Assessing the fitness for use of real-world electronic health records and log data with and without the application of privacy preserving technologies

Jason A. Thomas

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Adam Wilcox, Chair

Gang Luo

Matthew Trunnell

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Assessing the fitness for use of real-world electronic health records and log data with and without the application of privacy preserving technologies

Jason A. Thomas

Chair of the Supervisory Committee:

Professor Adam Wilcox, PhD

Biomedical Informatics and Medical Education

Over the past decade, electronic health record (EHR) adoption has led to an explosion in the volume of electronic health record and log data, then efforts to effectively harness the potential of these data for knowledge discovery (KD) and quality improvement (QI). In parallel, recent gains in artificial intelligence have produced powerful methods to analyze, use, and even create synthetic data which are statistically or mathematically reflective of real data yet are generated by a computer algorithm. However, limitations in data utility (e.g. bias, data quality, comprehensiveness) and accessibility (e.g. privacy, interoperability, availability), as well as limited means to measure and manage tradeoffs between the two are significant barriers to using these data effectively. Determining whether data are suitable to be used in a specific analysis or

context, known as “fitness for use” is not included in current frameworks for general health record data quality characterization nor evaluated by data quality assessment (DQA) tools. EHR log data use is particularly unrefined for QI and KD due to an absence of validated standards and methods. Thus, users of electronic health record and log data remain uninformed as to the fitness for use of their data at baseline and are unable to effectively assess subsequent tradeoffs between utility and privacy when applying privacy preserving technologies.

To address these challenges, we sought to assess the fitness for use of electronic health record and log data - both synthetic and real - across three use cases. First, we 1) developed a framework for data utility assessment of electronic health records, then 2) adapted open-source tools to make use of this framework which we then applied to assess the utility of real and synthetic EHR data for observational research related to COVID-19 and future influenza pandemics. Second, we evaluated whether synthetic data derived from a national COVID-19 data set could be used for geospatial and temporal epidemic analyses. To do so we conducted replication studies and computed general summary statistics on original and synthetic data, then compared the similarity of results between the two datasets. Third, we conducted a retrospective, observational analysis - with and without privacy preserving technology - of clinical workstation authentication behaviors from the UW Medicine health system to inform customized solutions that balance usability and security.

The three use cases studied advance our understanding of 1) the fitness for use of varied electronic health record and clinical workstation log data with and without privacy preserving technologies as well as 2) methods to conduct these assessments. As the use of synthetic data rises, so will the importance of fitness for use assessments on both original and synthetic data. Synthetic data that are broadly distributed will reach less expert users than those who have

access to the original data. Thus, in addition to helping those creating synthetic data manage tradeoffs, fitness for use assessments will provide guidance to synthetic data end-users on 1) the approximate similarity between the synthetic data and the original data as well as 2) the overall limitations of the likely inaccessible (to the end-users, at least at the time of analyzing the synthetic data) original data which have a downstream effect on the synthetic data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Melanie for all her support over these nearly four years. Adam Wilcox for his academic mentorship but also for being someone who I could trust to look out for my well-being, to be fair with me and others. Additionally, for his good-natured banter during our meetings about camping trips, sailing, working on cars and all sorts of other topics that made this journey all the more enjoyable. Matthew Trunnell for being so welcoming, generous with his time and energy, and for continuing to provide insight towards the next steps that I take. Gang Luo for his part (along with Adam) in coming up with the idea that led to Aims 1 & 2 of this dissertation. Franzi Roesner for stellar instruction then encouragement in my pursuit of completing aim 3. Cris Ewell for all his help with the conception of aim 3 and for providing access to all the data and personnel needed. Larry Kessler for helping me progress as a writer while clearly going the extra mile in the course he taught and for agreeing for be my GSR. My department (classmates, faculty & staff) for fostering a friendly, supportive environment. Andrew Teng and Calvin Apodaca of my 2017 cohort, who most closely shared this experience with me. My department chair, Peter Tarczy-Hornoch, for explicitly encouraging trainees to take all their vacation days, which I did years 1 & 2, leading to a much better work-life balance than I suspect students at most other programs have. The OHDSI community for outstanding academic open-source software development, open research, and guidance as I sought to gain expertise working with the OMOP CDM. The N3C for providing challenging and important work, highly valuable scientific feedback, and a platform to disseminate my research. The National Library of Medicine for three years of funding. Last, my family, friends, and specifically Steph & Matt, for their love and/or support along the way.

Chapter 1. INTRODUCTION

1.1 SIGNIFICANCE OF THE PROBLEM

The past decade has seen dramatic growth in the adoption and use of electronic health records in the United States due to the Meaningful Use federal incentive program.[1–7] Additional investments in expanding various secondary use activities of EHR data have followed, from the Patient Centered Outcomes Research Network (PCORNet) through the Electronic Medical Records and Genomics Network (eMERGE), All of Us Research program[8–17] and various organizations developing standards for both electronic health records and log data.[18,19] Both the National Library of Medicine (NLM) and the National Institutes of Health (NIH) have highlighted the value of data from EHRs in strategic planning, with the latter identifying that EHR "records present great opportunities for advancing medical research and improving human health—particularly in the area of precision medicine"[20], and the former stating the "NLM must create the controls for effective stewardship of data generated in the course of clinical care" and that "aggregated collections of such data, properly curated, will enable analyses of subpopulations based on aspects of their medical care and on demographic characteristics such as gender, age, race, and ethnicity."[21] However, EHR data remain challenging for secondary use due to barriers associated with utility (quality, bias, comprehensiveness)[22–31] and accessibility (e.g. privacy, interoperability, availability)[32] as well as limited means to measure and manage tradeoffs between the two. Data quality and barriers to access and sharing have been identified as two of the top five "Challenges Hindering the Use of Machine Learning in Drug Development" in a 2019 technology assessment conducted by the National Academy of Medicine and Science, Technology Assessment, and Analytics U.S. Government Accountability Office.[32]

At the same time, data mining, predictive analytics, and other artificial intelligence and advanced analytics methods have generated promise and hope in their applications to healthcare, that they can advance both the value and methods of EHR data use.[15,33–44] The result has been an emphasis on data science, which the NLM defines as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."[21,45] However, there have also been challenges in applying data mining and data science methods to EHR data.[46] Research publications using deep learning methods in healthcare have greatly increased since 2010, yet the number of studies using EHR data is disproportionately less in medical informatics than in other areas such as imaging, bioinformatics, public health and sensors.[43,46] There are 5 to 10 times fewer deep learning publications in medical informatics than any of the other listed fields in health informatics, illustrating the challenges EHR data present for analysis. Nascent efforts to harness EHR log data for tasks such as characterizing and reducing IT burden on healthcare providers have faced similar challenges, spurring the creation of a National Research Network for EHR Audit-log and Meta-data[19].

Overcoming data sharing barriers is critical for advancing secondary use of EHR data, since without sharing all analyses must be done within healthcare organizations where data are initially collected for care and the amount of data available for analysis is limited in size and diversity. The primary barriers are related to effectively protecting privacy and confidentiality while sharing data (84). While sharing of data across institutions is possible, the required agreements and regulatory reviews are sufficiently difficult to hamper most sharing efforts[47], especially when working to produce "generalizable knowledge" opposed to conducting "quality assessment and improvement activities."[48] De-identification of data has been used as a

primary method to reduce some of this burden, but is increasingly limited.[47,49–52] Newer privacy preserving technologies - some of which seek to achieve formal epsilon-differential privacy[53] - have gained traction as an alternative to traditional de-identification methods both inside and outside of medicine.[54–57] Differential privacy promises that "the analyst knows no more about any individual in the data set after the analysis is completed than she knew before the analysis was begun."[58] One promising privacy preserving technology that can be used with EHR data is the creation of synthetic data.[56,59–74] Synthetic data conserve the statistical distributions of the real data they are modeled on yet prevent the two from being linked together because synthetic data rows are not tied to paired rows within its source data. Synthetic data have been generated for a wide variety of biomedical use cases[54,58,75–79] and can be made differentially private by adding noise during the training of deep learning models used in the generative process.[57] The major challenge in differential privacy approaches with synthetic data is that the amount of noise introduced must be calibrated in a tradeoff of accuracy and utility of the data against privacy. This need to determine whether synthetic data are "fit for use" as substitutes to real data has recently been identified as a pressing need by the National Library of Medicine.[80]

The potential benefit of applying privacy preserving methods to data in medicine extends beyond patients. Increased interest in worker data in and outside of medicine[81–83] has brought with it concerns for balancing the use of these data in alignment with worker and employer interests. A 2018 international survey of 1,400 C-level executives and 10,000 workers reported that "62% of businesses are using new technologies and sources of workforce data today but only 30% of these leaders are confident that they are using new sources of workforce data in a highly responsible way."[84] The same survey showed that workers have concerns about use of

their data yet are willing to share their data in exchange for benefits, listing a customized work experience as the number one desired benefit.

Real-World Data utility issues encompassing quality, bias, and comprehensiveness limit the effective use of EHR data for knowledge discovery and quality improvement. In a 2017 survey of over 16,000 data scientists, “dirty data” was listed as the number one barrier faced at work[85] and a separate survey in 2016 of data scientists found that their least enjoyable task at work was collecting, cleaning, and organizing data yet those tasks took up nearly 80% of their time.[86] Within healthcare, data quality issues can undermine use of EHRs for knowledge discovery, precision medicine, comparative effectiveness research, and other research using secondary data analysis.[20,32,47,49,87–91] Bias in datasets used to train AI outside of healthcare are well documented and have resulted in high-profile public relations issues.[92,93] Bias in EHR data is also a major issue for its secondary use. Some examples of bias are that: sicker patients have more complete data[94], vulnerable populations have a higher probability of visiting multiple health care system for care[95–97], patients of lower socioeconomic status have limited healthcare access and are less likely to receive diagnoses and medications[95,98], female patients receive less aggressive coronary revascularization approaches compared to men[99]. Comprehensiveness of the EHR is limited in multiple ways by insufficient granularity or lack of data capture - especially in structured data - of social determinants of health[100,101], nutrition and exercise data[102,103], patient data from wearable health technology[104], and more. EHR log data use suffers from a lack of data standards and methodological transparency.[105]

While privacy preserving technologies provide opportunity to increase the accessibility of data by mitigating privacy concerns of data sharing, they simultaneously degrade the utility of these data. Formal epsilon- differential privacy degrades data quality by adding noise.[58]

Synthetic data reduce comprehensiveness by struggling to capture longitudinal relationships continuous variables.[72] Both technologies perform worse on smaller groups of data[76,106] which increases bias and may exacerbate preexisting health disparities.

Paradoxically, synthetic data generation can be used in non-privacy-related applications as a solution for, rather than a cause of, bias in datasets. To do so, synthetic data are injected into real data as augmentation to reduce bias and provide balancing of data. A hybrid dataset results. Hybrid datasets may reduce known health disparities[107] by reducing the bias of data used for training AI. In a recent study, hybrid data greatly improved classification of skin lesions from underrepresented patients with skin of color at a small cost in performance overall.[108] Similar hybrid data methods could be applied to EHR data. However, the means to assess biases within datasets is a critical prerequisite to adjusting for them with synthetic data.

Recent work has been done to characterize and assess dimensions of EHR data utility, leading to open tools to make these assessments. Electronic health record data quality characterization[109–111] has matured to yield a harmonized terminology and framework[110] for describing EHR data quality. Multiple organizations have produced DQA methods[112–115] adhering to this framework that assess common data model conformance and a limited number of overall data quality checks against rules such as "birthdate prior to 1850."[113] DQA has been conducted across distributed research networks[116–121]. Limited, if any, open tools currently exist to assess bias and comprehensiveness of electronic health records nor EHR audit-log data.

Publicly available, synthetic data quality has rarely been assessed for its representativeness of the population it was modeled on yet doing so is becoming more common. Chen et al. assessed limited clinical quality measures (n=4) within the 1.2 million person, Massachusetts-modeled Synthea dataset[59], finding that it performs well on demographics and

services offered, but poorly on heterogenous outcomes[122]. Zhang et al. evaluated the similarity between its limited synthetic data generated compared to real data it was modeled on for the distribution of diagnosis codes, interdimensional relationships of data, local data structure, and latent factorized representations of data.[72] A multi-site study assessed the similarity of five retrospective, observational EHR studies conducted on both real and synthetic data.[106] In the last couple years, more synthetic data have been assessed for utility[125-129] using a variety of the methods by which one can validate synthetic data.[130]

While each of these guidelines, tools and examples can identify and address specific data utility issues, they do not assess overall data utility in terms of fitness for use with extrinsic context in mind to enable knowledge discovery or other research. Instead, the tools that exist are data quality-focused and primarily measure atemporal plausibility and conformance.[123] Current best practices that do address the need for fit for use within a DQA context includes an assessment of data elements' quality for the specific goal of novel heart failure biomarker discovery.[124] Electronic health record log data have been identified as having great potential to improve health services research yet suffer from many similar data utility issues and a lack of methods to assess their fitness for use.[131] Ultimately, the literature is relatively rich with recently developed methods for assessments of data quality yet there is a dearth of methods to assess the fitness for use of electronic health records and log data before and after applying privacy preserving technologies.

1.2 Statement of The Study Purpose

Although there has been work done on data quality assessment of electronic health records, little work has been done to enable fitness for use assessments of electronic health records or EHR log data. Because the data quality and fitness for use issues are similar and both data are available

electronically due to EHRs, we study both data types. COVID-19 has increased the use and sharing of real-world EHR data for observational research, which heightens the importance of ensuring the data used are good enough for the task at hand and that patient and healthcare worker data is managed in a responsible way that protects their privacy. Emerging privacy-preserving technologies provide an opportunity to analyze and/or share clinical data while maintaining privacy yet these methods degrade the utility of the data which increases the importance of assessing synthetic and original data fitness for use. This research will attempt to enable the Data Utility Assessment (DUA) of EHR data - both records and logs - before and after the application of privacy preserving technologies. In this study, there are three main objectives:

Aim 1: Collect and curate a repository of clinical facts as raw input needed for our health records DUA framework.

Aim 2: Inform, develop, and evaluate a DUA framework and DUA tool to support secondary use of EHR data - synthetic and real - across diverse contexts, using standards.

Aim 3: Assess SSO authentication behaviors to inform customized solutions using log data - while maintaining privacy and promoting standards.

1.3 Content of the Dissertation

Our work is spread out across three separate use cases described in chapters 2-4. Each of these chapters includes an independent analysis that enables and/or conducts fitness for use assessments of clinical or log data - with and without the application of privacy preserving technology. Chapters 2 and 3 contain analyses on electronic health records whereas chapter 4 contains an analysis on clinical workstation authentication log data.

In Chapter 2 (the first use case), we developed and tested a new framework by which one can determine whether electronic health records are fit for use and also assessed the impact of privacy preserving technology on fitness for use. To do so, we made use of the Cochrane Database of Systematic Reviews (CDSR). We built up a repository of outcomes from the CDSR as individual research findings to be replicated within electronic health records formatted in the OMOP CDM. Due to the present relevance of the COVID-19 pandemic and its implications on observational research and disease surveillance going forward, we focused our replications on COVID-19 related outcomes.

In Chapter 3 (the second use case), we described methods and results focused on evaluating whether synthetic N3C data can be used for geospatial and temporal epidemic analyses. Our replication studies focused on what we deemed were important and common analyses to be performed, such as epidemic curves for key indicators and creation of public-facing dashboards. Our validation included replication of studies and general utility metrics for: analyses at the zip code level over time, construction of epidemic curves, and aggregate population characteristics. We believe these approaches balance the need to provide broad utility results for a wide range of analyses while also providing specific validation results relevant to analyses of common interest.

In Chapter 4 (the third use case), we used Imprivata Onesign SSO log data from the UW Medicine Health system (Seattle, WA USA) - comprised of a Trauma Level 1 hospital, academic medical center, and outpatient clinics - to inform customized SSO authentication protocols and report on the utility of observational SSO log data to do so. In addition to characterizing SSO behaviors broadly, we considered two potential SSO implementation changes and their simulated impacts stratified by user role and location. The first potential

change was variation of the challenge period from 1-12 hours in 1-hour increments. The second was requiring a challenge for each new workstation a user logs into with and without incorporating their prior workstation access history. Additionally, we piloted the creation and use of synthetic SSO log data to re-create portions of our analysis in an effort to protect the privacy of worker data.

In Chapter 5, we summarize the findings of the dissertation. In addition, we outline future directions for the research.

1.4 REFERENCES FOR CHAPTER 1

1 Office of the National Coordinator for Health Information Technology. "Percent of Hospitals, By Type, that Possess Certified Health IT," Health IT Quick-Stat #52. Office of the National Coordinator for Health Information Technology 2018. dashboard.healthit.gov/quickstats/pages/certified-electronic-health-record-technology-in-hospitals.php (accessed 29 Nov 2018).

2 Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. *N Engl J Med* 2010;**363**:501–4. doi:10.1056/NEJMp1006114

3 Adler-Milstein J, Everson J, Lee S-YD. Sequencing of EHR adoption among US hospitals and the impact of meaningful use. *J Am Med Inform Assoc JAMIA* 2014;**21**:984–91. doi:10.1136/amiajnl-2014-002708

4 Lin SC, Jha AK, Adler-Milstein J. Electronic Health Records Associated With Lower Hospital Mortality After Systems Have Time To Mature. *Health Aff Proj Hope* 2018;**37**:1128–35. doi:10.1377/hlthaff.2017.1658

5 Furukawa MF, Eldridge N, Wang Y, *et al.* Electronic Health Record Adoption and Rates of In-hospital Adverse Events. *J Patient Saf* Published Online First: 6 February 2016. doi:10.1097/PTS.0000000000000257

6 Hydari MZ, Telang R, Marella W. Saving Patient Ryan — Can Advanced Electronic Medical Records Make Patient Care Safer? Rochester, NY: : Social Science Research Network 2017. https://papers.ssrn.com/abstract=2503702 (accessed 3 Jan 2019).

7 Nuckols TK, Smith-Spangler C, Morton SC, *et al.* The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. *Syst Rev* 2014;**3**:56. doi:10.1186/2046-4053-3-56

8 Ball R, Robb M, Anderson SA, *et al.* The FDA's sentinel initiative—A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016;**99**:265–8. doi:10.1002/cpt.320

9 NIH award announcement - News | All of Us. https://allofus.nih.gov/news-events-and-media/announcements/nih-awards-55-million-build-million-person-precision-medicine (accessed 12 Dec 2018).

10 2011 Release: eMERGE network moves closer to tailored treatments based on patients' genomic information. Natl. Hum. Genome Res. Inst. NHGRI. https://www.genome.gov/27545093/2011-release-emerge-network-moves-closer-to-tailored-treatments-based-on-patients-genomic-information/ (accessed 27 Nov 2018).

11 Our Funding. 2014.https://www.pcori.org/about-us/financials-and-reports/our-funding (accessed 12 Dec 2018).

12 RFA-HG-07-005: Genome-Wide Studies in Biorepositories with Electronic Medical Record Data (U01). https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-07-005.html (accessed 27 Nov 2018).

13 Ross MK, Wei W, Ohno-Machado L. "Big Data" and the Electronic Health Record. *Yearb Med Inform* 2014;**9**:97–104. doi:10.15265/IY-2014-0003

14 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;**26**:1205–10. doi:10.1093/bioinformatics/btq126

15 Boland MR, Shahn Z, Madigan D, *et al.* Birth month affects lifetime disease risk: a phenome-wide method. *J Am Med Inform Assoc JAMIA* 2015;**22**:1042–53. doi:10.1093/jamia/ocv046

16 Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat Rev Genet* 2016;**17**:129–45. doi:10.1038/nrg.2015.36

17 PCORI in the Literature. https://www.pcori.org/literature/research-articles (accessed 16 Jan 2019).

18 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

19 National Research Network EHR Audit-log and MEta-data workgroups. https://medicine.ucsf.edu/center-clinical-informatics-and-improvement-research/national-research-network (accessed 27 Feb 2020).

20 NIH Strategic Plan for Data Science | Data Science at NIH. National Institutes of Health 2018. https://datascience.nih.gov/strategicplan (accessed 14 Nov 2018).

21 A Platform for Biomedical Discovery and Data-Powered Health: National Library of Medicine Strategic Plan 2017-2027. U.S. National Library of Medicine 2017.

22 Cowie MR, Blomster JI, Curtis LH, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017;**106**:1–9. doi:10.1007/s00392-016-1025-6

23 Just BH, Marc D, Munns M, *et al.* Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields. *Perspect Health Inf Manag* 2016;**13**.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4832129/ (accessed 10 Jan 2019).

24 Public Phenotypes | PheKB. https://phekb.org/phenotypes (accessed 15 Nov 2018).

25 Levine ME, Ryan PB. Lessons from CIRCE implementation of eMERGE phenotype definitions into actionable CDM v5 SQL queries. ;:2.

26 Meystre SM, Savova GK, Kipper-Schuler KC, *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008;:128–44.

27 Rosenbloom ST, Denny JC, Xu H, *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011;**18**:181–6. doi:10.1136/jamia.2010.007237

28 Wang Y, Wang L, Rastegar-Mojarad M, *et al.* Clinical information extraction applications: A literature review. *J Biomed Inform* 2018;**77**:34–49. doi:10.1016/j.jbi.2017.11.011

29 Carrell DS, Schoen RE, Leffler DA, *et al.* Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc JAMIA* 2017;**24**:986–91. doi:10.1093/jamia/ocx039

30 Weber GM, Adams WG, Bernstam EV, *et al.* Biases introduced by filtering electronic health records for patients with "complete data." *J Am Med Inform Assoc JAMIA* 2017;**24**:1134–41. doi:10.1093/jamia/ocx071

31 Hripcsak G, Knirsch C, Zhou L, *et al.* Bias Associated with Mining Electronic Health Records. *J Biomed Discov Collab* 2011;**6**:48–52.

32 United States Government Accountability Office. Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development. Published Online First: December 2019.https://www.gao.gov/products/GAO-20-215SP (accessed 27 Jan 2020).

33 Robert C. Machine Learning, a Probabilistic Perspective. *CHANCE* 2014;**27**:62–3. doi:10.1080/09332480.2014.914768

34 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nat Lond* 2015;**521**:436–44.

35 Dahl GE, Yu D, Deng L, *et al.* Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Trans Audio Speech Lang Process* 2012;**20**:30–42. doi:10.1109/TASL.2011.2134090

36 Hinton G, Deng L, Yu D, *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process Mag* 2012;**29**:82–97. doi:10.1109/MSP.2012.2205597

37 Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, *et al.*, eds. *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc. 2012. 1097–105.http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (accessed 8 Jan 2019).

38 Deng L. Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook [Perspectives]. *IEEE Signal Process Mag* 2018;**35**:180–177. doi:10.1109/MSP.2017.2762725

39 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;**25**:44. doi:10.1038/s41591-018-0300-7

40 Madani A, Arnaout R, Mofrad M, *et al.* Fast and accurate view classification of echocardiograms using deep learning. *Npj Digit Med* 2018;**1**:6. doi:10.1038/s41746-017-0013-1

41 Commissioner O of the. Press Announcements - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm (accessed 17 Jan 2019).

42 Gulshan V, Peng L, Coram M, *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016;**316**:2402–10. doi:10.1001/jama.2016.17216

43 Shickel B, Tighe PJ, Bihorac A, *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform* 2018;**22**:1589–604. doi:10.1109/JBHI.2017.2767063

44 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;**2**:719. doi:10.1038/s41551-018-0305-z

45 Big Data to Knowledge | NIH Common Fund. https://commonfund.nih.gov/bd2k (accessed 12 Dec 2018).

46 Ravì D, Wong C, Deligianni F, *et al.* Deep Learning for Health Informatics. *IEEE J Biomed Health Inform* 2017;**21**:4–21. doi:10.1109/JBHI.2016.2636665

47 Google and the University of Chicago Are Sued Over Data Sharing - The New York Times. https://www.nytimes.com/2019/06/26/technology/google-university-chicago-data-sharing-lawsuit.html (accessed 26 Jul 2019).

48 45 C.F.R. § 164.501. 2012.

49 National Committee on Vital and Health Statistics - Subcommittee on Privacy, Confidentiality and Security - Beyond HIPAA Working Meeting Summary. US Department of Health and Human Servies 2019. https://ncvhs.hhs.gov/wp-content/uploads/2019/10/2019-March-PCS-Beyond-HIPAA-meeting-summary-final.pdf

50 Terry NP. Regulatory Disruption and Arbitrage in Health-Care Data Protection. *Yale J Health Policy Law Ethics* 2017;**17**:143–208.

51 Gymrek M, McGuire AL, Golan D, *et al.* Identifying Personal Genomes by Surname Inference. *Science* 2013;**339**:321–4. doi:10.1126/science.1229566

52 Johnson KW, De Freitas JK, Glicksberg BS, *et al.* Evaluation of patient re-identification using laboratory test orders and mitigation via latent space variables. *Pac Symp Biocomput Pac Symp Biocomput* 2019;**24**:415–26.

53 Dwork C, McSherry F, Nissim K, *et al.* Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S, Rabin T, eds. *Theory of Cryptography*. Berlin, Heidelberg: : Springer 2006. 265–84. doi:10.1007/11681878_14

54 Abowd JM, Garfinkel SL. Disclosure Avoidance and the 2018 Census Test: Release of the Source Code. 2019. https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html (accessed 24 Jan 2020).

55 Bonomi L, Jiang X, Ohno-Machado L. Protecting patient privacy in survival analyses. *J Am Med Inform Assoc* doi:10.1093/jamia/ocz195

56 Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. *J Am Med Inform Assoc JAMIA* 2007;**14**:550–63. doi:10.1197/jamia.M2444

57 Beaulieu-Jones Brett K., Wu Zhiwei Steven, Williams Chris, *et al.* Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circ Cardiovasc Qual Outcomes* 2019;**12**:e005122. doi:10.1161/CIRCOUTCOMES.118.005122

58 Dwork C, Roth A. *The Algorithmic Foundations of Differential Privacy*. now 2014. https://ieeexplore.ieee.org/document/8187424 (accessed 28 Aug 2019).

59 Walonoski J, Kramer M, Nichols J, *et al.* Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc JAMIA* Published Online First: 30 August 2017. doi:10.1093/jamia/ocx079

60 Choi E, Biswal S, Malin B, *et al.* Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In: Doshi-Velez F, Fackler J, Kale D, *et al.*, eds. *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Boston, Massachusetts: : PMLR 2017. 286–305.http://proceedings.mlr.press/v68/choi17a.html

61 Begoli E, Brown K, Srinivas S, *et al.* SynthNotes: A Generator Framework for High-volume, High-fidelity Synthetic Mental Health Notes. In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018. 951–8. doi:10.1109/BigData.2018.8621981

62 MDClone | Unlock Healthcare Data.Transform Care. https://www.mdclone.com/ (accessed 2 Jul 2019).

63 Muniz-Terrera G, Mendelevitch O, Barnes R, *et al.* Virtual Cohorts and Synthetic Data in Dementia: An Illustration of Their Potential to Advance Research. *Front Artif Intell* 2021;**4**. doi:10.3389/frai.2021.613956

64 Lee SH. Natural language generation for electronic health records. *NPJ Digit Med* 2018;**1**:63.

65 Foraker R, Mann DL, Payne PRO. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* 2018;**3**:716–8. doi:10.1016/j.jacbts.2018.08.007

66 Dernoncourt F, Lee JY, Uzuner O, *et al.* De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc* 2017;**24**:596–606. doi:10.1093/jamia/ocw156

67 Making NHS data work for everyone | Reform. https://reform.uk/research/making-nhs-data-work-everyone (accessed 8 Jul 2019).

68 Medicare C for, Baltimore MS 7500 SB, Usa M. CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF). 2014.https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html (accessed 2 Jul 2019).

69 Dahmen J, Cook D. SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors* 2019;**19**:1181. doi:10.3390/s19051181

70 Buczak AL, Babin S, Moniz L. Data-driven approach for creating synthetic electronic medical records. *BMC Med Inform Decis Mak* 2010;**10**:59. doi:10.1186/1472-6947-10-59

71 Kartoun U. A Methodology to Generate Virtual Patient Repositories. Published Online First: 1 August 2016.http://arxiv.org/abs/1608.00570v1 (accessed 8 Jul 2019).

72 Zhang Z, Yan C, Mesa DA, *et al.* Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 2020;**27**:99–108. doi:10.1093/jamia/ocz161

73 Laderas T, Vasilevsky N, Pederson B, *et al.* Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. *bioRxiv* 2018;:232611. doi:10.1101/232611

74 Pollack AH, Simon TD, Snyder J, *et al.* Creating synthetic patient data to support the design and evaluation of novel health information technology. *J Biomed Inform* 2019;**95**:103201. doi:10.1016/j.jbi.2019.103201

75 Abadi M, Chu A, Goodfellow I, *et al.* Deep Learning with Differential Privacy. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna, Austria: : Association for Computing Machinery 2016. 308–18. doi:10.1145/2976749.2978318

76 boyd danah. Differential Privacy in the 2020 Decennial Census and the Implications for Available Data Products. *ArXiv190703639 Cs* Published Online First: 8 July 2019.http://arxiv.org/abs/1907.03639 (accessed 24 Jan 2020).

77 Wilson RJ, Zhang CY, Lam W, *et al.* Differentially Private SQL with Bounded User Contribution. *ArXiv190901917 Cs* Published Online First: 25 November 2019.http://arxiv.org/abs/1909.01917 (accessed 25 Jan 2020).

78 *google/differential-privacy*. Google 2020. https://github.com/google/differential-privacy (accessed 26 Feb 2020).

79 Microsoft and Harvard's Institute for Quantitative Social Science Collaboration Develops Open Data Differential Privacy Platform, Opens New Research. https://www.linkedin.com/pulse/microsoft-harvards-institute-quantitative-social-science-john-kahan (accessed 25 Jan 2020).

80 NOT-LM-19-003: Notice of Special Interest (NOSI): Computational and Statistical Methods to Enhance Discovery from Health Data. https://grants.nih.gov/grants/guide/notice-files/NOT-LM-19-003.html (accessed 10 Jul 2019).

81 Cohn JE. Ultrasonic bracelet and receiver for detecting position in 2d plane. 2017.https://patents.google.com/patent/US20170278051A1/en (accessed 25 Jul 2021).

