Ensuring patient privacy and accuracy of analytical methods to support evidence-based healthcare

Timothy R. Bergquist

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Sean D. Mooney, Chair

Justin Guinney

David Crosslin

Brian Shirts

Program Authorized to Offer Degree:
Department of Biomedical Informatics and Medical Education

University of Washington

Abstract

Ensuring patient privacy and accuracy of analytical methods to support evidence-based healthcare

Timothy R. Bergquist

Chair of the Supervisory Committee:
Professor Sean D. Mooney
Department of Biomedical Informatics and Medical Education

Over the past two decades, healthcare providers substantially increased their use of electronic health record (EHR) systems. EHRs are primed to become the core of the data driven healthcare system, with the potential to serve as a platform for population health analytics, predictive model development and implementation, and coordination with patients to manage their health information. However, research with EHRs introduces the risk of exposing patient records and business practices to nefarious actors. Creating infrastructure to deliver predictive methods to clinical records while protecting patient privacy is key to building a reliable healthcare analytics platform. In addition, the quality of data from these systems is not fully validated for all use cases, such as assessing population health. Validating the utility of EHRs for use as a population health platform is necessary to fully realize the vision of the data driven health system. Patient involvement in their health is essential to maximize positive patient outcomes. While many vectors exist for patients to access their health information, they are still limited in their ability to contribute to their health data. More solutions are needed to further promote patient involvement with their healthcare information. In this dissertation, I focus on three areas with four aims for building a safe, private, and accessable data analytics platform on the EHR. The aims are to: (1) Evaluate the University of Washington EHR as a generalizable public health repository; (2) Pilot a ”Model to data” framework as a method to deliver predictive analytic methods to clinical records; (3) Scale the ”Model to data” pipeline to host a community challenge, securely delivering outside models to EHRs; and (4) Develop a patient portal to enable patient interaction with their health data and the return of clinically actionable research results.

ACKNOWLEDGMENTS

While only one person will be receiving the degree, this work would not have been possible without the incredible support I received from family, friends, and faculty.

First and foremost, I'd like to thank my wife for her patience with the many late evenings working, and for her emotional and culinary support through this whole process. She has been incredibly encouraging and understanding over the past few years and this would not have been possible without her.

I would not have applied to the PhD program without the encouragement of my advisor Sean Mooney. From hiring me as an undergratuate summer intern, to providing financial support and opening doors for research collaborations, I will be forever grateful to Dr. Mooney for his encouragement, advise, and support over the past 5 years. I'd also like to extend my thanks to my other PhD committee memebers: Justin Guinney, David Crosslin, and Brian Shirts for their collaborations and encouragement.

The Mooney lab has been an incredible source of stimulating conversations, friendships, and laughs. Many people have passed through the lab, each leaving their mark. A special thanks to Chethen Jujjavarapu, Vikas Pejaver, Noah Hammarlund, Yao Yan, Jimmy Phuong, Steve Mooney, Sicheng Song, Don Smith, Abhi Pratap, Houda Benhlabib, Christian Bock,

Rishabh Jain, and many more who have made an impact during my time at the University of Washington.

I'd like to thank the members of the BIME class of 2016. The past 4.5 years have been so much fun because of you and I have so many great memories with these amazing folks. Special shout out to Chethan Jujjavarapu, Pascal Brandt, Ron Buie, Piotr Mankowski, Maggie Dorr, Sicheng Song, Aarti Swaminathan, and Lauren Snyder. Other students in BIME who have been a source of encouragement and enthusiam include Jason Thomas, Hannah Burkhardt, Reggie Casanova-Perez, Nic Dobbins, Abby Menschik, Calvin Apodaca, Will Kearns, Harkirat Sohi, and Aakash Sur.

I'd like to thank many of the staff and professors of the department of Biomedical Informatics and Medical Education for their continued support and teachings: Heidi Kelm, Lora Brewsaugh, Jill Fillmore, Shawn Banta, John Gennari, Kari Stephens, David Crosslin, Peter Tarczy-Hornoch, Neil Abernathy, Annie Chen, David Crosslin, Andrea Hartzler, Bill Lober, Gang Luo, David Masuda, Anne Turner, and Adam Wilcox. I'd like to the thank the leadership and staff of the Institute of Translational Health Sciences and Research IT. A special thanks to Tom Yu and Thomas Schaffter at Sage Bionetworks and Justin Prosser at RIT for their collaboration and teamwork to pull off the EHR DREAM Challenge.

I'd like to thank three of my life-long friends, Christopher, Stephen, and Michael Hunter for the many laughs and much needed moral support through this process. They've been there for me through thick and thin. I'd also like to thank everyone at EOPC, who have all been a great source of comfort, encouragement, and love.

And finally, this wouldn't have been possible without my family. Of course, I wouldn't be here if my mother hadn't birthed me, a fact that I've been reminded of many times over the course of my life. Considering I was homeschooled for most of my pre-college years, she probably gets as much credit for this degree as I do. Thank you to both my father and mother for their encouragement through this process and for the work ethic they instilled in me. To Stephanie (sister), Daniel (brother), and Joe (other brother), thank you for the encouragement and the laughs.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

Chapter 1

INTRODUCTION

1.1 Background

The past two decades have seen a substantial rise in the use of electronic health record (EHR) systems [1]. While the primary drivers of EHR adoption were the 2009 Health Information Technology for Economic and Clinical Health Act and the data exchange capabilities of EHRs [2], secondary use of EHR data 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 history, including doctors' notes, medications prescribed, and billing codes [4]. The widespread adoption of this rich source of data has driven research excitement around the idea of data driven healthcare, where trends from the EHRs inform clinical practice.[5, 6]

The realization of this data driven health future is arriving on many fronts. Machine learning approaches can provide insights into large and complex EHR datasets in a more automated and scalable manner [7, 8]. Hospitals and clinics have already begun to implement predictive analytics solutions to optimize patient care, including models for 30-day readmissions, mortality, and sepsis.[9][10] Population health surveillance is already being carried out for specific diseases using clinical data collected in routine medical care.[11, 12, 13, 14, 15]

Because EHRs, in many cases, have been shown to accurately reflect the disease prevalence of the locally served population[12, 16], EHRs could serve as automated and scalable public health surveillance platforms. And recently, studies have looked at the effect of patients being given access to their clinical notes, where they are able to read the doctors notes as well as enter their own notes [17, 18, 19]. These studies have shown that patients responded positively to having increased access to their EHR data. All of these trends are pushing the EHR to be a one stop shop for healthcare management and coordination, population health surveillance, and predictive analytics development and implementation.

Achieving this vision of the data driven healthcare system is not barrier free. Healthcare institutions face the challenge of balancing patient privacy and EHR data utilization.[20] Regulatory policies such as the Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health Act (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 for the development and evaluation of clinical prediction models. Methods and infrastructure are needed to protect patient privacy while allowing researchers to use EHR data.

Another barrier is the inconsistent quality of EHR data for different use cases. For instance, studies showing that EHRs reflect accurate population health trends also find many exceptions to this claim, such as vaccination rates.[12, 16] In addition, these studies typically only look at diseases that are of interest to public health officials (ex. tuberculosis and diabetes) leaving other diseases out of their validations. If EHRs are to be generalized population health surveillance systems, more validation and assessment is needed.

And while many portals exist linking patients with their health records, there is limited opportunity for patients to interact with and edit their health data. Patient involvement in contributing to their health records will be important as we move toward a data driven health system. Informatics solutions are needed to better link patients with their health information to enable a more personalized approach to elucidating the state of their health.

1.2 Dissertation Aims

In this thesis, I addressed the highlighted deficiencies through the following aims:

1.2.1 Aim 1: Evaluate the University of Washington EHR as a generalizable public health repository.

I used statistical methods to detect events that cause traumatic injury at the University of Washington Medical Center. I compared statistical enrichments and trends to expected population health events to assess longitudinal accuracy of the clinical data when compared to gold standard events.

1.2.2 Aim 2: Pilot a Model to Data framework as a method to deliver predictive analytics methods to protect health information.

I carried out a pilot study that implemented the ”Model to data” framework, enabling a model developer to build a mortality prediction model using hidden EHR data.

1.2.3 Aim 3: Scale the Model to Data pipeline to host a worldwide community challenge, delivering outside models to electronic health records.

I leveraged the pilot study from Aim 2 to host a community challenge where challenge participants built 180-day mortality prediction models using the ”Model to data” infrastructure. Model accuracy was evaluated using standard predictive accuracy methods.

1.2.4 Aim 4: Develop a patient portal to enable the return of research study results from REDCap to patients.

I developed a WordPress plugin that linked patients with their research study results in the REDCap research data management platform. The plugin and its capabilities were showcased in the patient-driven research project FindMyVariant.

1.3 Dissertation Overview

This work serves to advance the knowledge about methods and tools for interacting with health data from the prespective of both the researcher and patient in a secure and ethical manner as well as advancing the body of literature enhancing trust in the EHR as a generalizable population health platform.

Chapter 2

EVALUATING THE EHR AS A GENERALIZABLE PUBLIC HEALTH REPOSITORY

2.1 Introduction

2.1.1 Electronic Health Records and Meaningful Use

The past decade has seen a substantial increase in the rate of Electronic Health Record (EHR) adoption in healthcare.[1] While the primary drivers of EHR adoption have been the 2009 HITECH act and the data exchange capabilities of EHRs,[2] secondary use of EHR data to improve patient safety and health is a key benefit of large-scale adoption.[3] EHRs contain a rich set of information about patients and their health experiences, including doctor's notes, medications prescribed, and billing codes.[4] As hospitals improve data capture quality and quantity, opportunities arise for meaningful use of the data outside the clinic.

2.1.2 Electronic Health Records and Public Health

Public health surveillance – monitoring disease prevalence, and the conditions and behaviors that affect prevalence – is a core component of preventive medicine. Surveillance is conventionally categorized as either 'active' (wherein a health authority contacts care providers or the public to assess conditions) or 'passive' (wherein care providers are mandated to report certain conditions to the health authority).[14] For example, the Center for Disease Control's Behavior Risk Factor Surveillance System (BRFSS),[21] in which trained interviewers contact tens of thousands of respondents by phone each year, is an active system. By contrast, the National Highway Transport Safety Administration's Fatality Analysis Reporting System, in which state transportation departments report motor vehicle crashes to a central system, is a passive system.

With the increasing adoption of EHRs, automated and scalable public health surveillance has become possible. Clinical data that is collected in routine medical care can be algorithmically processed for syndromic surveillance, a passive reporting technique wherein patient cases of a particular disease or condition relevant to population health (frequently, but not exclusively infectious disease) are automatically flagged and reported to appropriate authorities in real time. EHRs have been shown to be a reliable data source capable of facilitating syndromic surveillance.[11, 12, 13, 14, 15] The prevalence estimation of EHRs have also been shown to accurately reflect the known prevalence of a served region. For example, when compared to the gold standard BRFSS dataset, Klompas et al. found that an EHR-based diabetes prevalence detection algorithm was nearly as accurate as the BRFSS dataset.[12] Perlman et al. found that measures of smoking prevalence, obesity rates, hypertension, and diabetes that were derived from the EHR were as accurate as the gold standard BRFSS datasets.[16] The reliability of different conditions often differs by healthcare system, but as more sites adopt EHRs, the estimates should improve for more conditions.[16]

Previous efforts to use EHRs for public health reporting have revolved around using syndromic surveillance to electronically report cases to a data repository external to the

EHR. For instance, Klompas et al. developed a platform for integrating EHR data for use in public health called the Electronic medical record Support for Public Health (ESP).[22] The platform enabled automated systems to pull relevant records from the EHR, and then aggregate data for visualization and analysis in an application called RiskScape.[11] A more recent example of integrating clinical data into a repository for public health surveillance was the Public Health Community Platform (PHCP), an attempt by multiple public health organizations (APHL, ASTHO, JPHIT) to standardize and develop a platform for EHR to cloud-based public health data sharing and electronic case reporting.[22, 23] While the pilot study faced several challenges, it demonstrated long-term feasibility for widespread integration between clinical practice and public health.

