A Model-to-data Approach for Building Accurate Machine Learning Algorithms on EHR Data

Yao Yan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Sean Mooney, Chair

Justin Guinney

Trevor Cohen

Graham Nichol

Program Authorized to Offer Degree:

Molecular Engineering

University of Washington

Abstract

A Model-to-data Approach for Building Accurate Machine Learning Algorithms on EHR Data

Yao Yan

Chair of the Supervisory Committee:

Sean Mooney

Department of Biomedical Informatics and Medical Education

Over the past few decades, information about patients' diagnoses, medication, and procedures has been collected and transformed into standardized and shareable electronic health records (EHRs). Machine learning algorithms have proven efficient for mining predictive clinical patterns from EHRs and thus can be used to guide next-generation personalized medicine and enable effective clinical decision support. However, privacy concerns often limit access to individual patient data, hampering researchers' capability to develop machine learning models and conduct model generalizability assessments. Creating an infrastructure that enables secure utilization of patient data with adequate privacy control is the key to bridging researchers and data, thereby unlocking the full potential of the data. In addition, facilitating the contribution of data from multiple sites and enabling federated evaluation are essential for developing robust and generalizable models and overcoming the barrier to clinical implementation. A 'model to data' approach, in which researchers build and submit models to be evaluated by a trusted party

without direct access to data, can reduce the risk posed by direct data sharing, lower the barrier to federated evaluation, and open up data for utilization by the broader data science community.

In this dissertation, I focus on the implementation of a ‘model to data’ approach for enabling secure utilization of multi-modal patient data, and on synthetic EHR data generation as a complement to this approach. The 4 aims of my dissertation are (1) Piloting a 'model to data' approach to enable patient mortality prediction; (2) Implementing the 'model to data' approach in a crowdsourced benchmarking challenge for COVID-19 outcome prediction; (3) Enabling clinical notes sharing and de-identification through the NLP sandbox; and (4) Benchmarking generative adversarial network (GAN)-related synthetic EHR generation on real-world patient data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Four years ago, in my zest for adventure, I came to the U.S. to do my Ph.D. I was fresh out of college and felt excited about moving to a new place. I was passionate about research but had a very vague idea of what committing years to do a Ph.D. would entail and how this journey would shape me. That was 2018 when a pandemic sounded unimaginable and when going back home was only a 10-hour flight for me.

So much has changed, all in the blink of an eye.

Looking back on the past years, memories flood in intertwined with the pain and gain from growth, stress, and pride from research, and with laughter and tears from life experiences. This journey has transformed me into the person and young scholar that I wanted to be 4 years ago. I would not have been able to make it here without the support and love from so many people around me.

First and foremost, I want to express my deepest gratitude to my Ph.D. advisors Dr. Sean Mooney and Dr. Justin Guinney for their invaluable mentorship and tremendous support throughout the past 4 years. Dr. Mooney and Dr. Guinney devoted a substantial amount of time to helping me improve my writing and presentation skills and guiding me to think critically and innovatively. I want to thank Dr. Mooney for introducing me to the world of electronic health records, patiently pointing me to resources and networks, and generously sharing his knowledge and vision. He inspired me to think big and far, a mindset that I will keep practicing throughout my life. I want to thank Dr. Guinney, whom I admire from the bottom of my heart, for his meticulous scholarship

and sharp thinking. Dr. Guinney provided insightful and constructive guidance throughout my research and helped me to cultivate an appreciation of research. Dr. Mooney and Dr. Guinney provided me with the freedom and flexibility to pursue what truly interested me and they have been a constant source of support and inspiration over the past years. They shaped me into the scholar I am today. I feel extremely lucky and grateful to have studied study under Dr. Mooney and Dr. Guinney.

I would like to thank my other thesis committee members: Dr. Graham Nichol, Dr. Trevor Cohen, and Dr. Sara Mostafavi, who have been immensely supportive of my research and the development of this dissertation. I owe special thanks to Dr. Graham Nichol for offering guidance since the first year of my Ph.D.; as a physician, he always inspires me to think about clinical implementation. I also thank Dr. Trevor Cohen for offering insightful comments and Dr. Sara Mostafavi for her constructive feedback and remarkable insights for my research.

I want to thank my dear colleagues and alumni in the Mooney lab: Tim Bergquist, Chethan Jujjavarapu, Jimmy Phuong, Vikas Pejaver, Noah Hammarlund, Steve Mooney, Sicheng Song, Don Smith, Abhi Pratap, Houda Benlhabib, Yile Chen, Su Xian, Arjun Chakraborty. Special thanks to Tim for being an amazing collaborator and to Chethan, Jimmy, and Tim for their support and care.

I would like to thank all the excellent collaborators I met throughout the years: Kathleen Muenzen, Nic Dobbins, Dr. Meliha Yetisgen, Justin Prosser (University of Washington), Dr. Chao Yan, Dr. Ziqi Zhang, Dr. Zhiyu Wan, Dr. Brad Malin (Vanderbilt University), Dr. Sijia Liu, Dr. Hongfang

Liu (Mayo Clinic), George Kowalski (Medical College of Wisconsin), Connor Boyles, Dr. Jineta Banerjee, Dr. Haley Hunter-Zinck, Dr. Laura Heath, Dr. Elias Chaibub Neto, Vijay Yadav (Sage Bionetworks).

I am tremendously grateful to Sage Bionetworks for funding my research and providing me with a supportive and collaborative environment. I feel extremely fortunate to meet and learn from my wonderful colleagues at Sage. Special thanks to Dr. Larsson Omberg, Dr. Lara Mangravite, Dr. Julie Bletz, Dr. Michael Mason and Dr. James Eddy for their great mentorship. I also want to thank Jiaxin Zheng, Yooree Chan, Mialy DeFelice, Xindi Guo, Thomas Yu, Phil Snyder, and Verena Chung for their support.

This would not be possible without my family's support. My parents and grandparents gave me all they have and raised me with love and care. I deeply thank them for always encouraging me to pursue what I am passionate about and for being so selfless in supporting me. I have not seen my family in China for more than 3 years, but I feel love and support from them all the time.

Last but not least, I want to thank my boyfriend Seth for his love, care, and support, and for the countless moments of laughter and happiness he brought to my life. He has been a great emotional support for me all the time, and his optimism, empathy, and patience were the power source for me to go through the ups and downs in the past years. Of course, special thanks to Seth's family for also being so supportive and loving, and to my furry friends Rocky and Maizy for their company and entertainment.

Chapter 1

INTRODUCTION

1.1 Background

Healthcare institutions substantially increased their use of electronic health records (EHR) in the past decade.[1–3] While the primary drivers of EHR adoption were to support clinical care, financial billing and insurance claims, secondary use of EHR data for supporting clinical research has become more common.[2,4–7] EHR data are multimodal and come in different formats; structured tabular data, unstructured clinical notes, and radiology images are just a few examples. The data contain rich information about patients' demographics, health conditions, and clinical trajectories. The digitalization of health records opens up rich opportunities for machine learning (ML) algorithms to identify disease patterns, enable clinical decision-making, and personalize treatment plans for individual patients.[8–13] Meanwhile, the continuous aggregation of large volumes of patient data can be used to validate and improve the accuracy of predictive models.[11,14–16] Given the size and diversity of the data, ML approaches present opportunities for improving healthcare in a more automated and scalable way. In past years, some ML solutions have moved past testing to deployment, winning administrative approval and clearing regulatory hurdles. For example, the Center for Medicare and Medicaid Services has approved insurance reimbursement for the use of two Artificial Intelligence (AI) systems for medical image diagnosis.[16] Other examples include the FDA-approved deep learning platforms Arterys for finding

lesions within pulmonary and liver computed tomography (CT) and Babylon Health for general health recommendations, which performs ~4,000 clinical consultations per day. [17]

Challenges for Data Sharing

Data sharing is the key to embracing the digital health ecosystem and realizing data's full potential. However, healthcare institutions are faced with the competing demands of data utilization and privacy. [18] A balance needs to be struck between data access and privacy control. Due to the privacy and sensitivity of patient data, restrictions are in place for the sharing of this data; a well-known example is the United States Health Insurance Portability and Accountability Act (HIPAA). [19] These restrictions, though necessary, hinder data accessibility for researchers, and limit their ability to develop models and externally validate their models.[20] In most common research cases, access to patient data is limited to researchers affiliated with health institutions, while many researchers with no healthcare institution affiliations are relegated to smaller public de-identified datasets or inferior synthetic data. De-identified datasets obscure identifiers to satisfy HIPAA requirements, however, the data are still susceptible to re-identification risk. Besides, the removal of date or location PHI limits researchers' capability of conducting temporal or geographical analyses. Synthetic data attempt to maintain the information and format of real data without compromising privacy, but to date, no synthetic data can retain all correlation and temporal information presented in the original clinical repository. Even for researchers connected with a healthcare institution, the turnaround time to get administrative and governance approval can often lead to a delay between the data being available and the study being conducted and result in missed research opportunities. Research collaborations are also often bounded by highly restricted data use agreements or business associate agreements, limiting the scope, duration, quantities, and types

of EHR data that can be shared. This friction has impacted the capability of researchers to develop and validate predictive models.

'Model to data' Approach

In the traditional paradigm of data sharing, data owners directly transfer data to users. However, this paradigm faces challenges when it comes to the sharing of sensitive and private EHR, as mentioned in the section on *Challenges for data sharing.* The 'Model to data' approach is an alternative paradigm that minimizes the risk of compromising privacy by avoiding direct data exposure to users. [21] Under this approach, model developers containerize their models by packaging code with operating system libraries and dependencies using containerization software like Docker and send their containerized models to the data hosts which operate and evaluate the performance of models on behalf of the developers. In this way, the data are utilized in a privacy-protection mode such that only the models, not the model developers, have direct access to the data. Under this approach, data owner enjoys the benefits of data sharing (e.g., recognition, leading research questions, etc.), while preserving the full control of the data. Researchers who otherwise do not have access can develop and evaluate models on private dataset. [22] This approach still has its limitation as (1) the information returned to researchers, though highly restricted, theoretically allows information leakage and (2) no direct data interaction and restricted feedback handicap researchers in parameter tuning and feature selection.

Prior to the development of this dissertation, the ‘model to data’ approach has been applied to enable secure sharing of sensitive biomedical data in the Multiple Myeloma DREAM Challenge and the NCI-DREAM Proteogenomic Challenge, [22] and radiology images in Digital

Mammography DREAM Challenge, [14] but there has been limited evaluation of this approach to analysis of structured and unstructured EHR data.

1.2 Dissertation Aims

In the research conducted for this thesis, I validated, implemented, and scaled up the 'model to data' paradigm for EHR data sharing. In doing so, I demonstrated that the 'model to data' approach could alleviate the tension between data sharing and privacy, and engage the broad data science community to develop predictive models to improve clinical decision-making.

1.2.1 Aim 1 Piloting a 'Model to data' Approach to Enable Patient Mortality Prediction

The development of predictive models for clinical application requires the availability of EHR data, which is complicated by patient privacy concerns. I showcased and implemented the 'model to data' approach as a new mechanism to make private clinical data available for model development without compromising data privacy. In this pilot study, an external researcher without direct data access utilized structured EHR data to build high-performing ML models for all-cause mortality prediction. In this aim, I demonstrated the feasibility and scalability of the ‘model to data’ paradigm and identified weaknesses and corresponding improvements.

1.2.2 Aim 2 Implementing the 'Model to data' Approach in a Crowdsourced Benchmarking Challenge for COVID-19 Outcome Prediction

Lack of COVID-19 patient data has hindered the data science community in developing models to aid in the response to the pandemic. In Aim 2, I leveraged the 'model to data' approach to organize a crowdsourced challenge for addressing pressing clinical questions which arose during the COVID-19 pandemic, including prediction of the likelihood of a COVID-19 diagnosis as well as of short-term outcomes (likelihood of hospitalization within 21 days.). In this aim, the 'model to data' approach was scaled up from involving one participant in Aim 1 to a community challenge with hundreds of participants. The COVID-19 challenge was operated in a continuous benchmarking fashion where the challenge datasets were regularly updated for participants to track model performance over the accumulation of patient datasets. I conducted a post-challenge analysis on top-performing models to research how model performance might vary across different demographic traits, how historical patient data could impact performance results, and whether an ensemble strategy of aggregating heterogeneous predictions could outperform individual models. The post-challenge analysis indicated high utility but existence of bias in the models.

1.2.3 Aim 3 Enabling Clinical Notes Sharing and De-identification through a NLP Sandbox

The evaluation of natural language processing (NLP) models for clinical text de-identification relies on the availability of clinical notes, which is often restricted due to privacy concerns. NLP Sandbox is an approach for alleviating the lack of data and evaluation frameworks for NLP models by adopting the 'model to data' approach. This enables unbiased federated model evaluation without the need for sharing sensitive data from multiple institutions. Synapse collaborative

platform, containerization software, and OpenAPI [23] generator were leveraged to build the NLP Sandbox. Two state-of-the-art NLP de-identification models, Philter [24] and NeuroNER [25] were evaluated using data from three sites in the NLP Sandbox and further validated using data from one external validation site. In Aim 3, I demonstrated the feasibility of using the 'model to data' approach in the context of evaluating models built on clinical notes. I conducted a multi-site evaluation of clinical text de-identification models without the sharing of clinical notes and identified the need for standardized and extensible models and data schemas for the scalability and flexibility of the NLP Sandbox.

1.2.4 Aim 4 Benchmarking GAN-related Synthetic EHR Generation on Real-world Patient Data

In **Aim 4**, I explored a complement to the 'model to data' approach - synthetic data generation. Under the 'model to data' approach, developers build models without access to data, leading to difficulties and inconvenience for feature selection and parameter tuning. Implementing synthetic data generation could improve the dilemma of the 'model to data' approach by allowing model developers to get hands-on experience with data that preserve the utility of real data without compromising patient privacy. Generative Adversarial Networks (GANs) have shown increasing potential for generating high-dimension synthetic patient data. However, the nature of the GAN comes with a tension between utility and privacy, often known as the utility-privacy trade-off, leading to difficulties in fair evaluation of the models. In this Aim, I conducted a benchmarking study to evaluate state-of-the-art GAN-related methods utilizing metrics in both utility and privacy aspects. Datasets from two institutions were leveraged to test the generalizability of the GAN

models on datasets of different sizes and data types. Several use cases were provided to shed light on model selection and interpretation.

1.3 Dissertation Overview

This dissertation serves to advance knowledge of the 'model to data' approach for EHR data sharing and utilization. In this work, I demonstrated the implementation of the 'model to data' approach for sharing multi-modal EHR data and enabling ML model benchmarking through crowdsourced approaches. I further provide synthetic data generation as an improvement for the current 'model to data' approach.

Chapter 2

PILOTING A 'MODEL TO DATA' APPROACH TO ENABLE PATIENT MORTALITY PREDICTION

2.1 Introduction

2.1.1 Electronic Health Records and the Future of Data-driven Healthcare

Healthcare providers substantially increased their use of EHR systems in the past decade.[26] While the primary drivers of EHR adoption were the 2009 HITECH act and the data exchange capabilities of EHRs, secondary use of EHR data intended to improve clinical decision support and healthcare quality also contributed to large-scale adoption.[3] EHRs contain a rich set of information about patients and their health histories, including doctors' notes, medications prescribed, and billing codes.[27] The prevalence of EHR systems in hospitals enables the accumulation and utilization of large clinical data to address specific clinical questions. Researchers have already begun to implement predictive analytics solutions to optimize patient care, including models for 30-day readmissions, mortality, and sepsis.[28] As hospitals improve data capture quality and quantity, opportunities for more granular and impactful prediction questions will become more prevalent.

2.1.2 Hurdles to Clinical Data Access

Healthcare institutions face the challenge of balancing patient privacy and EHR data utilization.[18] Regulatory policies such as HIPAA and HITECH place the onus and financial burden of ensuring the security and privacy of patient records on the healthcare institutions hosting the data. A consequence of these regulations is the difficulty of sharing clinical data within the research community. While these data host/researcher relationships are important and lead to impactful collaborations, they are often limited to intra-institution collaborations, relegating many researchers to smaller public datasets. One exception to this is the Patient Level Prediction (PLP) working group in the Observational Health Data Sciences and Informatics (OHDSI) community, which developed a framework for building and externally validating ML models.[20] While the PLP group has successfully streamlined the process to externally validate model performance, there is still an assumption that the model developers have direct access to an EHR dataset that conforms to the Observational Medical Outcomes Partnerships Common Data Model (OMOP CDM) [29,30] on which they can develop their models. In order to support model-building and testing more widely in the research community, new governance models and technological systems are needed to minimize the risk of re-identification of patients, while maximizing the ease of access and use of the clinical data.

2.1.3 Methods for Sharing Clinical Data

De-identification of EHR data and the generation of synthetic EHR data are two solutions to enable clinical data sharing. De-identification methods focus on removing or obfuscating the 18 identifiers that make up the protected health information (PHI) as defined by HIPAA.[19] De-identification reduces the risk of information leakage but may still leave a unique fingerprint of

information that is susceptible to re-identification.[19,31] De-identified datasets, for example MIMIC-III, are available for research and have led to innovative research studies.[32–34] However, these datasets are either limited in size (MIMIC-III only includes 38,597 distinct adult patients and 49,785 hospital admissions), scope (MIMIC-III is specific to ICU patients), and availability (DUAs are required to use MIMIC-III).

Generated synthetic data attempt to preserve the structure, format, and distributions of real EHR datasets but do not contain identifiable information about real patients.[35] Synthetic data generators, such as medGAN,[33] can generate EHR datasets consisting of high-dimensional discrete variables (both binary and count features), although the temporal information of each EHR entry is not maintained. Methods such as Observational Medical Dataset Simulator(OSIM2) are able to maintain this temporal information but only simulate a subset of the data specific to a use-case (e.g., drug and treatment effects).[36] Synthea uses publicly available data to generate synthetic EHR data but is limited to the 10 most common reasons for primary care encounters and 10 chronic diseases that have the highest morbidity in the United States.[37] To our knowledge, no existing method can generate an entire synthetic repository while preserving complete longitudinal and correlational aspects of all features from the original clinical repository.

2.1.4 'Model to data' Framework

The 'model to data' framework, as described in chapter 1, is an alternative to traditional data sharing methods that allows ML research on private biomedical data. [21] The focus of 'model to data' is to enable the development of analytic tools and predictive models without granting researchers direct, physical access to the data. Instead, a researcher sends a containerized model to the data hosts who are then responsible for running the model on the researcher’s behalf. In contrast to the

methods previously described, where the shared or synthetic data were limited in both scope and size, a 'model to data' approach grants a researcher the ability to use all available data from identified datasets, as those data stay at the host sites, while not giving direct access to the researcher. This strategy enables the protection of confidential data while allowing researchers to leverage complete clinical datasets. The 'model to data' framework relies on modern containerization software such as Docker [38] or Singularity [39] for model portability, which serves as a "vehicle" for sending models designed by a model developer to a secure, isolated, and controlled computing environment where it can be executed on sensitive data. The use of such containerization software not only facilitates the secure delivery and execution of models, but it opens up the ability for integration into cloud environments (*e.g.,* Amazon Web Services (AWS), Google Cloud) for cost-effective and scalable data analysis.

The 'model to data' approach has been successful in a series of recent community challenges but has not yet been shown to work with large, EHR datasets.[40] In Aim 1, I implemented a pilot study of a 'model to data' framework implementation enabling the intake and ingestion of containerized clinical prediction models by a large healthcare institution (the University of Washington health system, UW Medicine, Seattle, WA) to their on-premises secure computing infrastructure. The main goals of this pilot are to demonstrate (1) the operationalization of the 'model to data' approach within a large health system, (2) the ability of the 'model to data' framework to facilitate predictive model development by a researcher (here referred to as the model developer) who does not have direct access to UW Medicine EHR data, and (3) the feasibility of a 'model to data' community challenge for evaluating clinical algorithms on remotely stored and protected patient data.

2.2 Methods

2.2.1 Pilot Data Description

The UW Medicine enterprise data warehouse (EDW) includes patient records from medical sites across the UW Medicine system including the University of Washington Medical Center, Harborview Medical Center, and Northwest Hospital and Medical Center. The EDW gathers data from over 60 sources across these institutions including laboratory results, microbiology reports, demographic data, diagnosis codes, and reported allergies. An analytics team at the University of Washington has transformed the patient records from 2010 to the present day into a standardized data format, the Observational Medical Outcomes Partnerships Common Data Model (OMOP CDM v5.0). For this pilot study, we selected all patients who had at least one visit in the UW OMOP repository, which represented 1.3 million patients, 22 million visits, 33 million procedures, 5 million drug exposure records, 48 million condition records, 10 million observations, and 221 million measurements.

2.2.2 Scientific Question for the Pilot of the 'Model to data' Approach

For this 'model to data' demonstration, the scientific question we asked the model developer to address was, *Given the past electronic health records of each patient, predict the likelihood that he/she will die within the next 180 days following his/her last visit.* Patients who had a death record and whose last visit records were within 180 days of the death date were defined as positives. Negatives were defined as patients whose death records were more than 180 days away from the

last visit or who did not have a death record and whose last visit was at least 180 days prior to the end of the available data.

We selected all-cause mortality as the scientific question due to the abundance and availability of patient outcomes from the Washington state death registry. As UW has linked patient records with state death records, the gold standard benchmarks are not constrained to events happening within the clinic. Moreover, mortality prediction has been thoroughly studied.[2,11,41] 180-day mortality prediction is meaningful for identifying patients who will benefit from palliative care. For these reasons, patient mortality prediction represents a well-defined proof-of-concept study to showcase the potential of the 'model to data' evaluation platform.

2.2.3 Defining the Training and Evaluation Datasets

For this study, we split the data into two sets: the Training and Evaluation Datasets. In a *live* healthcare setting, EHR data is constantly changing and evolving along with clinical practice, and prospective evaluation of predictive models is important to ensure that the clinical decision support (CDS) recommendations generated from model predictions are robust to these changes. We defined the Evaluation Dataset as patients who had more recently visited the clinic prior to our last death record and the Training Dataset as all the other patients. An intent of these methods was that the longitudinal properties of the data would be approximately maintained.

The last death record in the available UW OMOP repository at the time of this study was February 24, 2019. Any record or measurement that was found after this date was excluded from the pilot dataset and this date was defined as "End of Data". When building the Evaluation Dataset, we

considered the date 6 months prior to the End of Data (August 24, 2018) the end of the "Evaluation Window" and the beginning of the evaluation window to be 9 months prior to the evaluation window start (November 24, 2017). We chose a 9-month evaluation window size since this resulted in an 80/20 split between the Training/Evaluation Datasets. We defined the Evaluation Window as the period of time in which, if a patient had a visit, we included that patient and all their records in the Evaluation Dataset. Patients who had visits outside the window, but none within the window, were included in the Training Data. Visit records that fell after the Evaluation Window end were removed from the Evaluation Dataset (**Figure 2.1**, patient 7) and from the Training Dataset for patients who didn't have a confirmed death (**Figure 2.1**, patient 3). We only defined the true positives for the Evaluation Dataset and created a gold standard of these patients' mortality status based on their last visit date and the death table. However, we gave the model developer the flexibility to select prediction dates for patients in the Training Dataset and to create corresponding true positives and true negatives for training purposes.

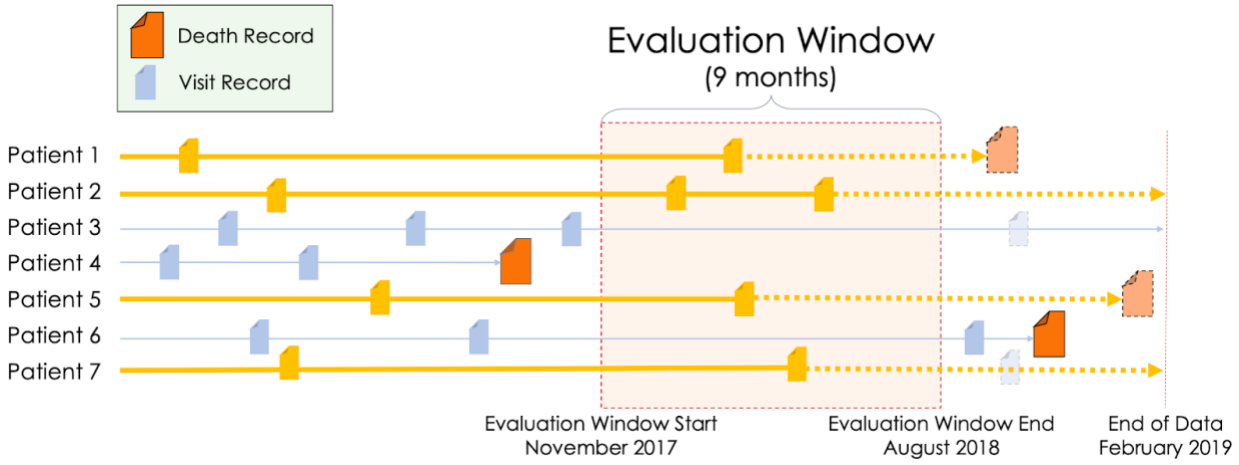


Figure 2.1 Training/Evaluation Dataset Split

Any patient with at least one visit within the Evaluation Window was included in the Evaluation Dataset (gold). All other patient records were added to the Training Dataset (blue). Visits that were

after the Evaluation Window end were excluded from the Evaluation Dataset (light/transparent blue). A 9-month Evaluation Window was chosen as the timeframe as that resulted in an 80/20 split between the Training Dataset and the Evaluation Dataset.

2.2.4 Model Evaluation Pipeline

Docker Containerized Models

Docker is a tool designed to facilitate the sharing of software and dependencies in a single unit called an image.[38] These images make package dependency, language compilation, and environmental variables easier to manage. This technology enables the simulation of an operating system that can be run on any computer that has the Docker engine or compatible container runtime installed. These containers can also be completely isolated from the internet or the server on which they are hosted, an important feature when bringing untrusted programs to process protected data. For this study, the model developer built mortality prediction Docker images, which included dependencies and instructions for running models in the Docker container.

Synapse Collaboration Platform

Synapse is an open-source software platform developed by Sage Bionetworks for researchers to share data, compare and communicate their methodologies, and seek collaboration.[42] Synapse is composed of a set of shared Representational State Transfer (REST)-based web services that support both a website to facilitate collaboration among scientific teams and integration with analysis tools and programming languages to allow computational interactions.[43] The Synapse platform provides services that enable submission of files or Docker images to an evaluation queue, which have previously been used to manage containerized models submitted to DREAM

challenges.[42] We used an evaluation queue to manage the model developer's Docker image submissions.

Submission Processing Pipeline

To manage the Docker images submitted to the Synapse collaboration platform, we used a Common Workflow Language (CWL) pipeline, developed at Sage Bionetworks. The CWL pipeline monitors an evaluation queue on Synapse for new submissions, automatically downloading and running the Docker image when the submission is detected. Executed commands are isolated from network access by Docker containers run on UW servers.

UW On-premises Server Infrastructure

We installed this workflow pipeline in a UW Medicine environment running Docker v1.13.1. UW Research Information Technology uses CentOS 7 (Red Hat Linux) for their platforms. The OMOP data were stored in this environment and were completely isolated behind UW's firewalls. The workflow pipeline was configured to run up to four models in parallel. Each model had access to 70 GB of RAM, 4 vCPU, and 50 GB of SSD.

2.2.5 IRB Considerations

We received an Institutional Review Boards (IRB) non-human subjects research designation from the University of Washington Human Subjects Research Division to construct a dataset derived from all patient records from the EDW that had been converted to the OMOP v5.0 Common Data Model. (IRB number: STUDY00002532) Data were extracted by an honest broker, the UW Medicine Research IT data services team, and no patient identifiers were available to the research team. The model developer had no access to the UW data.

2.3 Results

2.3.1 Model Development, Submission, and Evaluation

For this demonstration, a model developer built a Dockerized mortality prediction model. The model developer was a graduate student from the University of Washington who did not have access to the UW OMOP clinical repository. This model was first tested on a synthetic dataset (SynPUF),[44] by the model developer to ensure that the model did not fail when accessing data, training, and making predictions. The model developer submitted the model as a Docker image to Synapse via a designated evaluation queue, where the Docker image was uploaded to a secure Docker Hub cloud storage service managed by Sage Bionetworks. The CWL pipeline at the UW secure environment detected this submission and pulled the image into the UW computing environment. Once in the secure environment, the pipeline verified, built, and ran the image through two stages: the training and inference stages. During the training stage, a model was trained and saved to the mounted volume "model" and during the inference stage, a "predictions.csv" file was written to the mounted volume "output" with mortality probability scores (between zero and one) for each patient in the Evaluation Dataset (**Figure 2.2**). Each stage had a mounted volume "scratch" available for storing intermediate files such as selected features (**Figure 2.2**). The model developer specified commands and dependencies (e.g., python packages) for the two stages in the Dockerfile, train.sh, and infer.sh. The Training and Evaluation Datasets were mounted to read-only volumes designated "train" and "infer" (**Figure 2.2**).