82 Solutions. Humanyze. https://humanyze.com/solutions/ (accessed 25 Jul 2021).

83 Cram WA, Wiener M. Technology-mediated Control: Case Examples and Research Directions for the Future of Organizational Control. *Commun Assoc Inf Syst* 2020;**46**. doi:10.17705/1CAIS.04604

84 Shook E, Knickrehm M, Sage-Gavin E. Putting Trust to Work: Decoding Organizational DNA: Trust, Data and Unlocking Value in the Digital Workplace. https://www.accenture.com/_acnmedia/thought-leadership-assets/pdf/accenture-wf-decoding-organizational-dna.pdf (accessed 30 Jun 2021).

85 The State of ML and Data Science 2017. Kaggle. https://www.kaggle.com/surveys/2017 (accessed 12 Dec 2018).

86 2016 Data Science Report. Crowd Flower https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf (accessed 12 Dec 2018).

87 Ohno-Machado L. Mining electronic health record data: finding the gold nuggets. *J Am Med Inform Assoc* 2015;**22**:937–937. doi:10.1093/jamia/ocv119

88 Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;**24**:198–208. doi:10.1093/jamia/ocw042

89 Bayley KB, Belnap T, Savitz L, *et al.* Challenges in Using Electronic Health Record Data for Cer: Experience of 4 Learning Organizations and Solutions Applied. *Med Care* 2013;**51**. doi:10.1097/MLR.0b013e31829b1d48

90 Green, Steven M. Congruence of Disposition After Emergency Department Intubation in the National Hospital Ambulatory Medical Care Survey. *Ann Emerg Med*;**61**. doi:10.1016/j.annemergmed.2012.09.010

91 Brennan L, Watson M, Klaber R, *et al.* The importance of knowing context of hospital episode statistics when reconfiguring the NHS. *BMJ* 2012;**344**:e2432. doi:10.1136/bmj.e2432

92 Kasperkevic J. Google says sorry for racist auto-tag in photo app. The Guardian. 2015.https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app (accessed 2 Mar 2020).

93 The accent gap: How Amazon's and Google's smart speakers leave certain voices behind. Wash. Post. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/ (accessed 17 May 2019).

94 Rusanov A, Weiskopf NG, Wang S, *et al.* Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak* 2014;**14**:51. doi:10.1186/1472-6947-14-51

95 Arpey NC, Gaglioti AH, Rosenbaum ME. How Socioeconomic Status Affects Patient Perceptions of Health Care: A Qualitative Study. *J Prim Care Community Health* 2017;**8**:169–75. doi:10.1177/2150131917697439

96 Ng JH, Ye F, Ward LM, *et al.* Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. *Health Aff Proj Hope* 2017;**36**:548–52. doi:10.1377/hlthaff.2016.1044

97 Ramoni M, Sebastiani P. Robust Learning with Missing Data. *Mach Learn* 2001;**45**:147–70. doi:10.1023/A:1010968702992

98 Hyun KK, Brieger D, Woodward M, *et al.* The effect of socioeconomic disadvantage on prescription of guideline-recommended medications for patients with acute coronary syndrome: systematic review and meta-analysis. *Int J Equity Health* 2017;**16**. doi:10.1186/s12939-017-0658-z

99 Jabagi H, Tran DT, Hessian R, *et al.* Impact of Gender on Arterial Revascularization Strategies for Coronary Artery Bypass Grafting. *Ann Thorac Surg* 2018;**105**:62–8. doi:10.1016/j.athoracsur.2017.06.054

100 Freij M, Dullabh P, Lewis S, *et al.* Incorporating Social Determinants of Health in Electronic Health Records: Qualitative Study of Current Practices Among Top Vendors. *JMIR Med Inform* 2019;**7**:e13849. doi:10.2196/13849

101 Integrating Data On Social Determinants Of Health Into Electronic Health Records | Health Affairs. https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2017.1252 (accessed 27 Feb 2020).

102 Powell HS, Greenberg DL. Screening for unhealthy diet and exercise habits: The electronic health record and a healthier population. *Prev Med Rep* 2019;**14**:100816. doi:10.1016/j.pmedr.2019.01.020

103 Kight CE, Bouche JM, Curry A, *et al.* Consensus Recommendations for Optimizing Electronic Health Records for Nutrition Care. *Nutr Clin Pract* 2020;**35**:12–23. doi:10.1002/ncp.10433

104 Dinh-Le C, Chuang R, Chokshi S, *et al.* Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions. *JMIR MHealth UHealth* 2019;**7**. doi:10.2196/12861

105 Adler-Milstein J, Chen Y, Hribar M, *et al.* Advancing Common Approaches to Working with EHR Log Data. 2019.

106 Benaim AR, 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. doi:10.2196/16492

107 Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol* 2018;**154**:1247–8. doi:10.1001/jamadermatol.2018.2348

108 Gohorbani A, Natarajan V, Coz DD, *et al.* DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. 2019. https://arxiv.org/abs/1911.08716 (accessed 26 Feb 2020).

109 Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 2013;**46**. doi:10.1016/j.jbi.2013.06.010

110 Kahn MG, Callahan TJ, Barnard J, *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs* 2016;**4**. doi:10.13063/2327-9214.1244

111 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc JAMIA* 2013;**20**:144–51. doi:10.1136/amiajnl-2011-000681

112 ACHILLES for data characterization – OHDSI. https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/ (accessed 12 Jan 2019).

113 Blacketer C, Londhe A, Sena A, *et al. Data Quality Dashboard*. Observational Health Data Sciences and Informatics https://ohdsi.github.io/DataQualityDashboard/ (accessed 21 Jan 2020).

114 Estiri H, Stephens KA, Klann JG, *et al.* Exploring completeness in clinical data research networks with DQe-c. *J Am Med Inform Assoc* 2018;**25**:17–24. doi:10.1093/jamia/ocx109

115 DQUEEN – OHDSI. https://www.ohdsi.org/2019-us-symposium-showcase-106/ (accessed 27 Jan 2020).

116 Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS Wash DC* 2016;**4**:1239. doi:10.13063/2327-9214.1239

117 Huser, Vojtech. OHDSI Data Quality Study. http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy (accessed 12 Jan 2019).

118 Khare R, Utidjian L, Ruth BJ, *et al.* A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc* 2017;**24**:1072–9. doi:10.1093/jamia/ocx033

119 Raebel MA, Haynes K, Woodworth TS, *et al.* Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf* 2014;**23**:609–18. doi:10.1002/pds.3580

120 Data Quality Review and Characterization Programs v4.2.1 | Sentinel Initiative. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/data-quality-review-and-characterization (accessed 12 Jan 2019).

121 Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMs Gener Evid Methods Improve Patient Outcomes* 2018;**6**:3. doi:10.5334/egems.199

122 Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019;**19**. doi:10.1186/s12911-019-0793-0

123 Callahan TJ, Bauck AE, Bertoch D, *et al.* A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *EGEMs Gener Evid Methods Improve Patient Outcomes* 2017;**5**:8. doi:10.5334/egems.223

124 Lee K, Weiskopf N, Pathak J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA Annu Symp Proc* 2018;**2017**:1080–9.

125 Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019;**19**:44. doi:10.1186/s12911-019-0793-0

126 Foraker RE, Yu SC, Gupta A, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* Published Online First: 14 December 2020. doi:10.1093/jamiaopen/ooaa060

127 El Emam K, Mosquera L, Jonker E, *et al.* Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 2021;**4**. doi:10.1093/jamiaopen/ooab012

128 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic

healthcare data: Preserving data utility and patient privacy. *Comput Intell*;**n/a**. doi:https://doi.org/10.1111/coin.12427

129 Hittmeir M, Ekelhart A, Mayer R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Canterbury, CA, United Kingdom: : Association for Computing Machinery 2019. 1–6. doi:10.1145/3339252.3339281

130 Emam KE. Seven Ways to Evaluate the Utility of Synthetic Data. *IEEE Secur Priv* 2020;**18**:56–9. doi:10.1109/MSEC.2020.2992821

131 Adler-Milstein J, Adelman JS, Tai-Seale M, *et al.* EHR audit logs: A new goldmine for health services research? *Journal of Biomedical Informatics* 2020;**101**:103343. doi:10.1016/j.jbi.2019.103343

Chapter 2. CREATION AND TESTING OF A NOVEL FITNESS FOR USE FRAMEWORK FOR ASSESSING THE UTILITY OF SYNTHETIC AND REAL ELECTRONIC HEALTH RECORDS

2.1 ABSTRACT

Objective: Evaluate the fitness for use of privacy preserving and non-privacy preserving electronic health record (EHR) data to enable influenza pandemic surveillance or research.

Materials and Methods: A total of 50 reviews from the Cochrane Database of Systematic Reviews were assessed for their feasibility to be replicated in the EHR along with other factors such as their evidence level per the GRADE criteria. Findings (n=31) from a COVID-19 related Cochrane review were replicated through the use of open-source Observational Health Data Science and Informatics (OHDSI) software applied to three EHR databases: the UW Medicine COVID-19 Research data set (UWM-CRD) from September 2020, the UWM-CRD from May 2021, and a public synthetic data (Synthea COVID-19) set created for COVID-19 research. We compared the similarity of our replications to the results published by Cochrane.

Results: Of the 50 CDSR reviews assessed, 15 reviews (30%) with 37 individual findings were found to be suitable for inclusion within our repository of clinical findings for future replication. All COVID-19 related CDSR outcomes were able to be calculated in the UWM-CRD 2021 and Synthea OMOP databases. In contrast, only about ⅓ of the CDSR outcomes were able to be calculated in the UWM-CRD 2020 dataset. Nearly all the replicated results across all three databases showed lower prevalence or incidence than the Cochrane review's results. UWM-CRD 2021 results were most similar to the CDSR results whereas Synthea COVID-19 had roughly ⅓ of outcomes falling below (mostly due to counts of zero) the range reported in the CDSR review.

Discussion: Results from our broad review of the CDSR demonstrate that the EHR may be a rich source of real-world data to supplement systematic review findings and, in the process, can enable fitness for use assessments. Barriers to using real-world data and potential causes of underreporting cohorts were identified, however.

Conclusion: We demonstrated the feasibility of replicating CDSR reviews using electronic health record data for both synthetic and real data as a method to assess their fitness for use. Our EHR databases did have lower values for prevalence and incidence of cardiovascular events compared to the CDSR review's weighted averages for the vast majority of outcomes. We observed heterogeneity between databases. The Synthea COVID-19 data set released in the Spring of 2020 may not be fit for use for analysis of cardiovascular outcomes.

2.2 INTRODUCTION

Electronic health records (EHRs), which are now used in nearly all of the United States' hospitals[1], provide a rich source of real-world data (RWD) to create real-world evidence (RWE) from. In contrast to traditional clinical trials, real-world data are generated by the routine provision of medical care to patients; real-world evidence is the clinical knowledge created from the analysis of real-world data[2]. In acknowledgement of the potential for RWE to inform health care decisions, the United States Food and Drug administration created a framework for the use of real-world evidence in 2018 - including that from Electronic Health Records - to inform their decision-making[2]. With increasing importance placed on the use of real-world evidence and the key role of EHR data in national-scale research such as the National COVID Cohort Collaborative (N3C)[3] and the All of Us Research program[4], scrutinizing the "fitness for use" - meaning whether the data are suitable to be used in a specific analysis or context or not[5] - of real-world data becomes more important.

Real-World Data utility issues encompassing quality, bias, and comprehensiveness limit the effective use of EHR data. Within healthcare, data quality issues can undermine use of EHRs for knowledge discovery, precision medicine, comparative effectiveness research, and other research using secondary data analysis[6–12]. Biases are seen in that sicker patients have more complete data[13], vulnerable populations have a higher probability of visiting multiple health care system for care[14–17], patients of lower socioeconomic status have limited healthcare access and are less likely to receive diagnoses and medication[14,17], and female patients receive less aggressive coronary revascularization approaches compared to men.[18] Comprehensiveness of the EHR is limited in multiple ways by insufficient granularity or lack of data capture - especially in structured data - of social determinants of health[19,20], nutrition and exercise data[21,22], and patient data from wearable health technology.[23] All these issues can affect the fitness for use of EHR data.

Despite the importance of assessing fitness for use, determining whether EHR data are fit for use is not included in current frameworks for general health record data quality characterization[24,25] nor evaluated by data quality assessment (DQA) tools.[26–28] Recent work has been done to characterize and assess dimensions of EHR data utility, leading to open tools to make these assessments. Electronic health record data quality characterization[24,25,29] has matured to yield a harmonized terminology and framework[24] for describing EHR data quality. Multiple organizations have produced DQA methods[26,27,30] adhering to this framework that assess common data model conformance, completeness and a limited number of overall data quality checks against rules such as age less than zero.[27] DQA has been conducted across distributed research networks.[31–36]

New technologies such as synthetic data aim to reduce the barriers to data accessibility yet they can degrade data utility.[37–41] Studies have been conducted on synthetic data sets to compare their utility in comparison to the original "ground truth" data they were modeled on[41–49]. Yet these assessments still don't help EHR data analysts to understand the fitness for use of their original data at baseline let alone manage subsequent tradeoffs between utility and privacy when applying privacy preserving technologies (PPT). COVID-19 catalyzed interest in EHR data analysis and sharing[3], further increasing the need to assess the fitness for use of both real and synthetic EHR data formatted in common data models such as the observational health data sciences and informatics' (OHDSI) observational medical outcomes partnership (OMOP) common data model (CDM).[50]

A potential method to assess the fitness for use of EHR data and their synthetic derivatives involves creating a feedback loop between traditional clinical trials and/or systematic reviews and the real-world data provided by EHRs. A broad library of findings from traditional evidence sources could be built up to then compared with EHR data to assess the fitness for use of EHR data along axes such as whether the necessary data elements and patients exist within the EHR and if so, whether the same results can be obtained through analyzing the EHR data. Limited related work has been conducted so far. Bartlett et al., 2019 explored the feasibility of replicating clinical trial evidence with real-world data and found that 15% of the trials could feasibly be replicated using claims and/or EHR data.[51] Chen et. al, 2019 compared the rates of four clinical quality measures in a synthetic data set to publicly reported rates that would be expected in that synthetic population, finding that the synthetic data performed well modeling demographics but not on heterogeneous outcomes.[52]

2.2.1 *Objective*

In this study, we sought to develop and test a new framework by which one can determine whether electronic health records are fit for use and also assess the impact of privacy preserving technology on fitness for use. To do so, we made use of the Cochrane Database of Systematic Reviews (CDSR). We built up a repository of outcomes from the CDSR as individual research findings to be replicated within electronic health records formatted in the OMOP CDM. Due to the present relevance of the COVID-19 pandemic and its implications on observational research and disease surveillance going forward, we focused our replications on COVID-19 related outcomes.

2.3 METHODS

2.3.1 *Cochrane Review Search*

Prior to the onset of COVID-19, we conducted a cursory review of the Cochrane Database of Systematic Reviews (CDSR)[53] to evaluate the abundance of potential clinical findings for our repository and their suitability to be mined in an Electronic Health Record as part of a data utility assessment (DUA). A single reviewer (author JAT) assessed the most recent 2-4 reviews for 24 randomly selected Cochrane Review Groups and Topics (e.g. *Common Mental Disorders*) to determine whether or not review findings met the following criteria created by the study team: (1) had sufficient (≥low) evidence, (2) could be mined from an EHR, (3) could be mined from *structured* data in an EHR, (4) were pediatric specific, and (5) overall suitability for preliminary inclusion within our repository of clinical findings. The criteria were designed to select for findings with evidence high enough to be expected to change less over time and to collect other important metadata (e.g. whether the finding could be mined from structured data) to provide

insight into reasons for exclusion and/or applicability of findings to a dabatase of interest (e.g. pediatric population or not). A subset was reviewed by a second reviewer (ABW) for validation. We created a pilot version of the repository data visualization using R Shiny.[54] Interactive features included within the visualization are free-text search and filtering by multiple drop-down boxes. In response to the COVID-19 pandemic, we decided to focus actual replications on CDSR findings related to COVID-19. To gather COVID-19 related findings for replication, we searched the CDSR further for reviews that contained the phrase "COVID-19" in the title, abstract, or its list of keywords through the July, 2021 Cochrane issue.

2.3.2 *Electronic Health Record Data sets*

The UW Medicine COVID Research Data set (UWM-CRD) was created in March 2020 to enable observational COVID-19 related research. Authors ABW, GL, JAT, were involved in the creation of the UWM-CRD along with many other personnel from UW Medicine and the University of Washington (Seattle, USA). The UWM-CRD is formatted in the OMOP CDM v5.3.1 and hosted in Microsoft SQL Server. The dataset includes UW Medicine patients who have been tested for COVID-19 that have information within the UW Medicine Electronic Health Record. The dataset is a subgroup of the non-OMOP clinical data warehouse of UW Medicine, comprised of Harborview Medical Center, UW Medical Center - Montlake, and Northwest hospital in Seattle WA, and is based on its current electronic health record systems, with data spanning over 10 years and including roughly 5 million patients. As part of the iterative extract, transform, load (ETL) process, OHDSI's Data Quality Dashboard was run (by author JAT) on the UWM-CRD to identify data quality issues which were remedied by the broad UW Medicine ETL team. The UWM-CRD is one of the data sets within the National Covid Cohort Collaborative and has been used in OHDSI network studies.[55,56]

Synthea data are synthetic data created using PADARSER, the Publicly Available Data Approach to the Realistic Synthetic EHR[57], which creates synthetic data that use publicly available statistics and clinical practice guidelines in an attempt to create realistic EHR data.[58] A Synthea dataset (n = 12,359 patients) simulating a COVID-19 outbreak over Jan-March 2020 and formatted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v5.3.1[50] was downloaded from the Observational Health Data Sciences and Informatics (OHDSI) forums.[59] We then loaded the data into a PostgreSQL database using the SQL scripts OHDSI provides for v5.3.1.[60] We loaded the default vocabulary files as of July 09, 2021 from the ATHENA OHDSI vocabularies repository.[61] For the purposes of this study, the Synthea database will be referred to as "Synthea COVID-19." More detailed information about each database is shown in Table 2.1.

Table 2.1. Data set summaries

Data Set	**Date Range**	**Description**
UWM-CRD 2020	2010 - 09/20/2020	The dataset includes UW Medicine patients who have been tested for COVID-19 that have information within the UW Medicine Electronic Health Record. The dataset is a subgroup of the non-OMOP clinical data warehouse of University of Washington Medical Center, comprised of Harborview Medical Center, UW Medical Center - Montlake, and Northwest hospital in Seattle WA, and is based on its current electronic health record systems, with data spanning over 10 years and including roughly 5 million patients.
UWM-CRD 2021	2010 - 05/19/2021	The dataset includes UW Medicine patients who have been tested for COVID-19 that have information within the UW Medicine Electronic Health Record. The dataset is a subgroup of the non-OMOP clinical data warehouse of University of Washington Medical Center, comprised of Harborview Medical Center, UW Medical Center - Montlake, and Northwest hospital in Seattle WA,

		and is based on its current electronic health record systems, with data spanning over 10 years and including roughly 5 million patients.
Synthea COVID-19	01/01/2020 - 03/29/2020	Synthea COVID-19 data set distributed on the OHDSI forums March 2020.[59] The data were created by 1) starting with the Synthea COVID-19 github branch as of late March 2020[62] which enriched the data with COVID-19 related diagnosis and tests SNOMED and LOINC codes that were known of at that time, 2) creating a Massachusetts data set, then 3) converted to the OMOP CDM using ETL-Synthea[63]

2.3.3 *Replication of Cochrane Reviews*

To replicate outcomes from the CDSR, we used computable phenotypes from the OHDSI open-source tool ATLAS[64,65] which enables auto-generation of SQL queries and translation into multiple SQL dialects. Computable phenotypes previously developed by and used in OHDSI network studies such as CHARYBDIS (Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2)[55] were used, when available, by running the CHARYBDIS software package on all databases to generate cohorts. The full list of CHARYBDIS phenotypes can be found within the study package[66] with hyperlinks to their cohort definitions in ATLAS.[64] Some replications included the use of condition_era groups which roll up conditions found in the condition_era table into the highest ancestor as described in the OHDSI FeatureExtraction package.[67] The presence of standardized derived elements tables (drug_era, dose_era, condition_era tables) allow selection of patients based off the timeframe a patient a patient has a diagnosis or is taking a drug.[68]

For each replication and database, CHARYBDIS phenotypes were used to create cohorts for or calculate each of the following replication components: the patients and/or population (e.g.

COVID-19 diagnosis or tested positive), setting (e.g. hospitalized), the target (e.g. deep vein thrombosis), comparison (if applicable), and outcome (e.g. 30-day incidence). We performed a qualitative analysis to investigate why the calculated outcomes in each database may differ from the outcome result in the CDSR.

2.4 RESULTS

2.4.1 *Cochrane Review Search*

Our review evaluated 50 CDSRs yielding a total of 15 reviews (30%) with 37 individual findings suitable for inclusion within our repository of clinical findings. A total of three CDSRs (6%) of the fifty reviewed were excluded solely due to the need for analysis of unstructured data. Of the 15 CDSRs suitable for our repository, seven (46%) were conducted on solely pediatric (≤18 years old) populations. Detailed information on each Cochrane review that was assessed can be found in supplemental table 2.5. The flowchart of our assessment can be seen in Figure 2.1.

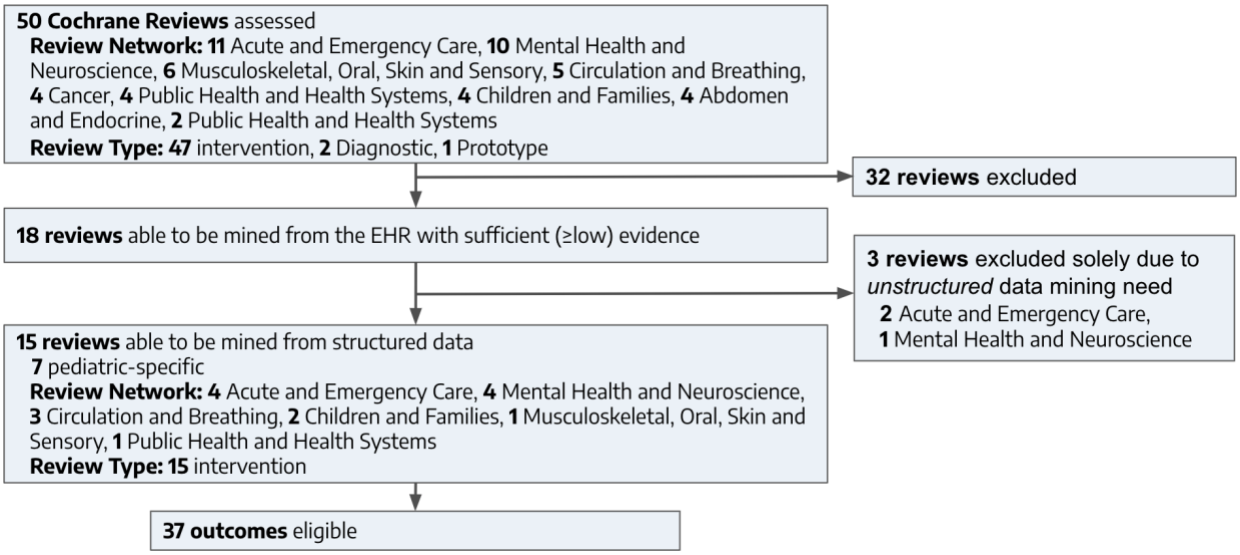

Figure 2.1. Study flowchart of the review process and results from our search of the CDSR for replicable reviews within the EHR.

We found 35 COVID-19 related reviews within the CDSR as a result of our keyword search. Of those, "COVID-19 and its cardiovascular effects: a systematic review of prevalence studies"[69] was selected for replication since it provided 31 outcomes to be replicated, incorporated over 200 studies into its review, is broadly relevant to covariates studied in COVID-19[70] and non-COVID-19 cardiology research (to enable fitness for use assessments for cardiology research), and will be updated by the authors as more evidence is accumulated over time. The countries of origin in the review consisted of China (47.7%), USA (20.9%), Italy 9.5%. The majority of studies were retrospective (89.5%) while the remaining were prospective (9.1%) or randomized clinical trials(1.4%). All other study designs were excluded by the authors. The main criteria for the study were that the study was written in English, had >100 participants and was peer reviewed. The details of this study are presented in Table 2.2.

Table 2.2. CDSR Study used in replication with the EHR

Title	**Search Period**	**Main Criteria**	**No. of studies**	**Country of origin**	**Study type**
COVID-19 and its cardiovascular effects: a systematic review of prevalence studies	December 2019 to 24 July 2020	Prospective and retrospective cohort studies, controlled before-and-after, case-control and cross-sectional studies, and randomised controlled trials (RCTs). Peer-reviewed studies only with >100 participants. Written in English.	220	China (47.7%), USA (20.9%), Italy 9.5%	Retrospective (89.5%), RCT (1.4%) prospective (9.1%)

2.4.2 *Characteristics of each database*

As part of the replication process, database characteristics were generated using the CHARYBDIS package and reported in Table 2.3. UWM-CRD 2021 had the largest number of

patients with a COVID-19 diagnosis or SARS-CoV-2 positive test without requiring prior history on those patients (n = 7,609) followed by UWM-CRD 2020 (n = 3,245) and the Synthea COVID-19 (n = 1,835). Notably, the Synthea COVID-19 database has zero count for all tested positive (versus tested positive or diagnoses) cohorts despite having patients with COVID-19 diagnoses. All patients within the UWM-CRD 2021 database were tested for SARS-COV-2 which is expected since COVID-19 testing is a mandatory inclusion criterion for the UWM-CRD. In contrast, a minority (n = 1,826; 14.8%) of those in the Synthea COVID-19 database were tested for SARS-CoV-2.

Table 2.3. COVID-19 and influenza-related cohort counts across the three databases

	Cohorts	**UWM-CRD 2020**	**UWM-CRD 2021**	**Synthea COVID-19**
CHARYBDIS Cohort ID	Total Patients (n)	NaN*	125,340	12,359
127	Persons tested for SARS-CoV-2 with no required prior observation	83,921	125,340	1,826
126	Persons tested for SARS-CoV-2 with at least 365d prior observation	53,581	78,896	1,800
133	Persons with a COVID-19 diagnosis or a SARS-CoV-2 positive test with no required prior observation	3,245	7,609	1,835
131	Persons tested with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with no required prior observation	3,177	6,963	1,825
129	Persons tested positive for SARS-CoV-2 with no required prior observation	3,140	4,886	0
132	Persons with a COVID-19 diagnosis or a SARS-CoV-2 positive test with at least 365d prior observation	1,848	4,690	1,809
130	Persons tested with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation	1,797	4,168	1,799
128	Persons tested positive for SARS-CoV-2 with at least 365d prior observation	1,777	3,042	0
135	Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with no required prior observation	733	2,166	701
139	Persons hospitalized with a SARS-CoV-2 positive test with no required prior observation	676	1,300	0

134	Persons hospitalized with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation	534	1,436	694
138	Persons hospitalized with a SARS-CoV-2 positive test with at least 365d prior observation	494	957	0
112	Persons with Influenza diagnosis or positive test 2017-2018 with no required prior observation	364	569	NaN
137	Persons hospitalized and requiring intensive services with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with no required prior observation	117	201	273
141	Persons hospitalized and requiring intensive services with a SARS-CoV-2 positive test with no required prior observation	112	171	0
136	Persons hospitalized and requiring intensive services with a COVID-19 diagnosis record or a SARS-CoV-2 positive test with at least 365d prior observation	68	149	0
114	Persons hospitalized with influenza diagnosis or positive test 2017-2018 with no required prior observation	68	126	NaN

*Total patients NaN due to the UWM-CRD database changing over time and the total patient was not recorded at the time cohorts were generated.

2.4.3 *Replications of CDSR findings*

Results from the replication of CDSR findings are shown in Table 2.4. All CDSR outcomes were able to be calculated in the UWM-CRD 2021 and Synthea COVID-19 databases. In contrast, only about ⅓ of the CDSR outcomes were able to be calculated in the UWM-CRD 2020 dataset. Nearly all the replicated results showed lower prevalence or incidence than the CDSR weighted averages across all three databases. Only COVID-19+ prevalence of obesity within the last 365 days prior to index in Synthea COVID-19, COVID-19+ 30-day incidence of myocardial infarction or acute coronary syndrome, COVID-19+ and hospitalized 30-day incidence of vasopressor support, and COVID-19+ and intensive care death 30-day incidence of death in the UWM-CRD 2021, and COVID-19+ and hospitalized incidence of ECMO in the UWM-CRD 2020 were equal to or greater than the CDSR weighted average. Compared to the CDSR findings' ranges, only two of the UWM-CRD 2021 and about ⅓ of the Synthea COVID-19 replication results were outside of the ranges. Results outside of the ranges are bolded in Table 2.4. Roughly ⅔ of the Synthea COVID-19 results for prevalence and incidence were 0.0%.