2.1.3 The EHR as a Generalizable Population Health Surveillance Platform

While syndromic surveillance typically focuses on the detection and prevalence estimation of specific conditions, EHR databases can act as a generalized population health surveillance system, giving insight into previously unmonitored diseases. For instance, Melamed et al. showed the utility of EHRs to link diseases to seasonal trends.[24] Other seasonal detection methods using EHR data have been used to model seasonal influenza outbreaks, seasonal blood pressure controls, and seasonal effects on early child development.[25, 26, 27] While these studies show that EHRs can be used for accurate population health trends, each of these have looked at only one category of disease at a time.

In this paper, we explore the utility of the EHR as a generalizable event and trend detection platform. In contrast to previous studies, we don't look for seasonal trends of specific diseases, but rather look for unusual coding trends for all traumatic injuries because they have known seasonal trends [24, 25, 26] and gold standard events by which we can validate a generalizable event detection method (e.g., we expect the 4th of July to have a spike in firework accidents). Our goal is to test whether a general event detection method can use a live EHR system to alert public health officials to possible actionable environmental events. We look at deviations from seasonal and temporal trends in medical information collected in routine clinical care, conceptualizing these deviations as events of potential interest to authorities tasked with monitoring population health. We externally validate flagged code/time period combinations, confirming that a holiday or rare event was likely the cause of the unusual injury pattern.

Throughout this paper, we use the term "detection" to refer to the association of statistical trauma trends with individual dates or seasons (e.g., can we "detect" winter or July 4th based on relative diagnosis code frequencies?). We look for diagnosis codes that are statistically "enriched" (a greater proportion of overall visits than would be expected due to chance alone) for different periods of time. We define a code as "enriched" when that code is significantly associated with a given period of time.[28] For instance, we expect injuries from snow sports like skiing, snowboarding, and snowmobiling to be "enriched" in the winter months. We compare trends found to expected trends from literature and common knowledge to test the validity of this event detection technique.

2.2 Methods

2.2.1 Data Source

We obtained a data set (diagnoses by date) from the UW Medicine (the University of Washington Health System) enterprise data warehouse (EDW). The EDW includes patient data from over 4.5 million patients spanning 25 years, and representing various clinical sites across the UW Medicine system including University of Washington Medical Center, Harborview Medical Center, and Northwest Hospital and Medical Center.

"Injury and poisoning" is a category of clinical affliction that includes any traumatic injury or poisoning and is coded as E-codes (E000-E999) or 800-999 codes using the ICD-9-CM diagnosis coding standard or S00-T99 or V00-Y99 codes using the ICD-10-CM coding standard, as defined in the CDC's guidelines for traumatic injury and poisoning.[29, 30] From the EDW, we selected records of all visits between January 1, 1994 and May 2, 2017 for patients who were over the age of 18 as of May 2, 2017 and where, for each visit, at least one ICD-9-CM code or ICD-10-CM code in the "Injury and poisoning" category was recorded. For each patient record, we collected patient visit information which included de-identified patient ID, diagnosis coding method (ICD-9-CM or ICD-10-CM), visit number identifier, admission date and time, diagnosis codes (ICD-9-CM or ICD-10-CM), and diagnosis code description. These data represent just over 3,000,000 unique trauma-related visits to the UW medical system made by over 650,000 unique individuals.

2.2.2 Data Cleaning

UW Medicine adopted the ICD-10-CM billing code system in mid-2015. In order to ensure we had consistent data throughout, we mapped ICD-10-CM codes to their ICD-9-CM equivalents, using the Center for Medicare and Medicaid Services (CMS) General Equivalence Mappings.[31] Since ICD-10-CM has more detailed coding descriptions than ICD-9-CM, there is a potential for data loss when converting from ICD-10-CM to ICD-9-CM. While this may be an issue in some studies, we were more interested in the high level view of UW's patient population, and this data loss was not a major concern for this study. We used a custom tool, DxCodeHandler (https://github.com/UWMooneyLab/DxCodeHandler), to handle code conversion, ICD hierarchy traversal, and diagnosis code manipulation (Additional File 1).

2.2.3 Obtaining Count Data

Per our selection criteria, each patient visit included one or more ICD-9-CM or ICD-10-CM billing codes representing the billing information for the patient visit. We attributed all codes appearing in a visit to the day that visit occurred such that each day was considered a collection of independent code counts. We also included all higher level categories in the ICD hierarchy along with the low level codes. For example, a day that had the code E880.0 (Accidental Fall on or from Escalator) would also have E880 (Accidental Fall from Stairs or Steps), E880-E888 (Accidental Falls), and E000-E999 (External Causes of Injury or Poisoning) counted on that day. This incorporation of multiple category levels was necessary because some real world events enrich different classes of injury such as large classes of injury (e.g. 800-829, Fractures), mid-level classes of injury (e.g. 989, Toxic Effect of Non-medicinal Substances), or specific injury types (e.g. 854.06, Intracranial injury with loss of consciousness).

2.2.4 Binomial Test and Hypothesis Testing

For each diagnosis code, both billable and parent codes, we tested the null hypothesis that the prevalence of each diagnosis code, when calculated against all trauma visits, was consistent across time. We tested this hypothesis using a binomial test, where we tested whether a diagnosis code is more or less prevalent in a given time period when compared to the expected prevalence if the null hypothesis were true. If a code-time period pair had a p-value less than the Bonferroni cutoff, we said that the code is enriched for that tested time period. We used an $\alpha = 0.01$ when calculating the Bonferroni cut off for each experiment. We ran this test for every code that appears more than 10 times in our dataset for all four seasons and for all 365 (non-leap year) days. For each code-time period pair, we generated a score by calculating the -log(p-value) from the binomial test.

2.2.5 Enrichment of Seasons

To find seasonal statistical enrichment of ICD-9-CM billing codes we summed daily counts of each of the 4,582 poisoning and injury billing codes within each season. We defined winter as December - February, spring as March - May, summer as June - August, and autumn as

September - November. For each season/code pair, we performed a binomial test, treating the sum of all codes in that season as the trials, and the count of the code in question for that season as the successes. The expected rate of appearance for each code in question was established by calculating its proportion of all trauma visits across all seasons and years. Thus, the p-value from this test is interpretable as the probability that these many codes or more would be seen in a given season under the null hypothesis that codes are evenly distributed across the year. We used a Bonferroni correction at $n = 18,328(4x4582)$. We also filtered out codes that appeared less than 10 times over the course of the 24-year period.

2.2.6 Enrichment of Dates

We used an analogous method to detect code enrichments for days of the year. Again, we computed the sum of codes occurring on each of the 365 (non-leap-day) days of the year. For each code/day pair, we performed a binomial test using the total number of codes used on that day as the number of trials, and the number of times the specific code of interest was used as the number of successes. The expected rate was derived from the baseline rate of appearance for the code of interest per day across the entire year when compared to the total number of trauma visits on that given day. We calculated a Bonferroni cutoff at $n = 1,672,430(4582x365)$. We counted codes as enriched if the p-value was less that the Bonferroni correction and the daily rate of the code was greater than the baseline expected rate of the code (we did not look at depletions). We also filtered out codes that appeared less than 10 times over the course of the 24 year dataset period.

2.2.7 IRB Considerations

We received an IRB non-human subjects research designation from the University of Washington Human Subjects Research Division to construct a dataset derived from all patient diagnoses from the EDW over the age of 18. (IRB number: STUDY00000669) Data was extracted by an honest broker, the UW Medicine Research IT data services team, and no patient identifiers were available to the research team.

2.3 Results

2.3.1 Statistical Enrichment of Seasons

We detected patterns of seasonal enrichment consistent with our expectations about seasonal behavior. For example, in winter, we found enrichment of not only accidents from snow sports such as skiing and snowboarding, among others, but also cold weather-related ailments such as frostbite and hypothermia. Other codes that may be related to snow sport accidents such as head injuries, sprains, and strains were also enriched (Table 2.1). Spring begins to have more fair weather activities such as outdoor related ailments like allergies and sporting accidents (Table 2). Summer sees disproportionate numbers of accidents related to outdoor activities in warm weather such as bites and stings from bugs, firework accidents, bicycle accidents, and water transport accidents (Table 3). While autumn is the least distinctive of the seasons, it has a unique enrichment for vehicle accidents (Table 4). This may be because autumn contains high traffic holidays (Thanksgiving, Labor Day) and increased levels of rain in Seattle.

ICD 9 Code	Description	Winter Code Counts	Average Counts in Other Seasons	Percent Increase	P Value	Scores
E885.4	Fall From Snowboard	831	115	622.61	0.00	750.00
E885.3	Fall From Skis	593	122	386.07	2.59E-221	507.92
991	Effects of Reduced Temperature	2027	995.33	103.65	5.16E-217	498.02
E885	Fall on Same Level From Slipping, Tripping, or Stumbling	10738	9019	19.06	6.67E-140	320.46
996-999	Complications of Surgical and Medical Care	134022	135432	-1.04	3.24E-137	314.28
995.29	Unspecified Adverse Effect of Other Drug, Medicinal and Biological Substance	5019	3751.33	33.79	1.61E-134	308.07
996	Complications Peculiar to Certain Specified Procedures	88630	88559	0.08	4.72E-122	279.36
E930-E949	Adverse Effects From Substance in Therapeutic use	16422	15087.67	8.84	1.60E-92	211.37
E820	Nontraffic Accident Involving Motor-driven Snow Vehicle	239	51.33	365.61	7.72E-87	198.28
995.2	Other and Unspecified Adverse Effect of Drug, Medicinal and Biological Substance (due) to Correct Medicinal Substance Properly Administered	8648	7517.33	15.04	7.82E-86	195.97
991.2	Frostbite of Foot	372	118.67	213.47	1.25E-85	195.5
E003.2	Activities Involving Snow (alpine) (downhill) Skiing, Snow Boarding, Sledding, Tobogganing and Snow Tubing	148	19.67	652.41	1.50E-80	183.8
E901.0	Accident due to Excessive Cold due to Weather Conditions	334	104.67	219.1	6.42E-79	180.04
E003	Activities Involving Snow and ice	169	27.67	510.77	1.81E-78	179.01
E901	Excessive Cold	474	201.33	135.43	1.25E-69	158.65
E880-E888	Accidental Falls	38739	38299.67	1.15	1.37E-68	156.26
991.6	Hypothermia	760	414	83.57	1.01E-64	147.36
E885.9	Fall From Other Slipping, Tripping, or Stumbling	8953	8067.67	10.97	3.03E-63	143.95
990-995	Other and Unspecified Effects of External Causes	29795	29270	1.79	1.69E-60	137.63

Table 2.1: The top 20 most enriched codes for winter. Enriched codes include accidents from snow sports such as skiing and snowboarding as well as cold weather-related ailments such as frostbite and hypothermia. Other codes that may be related to snow sport accidents such as head injuries, sprains, and strains were also enriched. We report by percent increase as well as -log(p). We compare the number of codes found in winter to the average code counts of the other three seasons.