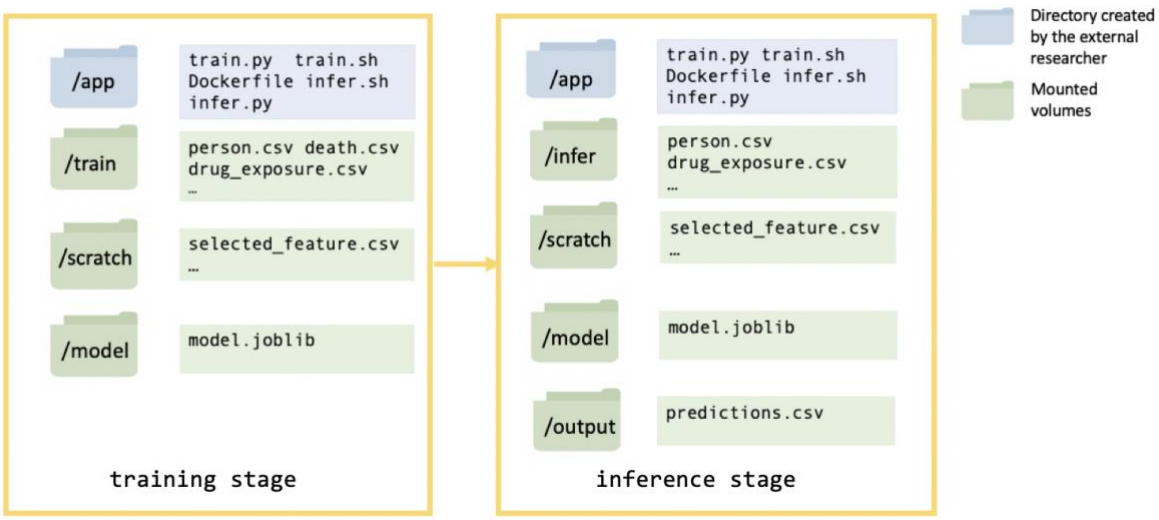


Figure 2.2 Docker container structure.

Showing the Docker container structure for the training stage and inference stage of running the Docker image.

After checking that the "predictions.csv" file had the proper format and included all the patients in the Evaluation Dataset, the pipeline generated an Area Under the Receiver Operator Curve (AUROC) score and returned this to the model developer through Synapse. If the Docker model failed, a UW staff member would evaluate the saved log files to assess potential errors. Filtered error messages were sent to the model developer for debugging purposes. See **Figure 2.3** for the full workflow diagram.

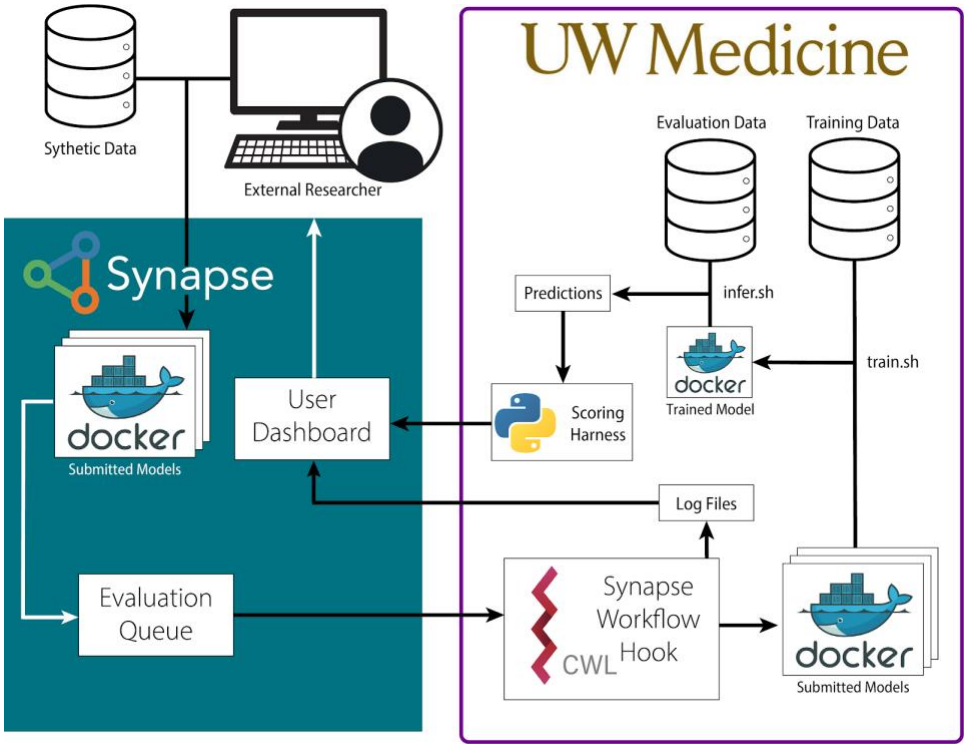


Figure 2.3 Workflow diagram.

Dockerized models were submitted to Synapse by a model developer to an evaluation queue. The Synapse Workflow Hook pulled in the submitted Docker image and built it inside the protected UW environment. The model trained on the available EHR data and then made inferences on the Evaluation Dataset patients, outputting a prediction file with mortality probability scores for each patient. The prediction file was compared to a gold standard benchmark. The model's performance was measured by AUROC and was returned to the model developer.

2.3.2 Model Developer's Perspective

The model developer built and submitted models, using three sets of features: (1) basic demographic information (age on the last visit date, gender, and race); (2) basic demographic information and binary indicators for five common chronic diseases (cancer, heart disease, type-II diabetes, chronic obstructive pulmonary disease, and stroke); [45] (3) the 1000 most common

concept_ids selected from the procedure_occurrence, condition_occurrence and drug_exposure domains in the OMOP dataset. For model 2, the developer used the OMOP vocabulary search engine, Athena, to identify 404 clinical condition-concept-ids associated with cancer, 76 condition-concept-ids with heart disease, 104 condition-concept-ids with type-II diabetes, 11 condition-concept-ids with chronic obstructive pulmonary disease and 153 condition-concept-ids with stroke (**Table 2.1**). Logistic regression was used on the three sets of features respectively. All model scripts are available here: https://github.com/yy6linda/Jamia_ehr_predictive_model.

	Training set	Evaluation set
Total number of patients	956,212	336,548
Number of patients with cancer	66,203 (6.9%)	42,195 (12.5%)
Number of patients with heart disease	31,352 (3.3%)	23,108 (6.9%)
Number of patients with type II diabetes	40,938 (4.3%)	28,234 (8.4%)
Number of patients with chronic obstructive pulmonary disease	13,777 (1.4%)	8,302 (2.5%)
Number of patients with stroke	5,216 (0.6%)	3,927 (1.2%)
Other Patients	834,591 (87.3%)	257,884 (76.6%)

Table 2.1 Patient profile for the Training/Evaluation Dataset.

The number of patients in the UW Medicine OMOP repository who have been diagnosed with cancer, heart disease, type II diabetes, or chronic obstructive pulmonary disease (features used in model 2)

Model Performance

The submitted models were evaluated at the University of Washington by comparing the output predictions of the models to the true 180-day mortality status of all the patients in the Evaluation

Dataset. The implementation of the logistic regression model, Model 1, using only demographic information, had an AUROC of 0.693. Model 2, using demographic information and 5 common chronic diseases, yielded an AUROC of 0.861. Model 3, using demographic information and the most common 1000 condition/drug/procedure('cdp') concepts, yielded an AUROC of 0.921. (**Figure 2.4**)

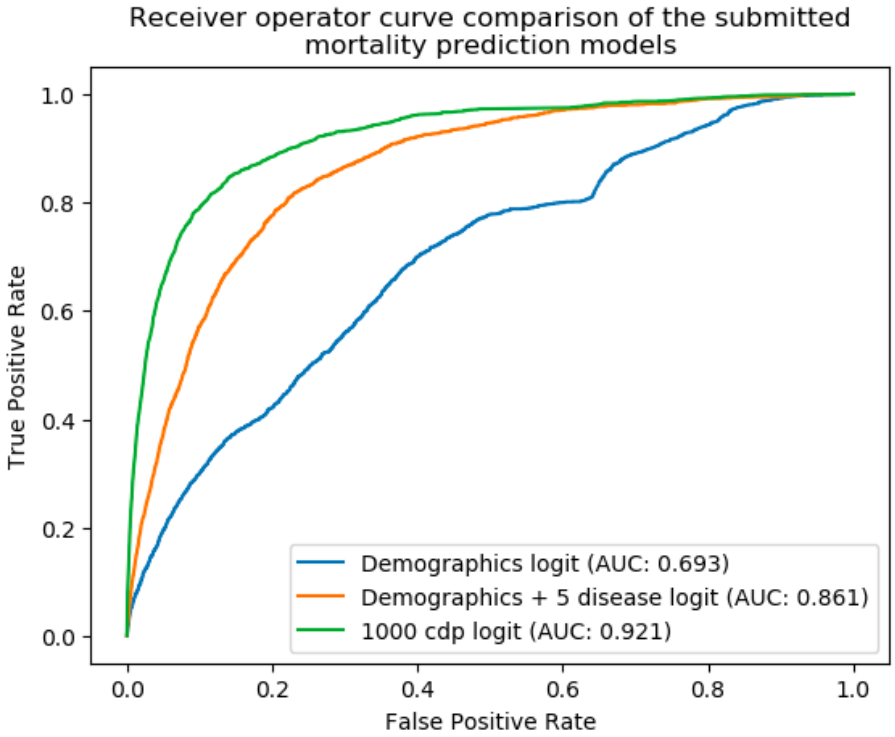


Figure 2.4 A comparison of the receiver operator curves for the three mortality prediction models.

2.3.3 Benchmarking the Capacity of Fixed Computing Resources for Model Running

We tested the capability of running models through the pipeline on increasingly large feature sets using two ML algorithms: logistic regression and neural networks. The models ran under fixed

computational resources: 70 GB of RAM and 4 vCPU (a quarter of the total available UW resources made available for this project). This tested the feasibility of running multiple (here, four) concurrent, high-performance models on UW infrastructure for a community challenge. 6934 of the features used for this scalability test were selected from condition_concept_ids that have more than 20 occurrences within 360 days from the last visit dates of patients in the Training Dataset. The two selected algorithms were applied to a subset of the features of 1,000, 2,000, 3,000, 4,000, 5,000 and 6,000 selected condition_concept_ids. We used the python sklearn package to build a logistic regression model and keras frameworks to build a three-layer neural network model (dimension 25 * 12 * 2). For both models, we trained and inferred using the 6 different feature set sizes. We report here the run times and max memory usage (**Figure 2.5**). While run times scale linearly with the number of features, maximum memory usage scales in a slightly super-linear fashion.

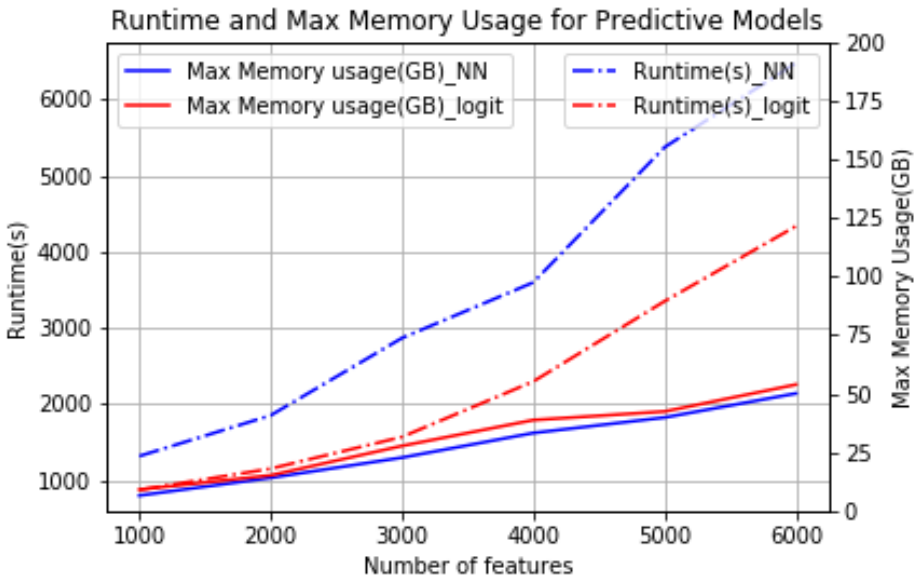


Figure 2.5 Runtime and Max Memory Usage for training predictive models in the scalability test.

2.4 Discussion

In this pilot project, we implemented the 'model to data' framework in the context of an institutional enterprise data warehouse and demonstrated how a model developer can develop clinical predictive models without having direct access to patient data. This 'model to data' evaluation platform relied on a mutually agreed upon set of expectations between the data-hosting institution and the model developer, including the use of (1) a common data model (here, OMOP), (2) a standard containerization platform (here, Docker), (3) predetermined input and output file formats, (4) standard evaluation metrics and scripts, and (5) a feedback exchange mechanism (here, Synapse). While we focused on the specific task of mortality status prediction in this pilot study, our platform would naturally be generalizable to other prediction questions or data models. A well-documented common data model (here OMOP v5.0) is essential to the successful operation of the 'model to data' approach. This framework, however, is not limited to the designated OMOP version, nor the OMOP CDM, and could be expanded to the PCORnet Common Data Model [46], I2B2 [47], or any other clinical data model. The focus of the 'model to data' framework is to deliver containerized algorithms to private data, of any standardized form, without exposing private data to model developers. With increased computational resources, our platform could scale up to handle submissions of multiple prediction models from multiple researchers. Our scalability tests show that complex models on wide feature sets can be trained and evaluated in this framework even with limited resources (70 GB RAM per submission). These resources, including more RAM, CPUs, and GPUs, could be expanded in a cloud environment and parallelized across multiple models. This scalability makes the 'model to data' approach particularly appealing in certain contexts as discussed in the following sections.

'Model to data' as a Mechanism to Standardize Sharing, Testing, and Evaluation of Clinical Prediction Models

Typically, most clinical prediction models have been developed and evaluated in isolation on site-specific or network-specific datasets, without additional validation on external health record data from other sites.[2] By implementing an evaluation platform for common clinical prediction problems, it would be possible to compare the performance of models implementing different algorithms on the same data and to test the generalizability of the same model across different sites, assuming those sites are using the same common data model. This framework also motivates researchers to containerize models for future reproduction. In the long term, we envision that the 'model to data' approach will enable researchers to test their predictive models on protected health data without worrying about the identification of patients and will inspire the ubiquitous use of Dockerized containers as a standard means to deliver and customize predictive models across institutions. The proposed pipeline is not dependent on the clinical question under investigation nor whether the study question involves all steps of model development (training, inference/test, open-ended prospective and non-performance-related evaluation).

'Model to data' as a Mechanism for Enabling Community Challenges

Community challenges are a successful research model where groups of researchers develop and apply their prediction models in response to a challenge question(s), for which the gold standard truth is known only to the challenge organizers. There have been a large number of successful biomedical community challenges including DREAM,[42], CAFA,[48,49] CAGI,[50], and CASP[51]. A key feature of some of these challenges is the prospective evaluation of prediction models, an often unmet need in clinical applications. The 'model to data' approach uniquely enables

such an evaluation of live EDW data. Based on our observations in this pilot study, the next stage is to scale up our platform to initiate an EHR mortality community challenge, in which participants from different backgrounds will join us in developing mortality prediction models.

Lessons Learned and Limitations

During the iterative process of model-training and feedback exchange with the model developer, we discovered issues that will have to be addressed in future implementations. (1) The model developer had multiple failed submissions due to discrepancies between the synthetic data and real data. Devoting effort toward improving the structural alignment between the synthetic data and the UW data will help alleviate this barrier. Correcting differences in data type, column names, and concept availability would allow model developers to catch common bugs early in the development process. (2) Providing manually filtered log files (filtered by UW staff) that are generated by the submitted models when running on UW data as an iterative process can be cumbersome. We propose that prior to running submitted models on the UW data, models should first be run on the synthetic dataset hosted in an identical remote environment which would allow the model developer to access all log files to support debugging. This would be intended to allow any major errors or bugs to be caught prior to the model running on the real data. (3) Inefficiently written prediction models and their containers burdened servers and system administrators. The root cause of this issue was the model developer's inability to estimate the computing resources (RAM, CPU) and time needed to run the submitted models. We can use the same synthetic data environment as solution 2 to estimate run time and RAM usage on the full dataset prior to running the model on the real data. (4) The model developer was unaware of the data distributions or even the terminologies for certain variables making feature engineering difficult. Making a data

dictionary with the more commonly used concept codes from the UW data available to the model developer will enable smarter feature engineering.

The presented pilot predictive models are relatively simple. However, the 'model to data' framework is also compatible with more complicated ML algorithms and feature engineering. Future researchers can Dockerise their complicated predictive models with more advanced feature engineering and send them through our pipeline as Docker images. Our pipeline can execute these Docker images on real data and return scores. Model interpretation, such as feature importance scores, is also feasible under this framework if the feature importance calculation is embedded in the Docker models and output to a designated directory in the Docker container. After checks for information leakage, the UW IT team would be able to share that information for the model developer to further interpret their models. However, the remote nature of the 'model to data' framework limits the opportunities for manual hyperparameter tuning which usually requires direct interaction with data. Hyperparameters are model parameters pre-defined before the models' training stages. (e.g., learning rate, number of layers in neural networks, etc.) However, automated methods to tune the hyperparameters work with the proposed pipeline. The emergence of AutoML as well as other algorithms including grid & random search, reinforcement learning, evolutionary algorithms, and Bayesian optimization allows hyperparameter optimization to be automated and efficient. [52]

In late 2019, we extended this work to enable the first EHR DREAM Challenge: Patient Mortality Prediction as a further demonstration for this pilot study. The patient mortality prediction challenge attracted 345 registered participants, coalescing into 25 independent teams, spread over 3

continents and 10 countries. The top-performing team achieved a final AUROC of 0.947 (95% CI 0.942, 0.951) and an area under the precision-recall curve (AUPRC) of 0.478 (95% CI 0.458, 0.499) on patients prospectively collected over a one-year observation of UW health system.

2.5 Conclusions

We demonstrate the potential impact of the 'model to data' framework to bring clinical predictive models to private data. This is achieved by operationalizing this framework to enable a model developer to build mortality prediction models using protected UW Medicine EHR data without gaining access to the dataset or the clinical environment. This work serves as a demonstration of the 'model to data' approach in a real-world clinical analytics environment. This enables future predictive analytics sandboxing activities, and the development of new clinical-predictive methods safely.

Chapter 3

IMPLEMENTING THE 'MODEL TO DATA' APPROACH IN A CROWDSOURCED BENCHMARKING CHALLENGE FOR COVID-19 OUTCOME PREDICTION

3.1 Introduction

First reported in December 2019, the novel Coronavirus SARS-CoV-2 caused a global pandemic resulting in strained hospital capacity and the deaths of 1,020,240 patients in the U.S. (as of July 17 2022) alone.[53] As the cumulative case counts rise, patient-level health data become a viable and crucial resource for researchers to understand disease patterns and design evidence-based interventions against the disease.[54] ML approaches applied to COVID-19 patient EHR data have shown value in outbreak prediction,[55,56] early screening,[9,57] contact tracing of infected patients,[58,59] health outcome prediction to improve diagnosis and treatment,[60] and to prioritize healthcare resources for patients who are at a higher risk for health complications.[8,12]

Patient data must be acknowledged to be private and sensitive, and restrictions are in place for the sharing of this data. These necessary restrictions hinder data accessibility for researchers, limiting their ability to develop models and externally validate them. In cases where researchers have

access to patient health data, models developed by isolated teams with no objective evaluation oversight can lead to self-assessment bias and overfit models.[57]

To overcome these challenges, we leveraged the 'model to data' approach as mentioned in Chapter 1 to launch the COVID-19 EHR DREAM Challenge in the hopes of supporting an informed response to the COVID-19 pandemic. We had previously demonstrated the feasibility and utility of this approach in the Patient Mortality EHR DREAM Challenge, leading to the unbiased assessment of ML models applied to EHRs to predict patient mortality, as mentioned in Chapter 2.[61,62] In our COVID-19 Challenge, we asked participants to address two clinically pressing questions:

- Question 1 (Q1) Of patients who received a test for COVID-19, who will test positive?
- Question 2 (Q2) Of patients who test positive for COVID-19 in an outpatient setting, who is at risk for hospitalization within 21 days?

The questions were motivated by the need to triage patients prior to widespread diagnostic and treatment capabilities. We evaluated models' performance and generalizability to patient subgroups stratified by age, gender, race, ethnicity, and time of COVID test.

3.2 Methods

3.2.1 Data

We ran the COVID-19 EHR DREAM challenge as a continuous benchmarking exercise where the datasets were updated every 2-5 weeks to incorporate new patients and update existing patients' clinical trajectories. We curated two sub-datasets (diagnostic Q1 challenge dataset and prognostic Q2 challenge dataset) separately for the two challenge questions for the purpose of model training

and evaluation. The Q1 challenge dataset - involving all patients who received a COVID-19 test by the date when each data update was conducted - was split into training and evaluation datasets using a temporal ordering where the EHR data for the most recent 20% of patients tested for COVID-19 were included in the evaluation dataset and the remaining 80% of patients were included in the training dataset. The Q2 challenge dataset included only patients who tested positive for COVID-19 at an outpatient setting by the date when the data update was conducted. These data were randomly split into training (70%) and evaluation datasets (30%) in order to maintain the same true positive prevalence (hospitalization within 21 days versus all patients who tested positive during an outpatient visit). Gold standards for the training data were provided for model training and gold standards for the evaluation dataset were hidden by the challenge organizers for scoring. For both the Q1 challenge dataset and Q2 challenge dataset, EHR data after each patient's COVID-19 test were removed. We incorporated new patients (i.e., patients who were not included in previous data updates) into the evaluation datasets and made sure no patient data existing in the training dataset of previous versions were included in later evaluation dataset versions. The Q1 challenge dataset has 6 versions, over 30 weeks and the Q2 challenge dataset has 4 versions over 18 weeks. (**Figure 3.1a**)

In contrast to the last version of the Q1 and Q2 challenge datasets which were both updated in November 2020, we gathered all the data that had accumulated by January 2021 and referred to this dataset as the "cumulative dataset". This represented 108,500 patients that tested for COVID-19, 4,980 that tested positive, 3,100 that tested positive during an outpatient visit and 170 that were hospitalized within 21 days after testing positive during that outpatient visit. We split the cumulative dataset such that 50% of patients who had most recently tested for COVID-19 were

incorporated into the cumulative evaluation dataset (patients who were tested between July 2020 and January 2021), and the other 50% were incorporated into the cumulative training dataset (patients tested between March 2020 and July 2020). The cumulative evaluation dataset was split evenly and prospectively into three sub-evaluation datasets based on the patients' COVID-19 measurement date to eval_1 (July 2020 - September 2020), eval_2 (September 2020 - November 2020) and eval_3 (November 2020 - January 2021). The cumulative dataset was used for post-challenge model analysis and training ensemble models.

We built an ensemble validation dataset to evaluate the performance of ensemble models. This dataset comprised 12,870 patients who had been tested for COVID-19 between January 2021 and March 2021, among which 278 had tested positive, 208 tested positive in outpatient settings and 16 were hospitalized within 21 days.

		cumulative training dataset	cumulative evaluation dataset	eval_1	eval_2	eval_3	ensemble validation dataset
	Time period of data collection	Mar 2020-July 2020	July 2020 - Jan 2021	July 2020 - Sep 2020	Sep 2020 - Nov 2020	Nov 2020 - Jan 2021	Jan 2021 - Mar 2021
	Age(%)						
	0-17	3.05	4.51	5.60	3.75	4.18	5.58
	18-24	9.06	9.91	9.82	11.50	8.40	6.93
	25-49	39.89	37.73	38.03	34.92	40.25	38.23
	50-64	25.21	24.34	24.27	25.00	23.76	23.9
	65-99	22.79	23.51	22.28	24.83	23.40	25.36
	Gender(%)						
	Female	51.47	51.07	50.76	51.68	50.77	52.06
	Male	48.47	48.90	49.21	48.31	49.17	47.9
	Other/Nan	0.06	0.03	0.03	0.01	0.06	0.04
	Race(%)						
	White	65.61	65.83	62.92	68.22	66.29	68.22
	Asian	9.02	10.41	10.01	10.55	10.67	11.21
	Black	9.65	8.73	9.03	8.09	8.81	8.17
	Other/Nan	15.71	15.03	17.77	13.13	14.24	12.4
	Covid-19 test(%)						
	Positive	5.16	4.02	3.93	2.45	5.68	2.16
	Negative	94.84	95.98	96.07	97.55	94.32	97.84
Q1	Number of patients tested for COVID-19	54600	53936	17932	18020	17984	12870
	Patients hospitalized within 21 days(%)						
	Positive	5.73	5.10				7.66
	Negative	94.27	94.90				92.34
Q2	Number of patients tested positive for COVID-19 in outpatient setting(count)	1554	1552	---	---	---	208

Table 3.1 Dataset Demographics decomposition.

Datasets include cumulative dataset, temporal-split cumulative evaluation dataset (eval_1, eval_2 and eval_3) and ensemble validation dataset.

3.2.2 Challenge Infrastructure and Workflow

We implemented the 'model to data' approach for the COVID-19 challenge to facilitate the delivery of participants' models to the sensitive challenge datasets. All work was reviewed and approved by the UW IRB and UW Medicine leadership. COVID-19 patient datasets were hosted on a UW Medicine IT provisioned secure server. Challenge participants never had direct access to the patient data; instead, they were required to build and submit Dockerized (containerized) models. A synthetic dataset was provided to the participants to help them become familiar with the format of the data and to aid in technical debugging. Models submitted by participants went through a validation process in an Amazon Web Service (AWS) cloud environment, running against synthetic data. Once validated, the models would be transferred to the UW environment to train and evaluate on real patient data. AUROC and AUPRC were two performance metrics we used to assess models. Synapse collaboration platform was used to receive submissions and host the challenge leaderboard. (**Figure 3.1b**)

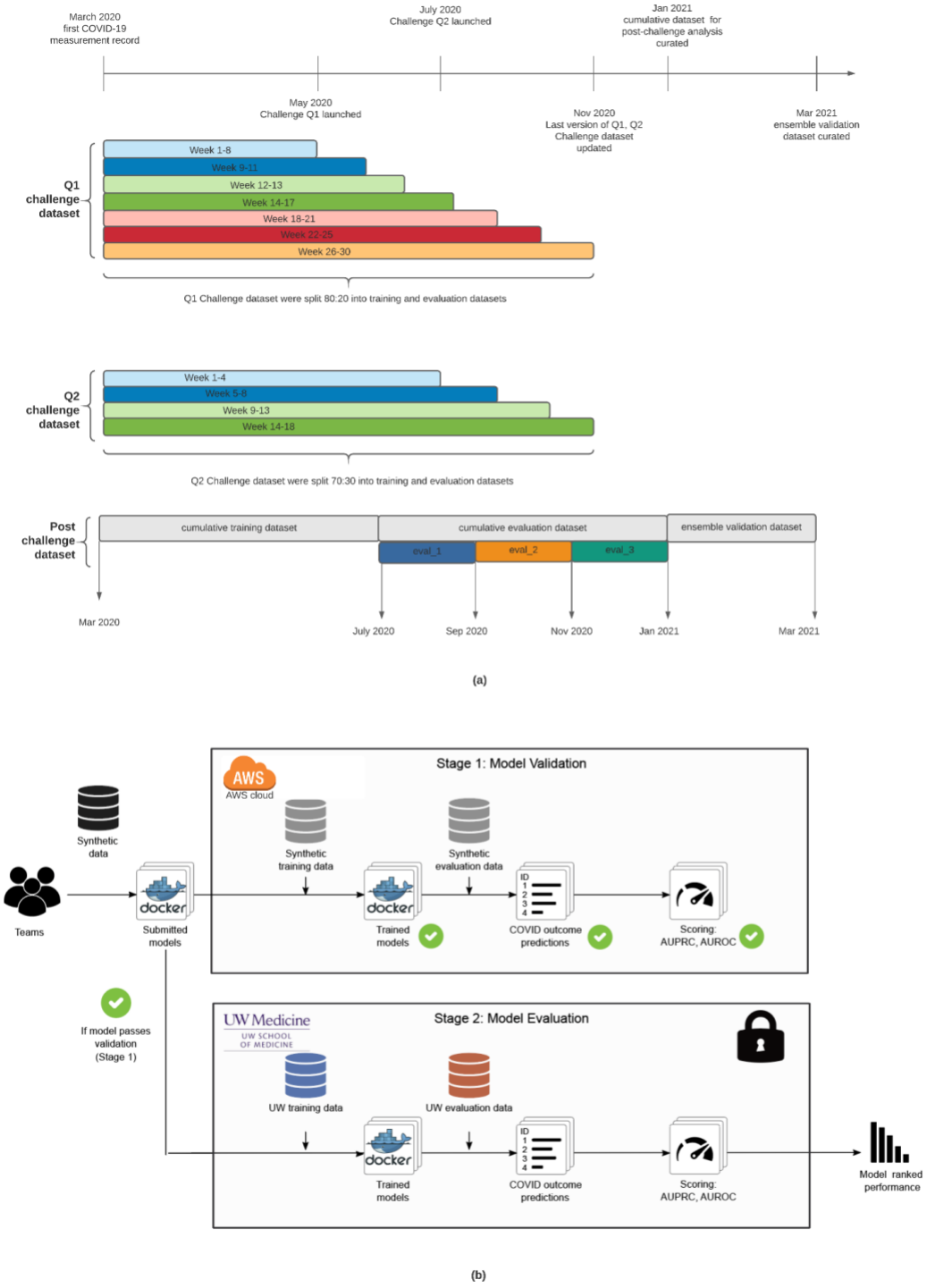


Figure 3.1 Visualization of the challenge timeline and infrastructure.