Table 2.4. Outcome results in EHR databases compared to the CDSR

Cohort ID	Domain	Replication Target	CDSR Target	CDSR weighted average (range)	UWM-CRD 2020	UWM-CRD 2021	Synthea COVID-19
Prevalence (-365 through -1 days relative to index)							
133	Cohort	Prevalent hypertension	Hypertension	36.1% (4.5 - 100%)	20.6%	23.8%	23.9%
133	Cohort	Prevalent obesity	Obesity	21.6% (0.2 - 57.6%)	6.3%	11.8%	45.3%
133	Cohort	Prevalent Type 2 Diabetes Mellitus	Diabetes	22.1% (0.0% to 100%)	10.8%	12.0%	4.4%
133	Condition_era group	Ischemic heart disease	Ischemic heart disease	10.5% (1.0% to 28.2%)	NaN	2.7%	**0.0%**
133	Cohort	Prevalent heart disease	Cardiovascular disease	23.5% (0.7% to 68.7%)	11.9%	15.5%	8.6%
133	Condition_era group	Heart failure	Heart failure	6.5% (0.0% to 28.0%)	NaN	3.6%	0.2%
133	Condition_era group	Sequela of cerebrovascular accident	Cerebrovascular accident	5.1% (0.5% to 19.6%)	NaN	0.6%	**0.2%**
133	Condition_era group	Atrial fibrillation	Atrial fibrillation	11.1% (1.0% to 22.8%)	NaN	3.0%	**0.1%**

133	Condition_era group	Heart valve disorder	Valve disease	3.7% (1.8% to 6.8%)	NaN	2.3%	**0.0%**
Incidence (index to 30 days)							
133	Condition_era group	Myocardial infarction	Myocardial infarction or Acute coronary syndrome	1.7% (0.0% to 3.6%)	NaN	1.7%	0.0%
133	Cohort	Stroke (ischemic or hemorrhagic) events	Stroke	1.2% (0.0% to 9.6%)	0.6%	0.6%	0.0%
133	Condition_era	Heart failure	Heart failure	6.8% (0.0% to 24.0%)	NaN	3.6%	0.0%
133	Cohort	Venous thromboembolic (pulmonary embolism and deep vein thrombosis) events	Venous thromboembolism	7.4% (0.0% to 46.2%)	0.9%	1.2%	0.0%
133	Condition_era	Deep venous thrombosis	Deep vein thrombosis	6.1% (0.0% to 46.2%)	NaN	0.6%	0.0%
133	Condition_era	Pulmonary embolism	Pulmonary embolism	4.3% (0.0% to 23.8%)	NaN	1.0%	0.0%
133	Condition_era group	Blood coagulation disorder	Coagulopathy	8.0% (0.5% to 38.0%)	NaN	0.7%	**0.0%**
133	Condition_era group	Cardiac arrhythmia	Arrhythmia	9.3% (0.0% to 30.3%)	NaN	4.2%	0.1%
133	Condition_era group	Supraventricular arrhythmia	Arrhythmia: Supra-ventricular	8.5% (0.0% to 24.7%)	NaN	3.3%	0.1%

133	Cohort	ventricular arrhythmia or cardiac arrest during hospitalization	Arrhythmia: Ventricular	2.7% (0.0% to 12.4%)	0.4%	0.7%	0.0%
133	Condition_era group	Heart block	AV-block	1.3% (0.0% to 2.6%)	NaN	0.8%	0.0%
133	Condition_era group	Long QT syndrome	Prolonging QT	7.6% (0.0% to 20.0%)	NaN	<0.1%	0.0%
135	Condition_era group	Shock	Shock[1]	17.1%(0.2% to 67.0%)	NaN	7.0%	**0.0%**
135	Drug_era group	Vasoproctectives	Vasopressor support[1]	20.9% (3.0% to 71.0%)	NaN	22.0%	**0.4%**
135	Condition_era group	Acute renal failure syndrome	RRT[1]	5.1% (0.0% to 50.0%)	NaN	13.5%	0.0%
135	Cohort	ECMO during hospitalization	ECMO[1]	1.1% (0.0% to 8.1%)	1.5%	0.6%	0.0%
135	Condition_era group	Myocarditis	Myocarditis[2]	2.6% (0.0 to 12.5)	NaN	<0.5%	0.0%
135	Condition_era group	Myocardial necrosis	Cardiac injury[2]	27.6% (0.6% to 100%)	NaN	6.3%	**0.0%**
135	Condition_era group	Left ventricular abnormality	LV dysfunction[2]	13% (4.0% to 30.0%)	NaN	**<0.5%**	**0.0%**
135	Condition_era group	Right ventricular abnormality	RV dysfunction[2]	14.2% (3.6% to 25.0%)	NaN	**<0.5%**	**0.0%**

133	Cohort	Death	Death due to any cause	6.1% (0.0% to 100%)	3.3%	1.9%	0.1%
137	Cohort	Death	Death due to any cause	32.0% (8.7% to 72%)	NaN	33.3%	**0.0%**

Prevalence: prevalence of pre-existing disease: weighted mean (range)
CHARYBDIS Cohort ID corresponds to the counts and longer descriptions seen in Table 2.3.
[1]These targets were primarily from source articles studying patients receiving intensive services.
[2]These targets were primarily from imaging and lab tests in the source articles

2.5 DISCUSSION

Results from our broad review of the CDSR demonstrate that the EHR may be a rich source of real-world data to supplement systematic review findings. We found a higher percentage of potentially replicable findings from the CDSR as compared to an assessment by Bartlett et al. of 220 individual clinical trials published in high-impact journals articles, which found that 15% of the clinical trial findings could theoretically be discovered in structured EHR or claims data[51]. Reasons for the higher percentage from our review may be due to: 1) the CDSR frequently assessing multiple clinical outcomes as endpoints within a single review and 2) potential bias within the CDSR dataset towards older or more established interventions with enough publications to warrant a review - as opposed to potentially more cutting edge & translational studies - that are more likely to be already used in practice. This work demonstrates the feasibility of identifying a sufficient number of clinical findings for a fitness for use tool using our approach beyond COVID-19 specific research, even assuming the 15% level found by Bartlett et al, 2019.[51]

Our replications of the CDSR findings revealed that all the outcomes could be calculated by the stock OHDSI software and that the UWM-CRD 2021 data most closely matched the CDSR results with only two of the outcomes falling outside of the documented ranges while Synthea COVID-19 had about ⅓ of outcomes fall outside of the ranges. The limited number of cohorts calculated on the UWM-CRD 2020 was due to our OMOP ETL process being incomplete at that time; standardized derived elements (e.g. the condition_era table)[68] were not yet generated. The impact of missing standardized derived elements on our analysis highlights the importance of prioritizing the ETL of these tables when sites are building out an OMOP data warehouse. Our results also emphasize the value of iteratively running these replications to

capture changes to an EHR database ranging from data quality fixes to the natural growth of the database over time.

For the outcomes that were able to be calculated across all three databases, our prevalence and incidence rates were nearly all lower than the CDSR results, which is similar to findings of EHR observational network studies on COVID-19. In two OHDSI network studies using the CHARYBDIS software package, potential underreporting of symptoms and covariates was identified yet the results between sites were heterogeneous.[55,56] For results in this study with a value of 0.0%, which was common in the Synthea COVID-19 database, data quality issues such as a lack of mapping the required concepts source concepts to the OMOP CDM or utility issues such as a lack of realism in the data could be the root cause.

A few examples from the replication targets illustrate the barriers to using real-world data and potential causes of underreporting. Myocarditis, cardiac injury, left ventricular (LV) dysfunction, and right ventricular (RV) dysfunction CDSR results were primarily sourced from studies using imaging and laboratory tests to calculate incidence. The source article documenting LV dysfunction and RV dysfunction incidence relied upon granular echocardiogram findings such as wall motion abnormalities.[71] These findings are commonly locked in text and, despite efforts to extract them with natural language processing, the main focus of these extractions has been limited to left ventricular ejection fraction values.[72,73] The electrocardiography results likely suffer from a similar problem with information ‘locked’ in electrocardiograms. All the arrhythmias and AV-block have lower incidence in the UWM-CRD 2021 database than the CDSR review, yet the intra-database proportions of these targets are roughly similar which suggests a denominator issue across all the UWM-CRD electrocardiology results. Transient events recorded in labs or test values likely increased the chances of underreporting as well. The

prolonging QT target would most reliably be calculated by analyzing the QRS duration of each electrocardiogram directly rather than relying upon the condition of "long QT syndrome" to be established and documented. Similarly, the coagulopathy phenotyping would have been improved by including patients with an abnormal prothrombin time test.

2.6 Limitations and future work

Our study was limited in its size and scope. Our broad review of the CDSR to assess the feasibility of replication consisted of evaluating only 50 total reviews and their outcomes. The replications conducted were on more than 30 outcomes yet they were from a single Cochrane Review with a focus on COVID-19 and cardiovascular events. Thus, our replications are best suited to aiding the assessment of fitness for use of EHR data in either or both of these domains. The Synthea COVID-19 data assessed were limited in size and have been improved upon in future iterations. In future work, we will increase the quantity of synthetic data sets assessed, the quantity and variety of domains of Cochrane reviews to be replicated, and expand beyond prevalence and incidence to study other outcomes such as the effects of interventions. Additionally, we plan to build upon existing OHDSI software to build an R package that 1) assesses solely the replications of interest with the results automatically compared to the results of the Cochrane reviews being replicated and 2) enables user-driven filtering of results relative to the data task at hand (e.g. filtering on relevant MeSH terms and domain-specific concepts).

2.7 Conclusion

We demonstrated the feasibility of replicating CDSR reviews using electronic health record data for both synthetic and real data as a method to assess their fitness for use. Our EHR databases did consistently have lower values for prevalence and incidence of cardiovascular

events compared to the CDSR review's weighted averages, for the vast majority of outcomes. This may be a result of the data used in the Cochrane review being more complete because it used more assessments missed by EHR data or used concepts not directly collected or collected only in narrative form within the EHR. We observed heterogeneity between databases. UWM-CRD 2021 results were most similar to the CDSR results whereas Synthea COVID-19 had roughly ⅓ of outcomes falling below the range reported in the CDSR review, primarily due to outcomes that had a 0.0% incidence or prevalence in the Synthea COVID-19 database. This suggests the Synthea COVID-19 data set released in the Spring of 2020 may not be fit for use for analysis of cardiovascular outcomes.

2.8 REFERENCES FOR CHAPTER 2

1 Pedersen CA, Schneider PJ, Scheckelhoff DJ. ASHP national survey of pharmacy practice in hospital settings: Prescribing and transcribing—2016. *Am J Health Syst Pharm* 2017;**74**:1336–52. doi:10.2146/ajhp170228

2 Framework for FDA's Real-World Evidence Program. U.S. Food & Drug Administration 2018. https://www.fda.gov/media/120060/download (accessed 16 Jul 2021).

3 Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021;**28**:427–43. doi:10.1093/jamia/ocaa196

4 The "All of Us" Research Program. *N Engl J Med* 2019;**381**:668–76. doi:10.1056/NEJMsr1809937

5 Juran JM, Godfrey AB, editors. *Juran's quality handbook*. 5th ed. New York: : McGraw Hill 1999.

6 NIH Strategic Plan for Data Science | Data Science at NIH. National Institutes of Health 2018. https://datascience.nih.gov/strategicplan (accessed 14 Nov 2018).

7 United States Government Accountability Office. Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development. Published Online First: December 2019.https://www.gao.gov/products/GAO-20-215SP (accessed 27 Jan 2020).

8 Ohno-Machado L. Mining electronic health record data: finding the gold nuggets. *J Am Med Inform Assoc* 2015;**22**:937–937. doi:10.1093/jamia/ocv119

9 Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;**24**:198–208. doi:10.1093/jamia/ocw042

10 Bayley KB, Belnap TM, Savitz L, *et al.* Challenges in Using Electronic Health Record Data for CER: Experience of 4 Learning Organizations and Solutions Applied. *Med Care* Published Online First: August 2013. doi:10.1097/MLR.0b013e31829b1d48

11 Green, Steven M. Congruence of Disposition After Emergency Department Intubation in the National Hospital Ambulatory Medical Care Survey. *Ann Emerg Med*;**61**. doi:10.1016/j.annemergmed.2012.09.010

12 Brennan L, Watson M, Klaber R, *et al.* The importance of knowing context of hospital episode statistics when reconfiguring the NHS. *BMJ* 2012;**344**:e2432. doi:10.1136/bmj.e2432

13 Rusanov A, Weiskopf NG, Wang S, *et al.* Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak* 2014;**14**:51. doi:10.1186/1472-6947-14-51

14 Arpey NC, Gaglioti AH, Rosenbaum ME. How Socioeconomic Status Affects Patient Perceptions of Health Care: A Qualitative Study. *J Prim Care Community Health* 2017;**8**:169–75. doi:10.1177/2150131917697439

15 Ng JH, Ye F, Ward LM, *et al.* Data On Race, Ethnicity, And Language Largely Incomplete For Managed Care Plan Members. *Health Aff Proj Hope* 2017;**36**:548–52. doi:10.1377/hlthaff.2016.1044

16 Ramoni M, Sebastiani P. Robust Learning with Missing Data. *Mach Learn* 2001;**45**:147–70. doi:10.1023/A:1010968702992

17 Hyun KK, Brieger D, Woodward M, *et al.* The effect of socioeconomic disadvantage on prescription of guideline-recommended medications for patients with acute coronary syndrome: systematic review and meta-analysis. *Int J Equity Health* 2017;**16**. doi:10.1186/s12939-017-0658-z

18 Jabagi H, Tran DT, Hessian R, *et al.* Impact of Gender on Arterial Revascularization Strategies for Coronary Artery Bypass Grafting. *Ann Thorac Surg* 2018;**105**:62–8. doi:10.1016/j.athoracsur.2017.06.054

19 Freij M, Dullabh P, Lewis S, *et al.* Incorporating Social Determinants of Health in Electronic Health Records: Qualitative Study of Current Practices Among Top Vendors. *JMIR Med Inform* 2019;**7**:e13849. doi:10.2196/13849

20 Integrating Data On Social Determinants Of Health Into Electronic Health Records | Health Affairs. https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2017.1252 (accessed 27 Feb 2020).

21 Powell HS, Greenberg DL. Screening for unhealthy diet and exercise habits: The electronic health record and a healthier population. *Prev Med Rep* 2019;**14**:100816. doi:10.1016/j.pmedr.2019.01.020

22 Kight CE, Bouche JM, Curry A, *et al.* Consensus Recommendations for Optimizing Electronic Health Records for Nutrition Care. *Nutr Clin Pract* 2020;**35**:12–23. doi:10.1002/ncp.10433

23 Dinh-Le C, Chuang R, Chokshi S, *et al.* Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions. *JMIR MHealth UHealth* 2019;**7**. doi:10.2196/12861

24 Kahn MG, Callahan TJ, Barnard J, *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.

eGEMs 2016;**4**. doi:10.13063/2327-9214.1244

25 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc JAMIA* 2013;**20**:144–51. doi:10.1136/amiajnl-2011-000681

26 ACHILLES for data characterization – OHDSI. https://www.ohdsi.org/analytic-tools/achilles-for-data-characterization/ (accessed 12 Jan 2019).

27 Blacketer C, Londhe A, Sena A, *et al. Data Quality Dashboard*. Observational Health Data Sciences and Informatics https://ohdsi.github.io/DataQualityDashboard/ (accessed 21 Jan 2020).

28 Blacketer C, Defalco FJ, Ryan PB, *et al.* Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality. *medRxiv* 2021;:2021.03.25.21254341. doi:10.1101/2021.03.25.21254341

29 Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform* 2013;**46**. doi:10.1016/j.jbi.2013.06.010

30 DQUEEN – OHDSI. https://www.ohdsi.org/2019-us-symposium-showcase-106/ (accessed 27 Jan 2020).

31 Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets. *EGEMS Wash DC* 2016;**4**:1239. doi:10.13063/2327-9214.1239

32 Huser, Vojtech. OHDSI Data Quality Study. http://www.ohdsi.org/web/wiki/doku.php?id=research:dqstudy (accessed 12 Jan 2019).

33 Khare R, Utidjian L, Ruth BJ, *et al.* A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc JAMIA* 2017;**24**:1072–9. doi:10.1093/jamia/ocx033

34 Raebel MA, Haynes K, Woodworth TS, *et al.* Electronic clinical laboratory test results data tables: lessons from Mini-Sentinel. *Pharmacoepidemiol Drug Saf* 2014;**23**:609–18. doi:10.1002/pds.3580

35 Data Quality Review and Characterization Programs v4.2.1 | Sentinel Initiative. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/data-quality-review-and-characterization (accessed 12 Jan 2019).

36 Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *EGEMs Gener Evid Methods Improve Patient Outcomes* 2018;**6**:3. doi:10.5334/egems.199

37 Mukherjee S, Xu Y, Trivedi A, *et al.* privGAN: Protecting GANs from membership inference attacks at low cost. *ArXiv200100071 Cs Stat* Published Online First: 13 December 2020.http://arxiv.org/abs/2001.00071 (accessed 17 Mar 2021).

38 Petti S, Flaxman A. Differential privacy in the 2020 US census: what will it do? Quantifying the accuracy/privacy tradeoff. *Gates Open Res* 2020;**3**:1722. doi:10.12688/gatesopenres.13089.2

39 Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;**25**:37–43. doi:10.1038/s41591-018-0272-7

40 Wu L, He H, Zaïane OR. Utility of privacy preservation for health data publishing. In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013. 510–1. doi:10.1109/CBMS.2013.6627853

41 Muniz-Terrera G, Mendelevitch O, Barnes R, *et al.* Virtual Cohorts and Synthetic Data in Dementia: An Illustration of Their Potential to Advance Research. *Front Artif Intell* 2021;**4**. doi:10.3389/frai.2021.613956

42 Thomas JA, Foraker RE, Zamstein N, *et al.* Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: Results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *medRxiv* Published Online First: 2021. doi:10.1101/2021.07.06.21259051

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* Published Online First: 14 December 2020. doi:10.1093/jamiaopen/ooaa060

44 Analyses of Original and Computationally-Derived Electronic Health Record Data: The National COVID Cohort Collaborative. JMIR Prepr. https://preprints.jmir.org/preprint/30697 (accessed 7 Jun 2021).

45 Benaim AR, 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. doi:10.2196/16492

46 Zhang Z, Yan C, Mesa DA, *et al.* Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 2020;**27**:99–108. doi:10.1093/jamia/ocz161

47 El Emam K, Mosquera L, Jonker E, *et al.* Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 2021;**4**. doi:10.1093/jamiaopen/ooab012

48 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Comput Intell*;**n/a**. doi:https://doi.org/10.1111/coin.12427

49 Hittmeir M, Ekelhart A, Mayer R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Canterbury, CA, United Kingdom: : Association for Computing Machinery 2019. 1–6. doi:10.1145/3339252.3339281

50 OMOP CDM v5.3.1. https://ohdsi.github.io/CommonDataModel/cdm531.html (accessed 17 Jul 2021).

51 Bartlett VL, Dhruva SS, Shah ND, *et al.* Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Netw Open* 2019;**2**:e1912869–e1912869. doi:10.1001/jamanetworkopen.2019.12869

52 Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019;**19**. doi:10.1186/s12911-019-0793-0

53 Cochrane Database of Systematic Reviews. https://www.cochranelibrary.com/cdsr/reviews (accessed 18 Jul 2021).

54 *Easy interactive web applications with R. Contribute to rstudio/shiny development by creating an account on GitHub*. RStudio 2019. https://github.com/rstudio/shiny (accessed 24 May 2019).

55 Prieto-Alhambra D, Kostka K, Duarte-Salles T, *et al.* Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. *Res Sq* Published Online First: 1 March 2021. doi:10.21203/rs.3.rs-279400/v1

56 Golozar A, Lai LY, Sena AG, *et al.* Baseline phenotype and 30-day outcomes of people tested for COVID-19: an international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain and the United States. *medRxiv* 2020;:2020.10.25.20218875. doi:10.1101/2020.10.25.20218875

57 Dube K, Gallagher T. Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons J, MacCaull W, eds. *Foundations of Health Information Engineering and Systems*. Berlin, Heidelberg: : Springer 2014. 69–86. doi:10.1007/978-3-642-53956-5_6

58 Walonoski J, Kramer M, Nichols J, *et al.* Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc JAMIA* Published Online First: 30 August 2017. doi:10.1093/jamia/ocx079

59 Shamberger M. Synthetic data with simulated covid outbreak - Developers. OHDSI Forums. 2020.https://forums.ohdsi.org/t/synthetic-data-with-simulated-covid-outbreak/10256 (accessed 17 Jul 2021).

60 *OHDSI/CommonDataModel/PostgreSQL*. Observational Health Data Sciences and Informatics 2021. https://github.com/OHDSI/CommonDataModel (accessed 17 Jul 2021).

61 Athena. https://athena.ohdsi.org/search-terms/start (accessed 23 Jul 2021).

62 synthea/src/main/resources/modules/covid19 at covid19 · synthetichealth/synthea. GitHub. https://github.com/synthetichealth/synthea (accessed 22 Jul 2021).

63 OHDSI/ETL-Synthea: Conversion from Synthea CSV to OMOP CDM. GitHub. https://github.com/OHDSI/ETL-Synthea (accessed 22 Jul 2021).

64 ATLAS: Cohort Definitions. https://atlas.ohdsi.org/#/cohortdefinitions (accessed 18 Jul 2021).

65 *OHDSI/Atlas*. Observational Health Data Sciences and Informatics 2021. https://github.com/OHDSI/Atlas (accessed 18 Jul 2021).

66 Duarte-Salles, Prats-Uribe A, Prieto-Alhambra D, *et al. Charybdis Phenotype Library*. OHDSI Studies 2021. https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis/blob/7907813db56478a12daf67d979809f1337674deb/documents/CharybdisPhenotypeLibrary.csv (accessed 18 Jul 2021).

67 Schuemie M, Suchard M, Ryan P, *et al. FeatureExtraction - Generating Features for a Cohort*. https://ohdsi.github.io/FeatureExtraction/ (accessed 19 Jul 2021).

68 Standardized Derived Elements - OMOP Common Data Model v5.3.1. observational health data sciences and informatics https://ohdsi.github.io/CommonDataModel/cdm531.html#Standardized_Derived_Elements (accessed 21 Jul 2021).

69 Pellicori P, Doolub, G, Wong, CM, Lee, KS, Mangion, K, Ahmad, M, Berry, C, Squire, I, Lambiase, PD, Lyon, A, McConnachie, A, Taylor, RS, Cleland J. COVID-19 and its cardiovascular effects: a systematic review of prevalence studies. *Cochrane Database Syst Rev* Published Online First: 2021. doi:10.1002/14651858.CD013879

70 CDC. Science Brief: Evidence used to update the list of underlying medical conditions that increase a person's risk of severe illness from COVID-19. Cent. Dis. Control Prev. 2020.https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/underlying-evidence-table.html (accessed 20 Jul 2021).

71 Sud K, Vogel B, Bohra C, *et al.* Echocardiographic Findings in Patients with COVID-19 with Significant Myocardial Injury. *J Am Soc Echocardiogr* 2020;**33**:1054–5. doi:10.1016/j.echo.2020.05.030

72 Patterson OV, Freiberg MS, Skanderson M, *et al.* Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc Disord*

2017;**17**:151. doi:10.1186/s12872-017-0580-8

73 Johnson SB, Adekkanattu P, Campion TR, *et al.* From Sour Grapes to Low-Hanging Fruit: A Case Study Demonstrating a Practical Strategy for Natural Language Processing Portability. *AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci* 2018;**2017**:104–12.

2.9 SUPPLEMENT

Table 2.5. Characteristics of non-COVID-19 related Cochrane reviews assessed for replicability in electronic health records

Cochrane Database ID	Review Title	Date	Cochrane Network	Cochrane Group	Eligible in Repository	Eligible Outcomes	Excluded due to NLP need	Pediatric Only	Type	MeSH_Keywords	Doi_link
CD012941	Probiotics for preventing acute otitis media in children	18-Jun-19	Acute and Emergency Care	Acute Respiratory Infections	Yes	1	No	Yes	Intervention	Acute Disease;Anti_Bacterial Agents [therapeutic use];Disease Susceptibility;Otitis Media [epidemiology, *prevention & control];Probiotics [adverse effects, *therapeutic use];Randomized Controlled Trials as Topic [statistics & numerical data];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012941.pub2/full
CD001480	Pneumococcal conjugate vaccines for preventing acute otitis media in children	28-May-19	Acute and Emergency Care	Acute Respiratory Infections	Yes	1	No	Yes	Intervention	*Pneumococcal Vaccines [therapeutic use];Acute Disease;Otitis Media [microbiology, *prevention & control];Otitis Media with Effusion [drug therapy];Vaccines, Conjugate [therapeutic use];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001480.pub5/full
CD010473	Continuous positive airway pressure (CPAP) for acute bronchiolitis in children	31-Jan-19	Acute and Emergency Care	Acute Respiratory Infections	Yes	1	No	Yes	Intervention	Acute Disease;Bronchiolitis [blood, *therapy];Carbon Dioxide;Continuous Positive Airway Pressure [*methods, statistics & numerical data];Length of Stay;Oxygen [blood];Partial Pressure;Randomized Controlled Trials as Topic;Respiration, Artificial [statistics & numerical data];Respiratory Rate;Selection Bias;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010473.pub3/full
CD010126	Inhaled corticosteroids in children with persistent asthma: effects of different drugs and	10-Jun-19	Circulation and Breathing	Airways	Yes	1	No	Yes	Intervention	Administration, Inhalation;Adrenal Cortex Hormones [administration & dosage, *pharmacology];Anti_Asthmatic Agents [administration & dosage, *pharmacology];Asthma [*drug therapy];Beclomethasone [administration & dosage, pharmacology];Body Height	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010126.pub2/full

	delivery devices on growth									[drug effects];Budesonide [administration & dosage, pharmacology];Fluticasone [administration & dosage, pharmacology];Growth [*drug effects];Metered Dose Inhalers;Randomized Controlled Trials as Topic;Time Factors;	
CD006715	Epidural analgesia for adults undergoing cardiac surgery with or without cardiopulmonary bypass	1-Mar-19	Acute and Emergency Care	Anaesthesia	Yes	4	No	No	Intervention	*Cardiac Surgical Procedures [adverse effects, mortality];Analgesia, Epidural [*adverse effects, methods, mortality];Anesthesia, General [*adverse effects, methods, mortality];Arrhythmias, Cardiac [prevention & control];Coronary Artery Bypass [adverse effects, mortality];Myocardial Infarction [*etiology];Randomized Controlled Trials as Topic;Respiration Disorders [etiology];Stroke [*etiology];	https://doi.org/10.1002/14651858.CD006715.pub3
CD001026	Antidepressants plus benzodiazepines for adults with major depression	3-Jun-19	Mental Health and Neuroscience	Common Mental Disorders	Yes	3	No	No	Intervention	*Antidepressive Agents [therapeutic use];*Benzodiazepines [therapeutic use];*Depressive Disorder, Major [drug therapy];Anxiety [drug therapy];Drug Therapy, Combination;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001026.pub2/full
CD012943	Inhaled corticosteroids for cystic fibrosis	4-Jul-19	Children and Families	Cystic Fibrosis and Genetic Disorders	Yes	4	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD012943.pub2
CD013069	Cognitive training for people with mild to moderate dementia	25-Mar-19	Mental Health and Neuroscience	Dementia and Cognitive Improvement	Yes	3	No	No	Intervention	*Cognition;Activities of Daily Living;Cognitive Dysfunction [rehabilitation, *therapy];Dementia [complications, rehabilitation, *therapy];Randomized Controlled Trials as Topic;Task Performance and Analysis;Therapy, Computer_Assisted [methods];	https://doi.org/10.1002/14651858.CD013069.pub2
CD004328	Tranexamic acid for patients with nasal	31-Dec-18	Musculoskeletal, Oral, Skin	ENT	Yes	2	No	No	Intervention	Administration, Oral;Administration, Topical;Antifibrinolytic Agents [administration & dosage, adverse effects, *therapeutic use];Blood Transfusion	https://doi.org/10.1002/14651858.CD004328.pub3

	haemorrhage (epistaxis)		and Sensory							[statistics & numerical data];Epinephrine [therapeutic use];Epistaxis [*drug therapy];Length of Stay;Lidocaine [therapeutic use];Phenylephrine [therapeutic use];Placebos [therapeutic use];Randomized Controlled Trials as Topic;Recurrence;Secondary Prevention [statistics & numerical data];Tranexamic Acid [administration & dosage, adverse effects, *therapeutic use];	
CD010541	Surgery for epilepsy	25-Jun-19	Mental Health and Neuroscience	Epilepsy	Yes	2	No	No	Intervention	Analysis of Variance;Anticonvulsants [therapeutic use];Epilepsies, Partial [drug therapy, *surgery];Hippocampus [surgery];Prognosis;Randomized Controlled Trials as Topic;Retrospective Studies;Treatment Outcome;	https://doi.org/10.1002/14651858.CD010541.pub3
CD012065	Topiramate versus carbamazepine monotherapy for epilepsy: an individual participant data review	24-Jun-19	Mental Health and Neuroscience	Epilepsy	Yes	2	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD012065.pub3
CD004317	Strategies to improve adherence and continuation of shorter-term hormonal methods of contraception	23-Apr-19	Children and Families	Fertility Regulation	Yes	2	No	No	Intervention	*Counseling;*Family Planning Services;Contraception [*methods];Contraceptive Agents, Female [*administration & dosage];Contraceptives, Oral, Hormonal;Pregnancy, Unplanned;	https://doi.org/10.1002/14651858.CD004317.pub5
CD005351	Non-invasive positive pressure ventilation (CPAP or bilevel NPPV) for cardiogenic pulmonary oedema	5-Apr-19	Circulation and Breathing	Heart	Yes	3	No	No	Intervention	*Hospital Mortality;Continuous Positive Airway Pressure [adverse effects, *methods];Intensive Care Units;Intubation, Intratracheal [statistics & numerical data];Length of Stay;Noninvasive Ventilation;Pulmonary Edema [*therapy];Randomized Controlled Trials as Topic;	https://doi.org/10.1002/14651858.CD005351.pub4

