purposes through the Health Insurance Review and Assessment (HIRA) service (the national health insurer).[9,10] The NIH is providing data through a secure enclave as part of the National COVID Cohort Collaborative (N3C).[11] The Observational Health Data Sciences and Informatics (OHDSI) organization has made large datasets from participating organizations available through a federated analysis model.[12] The Government of Ontario is similarly making population level administrative and clinical databases available to the research community.[13]

However, privacy concerns have historically acted as a barrier to local and global sharing of public health data.[14,15] These concerns are growing in the context of making COVID-19 data more accessible,[16–19] and some governments have begun to reduce the amount of information being shared about COVID cases.[16,20–25] It is known that privacy concerns make individuals reluctant to seek care and adopt other privacy protective behaviors,[26] and providers can be reluctant to report cases to public health authorities due to concerns about patient privacy,[27–32] even in the context of a pandemic.[32]

Privacy enhancing technologies can address this risk by creating databases with perturbed data that can be shared with a very small risk of identifying individual patients. Data synthesis is one approach for achieving that.[33,34] It has long been recognized that synthetic data is a key approach for data dissemination complementing more traditional disclosure control methods,[35] and has been highlighted as a key privacy enhancing technology to enable data access for the coming decade.[36]

A number of recent efforts have made large COVID-19 datasets available specifically through data synthesis. The Clinical Practice Research Datalink (CPRD) database in the UK has made available a COVID-19 symptoms and risk factors synthetic dataset based on primary care encounters in the UK.[37,38] The NIH's N3C is also developing synthetic datasets for broader sharing with researchers.[11,39]

Multiple researchers and analyses have noted that synthetic data does not have an elevated identity disclosure (privacy) risk because there is no unique or one-to-one mapping between the records in the synthetic data with the records in the original data.[35,40–47] Therefore, a key remaining question is whether a synthetic version of COVID-19 datasets can provide reasonably good data utility and act as a proxy for real data. If that is the case, then synthetic COVID-19 datasets can be shared more broadly for secondary analysis and research.

This paper focuses on an assessment of the utility of a synthetic variant of the Ontario COVID-19 case dataset using a commonly applied data synthesis approach: sequential trees. Utility was defined as the ability to replicate patterns and analysis conclusions from the synthetic data that were in the original data.[48] Specifically, we evaluate the extent to which synthetic data can replicate the accuracy and functional relationships of a gradient boosted tree (GBT) classification model predicting death for 90 514 Ontario cases.

MATERIALS AND METHODS

The objective was to construct a prediction model of COVID-19 mortality in Ontario using the real data and compare that to the same model developed on the synthetic data. The outcome was a binary indicator of death over the study period. The predictors were individual and community variables reflecting factors that have been shown in the literature to affect COVID-19 mortality.

The Supplementary Material contains a review of factors that have been found to affect COVID-19 mortality and that we consider in our analysis.

Our primary analysis uses a machine learning technique. It has been argued, specifically in the context of COVID-19 mortality prediction, that machine learning models are better at fully using the information in clinical datasets compared to traditional regression methods.[49]

Data set

The dataset we used was obtained on November 15 from Esri Canada's COVID-19 dashboard,[50] which is collected from the Public Health Agency of Canada and curated. The last case was reported on that day. The full dataset consisted of 306 816 Canadian cases. Because the values were incomplete for some provinces, our analysis focused only on Ontario with 100 368 records at the time the data was obtained. The fields in that dataset are shown in Table 1.

This case data were linked with community information for each of the health regions in Ontario.[51] The variables related to the health region are shown in Table 2. These variables were also considered in our mortality prediction model. Following recommended practices, the selection of these predictors was informed by previous literature,[52] and a literature review is provided in the Supplementary Appendix.

There are precedents for using population or community metrics in prediction models in the context of COVID-19. For example, a model of student transmission of the disease was constructed and population values from prior publications were used to instantiate it.[53] Similarly, a mortality model used population values from the China CDC.[54] In both cases, individual-level data were created through simulation, using the population values to define sampling distribution parameters. In another study, a baseline model was developed using individual-level data on a related outcome (hospitalizations with

Table 1. Fields in the Canadian COVID-19 case dataset used for our study

Variables	Definitions
Date reported	Number of days since January 1, 2020
Health region	34 unique regions
Age group	Decades from 20 to 80+ (ordinal)
Gender	
Exposure	Close contact, outbreak, travel, not reported
Case status	Recovered, deceased, active

Table 2. Fields included on the health region community

Variables	Definitions
Proportion living in rural areas	Rural areas are defined as all territory lying outside population centers (population centers have a population of at least 1000 and a density of 400 or more persons per square kilometer)
Proportion of immigrants	An immigrant as a person who is, or who has ever been, a landed immigrant or permanent resident. Such a person has been granted the right to live in Canada permanently by immigration authorities. Immigrants who have obtained Canadian citizenship by naturalization are included in this group.
Proportion of aboriginal population	Aboriginal identity is based on whether the person identified with the Aboriginal peoples of Canada. This includes those who are First Nations, Métis or Inuk (Inuit) and/or those who are registered or treaty Indians (i.e. registered under the Indian Act of Canada) and/or those who have membership in a First Nation or Indian band.
Prevalence of diabetes	Population age 12 and older who reported having been diagnosed by a health professional as having type 1 or type 2 diabetes; includes females age 15 and older who reported having been diagnosed with gestational diabetes.
Prevalence of COPD	Population age 35 and older who reported being diagnosed by a health professional with chronic bronchitis, emphysema or chronic obstructive pulmonary disease (COPD).
Prevalence of high blood pressure	Population age 12 and older who reported that they have been diagnosed by a health professional as having high blood pressure.
Family medicine physicians per 100 000 population	The number of family medicine physicians per 100 000 population.
Proportion reporting Moderate-to-severe Food Insecurity	Food security is commonly understood to exist in a household when all people, at all times, have access to sufficient safe and nutritious food for an active and healthy life. Conversely, food insecurity occurs when food quality and/or quantity are compromised and is typically associated with limited financial resources.