Dx Code	Descriptions	Spring Code Counts	Average Count in Other Seasons	Percent Increase	P Value	Scores
840-848	Sprains and Strains of Joints and Adjacent Muscles	138376	132163.33	4.7	1.04E-66	151.93
995.3	Allergy, Unspecified	6304	5087	23.92	5.55E-61	138.74
990-995	Other and Unspecified Effects of External Causes	31010	28865	7.43	3.84E-36	81.55
905-909	Late Effects of Injuries, Poisonings, Toxic Effects, and Other External Causes	50277	47594	5.64	9.62E-35	78.33
995	Certain Adverse Effects not Elsewhere Classified	27996	26012	7.63	2.36E-34	77.43
980.9	Toxic Effect of Unspecified Alcohol	328	167.67	95.62	7.71E-28	62.43
844	Sprains and Strains of Knee and leg	18141	16806.33	7.94	1.71E-24	54.73
908.6	Late Effect of Certain Complications of Trauma	648	431	50.35	2.05E-22	49.94
E917.0	Striking Against or Struck Accidentally by Objects or Persons in Sports	2471	2020.33	22.31	3.06E-22	49.54
842	Sprains and Strains of Wrist and Hand	12683	11674.33	8.64	2.29E-20	45.22
848	Other and Ill-defined Sprains and Strains	16380	15278.67	7.21	8.61E-19	41.60
854	Intracranial Injury of Other and Unspecified Nature	17691	16558	6.84	2.01E-18	40.75
854	Without Mention of Open Intracranial Wound	17515	16401.67	6.79	5.34E-18	39.77
905	Late Effects of Musculoskeletal and Connective Tissue Injuries	23970	22703.33	5.58	4.87E-17	37.56
905.4	Late Effect of Fracture of Lower Extremities	9711	8915	8.93	7.21E-17	37.17
842.12	Sprain of Metacarpophalangeal (joint) of Hand	1659	1344.33	23.41	1.05E-16	36.79
842.1	Hand	5902	5296.67	11.43	2.50E-16	35.93
919.9	Other and Unspecified Superficial Injury of Other, Multiple, and Unspecified Sites, Infected	78	27.33	185.4	2.06E-15	33.82
854	Intracranial Injury of Other and Unspecified Nature Without Mention of Open Intracranial Wound, Unspecified State of Consciousness	13228	12381.67	6.84	3.94E-14	30.86
996	Complications Peculiar to Certain Specified Procedures	90209	88032.67	2.47	9.98E-14	29.94

Table 2.2: The top 20 most enriched codes for spring. Enriched codes include allergies, sprains and strains, and sports related injury. We report by percent increase as well as -log(p). We compare the number of codes found in spring to the average code counts of the other three seasons.

Dx Code	Descriptions	Summer Code Counts	Average Count in Other Seasons	Percent Increase	P Value	Scores
E826-E829	Other Road Vehicle Accidents	5872	3166	85.47	5.54E-314	721.3
919	Superficial Injury of Other Multiple and Unspecified Sites	7846	4621	69.79	2.28E-301	692.25
E923.0	Accident Caused by Fireworks	480	44	990.91	2.97E-296	680.48
910-919	Superficial Injury	30366	22597.33	34.38	1.10E-290	667.65
919.4	Insect Bite, Nonvenomous, of Other, Multiple, and Unspecified Sites, Without Mention of Infection	2483	1067.33	132.64	7.41E-251	575.95
E826.1	Pedal Cycle Accident Injuring Pedal Cyclist	3933	2021.33	94.57	3.24E-246	565.26
997.91	Complications Affecting Other Specified Body Systems, Hypertension	1040	297.67	249.38	5.61E-220	504.84
940-949	Burns	45311	36094.67	25.53	3.83E-209	479.9
989.5	Toxic Effect of Venom	3019	1535.67	96.59	1.57E-195	448.56
E905.3	Sting of Hornets, Wasps, and Bees Causing Poisoning and Toxic Reactions	759	188	303.72	4.56E-195	447.49
800-829	Fractures	264689	231748.33	14.21	3.25E-171	392.56
E905	Venomous Animals and Plants as the Cause of Poisoning and Toxic Reactions	1006	350	187.43	1.06E-156	359.14
E830-E838	Water Transport Accidents	629	163	285.89	4.56E-153	350.78
989	Toxic Effect of Other Substances, Chiefly Nonmedicinal as to Source	3464	2016	71.83	8.47E-141	322.53
959.8	Other Specified Sites, Including Multiple Injury	17695	13440	31.66	1.21E-140	322.17
E923	Accident Caused by Explosive Material	927	335.67	176.16	1.66E-134	308.04
997.9	Complications Affecting Other Specified Body Systems	1262	535.67	135.59	3.50E-132	302.69
E826	Pedal Cycle Accident	5176	2631.67	96.68	0.00E+00	750
E900-E909	Accidents due to Environmental Factors	5367	3562.33	50.66	1.22E-116	266.9
E906.4	Bite of Nonvenomous Arthropod	1548	762	103.15	2.52E-111	254.66

Table 2.3: The top 20 most enriched codes for summer. Enriched codes include accidents related to outdoor activities in warm weather such as bites and stings from bugs, burns, firework accidents, bicycle accidents, and water transport accidents. We report by percent increase as well as -log(p). We compare the number of codes found in summer to the average code counts of the other three seasons.

Dx Code	Descriptions	Fall Code Counts	Average Count in Other Seasons	Percent Increase	P Value	Scores
E819.0	Motor Vehicle Traffic Accident of Unspecified Nature Injuring Driver of Motor Vehicle Other Than Motorcycle	1388	832.33	66.76	5.40E-70	159.49
E810-E819	Motor Vehicle Traffic Accidents	41860	39083.67	7.1	1.18E-48	110.36
E819	Motor Vehicle Traffic Accident of Unspecified Nature	17872	16481	8.44	5.12E-29	65.14
E819.1	Motor Vehicle Traffic Accident of Unspecified Nature Injuring Passenger in Motor Vehicle	777	518.33	49.9	1.53E-26	59.44
825	Fracture of one or More Tarsal and Metatarsal Bones	17527	16296	7.55	1.39E-23	52.63
900.9	Injury to Unspecified Blood Vessel of Head and Neck	557	366	52.19	8.43E-21	46.22
847	Sprain of Neck	18864	17707	6.53	8.62E-20	43.90
825	Fracture of Calcaneus, Closed	6243	5581.67	11.85	2.96E-19	42.66
E863.1	Accidental Poisoning by Insecticides of Organophosphorus Compounds	17	0.67	2437.31	1.43E-18	41.09
E980.9	Poisoning by Other and Unspecified Solid and Liquid Substances, Undetermined Whether Accidentally or Purposely Inflicted	565	383.67	47.26	2.74E-18	40.44
E949.6	Other and Unspecified Viral and Rickettsial Vaccines Causing Adverse Effects in Therapeutic use	37	6.33	484.52	6.33E-17	37.30
836	Dislocation of Knee	8608	7889.33	9.11	9.74E-17	36.87
999.9	Other and Unspecified Complications of Medical Care	1538	1241.33	23.9	1.54E-16	36.41
830-839	Dislocation	21369	20276.33	5.39	3.83E-16	35.50
E912	Inhalation and Ingestion of Other Object Causing Obstruction of Respiratory Tract or Suffocation	219	121.33	80.5	1.02E-15	34.52
E812	Other Motor Vehicle Traffic Accident Involving Collision With Motor Vehicle	13813	12965.67	6.54	6.65E-15	32.64
850.9	Concussion, Unspecified	2125	1793	18.52	7.41E-15	32.54
E812.0	Other Motor Vehicle Traffic Accident Involving Collision With Motor Vehicle Injuring Driver of Motor Vehicle Other Than Motorcycle	8037	7412	8.43	7.01E-14	30.29
E849.5	Street and Highway Accidents	7876	7266.67	8.39	1.66E-13	29.43
E881	Fall on or From Ladders or Scaffolding	1663	1386.33	19.96	1.97E-13	29.25

Table 2.4: The top 20 most enriched codes for autumn. Enriched codes include motor vehicle accidents and sprains of neck. We report by percent increase as well as -log(p). We compare the number of codes found in autumn to the average code counts of the other three seasons.

2.3.2 Statistical Enrichment for Days of the Year

To complement our seasonal analyses, we explored enrichment of diagnosis codes for all 365 days of the year. Each date that had a code scored below the Bonferroni threshold was flagged as having possible significance. We detected 100 days that had at least one code flagged as enriched. We generated an enrichment score for each of the dates by calculating the -log(p-value) of the lowest p-value for the date. The top 15 dates with the highest scoring codes are shown in Figure 2.1. The days in which enrichment of many codes is common are a mixture of holidays and one time events. For example, there was enrichment of codes related to fights, firework accidents, and alcohol poisoning on January 1st (Table 5). Analogously, there was a large increase in the number of firework related accidents and burns on the 4th and 5th of July as well as an increase in the number of off-road vehicle accidents and poisoning by alcohol (Table 6-7). We also observe an increase in alcohol poisoning, vehicle accidents, and an increase in possible self-harm on Christmas Eve (Table 8). For tables 5-8, we limit the reporting of codes to those that had more than 30 appearances over the 24 years of data. This reduces false positives arising from extremely rare codes that appeared during the baseline period. We also report by percent increase rather than -log(p) for better interpretability.

Dx Code	January 1st Average Code Count	Daily Average Code Count	% Increase	Description
E923.0	1.74	0.07	2469.74	Accident caused by fireworks
E923	2.39	0.22	981.77	Accident caused by explosive material
E965	1.39	0.42	235.15	Assault by firearms and explosives
854.06	1.57	0.49	217.75	Intracranial injury with loss of consciousness of unspecified duration
E922.9	1.52	0.54	180.57	Accident caused by unspecified firearm missile
E922	1.87	0.69	171.55	Accident caused by firearm and air gun missile
E860	5.39	2.01	168.36	Accidental poisoning by alcohol, not elsewhere classified
E860-E869	5.96	2.23	167.25	Accidental Poisoning By Other Substance
E860.0	5.17	1.95	165.22	Accidental poisoning by alcoholic beverages
980.8	1.57	0.61	154.88	Toxic effect of other specified alcohols

Table 2.5: The top 10 most enriched codes for January 1st. As expected for New Year's Day, the most enriched codes were related to firework accidents, alcohol, and assaults. To reduce the false positive rate of the code enrichment from extremely rare codes that appeared during the baseline period, the enriched codes were only counted if they appeared more than 10 times over the 24 year period. We also report by percent increase rather than -log(p) for better interpretability.

Dx Code	Descriptions	Fall Code Counts	Average Count in Other Seasons	Percent Increase
Dx Code	July 4th Average Code Count	Daily Average Code Count	% Increase	Description
E923.0	4.26	0.06	6913.3	Accident caused by fireworks
E923	4.91	0.21	2194.4	Accident caused by explosive material
E820-E825	1.70	0.69	147.0	Motor Vehicle Non-traffic Accidents
980.8	1.43	0.61	133.5	Toxic effect of other specified alcohols
948.00	2.87	1.54	85.8	Burn involving less than 10 percent of body surface with third degree burn
948.0	2.87	1.56	83.8	Burn involving less than 10 percent of body surface
948	4.09	2.24	82.5	Burns classified according to extent of body surface involved
E819.2	3.00	1.70	76.2	Motor vehicle traffic accident of unspecified nature injuring motorcyclist
851	2.09	1.22	71.1	Cerebral laceration and contusion
851.8	1.35	0.79	69.9	Other and unspecified cerebral laceration and contusion, without mention of open intracranial wound

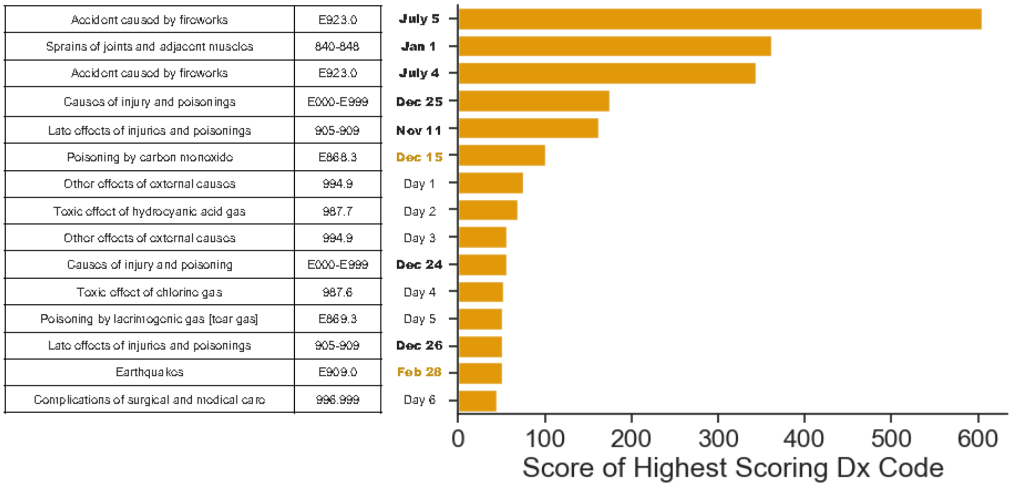
Table 2.6: The top 10 most enriched codes for July 4th. As expected for Independence Day, the most enriched codes were related to firework accidents, burns, and alcohol poisoning. To reduce the false positive rate of the code enrichment from extremely rare codes that appeared during the baseline period, the enriched codes were only counted if they appeared more than 10 times over the 24 year period. We also report by percent increase rather than -log(p) for better interpretability.