CD006150	Adjunctive corticosteroids for Pneumocystis jiroveci pneumonia in patients with HIV infection	2-Apr-15	Public Health and Health Systems	HIV/AIDS	Yes	4	No	No	Intervention	*Pneumocystis carinii;AIDS_Related Opportunistic Infections [*drug therapy];Adrenal Cortex Hormones [*therapeutic use];Chemotherapy, Adjuvant;Hypoxia [etiology, therapy];Pneumonia, Pneumocystis [*drug therapy];Randomized Controlled Trials as Topic;Respiration, Artificial;	https://doi.org/10.1002/14651858.CD006150.pub2
CD000028	Pharmacotherapy for hypertension in adults 60 years or older	5-Jun-19	Circulation and Breathing	Hypertension	Yes	4	No	No	Intervention	*Antihypertensive Agents [therapeutic use];*Hypertension [drug therapy];Coronary Disease [prevention & control];Randomized Controlled Trials as Topic;Stroke [prevention & control];	https://doi.org/10.1002/14651858.CD000028.pub3
CD010406	Corticosteroids as adjunctive therapy in the treatment of influenza	24-Feb-19	Acute and Emergency Care	Acute Respiratory Infections	No	0	No	No	Intervention	Adrenal Cortex Hormones [adverse effects, *therapeutic use];Chemotherapy, Adjuvant [adverse effects];Cross Infection [etiology, mortality];Hospital Mortality;Influenza A Virus, H1N1 Subtype;Influenza, Human [*drug therapy, mortality];Intensive Care Units [statistics & numerical data];Observational Studies as Topic;Randomized Controlled Trials as Topic;Respiration, Artificial [statistics & numerical data];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010406.pub3/full
CD010473	Adverse events in people taking macrolide antibiotics versus placebo for any indication	18-Jan-19	Acute and Emergency Care	Acute Respiratory Infections	No	0	No	No	Intervention	Abdominal Pain [chemically induced];Anti_Bacterial Agents [*adverse effects];Bile Duct Diseases [chemically induced];Diarrhea [chemically induced];Hearing Loss [chemically induced];Heart Diseases [chemically induced];Macrolides [*adverse effects, therapeutic use];Nausea [chemically induced];Numbers Needed To Treat;Placebos;Randomized Controlled Trials as Topic;Taste Disorders [chemically induced];Vomiting [chemically induced];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011825.pub2/full
CD013024	Head_to_head oral prophylactic antibiotic therapy for	24-May-19	Circulation and Breathing	Airways	No	0	No	No	Intervention	Anti_Bacterial Agents [*therapeutic use];Antibiotic Prophylaxis [*methods];Disease Progression;Pulmonary Disease, Chronic	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.C

	chronic obstructive pulmonary disease									Obstructive [*drug therapy];Quality of Life;Treatment Outcome;	D013024.pub2/full
CD012212	Supplemental perioperative intravenous crystalloids for postoperative nausea and vomiting	29-Mar-19	Acute and Emergency Care	Anaesthesia	No	0	Yes	No	Intervention	Administration, Intravenous;Anesthesia, General [*adverse effects];Crystalloid Solutions [administration & dosage, *therapeutic use];Postoperative Nausea and Vomiting [chemically induced, epidemiology, *prevention & control];Randomized Controlled Trials as Topic;Time Factors;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012212.pub2/full
CD011686	Triage tools for detecting cervical spine injury in pediatric trauma patients	7-Dec-17	Musculoskeletal, Oral, Skin and Sensory	Back and Neck	No	0	No	Yes	Diagnostic	*Decision Support Techniques;Cervical Vertebrae [diagnostic imaging, *injuries];Checklist;Cohort Studies;Magnetic Resonance Imaging;Radiography;Reference Standards;Spinal Injuries [*diagnosis, diagnostic imaging, etiology];Tomography, X_Ray Computed;Triage [*methods];Wounds, Nonpenetrating [*complications, diagnostic imaging];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011686.pub2/full
CD011674	Back Schools for chronic non-specific low back pain	3-Aug-17	Musculoskeletal, Oral, Skin and Sensory	Back and Neck	No	0	No	No	Intervention	Chronic Pain [*therapy];Disability Evaluation;Exercise Therapy [*methods];Low Back Pain [*therapy];Pain Measurement;Patient Education as Topic [*methods, organization & administration];Randomized Controlled Trials as Topic;Time Factors;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011674.pub2/full
CD004962	Conservative management following closed reduction of traumatic anterior dislocation of the shoulder	10-May-19	Acute and Emergency Care	Bone, Joint, and Muscle Trauma	No	0	No	No	Intervention	*Conservative Treatment;Immobilization [adverse effects, *methods];Joint Instability [etiology];Randomized Controlled Trials as Topic;Shoulder Dislocation [complications, *therapy];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004962.pub4/full

CD012424	Exercise for preventing falls in older people living in the community	31-Jan-19	Acute and Emergency Care	Bone, Joint, and Muscle Trauma	No	0	Yes	No	Intervention	*Exercise;*Independent Living;Accidental Falls [*prevention & control, statistics & numerical data];Dance Therapy [statistics & numerical data];Exercise Therapy [*statistics & numerical data];Fractures, Bone [epidemiology, prevention & control];Gait;Postural Balance;Quality of Life;Randomized Controlled Trials as Topic;Resistance Training [statistics & numerical data];Tai Ji [statistics & numerical data];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012424.pub2/full
CD011518	Mindfulness-based stress reduction for women diagnosed with breast cancer	27-Mar-19	Cancer	Breast Cancer	No	0	No	No	Intervention	*Mindfulness;Anxiety [psychology];Breast Neoplasms [*psychology];Depression [psychology];Fatigue [psychology];Quality of Life;Randomized Controlled Trials as Topic;Sleep Wake Disorders [psychology];Stress, Psychological [*therapy];Time Factors;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011518.pub2/full
CD009219	Medical interventions for the prevention of platinum-induced hearing loss in children with cancer	7-May-19	Cancer	Childhood Cancer	No	0	No	Yes	Intervention	Antineoplastic Agents [*adverse effects, therapeutic use];Carboplatin;Cisplatin;Hearing Loss [*chemically induced, *prevention & control];Neoplasms [drug therapy];Organoplatinum Compounds [*adverse effects, therapeutic use];Oxaliplatin;Randomized Controlled Trials as Topic;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD009219.pub5/full
CD012442	Anti-GD2 antibody-containing immunotherapy postconsolidation therapy for people with high-risk neuroblastoma treated with autologous haematopoietic stem cell transplantation	24-Apr-19	Cancer	Childhood Cancer	No	0	No	Yes	Intervention	*Hematopoietic Stem Cell Transplantation;Antibodies, Monoclonal [*therapeutic use];Consolidation Chemotherapy;Disease_Free Survival;Immunologic Factors [therapeutic use];Immunotherapy;Neuroblastoma [immunology, mortality, *therapy];Randomized Controlled Trials as Topic;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012442.pub2/full

CD008237	Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy	17-Aug-18	Abdomen and Endocrine	Colorectal Cancer	No	0	No	No	Intervention	*Clinical Competence;*Virtual Reality;Endoscopy, Gastrointestinal [*education];Health Personnel [*education];Randomized Controlled Trials as Topic;Simulation Training [*methods];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD008237.pub3/full
CD010168	Abdominal drainage to prevent intra-peritoneal abscess after open appendectomy for complicated appendicitis	9-May-18	Abdomen and Endocrine	Colorectal Cancer	No	0	No	No	Intervention	Abdominal Abscess [*prevention & control];Appendectomy [*adverse effects];Appendicitis [complications, *surgery];Drainage [*methods];Emergencies;Length of Stay;Peritoneal Diseases [*prevention & control];Postoperative Complications [*prevention & control];Randomized Controlled Trials as Topic;	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010168.pub3/full
CD011271	Melatonin and agomelatine for preventing seasonal affective disorder	17-Jun-19	Mental Health and Neuroscience	Common Mental Disorders	No	0	No	No	Intervention	Acetamides [*therapeutic use];Antidepressive Agents [*therapeutic use];Melatonin [agonists, *therapeutic use];Placebos [therapeutic use];Seasonal Affective Disorder [*prevention & control];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD011271.pub3/full
CD009912	Psychosocial interventions for informal caregivers of people living with cancer	17-Jun-19	Public Health and Health Systems	Consumers and Communication	No	0	No	No	Intervention	*Anxiety [therapy];*Caregivers [psychology];*Depression [therapy];*Neoplasms [nursing];Health Status;Quality of Life;Randomized Controlled Trials as Topic;	https://doi.org/10.1002/14651858.CD009912.pub2
CD012533	Telephone interventions, delivered by healthcare professionals, for providing education and psychosocial support for informal caregivers of adults with	14-May-19	Public Health and Health Systems	Consumers and Communication	No	0	No	No	Intervention	*Chronic Disease [psychology];*Psychosocial Support Systems;*Telephone;Adaptation, Psychological;Anxiety [psychology];Caregivers [*psychology];Depression [psychology];Family;Mental Health;Quality of Life;Randomized Controlled Trials as Topic;Stress, Psychological [*psychology];	https://doi.org/10.1002/14651858.CD012533.pub2

	diagnosed illnesses										
CD012943	Interventions for treating neuropathic pain in people with sickle cell disease	5-Jul-19	Children and Families	Cystic Fibrosis and Genetic Disorders	No	0	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD012943.pub2
CD009537	Interventions for preventing delirium in older people in institutional long-term care	23-Apr-19	Mental Health and Neuroscience	Dementia and Cognitive Improvement	No	0	No	No	Intervention	*Long_Term Care;Activities of Daily Living;Delirium [chemically induced, epidemiology, *prevention & control];Frail Elderly;Incidence;Institutionalization;Medication Reconciliation;Quality of Life;Randomized Controlled Trials as Topic;	https://doi.org/10.1002/14651858.CD009537.pub3
CD013135	A realist review of which advocacy interventions work for which abused women under what circumstances	29-Jun-19	Mental Health and Neuroscience	Developmental, Psychosocial and Learning Problems	No	0	No	No	Prototype	NaN	https://doi.org/10.1002/14651858.CD013135.pub2
CD008223	Social skills training for attention deficit hyperactivity disorder (ADHD) in children aged 5 to 18 years	21-Jun-19	Mental Health and Neuroscience	Developmental, Psychosocial and Learning Problems	No	0	No	No	Intervention	*Attention Deficit Disorder with Hyperactivity [therapy];*Behavior Therapy;*Social Skills;Cognitive Behavioral Therapy;Interpersonal Relations;	https://doi.org/10.1002/14651858.CD008223.pub3
CD012287	Family-based prevention programmes for alcohol use in young people	19-Mar-19	Mental Health and Neuroscience	Drugs and Alcohol	No	0	No	No	Intervention	*Family Health;*Family Therapy [methods];*Program Evaluation;Alcohol Drinking [epidemiology, *prevention & control];Prevalence;Randomized Controlled Trials as Topic;	https://doi.org/10.1002/14651858.CD012287.pub2
CD008940	Pharmacotherapies for cannabis dependence	28-Jan-19	Mental Health and	Drugs and Alcohol	No	0	Yes	No	Intervention	Acetylcysteine [adverse effects, therapeutic use];Anticonvulsants [adverse effects, therapeutic use];Antidepressive	https://doi.org/10.1002/14

			Neuroscience							Agents [adverse effects, therapeutic use];Buspirone [adverse effects, therapeutic use];Dronabinol [adverse effects, therapeutic use];Marijuana Abuse [*drug therapy];Randomized Controlled Trials as Topic;Serotonin Receptor Agonists [adverse effects, therapeutic use];Serotonin Uptake Inhibitors [therapeutic use];	651858.CD008940.pub3
CD011156	Pay for performance for hospitals	5-Jul-19	Public Health and Health Systems	Effective Practice and Organisation of Care	No	0	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD011156.pub2
CD000125	Local opinion leaders: effects on professional practice and healthcare outcomes	24-Jun-19	Public Health and Health Systems	Effective Practice and Organisation of Care	No	0	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD000125.pub5
CD011811	Plasma interleukin-6 concentration for the diagnosis of sepsis in critically ill adults	30-Apr-19	Acute and Emergency Care	Emergency and Critical Care	No	0	No	No	Diagnostic	Biomarkers [blood];Critical Illness;Diagnosis, Differential;Interleukin_6 [*blood];Sepsis [*diagnosis];	https://doi.org/10.1002/14651858.CD011811.pub2
CD013315	Interventions for preventing high altitude illness: Part 3. Miscellaneous and non-pharmacological interventions	23-Apr-19	Acute and Emergency Care	Emergency and Critical Care	No	0	No	No	Intervention	Acetazolamide [therapeutic use];Altitude Sickness [*prevention & control];Brain Edema [prevention & control];Hypertension, Pulmonary [prevention & control];Medroxyprogesterone [therapeutic use];Plant Extracts [therapeutic use];Randomized Controlled Trials as Topic;	https://doi.org/10.1002/14651858.CD013315
CD012173	Restriction of salt, caffeine and alcohol intake for the	31-Dec-18	Musculoskeletal, Oral, Skin and Sensory	ENT	No	0	No	No	Intervention	*Caffeine;*Central Nervous System Stimulants;*Diet, Sodium_Restricted;*Sodium Chloride, Dietary;Meniere Disease [*therapy];Syndrome;	https://doi.org/10.1002/14651858.CD012173.pub2

	treatment of Ménière's disease or syndrome										
CD013000	Interventions for orbital lymphangioma	15-May-19	Musculoskeletal, Oral, Skin and Sensory	Eyes & Vision	No	0	No	No	Intervention	*Lymphangioma [drug therapy, surgery];*Orbital Neoplasms [drug therapy, surgery];Antibiotics, Antineoplastic [therapeutic use];Treatment Outcome;	https://doi.org/10.1002/14651858.CD013000.pub2
CD011150	Intrastromal corneal ring segments for treating keratoconus	14-May-19	Musculoskeletal, Oral, Skin and Sensory	Eyes & Vision	No	0	No	No	Intervention	Corneal Stroma [*surgery];Corneal Transplantation [methods];Keratoconus [*surgery];Prostheses and Implants;Prosthesis Implantation [*methods];	https://doi.org/10.1002/14651858.CD011150.pub2
CD012521	Interventions using social networking sites to promote contraception in women of reproductive age	1-Mar-19	Children and Families	Fertility Regulation	No	0	No	No	Intervention	*Health Knowledge, Attitudes, Practice;*Online Social Networking;Condoms [statistics & numerical data];Contraception [*statistics & numerical data];Contraception Behavior [*statistics & numerical data];Randomized Controlled Trials as Topic;Sexual Health [*statistics & numerical data];	https://doi.org/10.1002/14651858.CD012521.pub2
CD009825	Mediterranean-style diet for the primary and secondary prevention of cardiovascular disease	13-Mar-19	Circulation and Breathing	Heart	No	0	No	No	Intervention	*Diet, Mediterranean;Blood Pressure;Cardiovascular Diseases [blood, mortality, *prevention & control];Cholesterol [blood];Cholesterol, HDL [blood];Cholesterol, LDL [blood];Primary Prevention [*methods];Randomized Controlled Trials as Topic;Secondary Prevention [*methods];	https://doi.org/10.1002/14651858.CD009825.pub3
CD004039	Plasma expanders for people with cirrhosis and large ascites treated with abdominal paracentesis	28-Jun-19	Abdomen and Endocrine	Hepato_Biliary Group	No	0	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD004039.pub2

CD013106	Radix Sophorae flavescentis versus other drugs or herbs for chronic hepatitis B	24-Jun-19	Abdomen and Endocrine	Hepato_Biliary Group	No	0	No	No	Intervention	NaN	https://doi.org/10.1002/14651858.CD013106.pub2
CD007497	Low dose versus high dose stavudine for treating people with HIV infection	28-Jan-15	Public Health and Health Systems	HIV/AIDS	No	0	No	No	Intervention	Anti_HIV Agents [*administration & dosage, adverse effects];Developing Countries;HIV Infections [*drug therapy, virology];HIV_1;Randomized Controlled Trials as Topic;Stavudine [*administration & dosage, adverse effects];Viral Load [drug effects];	https://doi.org/10.1002/14651858.CD007497.pub2
CD012873	Sequencing of anthracyclines and taxanes in neoadjuvant and adjuvant therapy for early breast cancer	18-Feb-19	Cancer	Breast Cancer	No	0	No	No	Intervention	Anthracyclines [*administration & dosage, adverse effects];Antibiotics, Antineoplastic [*administration & dosage, adverse effects];Antineoplastic Agents [*administration & dosage, adverse effects];Breast Neoplasms [*drug therapy, mortality, pathology];Chemotherapy, Adjuvant;Cyclophosphamide [administration & dosage, adverse effects];Disease_Free Survival;Docetaxel [administration & dosage, adverse effects];Doxorubicin [administration & dosage, adverse effects];Drug Administration Schedule;Epirubicin [administration & dosage, adverse effects];Fluorouracil [administration & dosage, adverse effects];Neoadjuvant Therapy;Nervous System [drug effects];Neutropenia [chemically induced];Paclitaxel [administration & dosage, adverse effects];Quality of Life;Randomized Controlled Trials as Topic;Taxoids [*administration & dosage, adverse effects];	https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012873.pub2/full

Chapter 3. DEMONSTRATING AN APPROACH FOR EVALUATING SYNTHETIC GEOSPATIAL AND TEMPORAL EPIDEMIOLOGIC DATA UTILITY: RESULTS FROM ANALYZING >1.8 MILLION SARS-COV-2 TESTS IN THE UNITED STATES COVID COHORT COLLABORATIVE (N3C)

3.1 ABSTRACT

Objective: To evaluate whether synthetic data derived from a national COVID-19 data set could be used for geospatial and temporal epidemic analyses.

Materials and Methods: Using an original data set (n=1,854,968 SARS-CoV-2 tests) and its synthetic derivative, we compared key indicators of COVID-19 community spread through analysis of aggregate and zip-code level epidemic curves, patient characteristics and outcomes, distribution of tests by zip code, and indicator counts stratified by month and zip code. Similarity between the data was statistically and qualitatively evaluated.

Results: In general, synthetic data closely matched original data for epidemic curves, patient characteristics, and outcomes. Synthetic data suppressed labels of zip codes with few total tests (mean=2.9±2.4; max=16 tests; 66% reduction of unique zip codes). Epidemic curves and monthly indicator counts were similar between synthetic and original data in a random sample of the most tested (top 1%; n=171) and for all unsuppressed zip codes (n=5,819), respectively. In small sample sizes, synthetic data utility was notably decreased.

Discussion: Analyses on the population-level and of densely-tested zip codes (which contained most of the data) were similar between original and synthetically-derived data sets. Analyses of sparsely-tested populations were less similar and had more data suppression.

Conclusion: In general, synthetic data were successfully used to analyze geospatial and temporal trends. Analyses using small sample sizes or populations were limited, in part due to purposeful data label suppression - an attribute disclosure countermeasure. Users should consider data fitness for use in these cases.

3.2 INTRODUCTION

3.2.1 *Background and significance*

COVID-19 has illustrated the need to disseminate accurate, timely, and useful epidemiologic public health data - especially data related to ongoing pandemics or pandemic preparedness. It has also highlighted the need to protect the privacy of individuals.[1,2] The National COVID Cohort Collaborative (N3C) was created to share and harmonize individual-level electronic health record (EHR) data into a single data set.[3] The N3C has received, ingested, harmonized, and characterized[4] data from across the United States (US). To balance data access and privacy, N3C created two levels of data sets: (1) the limited data set (LDS) which has 16 HIPAA Privacy Rule[5,6] direct identifiers stripped out except dates and zip codes, and (2) synthetic data which are computationally derived from the LDS to mimic the LDS data statistical distributions, covariance, and higher order interactions. Synthetic data generation can potentially protect privacy because synthetic data rows are not directly tied to the original source data.[7–11] Pending a pilot study and privacy validation, synthetic data sets are the only data under consideration to be shared outside of the N3C enclave.[3]

Applying privacy-preserving methods to data comes at varying cost to utility, producing a privacy-utility trade-off.[9,12–15] De-identification removes granular geographic information such as street-level address. Obscuring dates reduces the utility of temporal data for some analyses, such as epidemic curves. However, these geographic and temporal data are critical components needed to measure key indicators of COVID-19 community spread[16] used to inform pandemic management decisions such as determining when to reopen schools[17] and businesses.[18] Thus, synthetic data are likely the only privacy-preserving N3C data that can be used to analyze some of the most critically-important data related to pandemic management and preparedness while also providing citizens more transparency into the underlying data. However, previous research has reported deficits in how well synthetic data mimic original data including limitations in their: ability to capture longitudinal relationships, model multiple data types, and perform well on small sample sizes.[10,19,20] Due to the combination of potential widespread synthetic data dissemination, heightened research interest in COVID-19[21], and the rise of "citizen science"[22–24], the user base and applications of pandemic-related synthetic data will likely be heterogenous and broad. Therefore, it is important to evaluate N3C synthetic data in a manner that can inform users with a wide range of intended use cases and definitions for synthetic data fitness for use.[25]

The utility of synthetic health data has been evaluated in other work[15,19,20,26–30] outside of N3C which applied a variety of the ways one can validate synthetic data.[31] However, N3C synthetic data utility has only been evaluated once before. Recently, the N3C synthetic data validation task team evaluated the utility of N3C synthetic data (MDClone, Beer Sheva, Israel) across three use cases, one of which had a geospatial and temporal focus.[32] Foraker et al. (2021) found the synthetic data had high utility for construction of a single

aggregate epidemic curve of COVID-19 cases. However, it showed that rural zip codes with smaller population counts were more likely than urban zip codes to have zip code labels censored (suppressed) in the synthetic data, which is where a categorical variable's value is replaced with the word 'censored' to protect privacy of patients with particularly uncommon, and thus identifiable, features. To date, no analyses have been conducted on the N3C synthetic data to assess utility for analyses by individual zip codes and/or aggregate indicators beyond case counts (e.g. percent positive) over time.

3.2.2 *Objective*

In this paper, we describe the N3C Synthetic data validation task team methods and results focused on evaluating whether synthetic N3C data can be used for geospatial and temporal epidemic analyses. Our replication studies focused on what we deemed were important and common analyses to be performed, such as epidemic curves for key indicators and creation of public-facing dashboards.[33–35] Our validation included replication of studies and general utility metrics[31] for: analyses at the zip code level over time, construction of epidemic curves, and aggregate population characteristics. We believe these approaches balance the need to provide broad utility results for a wide range of analyses while also providing specific validation results relevant to analyses of common interest.

3.3 Materials and Methods

3.3.1 *Data*

The N3C data analyzed include individual-level EHR data enriched with social determinants of health (SDOH) at the 5-digit zip code level. The data have been harmonized into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) v5.3.1[3,36]

and are the same data sets described in a previous N3C synthetic data validation use case.[32] The N3C LDS as of November 30, 2020 - which included 34 data source partners - was used as the data source. MDClone received a copy of the LDS then transformed these data from the N3C harmonized data model into MDClone's data model. Afterwards, the required data needed for the study team's analyses were extracted by MDClone from the transformed LDS for use as the "original" data set. A synthetic derivative of this transformed original data set was then created by MDClone. MDClone provided both the original and synthetic data sets to the research team for evaluation within the N3C secure enclave environment (flowchart Figure 3.8 in supplement).

Both the original and synthetic data were formatted as a single table adhering to the same schema, with each row representing a single COVID-19 test. The table had the following columns: test result (positive/negative; only each patient's first negative and/or first positive test included), age at confirmed test result; admission start date days from reference if admission occurred within ±7 days of COVID-19 positive test result; death (null/yes) during admission; admission length of stay (LOS); patient's state of residence; source partner with which the patient was affiliated; and patient's 5-digit zip code. The data also included the following SDOH columns determined by the patient's zip code: total population in zip code; percent of residents under the poverty line; percent without health insurance; and median household income.

As in Foraker et al. (2021), we used consistent definitions for censored and uncensored zip codes. Censored zip codes were those present within the original data not found (n=11,222) within the synthetic data set either because the zip code was suppressed by labeling the zip code 'censored' or removed within the synthetic data set to protect privacy. Conversely, uncensored zip codes were defined as discrete zip codes found in the original and the synthetic data (n=5,819).

3.3.2 *Analysis*

All analyses were conducted solely by one author (JAT). All code was written in Python (v3.6.10) and - as required by N3C - ran within the secure N3C enclave using the Palantir Foundry Analytic Platform (Palantir Technologies, Denver, CO). The entirety of code used in this analysis is contained within a single Foundry Code Workbook using a saved Spark environment to preserve required software versions and dependencies. The code workbook and source data have been stored within the N3C enclave so that they may inform and be reused in future validation work.

3.3.3 *Summary of data*

Descriptive statistics were calculated and reported for age, number of unique zip codes present, LOS, and admission date after positive test stratified by patients who were tested, positive, admitted, and who died during admission. Number of unique zip codes present excluded null or censored zip codes. The difference between original and synthetic values was reported as the raw synthetic difference (synthetic - original). The difference as a percentage of the original value was reported as synthetic difference percentage (raw synthetic difference/original).

3.3.4 *Aggregate epidemic curves*

We constructed aggregate epidemic curves using each data set spanning January 1st through November 30th 2020. The following key indicators were calculated and visualized: tests, cases (reproduced from Foraker et al., 2021, to view others in context), percent positive, admissions, and deaths during admission. Each indicator had the following daily metrics calculated: count (discrete indicators) or value (continuous indicators), 7-day midpoint moving average, 7-day slope (count or daily value - its value six days prior). To assess the statistical difference between

original and synthetic epidemic curves, we conducted the paired two-sided t-test (scipy v1.5.3, stats.ttest_rel) and two-sided wilcoxon signed-rank test (scipy v1.5.3, stats.wilcoxon) for all metrics across all indicators, treating each data set's daily results as a pair.

3.3.5 *Distribution of tests; censoring of zip codes*

To assess the distribution of tests by zip code and threshold of zip code censoring, we calculated the total number of tests per zip code in the original and synthetic data. In the synthetic data, we excluded rows with a censored (n=44,337; 2.4%) or null (n=444,092; 23.9%) zip code. In the original data, we excluded rows with a null (n=444,380; 24.0%) zip code. We computed the 99th, 97.5th, and 90th percentiles of tests per zip code in the original data. The distributions of tests by zip code were plotted as a histogram with the synthetic and original data overlaid. Additionally, we calculated the distribution of tests by zip code in the original data that were censored in the synthetic data, then plotted the result as a histogram. We then calculated the difference in patients' SDOH values within the original data, comparing patients whose zip codes were censored within the synthetic data to those whose zip codes were not censored.

3.3.6 *Top 1% paired zip codes' epidemic curves*

Next, we assessed synthetic epidemic curves' performance at the zip code level, focusing on zip codes with relatively abundant data. We created a list of zip codes from the original data in the 99th percentile (n=171) by total number of tests, then removed any zip codes without an uncensored matched zip code pair in the synthetic data (n=0). We randomly sampled ten zip codes from the list and constructed epidemic curves for these zip codes' original and synthetic data. Each epidemic curve was constructed using the same date range, methods, and metrics as the aggregate epidemic curves described above with the following change: we only assessed tests

and admissions indicators due to the infrequency of death during admission at the zip code level and manuscript space limitations.

3.3.7 *Monthly zip code pairwise synthetic error*

We compared the difference in monthly counts of tests, cases, and admissions between the original data and paired uncensored synthetic zip codes. To do so, we calculated each data set's number of tests, cases and admissions for every zip code stratified by month for each month the zip code had ≥ 1 test. Then, the data sets were outer merged on month and zip code (Figure 3.1). Synthetic error, defined as the difference between the synthetic monthly count and the original data monthly count value, was computed for every zip code month pair. The distribution of synthetic error was visualized for tests, cases, and admissions.

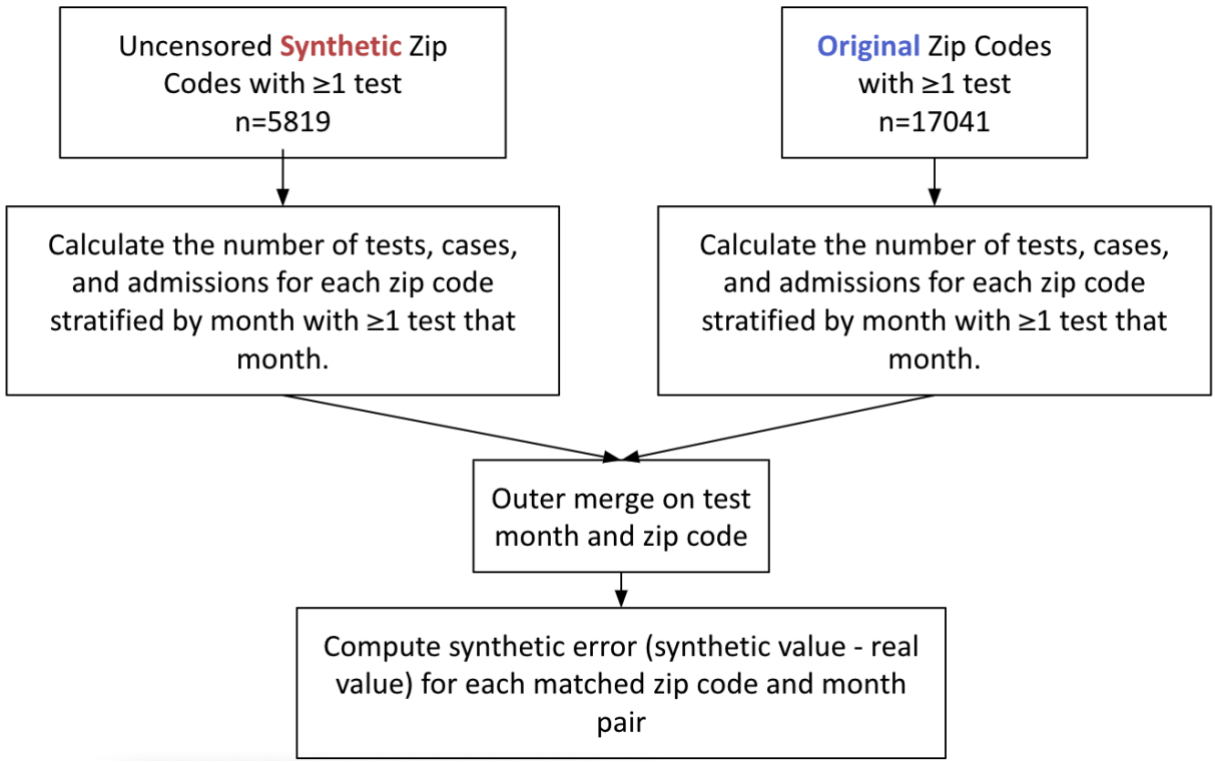

Figure 3.1. Workflow of synthetic error stratified by zip code and month analysis. Workflow of synthetic error experiment showing synthetic data on the left, original data on the right which are then merged to allow the calculation of synthetic error to be made.

3.3.8 *Visualizations*

All visualizations (Plotly v4.14.1, Plotly Technologies Inc.) were interactive, allowing N3C enclave users to zoom in/out, pan, and hover to see values and/or labels. In this manuscript, static figures are presented. Log scales were avoided when possible and, when used, annotated to draw attention to the scale.

Visualizations that overlaid both datasets adhered to consistent style conventions. We encoded synthetic and original data sources as red and blue, respectively. Vertical overlaid bars were set to an opacity of 0.35 to 1) provide contrast between two datasets and 2) allow additional tracings, such as 100% opacity 7-day moving averages used in epidemic curves, to be seen on top of the bars.

All visualizations were created using colorblind-safe color mappings. Categorical mappings encoding values besides data source (synthetic or original) used hexadecimal color codes found in the seaborn colorblind palette.[37,38] Each visualization was qualitatively tested for colorblind deuteranopia, protanopia, and tritanopia interpretability by one member of the research team (JAT) using Color Oracle.[39]

3.4 RESULTS

There were nearly two million tested patients (Original n=1,854,968; Synthetic n=1,854,950) in each data set. As seen in Table 3.6, the overall central tendencies of variables of interest overall were similar between the synthetic data and original data, especially for age and percent positive/admitted/died. The raw synthetic difference was zero, rounded to two decimal points, roughly one third (18/50 rows in Table 3.6) of the time. The variable with the greatest synthetic

difference was unique zip codes, with between a 65-98% reduction in unique zip codes. Median LOS and IQR for admitted patients were exactly the same, yet the mean LOS was 6.48 (±290.81) and 8.32 (±10.66) days for original and synthetic values, respectively. The extreme LOS standard deviation observed in the original data was due to an erroneous outlier. A single row in the original data had an extreme negative LOS [~-44,000 days; ~-120 years] and 11 rows with a LOS=-1. The synthetic data also had negative LOS values (n<10) but the values were greatly attenuated, ranging from -1 to roughly -175.