Diagram for the challenge timeline and patients' COVID-19 measurement date range in the datasets for Question 1 and Question 2. (a) The plot includes both challenge datasets (used in the

challenge) and cumulative dataset (used in post-challenge analysis). (b) Challenge operation workflow. When a participant submitted a model to the challenge platform, Synapse, the model underwent a validation procedure on an Amazon Web Service (AWS) cloud environment, in which the model was run against the synthetic dataset (Stage 1 model validation). If the model passed all the tests, the model was then pulled into a UW secure environment where it was trained and then applied to patient data (the holdout set from the full patient dataset) to generate predictions (Stage 2 model evaluation)

3.2.3 Post-challenge Model Analysis

To evaluate and compare models submitted to different versions of challenge datasets, we re-trained and evaluated Q1 and Q2 models separately on the cumulative dataset.

To evaluate the top 10 models from Q1 on different strata of the patient population (time of testing, age, gender, race, ethnicity), we trained these models on the cumulative training dataset and evaluated their performance on subsets in the cumulative evaluation dataset. For each stratum, we generated an AUROC score with a bootstrapped approach (distribution=1000; sample size 10000 with replacement). One-tailed t-tests were used to examine if the top 10 models' performances were consistently different.

Valid submissions to Q2 from 7 independent teams were also re-trained and evaluated on the cumulative training dataset. The analysis for Q2 models focuses on two aspects: (1) if the model was used to predict 21-day hospitalization for all patients who tested positive for COVID-19 regardless the type of visits, would it be more or less accurate than predictions made on patients who were at an outpatient visit when tested positive for COVID-19 and (2) if we limited the

amount of patients' pre-COVID-19 clinical history data available to model training, how would that impact model's performance? We generated AUROCs and bootstrapped distributions (n=1000; sample size of 1000 with replacement) using one-tailed t-tests to assess performance differences.

3.2.4 Ensemble Models

It has been shown that aggregating heterogeneous predictions from different models can improve individual model performance.[14,63] We trained ensemble models for Q1 and Q2 separately using the top individual models (mentioned above). Trained on the cumulative training dataset, each individual model outputs a confidence interval between 0 and 1 indicating the likelihood of a patient testing positive (Q1) or being hospitalized within 21 days (Q2). A logistic regression model with 10-fold cross-validation aggregated individual models' confidence intervals for the cumulative evaluation dataset to build an ensemble model. The ensemble validation dataset was used to assess ensemble models' performance. (**Figure 3.2**)

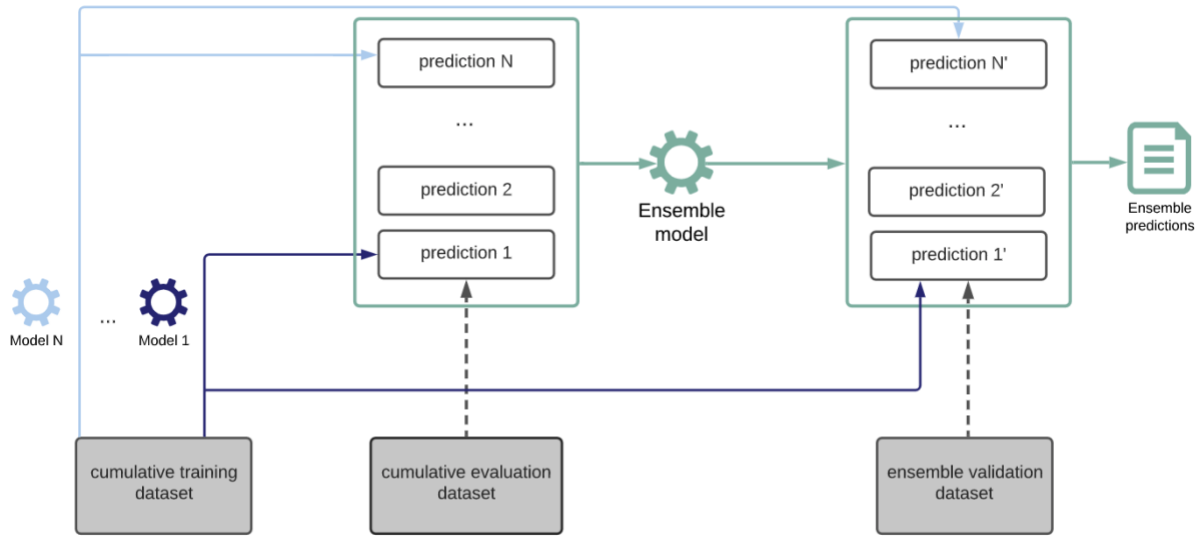


Figure 3.2 Ensemble model diagram.

3.3 Results

3.3.1 Challenge Summary

We hosted a continuously benchmarked community challenge to stimulate the development of ML methods for addressing clinical questions around COVID-19. This challenge had 482 registered participants with 90 teams successfully contributing submissions to at least one of the challenge questions. We had 369 valid submissions scored on Q1 challenge dataset and 232 for Q2. Over the course of this challenge, Q1 ran for 30 weeks, with the challenge dataset increasing from 9,100 patients to 89,600 patients through six data updates. Q2 ran for 18 weeks with the challenge dataset increasing from 1,700 patients to 2,200 patients through 4 data updates. For Q1, the AUROC and AUPRC of the best-performing model were 0.827 and 0.303 respectively, on the dataset version "Week 18-21". For Q2, the best AUROC and AUPRC were 0.982 and 0.897 respectively, for the dataset version "Week 1-4". However, these scores were observed in the first version of the Q2 challenge dataset which was small and the top team made multiple submissions in the first four weeks, presenting a high risk of overfitting. The best Q2 scores after the first challenge dataset version were an AUROC of 0.804 and AUPRC of 0.166 on dataset "Week 9-13". (**Figure 3.3**)

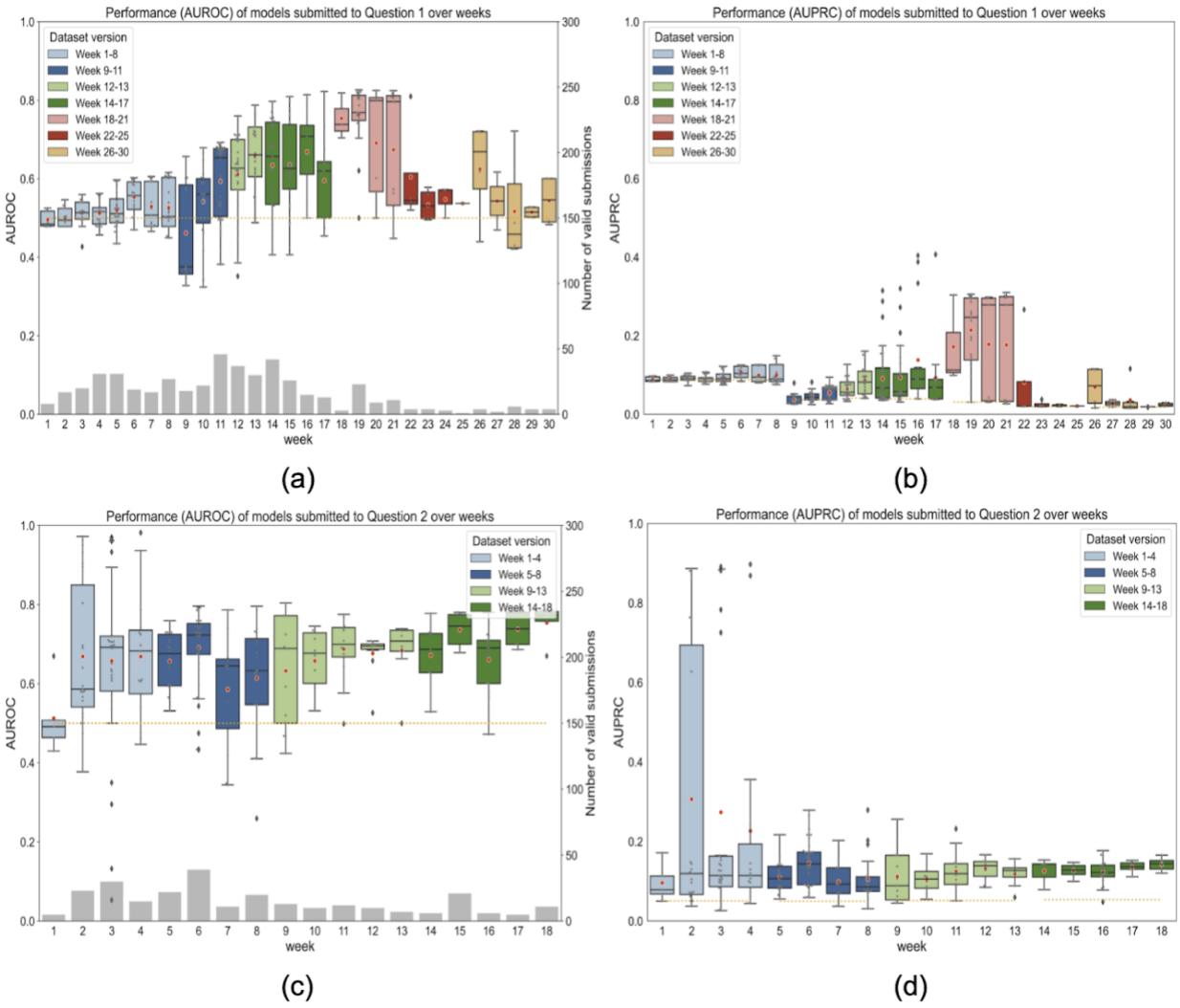


Figure 3.3 Performance of models submitted to challenge questions.

AUROC(a) and AUPRC(b) of models submitted to Question 1 every week. AUROC(c) and AUPRC(d) of models submitted to Question 2 every week. Grey bars in the plot a and c show the number of valid submissions to Question 1 and Question 2 weekly. Different colors represent different versions of challenge datasets. Datasets were named by the week in the challenge when they were in use. The yellow dash line is the performance baseline; for AUROC it is always 0.5 and for AUPRC it is the prevalence of positive patients in each evaluation dataset.

3.3.2 Post-challenge Analysis Results

The best performance for Q1 on the cumulative dataset - defined as data for patients tested for COVID-19 from March 2020 to January 2021 - was an AUROC of 0.776 (95% CI 0.775-0.777) and an AUPRC of 0.297. We then applied the top 10 re-trained models to longitudinally ordered subsets of the cumulative evaluation dataset (datasets eval_1, eval_2, and eval_3) to understand how models trained on previous patient data will generalize to future patients. The results showed for all the top 10 models, the performance on the eval_1 dataset was significantly better than on eval_2 and eval_3. (P<.001, **Figure 3.4**).

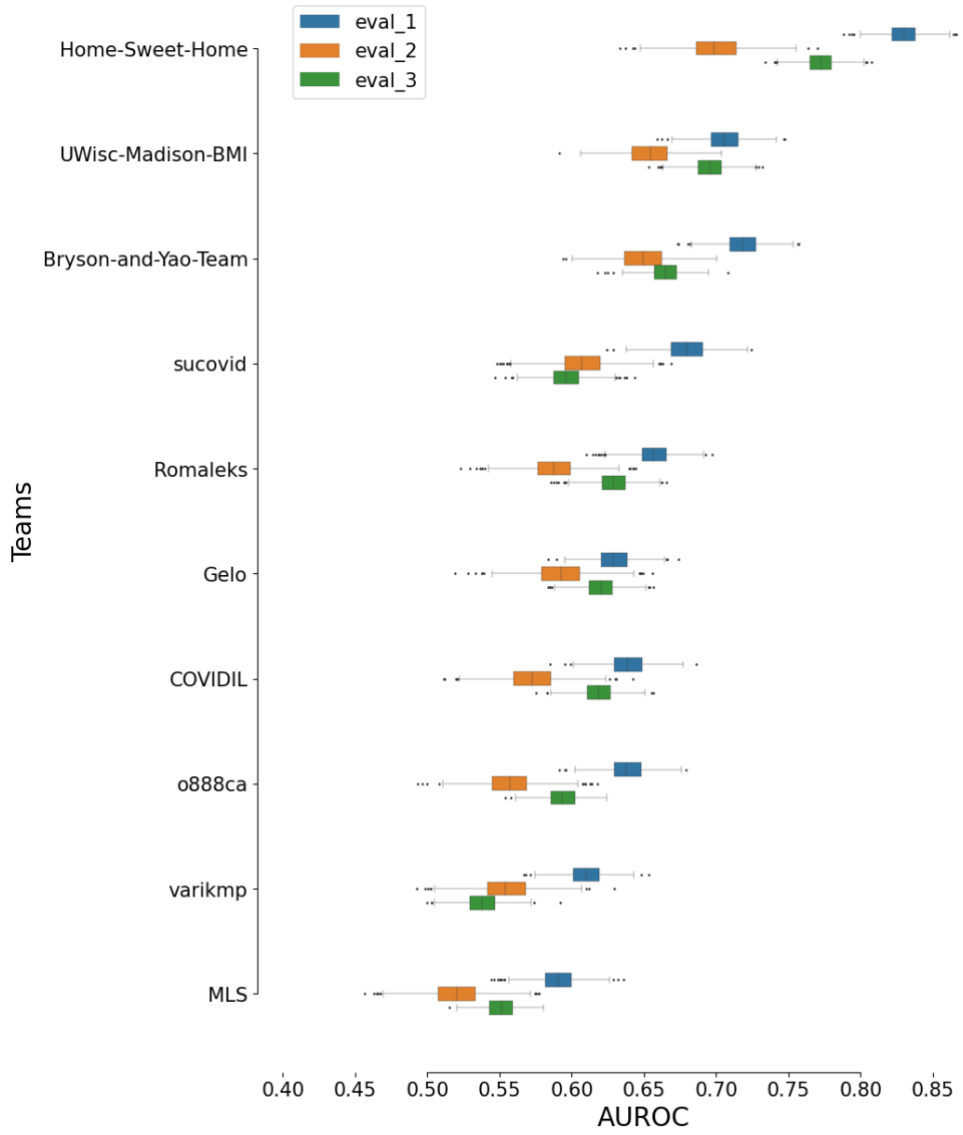


Figure 3.4 Performance of Question 1 models on prospective datasets.

Models were trained using patients tested for COVID-19 before July 2020 (training data, 50 % of patients) and made predictions on patients tested for COVID-19 after July 2020 (evaluation data 50 % of patients). Evaluation datasets were split into three sub-evaluation datasets based on the date when patients were tested.

We next explored how the performances of models might vary across different demographic traits. Splitting the cumulative evaluation dataset based on gender, 7 of the top 10 teams had a significantly better model performance on female subgroups compared to male subgroups (**Figure 3.5a**). When splitting by patient age, 8 of the top 10 teams had the lowest prediction performance on the youngest group (0-17) and 9 had the highest prediction performance on the 25-49 patient group ($P<.001$) (**Figure 3.5b**). The Pearson correlation coefficient of the top 10 models' average AUROC for each age subgroup to the subgroup dataset size was 0.849.

The top ten models did not show a consistent pattern of model performance on the sub-datasets split based on ethnicity ('hispanic or latino' and 'not hispanic or latino') or race ('Black', 'White', 'Asian' and 'Other'). Among the top three teams, the first team 'Home-Sweet-Home' outperformed the second ('UWisc-Madison-BMI') and third team ('Bryson-and-Yao-Team') in all race groups. However, the third team ('Bryson-and-Yao-Team') outperformed the second team ('UWisc-Madison-BMI') in the 'White' group (**Figure 3.5c-d**).

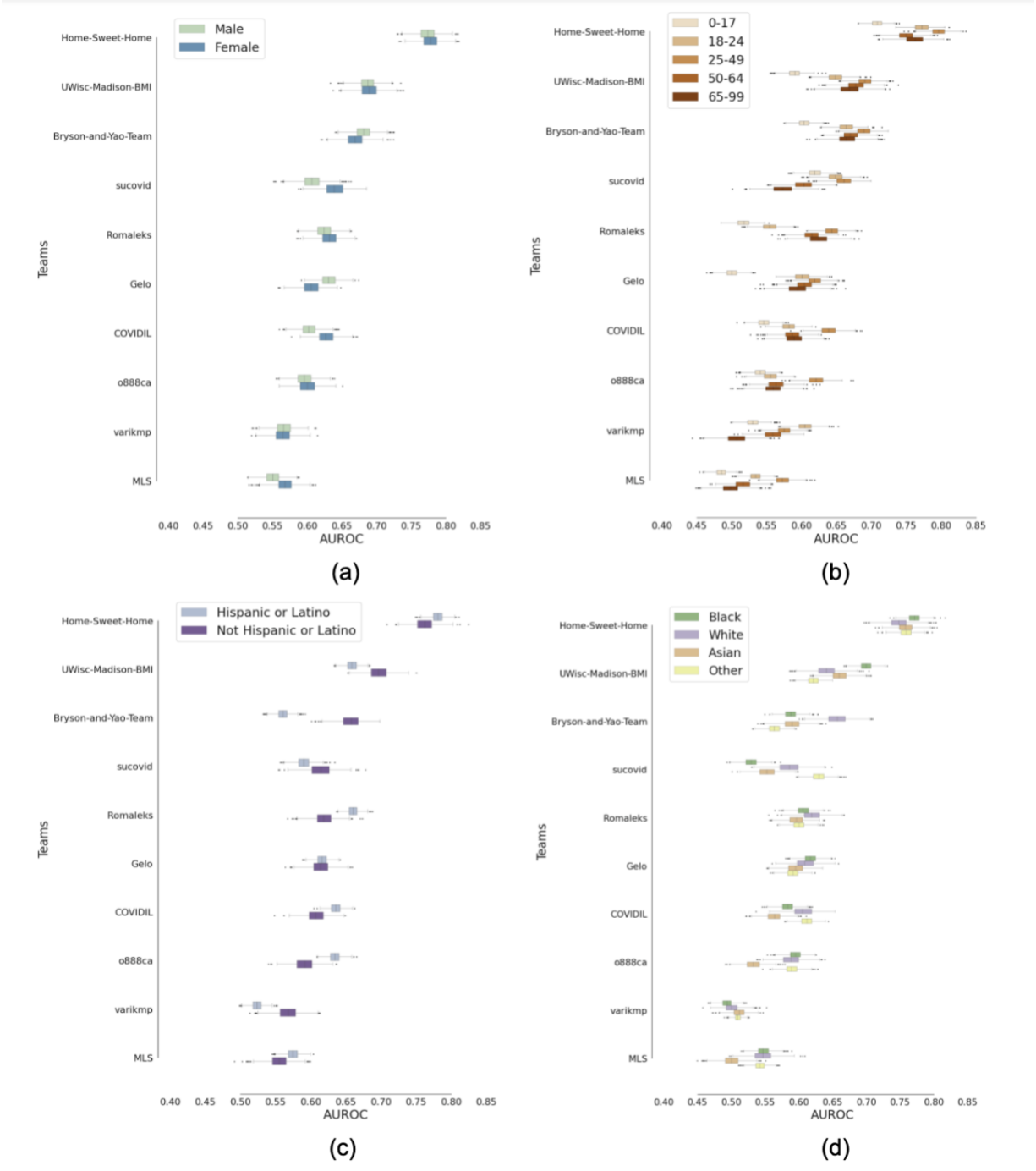


Figure 3.5 Performance of top 10 Question 1 models on subpopulation.

Subpopulation are stratified by gender(a), age(b), ethnicity(c) and race(d).

When Q2 models were re-trained and evaluated on the cumulative dataset, the best AUROC achieved was 0.796 (95%CI 0.794-0.798) and an AUPRC of 0.188. We asked whether the models could be generalized to patients who tested positive for COVID-19 during all visit types, not just outpatient settings. When the Q2 models were trained and applied on patients who tested positive in either inpatient or outpatient settings, 4 out of 7 models' performances dropped, and only 1 observed performance increased significantly (P<.001) compared to the prediction for only outpatient patients. This suggests that hospitalization prediction for patients who were tested for COVID-19 during non-outpatient visits, such as patients who were already in inpatient status for non-COVID-19 health conditions, were more difficult to predict correctly and patient data was noisier and clinically more ambiguous (**Figure 3.6a**).

We next tested whether truncating the length of the pre-COVID-19 EHR history made available to prediction models would affect model performance. We conducted this experiment by removing EHR records in 30-day increments up to 10 years before patients' COVID-19 measurement date in both training and evaluation datasets. We found that model performances did not consistently increase as more EHR clinical history was provided except for Ivanbrugere's model, which showed increasing performance as more clinical history became available up to 2-year data. (**Figure 3.6b**).

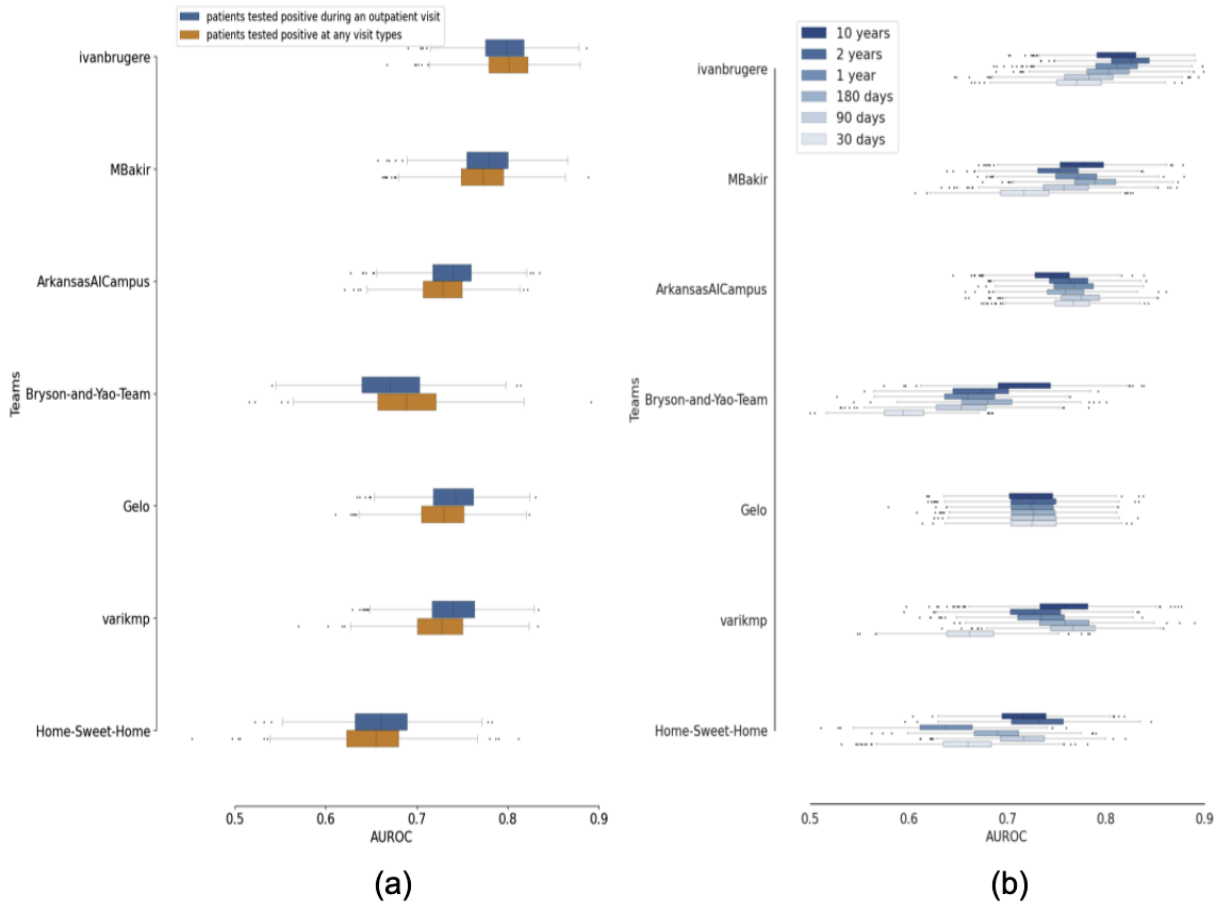


Figure 3.6 Model performance.

(a) Performance of Question 2 models for 21-day hospitalization prediction trained on all patients tested for COVID-19(orange) versus only patients tested for COVID-19 in outpatient settings (blue). (b) Performance of Question 2 models when trained on different lengths of EHR history prior to the COVID-19 test.

3.3.3 Top-performing Methods

We analyzed the top 3 teams' models for each question to shed light on the features and methodologies used by participants. The top teams used both data-driven approaches and pre-

selection based on clinical knowledge to select features. Boosting methods were the most popular top-performing algorithms. We asked physicians to review the top features selected by models to assess if the top features picked up by ML models were interpretable. Some features appeared to mechanistically relate to COVID-19 like loss of smell, cough, fever, leukocyte count for COVID diagnosis prediction, and oxygen saturation, asthma exacerbation diagnosis, acute renal failure, and abnormal coagulation studies for hospitalization prediction. Other features including serum CO2, hemoglobin/hematocrit, albumin, edema, etc. were selected by the models but did not have a known connection to COVID-19.

3.3.4 Ensemble Model Performance

We next developed an ensemble model to assess whether combining models could achieve better performance over any single model (see Methods). Applying the Q1 ensemble model combining the top 10 models to the ensemble validation dataset resulted in higher AUROC performance over any single model, with an AUROC of 0.714 (95%CI 0.713-0.715) and AUPRC of 0.106, compared to Q1 best individual model's AUROC of 0.699 (95%CI 0.698-0.700) and AUPRC of 0.112. When stratifying the ensemble validation dataset based on demographic profile, Q1 ensemble model outperformed the best individual model in 10 of the total 13 subgroups significantly (P<.001, **Figure 3.7)**. The Q2 ensemble model, which combined the top 7 teams, reached an AUROC of 0.740 (95%CI 0.739-0.742) and AUPRC of 0.286, compared to Q2 best individual model's AUROC of 0.772 (95%CI 0.771-0.774) and AUPRC of 0.193.

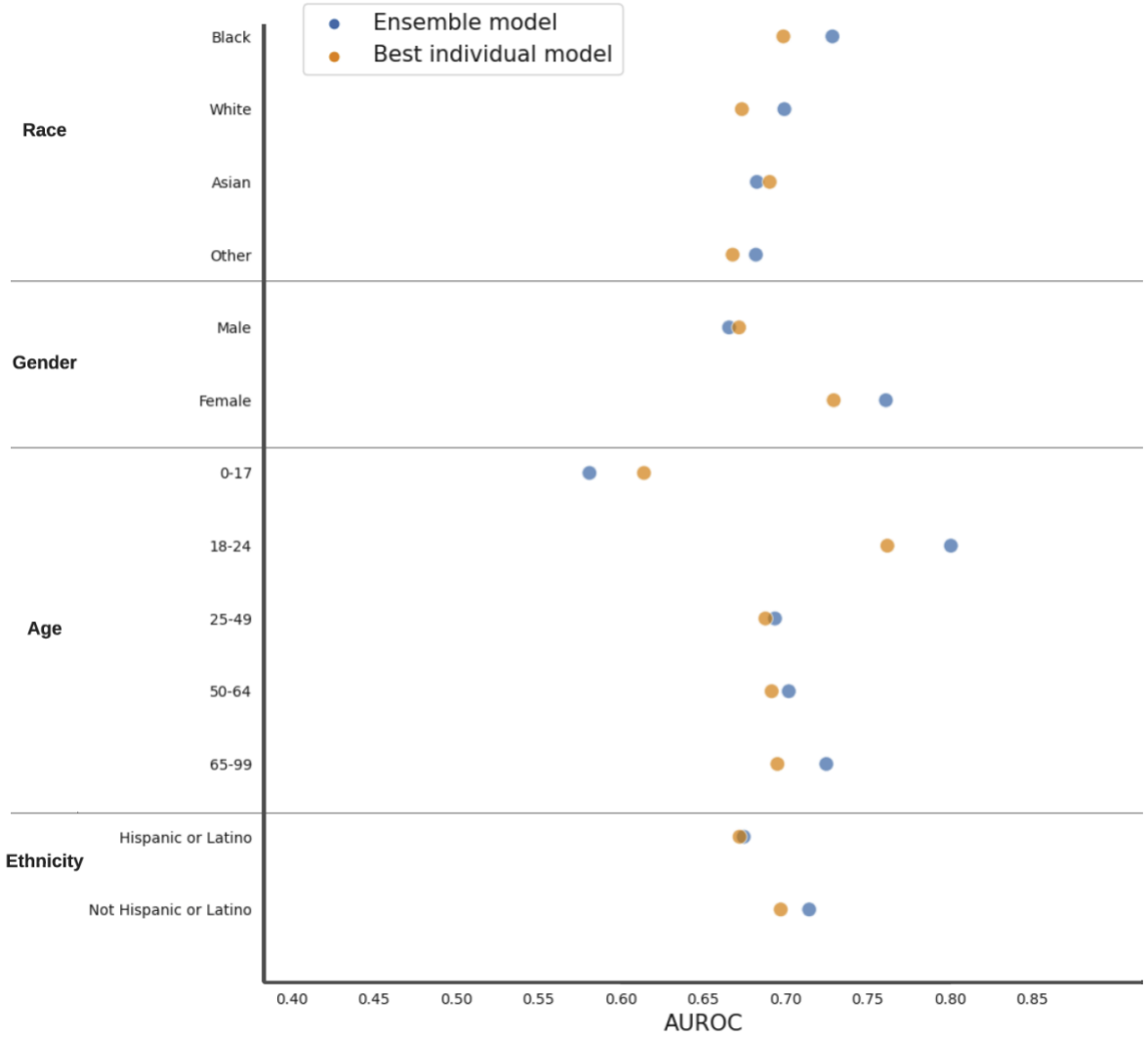


Figure 3.7 Model performance comparison.