The definitions are taken from the source document [51].

diagnoses of pneumonia and influenza) then its predictions were adjusted to match COVID-19 case fatality rates.[55]

In our study, simulating individual-level data from the community prevalence values would require an independence assumption among the covariates, which would weaken the overall model. Furthermore, the prevalence values we use can be seen as the individual-level likelihoods of a particular characteristic.

Cases where the case status was unknown or still active were removed. That way we only had recovered and deceased individuals. The final dataset had 90 514 observations. Of these, 3456 were deceased, which represents 3.82% of the Ontario dataset.

Synthesis method

The individual level variables in Table 1 were synthesized. The linking of the datasets with the community variables was performed on the synthetic data.

There are a number of data synthesis methods that have been used recently in the literature, such as Bayesian networks,[56–58] and Generative Adversarial Networks.[40,59,60] In this study, we used another approach that has been applied quite extensively for the synthesis of health and social sciences data, namely sequential classification and regression trees.[61–69] Classification and regression trees[70] have been proposed for data synthesis when implemented in a sequential manner.[71] Furthermore, existing evaluations have concluded that the privacy risks using sequential tree synthesis is low.[47,72] With these types of models, a variable is synthesized by using the values earlier in the sequence as predictors. Conceptually, sequential synthesis is similar to modeling multiple outcome variables using classifier chains[73] and regressor chains.[74] The details of the specific method we used are described elsewhere.[75]

Analysis methods

The same analysis was performed on the real and the synthetic datasets, and the results were compared. The analysis methods were selected to reflect common approaches that are used to model mortality. Our main analytical method uses a machine learning technique and the sensitivity analysis uses a regression technique. Both were operationalized to provide interpretable models, with an emphasis on selecting the most important variables and understanding the functional form of relationships.

The primary data analysis method is shown in Figure 1. Gradient boosted classification trees (GBT)[76] were used to build a predictive model of death. Five hundred bootstrap samples were used to compute 95% confidence intervals for all results reported. For each bootstrap sample, the records that were out-of-sample were used as the test dataset for that iteration. Five-fold cross validation was used to determine the optimal number of trees for the GBT model built within each bootstrap iteration. Because the dataset was imbalanced, under sampling of the majority class was used to create a balanced training dataset, within each bootstrap iteration.[77]

We compared the bootstrap confidence interval overlap between the two datasets. Confidence interval overlap has been proposed for evaluating the utility of privacy protective data transformations,[78] which is defined as the percentage average of the real and synthetic confidence intervals that overlap. The definition of this overlap is provided in the appendix. To interpret confidence interval overlap,

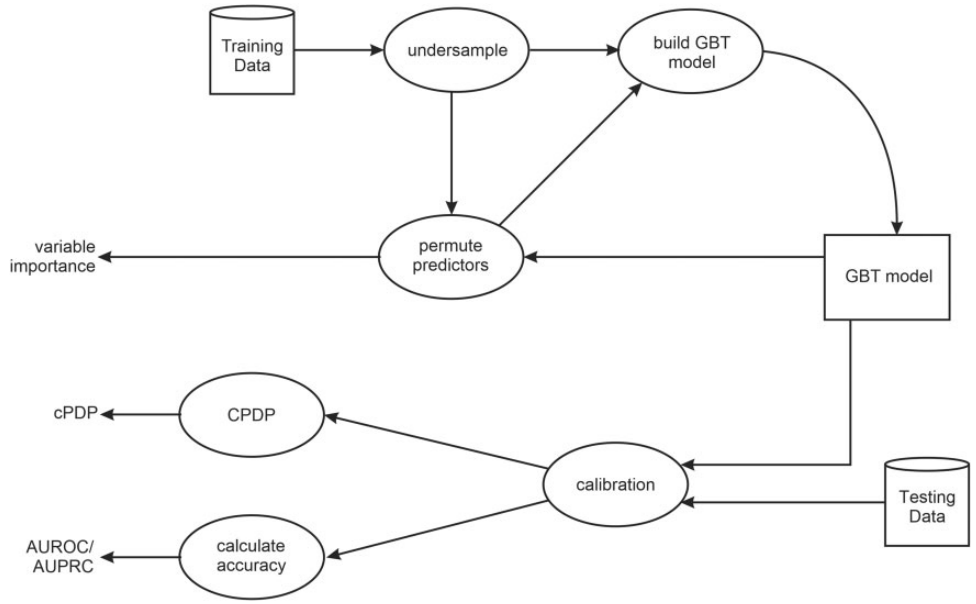

Figure 1. Process diagram for the analysis method. The diagram shows the steps for each iteration of the bootstrap sampling. The testing data is the out-of-sample subset in each bootstrap iteration. cPDP stands for conditional partial dependency plot.

we propose a minimal acceptable overlap of 37.5%, as explained and justified in the appendix.

Calibration

Probability calibration was performed as the predicted probabilities from boosted decision tree models do not correspond directly with the true probabilities of class membership. This discrepancy is amplified when data are under-sampled. Boosted decision trees can be viewed as additive logistic regression, meaning that the predictions made by boosting are trying to fit a logit of the true probabilities, rather than directly fitting the true probabilities.[76,79] Isotonic regression can be used to calibrate the probabilities of boosted decision trees to ensure that the predicted probabilities correspond with the true probabilities of class membership.[79,80] To calibrate the predicted probabilities $\widehat{y}_i$, using the true class labels y_i, the following regression model is fit: $y_i = m(\widehat{y}_i) + \varepsilon_i$ where m is an isotonic (non-decreasing) function. Calibrated probabilities are used for assessing model performance and conditional partial dependence.