Table 3.6. Testing and outcomes characteristics: comparison of original vs synthetic data

	Original	**Synthetic**	**Synthetic difference (raw)**	**Synthetic difference (%)**
Tests (n)	1854968	1854950	-18.00	0.00
Age (mean)	44	44	0.00	0.00
Age (stdev)	22.16	22.16	0.00	0.00
Age (median)	43.52	43.51	-0.01	-0.02
Age (IQR)	35.08	35.04	-0.04	-0.11
Unique zip codes (n)	17041	5819	-11222.00	-65.85
Positive (count)	195200	195198	-2.00	0.00
Positive (%)	10.52	10.52	0.00	0.00
Age (mean)	41.54	41.53	-0.01	-0.02
Age (stdev)	20.4	20.42	0.02	0.10
Age (median)	39.65	39.56	-0.09	-0.23
Age (IQR)	31.84	31.81	-0.03	-0.09
Unique zip codes (n)	6660	1798	-4862.00	-73.00
Negative (n)	1659768	1659752	-16.00	0.00
Negative (%)	89.48	89.48	0.00	0.00
Age (mean)	44.29	44.29	0.00	0.00
Age (stdev)	22.34	22.34	0.00	0.00
Age (median)	44.08	44.08	0.00	0.00
Age (IQR)	35.36	35.34	-0.02	-0.06
Unique zip codes (n)	16668	5805	-10863.00	-65.17
Admitted (n)	23044	23044	0.00	0.00
Admitted (%)	1.24	1.24	0.00	0.00
Age (mean)	57.87	57.85	-0.02	-0.03
Age (stdev)	19.77	19.74	-0.03	-0.15
Age (median)	59.98	60	0.02	0.03
Age (IQR)	28.2	28.22	0.02	0.07
Days after positive test (mean)	-0.07	-0.1	-0.03	42.86
Days after positive test (stdev)	1.77	1.74	-0.03	-1.69

Days after positive test (median)	-0.05	-0.04	0.01	-20.00
Days after positive test (iqr)	0.88	0.88	0.00	0.00
LOS (mean)	6.48	8.32	1.84	28.40
LOS (stdev)	290.81	10.66	-280.15	-96.33
LOS (median)	5	5	0.00	0.00
LOS (IQR)	8	8	0.00	0.00
Unique zip codes (n)	3132	1515	-1617.00	-51.63
Died (n)	2032	2032	0.00	0.00
Died (%)	0.11	0.11	0.00	0.00
Age (mean)	71.81	71.81	0.00	0.00
Age (stdev)	14.57	14.65	0.08	0.55
Age (median)	73.26	73.21	-0.05	-0.07
Age (IQR)	19.68	19.58	-0.10	-0.51
Days after positive test (mean)	-0.32	-0.32	0.00	0.00
Days after positive test (stdev)	1.39	1.36	-0.03	-2.16
Days after positive test (median)	-0.14	-0.11	0.03	-21.43
Days after positive test (IQR)	0.91	0.93	0.02	2.20
LOS (mean)	13.69	13.71	0.02	0.15
LOS (stdev)	12.93	13.05	0.12	0.93
LOS (median)	10	10	0.00	0.00
LOS (IQR)	13	13	0.00	0.00
Unique zip codes (n)	831	16	-815.00	-98.07

Aggregate epidemic curves are shown in Figure 3.2. In our statistical analysis, no differences were found between the aggregate epidemic curves besides the 7-day average of percent positive [(t-test p-value=0.025; wilcoxon p-value=0.072), Table 3.7].

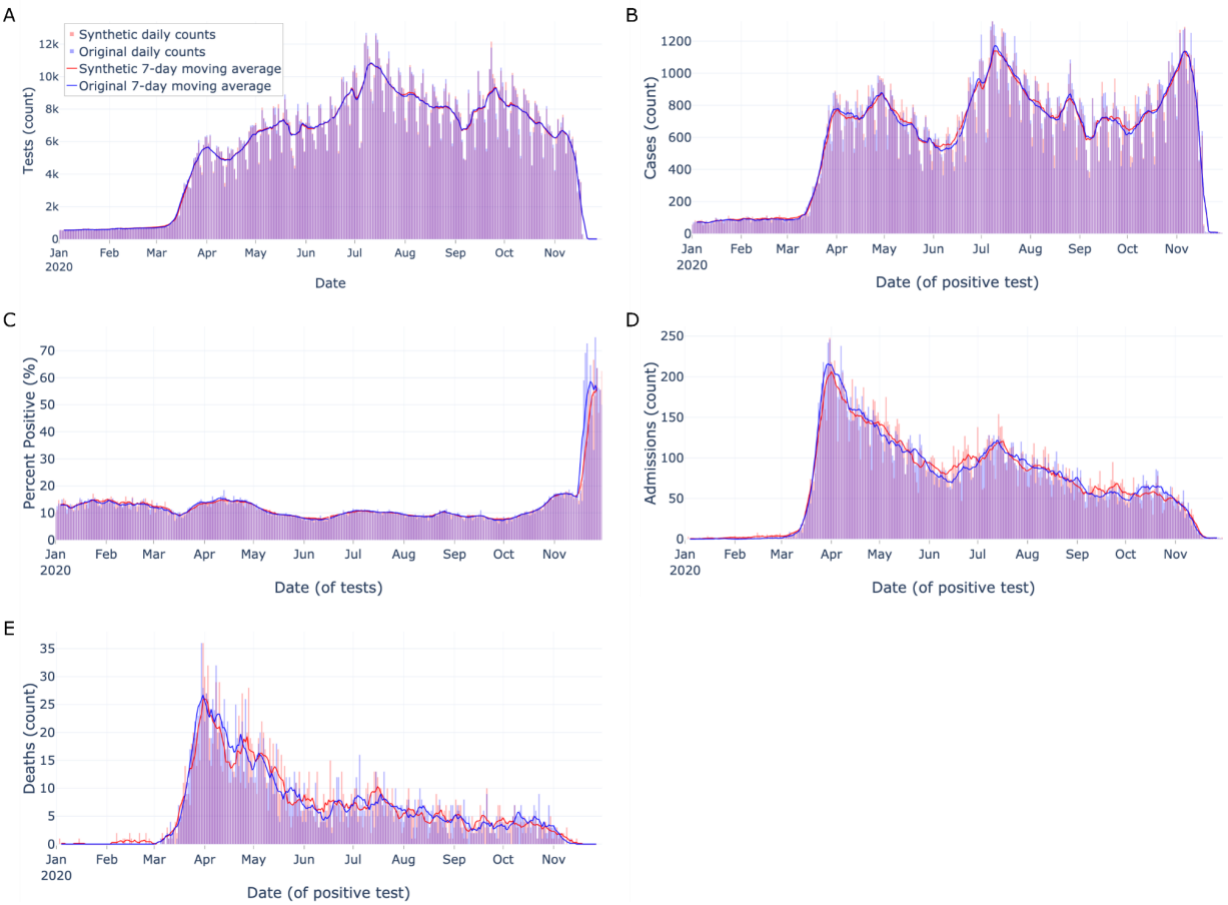

Figure 3.2. Aggregate epidemic curves of key indicators.

Aggregate epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line) for A) tests, B) cases, C) percent positive, D) admissions, and E) deaths during admission. Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). As counts get smaller from tests to deaths, the epidemic curves visually appear less similar.

Table 3.7. Tests for significant differences between aggregate original and synthetic epidemic curves.

Key Indicator	**Metric**	**Wilcoxon result**	**P-value**	**T-test stat**	**P-value**
Tests	Counts	25354.5	0.300	-0.007	0.994
	7-day average	25458.5	0.428	-0.025	0.980

	7-day slope	26075	0.735	-0.002	0.998
Cases	Counts	26288	0.496	-0.002	0.998
	7-day average	26005	0.775	-0.006	0.996
	7-day slope	25788.5	0.898	-0.002	0.998
Percent Positive	Counts	26407	0.426	-0.932	0.352
	7-day average	24038	0.072	-2.258	**0.025**
	7-day slope	27083	0.972	0.129	0.896
Admissions	Counts	21405	0.247	-0.007	0.995
	7-day average	24299	0.197	-0.030	0.976
	7-day slope	22825.5	0.894	-0.011	0.991
Deaths	Counts	13881	0.748	0	1
	7-day average	19171.5	0.247	-0.023	0.982
	7-day slope	16632	0.866	-0.011	0.992

Differences were observed between patients' SDOH values whose zip codes were uncensored in the synthetic data compared to patients whose zip codes were censored in the synthetic data (Table 3.8). The largest differences were found in the total population of zip code and age. Patients with uncensored zip codes lived in more populous zip codes (median median total population: uncensored=28,479, censored=7,935) and were younger (median age: uncensored=43.5, censored=48.7).

Table 3.8. SDOH and age of patients in the original data whose zip codes were censored vs. uncensored

SDOH	Censored Status	Mean	Standard deviation	Median	IQR
Age (years)	Uncensored	44.0	22.2	43.5	35.0
	Censored	46.4	22.0	48.7	40.1
	Uncensored Difference (raw)	-2.4	0.2	-5.2	-5.1
Median household income ($)	Uncensored	64092.6	23973.9	59324.0	29241.0
	Censored	63101.5	28964.1	55625.0	28857.0
	Uncensored Difference (raw)	991.1	-4990.2	3699.0	384.0
Percent under the poverty line	Uncensored	13.7	9.0	11.3	11.2
	Censored	13.3	9.6	11.2	10.9
	Uncensored Difference (raw)	0.4	-0.6	0.1	0.3
Percent without health insurance	Uncensored	8.7	5.1	7.6	7.0
	Censored (raw)	9.2	6.7	7.8	7.7
	Uncensored Difference (raw)	-0.5	-1.6	-0.2	-0.7
Total population of zip code	Uncensored	29758.7	17992.4	28479.0	25220.0
	Censored	15493.9	17967.1	7935.0	23119.3
	Uncensored Difference (raw)	14264.8	25.3	20544.0	2100.7

The randomly sampled top 1% paired zip codes' epidemic curves are presented in Figure 3.3 and Figure 3.4.

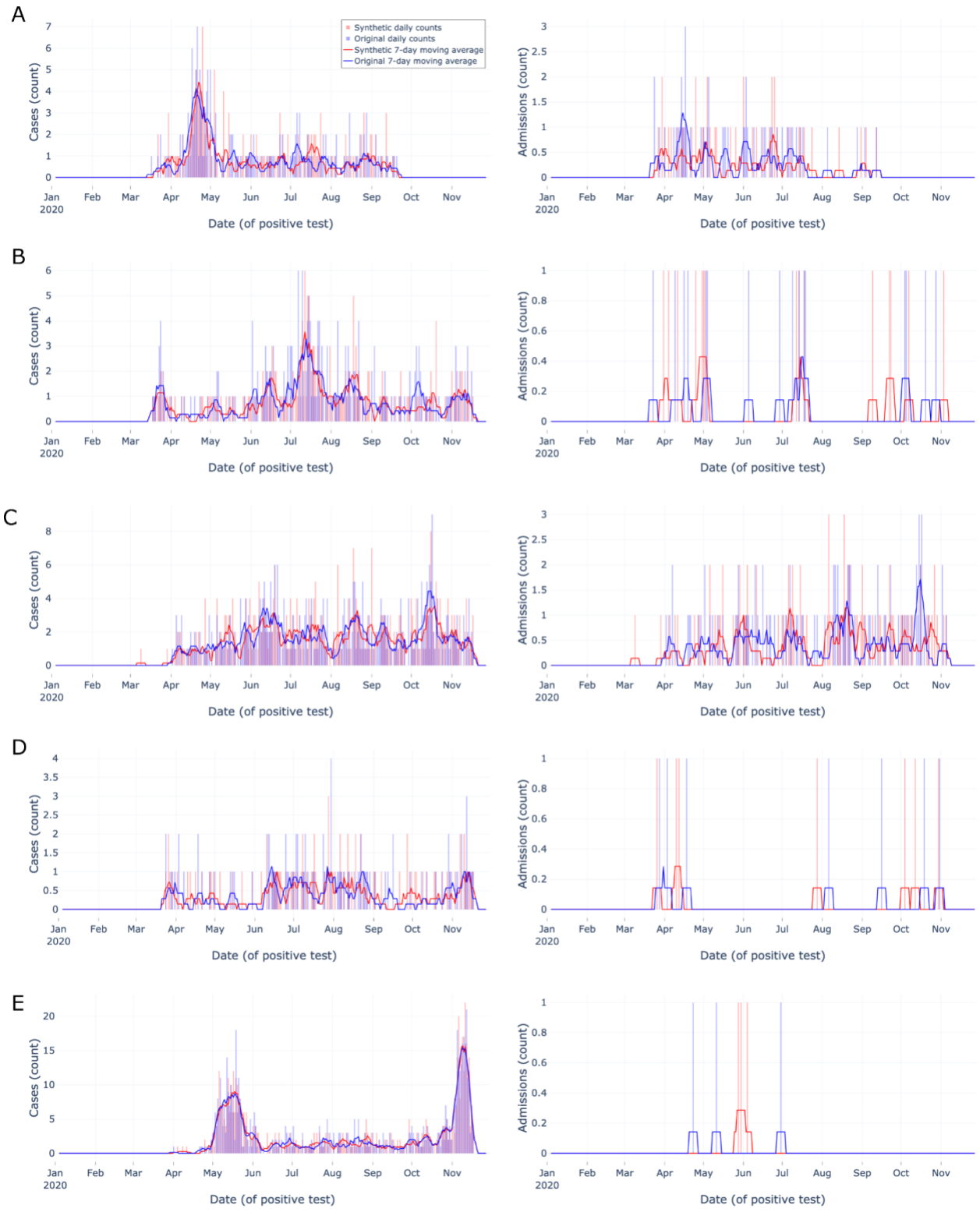

Figure 3.3. Zip code-level epidemic curves for random sample set #1 of the most tested (top 1%) zip codes.

Zip code-level epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line). Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). Each row (A-E) corresponds to a different randomly sampled zip code visualizing cases (left column) and admissions (right column). Synthetic data are more similar to original data when indicator density is higher, Overall, synthetic data closely match overall trends and closely match start and end dates.

Figure 3.4. Zip code-level epidemic curves for random sample set #2 of the most tested (top 1%) zip codes.

Zip code-level epidemic curves with counts (vertical bars) and 7-day moving averages (smoothed line). Color encodings include original data (light blue) and synthetic data (light red), with their overlap (purple). Each row (F-J) corresponds to a different randomly sampled zip code visualizing cases (left column) and admissions (right column). Synthetic data are more similar to original data when indicator density is higher, Overall, synthetic data closely match overall trends and closely match start and end dates.

3.4.1 *Distribution of tests by zip code and of censored zip codes:*

The 90th, 97.5th, and 99th percentiles for total tests by zip code in the original data were 125, 784, and 1,636 tests, respectively (see Figure 3.5A). Thus, a small minority of zip codes account for the vast majority of total tests. There were 15,108 (88.7%) unique zip codes in the original data with <100 total tests and 11,039 (64.7%) with <10 tests. Above this threshold (n≥10 tests), the synthetic data mimic the original data distribution closely (see Figure 3.5B). There were 17,041 unique zip codes and 5,819 unique uncensored zip codes in the original and synthetic data, respectively. The vast majority of censored zip codes are those that had <10 total tests in the original data (mean=2.9±2.4; median=2, IQR=3; max=16) as seen in Figure 3.7 of the supplement.

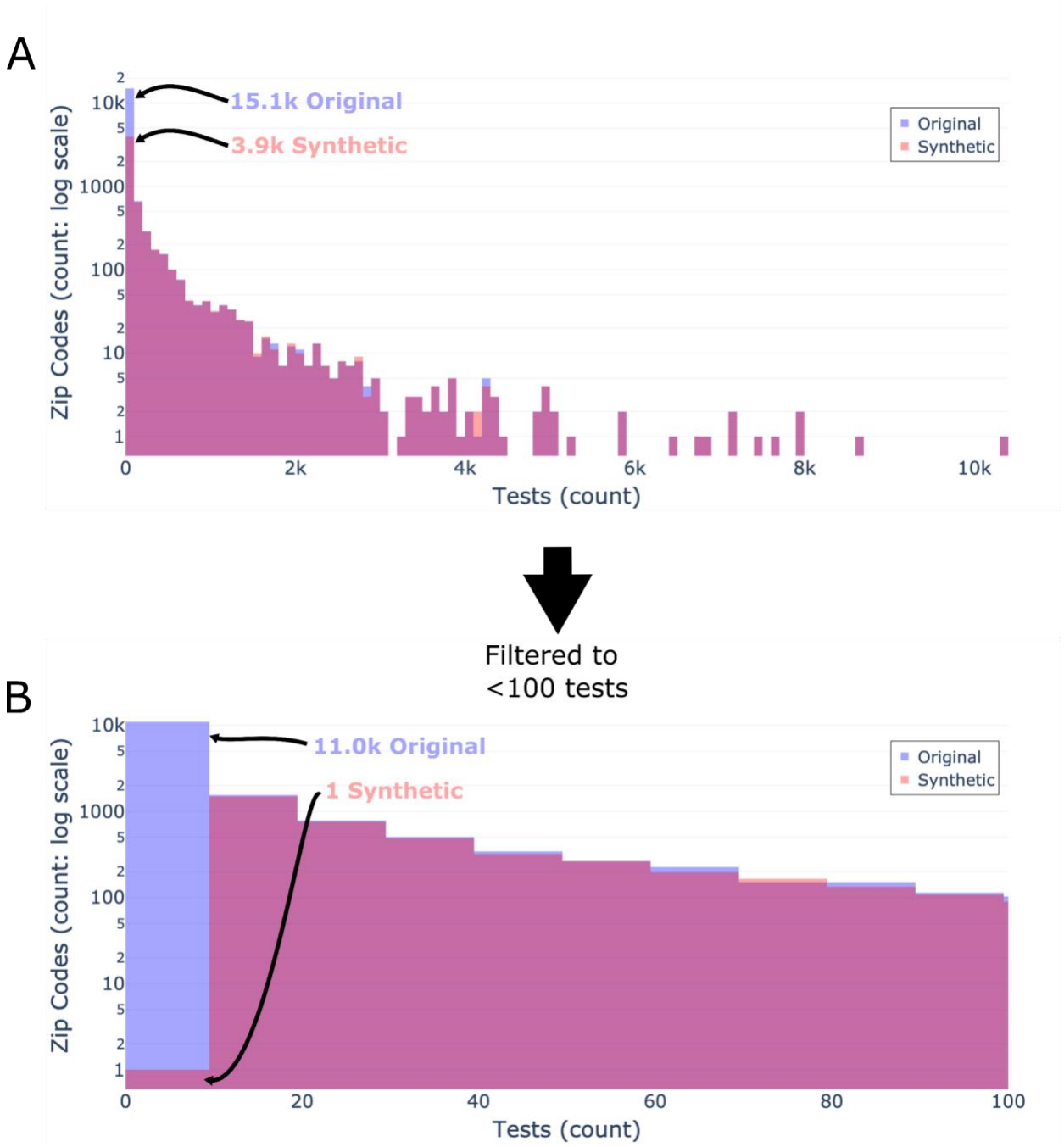

Figure 3.5. Distributions of total tests by zip code.

Original data (light blue) and synthetic data (light red), and their overlap (purple). A) All data binned by 100. B) Filtered data with a bin size of 10 to only show the distribution of tests by zip code in zip codes with <100 tests. Both y-axes use a log scale. As seen in panel A, the vast majority of tests are conducted in a minority of zip codes. As seen in panels A & B, the distribution of the synthetic data closely matches the original data at >10 tests per zip code.

3.4.2 *Monthly zip code pairwise synthetic error*

The absolute value of pairwise synthetic error stratified by month and zip code increased as the original data value of counts increased (see Figure 3.6; supplement Table 3.9). Thus, as sample size of data increased, so did the absolute synthetic error and vice versa. The synthetic error for tests ranged from an IQR=2 when the original value of tests was between 0 to 19 to IQR=9 when the original value of tests was between 250 and 1,705. All synthetic error for zip codes with an original bin value of zero count was positive. All other bins' synthetic error across key indicators was skewed negative, indicating that the synthetic data had lower counts than the original data.

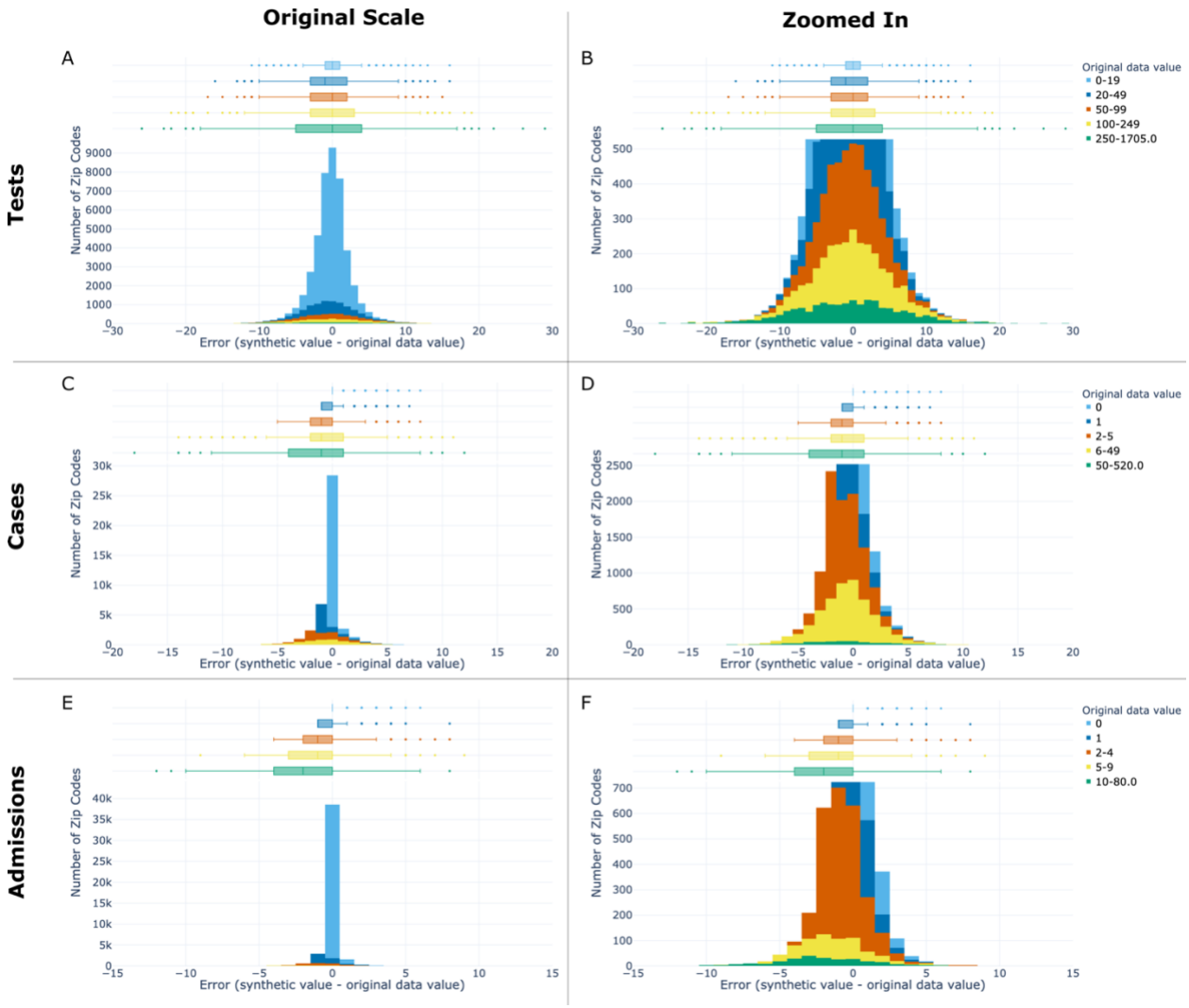

Figure 3.6. Synthetic error stratified by zip code and month.

Synthetic error distributions per zip code stratified by month for tests (top row), cases (middle row), and admissions (bottom row) shown both at original scale (left column) and zoomed in to the peak of each row's middle bin (legend showing bin ranges and color encodings seen on the far right of each row). Original data value denotes the monthly count in the original data for the key indicator of interest. Box plots of synthetic error are shown in the top 30% of each sub-plot (A-F), with a histogram of synthetic error shown in the bottom 70%. Within each sub-plot, the box plot and histogram have a shared x-axis corresponding to synthetic error and shared bins corresponding to the original data value. The y-axis shows the number of zip codes stratified by

month (e.g. zip code month pairs). Boxes in the box plots span from Q1 to Q3, with median marked inside the box. Fences span +/-1.5 times the IQR. Error increased as the size (count) of the original data increased, which allows users to estimate the level of error in their data of interest. The synthetic data systematically underestimate the monthly count of key indicators in zip codes with the most tests, cases, and deaths, and overestimate them in zip codes with the least.

3.5 DISCUSSION

Overall, analyses on the population-level and of densely-tested zip codes (which contained most of the data) were similar between original and synthetically-derived data sets. Analyses of sparsely-tested populations with smaller sample sizes were notably less similar and had more data suppression, which is in agreement with prior work.[19,32] Synthetic data most closely matched the original data on aggregate data tasks such as aggregate epidemic curves (Figure 3.2) and broad summary statistics (Table 3.6). At the aggregate level, only one metric (percent positive, 7-day average) across all indicators showed a significant difference between synthetic and original data aggregate epidemic curves (Table 3.7). Scarcity of data - as data collection used in this manuscript tapered off in November - is likely a contributing factor to the difference.

The summary statistics shown of both data sets' populations in Table 3.6 were similar. Major exceptions were the number of unique zip codes due to censoring in the synthetic data and attenuation in the synthetic data of a single extreme outlier (~-44,000 day LOS) caused by a data quality issue in the original data. Other erroneous negative LOS values persisted within the synthetic data, yet the bulk of the erroneous values remaining were a LOS=-1 which has been reported as a data quality issue attributed to daylight savings.[40,41] Thus, we show that

synthetic data can reduce the impact of data quality issues by removing or attenuating erroneous outliers in order to protect the privacy of rare, and thus identifiable, data.

At the zip code and month level, the synthetic data error performed well on an absolute level; the error increased as the size of the original data increased (Figure 3.6 & Supplementary Table 3.9). Therefore, the amount of synthetic error is predictable which gives users the ability to estimate the level of error in their data of interest. Additionally, the synthetic error relative to the original data value is likely small enough for most uses of synthetic data. For example, a zip code in the synthetic data with a monthly positive count of 6-49 is off from the original data by an average of -0.59±2.63. The overrepresentation of negative tests in the original data by 8.5-fold (Table 3.6) appears to bias synthetic error. Since it is impossible to have less than zero count, the synthetic data cannot add privacy-producing noise in the negative direction for zip code monthly counts equal to zero. Consequently, the synthetic data systematically underestimate the monthly count of key indicators in zip codes with the most tests, cases, and deaths, and overestimate them in zip codes with the least. Our results relate to Flaxman et al., 2020 which observed a similar effect resulting from a non-negativity constraint in the US Census' TopDown differential privacy algorithm.[12] The magnitude of the synthetic error skewing negative in a smaller concentration of zip codes increased as a key indicator became less frequent, which is fundamentally a signal problem in low-density data sets and is not specific to synthetic data generation.

The top 1% most tested zip codes' epidemic curves provide users with 10 qualitative examples of densely tested zip codes. Overall, the synthetic data closely matched the start and end dates of the original data and followed the overall trend of the original data over time (e.g. Figure 3.4A matched spike in late April). The ten examples show users the 99th percentile best-

case scenario of key indicator original data availability and synthetic data performance at the zip code level, yet the size and testing density of N3C data will likely continue to increase.

Our findings show the importance of understanding the characteristics and limitations of the original data since we found these biases affected synthetic data utility. Data biases resulting in poorer performance of software tools, clinical guidelines and other applications for groups underrepresented in source data has been previously reported for separate tasks.[12,42–45] Foraker et al. (2021) found that censored zip codes had greater missingness of SDOH values in the original data than uncensored zip codes. In our study, we found the bulk of patients in the N3C data live in a small minority of zip codes (Figure 3.5), likely those most adjacent to institutions contributing data. These zip codes are therefore more likely to be urban and less likely to have their zip code censored (Table 3.8). As a consequence, rural zip codes, which are already underrepresented in the original data, become even less available to directly analyze. Additionally, patients with censored zip codes were older, potentially due to older patients traveling from sparsely tested regions to receive care offered at distant academic medical centers which participate in N3C.

While our results demonstrate the utility of using synthetic data for a broad range of geospatial analyses, a caveat to synthetic data use is its utility to analyze rural N3C populations since nearly all zip codes with <10 tests were censored and much more likely to be rural within the original data. Suppression of non-zero counts <10 is a common convention within state and federal guidelines to avoid inadvertent disclosure of protected health information for publicly released data.[46–48] Analyses such as choropleth maps at the zip code level including sparsely tested regions would benefit from using the LDS to obtain access to all zip codes without suppression, or by generating and using a different MDClone synthetic dataset that reports

geospatial data at a lower level of granularity (e.g. 3-digit zip codes). Our results may inform future N3C discussions about data set balancing ranging from 1) creation of artificially balanced hybrid data sets to improve statistical models' performance on underrepresented data[42,49], 2) source partners sending a random sample of negative tests alongside all positive tests, or 3) expansion of data ingestion from rural regions.