Dx Code	July 4th Average Code Count	Daily Average Code Count	% Increase	Description
E923.0	4.26	0.06	6913.3	Accident caused by fireworks
E923	4.91	0.21	2194.4	Accident caused by explosive material
E820-E825	1.70	0.69	147.0	Motor Vehicle Non-traffic Accidents
980.8	1.43	0.61	133.5	Toxic effect of other specified alcohols
948.00	2.87	1.54	85.8	Burn involving less than 10 percent of body surface with third degree burn
948.0	2.87	1.56	83.8	Burn involving less than 10 percent of body surface
948	4.09	2.24	82.5	Burns classified according to extent of body surface involved
E819.2	3.00	1.70	76.2	Motor vehicle traffic accident of unspecified nature injuring motorcyclist
851	2.09	1.22	71.1	Cerebral laceration and contusion
851.8	1.35	0.79	69.9	Other and unspecified cerebral laceration and contusion, without mention of open intracranial wound

Table 2.7: The top 10 most enriched codes for July 5th. As expected for the day after Independence Day, the most enriched codes were related to firework accidents and burns as the injured persons from July 4th continue to appear in the hospital. To reduce the false positive rate of the code enrichment from extremely rare codes that appeared during the baseline period, the enriched codes were only counted if they appeared more than 10 times over the 24 year period. We also report by percent increase rather than -log(p) for better interpretability.

Dx Code	Descriptions	Fall Code Counts	Average Count in Other Seasons	Percent Increase
Dx Code	July 4th Average Code Count	Daily Average Code Count	% Increase	Description
E923.0	4.26	0.06	6913.3	Accident caused by fireworks
E923	4.91	0.21	2194.4	Accident caused by explosive material
E820-E825	1.70	0.69	147.0	Motor Vehicle Non-traffic Accidents
980.8	1.43	0.61	133.5	Toxic effect of other specified alcohols
948.00	2.87	1.54	85.8	Burn involving less than 10 percent of body surface with third degree burn
948.0	2.87	1.56	83.8	Burn involving less than 10 percent of body surface
948	4.09	2.24	82.5	Burns classified according to extent of body surface involved
E819.2	3.00	1.70	76.2	Motor vehicle traffic accident of unspecified nature injuring motorcyclist
851	2.09	1.22	71.1	Cerebral laceration and contusion
851.8	1.35	0.79	69.9	Other and unspecified cerebral laceration and contusion, without mention of open intracranial wound

Table 2.8: The top 10 most enriched codes for December 24th. The most enriched codes were related to alcohol poisoning, injury to spleen, and injury undetermined whether accidental of purposely inflicted. To reduce the false positive rate of the code enrichment from extremely rare codes that appeared during the baseline period, the enriched codes were only counted if they appeared more than 10 times over the 24 year period. We also report by percent increase rather than -log(p) for better interpretability.



Accident caused by fireworks	E923.0	July 5
Sprains of joints and adjacent muscles	840-848	Jan 1
Accident caused by fireworks	E923.0	July 4
Causes of injury and poisonings	E000-E999	Dec 25
Late effects of injuries and poisonings	905-909	Nov 11
Poisoning by carbon monoxide	E868.3	Dec 15
Other effects of external causes	994.9	Day 1
Toxic effect of hydrocyanic acid gas	987.7	Day 2
Other effects of external causes	994.9	Day 3
Causes of injury and poisoning	E000-E999	Dec 24
Toxic effect of chlorine gas	987.6	Day 4
Poisoning by lacrimogenic gas [tear gas]	E869.3	Day 5
Late effects of injuries and poisonings	905-909	Dec 26
Earthquakes	E909.0	Feb 28
Complications of surgical and medical care	996.999	Day 6

Figure 2.1: The top 15 highest scoring days of the year. The top 15 days with the highest scoring diagnosis codes. Each of the codes in the table are the most enriched codes on each of the days in the date column. The black bolded dates are either holidays or are dates that surround a holiday. The orange bolded dates are associated with known rare events that clearly explain the enrichment of their codes, namely the Nisqually Earthquake on Feb 28, 2001 and the Hanukkah Eve Windstorm on Dec 15, 2006. The other dates have unusual patterns of enriched codes such as chlorine gas poisoning and tear gas poisoning, but we could not find a readily available explanation to confirm some holiday, environmental, or social event on these days. Since these events appear to have happened on a single day in a single year and look to be associated with specific events, we have masked the dates due to the unknown specificity of these events and potential for identification of individuals involved in these events.

2.3.3 Rare Events as Case Studies

We detected enrichment of unusual codes on multiple days that did not seem linked to their respective day by either holiday or seasonal event. Upon further evaluation, we inferred that we had detected past environmental events that showed up as single day enrichments. Feb 28, Dec 15, May 31, and Nov 8 were four of the days in the top 15 highest scoring days that followed this pattern (Figure 2.1). Because these enriched days fell in single years, we were able to search for news stories published on or immediately after these days to see if we could find the cause of the increase in these unusual codes.

2.3.4 Nisqually Earthquake

In our analysis, February 28th was shown to have an increase in earthquake related accidents, ICD-9-CM code E909.0. On February 28, 2001, there was a magnitude 6.8 earthquake centered in Western Washington.[32, 33] All the earthquake codes found on February 28th in our dataset were from 2001, consistent with there being very few earthquake related accidents in the EHR except during the major earthquake.

2.3.5 Hanukkah Eve windstorm

Our event detection method also discovered a significant increase on December 15 of the ICD-9-CM code E868.3 (accidental poisoning by carbon monoxide from incomplete combustion of other domestic fuels). Nearly all the codes were found to have been coded in 2006. The Hanukkah Eve windstorm of Dec 15, 2006 led to widespread and lengthy power outages. In the aftermath, there were news stories about the increase in carbon monoxide poisonings due to people barbecuing and running generators in their homes without ventilation.[34, 35] Indeed, public health authorities responded with concerns that the dangers of carbon monoxide poisoning were not widely understood in select communities.[36]

2.3.6 Industrial Accidents

We detected two other single day enrichments: May 31 with an enrichment of E891.3 (Burning caused by conflagration) and Nov 8 with an enrichment of 987.6 (Toxic effect of chlorine gas). We were able to link these two enrichments to the May 31, 2004 monorail fire in Seattle [37] and the November 8, 1994 chlorine spill and fire at the Coastal Dock in Ballard, WA.[38]

2.4 Discussion

We explored the value of UW Medicine EHR data for detecting public health-related environmental and seasonal causes of traumatic injury. Our analysis finds that tests for seasonal and daily enrichment of the frequency of emergency room visits for trauma detects expected events, including both seasonal trends such as winter sports-related injuries, day-specific events such as July 4th burns, and rare events such as the Nisqually earthquake.

2.4.1 Interesting Anomalies

Non-enriched Holidays

While most of our results confirmed expected seasonal and date-specific trends, we were surprised not to find enrichment of alcohol related injuries on St. Patrick's Day or the day following, given that St. Patrick's Day is associated with increased alcohol consumption.[39, 40] This may indicate the effectiveness of extra police patrols deployed for that day. This could also be a false negative due to the conservative nature of Bonferroni corrections.

Prior studies have examined date-related events in relation to traumatic injury. One study found that on April 20th, a date associated with celebrating marijuana consumption, there was an increase in the number of car accidents.[41] While we did not observe a statistical enrichment in car accidents, our method did identify a statistical enrichment in burns (940-949), another potential consequence of marijuana use.[42] Future work could analyze clinical notes which might allow us to identify if this enrichment is attributable to elevated marijuana use.

Enrichment of Post-surgical Complications in Winter

We also saw unexpected trends in post-surgical complications, with those terms being enriched in the winter months at the very end and beginning of the year. One hypothesis is that there is a relative increase in the number of surgeries in November and December as people schedule elective surgeries before insurance deductibles reset in the new year. An alternate hypothesis is that people defer reporting minor surgical complications until after the end-of-year holidays. We were unable to explore these hypotheses for this study because our data was limited to visits including trauma codes and did not include surgical appointments. It is also important to note that we saw a relative increase in the number of surgical complications due to lower numbers of trauma visits in the winter, and not necessarily an absolute increase in the number of post-surgical complications (Figure 2.2). Since codes related to post-surgical complications are less specific and are more likely to appear during trauma visits than other codes discussed thus far, the effect of this "lowered baseline" is particularly noticeable.

Unlinked Events

There were multiple dates that had significant enrichment of codes on a date where nearly all the codes came from one year. For instance, there were a large number of visits with the code 994.9 (other effect of external causes) on one of the masked days. This code is too vague to understand the common injuries of patients and, at the time of this study, we did not have access to de-identified clinical notes from which to elicit the causes of these injuries. There was also no readily available source of news that we found to corroborate a large number of people being injured by any social or environmental event. We were not able to discern whether these dates were false positives, whether the codes were entered incorrectly, or whether there was a common event that caused these injuries. In this paper, we have masked the specific dates of these unlinked days to protect against the potential de-identification of patients since the circumstances surrounding these injuries are unknown.

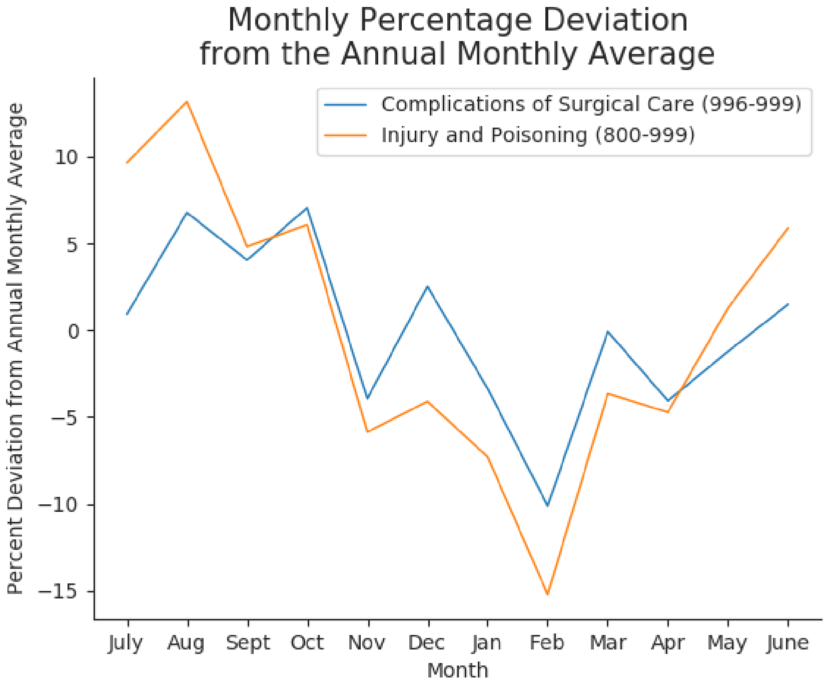


Figure 2.2: Comparison of the code count trend differences between 996 and 999 and 800–999The percent deviation from the annual monthly average code count for both the Complications of Surgical Care (996–999) diagnosis family and the broad category of Injury and Poisoning (800– 999). By calculating the average monthly code count for each family and the percent deviation per month from that expected average, we see that both code families follow a similar seasonal pattern of increase in the summer and decrease in the winter in terms of raw code count. While they follow the same pattern, Complications of Surgical Care doesn't decrease as much in the winter, and actually has a spike in December, which is why our method picks up this diagnosis family as enriched in the winter. Since the number of trauma visits is used to establish a baseline expected rate of each code count, our method is detecting relative enrichment and not absolute enrichment

2.4.2 Study Strengths and Limitations

Our study has several notable strengths. First, the UW Medicine system has used EHRs for a long time, affording us access to over 20 years of clinical data from a large urban health care system. Second, UW Medicine's location in Western Washington lends itself to year-round yet season-specific outdoor activities whose resulting injuries show up as specific trauma codes, including snow sports in the winter and boating in the summer. This access increased our ability to detect seasonal trauma trends.