Comparison between Question 1 ensemble model and Question 1 best individual model on demographics subgroups using ensemble validation dataset.

3.4 Discussion

In most common research cases, access to patient data is restricted to researchers affiliated with health institutions, and the turnaround time to get projects reviewed by IRBs can often lead to a delay between the data being available and the study being conducted. These delays and barriers result in missed opportunities for research and impact in time-critical scenarios like the COVID-19 pandemic. Our citizen science challenge provided a paradigm for sharing up-to-date patient data with those who otherwise would not have access. In this challenge, 482 participants from 7 countries were engaged in training predictive models that could aid clinical decisions and alleviate clinical burden as the COVID-19 pandemic overwhelmed healthcare institutions. After this study, we continued operating the platform to support the evaluation of methods on challenge datasets for a year till 2022.

We launched two questions in this challenge for predicting COVID-19 test results and hospitalization to assess the performance of methods, replicate results from other sites and identify key features for prediction. These two questions were most suitable for the beginning of the pandemic when test supplies and hospital resources needed to be prioritized. With this continuous benchmarking platform constructed, computational resources provisioned, and hundreds of data scientists engaged, we can point these resources at the next urgent questions, such as predictions of COVID-19 mortality risk, vaccine efficacy and the long-term effects of COVID-19.

We improved this EHR DREAM challenge from previously fixed datasets and time-limited submission quota to datasets that were updated and interrogated over time. The successful operation of the continuous benchmarking challenge demonstrated the flexibility and scalability

of the 'model to data' approach. This approach proved to have three benefits: (1) it protected the patient data while enabling model development on private data, (2) it forced model developers to standardize their models, enabling model transferability and reproducibility for rigorous evaluation, and (3) it enabled an unbiased third-party to evaluate these standardized models on previously unseen data.

We saw performance degradation on temporally evolving datasets indicating limitations in the models' generalizability on prospective datasets. However, this performance degradation was expected given the rapid changes to the challenge dataset caused by ever-changing clinical practice and variance in age distribution, and COVID-19 positive prevalence. We observed that model performance for the 25-49 age group was the best and the 0-17 age group was the worst among all age groups. This was consistent with the number of patients in the two age groups; 25-49 was the largest group and 0-17 was the smallest. However, with white patients making up the majority of the dataset, the model performance on the white group was not always better than the other race groups, indicating COVID-19 diagnosis prediction on white patients was difficult even with more training samples for race. We also identified that top teams could have inferior model performance on some sub-populations compared to other teams who ranked lower. This could be ameliorated with a model ensembled based on the strength of each team to maximize prediction accuracy. The Q1 ensemble model outperformed the best individual model in most demographic subgroups.

The high-performing models we received in the challenge indicate potential clinical utility. To achieve that, we will need to further test the generalizability of those models in a larger and multi-

site dataset, (e.g., National COVID Cohort Collaborative (N3C) data[64]) and incorporate the models into a live clinical workflow setting for providing clinical decision support.

3.5 Limitations

The continuous benchmarking challenge has also led to the identification of several "challenges". Data quality was difficult to maintain with regular updates. Data duplicates existed in some versions of the challenge dataset. In addition, compared to conventional challenges that have a fixed timeframe, models were more at risk of overfitting to the data as the number of allowed submissions increased over time. We also noticed that challenge models may be biased against one or more subpopulations, and it is not always the case that this is caused by the training data size. It could be caused by cultural and behavioral differences and requires further investigation.

3.6 Conclusions

We succeeded in operating a continuous benchmarking challenge to share up-to-date COVID-19 EHR patient data with a worldwide data science community. The benchmarking challenge provided an unbiased evaluation of models submitted by participants. Top models achieved high accuracy in predicting COVID-diagnosis results and hospitalization, indicating a potential for clinical implementation. Across submitted models, we observed discrepancies of performance in this temporally evolving dataset and among demographic sub-populations (gender, age, race, and ethnicity), indicating the existence of potential bias in ML approaches, which warrants attention prior to implementation of such models in clinical practice.

Chapter 4

ENABLING CLINICAL NOTES SHARING AND DE-IDENTIFICATION THROUGH THE NLP SANDBOX

4.1 Introduction

4.1.1 Challenges of Clinical Notes Access

Clinical notes contain rich information about a patient's medical history (symptoms, diagnoses, treatment plans, etc.) and may reveal important knowledge about a patient's lifestyle and disease progression. Compared to structured EHR data, clinical notes contain scattered and unstructured information, leading to difficulties in data collection, interpretation, and analysis.[65,66] NLP methods have been widely applied to clinical notes for information extraction and interpretation.[67–69] The implementation of NLP models in clinical settings often requires thorough evaluations of model performance using a large volume of clinical notes. However, due to the presence of PHI in clinical data, access to patient data is restricted, and most models are developed and evaluated using data from limited sources, such as single institutions. This limited access to data poses challenges for the generalizability of NLP models, as data from single institutions are not always representative of other health systems. [70,71]

There are several methods that may help address the lack of clinical notes for model evaluation. Advances in synthetic clinical text generation have partially contributed to larger datasets for model development and testing [72–76]. However, in accordance with the privacy-utility trade-off, synthetically generated text often lacks real-world utility. Publicly available de-identified datasets like MIMIC-III [34] and i2b2 [77]-[78] also help alleviate the shortage of high-utility clinical data, but they are not updated regularly, and publishing de-identified sets of clinical notes poses a risk for patient re-identification.[19,79]

4.1.2 Challenges of NLP model evaluation

Challenges faced when evaluating clinical NLP models include self-assessment bias [80] and the potential lack of model generalizability. Multi-site collaborations and data sharing can help address these challenges, but it often takes time to gain administrative and governance approval, leading to research delays. Even if multi-site collaborations are approved, the data can rarely be reused by other model developers due to privacy concerns. NLP model benchmarking frameworks like the ERASER [81], i2b2, GLUE [82], semEval [83], and kaggle [84] challenges have attempted to address this issue by releasing standardized note set, sets of evaluation metrics, and baseline models for comparison. However, these frameworks still use de-identified or non-sensitive datasets, and developers need to run their models on provided test datasets and upload the results to the platform for scoring, rather than uploading the actual models to the platform. As a result, the models are not automatically shared and immediately ready to be applied to new datasets.

4.1.3 'Model to data' Approach

As mentioned in previous chapters, The 'model to data' framework is a privacy-protected approach designed to lower the barrier to private data utilization. [21] Under this framework, model

developers train and evaluate models using private data, but without direct access to the data. The only information returned to participants is model performance scores; no private information leaves the data hosts. This approach has already been used to support several crowdsourced DREAM challenges, including challenges focused on EHR data, to enable the utilization of private clinical data in the form of structured tabular data and radiology images. [14,61,62,85]

We leveraged the 'model to data' approach to develop the NLP Sandbox (https://nlpsandbox.io) as a solution to the two problems outlined above: a lack of broadly shared, high utility clinical data, and a lack of NLP model sharing and generalizability. The NLP Sandbox is an NLP model evaluation system that enables federated evaluation and leverages knowledge and resources from multiple stakeholders. We identified three main stakeholders in the Sandbox design: (1) health institutions that contribute their datasets to the NLP Sandbox; (2) model developers who develop and submit NLP-based models to the sandbox environment; and (3) independent validation sites, which are external to the main NLP Sandbox infrastructure but contribute to model validation and implementation. In this project, we aimed to evaluate the utility of the NLP Sandbox for comparing clinical text de-identification models. Our goal is to inspire collaboration between the stakeholders and enable the utilization of private clinical notes by the broader data science community.

4.2 Materials and Methods

4.2.1 Datasets

The NLP Sandbox can be scaled to include an arbitrary number of distributed datasets. For evaluation purposes we included three datasets: (1) The 2014 i2b2 test dataset, (2) Mayo Clinic synthetic clinical notes, and 3) Medical College of Wisconsin (MCW) clinical notes (**Table 4.1**). The 2014 i2b2 test dataset includes 514 de-identified clinical notes. The MCW dataset includes 433 clinical notes, all of which are progress notes. The Mayo Clinic dataset includes 148 notes with 6 note types (progress notes, plan of care, telephone encounter notes, discharge summary, consultation notes, and emergency department notes). This dataset is synthetically modified by replacing the "name" PHI detected by the MCW de-identification tool (https://bitbucket.org/MCW_BMI/notes-deidentification) with synthetic surrogates generated from curated common English names and replacing the "date" PHI with shifted dates that mimic the original format.

In addition to the accessible datasets, UW is an independent validation site, meaning that NLP models can be evaluated by personnel at UW, but its data are otherwise private. Overall, the dataset includes 956 clinical notes with 10 note types (admit, discharge, emergency department, nursing, pain management, progress, psychiatry, radiology, social work, and surgery notes), generated from 2018 to 2019. [86] The MCW and UW datasets are both fully identified. A standardized data schema is used to annotate each of the four datasets with the following five PHI types: Date, ID, Person name, Location, and Contact.

4.2.2 Evaluation Standards

We conduct separate evaluations of the five PHI categories in the NLP Sandbox. The PHI gold standards have been converted to our data schema in JSON format. For example, a person name PHI gold standard is annotated as {“TextPersonNameAnnotations”:[{“noteId”: “110-01”,“start”: 60, “length”: 11 , “text”:”David Smith” }]}. We use the start position and the context of the PHI to conduct evaluation on two levels: (1) instance level, where the evaluation is based on each complete PHI entity (e.g., “David Smith”) and the model scores only if it captures both the text and location of the entity correctly, and (2) token level, where the PHI entity is broken into tokens (e.g., “David” and “Smith” for “David Smith”) and each token is evaluated independently. We report three evaluation metrics for each PHI category, both at the instance and token level: (1) Precision, (2) Recall, and (3) F1.

4.2.3 Containerized NLP Model

We demonstrate application of a model template to reduce the barriers for developers to adapt their NLP models to the NLP Sandbox schema. As mentioned in previous chapters, Docker makes version control, model dependencies, and environment variables easier to manage and is thus conducive to model standardization and sharing. The model template leverages the OpenAPI generator and Docker-compose, which can run multi-container applications. The model template also generates a Docker image for submission to the NLP Sandbox and provides step-by-step instructions for developers to build the Docker images and launch the web-based user interface (UI) for model testing.

4.2.4 Evaluation Workflow

After incorporating their NLP models into the model template, model developers can submit the Docker image to Synapse, an open-source software platform developed by Sage Bionetworks that supports collaborative data science and crowdsourced challenges. [43,87–89] Five separate submission queues corresponding to the five PHI types are provided to developers for model performance evaluation. Once the model is submitted to one evaluation queue, the NLP Sandbox orchestrator detects the submission, automatically downloads, and runs the Docker image.

The model is initially run in the test environment, where the annotated i2b2 test dataset is stored. The NLP Sandbox orchestrator sends a request to the i2b2 data node to retrieve clinical notes and annotation gold standards. Each data node consists of a Docker container with a MongoDB-backed database for storing and managing clinical notes and annotation gold standards. The security and stability of each data node are enhanced by Nginx[90], which is used as a reverse proxy and load balancer. Next, the NLP Sandbox orchestrator sends a request to the model and retrieves the annotation response. The NLP Sandbox orchestrator then evaluates the model annotation output against the gold standard annotation and generates scoring metrics. Finally, the orchestrator sends the scores to the leaderboard table on Synapse (**Figure 4.1**, Step 1). Once the model is successfully scored in the test environment, the orchestrator in the Mayo Clinic and MCW environments pulls the Docker images simultaneously to run and score the model on each dataset. The Mayo Clinic and MCW environments have similar submission infrastructure configurations as the test environment. Once model scoring is complete, the orchestrator sends the scores to the same leaderboard where the i2b2 scores are presented (**Figure 4.1**, Step 2).

Data Source		I2b2 test dataset	The Medical College of Wisconsin	The Mayo Clinic	The University of Washington
Total no. of notes		514	433	148	956
Total no. of annotations		10519	4703	2839	36843
No. of annotations(%)	Date	4980 (47.34%)	3484 (74.08%)	1274 (44.87%)	21351 (56.42%)
	Person name	2883 (27.41%)	809 (17.20%)	1565 (55.13%)	8141 (21.51%)
	Contact	218 (2.07%)	89 (1.89%)	/	1297 (3.43%)
	Id	625 (5.94%)	70 (1.49%)	/	779 (2.06%)
	Location	1813 (17.24%)	251 (5.34%)	/	6275 (16.58%)
No. of notes(%)	Date	514 (100.00%)	250 (57.74%)	148 (100.00%)	923 (96.55%)
	Person name	508 (98.83%)	289 (66.74%)	148 (100.00%)	773 (80.86%)
	Contact	165 (32.10%)	44 (10.16%)	/	409 (42.78%)
	Id	366 (71.21%)	65 (15.01%)	/	215 (22.49%)
	Location	420 (81.71%)	98 (22.64%)	/	740 (77.41%)

Table 4.1 NLP Sandbox Datasets.

4.2.5 Infrastructure Settings

The i2b2 and Mayo environments are both hosted on a cloud service managed by Sage Bionetworks. The MCW environment is maintained in an on-premises server behind their firewall. Since the UW dataset is not part of the NLP Sandbox infrastructure, it is used to demonstrate how models built and evaluated in the NLP Sandbox environment can be tested on other datasets outside of the NLP Sandbox. The UW data are stored in an on-premises server behind the UW firewall. Models are allotted 2 hours of total compute time, 7 GB of RAM, 4 CPU cores during each annotation and scoring process inside the NLP Sandbox and UW environment.

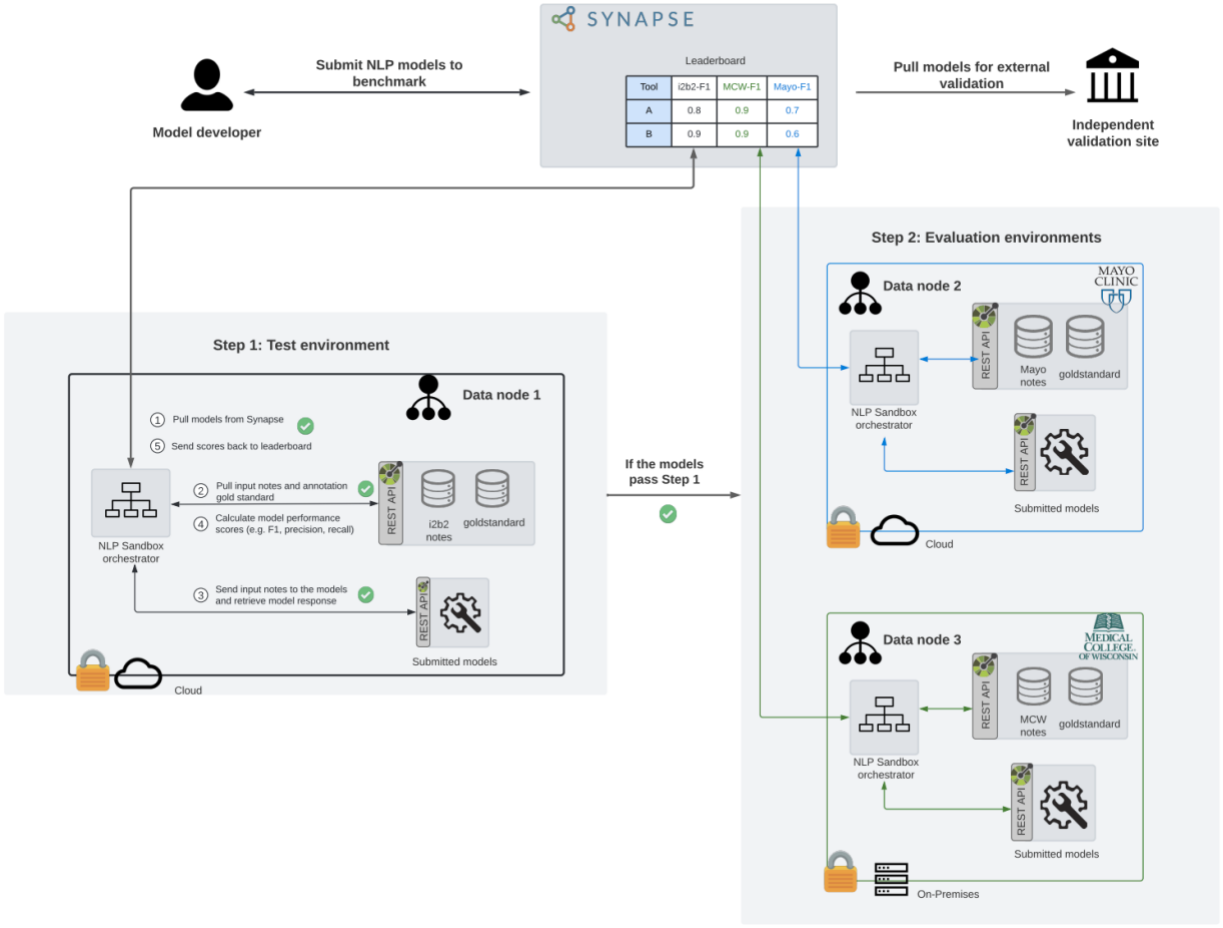


Figure 4.1 NLP Sandbox evaluation workflow.

A submission made to Synapse is first evaluated in the test environment (Step 1) on the i2b2 dataset. If the model passes Step 1, it will be sent to the Mayo and MCW environments for evaluation (Step 2). Only the scores are sent back to the Synapse leaderboard and made available to the developer.

4.2.6 Experiment Design

To test the usefulness of the NLP Sandbox, we conducted three experiments: (1) User experience from a model developer, where we invited a developer of Philter [24] to adapt their Philter Python algorithm into the NLP Sandbox framework for a user experience test. The Philter developer was not involved in developing the NLP Sandbox. The purpose of this test was to assess the framework and identify limitations in an unbiased fashion; (2) Multi-site evaluation for an open-sourced model, where we adapted NeuroNER [25] to the NLP Sandbox and demonstrated how model benchmarking could be accomplished; and (3) Model validation using data from an independent clinical site, where we demonstrated how models submitted to the NLP Sandbox could be further evaluated on new datasets without the involvement of the model developer.

External user experience test

In the first study, we aimed to evaluate the usability of the NLP Sandbox from the perspective of an external model developer. To simulate the experience of a typical user, the developer first created an account with Synapse and became a certified user. The developer was then asked to clone the NLP Sandbox model template from GitHub. The template contained separate modules corresponding to each PHI category. In each file, the data structure and format of input notes and annotation outputs were specified. To test the model, the developer created a Docker image following instructions in the model template GitHub page and tested the model locally through a private UI webpage that was automatically configured by OpenAPI generator during the Docker build phase.

To submit the model to the NLP Sandbox, the developer tagged the Docker image with a unique version number and pushed the tagged image to their private Docker repository on Synapse. The developer then submitted the tagged image to the five different submission queues for each PHI category. For each combination of PHI type and test set, the developer received an email notification with the F1, precision, and recall scores. If the submission failed in the test environment, the developer received an email notification with a link to the error logs generated from the i2b2 test environment. For more detailed troubleshooting, the developer worked with an NLP Sandbox developer to review the model's performance on the i2b2 dataset. Because the i2b2 dataset is public and fully de-identified, an NLP Sandbox developer returned false positive and false negative annotation results to the model developer for further model improvement. To ensure privacy protection and unbiased evaluation, no log files, false positives, or false negatives from the MCW or Mayo datasets were provided to the Philter developer. The developer went through several iterations of this process to improve the submission. Finally, the Philter developer was asked to provide an assessment of the model template adaptation and submission process.

Multi-site evaluation for an open-sourced model

In the second study, we wanted to explore the feasibility of benchmarking within the current NLP Sandbox framework. We currently do not make any identified datasets available for model training to model developers. However, if the developer has access to a private dataset, they can train their model using this dataset and submit the pre-trained model to the NLP Sandbox for model evaluation. We used NeuroNER, a neural network-based de-identification system, to demonstrate how a pre-trained ML model could be adapted to the NLP Sandbox.

To conduct the experiment, we first incorporated the NeuroNER package and i2b2 pre-trained model file into the NLP Sandbox model template. Then, we submitted the Docker image to Synapse and evaluated NeuroNER using the i2b2 test, Mayo, and MCW datasets. Following the evaluation process, the model performance scores for the three sites were returned to the Synapse leaderboard.

Model validation using data from an independent site

To mimic the further evaluation of models submitted to the NLP Sandbox at external sites, we used a private dataset from UW and an on-premises server behind the UW firewall to conduct an independent validation of submitted models. We first translated the UW data into the NLP Sandbox data schema and conducted a thorough data quality check to ensure that there were no annotation duplicates and that the start and end position of each annotation in the gold standard matched those in the original notes. We launched the data node service on the UW server to host the private clinical notes. We then used Docker commands to pull the submitted NeuroNER and Philter Docker images from Synapse and run them in the UW environment. We created a Python package called NLPsandboxclient to pull clinical notes from the data node, process them using the submitted models, and evaluate model performance by generating F1, precision, and recall scores for each PHI category.

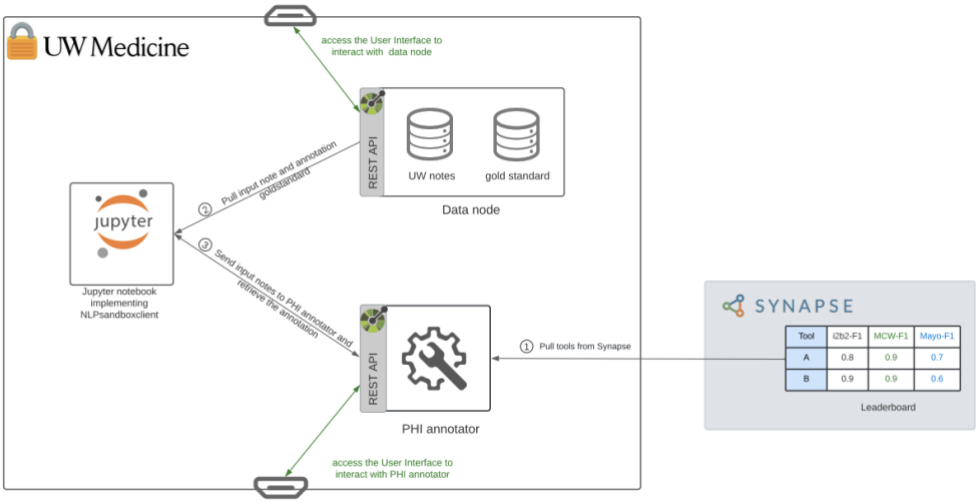


Figure 4.2 Infrastructure of the independent validation site (UW) environment.

The server hosting UW clinical notes is behind the UW firewall. For each model validation test, we pulled a Docker image submission from Synapse and launched it inside the UW environment as a PHI annotator. NLPsandboxclient was imported into a Jupyter notebook to pull notes from the data node, interact with the PHI annotator, and conduct the evaluation. Both the PHI annotator and data node provided UI webpages for interaction.

4.3 Results

4.3.1 Model Developer User Experience

The developer reported no difficulties in setting up a user account with Synapse, but they mentioned that cloning and modifying the de-identification model GitHub template required both command-line experience and knowledge of the Git version control tool. The developer worked with an NLP Sandbox developer to complete both the initial model integration and Docker image submission successfully. During model integration, the developer reported that the UI webpage

was helpful for identifying code integration errors and assessing whether each PHI annotator was functioning as expected. Although detailed stack trace messages did not appear in the UI, high-level error codes were displayed and helped guide debugging efforts. The developer was also able to generate print statements to retrieve more detailed debugging information.

In the original version of Philter, de-identification rules were applied sequentially as defined in a JSON configuration file. The Philter package utilizes a mix of rules including regular expressions, exclude regular expressions, include lists, exclude lists, and advanced pattern matching using previously identified PHI. To adapt the rule-based Philter algorithm into the NLP Sandbox model schema, the developer initially incorporated only exclude regular expressions and lists directly into the Python PHI annotation files. Because the NLP Sandbox evaluates the model based on its performance in each PHI category independently, the developer was required to apply different de-identification rules in separate annotation modules for each PHI type. This resulted in low precision and recall scores compared to previously reported scores on the 2014 i2b2 test corpus. Previously reported Philter recall scores on the 2014 i2b2 corpus were above 0.96, and precision scores were above 0.78 across all PHI categories. In contrast, recall scores evaluated within the NLP Sandbox environment ranged from 0.23-0.99 on the 2014 i2b2 corpus, and precision ranged from 0.27-0.79 for different PHI categories.

To rescue recall and precision as much as possible, the developer imported a modified version of the Philter package into the NLP Sandbox model template to leverage additional de-identification rules. However, these changes did not significantly improve performance, indicating that poor performance was not necessarily caused by unused algorithmic rules, but was instead caused by

differences in evaluation standards. In previously published evaluations of Philter, the evaluation was category-agnostic, and overlapping identification of PHI by different rules did not impact precision as long as the PHI was obscured. In the current NLP Sandbox evaluation, the evaluation is conducted separately for each PHI category. If a token is categorized as a PHI type different from the gold standard, this could negatively impact performance.

Overall, the developer recommended that clearer and more thorough documentation of the development process be posted to the GitHub README, including documentation of how to incorporate algorithms wholesale into the model schema. Additionally, automatically returning i2b2 false positives and false negatives to developers may help them diagnose issues with model integration.

4.3.2 Model performance

After several iterations of model adaptation and submission, the Philter developer received scores in all five PHI categories. Philter achieved its highest F1, precision, and recall on date annotation across datasets, and its lowest performances on ID and location. Token-level evaluation scores exceeded instance-level evaluation scores in most categories, with the exception of ID annotation on the MCW dataset. (**Table 4.2**)

NeuroNER also achieved its highest overall scores for date annotation and its lowest scores for ID annotation. Overall, NeuroNER outperformed Philter. The performance of NeuroNER on the i2b2 test dataset was significantly higher than on other datasets for all PHI categories except the date category. (**Table 4.3**)

Dataset	Evaluation standard	PHI Category(P/R/F1)				
		Date	Person_name	Id	Contact	Location
i2b2	instance	0.77/0.88/0.82	0.29/0.68/0.41	0.34/0.90/0.49	0.31/0.99/0.47	0.27/0.23/0.25
	token	0.79/0.89/0.84	0.51/0.86/0.64	0.37/0.94/0.53	0.43/0.99/0.60	0.52/0.74/0.61
Mayo	instance	0.95/0.99/0.97	0.25/0.55/0.34	/	/	/
	token	0.95/0.99/0.97	0.71/1.00/0.83	/	/	/
MCW	instance	0.76/0.94/0.84	0.19/0.63/0.29	0.13/0.90/0.23	0.25/0.85/0.39	0.09/0.14/0.11
	token	0.77/0.94/0.85	0.41/0.86/0.56	0.10/0.89/0.18	0.30/0.87/0.45	0.28/0.51/0.36
UW	instance	0.85/0.91/0.88	0.19/0.62/0.29	0.22/0.83/0.35	0.38/0.64/0.48	0.15/0.14/0.14
	token	0.87/0.92/0.89	0.38/0.84/0.52	0.21/0.79/0.33	0.39/0.81/0.53	0.27/0.43/0.33

Table 4.2 Philter performance on all NLP Sandbox test datasets.

Evaluation metrics in each cell are represented as precision/recall/F1. The darker red colors correspond to higher F1 scores, and the darker blue colors correspond to lower F1 scores.

Dataset	Evaluation standard	PHI Category(P/R/F1)				
		Date	Person_name	Id	Contact	Location
i2b2	instance	0.92/0.91/0.91	0.91/0.88/0.89	0.67/0.65/0.66	0.86/0.93/0.89	0.86/0.79/0.82
	token	0.93/0.92/0.92	0.96/0.93/0.94	0.65/0.69/0.67	0.91/0.95/0.93	0.96/0.86/0.91
Mayo	instance	0.92/0.99/0.95	0.24/0.20/0.22	/	/	/
	token	0.91/1.00/0.95	0.51/0.76/0.61	/	/	/
MCW	instance	0.89/0.95/0.92	0.76/0.80/0.78	0.51/0.70/0.59	0.42/0.76/0.54	0.37/0.36/0.36
	token	0.89/0.95/0.92	0.87/0.85/0.86	0.45/0.67/0.54	0.43/0.75/0.55	0.60/0.40/0.48
UW	instance	0.92/0.93/0.92	0.79/0.68/0.73	0.40/0.42/0.41	0.60/0.76/0.67	0.62/0.47/0.53
	token	0.93/0.94/0.93	0.86/0/77/0.81	0.36/0.40/0.38	0.66/0.76/0.71	0.68/0.48/0.56

Table 4.3 NeuroNER performance on all NLP Sandbox test datasets.

Evaluation metrics in each cell are represented as F1/precision/recall. The darker red colors correspond to higher F1 scores, and the darker blue colors correspond to lower F1 scores.

4.4 Discussion

In this pilot study, we achieved our goal of evaluating NLP models for de-identification using multiple private datasets. We demonstrated how health institutions, model developers, and independent validation sites all play crucial roles in the development and applications of the NLP Sandbox. The successful operation of the NLP Sandbox requires (1) A standardized but extensible model and data schema, (2) The participation of model developers, (3) The involvement of health institutions contributing their private data for evaluation, and (4) The engagement of external sites to extend the usage of submitted models for validation and implementation outside of the NLP Sandbox environment.