Whether these synthetic data are "good enough" hinges on a fitness for use determination to be made by each user. The authors believe the data will be useful enough for a wide variety of use cases. Educational software engineering projects or pandemic preparedness tool development could be especially well-served by these data. A major limitation of the data, however, is that they are output in a different data model than the OMOP CDM.[36] Thus, tools built on the synthetic data would not be transferable to run on the LDS without modification. Other users may find the synthetic data well suited to rapid, iterative hypothesis generation/testing without the delays of acquiring the relatively more restricted LDS.[3]

3.6 LIMITATIONS AND FUTURE WORK

To date, no privacy analysis has been published on these synthetic data to provide context for its utility in relation to its privacy. The data used in this manuscript do not reflect the current size nor state of the N3C LDS. Other statistical techniques such as equivalence testing, bhattacharyya distance[50,51], or adversarial challenges[28] could be used in the future to compare similarity between epidemic curves. The Wilcoxon signed-rank and paired t-tests assume the null hypothesis that the original and synthetic datasets are equivalent. Equivalence testing, which flips the null hypothesis, may be better suited. Equivalence testing was not used in this manuscript due to the challenge of selecting an equivalence bound without knowing what threshold(s) data end-users would find most applicable. Future work conducting equivalence

testing specific to well-defined, high-impact use cases may be merited. However, the work required to do so in an ad hoc manner may suggest the LDS is a better alternative in those cases.

3.7 CONCLUSION

Overall, the synthetic data are promising for a wide range of use cases including: population level summary statistics, epidemic curves for the data in aggregate and for the most densely tested zip codes, and analyses necessitating monthly counts of key indicators for the top third of zip codes by number of tests. However, analyses requiring unsuppressed zip code analyses on populations with <10 tests may be better served by the LDS. Biases found in the original data - namely an underrepresentation of positive tests and tests in rural zip codes - were reflected in the synthetic data. Therefore, it is important to understand the limitations and biases of the original data in addition to the synthetic data impacted downstream from it. We expect the user base of N3C synthetic data to be heterogeneous and the use cases of the data to be broad, resulting in a wide range of fitness for use definitions. To date, there is no published evaluation that quantifies the privacy afforded by this synthetic dataset specifically - nor of the MDClone system itself broadly - to contextualize this synthetic dataset's utility in relation to a privacy-utility tradeoff; such evaluations are beyond the scope of this work. Future privacy evaluations of MDClone will not necessarily reflect the privacy of the synthetic data analyzed in this study unless the same dataset and/or the same MDClone system version and parameters are evaluated. Our evaluation of the N3C synthetic data utility provides users the ability to assess whether the synthetic data are fit for use through its combination of general-purpose data utility assessments and visualized replications of analyses of common interest.

3.8 ACKNOWLEDGEMENTS

3.8.1 *Funding*

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave and supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource[2].

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources.

We gratefully acknowledge contributions from the following N3C core teams (leads noted with *)

- Principal Investigators: Melissa A. Haendel*, Christopher G. Chute*, Kenneth R. Gersing, Anita Walden
- Workstream, subgroup and administrative leaders: Melissa A. Haendel*, Tellen D. Bennett, Christopher G. Chute, David A. Eichmann, Justin Guinney, Warren A. Kibbe, Hongfang Liu, Philip R.O. Payne, Emily R. Pfaff, Peter N. Robinson, Joel H. Saltz, Heidi Spratt, Justin Starren, Christine Suver, Adam B. Wilcox, Andrew E. Williams, Chunlei Wu
- Key liaisons at data partner sites

- Regulatory staff at data partner sites
- Individuals at the sites who are responsible for creating the datasets and submitting data to N3C
- Data Ingest and Harmonization Team: Christopher G. Chute*, Emily R. Pfaff*, Davera Gabriel, Stephanie S. Hong, Kristin Kostka, Harold P. Lehmann, Richard A. Moffitt, Michele Morris, Matvey B. Palchuk, Xiaohan Tanner Zhang, Richard L. Zhu
- Phenotype Team (Individuals who create the scripts that the sites use to submit their data, based on the COVID and Long COVID definitions): Emily R. Pfaff*, Benjamin Amor, Mark M. Bissell, Marshall Clark, Andrew T. Girvin, Stephanie S. Hong, Kristin Kostka, Adam M. Lee, Robert T. Miller, Michele Morris, Matvey B. Palchuk, Kellie M. Walters
- Project Management and Operations Team: Anita Walden*, Yooree Chae, Connor Cook, Alexandra Dest, Racquel R. Dietz, Thomas Dillon, Patricia A. Francis, Rafael Fuentes, Alexis Graves, Julie A. McMurry, Andrew J. Neumann, Shawn T. O'Neil, Usman Sheikh, Andréa M. Volz, Elizabeth Zampino
- Partners from NIH and other federal agencies: Christopher P. Austin*, Kenneth R. Gersing*, Samuel Bozzette, Mariam Deacy, Nicole Garbarini, Michael G. Kurilla, Sam G. Michael, Joni L. Rutter, Meredith Temple-O'Connor
- Analytics Team (Individuals who build the Enclave infrastructure, help create codesets, variables, and help Domain Teams and project teams with their datasets): Benjamin Amor*, Mark M. Bissell, Katie Rebecca Bradwell, Andrew T. Girvin, Amin Manna, Nabeel Qureshi

- Publication Committee Management Team: Mary Morrison Saltz*, Christine Suver*, Christopher G. Chute, Melissa A. Haendel, Julie A. McMurry, Andréa M. Volz, Anita Walden
- Publication Committee Review Team: Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farrukh M Koraishy, Federico Mariona, Saidulu Mattapally, Amit Saha, Satyanarayana Vedula
- Synthetic Data Domain Team including Yujuan Fu, Nisha Mathews, Ofer Mendelevitch

Data was provided from the following institutions: Stony Brook University — U24TR002306 • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • University of Mississippi Medical Center — U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center — U54GM115458: Great Plains IDeA-Clinical & Translational Research • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • University of Massachusetts

Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • University of Washington — UL1TR002319: Institute of Translational Health Sciences • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • The

University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • Virginia Commonwealth University — UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science Institute • University of Virginia — UL1TR003015: iTHRIVL Integrated Translational health Research Institute of Virginia • Carilion Clinic — UL1TR003015: iTHRIVL Integrated Translational health Research Institute of Virginia • University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • University of Arkansas for Medical Sciences — UL1TR003107: UAMS Translational Research Institute • Nemours — U54GM104941: Delaware CTR ACCEL Program • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Tulane University — UL1TR003096: Center for Clinical and Translational Science • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine

(ITM) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks

Additional data partners who have signed DTA and data release pending: The Rockefeller University — UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • University of Texas Health Science Center at San Antonio — UL1TR002645: Institute for Integration of Medicine and Science • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • NorthShore University HealthSystem — UL1TR002389: The Institute for Translational Medicine (ITM) • Yale New Haven Hospital — UL1TR001863: Yale Center for Clinical Investigation • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Cincinnati Children's Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Boston

University Medical Campus — UL1TR001430: Boston University Clinical and Translational Science Institute • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • Aurora Health Care — UL1TR002373: Wisconsin Network For Health Research • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • New York University — UL1TR001445: Langone Health's Clinical and Translational Science Institute • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • Ochsner Medical Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • HonorHealth — None (Voluntary) • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute

Authorship was determined using ICMJE recommendations. The project described was supported by the National Institute of General Medical Sciences, 5U54GM104942-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Any opinions expressed in this document are those of the authors and do not necessarily reflect the views of NCATS, individual N3C team members, or affiliated organizations and institutions.

3.8.2 *Contributor Statements*

3.8.2.1 Masthead Authors

Authors ABW, JAT, NZ, RF, contributed to study conception and design. Author NZ contributed to the generation of the data. Author JAT conducted the experiment and data analysis. Author JAT wrote the manuscript with input from all authors. Authors ABW, RF led the N3C Synthetic Data Validation Task Team with support from author PP who led the broader N3C Synthetic Data Workstream.

3.8.2.2 Other Consortial Authors*

Christopher G. Chute[1,2,3,4,5,6,7,8,9,10], Jon D. Morrow[1,12,13,2,7,9], Melissa A. Haendel[14,6,10,11]

[1]clinical data model expertise, [2]data curation, [3]data integration, [4]data quality assurance, [5]funding acquisition, [6]governance, [7]critical revision of the manuscript for important intellectual content, [8]N3C Phenotype definition, [9]project evaluation, [10]project management, [11]regulatory oversight / admin, [12]clinical subject matter expertise, [13]data analysis, [14]funding acquisition

**Consortial authorship and corresponding contributions were self-reported as part of the N3C authorship committee review process.*

3.8.3 *Competing Interest Statement*

All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/downloads/coi_disclosure.docx and declare: Authors JAT, ABW, RF, and PP received financial support from the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number U24TR002306 disbursed to their affiliated institutions for the submitted work; author NZ is an employee of MDClone; this manuscript underwent National Covid Cohort Collaborative (N3C) publication review described at https://covid.cd2h.org/publication-review; the institution RF and PP are affiliated with (Washington University in St. Louis) is a customer of MDClone; all authors declare no other relationships or activities that could appear to have influenced the submitted work.

3.8.4 *Human Subjects Protections*

This study was approved by the Washington University and University of Washington Internal Review Boards.

3.9 SUPPLEMENT

Table 3.9. Zip code month pairs' synthetic error central tendencies and counts stratified by indicator and bin size.

Indicator	**Number of zip codes stratified by month**	**Bin value original count**	**Synthetic Error mean (stdev)**	**Synthetic Error median (IQR)**
Tests	33328	0-19	-0.14 (±1.9)	0 (2)
Tests	5283	20-49	-0.54 (±3.31)	-1 (5)
Tests	2697	50-99	-0.4 (±4.11)	0 (5)
Tests	2230	100-249	-0.28 (±5.17)	0 (6)
Tests	1102	250-1705	-0.59 (±7.29)	0 (9)
Positives	26707	0	0.07 (±0.37)	0 (0)
Positives	6499	1	-0.55 (±0.92)	-1 (1)
Positives	6264	2-5	-0.78 (±1.76)	-1 (2)
Positives	4715	6-49	-0.59 (±2.63)	-1 (3)
Positives	455	50-520	-1.13 (±4.22)	-1 (5)
Admissions	37963	0	0.04 (±0.25)	0 (0)
Admissions	3837	1	-0.43 (±0.82)	-1 (1)
Admissions	2078	2-4	-0.66 (±1.42)	-1 (2)
Admissions	499	5-9	-1.37 (±2.29)	-1 (3)
Admissions	263	10-80	-2.16 (±3.33)	-2 (4)

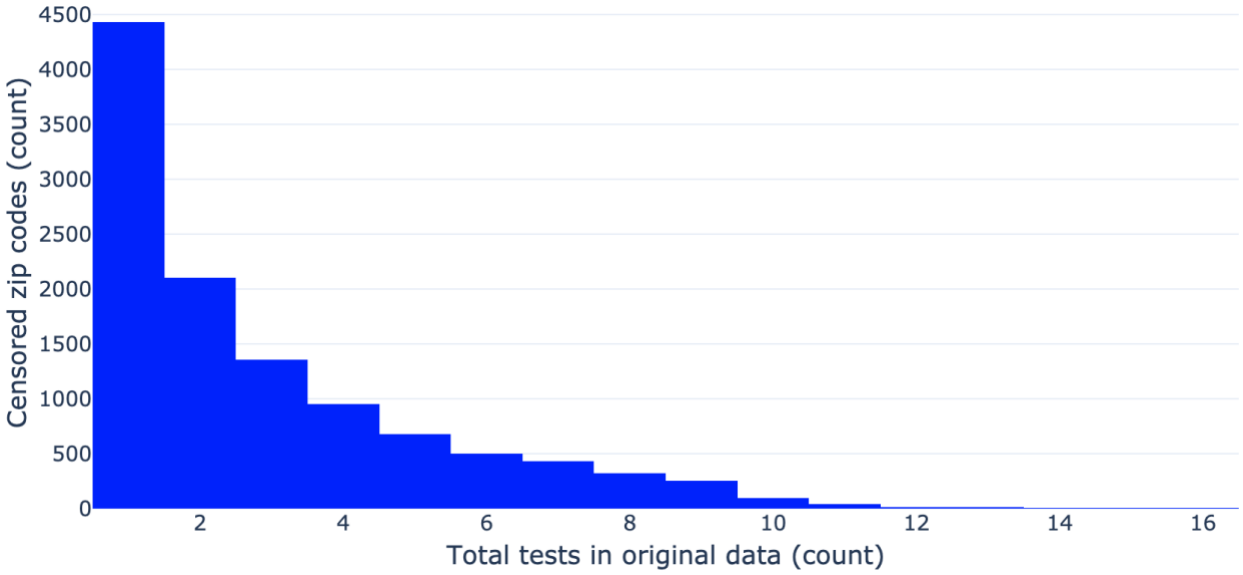

Figure 3.7. Distribution of total tests per zip code in original data which were censored within the synthetic data.

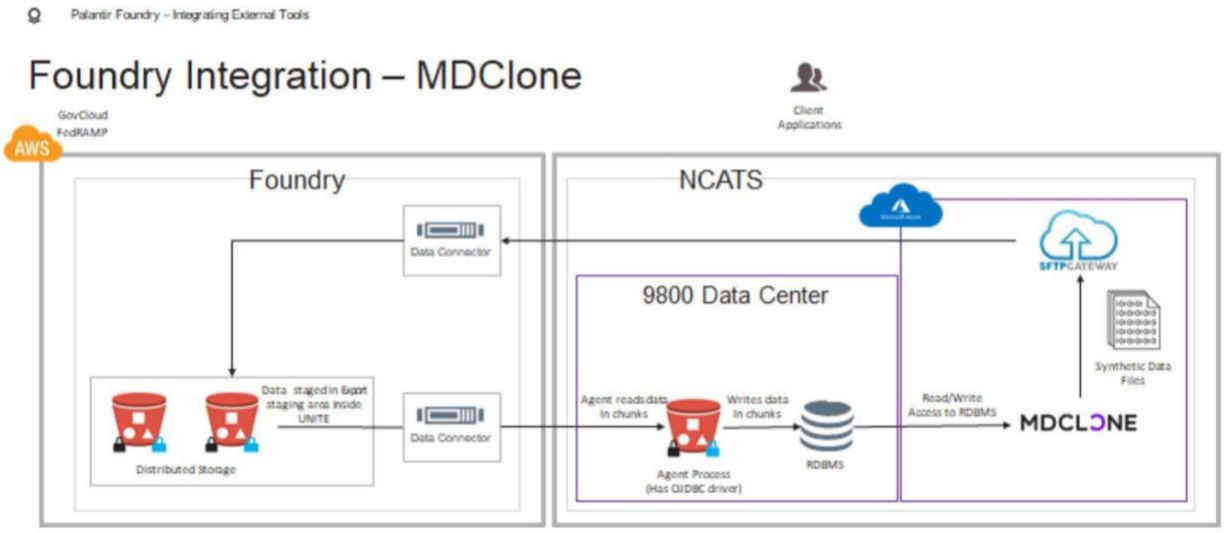

Figure 3.8. MDClone data synthesis workflow

3.10 REFERENCES FOR CHAPTER 3

1 Azzopardi-Muscat N, Kluge HHP, Asma S, *et al.* A call to strengthen data in response to COVID-19 and beyond. *J Am Med Inform Assoc* 2021;**28**:638–9. doi:10.1093/jamia/ocaa308

2 Subbian V, Solomonides A, Clarkson M, *et al.* Ethics and informatics in the age of COVID-19: challenges and recommendations for public health organization and public policy. *J Am Med Inform Assoc JAMIA* 2021;**28**:184–9. doi:10.1093/jamia/ocaa188

3 Melissa H, Christopher C, Kenneth G. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. *J Am Med Inform Assoc JAMIA* Published Online First: 17 August 2020. doi:10.1093/jamia/ocaa196

4 The National COVID Cohort Collaborative: Clinical Characterization and Early Severity Prediction | medRxiv. https://www.medrxiv.org/content/10.1101/2021.01.12.21249511v3 (accessed 1 Mar 2021).

5 HIPAA Privacy Rule and Its Impacts on Research. https://privacyruleandresearch.nih.gov/pr_08.asp (accessed 17 Mar 2021).

6 45 CFR 164.514 - Other requirements relating to uses and disclosures of protected health information. - Content Details - CFR-2011-title45-vol1-sec164-514. https://www.govinfo.gov/app/details/CFR-2011-title45-vol1/CFR-2011-title45-vol1-part164 (accessed 17 Mar 2021).

7 Raab GM, Nowok B, Dibben C. Guidelines for Producing Useful Synthetic Data. *ArXiv171204078 Stat* Published Online First: 11 December 2017.http://arxiv.org/abs/1712.04078 (accessed 17 Mar 2021).

8 Snoke J, Raab GM, Nowok B, *et al.* General and specific utility measures for synthetic data. *J R Stat Soc Ser A* 2018;**181**:663–88.

9 Mukherjee S, Xu Y, Trivedi A, *et al.* privGAN: Protecting GANs from membership inference attacks at low cost. *ArXiv200100071 Cs Stat* Published Online First: 13 December 2020.http://arxiv.org/abs/2001.00071 (accessed 17 Mar 2021).

10 Beaulieu-Jones Brett K., Wu Zhiwei Steven, Williams Chris, *et al.* Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circ Cardiovasc Qual Outcomes* 2019;**12**:e005122. doi:10.1161/CIRCOUTCOMES.118.005122

11 Foraker R, Mann DL, Payne PRO. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* 2018;**3**:716–8. doi:10.1016/j.jacbts.2018.08.007

12 Petti S, Flaxman A. Differential privacy in the 2020 US census: what will it do? Quantifying

the accuracy/privacy tradeoff. *Gates Open Res* 2020;**3**:1722. doi:10.12688/gatesopenres.13089.2

13 Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;**25**:37–43. doi:10.1038/s41591-018-0272-7

14 Wu L, He H, Zaïane OR. Utility of privacy preservation for health data publishing. In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013. 510–1. doi:10.1109/CBMS.2013.6627853

15 Muniz-Terrera G, Mendelevitch O, Barnes R, *et al.* Virtual Cohorts and Synthetic Data in Dementia: An Illustration of Their Potential to Advance Research. *Front Artif Intell* 2021;**4**. doi:10.3389/frai.2021.613956

16 CDC. Transitioning from CDC's Indicators for Dynamic School Decision-Making (released September 15, 2020) to CDC's Operational Strategy for K-12 Schools through Phased Mitigation (released February 12, 2021) to Reduce COVID-19. Cent. Dis. Control Prev. 2020.https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/indicators.html (accessed 21 Mar 2021).

17 CDC. Operational Strategy for K-12 Schools through Phased Mitigation. Cent. Dis. Control Prev. 2020.https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/operation-strategy.html (accessed 16 Feb 2021).

18 State-By-State Summary of Public Health Criteria in Reopening Plans. Natl. Gov. Assoc. https://www.nga.org/coronavirus-reopening-plans/ (accessed 2 Mar 2021).

19 Benaim AR, 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. doi:10.2196/16492

20 Zhang Z, Yan C, Mesa DA, *et al.* Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc JAMIA* 2019;**27**:99–108. doi:10.1093/jamia/ocz161

21 Teixeira da Silva JA, Tsigaris P, Erfanmanesh M. Publishing volumes in major databases related to Covid-19. *Scientometrics* 2021;**126**:831–42. doi:10.1007/s11192-020-03675-3

22 Guerrini CJ, Majumder MA, Lewellyn MJ, *et al.* Citizen science, public policy. *Science* 2018;**361**:134–6. doi:10.1126/science.aar8379

23 Katapally TR. A Global Digital Citizen Science Policy to Tackle Pandemics Like COVID-19. *J Med Internet Res* 2020;**22**:e19357. doi:10.2196/19357

24 Roche J, Bell L, Galvão C, *et al.* Citizen Science, Education, and Learning: Challenges and Opportunities. *Front Sociol* 2020;**5**. doi:10.3389/fsoc.2020.613814

25 Juran JM, Godfrey AB, editors. *Juran's quality handbook*. 5th ed. New York: : McGraw Hill 1999.

26 Chen J, Chun D, Patel M, *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 2019;**19**:44. doi:10.1186/s12911-019-0793-0

27 Foraker RE, Yu SC, Gupta A, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* Published Online First: 14 December 2020. doi:10.1093/jamiaopen/ooaa060

28 El Emam K, Mosquera L, Jonker E, *et al.* Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 2021;**4**. doi:10.1093/jamiaopen/ooab012

29 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Comput Intell*;**n/a**. doi:https://doi.org/10.1111/coin.12427

30 Hittmeir M, Ekelhart A, Mayer R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Canterbury, CA, United Kingdom: : Association for Computing Machinery 2019. 1–6. doi:10.1145/3339252.3339281

31 Emam KE. Seven Ways to Evaluate the Utility of Synthetic Data. *IEEE Secur Priv* 2020;**18**:56–9. doi:10.1109/MSEC.2020.2992821

32 Analyses of Original and Computationally-Derived Electronic Health Record Data: The National COVID Cohort Collaborative. JMIR Prepr. https://preprints.jmir.org/preprint/30697 (accessed 7 Jun 2021).

33 CDC. COVID Data Tracker. Cent. Dis. Control Prev. 2020.https://covid.cdc.gov/covid-data-tracker (accessed 23 Mar 2021).

34 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4. doi:10.1016/S1473-3099(20)30120-1

35 Roser M, Ritchie H, Ortiz-Ospina E, *et al.* Coronavirus Pandemic (COVID-19). *Our World Data* Published Online First: 5 March 2020.https://ourworldindata.org/coronavirus (accessed 23 Mar 2021).

36 Observational Health Data Sciences and Informatics. OMOP CDM v5.3.1. https://ohdsi.github.io/CommonDataModel/cdm531.html (accessed 26 Mar 2021).

37 Waskom M, team the seaborn development. *mwaskom/seaborn*. Zenodo 2020. doi:10.5281/zenodo.592845

38 Choosing color palettes — seaborn 0.11.1 documentation. https://seaborn.pydata.org/tutorial/color_palettes.html (accessed 24 Mar 2021).

39 Jenny B, Kelso NV. Color oracle. *Color Oracle Des Color Impair* 2011.

40 Ehlers A, Dyson RL, Hodgson CK, *et al.* Impact of Daylight Saving Time on the Clinical Laboratory. *Acad Pathol* 2018;**5**. doi:10.1177/2374289518784222

41 Thomas JA, Wilcox AB, Joo EJ. Readmissions: Data Quality and Prediction. *2018 Natl Libr Med Train Conf - Poster* Published Online First: 2018.https://osf.io/vk65x/ (accessed 15 Apr 2021).

42 Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol* 2018;**154**:1247. doi:10.1001/jamadermatol.2018.2348

43 Gijsberts CM, Groenewegen KA, Hoefer IE, *et al.* Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events. *PloS One* 2015;**10**:e0132321. doi:10.1371/journal.pone.0132321

44 Grother P, Ngan M, Hanaoka K. Face recognition vendor test part 3:: demographic effects. Gaithersburg, MD: : National Institute of Standards and Technology 2019. doi:10.6028/NIST.IR.8280

45 Kessler MD, Yerges-Armstrong L, Taub MA, *et al.* Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. *Nat Commun* 2016;**7**:12521. doi:10.1038/ncomms12521

46 Washington State Department of Health. Guidelines for working with small numbers. 2001.

47 Klein R, Proctor S, Boudreault M, *et al.* Healthy people 2010 criteria for data suppression. Centers for Disease Control Statistical Notes Number 24. 2002.

48 McCallister E, Grance T, Scarfone K. Guide to protecting the confidentiality of personally identifiable information (pii): recommendations of the National Institute of Standards and Technology. Special Publication 800-122, National Institute of Standards and Technology. 2010.

49 Ghorbani A, Natarajan V, Coz D, *et al.* DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. In: *Machine Learning for Health Workshop*. PMLR 2020. 155–70.http://proceedings.mlr.press/v116/ghorbani20a.html (accessed 29 Mar 2021).

50 Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*. 2000. 142–9 vol.2. doi:10.1109/CVPR.2000.854761

51 Kaloskampis I, Joshi C, Cheung C, *et al.* Synthetic data in the civil service. *Significance*

2020;**17**:18–23. doi:https://doi.org/10.1111/1740-9713.01466

Chapter 4. ASSESSING SINGLE SIGN-ON AUTHENTICATION BEHAVIORS TO INFORM CUSTOMIZED SOLUTIONS USING REAL AND SYNTHETIC LOG DATA

4.1 ABSTRACT

Objective: To investigate the potential for and impact of customized shared clinical workstation access policies by user role and location and 2) assess the utility of single sign-on (SSO) log data, and its synthetic derivative for this task.

Materials and Methods: SSO log data from January-March 2020 were analyzed. We assessed the impact of two potential policy changes in the second half of Q1. First, adjusting the duration (default=4 hours) SSO can rely on a password before 'challenging' a user to provide it again. Second, requiring a challenge for every ‘new’ workstation accessed with and without users’ prior history. Policy burden was stratified by user role (e.g. physician) and workstation location (e.g. emergency department). Synthetic data were generated from the SSO data, then used for analysis as well.

Results: Relative to the default, the percent of logins requiring a challenge ranged from an average increase of 18.13 and decrease of -6.67 percentage points for 1-hour and 12-hour challenge periods, respectively. Requiring a challenge for each new workstation increased challenges by an average of 4.19 and 8.55 percentage points with and without prior user history. For both policy changes, burden varied greatly between roles and locations. Synthetic data performed well on conformance to the real data but its results were not similar, or its features were at least muted, compared to the real data.

Discussion: The burden or relief from policy changes investigated, if implemented, would be unevenly distributed. Incorporating past user behavior can greatly attenuate the burden of policies that rely on user-specific behaviors.

Conclusion: Shared clinical workstation login behavior varies between individuals, roles and locations. This heterogeneity should be factored into policy discussions and presents an opportunity to afford users customized authentication solutions. Synthetic data that were generated from a model trained for a limited number of epochs on a subset of the real performed well on conformance but poorly on replicating the results of analyses conducted on the real data. Overall, SSO log data and the roles data they were merged with showed that granular authentication behavior patterns and the effects of simulated policy changes can be gleaned from these data.

4.2 INTRODUCTION

4.2.1 *Background and Significance*

The association between burdensome health information technology (HIT) and negative impacts on clinicians is well documented[1–3] and has spurred efforts to reduce this burden.[4–7] In 2020, the Office of the National Coordinator (ONC) identified optimization of system log-on for end users as a key strategy for reducing clinician burden, specifically citing token based authentication (e.g. a proximity card) as a potential solution to do so while balancing security and privacy risks.[6] Proximity cards are a key part of many hospital single sign-on (SSO) implementations. SSO allows users to log-on or off by solely tapping their card to the clinical workstation as long as they supplied another form of authentication (e.g. a password) recently in addition to tapping the card. The length of time a user may use the card to log-on and off by "tapping in and tapping out (TITO)" before being challenged to re-authenticate is called the

'challenge period.' The impact of SSO implementation on clinician satisfaction, time savings, password sharing, and institutional financial savings has been studied - often through analysis of SSO log data - with promising results.[8–11]

However, analysis of SSO data brings with it concerns for balancing the use of these data in alignment with worker and employer interests. A 2018 international survey of 1,400 C-level executives and 10,000 workers reported that "62% of businesses are using new technologies and sources of workforce data today but only 30% of these leaders are confident that they are using new sources of workforce data in a highly responsible way."[12] The same survey showed that workers have concerns about use of their data yet are willing to share their data in exchange for benefits, listing a customized work experience as the number one desired benefit. Those survey responses suggest that the customization of authentication experience could be a good target for use of worker data. One such potential customization is the challenge period, which is set to four hours for all users at UW Medicine. From the literature it can be gleaned that the duration of the single sign-on challenge period in healthcare settings is variable across sites, with sites/health systems setting their challenge period to four[8,13], eight[14], twelve[10,11], or a not explicitly disclosed length of hours.[9]

While retrospective, observational SSO log data might be useful to inform granular changes in SSO implementation policies (e.g. challenge period duration) and the potential for customized solutions, their utility to do so is unknown and may suffer from the same limitations of EHR log data utility. EHR log data have varying levels of comprehensiveness and granularity across EHR systems which can pose challenges to their use.[15] To the authors' knowledge, no study has been done to evaluate the utility of SSO log data for this purpose nor of the potential for customized SSO implementations and the differential impact of these customizations.

4.2.2 *Objective*

In this paper, we sought to use Imprivata Onesign SSO log data from the UW Medicine Health system (Seattle, WA USA) - comprised of a Trauma Level 1 hospital, academic medical center, and outpatient clinics - to inform customized SSO authentication protocols and report on the utility of observational SSO log data to do so. In addition to characterizing SSO behaviors broadly, we considered two potential SSO implementation changes and their simulated impacts stratified by user role and location. The first potential change was variation of the challenge period from 1-12 hours in 1-hour increments. The second was requiring a challenge for each new workstation a user logs into with and without incorporating their prior workstation access history. Additionally, we piloted the creation and use of synthetic SSO log data to re-create portions of our analysis in an effort to protect the privacy of worker data.

4.3 METHODS

This study was approved by the University of Washington Institutional Review Board. All analyses were performed by a single author (JAT) using python (version 3.7.4).

4.3.1 *Data collection and cleaning*

We obtained Imprivata SSO log data extracted from the UW Medicine Health System (Seattle, WA, USA) which included: a Trauma Level I hospital (Harborview), an Academic Medical Center (UWMC), and outpatient neighborhood clinics. The dates of SSO data extracted spanned from the beginning of system testing (2015) until part of July 2020. The raw data were obtained in .csv file format. Each row represented an action in the SSO system with the following relevant columns: Datetime (down to the second), User ID, SSO Activity (e.g. Shared Workstation Login), Method (e.g. Proximity Card + Password), and Host (e.g. UWM-ED-**** which denotes

the site, then subsite location, then specific workstation). SSO data were cleaned to remove duplicate rows and combined into separate files by month. The total count of actions (rows) after cleaning was 65,072,552.

User data were extracted from and/or combined with multiple data sources. First, from UW Medicine's privileged user and management audit system which, for users with medical licenses, included 1) the first two characters (eg. 'MD', 'AP') of the user's Washington State medical license number denoting medical license type and 2) userID start date (when available). Medical license prefix information was scraped from the Washington State Department of Health Provider Credential Search Website Credential Type dropdown box.[16] We constructed a dictionary for the one-to-many mapping of medical license prefix to credential type, and selected the most recent license for each user when dates were available. Similar medical licenses (e.g. Physicians = Medical or Osteopathic doctors) were grouped together into higher-level roles or clusters. More information regarding these mappings including the way licenses were grouped can be seen in the supplement (Table 4.12). For the purposes of this study, users without medical licenses were excluded.

Location information was extracted from the host column of the SSO log data using the second part of the host's three hyphen separated strings (e.g. UWM-ED-**** = ED). Four categories were created: emergency department (ED), intensive care unit (ICU), other (all others besides ED and ICU), and all locations to denote all the combined locations. ED included all hosts with "ED" as the second part of the host string. ICU included all hosts with neonatal (NICU), trauma/surgical (TSICU), burn and pediatric (BPICU), Medical cardiac (MCICU), or standard (ICU) intensive care units as the second part of the host string.