However, our study also has limitations. First, as with any study of EHRs, we cannot rule out biases due to site-specific coding practices or changes in practitioner knowledge of the health record system. However, we have no reason to believe errors caused by these issues would vary by season or day. Second, the UWMC is mainly a referral institution, such that many patients visit the system only for specialty services. We also know that only around 31% of all patients visiting the UW medical system will have their next visit at a UW clinic.[43] This is mitigated in our study by the fact that we only considered trauma-related diagnosis codes and that UW Medicine is the only Level I trauma center in Washington, Alaska, Montana and Idaho. The impact of this known bias decreases since our study looks at individual admissions and does not require a full picture of each patient odyssey. The results of our study are not reliant on continuity of care. Nevertheless, further validation studies are needed to evaluate the representation of the UWMC data in the Seattle Region. Another future solution would be to run our method at more sites across Washington, feeding the live statistics into an aggregation mechanism for a more robust population view.

2.4.3 Using Electronic Health Records for Event Detection

Our method could be used in a live surveillance situation by alerting authorities and doctors when an unusual increase of cases with a particular diagnosis code show up across multiple hospitals with linked EHR systems. It could spark an investigation into what is causing the sudden increase but also could initiate public health policy development that previously would take longer to assess and carry out. While our method focused on traumatic injury, it could easily be expanded to include surveillance of all diagnosis codes. A limitation of using billing codes for surveillance is the delay that occurs between patient care and the billing process. While this delay is shorter than periodically collecting all the latest billing codes, a true real-time surveillance system isn't possible. A possible next step would be to train an NLP classifier based on the clinical note texts from each visit to "predict" the diagnosis codes that will be associated with a visit. While not a trivial pursuit, this would enable a near real-time surveillance system. Aside from predicting diagnosis codes, incorporating clinical notes into the method could more accurately cluster events and better inform detected trends. Natural language processing techniques could be used to find "enriched" keywords on the detected days to add context to the detected events in a data driven automated manner.

2.5 Conclusion

In conclusion, EHR data hold considerable potential for public health surveillance. We explored the potential to leverage UW Medicine's enterprise data warehouse to detect seasonal, holiday, and rare events using diagnosis codes for injuries and poisonings. Our method detected many of the trends for seasons and specific dates we expected, while identifying several intriguing new enrichments. Future research should focus on improving our trend and event detection method to differentiate between one-time effects like the Nisqually earthquake, and repeat events like Independence Day. Incorporating clinical notes into a detection method could more accurately cluster events and better inform detected trends. Expanding the method to all diagnosis codes could detect new non-trauma related events. Our findings add to the growing body of literature showing that EHRs hold considerable potential as generalizable population health surveillance platforms.

Chapter 3

PILOTING A "MODEL TO DATA" APPROACH IN THE CONTEXT OF AN EHR ENTERPRISE DATA WAREHOUSE.

3.1 Introduction

3.1.1 Electronic health records and the future of data-driven health care

Healthcare providers substantially increased their use of EHR systems in the past decade [1]. While the primary drivers of EHR adoption were the 2009 Health Information Technology for Economic and Clinical Health Act and the data exchange capabilities of EHRs [2], secondary use of EHR data 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 history, including doctors' notes, medications prescribed, and billing codes [4]. The prevalence of EHR systems in hospitals enables the accumulation and utilization of large clinical data to address specific clinical questions. Given the size and complexity of these data, machine learning approaches provide insights in a more automated and scalable manner [7, 8]. Healthcare providers have already begun to implement predictive analytics solutions to optimize patient care, including models for 30-day readmissions, mortality, and sepsis [9]. As hospitals improve data capture quality and quantity, opportunities for more granular and impactful prediction questions will become more prevalent.

3.1.2 Hurdles to clinical data access

Healthcare institutions face the challenge of balancing patient privacy and EHR data utilization [20]. Regulatory policies such as Health Insurance Portability and Accountability Act and Health Information Technology for Economic and Clinical Health Act place the onus and financial burden of ensuring the security and privacy of the 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. Research collaborations are often bound by highly restrictive data use agreements or business associate agreements limiting the scope, duration, quantities, and types of EHR data that can be shared [44]. This friction has slowed, if not impeded, researchers' abilities to build and test clinical models.[44] While these data host-researcher relationships are important and lead to impactful collaborations, they are often limited to intrainstitution collaborations, relegating many researchers with no healthcare institution connections to smaller public datasets or inferior synthetic data. One exception to this is the patient-level prediction working group in the Observational Health Data Sciences and Informatics community, which developed a framework for building and externally validating machine learning models [45]. 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 (OMOP) Common Data Model (CDM) [46, 47], 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 reidentification of patients, while maximizing the ease of access and use of the clinical data.

3.1.3 Methods for sharing clinical data

De-identification of EHR data and the generation of synthetic EHR data are 2 solutions to enable clinical data sharing. De-identification methods focus on removing or obfuscating the 18 identifiers that make up the protected health information as defined by the Health Insurance Portability and Accountability Act [48]. De-identification reduces the risk of information leakage but may still leave a unique fingerprint of information that is susceptible to reidentification [48, 49]. De-identified datasets like MIMIC-III are available for research and have led to innovative research studies [50, 51, 52]. However, these datasets are either limited in size (MIMIC-III [Medical Information Mart for Intensive Care-III] only includes $38,597$ distinct adult patients and $49,785$ hospital admissions), scope (MIMIC-III is specific to intensive care unit patients), and availability (data use agreements 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 [53]. Synthetic data generators, such as medGAN [51], can generate EHR datasets consisting of highdimensional discrete variables (both binary and count features), although the temporal information of each EHR entry is not maintained. Methods such as OSIM2 are able to maintain the temporal information but only simulate a subset of the data specific to a usecase (eg, drug and treatment effects) [54]. 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 [55]. 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.

3.1.4 "Model to data" framework

The "Model to Data" (MTD) framework, a method designed to allow machine learning research on private biomedical data, was described by Guinney et al [56] as an alternative to traditional data sharing methods. The focus of MTD 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, in which the shared or synthetic data were limited in both scope and size, an MTD approach grants a researcher the ability to use all available data from identified datasets, even 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 MTD framework relies on modern containerization software such as Docker [57] or Singularity [58] for model portability, which serves as a "vehicle," 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 containerization software not only facilitates the secure delivery and execution of models, but it opens up the ability for integration into cloud environments (eg, Amazon Web Services, Google Cloud) for costeffective and scalable data analysis. The MTD approach has been successful in a series of recent community challenges but has not yet been shown to work with large, EHR datasets [59]. Here, we present a pilot study of an MTD 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) to their on-premises secure computing infrastructure. The main goals of this pilot are to demonstrate (1) the operationalization of the MTD approach within a large health system, (2) the ability of the MTD 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 MTD community challenge for evaluating clinical algorithms on remotely stored and protected patient data.

3.2 Methods

3.2.1 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 transformed the patient records from 2010 to the present day into a standardized data format, OMOP CDM v5.0. For this pilot study, we selected all patients who had at least 1 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.

3.2.2 Scientific question for the pilot of the "Model to data" approach

For this MTD demonstration, the scientific question we asked the model developer to address was the following: Given the past EHRs of each patient, predict the likelihood that he/ she will pass away 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, the mortality prediction question has been thoroughly studied [60, 10, 61]. For these reasons, patient mortality prediction represents a well-defined proof-of-concept study to showcase the potential of the MTD evaluation platform.

3.2.3 Defining the training and evaluation datasets

For the purpose of this study, we split the data into 2 sets: the training and the 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 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. This way 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 180 days 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 because this resulted in an 80/20 split between the training and 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 (Fig 3.1, patient 7) and from the training dataset for patients who did not have a confirmed death (Fig 3.1, patient 3). We only defined the true positives for the

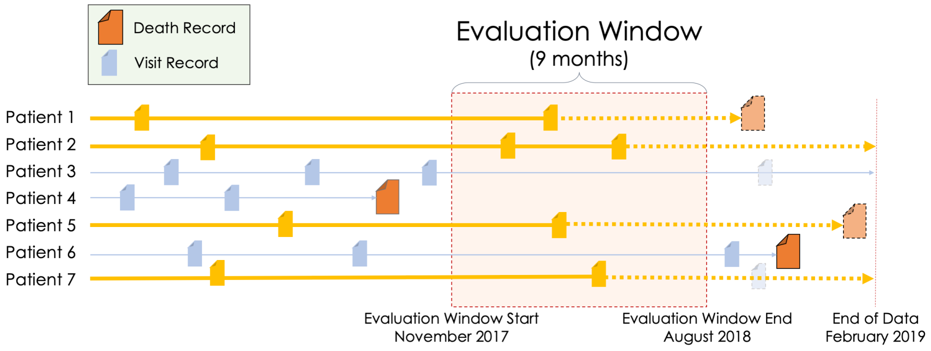


Figure 3.1: Defining the evaluation dataset. Any patient with at least 1 visit within the evaluation window was included in the evaluation dataset (gold). All other patient records were added to the training dataset (blue). Visits that were after the evaluation window end were excluded from the evaluation dataset and from the training dataset for patients who did not have a confirmed death (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.

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.

3.2.4 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 [57]. 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 unknown codes 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 (Seattle, WA) for researchers to share data, compare and communicate their methodologies, and seek collaboration [62]. Synapse is composed of a set of shared REST (representational state transfer)-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 [63]. The Synapse platform provides services that enable submissions of files or Docker images to an evaluation queue, which have previously been used to manage containerized models submitted to DREAM challenges [62]. We use 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 4 models in parallel. Each model had access to 70 GB of RAM, 4 vCPUs, and 50 GB of SSD.

3.2.5 Institutional review board considerations

We received an institutional review board (IRB) nonhuman 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 (institutional review board 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.

3.3 Results

3.3.1 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),30 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, in which 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 2 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 0 and 1) for each patients in the evaluation dataset (Fig 3.2). Each stage had a mounted volume "scratch" available for storing intermediate files such as selected features (Fig 3.2). The model developer specified commands and dependencies (eg, python packages) for the 2 stages in the Dockerfile, train.sh, and infer.sh. The training and evaluation datasets were mounted to read-only volumes designated "train" and "infer" (Fig 3.2).

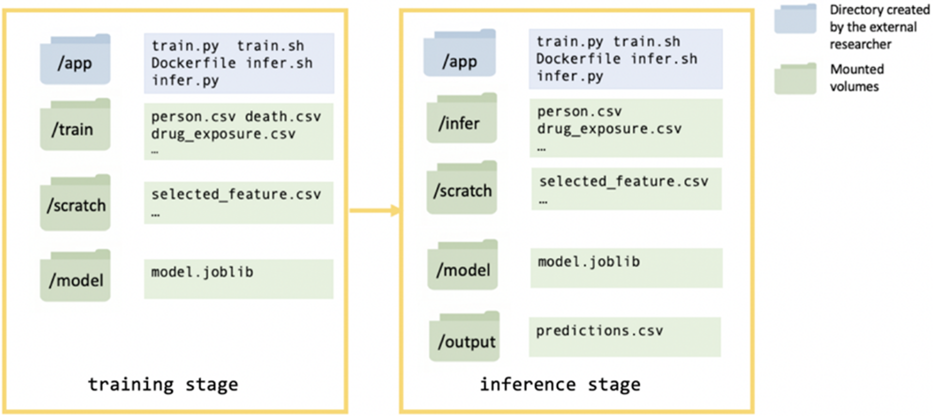


Figure 3.2: Schema showing the Docker container structure for the training stage and inference stage of running the Docker image.

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-operating characteristic curve (AUROC) score and returned this to the model developer through Synapse. When the Docker model failed, a UW staff member would look into the saved log files to assess the errors. Filtered error messages were sent to the model developer for debugging purposes. See Fig 3.3 for the full workflow diagram.