Model accuracy

We compared the real and synthetic datasets in terms of the death prediction model accuracy. Model accuracy was assessed using the Area Under the Receiver Operating Characteristic curve (AUROC).[81] The ROC plots the false positive rate against recall and is commonly used to evaluate the performance of binary classifiers in machine learning. For binary classification tasks a AUROC of 0.5 is the expected performance of a random classifier, whereas an AUROC of 1 is the expected performance of a perfect classifier.

Another metric which focuses on predictions of the positive class is the Area Under the Precision-Recall Curve (AUPRC).[82] Interpretation of AUPRC is dependent on the class distribution of the outcome. This means it is particularly important to evaluate AUPRC on the test data with true class distributions as the minimal achievable value is dependent on that distribution,[83] and the AUPRC value of a random classifier is the rate of the positive class,[83] which in our case is 0.0382.

Variable importance

We compared the variable importance in the models built using the real and synthetic datasets. One general purpose method for evaluating variable importance, or to determine which predictors are most relevant to predicting the outcome, is to use permutation.[84,85] If we let a training set of predictors be X with each row denoted by x_i, and the corresponding outcome variable by y_i, then we can permute a predictor variable j to get x_i^j. A model built from the training dataset is $f(.)$. The importance of a predictor variable can be given by using the model built on the training data to compute $\sum_{i=1}^{n}\left(L\left(y_i, f(x_i^j)\right) - L\left(y_i, f(x_i)\right)\right)$ where $L(.)$ is a loss function, such as prediction accuracy, and n is the total number of observations. This then gives us the importance of the variable j.

There is evidence that permuting a variable is biased towards predictors that are correlated with other predictors and that have many categories.[86,87] The reason is that, if we have two predictors that are positively correlated, say x_1 and x_2, then there will be no training examples where x_1 is large and x_2 is small, which means that the predictions made in that region will be extrapolations, resulting in high importance for these two variables.[88]

An alternative is to permute and reconstruct the model from the (undersampled) training data, and then compare the prediction accuracy on the original and permuted models[88,89] as follows: $\sum_{i=1}^{n}\left(L\left(y_i, f^j(x_i)\right) - L\left(y_i, f(x_i)\right)\right)$ where $f^j(.)$ is the model built with permuted variable j. This approach addresses the bias risk and the average difference in loss (or accuracy) across multiple permutations allows us to prioritize the variables.

Conditional partial dependence plots

To illustrate the relationships between the most important predictor variables and the outcome of interest, conditional partial dependence plots were constructed.[90] Traditionally, partial dependence plots for the j^{th} variable plot $\frac{1}{n}\sum_{i=1}^{n} f\left(x_i^{j=J[k]}\right)$ against $J[k]$ where $x_i^{j=J[k]}$ is the i^{th} observation where the j^{th} variable is set to $J[k]$ and

$J[]$ is the set of unique values for the j^{th} variable. Partial dependence plots have been subject to criticism as not all observations may plausibly be observed with $x_i^j = J[k]$, leading to poor predictions due to extrapolation.[88] Conditional partial dependence plots aim to minimize extrapolation by calculating partial dependence within conditional subgroups, and then pools the results across subgroups. It also isolates the effect of a variable so we can view its impact, within the model, on the outcome.

Sensitivity analysis

The data synthesis method that we used was based on decision trees, as was the primary modeling method for predicting mortality. There is the potential that using similar methods for synthesis and analysis creates a positive data utility bias in that the generative model learns specific patterns in the data that the analysis model is able to also detect in the generated data. The risk is that a different analysis method may not be able to detect the same pattern. To guard against this, we also built logistic regression mortality models using each of the real and synthetic datasets. Logistic regression is a common analytical approach in epidemiology and would not be biased from using data generated by a sequential tree synthesizer. This would allow us to directly test the sensitivity of the results to the analytical method used. The methods and results from this logistic regression model are included in the appendix.

Evaluating distinguishability

We also compared the multivariate distributions of the real and synthetic data using a distinguishability metric. We applied an omnibus comparison of multivariate distributions using a binary classifier.[91,92] This means that we build a discriminator model that attempts to distinguish between real and synthetic datasets. If it is not able to tell the difference, then that indicates that the real and synthetic data are similar to each other. The distinguishability metric we use is based on propensity scores.[93,94] Additional details to how we have adapted it to our specific context are described in the appendix, but the basic concept is that it is an interpretable mean squared error compared to guessing whether a record is real or synthetic.

Evaluating privacy

To evaluate the privacy risks of the synthetic data we tested for two types of disclosure. The first is attribute disclosure conditional on identity disclosure, which assesses the probability of mapping a synthetic record to a real person, and conditional on that learning something new about the individual.[47] The second is membership disclosure which assesses whether an adversary would reliably know whether a target individual was in the real dataset used for synthesis. The details of the methods used for each of these two evaluations are provided in the appendix.

RESULTS

Descriptive statistics

The summary statistics for the real dataset are shown in Table 3.

GBT model results

The AUROC value for the GBT model on the real data was 0.945 and the AUPRC was 0.340 as shown in Table 4. The baseline death rate was 3.82% and therefore the AUPRC is a considerable improvement over that. The GBT model built on the synthetic data yielded similar model accuracy results with a AUROC of 0.940 and AUPRC of 0.313 (CI overlap 45.50% and 52.02%, respectively, and they are both above our threshold).

The variable importance for the real data is shown in Figures 2 and 3 using each of the prediction accuracy measures. All CI values are above our overlap threshold. The variables with the largest impact on the outcome are from the individual characteristics. The community level characteristics did not have a significant effect on death. The most important variable is age, followed by date reported, exposure, and gender. By far the most important predictor of death is age with an approximately 6% increase in AUROC with its inclusion. The confidence intervals for the accuracy gain associated with gender and exposure cross zero when quantified using AUPRC or AUROC; for both the real and synthetic datasets. We therefore focus only on the effects of date and age as the two most important predictor variables.