4.3.2 *Characterization of SSO TITO activity and its rollout within the UW Medicine system*

Monthly SSO data were analyzed to calculate and visualize distributions of unique monthly active users and total logins. The distributions were put into context by overlaying the rollout start date of the system (July 2018) onto visualizations. Through our analysis of the rollout and overall characteristics of adoption, quarter 1 (Q1; January-March) 2020 was selected as the most stable time period for user-specific analyses. Q1 was selected due to 1) rollout being complete by this time, 2) its relatively stable unique monthly active user and TITO actions counts over this time period, and 3) the impact of COVID-19 confounding the data in Q2 and beyond.

4.3.3 *Summary of the data*

For the core analyses in this manuscript, only data of users with medical licenses were analyzed. A variety of summary statistics were calculated broken down by Q1 2020 overall and then by individual months within Q1. Summary statistics included the number of unique users stratified by role, the unique number of workstations accessed, total actions and actions stratified by site and workstation location as well as the number of Shared Workstation Logins stratified by whether the user used a proximity card and password, just the proximity card or just the password. For the remaining analyses, the data were filtered to the Q1 timeframe and users with medical licenses who worked (defined as having present data that calendar day) for seven or more days during the time period.

4.3.4 *User-specific behaviors*

The following metrics were calculated. First, the number of days worked which was defined as the number of days where data are present. Second, the number of unique workstations accessed. Third, the range from the first to last day worked was calculated as the date difference between

the first calendar date the user was present in the SSO log data to the last date. Fourth, the number of new workstations per workday was calculated by dividing unique workstations accessed by days worked.

Last, we analyzed the rate at which users reached the maximum number of workstations accessed by plotting the fraction of total workstations accessed compared to the fraction of days worked for each day worked by each user. The proportional relationship between the two was plotted in a heatmap with a Locally Weighted Scatterplot Smoothing (LOWESS) trendline superimposed on top of it.

4.3.5 *Simulation of potential changes in authentication policy*

Two potential changes to authentication policy were considered and their effects simulated within the second half of the Q1 time period. Policy change #1 required a challenge for each workstation a user logs into that the user has not before and updates each user's list of accessed workstations in an ongoing manner. We considered this policy change in two different contexts: 1) previous user workstation access history is ported into the algorithm used to determine whether the workstation is new *prior* to implementation and 2) no previous user workstation access history is ported into the algorithm used to determine whether the workstation is new prior to implementation. For the first context, we incorporated users' workstation access history from the first half of Q1. The burden of policy change #1 was compared to the default, unchanged observed SSO data (which uses a challenge period of 4 hours) by calculating the number of extra challenges resulting from the policy in each context divided by the total count of shared workstation logins. Thus, burden is the percentage point increase in shared workstation logins requiring a challenge relative to the default. Burden was stratified by grouped role and location.

Policy change #2 varied the challenge period from one hour to 12 hours in increments of one hour. The number of challenges was compared relative to the 4-hour challenge period by calculating the number of extra challenges resulting from the policy in each context divided by the total count of shared workstation logins. Thus, burden is the percentage point increase in shared workstation logins requiring a challenge relative to the 4-hour challenge period. Burden was again stratified by grouped role and location.

4.3.6 *Privacy Preserving Technologies*

We used the Synthetic Data Vault's[17] Deep Echo model[18] for time series to generate a synthetic data set from the dataset used to analyze the simulation of potential changes in authentication policy described above as our training data set. The training set was pre-processed by filtering to solely shared workstations logins and reducing its columns to solely Datetime, User ID, Method (e.g. password), Result (e.g. 'successful'), grouped role, and grouped location. Deep Echo parameters[18] were set in the following manner: sequence_index as the datetime column, entity_columns as the user column, and context_columns was set to solely the binned group role. The model was trained on a random 12.5% sample of users, specifically selecting a random sample that contained at least one user in each grouped role, for 20 epochs.

After the model was trained, the same number of users with shared workstation logins (n=3035) of the 3,177 in the real data were generated from the model to create the synthetic data set. Although the PAR model learns the distribution of the lengths of the sequences, the learned distribution yielded only roughly 0-30 rows (shared workstation logins) per user in preliminary synthetic data generations which was over an order of magnitude reduction in the range of row length per users in the real data. As a consequence, a specified row length of 100 per user was used for the final synthetic data set opposed to relying on the model's learned row length.

Additionally, dates were manually shifted three weeks forward to improve the similarity of the date distribution between the synthetic and real data so that both the first and second halves of Q1 and Q2 could be studied. The synthetic data were then used to assess the impact of policy #1 and #2, with its results compared to the original data.

4.4 RESULTS

4.4.1 *Characterization of SSO TITO activity and its rollout within the UW Medicine system*

After a period of testing, SSO was steadily rolled out in waves to the University of Washington Health System's hospitals and outpatient clinics starting July 2018 (red vertical line seen in Figure 4.1). A Dip that coincides with COVID-19's effects on UW Medicine can be seen starting in March 2020 and dropping further in April 2020.

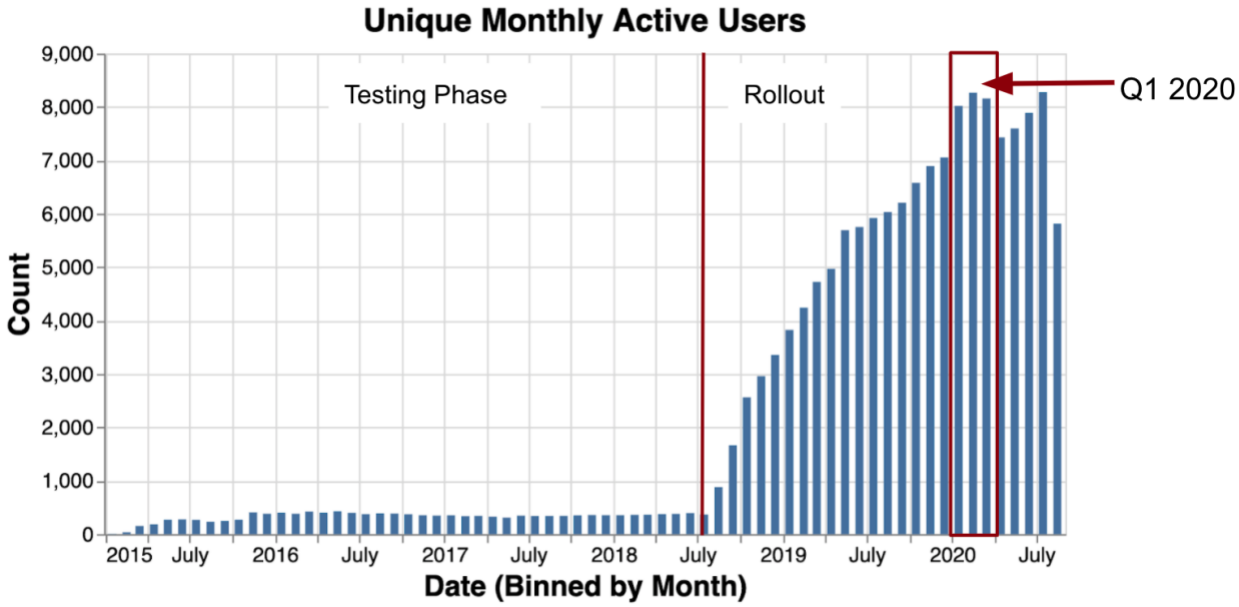

Figure 4.1. Unique monthly active users over the entire time period of data collection. Testing and rollout phases are divided by a red vertical bar. The effect of COVID-19 can be seen in the reduction of unique monthly active users late Q1 2020 which reduces further in Q2.

4.4.2 *Summary of the Data*

As seen in Table 4.10, there were a total of 4236 unique users with medical licenses that could be grouped into similar roles with ten or more users during Q1 2020. Grouping licenses by similar roles yielded 15 categories with at least 10 or more unique users. Only <10 users could not be grouped into a category of similar roles. By far the most common role categories were those consisting of physicians. Physicians and fellows (n = 1,428; 34%) were the most common role of all those grouped followed by RNs and LPNs [n = 968; 23% (this grouping consisted of <10 LPNs)], then physician residents (n = 587; 14%), and nurse practitioners (n = 321; 8%). In Q1 2020, these users totaled 817,221 shared workstations log-on, of which proximity card only was the most common method to log-on (n = 589,038; 72.1%) followed by proximity card + password (n = 120,291; 14.7%), then password only (n = 107,892; 13.2%). Therefore, challenges currently make up 14.7% of users log-ons.

Table 4.10. Characteristics of the dataset in Q1 2020, filtered to data resulting from users with medical licenses.

	Time Period			
	Q1 2020	**Jan 2020**	**Feb 2020**	**March 2020**
Unique Users (n)	4,236*	3,504	3,614	3,606
Physicians, fellows	1,428	1,109	1,169	1,177
Nurses: RNs & LPNs	968	862	860	866
Physician residents	587	503	500	484
Nurses: nurse practitioners (NP)	321	248	266	270
Occ/phys therapists, assistants	170	158	157	153
Counselors, therapists, social workers	168	132	136	140
Physician assistants	156	121	138	138

Pharm: pharmacists, interns, techs	117	88	95	93
Respiratory therapists	106	98	99	99
Medical assistants/techs	87	79	81	81
Speech language pathologists	40	36	34	37
Dieticians	33	27	31	27
Dentists, residents	24	17	16	15
Psychologists	21	14	19	15
Podiatrists	10	<10	<10	<10
Ungrouped	<10	<10	<10	<10
Unique workstations accessed (n)	3,972	3,683	3,798	3,606
Total system events (n)	3,141,374	1,090,173	1,079,516	900,181
Enable SSO (n)	956,552	337,089	327,522	269,883
Site: Harborview (relative %)	353,869 (37)	127,908 (38)	116,904 (36)	101,053 (37)
Site: UWMC (relative %)	428,433 (45)	153,429 (46)	141,187 (43)	123,582 (46)
Site: Outpatient Clinics (relative %)	174,250 (18)	55,752 (17)	69,431 (21)	45,248 (17)
Care location: ICU (relative %)	120,609 (13)	43,034 (13)	39,731 (12)	34,699 (13)
Care location: ED (relative %)	33,075 (3)	12,348 (4)	10,491 (3)	9,519 (4)
Care location: Other (relative %)	802,868 (84)	281,707 (84)	277,300 (85)	225,665 (84)
Shared Workstation Login	817,221	289,827	272,388	235,732
Proximity Card only (relative %)	589,038 (72)	212,046 (73)	197,482 (73)	165,133 (70)
Proximity Card + Password (Relative %)	120,291 (15)	41,436 (14)	39,304 (14)	36,863 (16)
Password only (relative %)	107,892 (13)	36,345 (13)	35,602 (13)	33,736 (14)

Locked	806,833	283,313	268,895	235,572
Shutdown Agent	18,694	6,447	6,630	5,229
Enroll Proximity Card	328	185	82	57
Replace proximity card	1,742	670	567	478
Enroll Password	590	308	165	102
Locked Out	98	36	34	26
Locked Out (Confirm ID)	29	2	11	16

*does not include ungrouped in the count

4.4.3 *User-Specific Behaviors*

The distributions of user behavior in unique workstations per workday, unique workstations accessed, and days worked was positively skewed and can be seen in Figure 4.2.

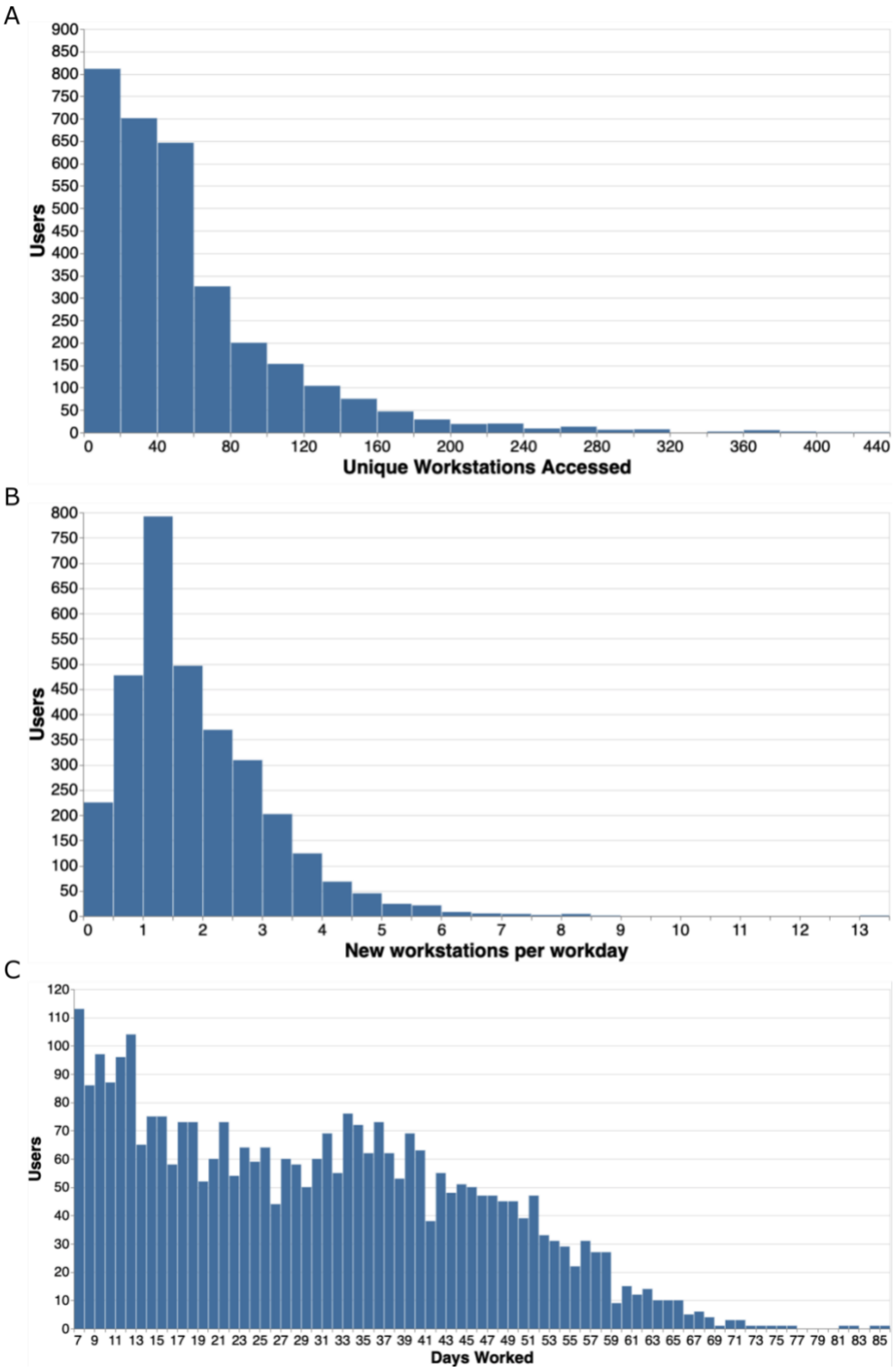

Figure 4.2. Distributions of individual user behavior in those with medical licenses who worked at least 7 days during Q2 2020.

Counts of users by A) unique workstations accessed, B) new workstations per workday, and C) days worked over this time period.

The rate at which users reached the maximum number of workstations accessed is shown in Figure 4.3.

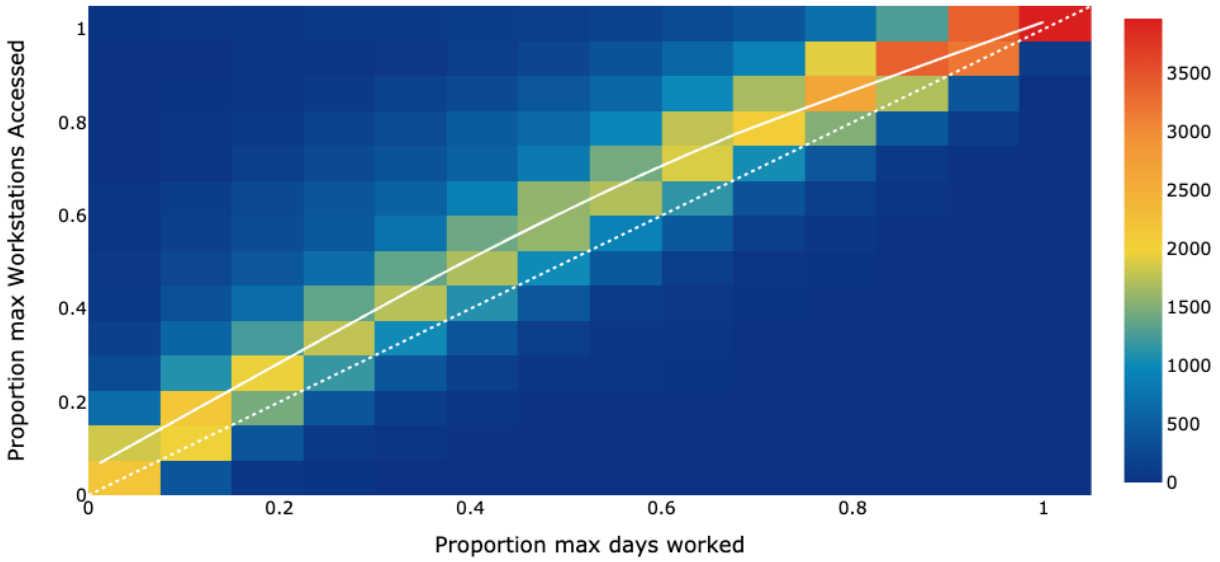

Figure 4.3. The rate at which unique workstations range is reached per user in those who worked 7 days or more Q1 and had had a medical license.
Dashed line shows y=x, solid white line shows the LOWESS of all data. Data are adjusted by the max number of days worked by each individual and the max number of workstations accessed. Bin size for both x and y axes is 0.075.

As seen in Table 4.11, users worked 28 days (IQR = 26), accessed 42 unique workstations (IQR = 50), had a range of days from their 1st to last day worked of 82 (IQR = 19.0) and accessed 1.57 (IQR = 1.53) new workstations per workday.

Table 4.11. Unique workstations accessed amongst users (n=3177) who worked at least seven days January-March 2020

	Mean (stdev)	**Median (IQR)**
Days worked[1]	29.55(15.71)	28.0(26.0)
Unique workstations accessed[2]	56.2(53.53)	42.0(51.0)

Range from 1st to last day worked[3]	75.48(17.09)	82(19.0)
New workstations per workday[4]	1.87(1.22)	1.57(1.53)

[1]Number of days where data are present
[2]Unique number of computers the user logged into
[3]Date difference between the 1st day the user had data and the last
[4]Unique workstations accessed/Workdays

4.4.4 *Simulation of potential changes in authentication policy*

Policy change #1 (challenges for new workstations)'s burden - measured by the percentage point change of shared workstation log-ons requiring a challenge compared to the policy not being implemented - was variable between roles and less so by location (Figure 4.4). Context #1 (½ a quarter's prior history) resulted in increases of burden ranging from a minimum of 0.75 for podiatrists and 14.88 percentage points for speech language pathologists. All grouped roles averaged a 4.2 percentage point increase. The emergency department (ED) had the lowest burden increase at 2.51 compared to the ICU with the highest at 4.82 percentage points. Context #2 (no prior history) resulted in increases ranging from a minimum of 2.16 for podiatrists, 22.54 for speech language pathologists and 7.16 percentage points for all grouped roles. Overall, context #2 roughly doubled the burden of context #1 across each location. However, the burden reduction of incorporating prior history in context #1 relative to context #2 was variable across roles.

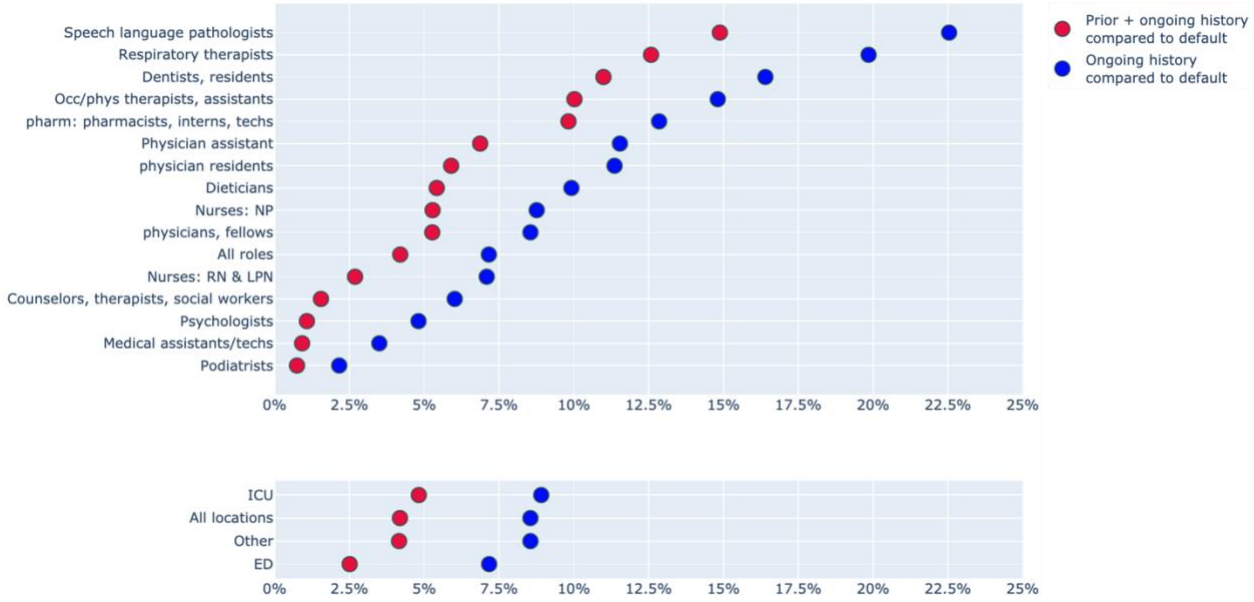

Figure 4.4. Real Data: Differences in burden over 2nd half of Q1 when not allowing solely proximity card to be used for shared workstation login to new workstations with and without prior user workstation access history in first half of Q1.
X-axis shows the percentage point change in shared workstation logins requiring a challenge relative to the default (4 hours).

Policy change #2 (variation of challenge period) burden - measured by the percentage point change of log-ons requiring a challenge compared to the 4-hour challenge period - was also variable between roles and less so between locations (Figure 4.5). Across all grouped roles, the 12-hour challenge period reduced burden by 6.67 and the 1-hour challenge period increased burden by 18.13 percentage points relative to the 4-hour challenge period. Shift length is associated with burden. Roles that do not typically work 12-hour shifts such as speech language pathologists, psychologists, and podiatrists experienced no or nearly no decrease in burden when increasing the challenge period from eight hours to twelve hours. Roles that do typically work 12-hour or longer shifts such as nurses, respiratory therapists and physician residents experienced incremental decreases in burden as the challenge period increased from eight to twelve hours with a noticeable decrease from eleven to twelve hours observed in nurses and

physician residents. ICU locations had the biggest range between 1 hour and 12 hours' burdens at +19.99 and -8.41 percentage points, respectively. Both the ED and ICU - and much less so other locations - experienced noticeable decreases in burden from 11 to 12 hours similar to the effect observed in roles that typically work 12-hour shifts.

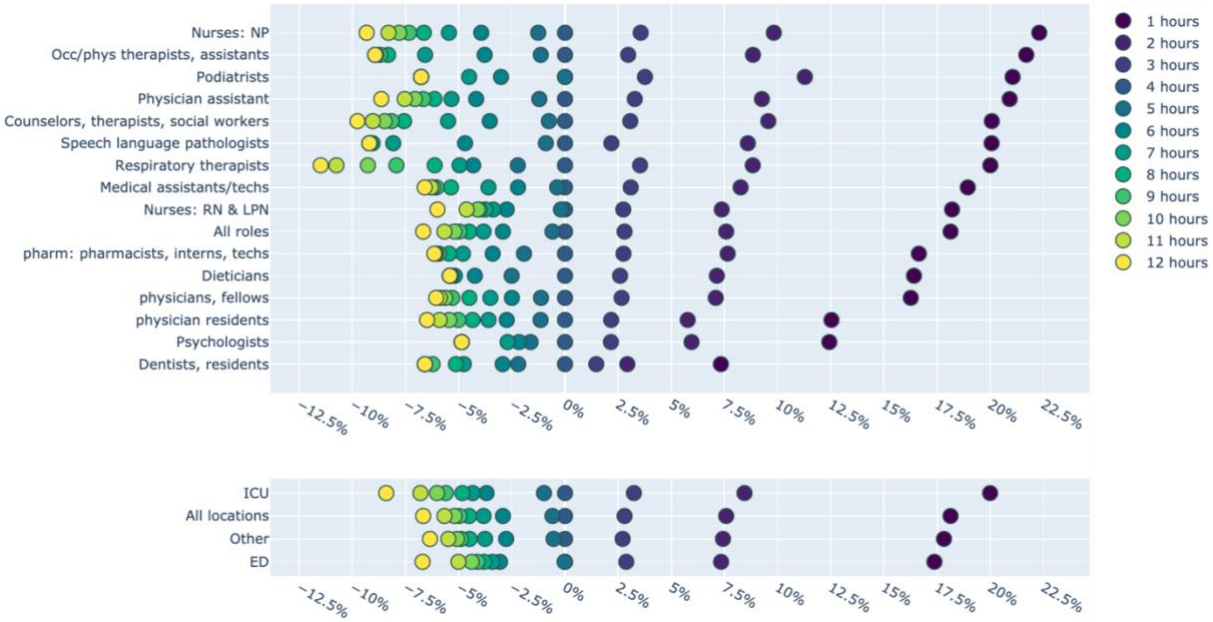

Figure 4.5. Real data: Differences in burden over 2nd half of Q1 when implementing longer or shorter challenge periods.
X-axis shows the percentage point change in shared workstation logins requiring a challenge relative to the default (4 hours).

4.4.5 *Privacy Preserving Technologies*

Model training used up to 65GB of memory at its peak consumption. The model's loss function was still decreasing without flattening out after 20 epochs so training was likely cut short of an optimal number of epochs. Automatic One-hot encoding of the locations was likely the cause of high memory use for the data set while training which is the reason a 12.5% sample was used rather than a larger sample.

The synthetic data's results for policy #1 (challenges for new workstations) were skewed to a higher burden for all roles and locations than the real data (Figure 4.6). In addition, the synthetic data results showed a much narrower difference between the two contexts (prior history vs. no prior history).

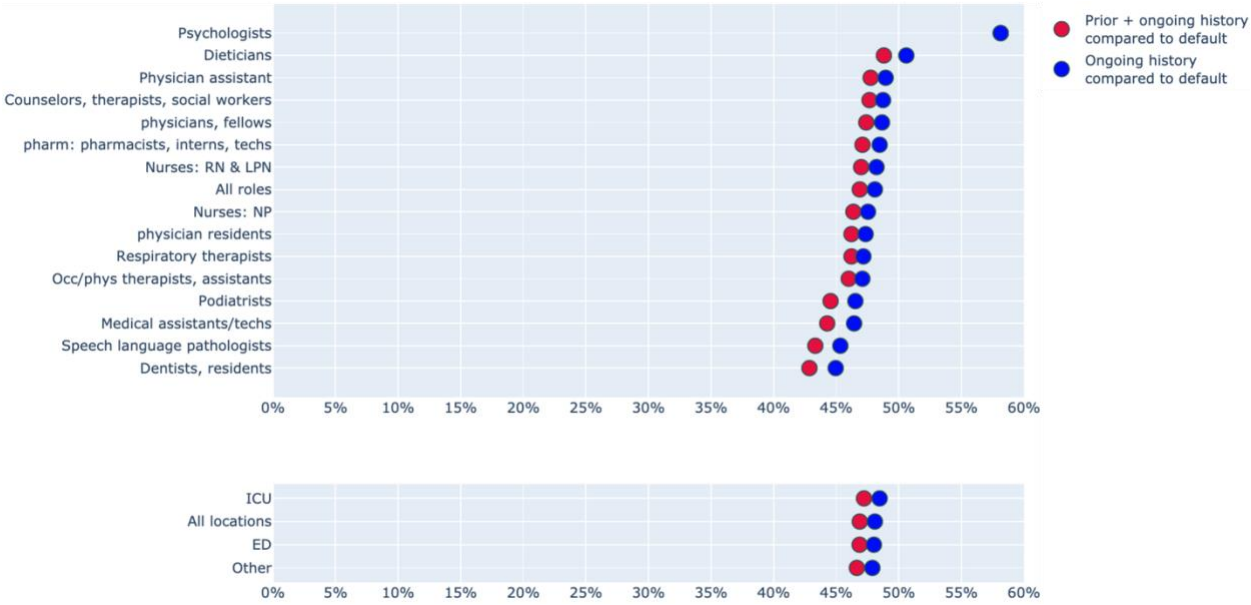

Figure 4.6. Synthetic data: Differences in burden over 2nd half of Q1 when not allowing solely proximity card to be used for shared workstation login to new workstations with and without prior user workstation access history in first half of Q1.
X-axis shows the percentage point change in shared workstation logins requiring a challenge relative to the default (4 hours).

The synthetic data's results for policy #2 (variation in challenge period) had less variation between roles and location than the real data (Figure 4.7). The synthetic data results had a larger range on average between the 12 hour and 1 hour burden reductions or increases and were skewed more towards a reduction in burden than the real data.