3.3.2 Model developer's perspective

The model developer built and submitted models, using 3 sets of features: (1) basic demographic information (age on the last visit date, gender, and race); (2) basic demographic information and binary indicators for 5 common chronic diseases (cancer, heart disease, type 2 diabetes, chronic obstructive pulmonary disease, and stroke)[64]; and (3) the 1000 most common concept_ids selected from the procedure_occurrence, condition_occurrence,

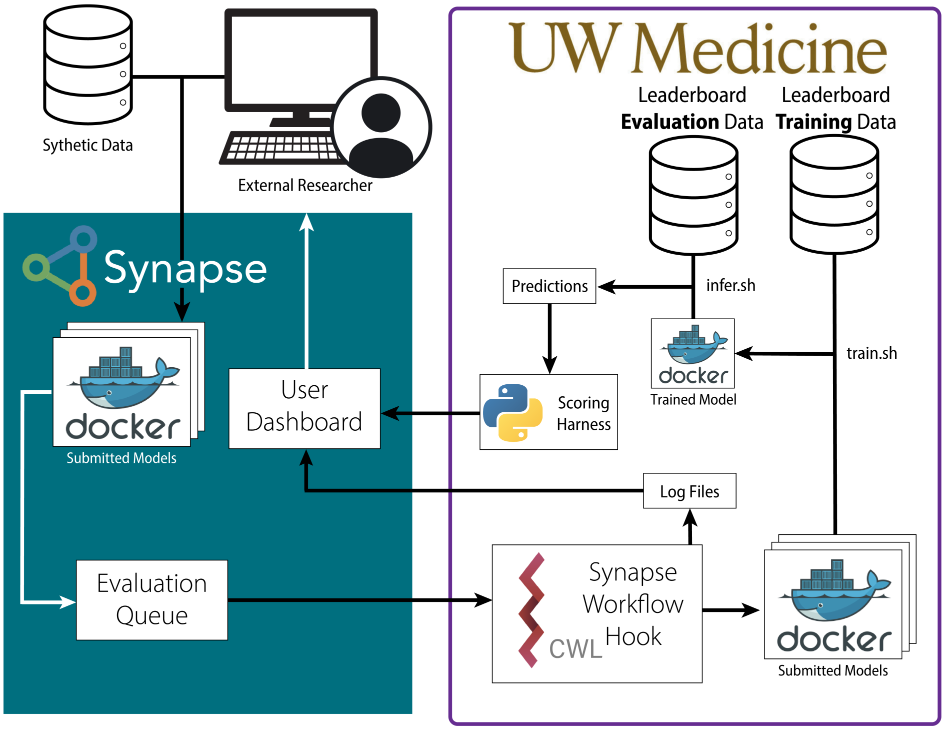


Figure 3.3: Diagram for submitting and distributing containerized prediction models in a protected environment. Dockerized models were submitted to Synapse by a model developer to an evaluation queue. The Synapse Workflow Hook pulled in the submitted Docker image and built it inside the protected University of Washington (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 with a gold standard benchmark. The model's performance, measured by area under the receiver-operating characteristic curve, was returned to the model developer. CWL: Common Workflow Language.

	Training set (n = 956,212)	Evaluation set (336,548)
Patients with cancer	66,203 (6.9)	42,195 (12.5)
Patients with heart disease	31,352 (3.3)	23,108 (6.9)
Patients with type 2 diabetes	40,938 (4.3)	28,234 (8.4)
Patients with chronic obstructive pulmonary disease	13,777 (1.4)	8,302 (2.5)
Patients with stroke	5,216 (0.6)	3,927 (1.2)
Other patients	834,591 (87.3)	257,884 (76.6)

Table 3.1: Number of patients in the University of Washington Medicine Observational Medical Outcomes Partnerships repository who have been diagnosed with cancer, heart disease, type 2 diabetes, or chronic obstructive pulmonary disease

and drug_exposure domains in the OMOP dataset. For model 2, the developer used the OMOP vocabulary search engine, Athena ("Athena" n.d.), to identify 404 clinical condition_concept_ids associated with cancer, 76 condition_concept_ids with heart disease, 104 condition_concept_ids with type 2 diabetes, 11 condition_concept_ids with chronic obstructive pulmonary disease, and 153 condition_concept_ids with stroke (Table 1). Logistic regression was used on the 3 sets of features respectively. All model scripts are available online (https:// github.com/yy6linda/Jamia_ehr_predictive_model).

3.3.3 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 concepts,

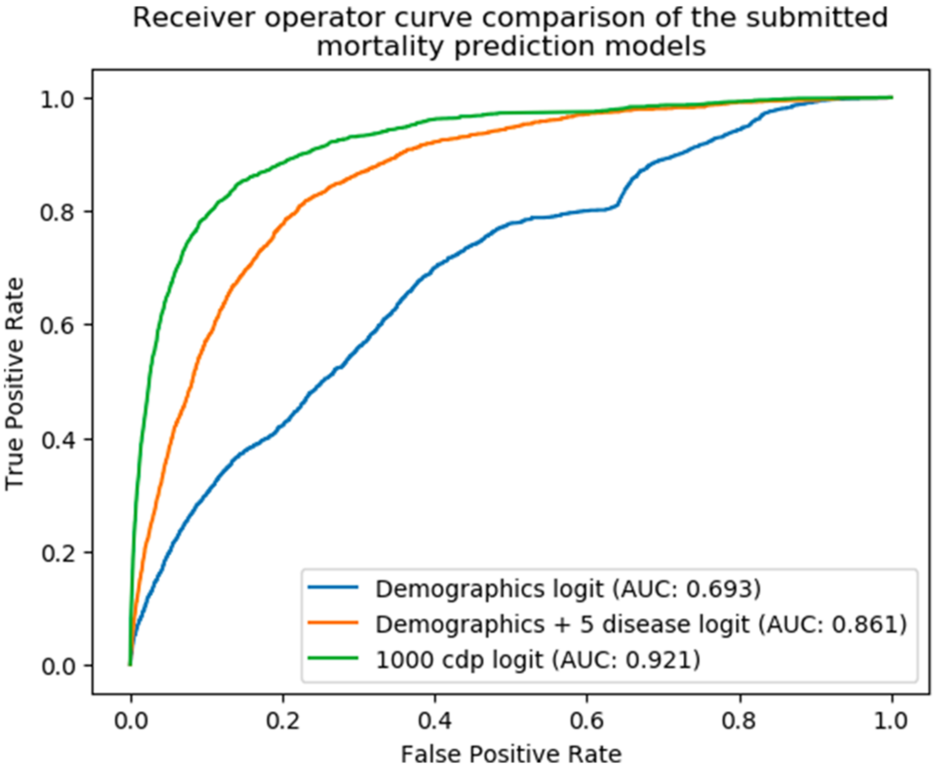


Figure 3.4: A comparison of the receiver-operating characteristic curves for the 3 mortality prediction models submitted, trained, and evaluated using the “Model to Data” framework. AUC: area under the curve; cdp: condition/procedure/drug.

yielded an AUROC of 0.921 (Fig 3.4).

3.3.4 Benchmarking the capacity of fixed computing resources for running predictive models

We tested the capability of running models through the pipeline on increasingly large feature sets using 2 machine learning algorithms: logistic regression and neural networks. The models ran under fixed computational resources: 70 GB of RAM and 4 vCPUs (a quarter of the total available UW resources made available for this project). This tested the feasibility of running multiple (here, 4) concurrent, high-performance models on UW infrastructure

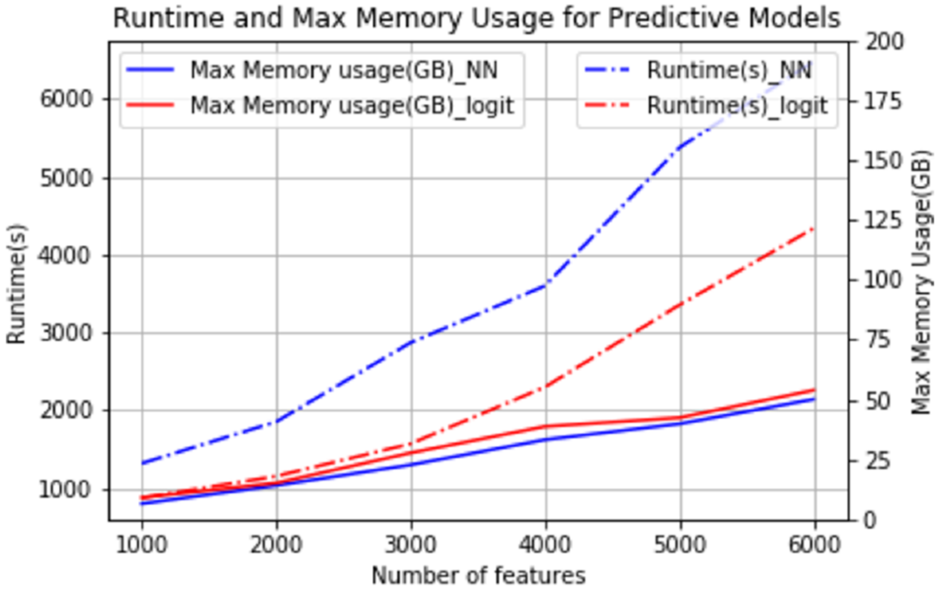


Figure 3.5: Runtime and max memory usage for training predictive models in the benchmarking test.

for a community challenge. A total of 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 2 selected algorithms were applied to a subset of the features of 1000, 2000, 3000, 4000, 5000, and 6000 selected condition_concept_ids. We used the python sklearn package to build a logistic regression model and keras frameworks to build a 3-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 (Fig 3.5). While run times scale linearly with the number of features, maximum memory usage scales in a slightly superlinear fashion.

3.4 Discussion

In this pilot project, we implemented the MTD 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 MTD 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 MTD 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 [65], i2b2 [66], or any other clinical data model. The focus of the MTD framework is to deliver containerized algorithms to private data, of any standardized form, without exposing the data. 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 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 MTD approach particularly appealing in certain contexts as discussed in the following sections.

3.4.1 MTD 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 [61]. 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 robustness 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

3.4.2 MTD 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 [62], CAFA [67, 68], CAGI [69], and CASP [70]. A key feature of some of these challenges is the prospective evaluation of prediction models, an often unmet need in clinical applications. The MTD approach uniquely enables such an evaluation on live EDW data. Based on our observations in this pilot study, we will scale up our platform to initiate an EHR mortality community challenge at the next stage, in which participants from different backgrounds will join us in developing mortality prediction models.

3.4.3 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. First, the model developer had multiple failed submissions due to discrepancies between the synthetic data and real data. Devoting effort toward improving the similarity 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. Second, 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 that would allow the return of all log files to support debugging. This would allow any major errors or bugs to be caught prior to the model running on the real data. Third, inefficiently written prediction models and their containers burdened servers and system administrators. The root cause of this issue was the model developer's difficulty in estimating 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. Fourth, 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 MTD framework is also compatible with more complicated machine learning algorithms and feature engineering. Future researchers can dockerize their complicated predictive models with more advanced feature engineering and send them through our pipeline as docker images. Our pipeline is able to 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 would be able to share that information for the model developer to further interpret their models. However, the remote nature of the MTD framework limits the opportunities for manual hyperparameter tuning which usually requires direct interaction with data. Hyperparameters are model parameters predefined before the models' training stages (eg, 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 and random search, reinforcement learning, evolutionary algorithms, and Bayesian optimization, allows hyperparameter optimization to be automated and efficient [71].

3.5 Conclusion

We demonstrate the potential impact of the MTD framework to bring clinical predictive models to private data 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 MTD approach in a real-world clinical analytics environment. We believe this enables future predictive analytics sandboxing activities and the development of new clinical predictive methods safely. We are extending this work to enable the EHR DREAM Challenge: Patient Mortality Prediction as a further demonstration.

Chapter 4

EVALUATION OF CROWD SOURCED MORTALITY PREDICTION MODELS ON AN ENTERPRISE DATA WAREHOUSE USING THE MODEL TO DATA FRAMEWORK.