Table 3. Summary statistics on the variables analyzed ($n = 90\,514$ Ontario cases)

Variable	Mean (SD)	Proportion
Date reported (days since January 1, 2020)	214.43 (82.66)	
Gender		
Male		48.5%
Age group		
<20		11.2%
[20–29]		20.8%
[30–39]		15.5%
[40–49]		13.8%
[50–59]		14.7%
[60–69]		9.4%
[70–79]		5.3%
80+		9.3%
Exposure		
Travel related		3.4%
Close contact		40%
Outbreak		24.6%
Not reported		32%
% living in rural areas	6.98 (12)	
% of immigrants	37.04 (14.59)	
% of aboriginal population	1.64 (2.04)	
Prevalence of diabetes	7.73 (1.45)	
Prevalence of COPD	3.26 (1.38)	
Prevalence of high blood pressure	17.29 (2.2)	
Family medicine physicians per 100 000 Population	112.57 (102.5)	
Proportion reporting moderate-to-severe food insecurity	7.99 (1.81)	

Table 4. Mean model accuracy results for the real and synthetic datasets with the 95% bootstrap confidence interval

Accuracy metric	Real data	Synthetic data	CI overlap
AUROC	0.945 (0.941–0.948)	0.940 (0.936–0.945)	45.50%
AUPRC	0.340 (0.314–0.368)	0.313 (0.286–0.342)	52.02%

The confidence interval overlap between the real and synthetic CIs is also shown in the last column.

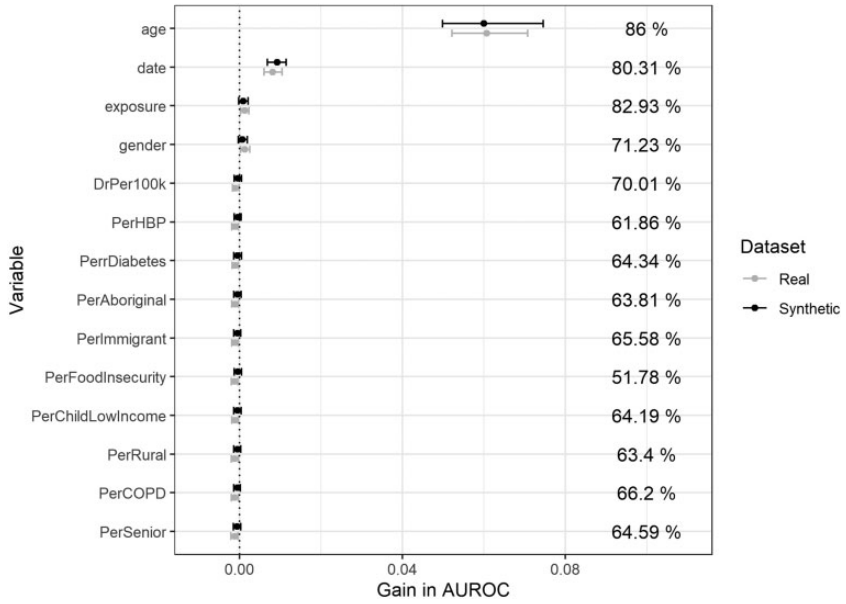

Figure 2. Variable importance using the permutation method with AUROC as the accuracy metric and the 95% bootstrap confidence interval. The values on the side are the confidence interval overlap values between the real and synthetic datasets.

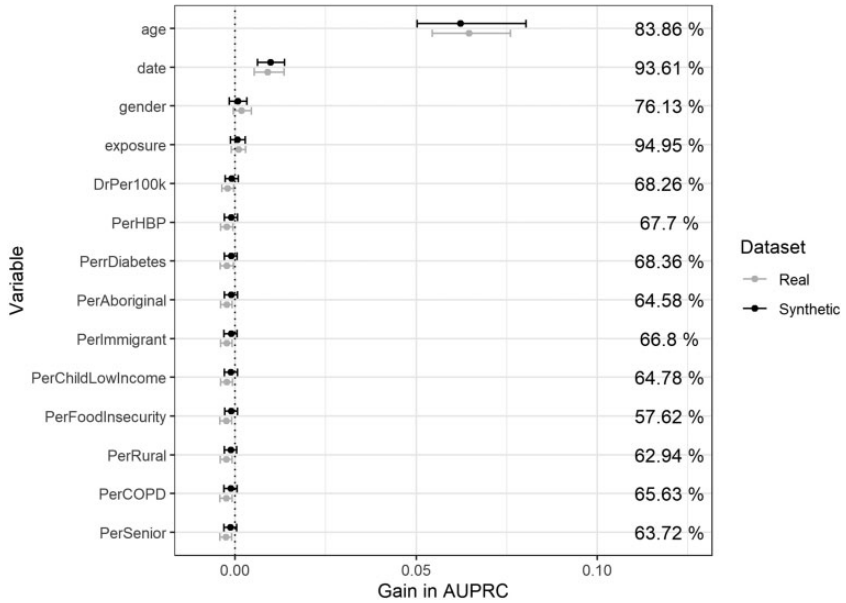

Figure 3. Variable importance using the permutation method with AUPRC as the accuracy metric and the 95% bootstrap confidence interval. The values on the side are the confidence interval overlap values between the real and synthetic datasets.

Figure 4 illustrates the conditional partial dependence observed across the date reported for the two GBT models. The 95% bootstrap confidence intervals for the synthetic data align well with those constructed from the real data. This indicates that the models produced using synthetic data will yield the same conclusions as those produced using the real data. This plot shows that, after factoring out other effects, this model captures an increasing probability of death over time, which does decrease and eventually plateau. There is an uptick that started at the tail end of the reporting period.