4.4.1 'Model to data' Framework Enabling Secure Data Utilization without Granting Data Access

NLP applications are growing in popularity in the healthcare industry. It is widely recognized that sharing clinical notes more broadly with the data science community can improve the development of Artificial Intelligence (AI) strategies to address clinically-relevant questions and improve healthcare quality. However, bounded by privacy concerns, health institutions are cautious and often progress slowly when it comes to sharing clinical data with both internal and external researchers. The NLP Sandbox offers a secure way for health institutions to open up their data for model evaluation without granting direct access to sensitive data. Under the NLP Sandbox 'model to data' framework, health institutions have full control over how their data is used in the NLP

Sandbox environment. They can choose to deposit their data in any secure computational environment (e.g. on-premises server, AWS, Azure, etc) behind their institution's firewalls to make their datasets available in NLP Sandbox data nodes. Only the models, rather than the model developers, can access the data for evaluation. In this way, no sensitive patient information leaves the health institution.

4.4.2 Federated and Unbiased Evaluation of NLP models in a Ready-to-go Setting

When pre-trained using the 2014 i2b2 training dataset, NeuroNER achieved the highest performance on test data from the same source while the performance did not generalize as well on datasets from different sources. Failure to test model generalizability and self-assessment bias are two main criticisms facing NLP model developers. However, they can be avoided if model developers use an integrated development and testing environment like the NLP Sandbox. Our 'model to data' framework lowers the barrier of access to clinical notes and can accelerate the development and evaluation of NLP models. Developers do not need to wait for months to get approval to access and implement their models on identified clinical notes from other institutions and can quickly retrieve multi-site performance metrics that can be used to inform further model development.

We also made efforts to lower the barrier for developers to adapt their models for submission to the NLP Sandbox. Our model template offers step-by-step instructions for developers to build a Docker image, test the incorporated model in an interactive UI, and submit the model for evaluation. In the first case study, we demonstrated that the NLP Sandbox can be used to iteratively return feedback to developers who wish to improve model performance.

4.4.3 Model and Data Standardization for Future Application

The NLP Sandbox utilizes a standardized data schema and applies quality checks on data converted to the schema. After converting their data to the NLP Sandbox schema, data owners external to the NLP Sandbox can pull the models submitted to the NLP Sandbox from Synapse and seamlessly test the models on their data without exposing their data to the developers. This allows future users to utilize models submitted to the NLP Sandbox and allows for continuous benchmarking of existing NLP Sandbox submissions on newly incorporated data nodes. Our schema is also flexible and can support tasks beyond PHI annotation, as demonstrated in a pilot test using the NLP Sandbox schema to support COVID-19 symptom annotation. We used synthetic data from Mayo Clinic for this task. The data are formatted using the NLP Sandbox COVID-19 annotation data schema (https://github.com/nlpsandbox/nlpsandbox-schemas/blob/main/openapi/commons/components/schemas/TextCovidSymptomAnnotation.yaml). The data include 94 notes and 533 annotations.

We provided a baseline model for this task (https://github.com/nlpsandbox/covid-symptom-annotator-example, v1.2.0). The baseline model achieved 0.71/0.79/0.64 (F1/precision/recall) in instance-level evaluation and 0.70/0.89/0.58 in token-level evaluation.

4.5 Conclusions

In this study, we demonstrated that the NLP Sandbox, as a 'model to data' evaluation system, enables the privacy-protected utilization of clinical notes and unbiased federated evaluation of NLP models. Model developers can receive a comprehensive model evaluation that can be used to improve model performance and generalizability. The standardization of data and model schemas

in the NLP Sandbox enables smooth implementation in the production setting and thus increases the potential for the further application of the models submitted to the NLP Sandbox.

Chapter 5

BENCHMARKING ON GAN-RELATED SYNTHETIC EHR GENERATION ON REAL-WORLD PATIENT DATA

5.1 Introduction

The rapidly evolving ability to collect and manage digitized personal data has made a revolutionary impact on humans' life quality by advancing opportunities for health-related research and applications at scale. [91–93] Throughout decades of medical practice, it has been widely recognized that broad and reliable data sharing is a key booster for the advancement of biomedicine and healthcare; [94] however, such sharing is often constrained (in many cases impeded) by the mandate to preserve privacy - a fundamental ethical and legal principle to defend in data usage - as well as the attendant concerns around the expensive consequences of violating the mandate. As such, when individual-level data sharing does occur, it is often accompanied by systematic assessment and tuning procedures to ensure privacy is sufficiently protected. [95]

As an alternative to operating on real data, generating fully synthetic health data has recently emerged in consort with advances in ML and has become an active research area to guide data sharing practice. [96] The fundamental power of synthetic data is in its potential ability to substantially reduce privacy risks - by breaking the one-to-one mapping between real individuals and synthetic records - while preserving the utility of real data. [35,97] For this reason, multiple research programs have been exploring leveraging synthetic data to support their data

dissemination and analytical needs, such as the U.S. National Institutes of Health-funded National COVID Cohort Collaborative (N3C). [64] Beyond the purpose of data sharing, health data synthesis has a substantial role to play as a solution to data augmentation and completion, where the limited size and imperfect representativeness of available real data can be addressed to enhance the performance of medical AI algorithms. [98–100]

As one of the major techniques applied to synthesize electronic health record (EHR) data, GANs demonstrate success in preserving data utility and privacy for various applications. [96,101,102] The architecture of GANs consists of two competing neural networks—a generator, which transforms data points from an ordinary distribution (e.g., a uniform or Gaussian distribution) to ones satisfying the target distribution through a deterministic function, and a discriminator, which approximates the statistical divergence between the generated data instances and the target distribution to provide feedback for optimizing network parameters. [103] Because of the nature of the adversarial learning mechanism, the GAN family is well known for allowing end-to-end training with a promising capability to approximate the real data effectively. In addition, GAN-based methods do not require any specific assumptions (or prior knowledge) about the target distribution to enable learning.

Despite the proposed merits of GANs, there lacks a systematic assessment framework and a corresponding benchmarking exercise to fairly evaluate and compare various GAN-based models for EHR data synthesis. This creates a challenging situation for users who need to select an optimal method to fulfill their specific needs in data synthesis. First, the empirical assessment of synthetic EHR data lacks consensus on evaluation metrics, and a well-organized system is acutely needed

to help discriminate candidate models according to distinct use cases. Second, although several related publications have conducted comparisons to demonstrate the superiority of their novel approach over existing methods, these comparisons are unsystematic and subject to inadvertently embedded self-evaluation biases. [80] Third, the common issue of stability in the procedure of GAN training renders the one-shot generation and comparison (i.e., performing model evaluation based on only one instance of model training) unreliable and thus could lead to a flawed decision-making process without sufficient attention paid to the variance of evaluation results.

Aim 4 addresses all these challenges by (1) establishing a novel benchmarking system for evaluating static patient profile synthesis (i.e., the static and tabular transformation of longitudinal EHRs), in which a set of metrics covering important requirements around data utility and privacy are systematically organized, and (2) demonstrating a benchmarking practice on the major GAN-based EHR synthesis methods towards the goal of synthesizing static patient profiles based on the developed system. We hope to guide the users through a process of scoping the data characteristics of interest to focus on use cases, conducting individual evaluations using the selected metrics, and identifying the optimal synthesis model(s), and corresponding synthetic data to be used.

5.2 Data

Two real patient datasets were extracted from two respective major academic medical institutions in the United States. Both institutions maintain their own EHR warehouses for secondary analysis purposes by transforming the raw EHR data into the OMOP Common Data Model[104]. The characteristics of these two datasets are described in **Table 5.1**.

VUMC Dataset. We focused on the COVID-19 positive cohort at Vanderbilt University Medical Center (VUMC). Specifically, we extracted the patients who tested positive (via a polymerase chain reaction test) in an outpatient visit before Feb 2021 (which is roughly the first year of the COVID-19 pandemic). For those who had multiple positive testing results, only one was randomly retained. We then collected the patients' history of diagnoses, medications, and procedures from their EHRs at VUMC between 2005 and the time of selected positive test result. Additionally, the most recent readings prior to the selected positive testing events on the 7 prevalent measures or laboratory tests were included, which are *diastolic and systolic blood pressures, pulse rate, temperature, pulse oximetry, respiration rate,* and *body mass index*. We enabled a clinically meaningful prediction task by incorporating an outcome variable—whether a patient would be hospitalized within 21 days from their COVID-19 positive results[85]. In total, there were 20,499 patients, and their data were retrieved for analysis.

UW Dataset. We focused on a general population stored in the University of Washington (UW) Medicine enterprise data warehouse, which manages EHR data from more than 60 medical sites across the UW Medicine system including the University of Washington Medical Center, Harborview Medical Center, and Northwest Hospital and Medical Center. For this study, we extracted patients with at least 10 visits within 2 years prior to the index event date, which is defined as the date of the latest recorded visit as of February 2019, and, thus, patient-specific. The corresponding data retrieval was dated back to January 2007. We also collected an outcome variable to construct a widely investigated prediction task—whether mortality occurred within 6 months after the index event for each patient. [61] The resultant dataset contained 188K patients, which is around 9 times large as the VUMC dataset.

To standardize data representation, we converted the categorical variables in both datasets, including diagnoses (encoded as International Classification of Diseases Ninth or Tenth Revision, or ICD9/10), procedures (encoded as Current Procedural Terminology Fourth Edition, or CPT4), medications (encoded as RxNorm Drug Terminology), and demographics (gender and race), to a binary format to denote the presence (or absence) of the corresponding concepts. We followed the convention of dimensionality reduction preprocessing, [105–107] by 1) mapping the ICD9/10 codes into Phenome-wide Association Studies (PheWAS) codes, or phecodes, which aggregate billing codes into clinically meaningful phenotypes, 2) generalizing the CPT4 codes using a hierarchical architecture of procedures, [108] and, 3) converting clinical RxNorm drugs to RxNorm drug ingredients. We retained those variables that had more than 20 occurrences in each cohort. Note that after data preprocessing, variables in the UW dataset are all binary, whereas the VUMC dataset contains 8 continuous variables (age and 7 laboratory tests). Such a difference was designed to highlight the impact of continuous variables on synthesis quality. The total numbers of model input variables for the UW and VUMC datasets are 2596 and 2670, respectively. Both UW and VUMC datasets have been split in a 70:30 ratio into Training and Evaluation datasets. Training datasets are used to train synthetic algorithms while Evaluation datasets are withheld from the training process and reserved for evaluation purposes.

Characteristic	VUMC Dataset		UW Dataset	
Age	26.0, 40.3, 55.8	41.0 ± 18.7		-
Race				
White	65.2%	13,366	69.9%	131,830
Black	8.8%	1,794	7.9%	14,956
Asian	1.9%	384	9.4%	17,646
American Indian or Alaska Native	0.0%	42	1.5%	2,836
Pacific Islander	0.0%	0	0.8%	1,563
Unknown	24.0%	4,913	10.5%	19,912
Gender				
Male	43.9%	8,990	45.3%	85,490
Female	56.1%	11,509	54.7%	103,253
Medical variables for generation				
Diagnosis (Phecode)	1,269			1,736
Procedure (Category)	67			66
Medication (RxNorm Ingredient)	1,245			860
# of unique codes	2,581			2,662

Average # of codes per patient	45.3			36.8
Min # of codes per patient	0			1
Max # of codes per patient	724			336
Vital Signs (#)		7		0
Diastolic Pressure	68.0, 75.0, 82.0	75.0 ± 10.7		-
Systolic Pressure	114.0, 124.0, 136.0	125.3 ± 15.9		-
Pulse	77.3, 90.0, 104.3	91.4 ± 18.6		-
Temperature	36.8, 37.1, 37.7	37.3 ± 0.6		-
Pulse Oximetry	95.1, 97.1, 99.0	97.1 ± 2.1		-
Respirations	16.0, 18.0, 23.9	19.6 ± 4.4		-
Body Mass Index	24.4, 30.3, 38.1	31.3 ± 8.7		-
Data split for prediction				
Training data				
Positive label	3.8%	541	3.8%	4,966
Negative label	96.2%	13,808	96.2%	127,158
Evaluation data				
Positive label	4.2%	260	3.8%	2,129
Negative label	95.8%	5,609	96.2%	54,490

Table 5.1 Cohort characteristics for synthetic data generation.

5.3 Methods

5.3.1 Benchmark Framework

Unlike previously published studies in generating synthetic EHR data, which conducted model comparisons based on one-shot model training, we integrated a mechanism into the benchmarking framework that involved multiple-time training and selection to guard against baking biases into model comparison. The unstable nature of GAN training has been known to be problematic for generating synthetic datasets of similar fidelity[109–111] but our mechanism accommodated this mode of uncertainty. Specifically, for each pair of models and real datasets, we implemented model training and data generation from scratch five times. We selected the top three datasets for the benchmarking analysis in terms of the average of the absolute prevalence rates differences (APD) in the dimension-wise distribution metric, which is deemed a basic utility measure and provides straightforward evidence of the usability of synthetic data. By doing so, the synthesized datasets which performed poorly at capturing the first-moment statistics of individual variables (though nothing abnormal happened in the GAN training process regarding the loss trajectories) were ruled out. (**Figure 5.1**, step 1)

The benchmarking framework (**Figure 5.1**, step 2) consists of two major components—*utility* and *privacy* - which are in alignment with the well-known utility-privacy trade-off. In contrast to the flat hierarchy used by several published studies[96,101], where the *utility* was equated to *resemblance* (i.e. the similarity or distance between two datasets) or in particular utilized to measure the performance gap between real and synthetic data in downstream outcome predictions, the term utility in our framework covers all evaluations regarding the characteristics of data except

for privacy risks. This is because numerous real-world use cases of synthetic data do not involve any downstream prediction tasks, but still require the synthetic data to be useful (in other words, they require that the dataset has utility).

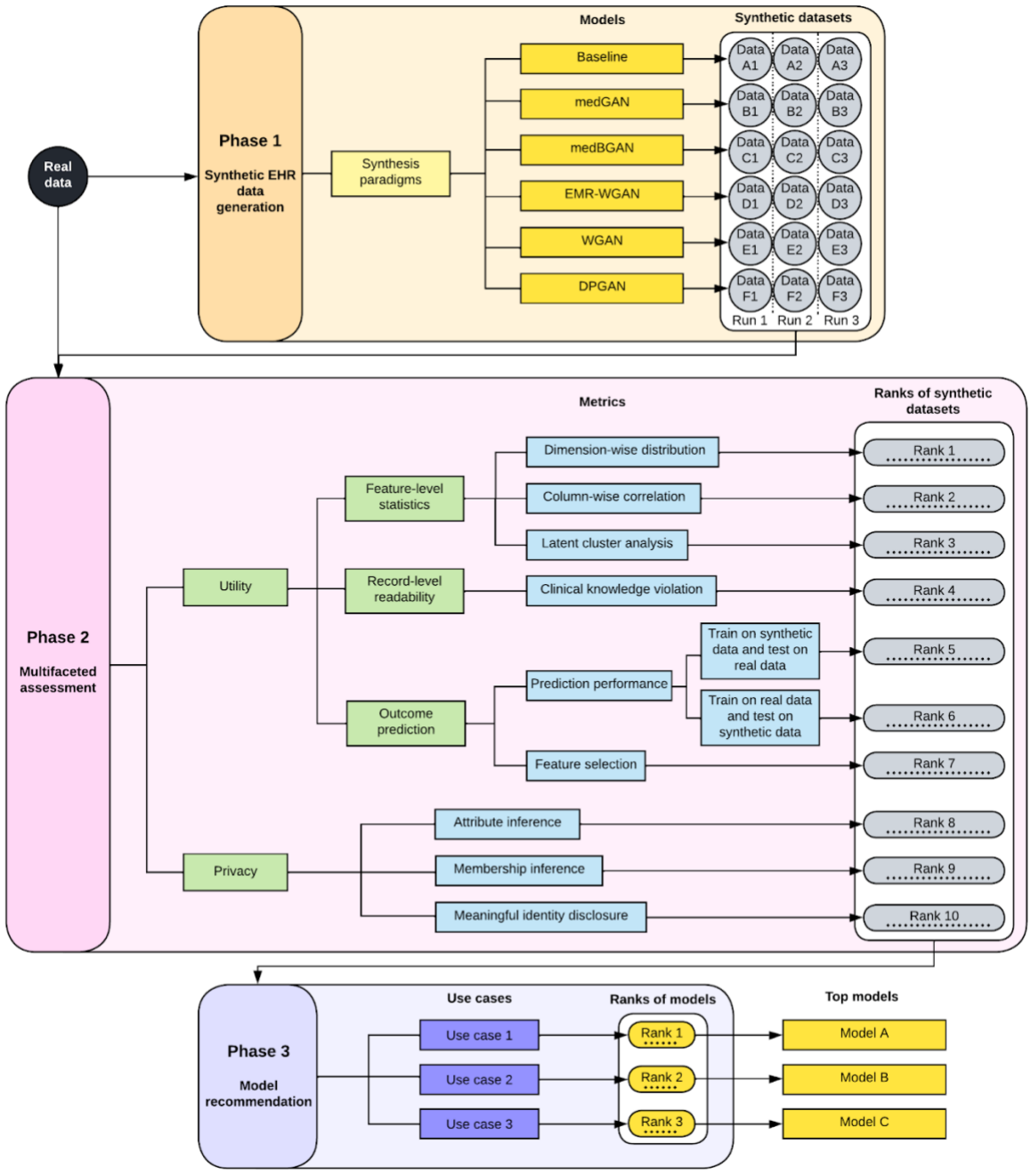


Figure 5.1 Overview of the benchmarking framework and the data pipeline.

Table 5.2 summarizes the focuses and the corresponding key measurements of all metrics included in the benchmarking framework. The examination of data utility of synthetic data generated by a model expands along with three aspects that are distinct but compensatory to each other: 1) feature-level statistics, the potentiality of capturing distributional characteristics in real data (*dimension-wise distribution,[105] column-wise correlation, [112]* and *latent cluster analysis[113]*), 2) record-level readability, the ability to avoid generating individual records that violate clinical knowledge or common sense (*clinical knowledge violation*), and 3) Outcome prediction, the ability to support downstream machine learning model development and interpretability (*model performance[114]* and *feature selection*). Among these, *feature selection* and *clinical knowledge violation* are the two new metrics that were neglected but are acutely important in real-world use cases. They are thus integrated into the framework, while others were selected from previous studies. Notably, the focuses of the three metrics in the first category differ in their levels of distributional characteristics, which are individual variables, correlations in paired variables, and the joint distribution of all variables. On the other hand, the three privacy metrics were incorporated to assess the privacy risks of synthetic datasets derived from distinct but common adversarial environments (where attack models differ): attribute inference[105] (inferring the unknown attribute values of interest given a set of real attribute values), membership inference[105] (inferring whether a real record was used to train a generative model), and

meaningful identity disclosure[115] (inferring the identity and sensitive attributes of a patient's record in the original dataset). The privacy risks were measured under the common assumption of attackers' knowledge—they only have access to synthetic data but not the generative models. Note that the term *attribute* in the context of privacy risks is equivalent to variable in general, and we use *record* to denote the vector of data points for a patient.

	Metric	Summary	Directionality
Utility	**Dimension-wise distribution**	The ability to capture marginal feature distributions in real data. This is calculated as the average of the absolute prevalence differences (APD) for binary features and the average of the Wasserstein distances (AWD) for continuous features between real and synthetic datasets.	↓
	Column-wise correlation	The ability to capture the relationship between two features in real data. This is calculated as the average of the cell-wise absolute differences of the Pearson correlation coefficient matrices derived from real and synthetic datasets.	↓
	Latent cluster analysis	The ability to capture the joint distribution of all features in real data. This is calculated as the deviation of a synthetic dataset in the underlying latent space from the corresponding real dataset in terms of unsupervised clustering.	↓
	TSTR Model performance	The ability to approximate the performance of the downstream task of machine learning model development. Given an outcome prediction task, this is calculated as the model performance, typically the area under the receiver operating characteristics curve (AUROC), in the scenario of training on synthetic dataset and testing on real dataset (TSTR).	↑

	TRTS Model performance	The ability to generate convincing and realistic data records for different labels. Given an outcome prediction task, this is calculated as the model performance, typically the AUROC, in the scenario of training on real dataset and testing on synthetic dataset (TRTS).	↑
	Feature selection	The ability to support model interpretability in downstream tasks. This is calculated as the number of shared important features for models trained on a synthetic dataset and the corresponding real dataset.	↑
	Clinical knowledge violation	The ability to learn the clinical common sense at the patient level. This is calculated as the proportion of generated records that violate clinical knowledge (e.g., a male patient is associated with a pregnancy code).	↓
Privacy	**Attribute inference**	The adversary's ability to infer sensitive attributes of a targeted record. Given demographics and some sensitive attributes of a targeted record, this is calculated as the weighted sum of F1 scores of the inferences of other sensitive attributes.	↓
	Membership inference	The adversary's ability to infer the membership of a targeted record. Given a set of attributes of a targeted record, this is calculated as the F1 score of the inference based on Euclidean distances between the targeted record and all synthetic records.	↓
	Meaningful identity disclosure	The adversary's ability to identify synthetic records with meaningful attributes. Given a population dataset with identities, this is calculated as the adjusted re-identification risk considering the linkage between the synthetic dataset and the real dataset, the linkage between the synthetic dataset and the population dataset, and the rareness of each sensitive attribute in the real dataset.	↓

Table 5.2 The summarization of the metrics and their focuses in evaluation.

The directions of the derived values are provided, where the upward arrows denote that a higher value corresponds to a better score and the downward arrows denote an opposite relationship.

Utility

Dimension-wise distribution. The distributional distance (or similarity) of each variable between a real and synthetic dataset has been commonly investigated as one of the basic measurements to quantify data utility. To do so, we calculated the average of the absolute prevalence rate difference (APD) for binary variables and the average of the variable-wise Wasserstein distances (AWD) for continuous variables. The prevalence of a concept was defined as the percentage of patients who exhibited this concept among the whole cohort. Note that these two measures are symmetric and were computed over absolute values of distances. Due to the fact that Wasserstein distances can be arbitrarily large, for each continuous variable, we normalized the corresponding Wasserstein distance into the range of [0,1] based on the distances derived from all candidate synthetic datasets in the benchmarking system. To make the average value readable, we multiplied the results by a factor of 1000, which was equivalent to the magnitude of the space of variables and made no impact on the ranking of models. For a dataset with both binary (note that categorical variables can be converted to binary variables for synthesis) and continuous variables, the results of APD and AWD need to be combined into a final score. One reasonable implementation we used in this study was to add up 1) the sum of the absolute prevalence rate difference for binary variables, and 2) the sum of the variable-wise (normalized) Wasserstein distances for continuous variables, and then calculated an average (i.e., divided by the total number of variables).

Column-wise correlation. It quantifies the degree to which a synthetic dataset retains the variable correlations in real data[112]. For each pair of the synthetic and real datasets, we first computed the Pearson correlation coefficients between all variables in each dataset, which yielded two correlation matrices of the same size. The summation of all cell-wise absolute differences between

the two correlation matrices was then calculated to quantify the fidelity loss of a synthetic dataset in correlations between variables.

Latent cluster analysis. This analysis measures the deviation of a synthetic dataset in the underlying latent space from the corresponding real dataset in terms of unsupervised clustering. [113] For each synthetic-real datasets pair, we concatenated them and performed dimension reduction by applying principal component analysis (PCA) to retain the dimensions in the projected space that can collectively explain 80% of the variance in the system. In the resultant latent space, *K*-means was then applied to find clusters, where *K* was determined based on the elbow method (which was set to 3 for both the VUMC and UW datasets). The following clustering metric was then calculated to quantify the deviation of synthetic data from real data:

$$\log\left(\frac{1}{K}\sum_{i=1}^{K}\left[\frac{n_i^R}{n_i}-0.5\right]^2\right)$$

where n_i^R and n_i denote the number of real data points and the total number of data points in the i -th cluster, respectively. A lower value of the metric implies that the density functions of the real and synthetic datasets in the latent space are more similar.

Clinical knowledge violation. Different from other metrics above, this metric focuses on the record-level utility, which quantifies the degree to which a generative model learns to synthesize clinically meaningful records in terms of the ability to capture common sense (e.g., a male patient does not associate with a pregnancy ICD code). Synthetic data with obvious violations against clinical knowledge can be less useful in use cases requiring record-level reasonableness. In this study, we leverage gender-related clinical knowledge for evaluation. For each gender, we selected

the most frequent 3 phecodes that were only associated with this gender and then computed the violation rate on each phecode as the odds of appearing in the opposite gender, e.g. the percentage of records with pregnancy code among all synthetic male patients. A higher value implies a lower capability of capturing clinical common sense derived from real data.

Prediction Performance and important variables. One of the most important use cases for high-utility synthetic data is to support downstream ML model development and evaluation. [116,117] To investigate the capability of a model to generate synthetic data to replicate the performance of downstream model development on real data, we incorporated a two-fold analysis. The first analysis, which is straightforward and has been widely utilized, compares the model performance in two distinct scenarios: (1) train a ML model using the synthetic dataset (obtained from a generative model learned from a real dataset) and then perform an evaluation based on another independent real dataset, and (2) train a model based on the independent real dataset and evaluate it using the synthetic dataset. Each scenario has its reference model trained from the corresponding real dataset to compare with. The first scenario adheres to how the synthetic data will be utilized, whereas the second plays a complementary role to assess how convincingly the synthetic dataset matches its labels. [114] We used the light gradient boosting machine (LightGBM) in this study due to its consistent superiority over traditional models in healthcare. [9,118,119] In the following evaluation, we used the reserved 30% evaluation dataset as the independent real dataset. The area under the receiver operating characteristic curve (AUROC) was utilized as the major performance measure. Bootstrapping was used to derive the 95% confidence intervals around the estimate of performance.

The second analysis, by contrast, measures the degree to which a synthetic dataset provides reliable insights into important variables in the target ML task. This was incorporated into the benchmarking system because the global-level (instead of patient-level) model explainability in medical AI becomes increasingly important for engendering trust and conducting algorithmic audits. [120,121] To do so, we counted the overlapping top important variables for models trained on a synthetic dataset and the corresponding real dataset. We used the SHapley Additive exPlanations (SHAP)[122] value to rank variables and defined the important variables as the top *M* features that retain 90% of the performance on real data, which is 20 for the VUMC dataset and 25 for the UW dataset.

Privacy

Attribute inference. An adversary tries to infer a set of sensitive attributes of a targeted record in the real training dataset given the synthetic dataset and some known attributes of the targeted record. The set of attributes known by the adversary usually includes demographic traits (such as age, gender, and race) and common diseases, while the sensitive attributes are the target's other diseases. We assume the adversary infers the sensitive attributes using a k-nearest neighbors algorithm - the adversary first finds a set of *k* records in the synthetic dataset that are the most similar to the targeted record based on the set of known attributes (the k records are known as the neighbors), and infer each unknown attribute for the targeted record using a majority rule classifier for the neighbors.

We calculated an F1 score for each unknown binary feature and accuracy for each unknown continuous attribute, where the accuracy was defined as the percentage of predictions that were

close enough to the true value given a closeness threshold. Afterward, the attribute inference risk was computed as a weighted sum of risks for attributes in which the weight is proportional to the corresponding information entropy for each attribute in the training dataset, and all weights were summed up to one.

All patient records in the training dataset were treated as targets. We set k as 1 and used the Euclidean distance as the distance measure for the k-nearest neighbor algorithm. We assumed that the adversary knows the demographic attributes (age, gender, and race for the VUMC dataset; gender, and race for the UW dataset) and 256 phecodes that are the most frequent in the training dataset. The adversary aims at inferring all the other phecodes and continuous attributes. The closeness threshold for calculating accuracy was set to 0.1.

Membership inference. The objective of the membership inference attack is to conclude whether a patient record was previously used for training in the generation process of a given synthetic dataset. Privacy is compromised under a successful membership attack, as a linked membership of a patient record to the original training dataset can disclose some sensitive traits of the record. The inclusion criteria for the training dataset may give away private information that is disease-specific (e.g., HIV), demographic-specific(e.g., a certain sexual orientation), or location-specific (e.g., a certain hospital). This information may not be included as a feature in the training dataset, as it is shared by all records in the dataset, thus, it can not be inferred in the aforementioned attribute inference attack.

In the membership inference test, [106] we assumed that the adversary was in possession of all attributes of a targeted patient record. We first calculated the Euclidean distance between each synthetic patient record and the target patient record. Given a distance threshold, the adversary claims that the target is in the training dataset if there existed at least one record with a distance smaller than the threshold. F1 score was used as the risk measure. We used all real patient records, including both the training dataset and the evaluation dataset, as the target records and normalized all continuous attributes into a range of zero and one. The distance threshold was set as 2 for our experiment.

Meaningful identity disclosure. Although a fully synthetic dataset seems to have no risk of identity disclosure, a synthetic dataset generated by an overfitting generative model may still lead to record linkage. El Emam et al. proposed a risk model[115] that quantified both the risk of identity disclosure and the ability of an adversary to infer new information. In this attack, an adversary can link the synthetic dataset to a population dataset upon quasi-identifiers (i.e., the common attributes in both datasets) and infer the values of sensitive attributes for patients in the population dataset by applying the majority rule on linked records in the synthetic dataset. It is also assumed that the adversary can generalize any attribute in any record to a certain level (i.e., an age, 20, can be generalized to an age group, [20 - 29]) and then conduct the record linkage attack.