4.1 Introduction

Applications of machine learning applied to patient data are being widely developed and implemented in healthcare scenarios [61, 72]. The performance of these methods as they are used in the clinic - and their associated impact on patient outcomes - are not well understood. An important risk in the design and implementation of machine learning algorithms is the self-assessment bias, where the implementer and evaluator are the same person or team, which can result in overfitting and poor generalization [73]. At the same time, health systems and journals are inundated with new methods and new personalized data platforms that are overwhelming the ability of healthcare providers to assess effective solutions. This is further exacerbated by different business practices along with vastly different data characteristics across healthcare institutions. Data can also contain hidden biases, reflecting social and institutional biases within the healthcare system [74]. Risks of biases in medicine have been well documented, and models built using biased data will promulgate these biases into practice through model recommendations [74, 75]. Addressing these issues requires a rigorous, unbiased framework that can evaluate algorithm performance using independent honest brokers, assess generalizability over time and across institutions, and report on performance disparities across sub-populations.

Access and use of EHR data for AI research is complicated by HIPAA laws and associated re-identification risks. To overcome these barriers to access, we have developed an approach, ”Model to data” (MTD), that delivers analytical models to protected data without sharing the data directly with model developers [56]. We previously piloted this method on an EHR dataset from the University of Washington and demonstrated the feasibility of accurate model development without the model developer having direct access to the patient data [76]. This approach has two benefits: (1) it protects patient data while allowing researchers to build machine learning methods and (2) it forces a more standardized and transferable approach to building models allowing the data host to perform rigorous evaluations of submitted models.

We leveraged this platform to implement the EHR DREAM Challenge: Patient Mortality Prediction to broadly assess machine learning approaches applied to a clinical data warehouse. We focused on the clinical question of predicting all-cause mortality since the clinical phenotype is clearly defined, University of Washington merges patient records with state death records to minimize missingness, and previous mortality prediction methods have been developed [64, 77, 78, 79]. DREAM Challenges are crowdsourced, biomedical competitions, where the challenge organizers solicit the broader research community to develop methods to answer a specific set of biomedical questions [62], and to assess these methods using hidden, gold standard datasets. Community challenges have proven to be a robust setting for objectively evaluating prediction models since they remove the researcher from the evaluation process [67, 68, 69, 80, 70], limiting the self assessment bias [73]. In this Challenge, we asked participants to predict whether patients would pass away within 180 days of their last visit to the UW medical system based on that patient's previous medical history. We evaluated these models, scoring them based on Area Under the Receiver Operator Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC).

4.2 Methods

4.2.1 The University of Washington clinical data repository

The UW Medicine enterprise data warehouse (EDW) includes patient records from clinical sites within the UW Medicine system including the University of Washington Medical Center (Montlake and Northwest Campuses), Harborview Medical Center, UW Neighborhood Clinics, and the Seattle Cancer Care Alliance. In aggregate, this network comprises more than 300 specialty and primary care clinics. The EDW gathers data from more than 60 sources, including laboratory results, microbiology reports, demographic data, diagnosis codes, medications prescribed, and procedures performed. An analytics team at the University of Washington transformed the patient records from 2010 to 2019 into a standardized data format, the Observational Medical Outcomes Partnerships Common Data Model (OMOP CDM v5.0) [46]. For the EHR DREAM Challenge, 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 covering approximately 10 years of patient histories.

Synthetic data

We derived a synthetic dataset from the SynPuf Synthetic OMOP dataset [81]. Starting with the original SynPuf dataset we adapted it to our challenge by randomly sampling concepts and terms that occurred more than 100 times in the structured EHR data from the University of Washington OMOP repository and then populated the tables of the original SynPuf dataset with these random samples to create a synthetic derivative that more closely represents the UW OMOP data. We also adjusted the size of the data to match the distribution of records across patients resulting in a synthetic dataset that represented 1,264,000 patients with 6300 true positives, 19,945,000 visits, and 189,605,000 measurements. This synthetic data was available to participants for local model development and debugging purposes, and was also used in the Challenge platform environment to validate that submitted models could be successfully executed. (Figure 4.1, Stage 1)

4.2.2 The EHR DREAM Challenge: Patient Mortality Prediction

The challenge infrastructure

The EHR DREAM Challenge was developed and run using a “Model to Data” (MTD) approach [56, 76]. This method relies on containerization software (Docker) [57], a common data model (OMOP) [46], a model intake mechanism (Synapse) [63], and a synthetic dataset for low risk technical validation of submitted models (Synpuf) [81]. Challenge participants

Demographic	Leaderboard Phase		Validation Phase			Post Challenge Resplit	
	Training (n=979,184)	Validation (n=284,883)	Training (n=942,381)	Validation (n=200,855)	Holdout Test (n=168,708)	Training (n=1,067,084)	Validation (n=273,597)
Age (%)							
0-17	6.12	6.38	6.18	7.31	6.04	5.62	9.42
18-34	23.77	22.18	23.84	24.61	20.98	22.07	29.2
35-64	46.82	45.31	46.68	44.84	45.58	47.36	41.12
65-99	22.86	26.03	22.85	23.13	27.33	24.55	20.09
100 +	0.35	0.09	0.37	0.1	0.06	0.34	0.1
Race (%)							
White	54.42	64.05	54.39	62.76	66.74	58.07	53.83
Asian	8.36	10.29	8.36	10.59	9.57	8.9	8.44
Black	6.3	7.41	6.22	6.82	8.39	6.81	5.51
Other/Nan	30.93	18.25	31.03	19.83	15.30	26.22	32.22
Gender (%)							
Female	52	54.2	51.92	53.82	54	52.59	51.6
Male	47.94	45.78	48.02	46.16	46	47.37	48.35
Other/Nan	0.05	0.01	0.05	0.02	0.01	0.04	0.05
Ethnicity (%)							
Hispanic	5.79	6.45	5.78	6.47	7.09	5.80	7.03
Not Hispanic	50.17	77.13	49.90	75.42	80.09	56.09	63.71
Other/Nan	44.04	16.42	44.31	18.11	12.81	38.11	29.26
Mortality Status (%)							
Passed	0.83	0.75	0.90	1.12	1.32	0.93	1.33
Alive	99.17	99.25	99.10	98.88	98.68	92.55	98.67

Table 4.1: Demographic makeup as a percentage of the individual data sizes across the different versions of data used in the DREAM Challenge. All values represent the percentage of the total number of patients in the dataset of interest. We include a 100+ category as a standalone age category because that age range is of questionable quality. This gives some idea to the quality of the data made available.

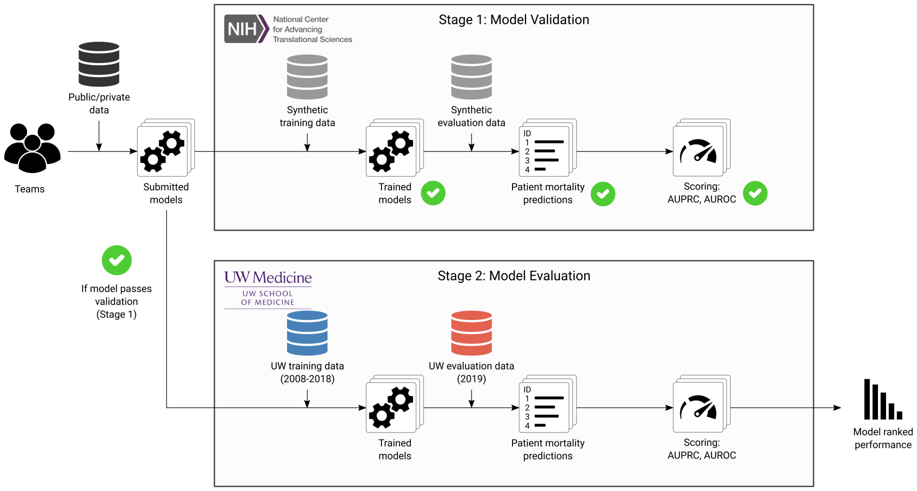

National Center for Advancing Translational Sciences
Stage 1: Model Validation
Public/private data
Synthetic training data
Synthetic evaluation data
Teams
Submitted models
Trained models
Patient mortality predictions
Scoring: AUPRC, AUROC
If model passes validation (Stage 1)
UW Medicine
UW SCHOOL OF MEDICINE
Stage 2: Model Evaluation
UW training data (2008-2018)
UW evaluation data (2019)
Trained models
Patient mortality predictions
Scoring: AUPRC, AUROC
Model ranked performance

Figure 4.1: "Model to data" architecture to evaluate the performance of EHR prediction models in the Patient Mortality DREAM Challenge. Models were developed on local environments using synthetic data that resembled the real private EHR data. Docker images were submitted through the Synapse collaboration platform to a submission queue. Images were pulled into the NCATS provided AWS cloud environment and run against a synthetic dataset for technical validation (Stage 1). Once validated, images were pulled into the UW Medicine secure infrastructure and run against the private EHR data. Model predictions were evaluated using Area Under the Receiver Operator Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC) which were returned to participants through Synapse.

were required to submit “containerized” models to be applied to protected data by the Challenge organizers. At no time during the Challenge did participants have direct access to real patient data. Participants were allowed to submit pretrained models using data they already have, such as their own institution’s clinical data warehouse. The containerized algorithms submitted by participants were able to use a training split of the UW dataset to train a predictive model de-novo, or to further optimize a pretrained model. To enable model debugging prior to evaluation on real clinical data, submitted models were first applied to the synthetic data to check for technical compliance (Figure 4.1, Stage 1: Model Validation). Log files generated by the models in the synthetic data environment were returned to participants for technical debugging purposes. Following successful execution on the synthetic data, the models were pulled into a University of Washington secure environment that was disconnected from the internet where they were trained on the UW OMOP repository. The UW OMOP repository did not contain patient identifiers. The models had no access to the internet during training and evaluation. (Figure 4.1, Stage 2: Model Evaluation) The trained models were tested on a holdout set and the Area Under the Recall Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC) were returned to participants via the Synapse platform. No logs, model parameters, or other information other than the performance metrics, were returned to participants after models were applied to the UW patient repository. Participants were allowed a total of 10 hours to train and test their models in this environment. The models were run on a server environment with access to 70 GB of RAM, 32 2.3 GHz CPU cores and no GPUs during this process.

Challenge Question

For this challenge, we asked participants to predict the 180-day all-cause mortality from the last patient visit at UW Medicine. True positives were defined as patients who had a death record in the first 180 days of their last visit record and true negatives were defined as patients who either had a death record more than 180 days from their last visit, or who did not have a death record. Death records were derived from the medical record and Washington State death records.

Timeline

The EHR DREAM Challenge lasted from September 9, 2019 to February 23, 2020 and was conducted in three phases: the open phase, the leaderboard phase, and the validation phase. During the open phase, participants could submit models for technical validation using only synthetic data in the Challenge cloud environment (Figure 4.1, Stage 1). During the leaderboard phase, technically validated models were applied to the leaderboard training data and evaluated against the leaderboard validation data (Table 4.1). We carried out this phase in 3 rounds, where each team was allowed 3 successful submissions per round. During the open and leaderboard phases, new data accumulated in the UW EHR. We gathered this data (called the "holdout test data") which represented patients who visited UW medical facilities during 2019. For the validation stage, models were trained on the leaderboard training data (excluding the patients transitioned to the holdout test data) and tested on this new prospectively-collected data. Scores generated from this holdout data were used to create final team rankings.

4.2.3 Model evaluation

Challenge Evaluation Metrics

The Area Under the Receiver Operator Curve (AUROC) was used as the primary metric for assessing model performance. A Bayes Factor, K, (bootstrapped distributions n = 10,000) was computed to determine if the AUROCs between two models were consistently different. If two models were found to have a small Bayes Factor ($K < 19$), we used the Area Under the Precision Recall Curve (AUPRC) as a tie-breaking metric. Both the AUROC and AUPRC were computed for all submissions and were used to rank teams on the Challenge leaderboard.