Figure 5 illustrates the conditional partial dependence observed across the age groups in the GBT models built using the real and synthetic datasets. The predicted probability of death increases monotonically with age group, with individuals greater than 80 years old having a mean predicted probability of death of 16.4% and 17.7% in the real and synthetic datasets, respectively. The GBT model built from synthetic data results in similar estimates, with a mean confidence interval overlap of 83.52% across all age groups, and the overlap for each age group exceeding our minimal threshold.

The sensitivity analysis results included in the appendix show that very similar logistic regression models would be constructed from the real and synthetic datasets, and the accuracy results are similar between the two and similar to the GBT model results.

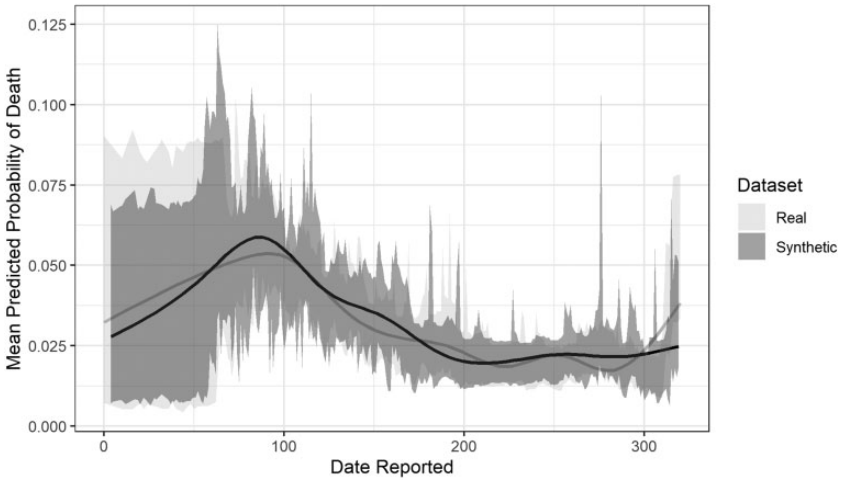

Figure 4. Conditional partial dependence plot for date reported with bootstrap confidence intervals on the real and synthetic datasets. The date reported is measured as the number of days since January 1, 2020.

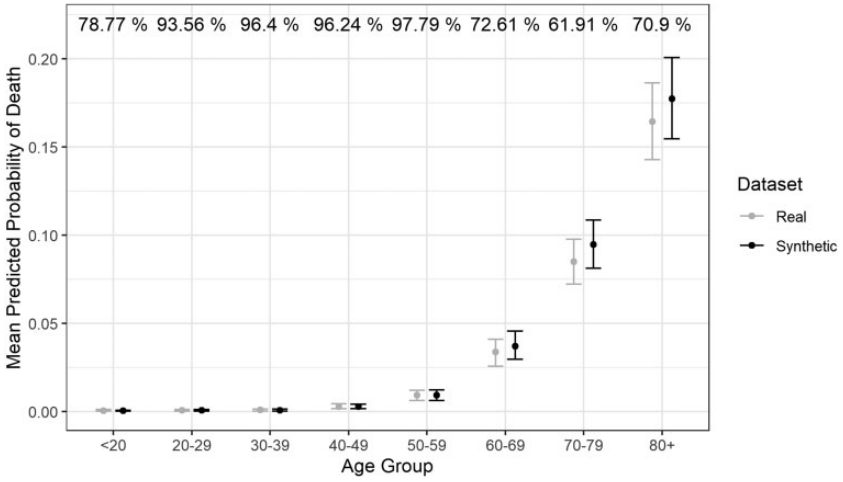

Figure 5. Conditional partial dependence plot for age with 95% bootstrap confidence intervals on the real and synthetic data. Confidence interval overlap is annotated at the top of the plot for each age group.

Distinguishability

The distinguishability between the real and synthetic datasets was 0.04 (on a scale from zero to one). This is quite low and indicates that the discriminator was not able to tell the difference between the real and synthetic datasets.

Privacy assessment

The probability of attribute disclosure conditional on identity disclosure for this dataset was 0.0585. This value is below the commonly used threshold of 0.09. This threshold has been recommended by the European Medicines Agency (EMA)[95] and Health Canada[97] for datasets to be considered to have a low risk of identification. The risk value for the original data was 0.3284. Therefore, the synthesis reduced this risk considerably.

For membership disclosure, the ability of an adversary to discriminate between a record that was used in synthesis (i.e. in the training dataset) and one that was not was evaluated using the standardized mean difference (SMD). The full dataset was split into a training dataset and a holdout, and the training dataset was synthesized. The distances between the training and synthesized dataset, and between the holdout and the synthesized dataset were computed. The SMD between the two distances was calculated at -0.063. This means that the distance between the training data and the synthetic data as slightly larger than holdout data and synthetic data. However, this value is below the commonly used 0.1 threshold which typically signifies a meaningful difference. This means that the likelihood of a successful membership disclosure is low. Further details about the methodology and justifications are provided in the appendix.

DISCUSSION AND CONCLUSIONS

Summary

We found that the analysis results between the real and synthetic datasets for the Ontario cohort of the Canadian COVID-19 case dataset were similar, and the conclusions from that analysis were the same. Gradient boosted classification trees were used to model the relationship between multiple factors and death. We found that age and the date since the start of 2020 were the biggest factors affecting the probability of death. These results are consistent with other reports from the literature.

We did not find a relationship between community characteristics associated with the public health regions where a case was reported and death (such as the percentage of immigrants and the percentage of individuals with diabetes, COPD, and high blood pressure). It is likely that such community-level measures over a large geographic region are not sufficiently associated with individual characteristics and therefore they are not sufficiently discriminatory with respect to the outcome in our models. This further emphasizes the importance of getting access to individual level data.

A sensitivity analysis performed to check for potential bias between the generator model and the analytic method did not reveal evidence of bias. Different types of logistic regression models produced consistent results between real and synthetic data, and with the GBT model.