The meaningful identity disclosure risk is calculated based on the marketer risk measure [123] and additionally considers the uncertainty and errors in the adversary's inference. It is calculated using the following equation:

$$max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s}\times\frac{1+\lambda_s}{2}\times I_s\times R_s\right),\frac{1}{n}\sum_{s=1}^{n}\left(\frac{1}{F_s}\times\frac{1+\lambda_s}{2}\times I_s\times R_s\right)\right)$$

In which *N* is the number of records in the population dataset. *s* is the index for a record in the real training dataset that the synthetic data are trained from. *n* is the number of records in the real training dataset. f_s is the number of records in the real training dataset that can match record *s* in terms of values on the quasi-identifiers (QIDs). F_s is the number of records in the population dataset that can match record *s* in the real training dataset in terms of values on the QIDs. λ_s is an adjustment factor based on error rates sampled from 2 triangular distributions [115]. I_s is a binary indicator of whether record s in the real training dataset matches a record in the synthetic dataset. R_s is a binary indicator of whether the adversary would learn something new. R_s is 1 if at least *L*% of the sensitive attributes satisfy the following criteria. For each categorical attribute, the criteria are: (1) there is at least one synthetic record that can match at least one real record on that sensitive attribute, and (2) $p_j < 0.5$, in which p_j is the proportion in the real training sample that has the same *j* value, and $j \in J$ in which *J* is the set of different values the sensitive variable can take. For each continuous attribute, the criterion is $p_s \times |X_s - Y_t| < 1.48 \times MAD$, in which p_s is the proportion in the real training sample that is in the same cluster with the real training record after a univariate k-means clustering, X_s is the sensitive attribute of the real training record, Y_t is the sensitive attribute of the synthetic record matching the real training record, and *MAD* is the median absolute deviation.

For the VUMC dataset, we assumed that the adversary can get access to a population dataset of 633,035 records, which includes the name and 10 QIDs of all patients that have visited VUMC before February 2021. The corresponding QIDs are three demographic attributes (namely, age,

sex, and race) and the seven most frequent phenotypic attributes. For the UW dataset, we assumed that the adversary could get access to a population dataset of 466,980 records including the name and 10 QIDs of all patients who have visited UW at least five times in the past 2 years prior to the index event date, which is defined as the date of the latest recorded visit as of February 2019. The corresponding QIDs are two demographic attributes (namely, sex and race) and eight most frequent phenotypic attributes. The parameter L is set to 1 which means at least 26 (27) attributes need to be inferred correctly and meaningfully for an attack to be regarded as a successful attack that brings risk to VUMC (UW) dataset. The difference between VUMC and UW datasets is because the VUMC dataset has 74 fewer attributes than the UW dataset has.

5.3.2 Ranking Strategies

We designed a novel ranking strategy that scores each candidate model based on the evaluation results of the aforementioned metrics across three independently generated datasets (**Figure 5.1**). Specifically, for each metric, we calculated the corresponding values for the selected synthetic datasets generated by all candidate models for comparisons. We then sorted these values with smaller ranks denoting better scores. The average rank of the three datasets representing the same model was taken as the score of this model on the given metric. By doing so, there were ten sets of model scores corresponding to the metrics in the benchmarking framework. The final score of a model is defined as the weighted sum of scores over all individual metrics with the weights determined based on specific use cases. With the final scores computed, the optimal model(s) for a specific use case can be naturally identified. The primary reason for operating the benchmarking on the ranks of individual synthetic datasets (rather than the mean value derived from multiple

datasets that represent a model) was to enable the consideration of the instability level of GAN training throughout the benchmarking pipeline.

5.3.3 Models for Benchmarking

In this study, we incorporated 5 major GAN-based models that were recently published and designed to synthesize static EHR profiles of patients[96], including medGAN[105], medBGAN[124], EMR-WGAN[106], WGAN[124], and DPGAN[125]. We excluded from our benchmarking system models that aimed to resolve a very specific need of EHR synthesis, such as MC-medGAN, which allows for a better representation of categorical features[126], CorGAN, which was designed for capturing correlations (if any) between physically adjacent variables, [127] HGAN, which considers value constraints between variables, [128] and others with minor adjustifications on the original GAN architecture. Additionally, we incorporated a simple approach that randomly sampled variable values across real records as an important baseline (referred to as sampling-based baseline, or Baseline for short) to complement the scope of benchmarking in terms of the variety of model behavior. **Figure 5.2** illustrates the architectures of all models incorporated into the benchmarking.

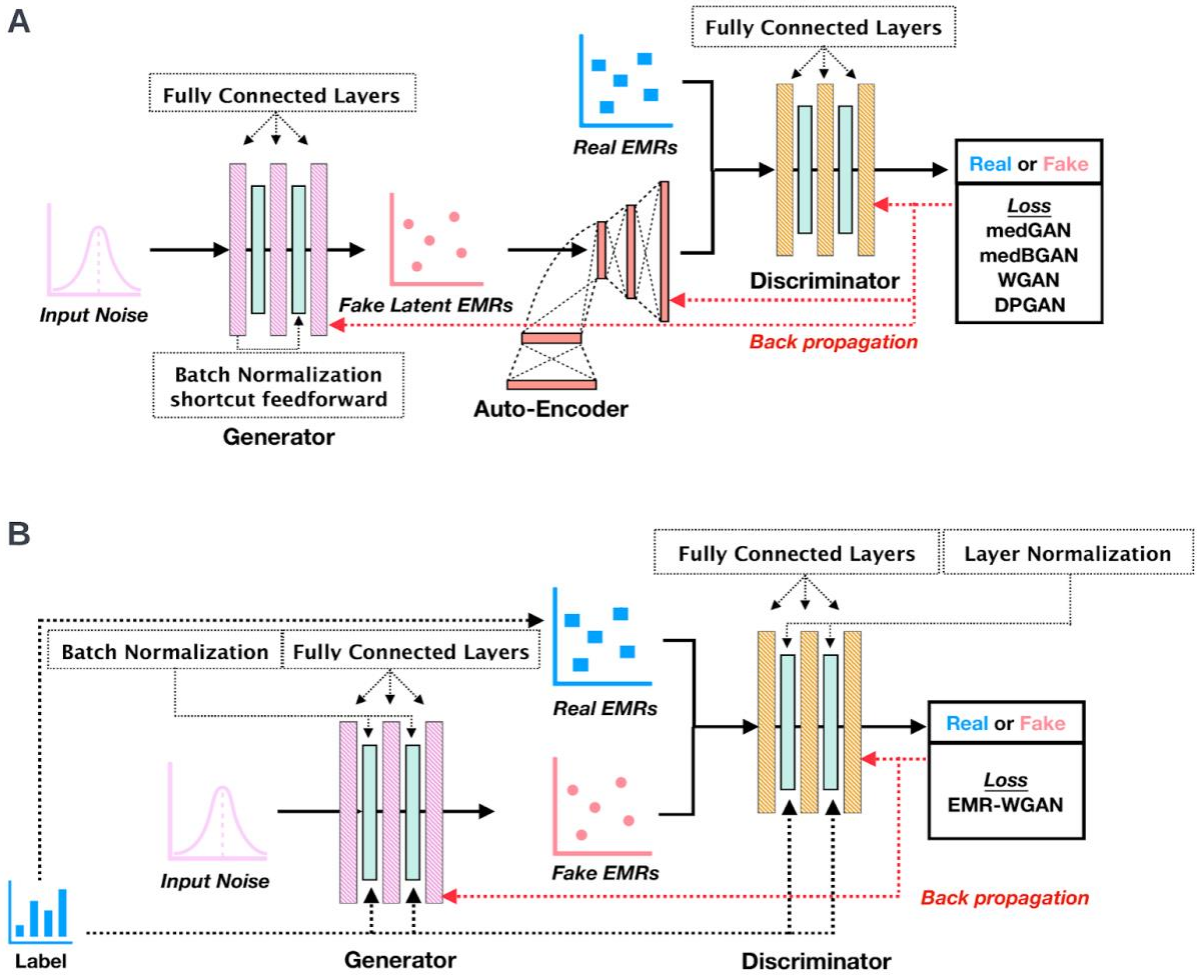


Figure 5.2 Architectures of the deep generative models.

A) medGAN, medBGAN, WGAN, and DPGAN. B) EMR-WGAN.

medGAN was an early attempt to extend the power of vanilla GANs[103] to synthesizing individual-level health data. [105] The discrete format of medical concepts presented an additional challenge for model training, in that the vanilla GANs' approximation of the discrete representations of health data rendered the training process suboptimal. To address this issue, medGAN leveraged a pre-trained autoencoder to project the discrete representations into a

compact continuous space to enhance the subsequent GANs training. Also, medGAN integrated a set of helpful learning techniques, including batch normalization and short connection, to stabilize the training process. Note that medGAN inherited the Jensen-Shannon Divergence (JS Divergence) between the real and synthetic data distributions as its optimization objective. These designs (except for the learning objective) were inherited by several later models as shown in **Figure 5.2**A.

medBGAN was built on the basis of medGAN with the same model architecture.[124] To enhance training stability and generation performance, particularly when dealing with discrete data, medBGAN replaced the discriminator's loss function of medGAN with a boundary-seeking loss function. In doing so, the generator is directed to generate samples at the decision boundary of the discriminator, which enables better generation performance, particularly for discrete data.

WGAN was proposed due to the fact that having JS divergence as the optimization objective can lead to diminishing gradients, which can subsequently impede the optimization of the generator. [124] To resolve this issue, Wasserstein divergence[129] was adopted as its training objective, which requires a Lipschitz constraint on the discriminator and ensures a much more accurate characterization of the distance between two distributions. The WGAN implementation in this paper used the parameter clipping strategy to satisfy the Lipschitz constraint.

EMR-WGAN upgraded the common model designs of medGAN and their successors by removing the autoencoder from the architecture[106] (**Figure 5.2**B), which can introduce barriers during model training when working with an advanced distance measure between two distributions (i.e., Wasserstein divergence). Additionally, EMR-WGAN introduced the layer normalization to the discriminator for further improvement of learning performance and used a gradient penalty

strategy to enforce the Lipschitz constraint, which alleviated the negative impact of parameter clipping on model capacity.

DP-GAN was a differentially private (DP) version of WGAN, [125] which achieved theoretically guaranteed privacy protection on synthetic health data via employing the differential privacy principle[130] in GAN training. It followed the differential private stochastic gradient descent (DP-SGD) mechanism to fulfill the DP requirement but modified the implementation of DP-SGD by replacing the gradient clipping with parameter clipping.

5.3.4 Model Training and Data Selection

For model training, we first normalized the continuous variables by mapping them into a range of [0,1], then trained generative models using the normalized real data, and finally mapped the generated data back to the original space as a post-training step. To ensure fairness, we constrained the shared hyperparameters between models to have the same values, including the architecture of the deep neural networks, learning rate, optimizer, initialization strategy, etc. The source code of this study can be found at https://github.com/yy6linda/synthetic-ehr-benchmarking.

Notably, we observed that the divergence loss of medGAN and medBGAN both demonstrated a pattern of heavy vibration before soaring quickly to a very large number; thus, we selected the epoch immediately before the increasing trend occurred, which usually corresponds to the lowest losses. By contrast, the losses of EMR-WGAN and WGAN demonstrated an apparent convergence pattern; however, it turned out that the quality of synthetic data in the convergent area can differ. To address this mode of variation, we considered multiple epochs in the convergent area for these two models.

5.3.5 Variations on Generation Strategies

We investigated two distinct synthesis strategies regarding the outcome variable of interest in a dataset, which can be leveraged to support a variety of downstream tasks, such as improving ML models and conducting algorithmic audits. The first synthesis strategy treated the outcome variable the same as other variables in model training, which led to a combined synthesis paradigm, whereas the second strategy was designed to train a model for each outcome in the space independently, leading to a separate synthesis paradigm. **Figure 5.3** illustrates the workflows of these two synthesis strategies. We applied the combined training strategy to both datasets and compared the two strategies on the UW dataset. We followed the convention to ensure that the synthesized data shared the same size as the corresponding real dataset and that the distribution of the outcome variable remained the same as well.

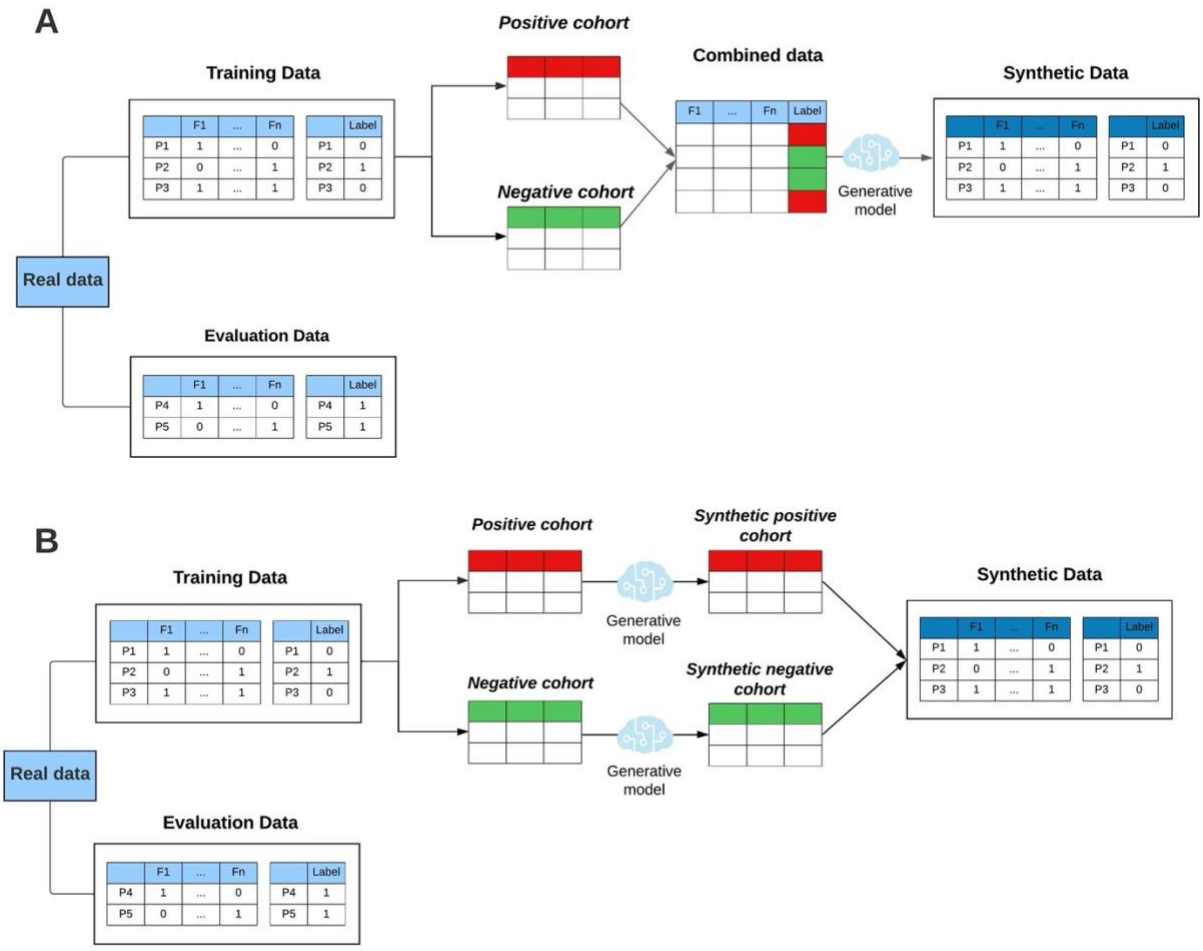


Figure 5.3 Data generation workflow.

A. combined synthesis paradigm; B. separate synthesis paradigm

5.3.6 Use Case Scoring

Considering the utility-privacy trade-off, we think it is important to take use cases into consideration when it comes to GAN-based model evaluation. Here, we provided several use cases to shed light on how to interpret results generated from our evaluation framework. (**Figure 5.1**, step 3) By default, a weight of 0.1 is associated with each metric. Through adjusting the weights

assigned to each metric, a weighted score combining both utility and privacy metrics for each model can be computed for various use cases.

(1) Synthetic data can be used for **educational purposes**. The target users for this use case are medical or informatics students who are interested in learning the common data model of patient data and extracting patient information from the data. We posit that the privacy risk is less acute in this use case as students are affiliated with institutions that own the data and thus data access is easier to control. Necessary governance and administrative contracts like data use agreements (DUAs) can be executed to further protect the data. In contrast to low privacy risk, educational use cases present high demand for the accuracy of medical records and correlations and consistency between patient records. Thus, we lowered the weight assigned to each of the privacy metrics to 0.05 and raised the weight for dimensional-wise distribution to 0.25, column-wise correlation and clinical knowledge violation to 0.15, with the remaining metrics set to 0.1(weight combined to 1).

(2) Synthetic data play an important role in **EHR DREAM challenges.** [61,62,85] Under the model-to-data framework, participants develop models without accessing the real data and then send their containerized models to the data hosts which operate and evaluate the models' performance on behalf of the participants. A synthetic dataset is often provided to participants to allow them to visualize the data format. However, the synthetic datasets in previous EHR DREAM challenges were not produced using any real data and thus were not representative of real patient data. Participants were discouraged from using model performance on synthetic data to inform feature selection and parameter tuning. We think high-quality synthetic data can fill that gap by providing participants hands-on experience with data of a similar distribution to real data. For model assessment under this use case,

we prioritized model prediction-related utility metrics (i.e., performance results and feature selection) and privacy; because the EHR DREAM challenge is open to the broad data science community, this use case presents heightened privacy risk. We kept the privacy combined weight as 0.3 and raised model performance-related weight to 0.5.

(3) **System development** often requires synthetic data as a placeholder for workflow testing and for estimating run time and computational resources. Under this use case, it is of the utmost importance that the synthetic data maintain the sparsity and size of the real data, which is highlighted in the dimension-wise distribution, while other utility metrics for downstream analysis are less necessary. Privacy still needs to be prioritized because the synthetic data need to be broadly shared with engineers who have access to the system. For weight selection in this case, 0.25 was assigned to dimension-wide distribution, 0.5 was assigned to privacy metrics combined, and with the rest of the utility metrics were lowered to 0.05. See **Table 5.3** for detailed weight setting.

	Dimension-wide distribution	Column-wise correlation	Latent cluster analysis	Model performance	Feature selection	Clinical knowledge violation	Attribute inference	Membership inference	Meaningful identity disclosure
Educational	0.25	0.15	0.1	0.1	0.1	0.15	0.05	0.05	0.05
EHR DREAM challenges	0.05	0.05	0.05	0.35	0.15	0.05	0.1	0.1	0.1
System design	0.25	0.05	0.05	0.05	0.05	0.05	1/6	1/6	1/6

Table 5.3 Weight setting for different use cases.

5.4 Results

5.4.1 Results for VUMC and UW Synthetic Data Using a Combined Synthesis Paradigm

The prevalence rates of categorical variables for both real and synthetic datasets were calculated (**Figure 5.4**). A consistently outstanding ability to capture the marginal distributions in real data was found for the Sampling-based Baseline (which corresponds to the lowest values of APD among all models) on both real datasets. Among the GAN-based synthetic datasets, those generated by EMR-WGAN demonstrated a clear pattern for both datasets that all dots were closely distributed along the diagonal line, suggesting a strong capability of preserving the first-moment statistics in real data. By contrast, medBGAN, medGAN, and DPGAN (with a descending order of utility in terms of APD values) were less useful to inform learning with the tendency to blur signals in real data. For the VUMC dataset, the synthetic datasets from WGAN showed a similar pattern to those generated by EMR-WGAN and the associated APD values were also close; however, this relation changed for the UW dataset with the APD values for WGAN much higher than those with EMR-WGAN. DPGAN, which enforced a differential privacy constraint on WGAN, exacerbated this issue by distorting the marginal distribution in real data and biasing heavily towards the synthetic data. Additionally, all models except for WGAN and DPGAN were consistently better on the UW dataset than on the VUMC dataset in terms of the APD values. These observations were baked into comparisons and were well reflected in the ranking results on this metric (**Table 5.4**).

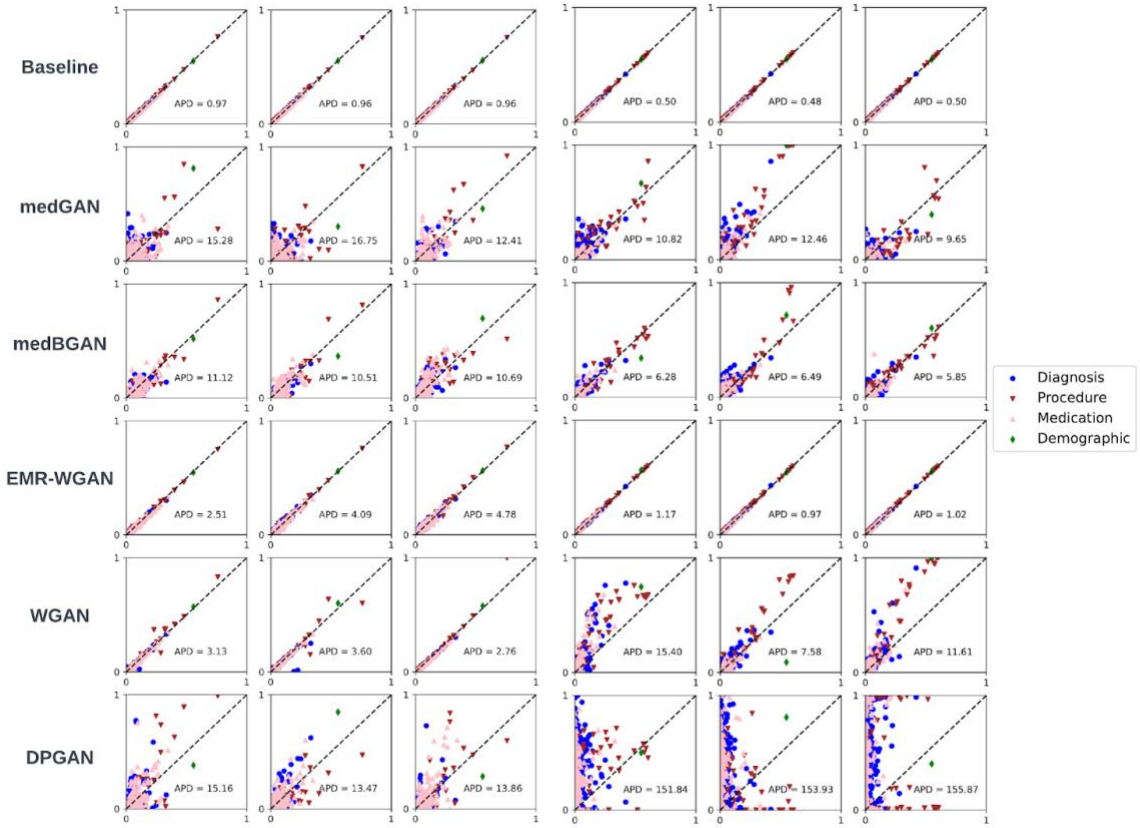


Figure 5.4 Dimension-wise distribution.

VUMC (left) and UW (right) datasets (in combined generation) with x- and y-axis denoting the prevalence rates of a variable in real and synthetic data, respectively. Average prevalence differences (APD) between real and synthetic datasets across all variables are marked. The dashed diagonal line denotes the perfect distribution.

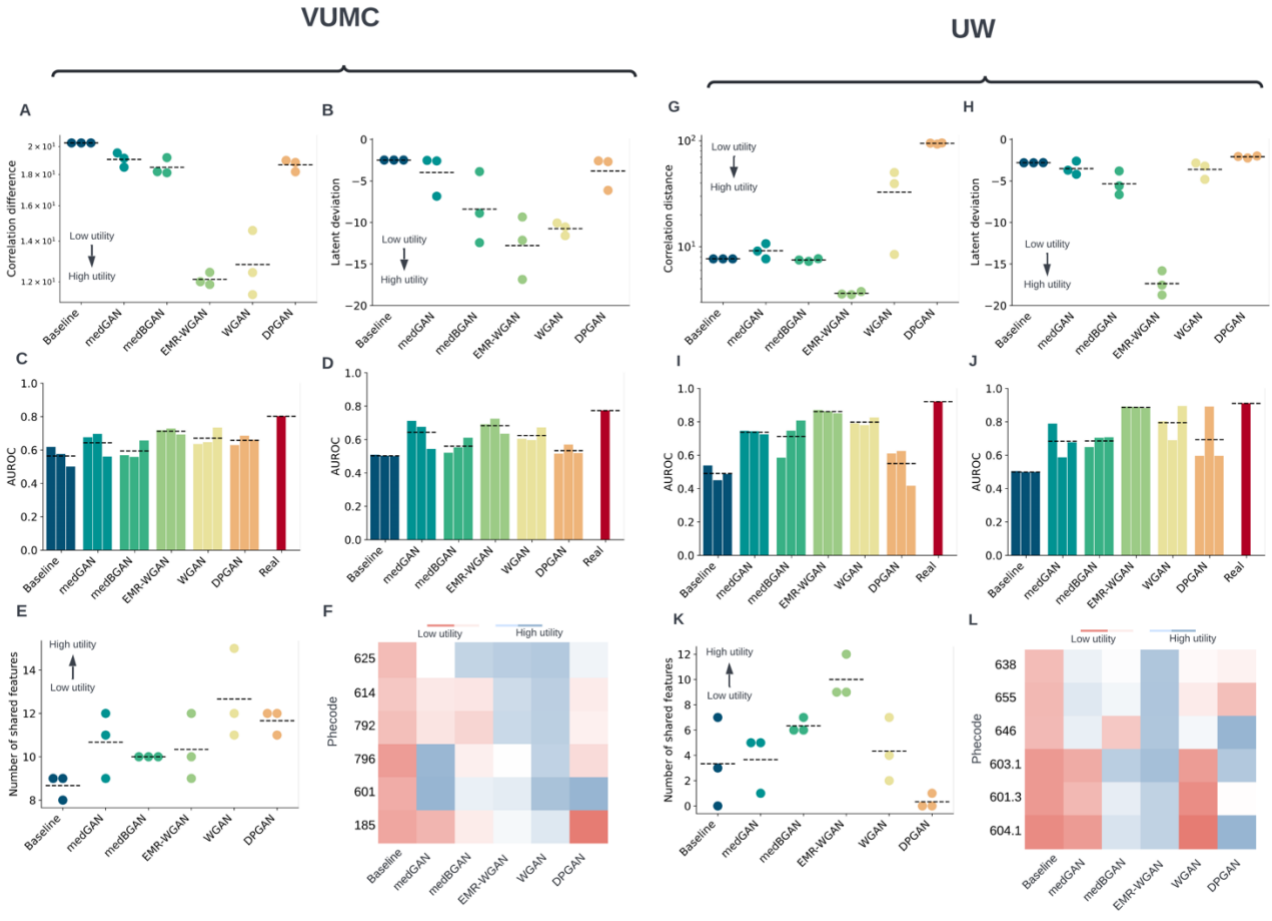


Figure 5.5 Data utility

VUMC (left, A-F) and UW (right, G-L) datasets. (A,G). Column-wise correlation, (B,H) Latent cluster analysis, (C,I) Model performance (TSTR) for training on synthetic data (trained from 70% real data) and testing on 30% real data, (D,J) Model performance (TRTS) for training on 30% real data and testing on synthetic data, (E,K) Overlapping top K features (25 for UW and 20 for VUMC), (F,L) Clinical knowledge violation against gender-specific phecodes as the y axis. The color of each cell in the heatmap was assigned based on the ratio of clinical knowledge violations for specific gender. Dashed lines indicate the mean values on three synthetic datasets. Note that

(C,D) and (I,J) correspond to the metric for model performance. Phecodes 625=Symptoms associated with female genital organs; 614=Inflammatory diseases of female pelvic organs; 792=Abnormal Papanicolaou smear of cervix and cervical HPV; 796=Elevated prostate specific antigen; 601=Inflammatory diseases of prostate; 185=Prostate cancer; 638=Other high-risk pregnancy; 655=Known or suspected fetal abnormality; 646=Other complications of pregnancy NEC; 603.1=Hydrocele; 601.3=Orchitis and epididymitis; 604.1=Redundant prepuce and phimosis/BXO.

EMR-WGAN demonstrated the highest average data utility for the UW dataset for all individual utility metrics except for dimension-wise distribution (**Figure 5.5** G-L), and outperformed other models in four (**Figure 5.5** A-D) of the six metrics for the VUMC dataset. On the other two metrics (i.e., feature selection and clinical knowledge violation), WGAN led the performance, suggesting its stronger support for interpretability of prediction model development and patient-level record plausibility. By contrast, Baseline was stably associated with the worst average data utility for the VUMC dataset (**Figure 5.5** A-F) and was ranked as one of the bottom two models for five metrics for the UW dataset (**Figure 5.5** H-L). This was also unsurprising because the sampling strategy neglected the joint distribution of real data. DPGAN, though never the worst among all models in average data utility (on individual metrics) for the VUMC dataset, demonstrated the lowest utility for the UW dataset regarding column-wise correlation, latent cluster analysis, and feature selection. The results based on the ranking strategy on individual metrics (**Table 5.4**) were roughly consistent with using the mean value of metric scores derived from multiple datasets to sort models; however, this can change when the utility variance from training a candidate model was large. For example, medGAN outperformed medBGAN with a higher averaged AUROCs in the

scenario of training on synthetic and testing on real data (**Figure 5.5** I) but corresponded to a flipped ranks (**Table 5.4**) due to the fact that two datasets generated by medBGAN outperformed all datasets generated by medGAN.

