Enabling COVID-19 Data Access Using Data Synthesis

Khaled El Emam[1,2,3], Harpreet Sood[4 5]

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

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

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

[4] London School of Economics, London, UK

[5] National Health Service, London, UK

Contact Information:

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

E: kelemam@ehealthinformation.ca

1. Background

COVID-19 has created a need for an unprecedented level of data sharing with researchers, health care providers, and public health organizations [1–3]. However privacy concerns have historically acted as a barrier to local and global sharing of public health data [4,5]. Data synthesis can enable access to COVID-19 data while still protecting patient privacy. Because there is no one-to-one mapping between synthetic data and real individuals, the privacy risks are limited[6]. In this position statement we address the question of whether synthetic COVID-19 data is useful ? We present an analysis of COVID-19 deaths in Canada performed on both real and synthetic data, and comparing the results. If we are able to obtain similar results and draw the same conclusions from real and synthetic data, then that adds to the weight of evidence that synthetic data can be analytically useful.

2. Methods

A dataset was obtained reflecting Canadian COVID-19 cases until mid-June 2020. Because the values were incomplete for some provinces, our analysis focused only on Ontario with 32,917 records. The fields in that dataset are shown in Table 1.

Variable	Definitions
Date Reported	Number of days since 1 January 2020
Health Region	
Age Group	Decades from 20 to 80+ (ordinal)
Gender	
Exposure	close contact, outbreak, travel, community, no epi link
Case Status	recovered, deceased, active

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

We used a sequential synthesis method based on machine learning techniques to create the synthetic variant of the dataset [7]. A logistic regression model with death as an outcome was then fitted on the real and synthetic datasets The predictors were date (represented as the number of days since January 2020), age group, and gender.

3. Results

The odds ratio values for the logistic regression model which predicts death are shown in Table 2. The parameter values are directionally as expected, with age having a large positive impact on the likelihood of death, and males being more likely to die than females. The likelihood of death has decreased over time. The conclusions are the same for the real and synthetic datasets.

Variable	Real Data	Synthetic Data
Age	3*	2.89*
Gender (Male)	1.7*	1.53*
Date Reported	0.9*	0.93*

Table 2: The logistic regression odds ratios for the three variables that were entered in the model for the real and synthetic datasets. An asterisk indicates significance at an alpha value of 0.05.

Figure 8 is a plot showing the 95% confidence intervals for the odds ratios of the logistic regression model and the confidence interval overlap. Confidence interval overlap is over 40%, indicating that the differences between the parameter estimates are not large.

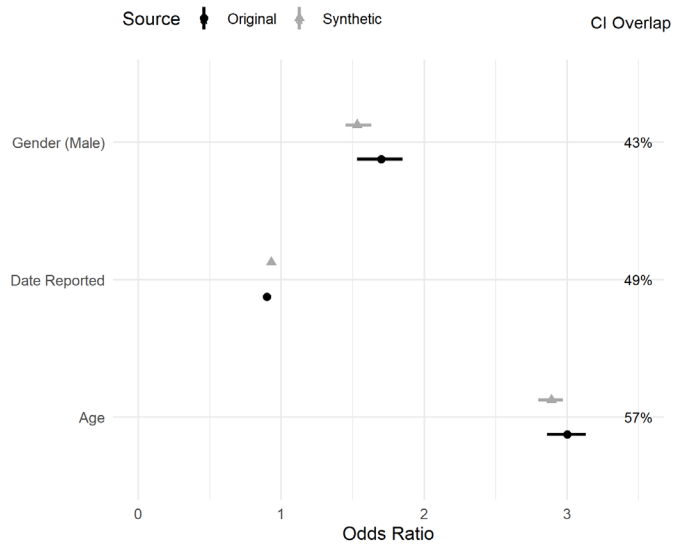

Figure 1: A forest plot showing the differences between the 95% confidence intervals for the odds ratio between the real and synthetic datasets.

4. Discussion and Conclusions

Our results indicate that the characteristics of the data and 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.

5. References

1. Layne, S., Hyman, J., Morens, D. & Taubenberger, J. New coronavirus outbreak: Framing questions for pandemic prevention. *Science Translational Medicine* **12**, (2020).
2. Downey, M. Sharing data and research in a time of global pandemic. *Duke University Libraries* https://blogs.library.duke.edu/bitstreams/2020/03/17/sharing-data-and-research-in-a-time-of-global-pandemic/ (2020).
3. Ng, A. Coronavirus pandemic changes how your privacy is protected. *CNET* https://www.cnet.com/news/coronavirus-pandemic-changes-how-your-privacy-is-protected/ (2020).
4. Panhuis, W. G. van *et al.* A systematic review of barriers to data sharing in public health. *BMC Public Health* **14**, 1144 (2014).
5. Kalkman, S., Mostert, M., Gerlinger, C., van Delden, J. J. M. & van Thiel, G. J. M. W. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Medical Ethics* **20**, 21 (2019).
6. Khaled El Emam, Lucy Mosquera & Jason Bass. A Method for Evaluating Idendity Disclosure Risk in Fully Synthetic Data. *(submitted for publication)* (2020).
7. Khaled E Emam, Lucy Mosquera & Mina Zheng. Optimizing the Synthesis of Clinical Trial Data Using Sequential Trees. *Journal of the American Medical Informatics Association* **(accepted)**, (2020).