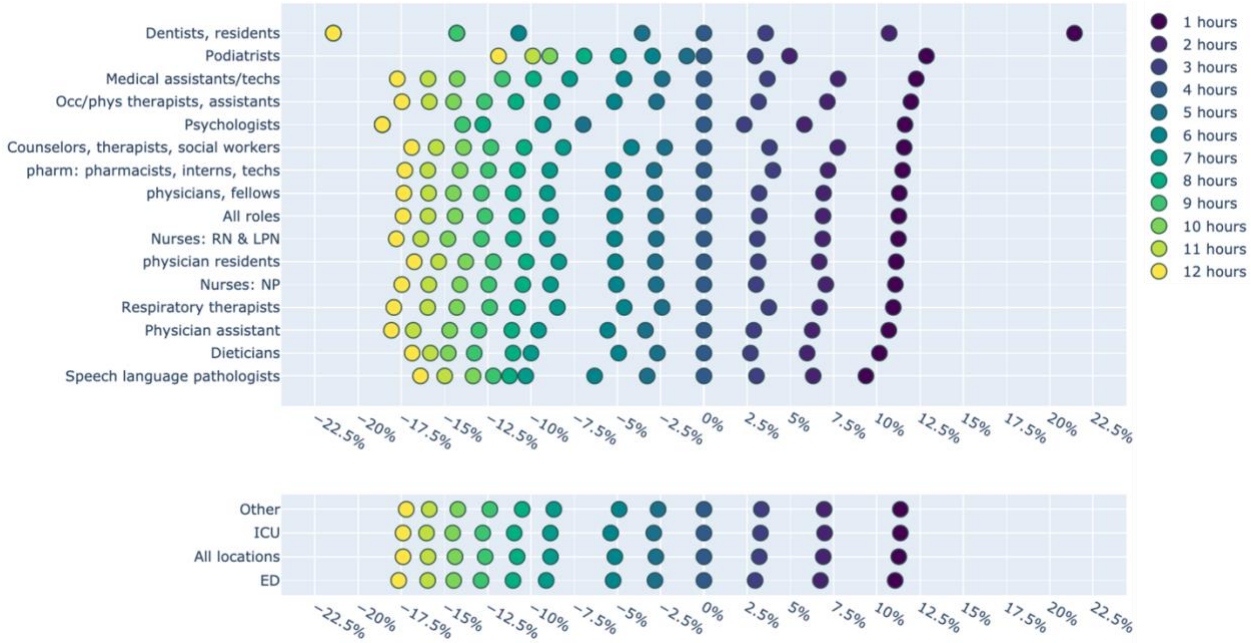

Figure 4.7. Synthetic data: Differences in burden over 2nd half of Q1 when implementing longer or shorter challenge periods.
X-axis shows the percentage point change in shared workstation logins requiring a challenge relative to the default (4 hours).

4.5 Discussion

Shift length's noticeable association with decrease in burden at the longer challenge periods (e.g. 8 to 12 hours) points to the potential for customization of the challenge period in consideration of users' expected shift lengths. In our data, nurses, physician residents, respiratory therapists, and physician assistants are the three groups who would stand to benefit the most from challenge period extensions beyond 8 hours. In contrast, other roles that likely work 8-hour shifts (in our health system) such as speech language pathologists, podiatrists, dieticians, psychologists, and occupational/physical therapists would see little to no benefit from increasing a challenge period beyond 8 hours. However, other health systems[8-11,13-14] should look at their own personnel's shift lengths as shift lengths by role may be variable between and within health systems.

Similarly, the ICU and ED had the most pronounced reduction in burden from challenge period

extensions beyond 8 hours which is likely due to a higher prevalence of 12-hour shifts inpatient vs. outpatient.

The results from policy change #1 show the importance of incorporating prior user behavior if this sort of policy or another approach, rules-based or not, was enacted to detect outlier behavior patterns then require challenges to verify authentication. The expected changes in burden by implementing this policy would not be distributed evenly across roles, location, nor of individuals.

In general, the synthetic data performed poorly in their ability to allow an analyst to reach the same conclusions when analyzing the synthetic data compared to the real data, especially in regards to policy change #1 (challenges for new workstations). The synthetic data results for policy change #1 show the locations were not modeled well since the burden of context #1 compared to #2 (prior history vs. no prior history incorporated) showed little to no difference. Thus, users in the synthetic data access nearly all new workstations in the latter half of Q2 opposed to (in the real data) returning to many of the same workstations. The synthetic data results for policy change #2 (variation of challenge periods) did show some of the features in the real data yet these features were muted. The synthetic data did perform well on conformance to the real data, however, which provides value to analysts whose use case is to build out software infrastructure and/or analysis pipelines on the synthetic data prior to acquiring the real data to conduct actual analyses upon.

Overall, SSO log data and the roles data they were merged with showed that granular behavior patterns and the effects of simulated policy changes can be gleaned from these data. Despite the granularity of SSO workstation information, the challenge in analyzing locations at a high-level of granularity was from limited access to clinically meaningful mappings between

workstations and the type of care provided at that time and at that location. The challenge to determine the type of care provided at each workstation and the likelihood of changes to these mappings over time - as was seen during COVID-19 ICU expansions - suggests that dynamic, automated methods should be developed to do so.

4.6 Limitations

Most analyses were limited to just users who had medical licenses, which excluded roughly half of the unique users found in the SSO data. Location data groupings were very high-level due to a lack of clinically meaningful mappings from the workstation to the type of care provided at that time at that location. Only one synthetic data generation model was tested and the model likely should have been trained for more epochs and with a larger sample size to generate more realistic synthetic data. Additionally, our evaluation of the impact of changes to authentication policy were retrospective and hypothetical opposed to a real-world implementation. Due to the impact of COVID-19, the time period of data analysis was restricted to one quarter of the year.

4.7 Conclusion

Shared clinical workstation login behavior varies between individuals, roles and locations. Characteristics such as shift length have an observable effect on the burden of changes to challenge periods. To reduce burden, policies that operate by detecting outlier behavior in user authentication should incorporate prior user history to the model or rules-based system before rolling out such a policy. These findings should be factored into policy discussions as they present an opportunity to afford users customized authentication solutions, which is the number one thing workers would like in exchange for employer use of their data. Synthetic data that were generated from a model trained for a limited number of epochs on a subset of the real

performed well on conformance but poorly on replicating the results of analyses conducted on the real data. Overall, SSO log data and the roles data they were merged with showed that granular authentication behavior patterns and the effects of simulated policy changes can be gleaned from these data.

4.8 ACKNOWLEDGEMENTS

The authors would like to thank other members of the UW Medicine security team including: Jesse Reith for providing us with guidance and delegating tasks necessary for us to obtain the TITO log and roles data, Neil Vernon for his expertise and detailed help regarding the identity and access management system, and Paul Adriance for help answering our questions about the TITO rollout and other aspects of endpoint management. Furthermore, we would like to thank Dr. Thomas Payne for his clinical guidance.

4.10 SUPPLEMENT

Table 4.12. Mapping of Washington State licenses to grouped roles

Grouped Role	**License Prefix**
physicians, fellows	MD
physicians, fellows	OP
physicians, fellows	FE
physicians, fellows	TR
physician residents	OL
physician residents	ML
Occ/phys therapists, assistants	OT
Occ/phys therapists, assistants	PT
Occ/phys therapists, assistants	OC
Occ/phys therapists, assistants	P1
Counselors, therapists, social workers	CG
Counselors, therapists, social workers	LH
Counselors, therapists, social workers	RC
Counselors, therapists, social workers	GT
Counselors, therapists, social workers	SW
Counselors, therapists, social workers	SA
Counselors, therapists, social workers	SC
Counselors, therapists, social workers	LW
Counselors, therapists, social workers	RE
Counselors, therapists, social workers	FX

Counselors, therapists, social workers	CP
pharm: pharmacists, interns, techs	PH
pharm: pharmacists, interns, techs	IR
pharm: pharmacists, interns, techs	VA
Physician assistant	OA
Physician assistant	TA
Physician assistant	PA
Respiratory therapists	LR
Medical assistants/techs	MR
Medical assistants/techs	PC
Medical assistants/techs	NC
Medical assistants/techs	CM
Medical assistants/techs	HC
Medical assistants/techs	NS
Dentists, residents	DE
Dentists, residents	GA
Dentists, residents	DF
Dentists, residents	DR
Podiatrists	PO
Dieticians	DI
Nurses: RN & LPN	RN
Nurses: RN & LPN	LP
Nurses: NP	AP

Psychologists	PY
Speech language pathologists	LL

*****Less than 10 personnel with the following rare licenses were not placed into grouped roles for analyses by role: LD (audiologist), NT (naturopathic physician license), MA (massage therapist license), MW (midwife license), PS (prosthetist license)

4.11 REFERENCES FOR CHAPTER 4

1 Yan Q, Jiang Z, Harbin Z, *et al.* Exploring the relationship between electronic health records and provider burnout: A systematic review. *J Am Med Inform Assoc* 2021;**28**:1009–21. doi:10.1093/jamia/ocab009

2 Kroth PJ, Morioka-Douglas N, Veres S, *et al.* Association of Electronic Health Record Design and Use Factors With Clinician Stress and Burnout. *JAMA Netw Open* 2019;**2**:e199609. doi:10.1001/jamanetworkopen.2019.9609

3 Nguyen OT, Jenkins NJ, Khanna N, *et al.* A systematic review of contributing factors of and solutions to electronic health record–related impacts on physician well-being. *J Am Med Inform Assoc* 2021;**28**:974–84. doi:10.1093/jamia/ocaa339

4 Gettinger A, Zayas-Cabán T. HITECH to 21st century cures: clinician burden and evolving health IT policy. *J Am Med Inform Assoc* 2021;**28**:1022–5. doi:10.1093/jamia/ocaa330

5 Safety-enhanced design | HealthIT.gov. https://www.healthit.gov/test-method/safety-enhanced-design#ccg (accessed 1 Jul 2021).

6 Office of the National Coordinator for Health Information Technology. Strategy on Reducing Regulatory and Administrative Burden Relating to the Use of Health IT and EHRs Final Report As Required by the 21st Century Cures Act Public Law 114-255, Section 4001. Washington, DC: : ONC 2020.

7 25x5 Symposium Home Page - Columbia DBMI. 2020.https://www.dbmi.columbia.edu/25x5/ (accessed 1 Jul 2021).

8 Fontaine J, Zheng K, Van De Ven C, *et al.* Evaluation of a proximity card authentication system for health care settings. *Int J Med Inf* 2016;**92**:1–7. doi:10.1016/j.ijmedinf.2016.04.015

9 Hope P, Zhang X. Examining user satisfaction with single sign-on and computer application roaming within emergency departments. *Health Informatics J* 2015;**21**:107–19. doi:10.1177/1460458213505572

10 Gellert GA, Crouch JF, Gibson LA, *et al.* Clinical impact and value of workstation single sign-on. *Int J Med Inf* 2017;**101**:131–6. doi:10.1016/j.ijmedinf.2017.02.008

11 Gellert GA, Ramirez R, Jacobs WJ, *et al.* Electronic Health Record Workstation Single Sign-on: A Quantification of Time Liberated for Nurses to Care for Patients. *JONA J Nurs Adm* 2020;**50**:462–7. doi:10.1097/NNA.0000000000000917

12 Shook E, Knickrehm M, Sage-Gavin E. Putting Trust to Work: Decoding Organizational DNA: Trust, Data and Unlocking Value in the Digital Workplace.

https://www.accenture.com/_acnmedia/thought-leadership-assets/pdf/accenture-wf-decoding-organizational-dna.pdf (accessed 30 Jun 2021).

13 Love at first tap: One sign to minimize nurses' time to clinical applications. http://3.138.46.196/handle/10755/20389 (accessed 5 Jul 2021).

14 James N, Marwaha S, Brough S, *et al.* Impact of Single Sign-on Adoption in an Assessment Triage Unit: A Hospital's Journey to Higher Efficiency. *JONA J Nurs Adm* 2020;**50**:159–64. doi:10.1097/NNA.0000000000000860

15 Adler-Milstein J, Adelman JS, Tai-Seale M, *et al.* EHR audit logs: A new goldmine for health services research? *J Biomed Inform* 2020;**101**:103343. doi:10.1016/j.jbi.2019.103343

16 Washington State Department of Health Provider Credential Search. https://fortress.wa.gov/doh/providercredentialsearch/

17 Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016. 399–410. doi:10.1109/DSAA.2016.49

18 Sala C, Zhang K, Hofmann F. *sdv-dev/DeepEcho*. The Synthetic Data Vault Project 2021. https://github.com/sdv-dev/DeepEcho (accessed 1 Jul 2021).

Chapter 5. CONCLUSION

5.1 OVERVIEW

The three use cases described in Chapters 2-4 address the study aims by advancing our understanding of 1) the fitness for use of varied electronic health record and clinical workstation log data with and without privacy preserving technologies as well as 2) methods to conduct these assessments. Assessing fitness for use and privacy-preserving technologies are interconnected because applying these technologies produces a privacy-utility tradeoff. To manage the tradeoff, one must be able to assess both privacy (out of scope for this work) and utility. As the use of synthetic data rises, so will the importance of fitness for use assessments on both original and synthetic data. Synthetic data that are broadly distributed will reach less expert users than those who have access to the original data. Thus, in addition to helping those creating synthetic data manage tradeoffs, fitness for use assessments will provide guidance to synthetic data end-users on 1) the approximate similarity between the synthetic data and the original data as well as 2) the overall limitations of the likely inaccessible (to the end-users, at least at the time of analyzing the synthetic data) original data which have a downstream effect on the synthetic data.

Synthetic data can be generated a variety of ways, however, so the results from one synthetic dataset's utility analysis do not necessarily relate to the performance of synthetic data generated for the same purpose but by a different method. Each use case studied in this dissertation analyzed synthetic data generated using distinct methods. Thus, our study results related to synthetic data must be interpreted in light of the synthetic data generation methods used which are outlined below.

The synthetic data in use case 1 (Chapter 2) were produced by Synthea. Synthea generates synthetic data using PADARSER, the Publicly Available Data Approach to the

Realistic Synthetic EHR[1], which creates synthetic data that use publicly available statistics and clinical practice guidelines in an attempt to create realistic EHR data.[2] MDClone generated the synthetic data in use case 2 (Chapter 3) using a computational derivation approach which takes real data then creates synthetic data modeled from the real data to match the co-variance and quantitative distributions of the real data.[3] In brief, the proprietary MDClone synthetic data generation process has three main steps and data are generated on-demand with their output in the tidy format.[4] First, its query tool is used for cohort identification to select a cohort of interest from the real data. Then, features of interest are selected to be extracted from each user within the cohort of interest resulting in a matrix of features of interest for patients within the cohort. Then, statistical models of groups of similar patients are created through the use of "a variation of a kernel density estimation of the multivariate probability density" which are used to create synthetic data representing the cohort (and their features) of interest.[3] In use case 3 (chapter 4), the synthetic data were generated using a computational derivation approach as well but the methods were dissimilar from MDClone's in multiple ways. In contrast to the MDClone methods, a different, open-source model was used which is applied to an entire real data set producing synthetic data with matching conformance to the real data set. The model used was a probabilistic autoregressive (PAR) model[5] contained in the Synthetic Data Vault's Deep Echo package.[6] The PAR model is particularly well suited for modeling time-series. Additionally, the model accepts user-defined parameters that identify groups of rows associated with a single entity (e.g. a specific SSO user in use case 3), and gives users the ability to generate a custom number of rows per entity as well as number of entities along with other customizations. Thus, not only were three separate use cases studied in this dissertation but also three separate synthetic datasets generated for different purposes by different methods.

The three dissertations aims were addressed through the analyses described in Chapters 2-4. The second chapter addressed aim 1 by creating a repository of clinical findings from the Cochrane Database of Systematic Reviews(CDSR)[7] that can augment traditional clinical trials and enable broad fitness for use determinations of synthetic and original EHR data by replicating these findings within EHR data. Of the 50 CDSR reviews assessed, 30% were eligible for our repository of findings which suggests the feasibility of replication.

The second and third chapter addressed Aim 2 by assessing the utility of real and synthetic electronic health records to conduct observational research in varied contexts. In chapter two, we replicated 31 COVID-19 related outcomes studied[8] within the Cochrane Database of Systematic Reviews on two real and one synthetic (Synthea COVID-19) database[9,10] by making use of the Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS)[11] open-source software package. We compared the results of replications within our databases to the results reported in the CDSR. Then, we performed a qualitative analysis to investigate why the calculated outcomes in each database may differ from the outcome result in the CDSR. We found that our real and synthetic EHR databases did have lower values for prevalence and incidence of cardiovascular events compared to the CDSR review's weighted averages for the vast majority of outcomes. We observed heterogeneity between databases. The Synthetic data set released in the Spring of 2020 likely was not fit for use for analysis of cardiovascular outcomes related to COVID-19. Improvements to the UWM-CRD database over time were observed in the improved 2021 UWM-CRD performance relative to the same database's results a year prior.

In chapter three, we analyzed real and synthetic individual-level electronic health record data from the National COVID Cohort Collaborative (N3C) to assess whether or not the

synthetic data could be used for geospatial and temporal epidemic analyses[12]. We decided to focus on analyses that were of common interest such as epidemic curves for key indicators and creation of public-facing dashboards. We conducted both replication of studies and also compared general summary statistics between the real and synthetic data. Overall, we found that synthetic data could successfully be used to analyze geospatial and temporal trends. However, we found that analyses using small sample sizes or populations were limited, in part due to purposeful data label suppression - an attribute disclosure countermeasure. More specifically, we found that a caveat to synthetic data use was its utility to analyze rural N3C populations since nearly all zip codes with <10 tests were censored and much more likely to be rural within the original data. In those cases, we believe users should consider data fitness for use.

In chapter four, we addressed aim 3 by using Imprivata Onesign SSO[13] log data - both real and synthetic - from the UW Medicine Health system to inform customized SSO authentication protocols and reported on the utility of observational SSO log data to do so. We characterized SSO behaviors broadly and also considered two potential SSO implementation changes and their simulated impacts stratified by user role and location. The first potential change was variation of the challenge period from 1-12 hours in 1-hour increments. The second was requiring a challenge for each new workstation a user logs into with and without incorporating their prior workstation access history. We piloted the creation and use of synthetic SSO log data to re-create portions of our analysis in an effort to protect the privacy of worker data. Overall, we found that shared clinical workstation login behavior varies between individuals, roles and locations. Characteristics such as shift length have an observable effect on the burden of changes to challenge periods. To reduce burden, policies that operate by detecting outlier behavior in user authentication should incorporate prior user history to the model or rules-

based system before rolling out such a policy. These findings should be factored into policy discussions as they present an opportunity to afford users customized authentication solutions, which is the number one thing workers would like in exchange for employer use of their data. Synthetic data that were generated from a model[6] trained for a limited number of epochs on a subset of the real performed well on conformance but poorly on replicating the results of analyses conducted on the real data. The synthetic data performed better at modeling users' general cadence of workstation access than they did modeling users' movements from one workstation to another. Regarding utility of the data, SSO log data and the roles data they were merged with showed that granular authentication behavior patterns and the effects of simulated policy changes can be gleaned from these data. However, role data for nearly half of the SSO users is reported by each users' manager in a free text field which led to these users being excluded from our analysis due to poor data quality.

5.2 LIMITATIONS

5.2.1 *Aim 1 Limitations*

Our study (the first half of Chapter 2) was limited in its size and scope. This was due to our strategic decision to conduct the additional yet timely, important, and relevant (to Aim 2) work of assessing the fitness for use of National COVID Cohort Collaborative synthetic data (Chapter 3). To accommodate the extra study, the scope of aim 1 was decreased. The final review of the CDSR to assess replication feasibility consisted of evaluating 50 total reviews and their outcomes.

5.2.2 *Aim 2 Limitations*

The replications conducted in Chapter 2 were on more than 30 outcomes yet they were from a single Cochrane Review with a focus on COVID-19 and cardiovascular events. Thus, our replications are best suited to aiding the assessment of fitness for use of EHR data in either or both of these domains. The Synthea COVID-19 data assessed were limited in size and have been improved upon in future iterations. To date, no privacy analysis has been published on the synthetic data assessed in Chapter 2 nor Chapter 3 to provide context for their utility in relation to their privacy. Both synthetic data sets tested are, to some extent, outdated since they have not been modeled on the most recent source data. In Chapter 3 we recognize that other statistical techniques may have been superior methods to detect significant differences between epidemic curves. Additionally, the Wilcoxon signed-rank and paired t-tests used in Chapter 3 assume the null hypothesis that the original and synthetic datasets are equivalent. Equivalence testing, which flips the null hypothesis, may be better suited. Equivalence testing was not used due to the challenge of selecting an equivalence bound without knowing what threshold(s) data end-users would find most applicable.

5.2.3 *Aim 3 Limitations*

Most analyses in Chapter 4 were limited to just users who had medical licenses, which excluded roughly half of the unique users found in the SSO data. Location data groupings were very high-level due to a lack of clinically meaningful mappings from the workstation to the type of care provided at that time at that location. Only one synthetic data generation model was tested and the model likely could have been trained for more epochs with a larger sample size to generate more realistic synthetic data. No privacy analysis was conducted on the synthetic data to quantify a privacy-utility tradeoff. Additionally, our evaluation of the impact of changes to authentication

policy were retrospective and simulated opposed to a real-world implementation. Due to the impact of COVID-19, the time period of data analysis was restricted to one quarter of the year. The National Research Network[14] did not develop standards for EHR log data over the course of our study which limited our ability to create a standardized analysis.

5.3 Future work and recommendations

5.3.1 *Immediate Considerations*

In future work, we will increase the quantity of synthetic data sets assessed, the quantity and variety of domains of Cochrane reviews to be replicated, and expand beyond prevalence and incidence to study other outcomes such as the effects of interventions. Regarding synthetic N3C data specifically, future work conducting equivalence testing specific to well-defined, high-impact use cases may be merited. Other statistical techniques such as equivalence testing, bhattacharyya distance[15,16], or adversarial challenges[17] could be used in the future to compare similarity between epidemic curves. Our results also may inform future N3C discussions about data set balancing ranging from 1) creation of artificially balanced hybrid data sets to improve statistical models' performance on underrepresented data, 2) source partners sending a random sample of negative tests alongside all positive tests, or 3) expansion of data ingestion from rural regions. The synthetic data used in Chapter 4 would be improved by training for a significantly longer duration and with a larger training data set.

5.3.2 *Long-Term Directions*

Over time, the repository of clinical findings (Chapter 2) must be built out larger in quantity of findings and in breadth of domains. The repository would likely benefit from incorporating high-

quality evidence beyond the CDSR. Our work in Chapter two will hopefully become less novel over time due to increased use of EHR real-world data to augment traditional clinical trials and increased use of systematic review replication to conduct fitness for use assessments of EHR data. Eventually, our work would be a good candidate for an OHDSI network study[18]. The realization of our framework in a software package could become a part of the OHDSI Methods Library[19] open-source software collection alongside the Data Quality Dashboard.[20]

Once the N3C synthetic data privacy analysis is complete, our utility results described in Chapter 3 must be compared to the privacy afforded by the data set. Since the code for the analysis is stored within the N3C enclave, the analysis can and should be rerun with multiple sets of the synthetic data generated with varying levels of privacy. Doing so will allow a true analysis of the privacy-utility tradeoff of these data and, in part, the MDClone system more broadly. Methods to compare the equivalence of two epidemic curves should also be studied further and improved upon. Currently, there are no clear gold standards to do so which makes analyzing the utility of geospatial synthetic data more challenging.

Chapter four highlights the necessity of creating EHR log data standards. Our work was presented in the fall of 2020 to National Research Network members who agreed that work should be done to enable dynamic mapping of workstation location to the type of care provided at that workstation. This would necessitate knowledge representation progress to 1) describe the expanse of unique care locations 2) define strict criteria for each location (e.g. what are the inclusion/exclusion criteria to be an ICU) and their relationships at varying levels of granularity. The dynamic mappings could then be likely be created by querying against the electronic health records accessed at each workstation to determine the care being provided at that time at that

workstation. Chapter four also points to the value of a prospective study that actually puts the potential policy changes into practice.

5.4 IMPLICATIONS

5.4.1 *Implications for broad determination of fitness for use of EHRs, log data both real and synthetic*

Our findings from Chapter 2 point to a lack of EHR database readiness at the beginning of the COVID-19 pandemic to conduct observational research using the OMOP CDM. The Synthea COVID-19 data set released in the Spring of 2020 was likely not fit for use in the context we assessed it in, and the UWM-CRD as of September 2020 lacked standardized derived elements which dramatically reduced its utility for the same task at hand (COVID-19 prevalence/incidence research and cardiovascular events). We showed that the UWM-CRD's fitness for use was improved over time, however, which demonstrates the value of our fitness for use assessment in general and to aid iterative data quality improvement efforts. Our results and methods may help improve broad fitness for use and pandemic preparedness by allowing institutions and/or developers to proactively, rather than reactively, assess and improve the fitness for use of their data by focusing on replications relevant to future analyses of high priority. Our framework could potentially be used to assess the fitness for use of data submitted to N3C.

5.4.2 *Implications for healthcare worker data, privacy and authentication*

Health systems should begin their own evaluations of their authentication policy to determine if their users could benefit from a more customized approach. Additionally, health systems should take a close look at whether or not they are responsibly using worker data and what solutions/processes (e.g. privacy preserving technologies) they might employ to balance privacy

with usability. At UW Medicine, leadership should consider a shift length-based approach for setting the challenge period. Since the current challenge period is 4 hours, leadership might consider leaving the challenge period at 4 hours for workers who do 8-hour shift and increasing the challenge period to 6 hours for those who do 12-hour shifts.

5.4.3 *Implications for National COVID Cohort Collaborative (N3C) synthetic data access and use*

Our evaluation[6] should provide N3C leadership confidence in the utility of these specific MDClone synthetic data - modeled on the N3C limited data set - to allow for geospatial and temporal analyses that do not require small sample sizes. However, we did uncover biases in the data that 1) N3C leadership should work to ameliorate through any combination of the three suggestions we provided and 2) users should be aware of. Additionally, our analysis will give users a sense for the how to map the results from our study onto their own fitness for use requirements of the data. Thus, they can determine whether the data are good enough or not for their task(s) at hand. Last, our study provides a foundational step towards building up rigorous methods to assess the utility of synthetic geospatial and temporal epidemiologic data. Our methods and the design of our data visualizations hopefully will reduce the burden for others to do similar and/or expanded analyses.

5.5 CONCLUSIONS

We demonstrated the feasibility of replicating CDSR reviews using electronic health record data for both synthetic and real data as a method to assess their fitness for use. Our EHR databases did have lower values for prevalence and incidence of cardiovascular events - likely due to a variety of the challenges in capturing outcomes in structured data - compared to the CDSR

review's weighted averages for the vast majority of outcomes. We observed heterogeneity between databases. The Synthea COVID-19 data set released in the Spring of 2020 may not be fit for use for analysis of cardiovascular outcomes.

In general, synthetic National COVID Cohort Collaborative (N3C) data were successfully used to analyze geospatial and temporal trends. Analyses using small sample sizes or populations were limited, in part due to purposeful data label suppression - an attribute disclosure countermeasure. Users should consider data fitness for use in these cases.

SSO log data and the roles data they were merged with showed that granular, heterogenous behavior patterns and the effects of simulated policy changes can be gleaned from these data. This heterogeneity should be factored into policy discussions and presents an opportunity to afford healthcare workers customized authentication solutions However, obtaining information regarding the type of care being provided at each workstation remains challenging and a target for future work. Synthetic SSO data performed poorly, particularly on modeling user behaviors accessing workstations likely due to a combination of: heterogenous and skewed source data, a large number of unique workstations, and limited training of the model.

5.6 REFERENCES FOR CHAPTER 5

1 Dube K, Gallagher T. Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons J, MacCaull W, eds. *Foundations of Health Information Engineering and Systems*. Berlin, Heidelberg: : Springer 2014. 69–86. doi:10.1007/978-3-642-53956-5_6

2 Walonoski J, Kramer M, Nichols J, *et al.* Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc JAMIA* Published Online First: 30 August 2017. doi:10.1093/jamia/ocx079

3 Foraker RE, Yu SC, Gupta A, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* Published Online First: 14 December 2020. doi:10.1093/jamiaopen/ooaa060

4 Wickham H. Tidy Data. *J Stat Softw* 2014;**59**:1–23. doi:10.18637/jss.v059.i10

5 PAR Model — SDV 0.12.0 documentation. https://sdv.dev/SDV/user_guides/timeseries/par.html (accessed 24 Aug 2021).

6 Sala C, Zhang K, Hofmann F. *sdv-dev/DeepEcho*. The Synthetic Data Vault Project 2021. https://github.com/sdv-dev/DeepEcho (accessed 1 Jul 2021).

7 About the Cochrane Database of Systematic Reviews | Cochrane Library. http://www.cochranelibrary.com/cdsr/about-cdsr (accessed 8 Jul 2019).

8 Pellicori P, Doolub, G, Wong, CM, Lee, KS, Mangion, K, Ahmad, M, Berry, C, Squire, I, Lambiase, PD, Lyon, A, McConnachie, A, Taylor, RS, Cleland J. COVID-19 and its cardiovascular effects: a systematic review of prevalence studies. *Cochrane Database Syst Rev* Published Online First: 2021. doi:10.1002/14651858.CD013879

9 synthea/src/main/resources/modules/covid19 at covid19 · syntheticheath/synthea. GitHub. https://github.com/syntheticheath/synthea (accessed 22 Jul 2021).

10 Shamberger M. Synthetic data with simulated covid outbreak - Developers. OHDSI Forums. 2020.https://forums.ohdsi.org/t/synthetic-data-with-simulated-covid-outbreak/10256 (accessed 17 Jul 2021).

11 Duarte-Salles, Prats-Uribe A, Prieto-Alhambra D, *et al. Charybdis Phenotype Library*. OHDSI Studies 2021. https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis/blob/7907813db56478a12daf67d979809f1337674deb/documents/CharybdisPhenotypeLibrary.csv (accessed 18 Jul 2021).

12 Thomas JA, Foraker RE, Zamstein N, *et al.* Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: Results from analyzing >1.8

million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *medRxiv* Published Online First: 2021. doi:10.1101/2021.07.06.21259051

13 Imprivata OneSign®. Single sign-on (SSO). www.imprivata.com/single-sign-on-sso.

14 National Research Network EHR Audit-log and MEta-data workgroups. https://medicine.ucsf.edu/center-clinical-informatics-and-improvement-research/national-research-network (accessed 27 Feb 2020).

15 Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*. 2000. 142–9 vol.2. doi:10.1109/CVPR.2000.854761

16 Kaloskampis I, Joshi C, Cheung C, *et al.* Synthetic data in the civil service. *Significance* 2020;**17**:18–23. doi:https://doi.org/10.1111/1740-9713.01466

17 El Emam K, Mosquera L, Jonker E, *et al.* Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 2021;**4**. doi:10.1093/jamiaopen/ooab012

18 Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

19 HADES. https://ohdsi.github.io/Hades/ (accessed 26 Jul 2021).

20 Blacketer C, Londhe A, Sena A, *et al. Data Quality Dashboard*. Observational Health Data Sciences and Informatics https://ohdsi.github.io/DataQualityDashboard/ (accessed 21 Jan 2020).