A distinguishability test between the real and synthetic data found that a classifier was not able to effectively tell the difference between the real and synthetic datasets. This further supports the modeling results above.

A privacy evaluation of attribute disclosure conditional on identity disclosure, and of membership disclosure, showed that the privacy risks of the synthetic data were low.

Given the increasing pressures to get access to data and growing concerns about individual patient privacy risks that this presents, the data synthesis method presented in this paper can address the privacy concerns and we have presented some evidence that the conclusions drawn will be comparable to the original data.

A recent article also found that a synthesis method similar to the one used in this study produces datasets that have high utility.[98] In that case utility was defined as prediction accuracy for a number of different machine learning models. Our study goes further by comparing more robust accuracy measures, variable importance, and model interpretability. Furthermore, our study is the first to consider the utility of synthetic COVID-19 data. As the weight of evidence on the utility of synthetic data increases, one would expect there to be broader acceptance of using synthetic data as a proxy for real data.

Limitations

Although our study used sequential classification and regression trees for data synthesis, other methods could also have been used and may have produced comparable results. We did not evaluate the utility of multiple methods as that was not the objective of the current study, but rather it was to see if a common synthesis method could produce useful synthetic data. The current study can serve as a baseline (dataset and methods) for future work comparing multiple synthesis methods.

Our results were performed on the Ontario cohort of 90 514 records within the Canadian COVID-19 case dataset. Further analysis on more complex COVID-19 datasets, such as those including co-morbidities and socio-economic factors at the individual level, should be performed to add more weight to our findings and further assess the utility of synthetic data. The current study shows good potential that justifies additional effort to evaluate the utility of complex synthetic datasets.

AUTHOR CONTRIBUTIONS

KEE contributed to designing the study, performing the analysis, and contributed to writing the paper. LM contributed to designing the study, performing the analysis, and contributed to writing the paper. EJ performed the literature review. HS contributed to the design of the study and the writing of the paper.

ETHICS APPROVAL

This project was reviewed by the Children's Hospital of Eastern Ontario Research Institute Research Ethics Board as protocol number CHEOREB#20/89X.

ACKNOWLEDGMENTS

This research was enabled in part by support provided by Compute Ontario (computeontario.ca) and Compute Canada (www.computecanada.ca). We wish to thank Alison Paprica for providing us feedback on an earlier version of this paper. 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.

CONFLICT OF INTEREST STATEMENT

This work was performed in collaboration with Replica Analytics Ltd. This company is a spin-off from the CHEO Research Institute. KEE is co-founder and has equity in this company. LM is employed by Replica Analytics Ltd. and has equity in this company.

DATA AVAILABILITY

The dataset used in this study can be accessed from the Esri Canada COVID-19 resources website: https://resources-covid19canada.hub.arcgis.com/.

REFERENCES

1. Layne S, Hyman J, Morens D, Taubenberger J. New coronavirus outbreak: Framing questions for pandemic prevention. *Sci Transl Med*. 2020; 12 (534): eabb1469.
2. Downey M. Sharing data and research in a time of global pandemic. Duke University Libraries, 2020. https://blogs.library.duke.edu/bitstreams/2020/03/17/sharing-data-and-research-in-a-time-of-global-pandemic/. Accessed April 8, 2020.
3. Fazlioglu M. *Privacy in the Wake of COVID-19*. Portsmouth, NH: IAPP; 2020.
4. Walport M, Brest P. Sharing research data to improve public health. *Lancet* 2011; 377 (9765): 537–9
5. Chan M, Kazatchkine M, Lob-Levyt J, *et al*. Meeting the demand for results and accountability: a call for action on health data from eight global health agencies. *PLOS Med*. 2010; 7 (1): e1000223.
6. Hajduk GK, Jamieson NE, Baker BL, Olesen OF, Lang T. It is not enough that we require data to be shared; we have to make sharing easy, feasible and accessible too!. *BMJ Glob Health*. 2019; 4 (4): e001550.
7. Adly AS, Adly AS, Adly MS. Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of COVID-19: scoping review. *J Med Internet Res*. 2020; 22 (8): e19104.
8. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; 319 (13): 1317–8.
9. #opendata4covid19. Ministry of Health and Welfare; Health Insurance Review & Assessment Service (HIRA). 2020. https://hira-covid19.net/. Accessed April 8, 2020.
10. #opendata4covid19 Website User Manual. Ministry of Health and Welfare; Health Insurance Review & Assessment Service (HIRA). 2020, https://rtrod-assets.s3.ap-northeast-2.amazonaws.com/static/tools/manual/COVID-19+website+manual_v2.1.pdf. Accessed April 8, 2020.