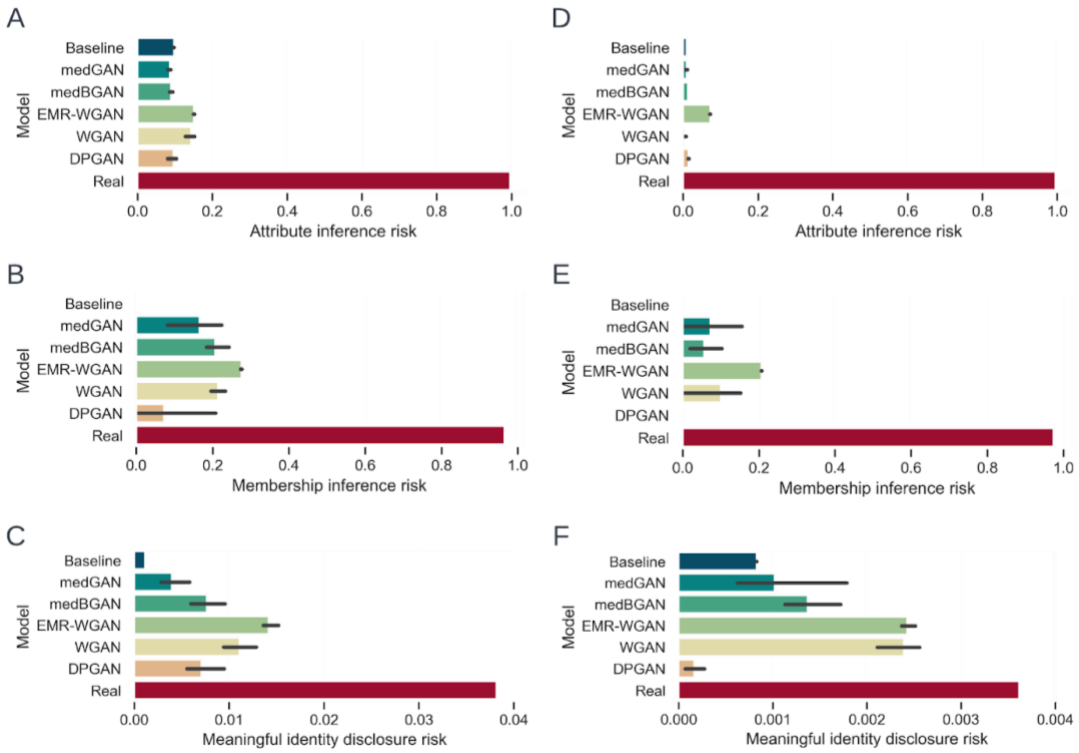


Figure 5.6 Privacy risks

VUMC (A-C) and UW(D-F) datasets. The 95% confidence intervals are marked. All synthetic datasets were generated using the combined generation approach.

It was observed from the privacy risk assessment that all generated synthetic datasets corresponded to a lower privacy risk than real data in terms of the three privacy metrics (**Figure 5.6**). This observation quantitatively confirmed the advantage of sharing synthetic data instead of real data. In general, EMR-WGAN posed the highest risks in all privacy metrics across the two real datasets. In terms of the adversarial concerns of membership inference and meaningful identity disclosure,

Baseline posed the lowest risk in general, except that DPGAN had the lowest meaningful identity disclosure risk for the UW dataset. Regarding the attribute inference risk, multiple models corresponded to a similarly low risk, medGAN and medBGAN for the VUMC dataset, and WGAN, medGAN, and Baseline for the UW dataset. These imply that the ranks of models can be different for different real datasets.

We summarized model scores based on the ranking strategy along individual metrics of data utility and privacy (**Table 5.4**). We observed clear evidence of the privacy-utility trade-off in both real datasets. When a generative model is associated with a set of higher data utility scores, such as EMR-WGAN, it usually corresponds to a set of lower scores for privacy; the opposite phenomenon can also be observed on the Baseline for the VUMC dataset and DPGAN for the UW dataset. Other models, which were ranked middle for data utility, often associated with middle privacy ranks. This phenomenon clearly holds - though not linearly - in this benchmarking study.

Dataset	Metric	Model					
		Baseline	medGAN	medBGAN	EMR-WGAN	WGAN	DPGAN
VUMC	Dimension-wise distribution	***2.0*** (1)	***16.3*** (6)	11.0 (4)	6.7 (3)	6.3 (2)	14.7 (5)
	Column-wise correlation	***17.0*** (6)	12.7 (5)	10.0 (3)	***3.3*** (1)	3.7 (2)	10.3 (4)
	Latent cluster analysis	***17.0*** (6)	12.3 (5)	7.0 (3)	***3.7*** (1)	5.0 (2)	12.0 (4)
	Model performance (TSTR)	***15.0*** (6)	9.0 (4)	13.7 (5)	***3.3*** (1)	7.3 (2)	8.7 (3)
	Model performance (TRTS)	***17.0*** (6)	6.0 (2)	10.3 (4)	***3.3*** (1)	7.3 (3)	13.0 (5)
	Feature selection	***15.3*** (6)	7.7 (3)	10.0 (5)	8.7 (4)	***3.3*** (1)	3.7 (2)
	Clinical knowledge violation	***17.0*** (6)	9.0 (3)	10.0 (4)	6.0 (2)	***2.0*** (1)	13.0 (5)
	Attribute inference	8.7 (4)	***4.0*** (1)	5.0 (2)	***16.0*** (6)	14.7 (5)	8.0 (3)
	Member inference	***1.0*** (1)	9.0 (3)	10.3 (4)	***17.0*** (6)	12.0 (5)	5.7 (2)
	Meaningful identity disclosure	***2.0*** (1)	5.3 (2)	10.3 (4)	***16.7*** (6)	13.3 (5)	9.0 (3)
UW	Dimension-wise distribution	***2.0*** (1)	12.3 (4)	8.0 (3)	5.0 (2)	12.7 (5)	***17.0*** (6)
	Column-wise correlation	8.0 (3)	10.3(4)	6.3 (2)	***2.0*** (1)	13.3 (5)	***17.0*** (6)
	Latent cluster analysis	13.0 (5)	10.3 (4)	5.7 (2)	***2.0*** (1)	9.0 (3)	***17.0*** (6)
	Model performance (TSTR)	***16.0*** (6)	10.0 (4)	9.0 (3)	***2.0*** (1)	5.7 (2)	14.3 (5)
	Model performance (TRTS)	***17.0*** (6)	11.0 (5)	9.7 (3)	***3.7*** (1)	5.7 (2)	9.7 (3)
	Feature selection	10.7 (4)	10.7 (4)	6.0 (2)	***1.7*** (1)	9.3 (3)	***15.3*** (6)
	Clinical knowledge violation	***16.3*** (6)	11.7 (4)	6.7 (2)	***2.0*** (1)	13.0 (5)	7.3 (3)
	Attribute inference	7.0 (3)	5.3 (2)	11.0 (4)	***17.0*** (6)	***2.7*** (1)	14.0 (5)
	Member inference	***1.0*** (1)	9.0 (3)	10.3 (4)	***16.7*** (6)	11.7 (5)	***1.0*** (1)
	Meaningful identity disclosure	6.7 (2)	7.0 (3)	10.0 (4)	15.0 (5)	***15.7*** (6)	***2.0*** (1)

Table 5.4 Rank under each metric.

The rank-derived scores for models with respect to the individual metrics. The best (i.e., lowest rank) and worst scores for each metric are marked as bold red and bold black, respectively. Tied ranks occurred in the privacy risk evaluation if the risk difference on a privacy metric between two synthetic datasets was less than 10^{-5}.

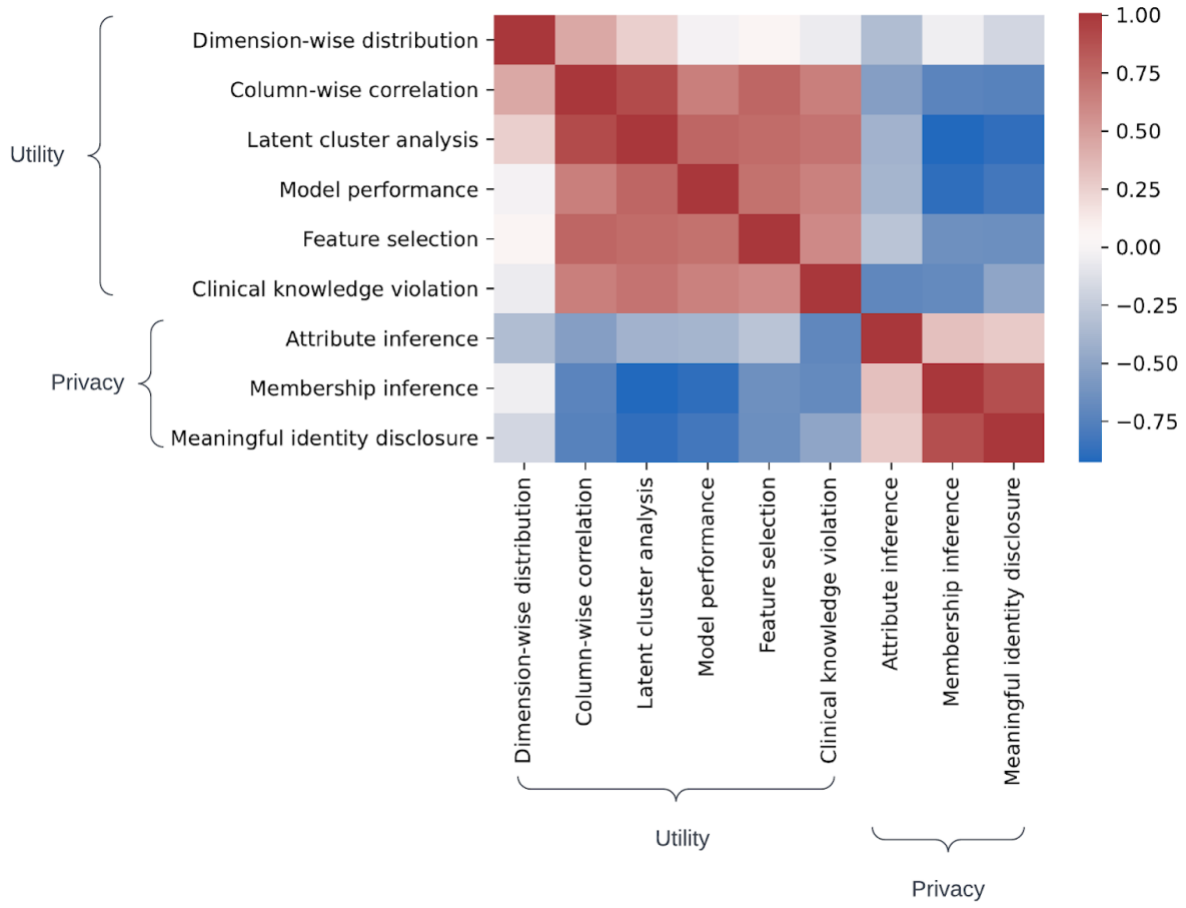


Figure 5.7 Correlation heatmap.

The heatmap of Pearson correlation coefficients of pairwise metrics on the rank averages across all candidate models.

To quantitatively investigate the correlation between different benchmarking metrics, a heatmap was generated using the rank-derived score under each metric (in **Table 5.4**) across all models using both the VUMC and UW datasets (**Figure 5.7**). The privacy-utility trade-off was evidently verified on a granular scale because all pairs of utility and privacy metrics were negatively correlated. Under the scope of utility metrics, a strong correlation was found between column-wise

correlation and latent cluster analysis, suggesting that if a GAN-based model demonstrated a high ability to retain correlations between variables in real data, it might be able to capture the joint distribution of real data as well, and vice versa. Complementary relationships (less strong correlation) were demonstrated between the rest of the utility metric pairs. In particular, the performance of generative models on dimension-wise distribution had very weak correlations with metrics for model prediction and clinical knowledge violation. Regarding privacy concerns, model rankings around membership inference were strongly correlated with meaningful identity disclosure, whereas attribute inference was weakly correlated with the other two privacy metrics.

5.4.2 Model Selection in the Context of Use Cases

GAN-based benchmarking makes sense only when it is contextualized through specific synthetic data use cases, where the users of synthetic datasets have their own preferences and priorities regarding which data characteristics to remain. In other words, the prioritization of benchmarking metrics for different use cases can differ. In this study, we considered three use cases that can benefit from the dissemination of synthetic EHR datasets -educational purposes, DREAM challenges, and system development - to illustrate the process and results of the use-case specific decision making for model selection. We carefully adjusted the weights assigned to the ranking results for individual metrics in each of the given use cases. See section *5.3.6* for a detailed description and justification of our weight setting and approach.

Use case	Dataset	Final ranks of models					
		1	2	3	4	5	6
Educational	VUMC	WGAN (18.00)	EMR-WGAN (21.35)	medBGAN (30.30)	DPGAN (32.20)	Baseline (32.75)	medGAN (33.45)
	UW	EMR-WGAN (14.55)	medBGAN (22.75)	Baseline (26.55)	medGAN (31.65)	WGAN (33.05)	DPGAN (40.25)
EHR DREAM challenges	VUMC	WGAN (23.75)	DPGAN (25.05)	EMR-WGAN (25.25)	medGAN (25.95)	medBGAN (32.25)	Baseline (34.10)
	UW	EMR-WGAN (19.10)	medBGAN (25.55)	WGAN (26.35)	medGAN (28.40)	Baseline (31.90)	DPGAN (35.80)
System design	VUMC	Baseline (19.53)	WGAN (27.95)	medBGAN (28.68)	medGAN (29.02)	DPGAN (29.48)	EMR-WGAN (33.58)
	UW	Baseline (18.43)	medBGAN (26.72)	medGAN (27.87)	EMR-WGAN (29.53)	DPGAN (31.90)	WGAN (32.05)

Table 5.5 Final ranks of generative models in the context of use cases.

Model ranks were determined by the final scores (in parentheses) from the benchmarking framework.

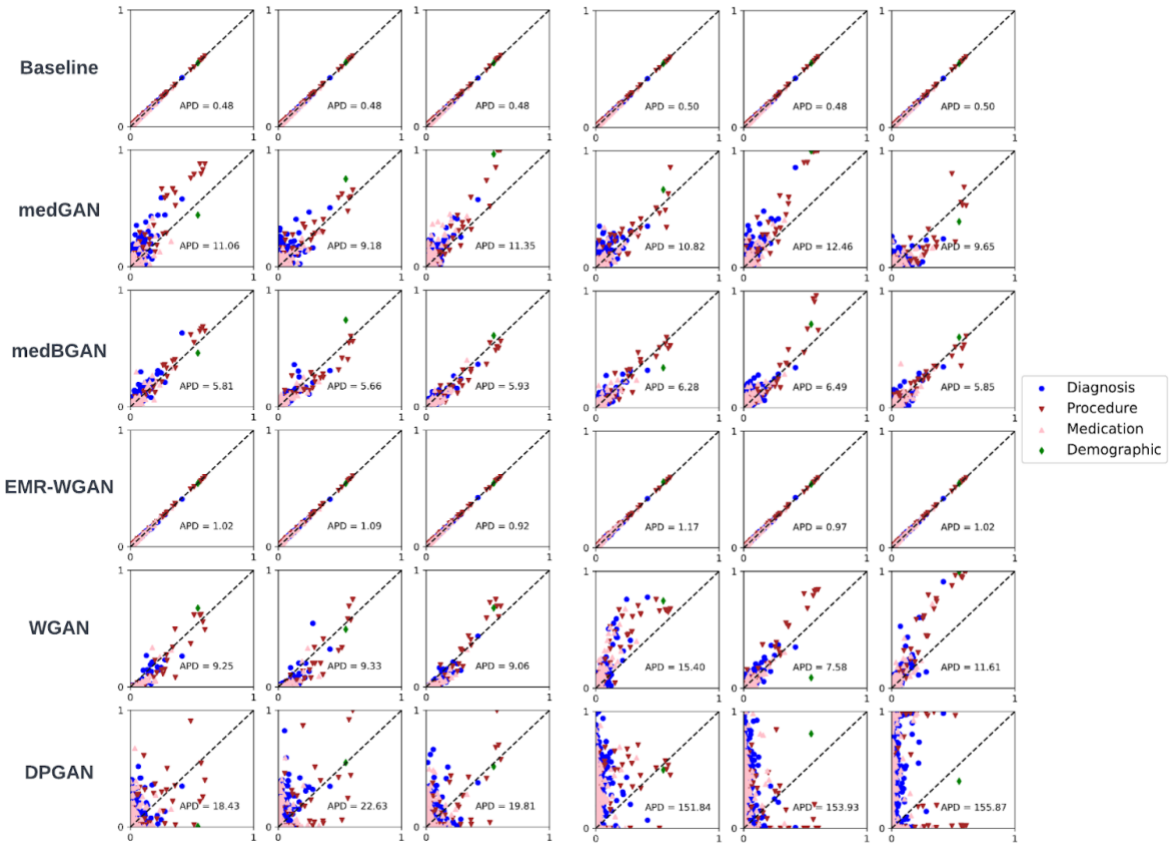
For educational purposes, we emphasized the accuracy of clinical knowledge present in the synthetic data while relatively relaxed the concerns around privacy. EMR-WGAN is the best model for this use case under the UW dataset while the best one for VUMC is WGAN.

The privacy and utility for downstream machine learning analysis are highlighted in DREAM challenges. Under this use case, EMR-WGAN has a superb performance on the UW dataset but the best one for VUMC data is WGAN.

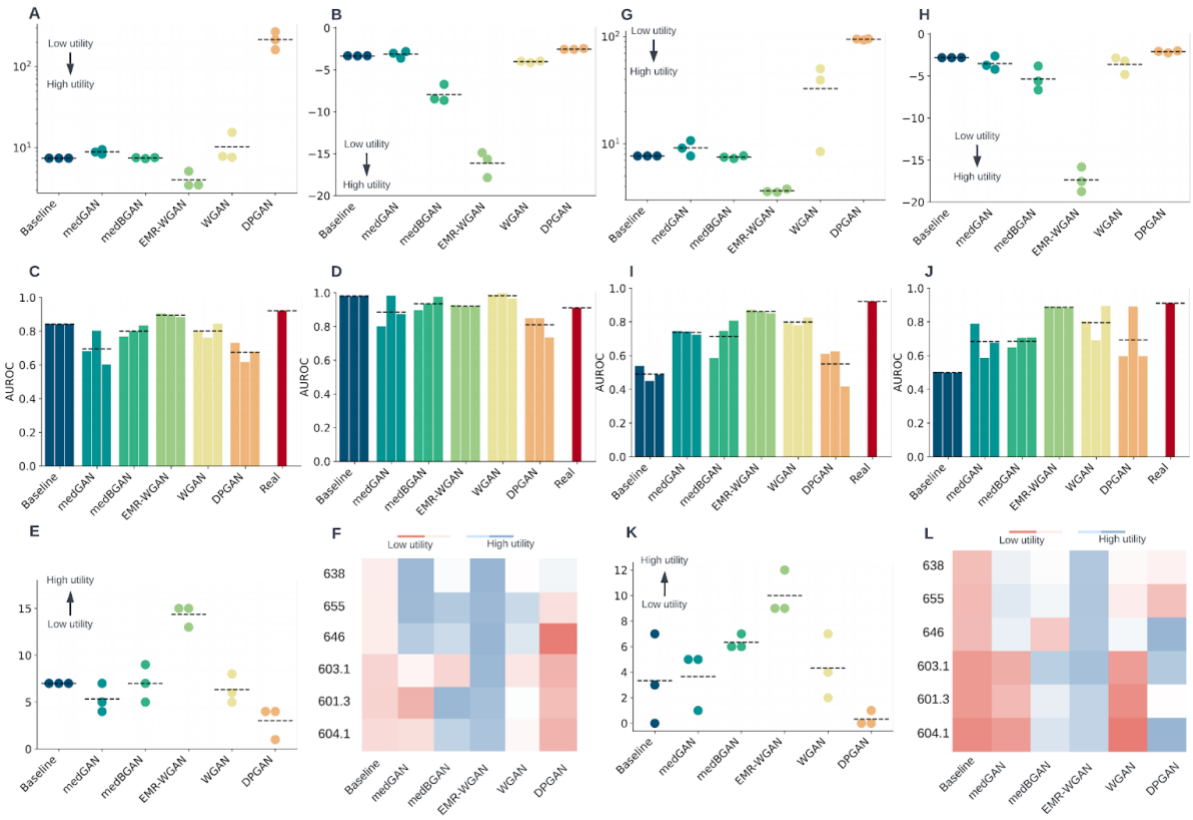
For the use case of system development, the first-order distribution and privacy of the synthetic datasets are emphasized, while the weights for the remaining tests are correspondingly reduced. Under this use case, the sampling-based baseline model is consistently the best for both the UW and VUMC datasets. For detailed rankings, see **Table 5.5**.

5.4.3 Variations on Generation Strategies

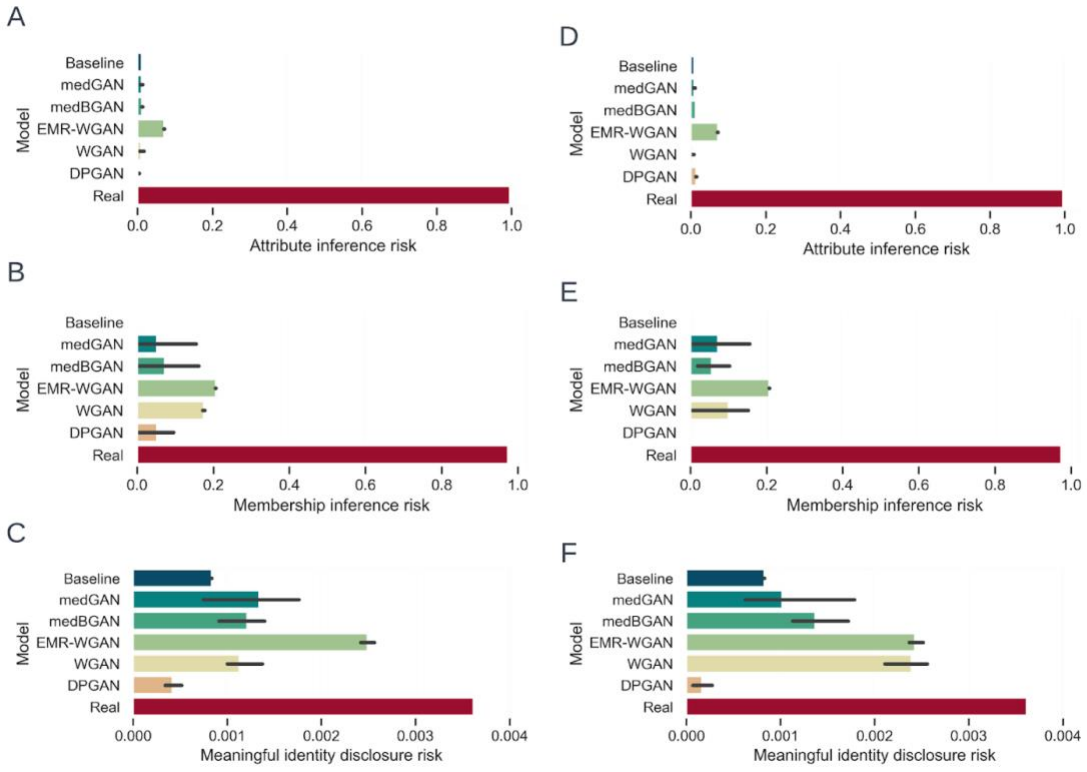
As mentioned in *5.3.5*, we investigated two distinct synthesis paradigms for generating synthetic datasets - a combined synthesis paradigm and a separate synthesis paradigm. We compared the two paradigms on the UW dataset, see the results in **Figure 5.8 A-C**. We followed the convention to ensure that the synthesized data shared the same size as the corresponding real dataset and that the distribution of the outcome variable remained the same as well.


Baseline
medGAN
medBGAN
EMR-WGAN
WGAN
DPGAN
APD = 0.48
APD = 0.48
APD = 0.48
APD = 0.50
APD = 0.48
APD = 0.50
APD = 11.06
APD = 9.18
APD = 11.35
APD = 10.82
APD = 12.46
APD = 9.65
APD = 5.81
APD = 5.66
APD = 5.93
APD = 6.28
APD = 6.49
APD = 5.85
APD = 1.02
APD = 1.09
APD = 0.92
APD = 1.17
APD = 0.97
APD = 1.02
APD = 9.25
APD = 9.33
APD = 9.06
APD = 15.40
APD = 7.58
APD = 11.61
APD = 18.43
APD = 22.63
APD = 19.81
APD = 151.84
APD = 153.93
APD = 155.87
Diagnosis
Procedure
Medication
Demographic

A


A
B
G
H
C
D
I
J
E
F
K
L
Low utility
High utility
AUROC
Baseline
medGAN
medBGAN
EMR-WGAN
WGAN
DPGAN
Real
638
655
646
603.1
601.3
604.1

B



C

Figure 5.8 Comparison of two synthesis paradigm using UW data.

(A)Dimension-wise distribution for the UW synthetic data using a separate synthesis paradigm(left) and the UW synthetic data using a combined synthesis paradigm(right). (B)Data utility (except for dimension-wise distribution) from using a separate synthesis paradigm (left, A-F) and a combined synthesis paradigm (right, G-L) for the UW datasets. (C). Privacy risks for the UW synthetic data using a separate synthesis paradigm(A-C) and UW synthetic data using a combined synthesis paradigm(D-F).

The separate synthesis paradigm leads to synthetic data with better utility in downstream machine learning model development while not compromising privacy. Under all the three use cases, the separate synthesis paradigm outperforms the combined synthesis paradigm for all 6 models with one exception: for the educational use case, the combined synthesis paradigm leads to better performance than the separate synthesis paradigm for DPGAN.

5.5 Discussion

Generalizable synthetic EHR benchmarking framework for unbiased and comprehensive evaluation

We implemented the benchmarking framework in the context of two institutional enterprise data warehouses and demonstrated how to evaluate the quality of different synthesis approaches under the framework through quantitative utility and privacy measures. We demonstrated that the framework is agnostic to data types, quantities, duration, and patient cohort types, and can thus be applied to inform the development of new synthesis algorithms and conduct comparative studies. The flexibility and generalizability of the framework are conducive to our two-site study and have led to discoveries that might have been missed in a single-site study. We observed performance discrepancies between the UW and VUMC data; for instance, WGAN and DPGAN did not generalize well to a larger dataset like UW. We think this performance discrepancy could be attributed to the dataset size difference between the two sites, as reducing the training data size by the separate synthesis paradigm leads to better first-order distribution resemblance (**Figure 5.8** A). This framework can be scaled up to benchmark models on multiple data sites. We are interested in ramping up this pilot study to a crowdsourced challenge.

Adjustable weights for model comparison and interpretation using output from the benchmarking framework

Our framework acknowledges the utility-privacy trade-off and provides examples to assess different synthetic data generation approaches through the lens of specific use cases. Through adjusting the weights assigned to each test, framework users can derive a model ranking that is tailored to their specific needs, as demonstrated in our use case session. Currently, our framework assigns weights based on the assumption that utility and privacy matter equally unless weight adjustment is applied. In contrast, most of the current studies did not hold the same assumption; for example, the original EMR-WGAN paper placed greater weight on utility based on the observation that the EMR-WGAN's privacy risk, though higher than some other GAN models, is still much lower than the real data. This provokes a different idea for interpreting the benchmarking results: we can improve the framework by introducing a threshold for privacy tests. If the risk for all candidate models is below the risk threshold, we can lower the weight assigned to the privacy test for model selection.

Chapter 6

CONCLUSIONS

In this chapter, I will summarize the main contribution of each Aim and acknowledge the limitations and opportunities for future work.

6.1 Summary of Contribution

EHR data contain rich information about patients' clinical trajectories, which can be used to enable personalized medicine and improve healthcare quality. However, the private and sensitive nature of EHR data prevents direct sharing and utilization of EHR data with the broader data science community. This dissertation focuses on approaches to facilitate the secure sharing and utilization of private datasets for developing ML algorithms. I investigated how the 'model to data' approach can better engage the data science community through data challenges to address clinically relevant questions using both structure and unstructured clinical data. Through the analysis of the hundreds of ML models received from community challenges, I identified the advantages of ML models to aid in clinical decision-making but also highlighted existing limitations and biases that predictive models can present. The 4 Aims of my study contributed to the secure utilization of clinical data and to the critical analysis and benchmarking of ML models in the clinical setting.

6.1.1 Aim 1 Summary

I showcased the 'model to data' approach as a new mechanism to make private clinical data available for the development of predictive models. Under this framework, researchers' direct interaction with patient data was replaced with indirect interaction via containerized models.

The 'model to data' framework was operationalized using the Synapse collaboration platform and an on-premises secure computing environment at UW hosting EHR data. Containerized mortality prediction models - developed by a model developer - were delivered to the University of Washington via Synapse, where the models were trained and evaluated. Model performance metrics were returned to the model developer.