11. National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocaa196/5893482?login=true (Accessed January 16, 2021).
12. COVID-19 Updates Page – OHDSI. https://www.ohdsi.org/covid-19-updates/. Accessed January 16, 2021.
13. The Ontario Health Data Platform (OHDP). https://computeontario.ca/covid-19-health/. Accessed September 24, 2020.
14. van Panhuis WG, Paul P, Emerson C, *et al*. A systematic review of barriers to data sharing in public health. *BMC Public Health* 2014; 14 (1): 1144.
15. Kalkman S, Mostert M, Gerlinger C, van Delden JJM, van Thiel GJMW. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Med Ethics* 2019; 20 (1): 21.
16. Park S, Choi GJ, Ko H. Information technology–based tracing strategy in response to COVID-19 in South Korea—privacy controversies. *JAMA* 2020; 323 (21): 2129.
17. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nat Med* 2020; 26 (4): 463–4.
18. Lewis P, Conn D, Pegg D. UK government using confidential patient data in coronavirus response. *The Guardian*, April 12, 2020.
19. Zastrow M. South Korea is reporting intimate details of COVID-19 cases: has it helped? *Nature* 2020. doi:10.1038/d41586-020-00740-y. https://www.nature.com/articles/d41586-020-00740-y
20. Rocha R. The data-driven pandemic: Information sharing with COVID-19 is 'unprecedented'. CBC News, Canada, March 17, 2020.
21. Rackley K. DHEC, state authorities address privacy issues, information about coronavirus case specifics. Aiken, SC, USA/; *Aiken Standard*. 2020.
22. Hinkle J. Framingham one of several cities and towns told by DPH to limit information about residents who test positive for coronavirus. Wicked Local - News, March 28, 2020. https://www.wickedlocal.com/news/20200328/framingham-one-of-several-cities-and-towns-told-by-dph-to-limit-information-about-residents-who-test-positive-for-coronavirus. Accessed April 8, 2020.
23. McCallum A. *Janesville and Rock County Officials Clash Over Sharing of COVID-19 Information*. Janesville, WI: GazetteXtra, 2020.
24. Hancock L. Ohio health director cites privacy concerns as local health departments withhold coronavirus details. cleveland.com, April 3, 2020. https://www.msn.com/en-us/news/us/ohio-health-director-cites-privacy-concerns-as-local-health-departments-withhold-coronavirus-details/ar-BB128Ztr. Accessed April 8, 2020.
25. Hill K. *Spokane Health Officials Providing More Information About COVID-19 Patients, But It Remains Unclear Where They're Being Treated*. Spokan, WA: The Spokesman-Review, April 6, 2020.
26. Malin BA, Emam KE, O'Keefe CM. Biomedical data privacy: problems, perspectives, and recent advances. *J Am Med Inform Assoc* 2013; 20 (1): 2–6.
27. Jones J, Meyer P, Garrison C, Kettinger L, Hermann P. Physician and infection control practitioner HIV/AIDS reporting characteristics. *Am J Public Health* 1992; 82 (6): 889–91.
28. Konowitz P, Petrossian G, Rose D. The underreporting of disease and physicians' knowledge of reporting requirements. *Public Health Rep* 1984; 99 (1): 31–5.
29. Marier R. The reporting of communicable diseases. *Am J Epidemiol*. 1977; 105 (6): 587–90.
30. AbdelMalik P, Boulos MNK, Jones R. The perceived impact of location privacy: a web-based survey of public health perspectives and requirements in the UK and Canada. *BMC Public Health* 2008; 8 (1): 156.
31. Drociuk D, Gibson J, Hodge J. Health information privacy and syndromic surveillance systems. *MMWR* 2004; 53: 221–225.
32. Emam KE, Mercer J, Moreau K, Grava-Gubins I, Buckeridge D, Jonker E. Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak. *BMC Public Health* 2011; 11 (1): 454.
33. Emam KE, Hoptroff R. The synthetic data paradigm for using and sharing data. *Cutter Executive Update* 2019; 19 (6).
34. El Emam K, Mosquera L, Hoptroff R. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. Sebastopol, CA: O'Reilly; 2020.
35. Reiter JP. New approaches to data dissemination: a glimpse into the future. *CHANCE* 2004; 17 (3): 11–5.
36. Jules P, Elizabeth R. 10 Privacy Risks and 10 Privacy Technologies to Watch in the Next Decade. Washington, DC: Future of Privacy Forum, January 2020.
37. Wang Z, Myles P, Tucker A. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility Patient Privacy. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), June 2019, pp. 126–131, doi:10.1109/CBMS.2019.00036.
38. Synthetic data at CPRD. https://www.cprd.com/content/synthetic-data. Accessed September 24, 2020.
39. N3C. Synthetic Data Workstream | N3C. https://covid.cd2h.org/N3C_synthetic_data. Accessed September 24, 2020.
40. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. *Proc Vldb Endow* 2018; 11 (10): 1071–83..
41. Hu J. Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data. arXiv:1804.02784 [stat], 2018. http://arxiv.org/abs/1804.02784. Accessed March 15, 2019.
42. Taub J, Elliot M, Pampaka M, Smith D. Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer J, Francisco Montes, eds. *Priv Stat Databases. Lecture Notes in Computer Science*. Vol. 11126. 2018:122–37.
43. Hu J, Reiter JP, Wang Q. Disclosure risk evaluation for fully synthetic categorical data. In: Domingo-Ferrer, J. (eds) *Priv Stat Databases*. Lecture Notes in Computer Science. Vol. 8744. Cham: Springer; 2014: 185–99. .
44. Wei L, Reiter JP. Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Statis J IAOSJI* 2016; 32 (1): 93–108.
45. Ruiz N, Muralidhar K, Domingo-Ferrer J. On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In: Domingo-Ferrer J, Francisco Montes, eds. *Priv Stat Databases*. Lecture Notes in Computer Science. Vol. 11126. 2018: 59–74.
46. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J Royal Statistical Soc A* 2005; 168 (1): 185–205.
47. Emam KE, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *JMIR* 2020; 22 (11): e23139.
48. El Emam K. Seven Ways to Evaluate the Utility of Synthetic Data. IEEE Security and Privacy, 2020;18 (4): 56–9.
49. Gao Y, Cai G-Y, Fang W, *et al*. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun* 2020; 11 (1): 5033.
50. Esri C. "COVID-19 Canada." https://resources-covid19canada.hub.arcgis.com/. Accessed September 24, 2020.
51. Your Health System: In Depth. Canadian Institute for Health Information, May 2020. https://yourhealthsystem.cihi.ca/hsp/indepth?lang=en&_ga=2.167615550.1636213197.1614266119-118357212.1614266118#/. Accessed October 2, 2020.
52. Wynants L, Van Calster B, Collins G, *et al*. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.
53. Harper PR, Moore JW, Woolley TE. Covid-19 transmission modelling of students returning home from university. *Health Systems* 2021; 1–10. doi:10.1080/20476965.2020.1857214.
54. Caramelo F, Ferreira N, Oliveiros B. Estimation of risk factors for COVID-19 mortality - preliminary results. medRxiv; : 2020.02.24.20027268. doi:10.1101/2020.02.24.20027268.
55. Barda N, Riesel D, Akriv A, *et al*. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat Commun* 2020; 11 (1): 4439. doi:10.1038/s41467-020-18297-9.
56. Kaur D, Sobiesk M, Patil S, *et al*. Application of Bayesian networks to generate synthetic health data. *J Am Med Informatics Assoc* 2020; Dec 23;ocaa303. doi:10.1093/jamia/ocaa303.
57. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *Npj Digit Med* 2020; 3 (1): 147. doi:10.1038/s41746-020-00353-9.

58. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X. PrivBayes: private data release via bayesian networks. *ACM Trans Database Syst* 2017; 42 (4): 1–25:41.
59. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J, Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv:1703.06490 [cs]. 2017. Available: http://arxiv.org/abs/1703.06490. Accessed September 1, 2020.
60. Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 2020; 27(1):99–108. doi:10.1093/jamia/ocz161.
61. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput Statist Data Anal* 2011; 55 (12): 3232–43.
62. Arslan RC, Schilling KM, Gerlach TM, Penke L. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *J Pers Soc Psychol* 2018; Aug 27. doi:10.1037/pspp0000208.
63. 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 Effect* 2019; 12 (4): 616–47.
64. Sabay A, Harris L, Bejugama V, Jaceldo-Siegl K. Overcoming small data limitations in heart disease prediction by using surrogate data. *SMU Data Science Rev* 2018; 1 (3): 12. https://scholar.smu.edu/datasciencereview/vol1/iss3/12.
65. Freiman M, Lauger A, Reiter J. Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau. US Census Bureau, Working paper, 2017.
66. Nowok B. Utility of synthetic microdata generated using tree-based methods. 2015. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_33_Session_2_-_Univ._Edinburgh__Nowok_.pdf. Accessed September 1, 2020.
67. Raab GM, Nowok B, Dibben C. *Practical Data Synthesis for Large Samples* 2016; 7 (3): 67–97. doi:10.29012/jpc.v7i3.407.
68. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1. *Statis J IAOS* 2017; 33 (3): 785–96.
69. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 2020; 9: e53275; doi:10.7554/eLife.53275.
70. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. Boca Raton: Taylor & Francis, 1984.
71. Reiter J. Using CART to generate partially synthetic, public use microdata. *J Official Stat* 2005; 21 (3): 441–62.
72. Mark E. Final Report on the Disclosure Risk Associated with the Synthetic Data produced by the SYLLS Team. Manchester University, 2014. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf. Accessed September 1, 2020.
73. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification in Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg,2009, pp. 254–269, doi:10.1007/978-3-642-04174-7_17.
74. Spyromitros-Xioufis E, Tsoumakas G, Groves W, Vlahavas I. Multi-target regression via input space expansion: treating targets as inputs. *Mach Learn* 2016; 104 (1): 55–98.
75. El Emam K, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. *J Am Med Informatics Assoc* 2020; 28 (1): 3–13.
76. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors. *Ann Statist* 2000; 28 (2): 337–407.
77. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018.
78. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. *Am Statist* 2006; 60 (3): 224–32.
79. Niculescu-Mizil A, Caruana RA. Obtaining Calibrated Probabilities from Boosting. arXiv:1207.1403 [cs, stat], July 2012, http://arxiv.org/abs/1207.1403. Accessed October 21, 2020.
80. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning; 2001: 609–16.
81. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction, 1 Edition*. Oxford: Oxford University Press, 2004.
82. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning, New York, NY, USA, 2006, pp. 233–240, doi:10.1145/1143844.1143874.
83. Boyd K, Santos Costa V, Davis J, Page CD. Unachievable region in precision-recall space and its effect on empirical evaluation. *Proc Int Conf Mach Learn* 2012; 2012: 349.
84. Breiman L. Random forests. *Machine Learn* 2001; 45 (1): 5–32.
85. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 2019; 20 (177): 1–81.
86. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformat* 2007; 8 (1): 25.
87. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* 2010; 11 (1): 110.
88. Hooker G, Mentch L. Please Stop Permuting Features: An Explanation and Alternatives. arXiv:1905.03151 [cs, stat], May 2019, http://arxiv.org/abs/1905.03151. Accessed October 19, 2020.
89. Mentch L, Hooker G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J Machine Learning Res* 2016; 17 (26): 1–41.
90. Molnar C, König G, Bischl B, Casalicchio G. Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach. arXiv:2006.04628 [cs, stat], Jun. 2020. http://arxiv.org/abs/2006.04628. Accessed November 6, 2020.
91. Jerome F. On Multivariate Goodness-of-Fit and Two-Sample Testing. Stanford University, 2003. Available: http://statweb.stanford.edu/~jhf/ftp/gof. Accessed on October 2, 2020.
92. Hediger S, Michel L, Näf J. On the Use of Random Forest for Two-Sample Testing. arXiv:1903.06287 [stat], April 2019, http://arxiv.org/abs/1903.06287. Accessed May 6, 2020.
93. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. *J R Stat Soc A* 2018; 181 (3): 663–88.
94. Woo M-J, Reiter JP, Oganian A, Karr AF. Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *J Priv Confidentiality* 2009; 1 (1). doi: 10.29012/jpc.v1i1.568.
95. 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. 2017. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf. Accessed October 1, 2019.
96. European Medicines Agency. European Medicines Agency policy on publication of data for medicinal products for human use: Policy 0070. October 2, 2014, [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf. Accessed September 1, 2020.
97. Health Canada. Guidance document on Public Release of Clinical Information. April 1, 2019. https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html. Accessed September 1, 2020.
98. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Med Inform* 2020; 8 (7): e18910.