The model developer was able to develop three mortality prediction models under the 'model to data' framework using simple demographic features (AUROC, 0.693), demographics and five common chronic diseases (AUROC, 0.861), and the 1000 most common features from the EHR's condition/procedure/drug domains (AUROC, 0.921).

My contribution to aim 1:

(1) Worked as a ML model developer, with no direct data access, to successfully develop accurate mortality prediction models using both data-driven and engineer features extracted from hidden EHR data.

(2) Identified crucial elements needed for the successful implementation of the 'model to data' approach. Pointed out challenges that both the model developer and the health system

information technology group encountered and proposed future efforts to improve implementation.

(3) Developed and conducted tests of infrastructure stability and robustness and identified computational resources needed for scaling up the pilot study from an individual participant to a community challenge.

This work has been published as a peer-reviewed journal paper. (https://academic.oup.com/jamia/article/27/9/1393/5868591)

6.1.2 Aim 2 Summary

In the interest of streamlining the response to the COVID-19 pandemic, ML was applied to predict the likelihood of diagnosis and severity of illness. However, the lack of COVID-19 patient data has hindered the data science community in developing models to aid in the response to the pandemic.

I operated a continuous, crowdsourced challenge using a 'model to data' approach to enable participants to securely use regularly updated COVID-19 patient data from UW from May 6 to December 23, 2020. A post-challenge analysis was conducted from December 24, 2020 to April 7, 2021 to assess the generalizability of models on the cumulative data set as well as subgroups stratified by age, sex, race, and time of COVID-19 test. By December 23, 2020, this challenge engaged 482 participants from 90 teams and 7 countries.

In the analysis using the cumulative data set, the best performance for COVID-19 diagnosis prediction was an AUROC of 0.776 (95% CI, 0.775-0.777) and an AUPRC of 0.297, and for hospitalization prediction, an AUROC of 0.796 (95% CI, 0.794-0.798) and an AUPRC of 0.188.

Analysis on top models submitted to the challenge showed consistently better model performance on the female group than the male group. Among all age groups, the best performance was obtained for the 25- to 49-year age group and the worst performance was obtained for the group aged 17 years or younger.

Models submitted by citizen scientists achieved high performance for the prediction of COVID-19 testing and hospitalization outcomes. Evaluation of challenge models on demographic subgroups and prospective data revealed performance discrepancies, providing insights into the potential bias and limitations of the models.

My contribution to aim 2:

(1) Leveraged the 'model to data' approach from a pilot study to a crowdsourced community challenge and enabled the broad data science community to build COVID-19 prediction models without risking the privacy of COVID-19 patients.
(2) Evaluated top-performing models from participants around the world on temporally evolving COVID-19 patient datasets and different demographic subgroups to shed light on the behavior of ML models on clinical data.
(3) Highlighted the need to investigate the bias and limitations present in high-performing ML models.

This work has been published as a peer-reviewed journal paper. (https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2784779)

6.1.3 Aim 3 Summary

The evaluation of NLP models for clinical text de-identification relies on the availability of clinical notes, which are often restricted due to privacy concerns. The NLP Sandbox is an approach for alleviating the lack of data and evaluation frameworks for NLP models by adopting a federated, 'model to data' approach. This approach enabled unbiased federated model evaluation without sharing sensitive data from multiple institutions.

The NLP Sandbox was built by leveraging the Synapse collaborative platform, containerization software, and OpenAPI generator. I evaluated two state-of-the-art NLP de-identification models using data from three institutions and further validated model performance using data from an external validation site. An external developer was able to incorporate their model into the NLP Sandbox template and provide user experience feedback. A multi-site evaluation of clinical text de-identification models without the sharing of data was conducted to showcase the feasibility of the NLP Sandbox to support federated evaluation. Standardized models and data schemas enable smooth model transfer and implementation.

My contribution to aim 3:

(1) Demonstrated application of the 'model to data' approach for enabling multi-site clinical note utilization by an external researcher without direct data access.

(2) Benchmarked state-of-the-art NLP de-identification models on multi-site data and conducted model generalizability test through federated evaluation.

(3) Identified crucial elements needed for lowering the participation barrier for model developers and health institutions and for generalizing the NLP sandbox to support tasks other than clinical text de-identification.

This work is currently under peer review.

6.1.4 Aim 4 Summary

Broad and reliable data sharing is one of the key boosters for the advancement of biomedicine and healthcare, and it is often accompanied by rigorous assessment and tuning procedures to ensure that data privacy is sufficiently protected. Instead of operating on real data, generating fully synthetic health data has recently emerged in conjunction with the advances in ML and has become an active research area to guide data sharing practice. GAN models have demonstrated success in maintaining the utility of real data without compromising their privacy. However, there is currently a lack of a comprehensive assessment framework and a benchmarking exercise for GAN model evaluation.

This study aimed to bridge this gap by benchmarking - in the same learning and assessment environment - the published synthetic EHR generators that simulate static patient profiles. I reported on 1) the assessment framework that covers multiple important metrics of data utility and privacy risks, and 2) the model benchmarking results using the datasets from two well-defined studies: the COVID-19 admission study from VUMC and UW, 3) benchmarking result interpretation and model recommendation based on specific use cases.

My contribution to aim 4:

(1). Established a comprehensive and generalizable framework for unbiased evaluation of state-of-the-art GAN-based methods.

(2). Demonstrate the flexibility and generalizability of the framework by conducting utility and privacy assessments on several GAN-based methods using datasets from two institutions.

(3). Shed light on model interpretation and comparison under various use cases using output from the evaluation framework

This work is currently under submission.

6.2 Limitations

6.2.1 Aim 1 and 2 Limitations

The goal of Aim 1 and Aim 2 is to implement and scale up the 'model to data' framework to support a crowdsourced EHR community challenge. At the end, close to 500 participants all around the world registered for the challenge and submitted hundreds of models for benchmarking. 'Model to data' framework lowered the barrier to data utilization through bypassing direct data access. Under the framework, participants who otherwise did not have suitable data could develop clinical predictive models on real patient data and receive model performance feedback. However, participants who were not researchers in the EHR domain (such as software engineers from tech companies or college students) had to go through a learning curve to understand the clinical common data model and code system. Synthetic data were provided for them to learn the format but because the synthetic data was not trained on UW patient cohorts, it was not representative of

the data used for model benchmarking. Participants were discouraged from using performance on synthetic data to inform model development during the DREAM challenges. A high-quality synthetic dataset could lower the challenge participation threshold and ease the pain of developing models without seeing the data.

Due to privacy concerns, the only information returned to participants was the performance metrics - AUROC and AUPRC. If a model experienced an error, a challenge organizer needed to pull out the log file and manually filter out PHI before sharing the logs with the developer. This placed a burden on challenge organizers. An automatic PHI-filtering and log-returning system is necessary to scale up future EHR challenges.

In aim 1 and aim 2, the datasets were split in a temporal order to mimic model implementation where models trained on previous data were applied to relatively recent data to make predictions. However, this approach has limitations as the dataset was evolving over time, and the previous data used for training might not be representative of the new evaluation data. For instance, the dominating COVID-19 variant was changing through the course of the pandemic leading to different symptoms and illness severity. The model trained on previous data might fail to catch important features that are crucial for prediction related to COVID-19 caused by a different variant. This could be remediated by curating specific training datasets related to each variant. But the variant information was not available in the challenge dataset.

In terms of model benchmarking, utility metrics, such as performance score, were priortized, but future EHR challenges will benefit from a wide spectrum of evaluation criteria including model robustness and fairness.

6.2.2 Aim 3 Limitations

To generalize the NLP Sandbox, work is still required on the part of data owners and model developers to develop suitable and standardized schemas and to adapt their data or model to fit the schemas.

I identified several aspects of the NLP Sandbox with room for improvement:

(1) Model template. The external model developer reported difficulties with incorporating model scripts into the NLP Sandbox. The Sandbox could benefit from additional detailed instructions and video tutorials to help developers incorporate their models into the provided template. In particular, instructions for wrapping models as packages, importing custom packages into the template, and making annotation outputs compatible with the NLP Sandbox data schema would be helpful.

(2) Data schema. The current data schema cannot suit all evaluation needs, and the involvement of the model developing community is needed for defining standardized schemas. Currently, NLP Sandbox implemented the schema used for the 2014 i2b2 challenge. For models like NeuroNER that were designed with i2b2 schemas in mind, they naturally fit and transitioned smoothly into the NLP Sandbox. On the other hand, Philter was not developed with the i2b2 data schema in mind, and instead prioritized identifying PHI regardless of the category and tolerated duplicated annotations to maximize recall. As a result, Philter's performance suffered in the NLP Sandbox. For example, Philter was penalized for capturing “David”, “Smith” and “David Smith” as separate

instances of PHI, and for identifying "David Smith" as both a location and a name. The flexibility of NLP Sandbox makes it easy to add separate evaluation queues that can be used to answer different questions, such as PHI category-agnostic performance in the future. Additionally, the decision to implement a common data schema enables data standardization but puts the burden on data owners to adapt their data to the selected schema. One future direction would be to allow data providers to specify the data schema they used to annotate the dataset and then tailor the NLP Sandbox evaluation to the unique schema of each dataset. Alternatively, it would be ideal to automate the generation of multiple data schemas for each dataset and users could compare model performance across different schemas.

(3) Model training. While model training is not currently enabled in the NLP Sandbox, rule-based and pre-trained NLP models are accepted for evaluation. From the NeuroNER experiment, I observed that training and testing models using data from the same site can lead to better performance, especially for PHI categories like location and ID where each site has its own format and specification. However, pre-trained models often carry fragments of the data that they are trained on. To achieve model training inside the NLP Sandbox, additional data transfer agreements and security measures must be implemented to ensure that models trained on sensitive datasets do not leak PHI in other data nodes.

(4) Data quality. While I applied data quality checks on the i2b2, MCW, Mayo, and UW datasets, it is possible that the gold standard annotations used for evaluation are not 100% correct or are not representative of authentic clinical notes. For example, I observed some quality issues in the

synthetic Mayo data (e.g., unusually long patient names), leading to low annotation performance for NeuroNER. Rigorous data curation and quality control can further improve the NLP Sandbox

6.2.3 Aim 4 Limitations

Aim 4 project established a comprehensive and flexible evaluation framework for GAN-based synthetic data generation approaches. I observed a consistent pattern of the separate-generation approach outperforming the combined generation approach on the UW dataset across different models under multiple use cases. However, extensive experiments on datasets from different sites are needed to test the generalizability of this conclusion.

The models' architectures were not modified for this benchmarking but some degree of parameter tuning was involved to adapt the models to the VUMC and UW datasets. It is observed that some models were more sensitive than others to parameter changes. Efforts were made to ensure the performance of each model, but limitations were still present as the data used in this benchmarking study were not the same as the original data involved in the model development. A community challenge for benchmarking synthetic data could resolve this problem by leaving the model adjustment in the hands of the original developers.

Some models were more unstable in the generation process than others, so we repeated the model training processes five times for each model and included the best 3 for each model in this analysis. But model robustness is currently not a criterion under the current evaluation framework due to the limitation of time and computational resources. In the future study, extensive experiments should be incorporated to investigate further the model stability and output consistency.

A ranking approach was applied to facilitate the combination of evaluation results across different tests under the framework, however, a limitation of the rank approach is the actual performance difference among the models was made invisible. Currently three use cases were provided to demonstrate how to interpret the results from the benchmarking frameworks based on specific application needs. But the example provided did not cover all the possible use cases and the weight setting for each use case was ad-hoc. Future users of the evaluation framework should adjust weights based on their requirements.

6.3 Future Work

6.3.1 Next-generation 'model to data' crowdsourced EHR challenges

'Model to data' crowdsourced challenges have proven effective for engaging the broader data science community and accelerating citizen science. Limitations exist in the current approach as mentioned in 6.2. Future EHR challenges can broaden and deepen their impact in several ways (1) **By leveraging multimodal EHR data.** Several EHR challenges covered in this dissertation (Aim 1-3) utilized data of a single type: either structured or unstructured. However, EHR data in different formats often contain complementary information about patients. Integrating multimodal data can lead to better clinical prediction; (2) **By operating continuously with real-time data feed and ensemble results.** As demonstrated in Aim 2, I achieved the goal of operating a community challenge for predicting COVID-19 patient outcomes in a continuous mode with regularly updated datasets. This can be further improved by allowing submissions to run on a live data stream and generating real-time ensemble prediction results by aggregating individual models; (3). **By enabling federated training and evaluation.** Federated learning has emerged as a promising

strategy to cope with poor model generalizability. As demonstrated in Aim 3, I achieved federated evaluation of NLP models on four data sites. The next step is to enable federated training and incorporate model generalizability as an evaluation metric for benchmarking; (4) **By providing challenge participants with high-quality synthetic data.** The nature of the 'model to data' approach reduces researchers' direct interaction with real data and thus raises the barrier to model development. Synthetic data can be shared with participants and used to derive actionable in, sights for model development and inform model tuning processes. In aim 4, I was able to benchmark which synthetic algorithms do best for this use case.

6.3.2 Implementation-oriented Challenge Design

Implementing Challenge submissions in a clinical setting is beyond the scope of this dissertation, but it is a critical direction for future work to explore. In that case, it is necessary to identify implementation sites prior to the Challenge, so that we can ensure data examples (e.g., synthetic data) provided to participants are in the same format and that models selected from the Challenge are compatible with the clinical workflow in the implementation sites. Evaluation metrics used in the Challenge need to be tailored to suit implementation needs. Potential evaluation metrics include: (1) Computational cost. This may refer to a cost in terms of time, money, computational resources, or some combination of them. (2). Scalability. This refers to a model's capacity to produce predictions in a given period of time. (3). Stability. This refers to a model's ability to produce the same prediction deterministically.

6.3.3 From Challenge to Implementation

Outside the scope of crowdsourced challenges, the work described in this dissertation can contribute to a broader general vision of clinical model implementation. Well-distributed synthetic

data with privacy control can ease the pain of restricted data access and streamline model development. Thorough examinations on temporal performance degradation and model bias are acutely in need before incorporating models into the clinical decision support. The future will benefit from the seamless and real-time dissemination and benchmarking of models on structured and unstructured clinical data within a large network of health care providers, model developers, and research scientists.

BIBLIOGRAPHY

1 Charles D, Gabriel M, Ma TSM. Adoption of electronic health record systems among U.s. non- federal acute care hospitals: 2008-2014. https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf (accessed 22 Apr 2021).

2 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.

3 Birkhead GS, Klompas M, Shah NR. Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. *Annu Rev Public Health* 2015;**36**:345–59.

4 Heisey-Grove D, Patel V. Physician motivations for adoption of electronic health records. https://hitconsultant.net/wp-content/uploads/2014/12/oncdatabrief-physician-ehr-adoption-motivators-2014.pdf (accessed 22 Apr 2021).

5 Si Y, Du J, Li Z, *et al.* Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *J Biomed Inform* 2020;:103671.

6 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.

7 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;**1**:18.

8 Gao Y, Cai G-Y, Fang W, *et al.* Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun* 2020;**11**:5033.

9 Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med* 2021;**4**:3.

10 Jung K, Kashyap S, Avati A, *et al.* A framework for making predictive models useful in practice. *J Am Med Inform Assoc* 2021;**28**:1149–58.

11 Avati A, Jung K, Harman S, *et al.* Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;**18**:122.

12 Assaf D, Gutman Y 'ara, Neuman Y, *et al.* Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Intern Emerg Med* 2020;**15**:1435–43.

13 Carrara M, Baselli G, Ferrario M. Mortality Prediction Model of Septic Shock Patients Based on Routinely Recorded Data. *Comput Math Methods Med* 2015;**2015**:761435.

14 Schaffter T, Buist DSM, Lee CI, *et al.* Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open* 2020;**3**:e200265.

15 Miotto R, Li L, Kidd BA, *et al.* Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016;**6**:26094.

16 Rajpurkar P, Chen E, Banerjee O, *et al.* AI in health and medicine. *Nat Med* 2022;**28**:31–8.

17 He J, Baxter SL, Xu J, *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;**25**:30–6.

18 Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. *Journal of Big Data* 2018;**5**:1.

19 Garfinkel SL. De-identification of personal information. National Institute of Standards and Technology 2015. doi:10.6028/NIST.IR.8053

20 Reps JM, Schuemie MJ, Suchard MA, *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018;**25**:969–75.

21 Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nat Biotechnol* 2018;**36**:391–2.

22 Yang M, Petralia F, Li Z, *et al.* Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. *Cell Syst* 2020;**11**:186–95.e9.

23 Home. OpenAPI Initiative. 2016.https://www.openapis.org/ (accessed 27 Jun 2022).

24 Norgeot B, Muenzen K, Peterson TA, *et al.* Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit Med* 2020;**3**:57.

25 Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: : Association for Computational Linguistics 2017. 97–102.

26 Charles D, Gabriel M, Searcy T, *et al.* Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2014. *ONC data brief* 2015;**23**.https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf

27 Jones SS, Rudin RS, Perry T, *et al.* Health information technology: an updated systematic

review with a focus on meaningful use. *Ann Intern Med* 2014;**160**:48–54.

28 Kaji DA, Zech JR, Kim JS, *et al.* An attention based deep learning model of clinical events in the intensive care unit. *PLoS One* 2019;**14**:e0211057.

29 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.

30 Klann JG, Joss MAH, Embree K, *et al.* Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PLoS One* 2019;**14**:e0212463.

31 Malin B, Sweeney L, Newton E. Trail re-identification: learning who you are from where you have been. *Proc LIDAP-WP12* Published Online First: 2003.https://dataprivacylab.org/dataprivacy/projects/trails/trails1.pdf

32 Desautels T, Calvert J, Hoffman J, *et al.* Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform* 2016;**4**:e28.

33 Choi E, Biswal S, Malin B, *et al.* Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv [cs.LG]. 2017.http://arxiv.org/abs/1703.06490

34 Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:160035.

35 Foraker R, Mann DL, Payne PRO. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* 2018;**3**:716–8.

36 Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. *AMIA Annu Symp Proc* 2011;**2011**:1176–85.

37 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* Published Online First: 30 August 2017. doi:10.1093/jamia/ocx079

38 Enterprise Container Platform | Docker. Docker. https://www.docker.com/ (accessed 9 Aug 2019).

39 Singularity. Sylabs.io. https://sylabs.io/ (accessed 18 Nov 2019).

40 Ellrott K, Buchanan A, Creason A, *et al.* Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol* 2019;**20**:195.

41 Ge W, Huh J-W, Park YR, *et al.* An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. *AMIA Annu Symp Proc* 2018;**2018**:460–9.

42 Saez-Rodriguez J, Costello JC, Friend SH, *et al.* Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* 2016;**17**:470–86.

43 Omberg L, Ellrott K, Yuan Y, *et al.* Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* 2013;**45**:1121–6.

44 Lambert CG, Amritansh, Kumar P. Transforming the 2.33M-patient Medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete. Published Online First: 23 September 2016.http://dx.doi.org/ (accessed 9 Aug 2019).

45 Weng SF, Vaz L, Qureshi N, *et al.* Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* 2019;**14**:e0214365.

46 Fleurence RL, Curtis LH, Califf RM, *et al.* Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014;**21**:578–82.

47 Murphy SN, Weber G, Mendis M, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010;**17**:124–30.

48 Radivojac P, Clark WT, Oron TR, *et al.* A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7.

49 Jiang Y, Oron TR, Clark WT, *et al.* An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 2016;**17**:184.

50 Cai B, Li B, Kiga N, *et al.* Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. *Hum Mutat* 2017;**38**:1266–76.

51 Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 2005;**15**:285–9.

52 He X, Zhao K, Chu X. AutoML: A Survey of the State-of-the-Art. arXiv [cs.LG]. 2019.http://arxiv.org/abs/1908.00709

53 The New York Times. Covid in the U.S.: Latest Map and Case Count. The New York Times. 2020.https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html (accessed 19 Nov 2020).

54 Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* 2020;**139**:110059.

55 Khakharia A, Shah V, Jain S, *et al.* Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning. *Annals of Data Science* Published Online First: 16 October 2020. doi:10.1007/s40745-020-00314-9

56 Yadav M, Perumal M, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. *Chaos Solitons Fractals* 2020;**139**:110050.

57 Wu J, Zhang P, Zhang L, *et al.* Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *medRxiv* 2020;:2020.04.02.20051136.

58 Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J Epidemiol Community Health* 2020;**74**:861–6.

59 Koetter P, Pelton M, Gonzalo J, *et al.* Implementation and Process of a COVID-19 Contact Tracing Initiative: Leveraging Health Professional Students to Extend the Workforce During a Pandemic. *Am J Infect Control* 2020;**48**:1451–6.

60 Jamshidi M, Lalbakhsh A, Talla J, *et al.* Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. *IEEE Access* 2020;**8**:109581–95.

61 Bergquist T, Yan Y, Schaffter T, *et al.* Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *J Am Med Inform Assoc* 2020;**27**:1393–400.

62 Bergquist T, Schaffter T, Yan Y, *et al.* Evaluation of crowdsourced mortality prediction models as a framework for assessing AI in medicine. *medRxiv* 2021;:2021.01.18.21250072.

63 Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods* 2016;**93**:92–102.

64 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.

65 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.

66 Capurro D, Yetisgen M, van Eaton E, *et al.* Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS (Wash DC)* 2014;**2**:1079.

67 Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13.

68 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;:17–21.

69 Soysal E, Wang J, Jiang M, *et al.* CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018;**25**:331–6.

70 Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Med Inform* 2020;**8**:e17984.

71 Khambete MP, Su W, Garcia JC, *et al.* Quantification of BERT Diagnosis Generalizability Across Medical Specialties Using Semantic Dataset Distance. *AMIA Jt Summits Transl Sci Proc* 2021;**2021**:345–54.

72 Melamud O, Shivade C. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. arXiv [cs.CL]. 2019.http://arxiv.org/abs/1905.07002

73 McLachlan S, Dube K, Gallagher T. Using the CareMap with Health Incidents Statistics for Generating the Realistic Synthetic Electronic Healthcare Record. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. 2016. 439–48.

74 Li J, Zhou Y, Jiang X, *et al.* Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *J Am Med Inform Assoc* 2021;**28**:2193–201.

75 Brekke PH, Rama T, Pilán I, *et al.* Synthetic data for annotation and extraction of family history information from clinical text. *J Biomed Semantics* 2021;**12**:11.

76 Chen X, Li Y, Jin P, *et al.* Adversarial Sub-sequence for Text Generation. arXiv [cs.LG]. 2019.http://arxiv.org/abs/1905.12835

77 Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *J Biomed Inform* 2015;**58 Suppl**:S20–9.

78 Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550–63.

79 Malin B, Sweeney L, Newton E. Trail re-identification: Learning who you are from where you have been. Published Online First: 2003.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.395.2183 (accessed 23 Apr 2021).

80 Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average? *Mol Syst Biol* 2011;**7**:537.

81 DeYoung J, Jain S, Rajani NF, *et al.* ERASER: A Benchmark to Evaluate Rationalized NLP Models. arXiv [cs.CL]. 2019.http://arxiv.org/abs/1911.03429

82 Wang A, Singh A, Michael J, *et al.* GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Published Online First: 27 September 2018.https://openreview.net/forum?id=rJ4km2R5t7 (accessed 12 Jan 2022).

83 SemEval. SemEval. https://semeval.github.io/ (accessed 28 Mar 2022).

84 Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/

(accessed 28 Mar 2022).

85 Yan Y, Schaffter T, Bergquist T, *et al.* A Continuously Benchmarked and Crowdsourced Challenge for Rapid Development and Evaluation of Models to Predict COVID-19 Diagnosis and Hospitalization. *JAMA Netw Open* 2021;**4**:e2124946.

86 Lee K, Dobbins NJ, McInnes B, *et al.* Transferability of neural network clinical deidentification systems. *J Am Med Inform Assoc* Published Online First: 29 September 2021. doi:10.1093/jamia/ocab207

87 Choobdar S, Ahsen ME, Crawford J, *et al.* Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**:843–52.

88 Sieberts SK, Schaff J, Duda M, *et al.* Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. *NPJ Digit Med* 2021;**4**:53.

89 Meyer P, Saez-Rodriguez J. Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. *Cell Syst* 2021;**12**:636–53.

90 Advanced Load Balancer, Web Server, & Reverse Proxy. NGINX. 2015.https://www.nginx.com/ (accessed 11 Apr 2022).

91 Tresp V, Marc Overhage J, Bundschus M, *et al.* Going Digital: A Survey on Digitalization and Large-Scale Data Analytics in Healthcare. *Proc IEEE* 2016;**104**:2180–206.

92 Topol EJ, Steinhubl SR, Torkamani A. Digital medical tools and sensors. *JAMA* 2015;**313**:353–4.

93 Elenko E, Underwood L, Zohar D. Defining digital medicine. *Nat Biotechnol* 2015;**33**:456–61.

94 Packer M. Data sharing in medical research. BMJ. 2018;**360**:k510.

95 Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet* 2020;**52**:646–54.

96 Ghosheh G, Li J, Zhu T. A review of Generative Adversarial Networks for Electronic Health Records: applications, evaluation measures and data sources. arXiv [cs.LG]. 2022.http://arxiv.org/abs/2203.07018

97 Chen RJ, Lu MY, Chen TY, *et al.* Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 2021;**5**:493–7.

98 Yang Y, Nan F, Yang P, *et al.* GAN-Based Semi-Supervised Learning Approach for Clinical Decision Support in Health-IoT Platform. *IEEE Access* 2019;**7**:8048–57.

99 Che Z, Cheng Y, Zhai S, *et al.* Boosting Deep Learning Risk Prediction with Generative

Adversarial Networks for Electronic Health Records. In: *2017 IEEE International Conference on Data Mining (ICDM)*. ieeexplore.ieee.org 2017. 787–92.

100 Waheed A, Goyal M, Gupta D, *et al.* CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access* 2020;**8**:91916–23.

101 Hernandez M, Epelde G, Alberdi A, *et al.* Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 2022;**493**:28–45.

102 Lan L, You L, Zhang Z, *et al.* Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Front Public Health* 2020;**8**:164.

103 Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets. *Adv Neural Inf Process Syst* 2014;**27**.https://proceedings.neurips.cc/paper/5423-generative-adversarial-nets

104 OMOP common data model. https://www.ohdsi.org/data-standardization/the-common-data-model/ (accessed 30 May 2022).

105 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*. PMLR 18--19 Aug 2017. 286–305.

106 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.

107 Zhang Z, Yan C, Lasko TA, *et al.* SynTEG: a framework for temporal structured electronic health data simulation. *J Am Med Inform Assoc* 2021;**28**:596–604.

108 CPT Hierarchy. http://medpricemonkey.com/cpt_hierarchy_list (accessed 30 May 2022).

109 Wang K, Gou C, Duan Y, *et al.* Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica* 2017;**4**:588–98.

110 Luo Y, Zhu L-Z, Wan Z-Y, *et al.* Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *J Neural Eng* 2020;**17**:056021.

111 DuMont Schütte A, Hetzel J, Gatidis S, *et al.* Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ Digit Med* 2021;**4**:141.

112 Goncalves A, Ray P, Soper B, *et al.* Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020;**20**:108.

113 Woo M-J, Reiter JP, Oganian A, *et al.* Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *JPC J Planar Chromatogr - Mod TLC* 2009;**1**. doi:10.29012/jpc.v1i1.568

114 Esteban C, Hyland SL, Rätsch G. Real-valued (Medical) Time Series Generation with

Recurrent Conditional GANs. arXiv [stat.ML]. 2017.http://arxiv.org/abs/1706.02633

115 El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J Med Internet Res* 2020;**22**:e23139.

116 Foraker RE, Yu SC, Gupta A, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 2020;**3**:557–66.

117 Guo A, Foraker RE, MacGregor RM, *et al.* The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation. *Frontiers in Digital Health* 2020;**2**. doi:10.3389/fdgth.2020.576945

118 Artzi NS, Shilo S, Hadar E, *et al.* Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* 2020;**26**:71–6.

119 Razavian N, Major VJ, Sudarshan M, *et al.* A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *NPJ Digit Med* 2020;**3**:130.

120 Liu X, Glocker B, McCradden MM, *et al.* The medical algorithmic audit. *Lancet Digit Health* 2022;**4**:e384–97.

121 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;**3**:e745–50.

122 Lundberg, Lee. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

123 Dankar FK, El Emam K. A method for evaluating marketer re-identification risk. In: *Proceedings of the 2010 EDBT/ICDT Workshops*. New York, NY, USA: : Association for Computing Machinery 2010. 1–10.

124 Baowaly MK, Lin C-C, Liu C-L, *et al.* Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc* 2019;**26**:228–41.

125 Xie L, Lin K, Wang S, *et al.* Differentially Private Generative Adversarial Network. arXiv [cs.LG]. 2018.http://arxiv.org/abs/1802.06739

126 Camino R, Hammerschmidt C, Radu State. Generating Multi-Categorical Samples with Generative Adversarial Networks. arXiv [stat.ML]. 2018.http://arxiv.org/abs/1807.01202

127 Torfi A, Fox EA. COR-GAN: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. Published Online First: 2020.https://onikle.com/articles/196767

128 Yan C, Zhang Z, Nyemba S, *et al.* Generating Electronic Health Records with Multiple Data Types and Constraints. *AMIA Annu Symp Proc* 2020;**2020**:1335–44.

129 Arjovsky, Chintala, Bottou. Wasserstein generative adversarial networks. *conference on machine* ...https://proceedings.mlr.press/v70/arjovsky17a.html

130 Dwork C. Differential Privacy: A Survey of Results. In: *Theory and Applications of Models of Computation*. Springer Berlin Heidelberg 2008. 1–19.