Re-split Data Validation

During the challenge, we used the validation phase holdout test set to build a final ranking of model performance. This holdout set only contained patients who appeared in the UW medical system in 2019. This left a 6-month longitudinal gap between the training data and holdout test data containing the leaderboard validation data which was not used in the final model scoring. We combined all the datasets (training, validation, and holdout test) and re-divided the dataset into an 80/20 split between training and testing data using the same prospective splitting method we used to split the initial leaderboard data. We trained the models on the 80% training data and evaluated the trained models against the 20% test data. This allowed us to compare the effect of the 6-month gap on model performance.

Subpopulation Accuracy Comparison

We evaluated model performance across various subpopulations which were defined by different demographic or clinical features including race, gender, ethnicity, age, and type of last visit. We compared model accuracy for each subpopulation against the accuracy of every other subpopulation within the same demographic or clinical group, calculating a paired Bayes Factor to evaluate the magnitude of accuracy differences. We ran this experiment on the prospectively gathered validation phase data.

4.2.4 Model Features

The top 5 performing teams were asked to output a list of features with accompanying weights in order to assess which features were most important in their models. We gathered a list of information that included the name of the feature, the "concept_id"s that were used to build those features, along with a weight or score that indicated the impact of that feature. Using these feature importance metrics, we randomized or removed the highest impact features from the prospective validation phase data and re-evaluated all the models, assessing the impact each feature had on each of the models. Feature importance was calculated as the percentage decrease in model performance, where 0% decrease meant there was no decrease in model performance and 100% decrease meant the model performance had decreased to 0.5 AUROC.

4.3 Results

The EHR DREAM Challenge on all-cause patient mortality prediction was held between September 9, 2019 and February 23, 2020 and resulted in 132 successful submissions from 25 teams over the course of the challenge. The overall design and workflow of the challenge is shown in Figure 4.1. Table 4.1 presents the training, validation, and holdout testing data sets that were used across the leaderboard and validation Phase. The validation phase training data contained 942,381 patients while the validation phase holdout testing data contained 168,708 patients, with mortality rates of 0.90% and 1.32% respectively. (Table 4.1)

During the leaderboard phase, of the 25 successfully validated models, ten teams exceeded $AUROC > 0.9$. AI4Life led the leaderboard phase - achieving an AUROC = 0.979 (0.977, 0.981) and AUPR = 0.614 (Table 4.3). In the final validation phase, 15 teams submitted successfully validated models, with three teams achieving an $AUROC > 0.9$ (Table 4.3, Figure 4.2). The top performing team, UW-biostat, achieved an AUROC=0.947 (0.924,0.952) and an AUPR=0.478. Between the leaderboard and validation phases, the average decrease in AUROC was 0.069 with the top 5 models decreasing by an average of 0.024 and the bottom 8 models decreasing by an average of 0.10. The top 5 models from the validation phase were ranked second, third, 11th, 10th, and 7th respectively at the end of the leaderboard phase but were ranked in the top 5 due to having the lowest decrease in performance.

Teams used a variety of machine learning techniques in their submitted models. Of the 15 validated models, 12 were boosted methods (LightGBM [82], XGBoost [83], CatBoost

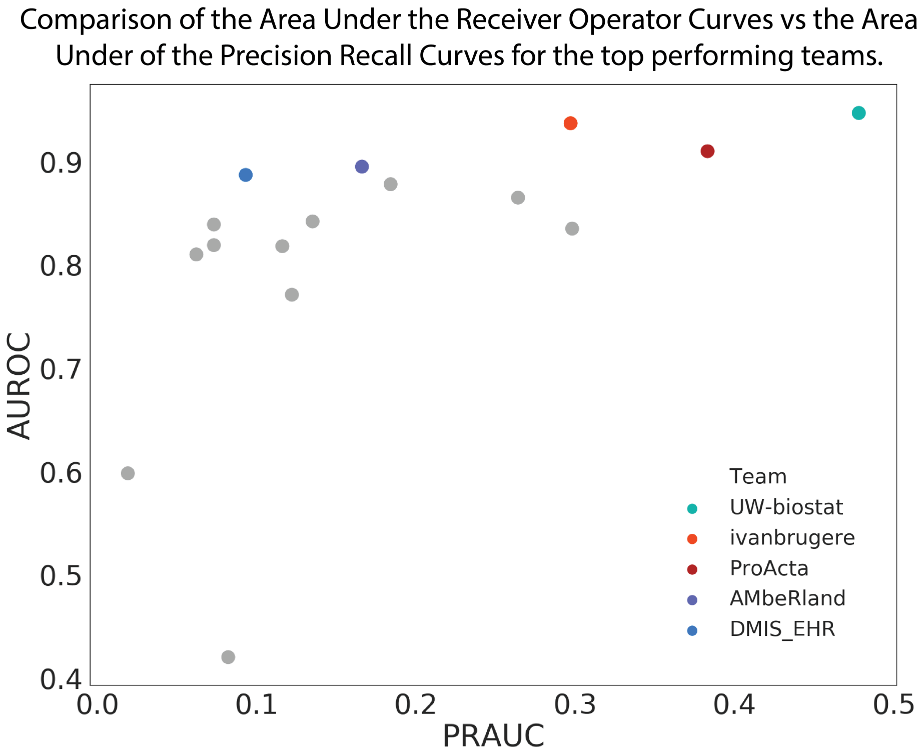


Figure 4.2: The Receiver Operator Curves and the Precision Recall Curves from all the models submitted in the validation phase. The top 5 models were not the top 5 models in the evaluation phase, but were more robust to longitudinal changes in the data. AUROCs and AUPRCs are reported in Table 4.3 for the leaderboard data, validation data, and the Resplit data. Comparison of leaderboard phase to validation phase scores are found in Figure 4.4.

Team	Leaderboard Phase			Validation Phase				Post Challenge Resplit		
	AUROC	AUROC 95% CI	AUPR	AUROC	AUROC 95% CI	Delong p value	AUPR	AUROC	AUROC 95% CI	AUPR
UW-biostat	0.972	(0.969, 0.975)	0.524	0.947	(0.942, 0.951)	1.70E-04	0.478	0.964	(0.961, 0.967)	0.43
Ivanbrugere	0.968	(0.964, 0.971)	0.474	0.938	(0.933, 0.942)	1.96E-07	0.3	0.956	(0.953, 0.96)	0.409
ProActa	0.943	(0.937, 0.948)	0.458	0.91	(0.903, 0.918)	2.84E-03	0.383	0.904	(0.898, 0.91)	0.43
AMbeRland	0.942	(0.937, 0.947)	0.288	0.897	(0.89, 0.903)	4.18E-02	0.163	0.929	(0.924, 0.934)	0.284
DMIS_EHR	0.915	(0.91, 0.92)	0.111	0.887	(0.88, 0.89)	5.95E-02	0.093	0.939	(0.936, 0.943)	0.347
PnP_India	0.958	(0.954, 0.963)	0.449	0.876	(0.87, 0.883)	1.26E-01	0.182	-	-	-
ultramangod671	0.882	(0.874, 0.891)	0.289	0.865	(0.856, 0.875)	2.45E-03	0.264	0.868	(0.86, 0.876)	0.37
HELM	0.951	(0.948, 0.955)	0.323	0.842	(0.834, 0.85)	5.65E-01	0.135	-	-	-
AI4Life	0.979	(0.977, 0.981)	0.614	0.831	(0.82, 0.841)	5.28E-01	0.302	0.971	(0.969, 0.974)	0.63
Georgetown - ESAC	0.938	(0.933, 0.942)	0.168	0.839	(0.832, 0.848)	1.64E-02	0.073	0.938	(0.933, 0.941)	0.272
LCSB_LUX	0.956	(0.952, 0.959)	0.307	0.82	(0.81, 0.829)	8.41E-01	0.116	0.936	(0.932, 0.94)	0.201
QiaoHezhe	0.925	(0.92, 0.93)	0.16	0.819	(0.81, 0.827)	2.92E-01	0.073	-	-	-
chk	0.903	(0.896, 0.908)	0.159	0.808	(0.8, 0.817)	1.38E-05	0.062	0.811	(0.804, 0.818)	0.061
moore	0.955	(0.951, 0.958)	0.313	0.771	(0.757, 0.784)	9.51E-45	0.122	0.947	(0.943, 0.95)	0.377
tgaudelet	0.904	(0.898, 0.91)	0.278	0.807	(0.798, 0.817)		0.201	0.158	(0.151, 0.166)	0.007

Table 4.2: Top 15 teams and the metrics for their highest performing models. 95% confidence intervals were calculated using bootstrapped (n=1000) distributions. The Delong test p value was generated by comparing each team's model with the team's model ranked below them. Leaderboard phase scores were generated using the models submitted during the final validation phase.

[84], Generalized Boosted Regression [85]), 2 were logistic regression, and 1 was a neural network. Of the top 5 models, 2 were LightGBM, 1 was logistic regression, 1 was CatBoost, and 1 was Generalized Boosted Regression. Each model used a different feature selection method ranging from randomly sampling all available concepts (Team IvanBrugere), carefully selecting a few features from the literature (Team LCSB_LUX), and using the structure of the concept ontologies to roll up low-level granular concepts into broad categories of disease and drugs as features (Team UW-biostat).

4.3.1 Top performing model

While UW-biostat's (University of Wisconsin-Madison, Biostatistics and Medical Informatics) model was not the highest scoring model during the leaderboard phase, their model had the highest score for the validation phase and had the smallest decrease in performance of any model between the two phases. The team used ontology-rollup to reduce feature dimensionality and used time binning and sample reweighting to capture longitudinal characteristics. For model development, they trained and tuned a LightGBM model to predict the mortality risk of each patient. To take into account potential data drift in EHRs [86, 87, 88, 89], the team upweighted more recent patients during optimization and training of their model. In order to validate the model's "future-proof" ability, they ordered the labeled patients by their last visit date from recent to early and used the top 15% of patients for validation.

4.3.2 Demographic Evaluation

To assess whether models generalized across patient subpopulations, we evaluated model accuracy across multiple demographic and clinical groups including race, gender, age, ethnicity, and last visit type. We generated bootstrapped distributions ($n = 10,000$) for each category in each model and ran paired permutation tests, calculating Bayes factors to assess the level of evidence for performance differences between subpopulations.

Models were consistently more accurate on Asian patients when compared to any other racial group (Figure 4.3, Table 3), despite Asian patients only making up 8.4 percent of the validation data and 9.6 percent of the validation phase training data (Table 4.3). Methods

Rank	UW-biostat	IvanBrugere	ProActa	AMbeRand
1	Age	Not for resuscitation	Age	Age
2	Average pulse	Temperature	Creatinine in Serum	Not for resuscitation
3	Average Diastolic Blood Pressure	Albumin in Plasma	Heart rate	Antineoplastic chemotherapy regimen
4	Average Systolic Blood Pressure	History of clinical finding in subject	Inpatient Visit	Lactate dehydrogenase (LD), (LDH)
5	Latest Systolic Blood Pressure value	Natriuretic peptide B [Mass/volume] in Serum or Plasma	Blood typing, serologic; ABO	Administration of antineoplastic agent
6	Year of last visit	Racial Variable (White)	Palliative care	Patient encounter procedure
7	Latest pulse measurement	Antineoplastic chemotherapy regimen	Albumin in Plasma	Secondary malignant neoplasm of lung
8	Latest Diastolic Blood Pressure Value	Protein in Plasma	Hematocrit of Blood by Automated count	Disorder of lung
9	Latest Glucose measurement	Heart rate	Neutrophils/100 leukocytes in Blood by Automated count	Dexamethasone
10	Indicator: Unknown conditions	Essential hypertension	Cholecalciferol	Bacterial culture

Table 4.3: The top 10 weighted features as reported from the top 4 performing teams. Features were ordered by their model weight and assigned a rank out of all available features. Feature names were either reported by the teams or were mapped using the OMOP concept table from the reported concept ids.

varied in their accuracies for other races with some models (e.g. UW-biostat, IvanBrugere, Proacta, AMbeRland, DMIS_EHR) scoring higher on White patients compared to Black patients, and others scoring higher on Black patients than White patients (PnP_India, HELM, Georgetown-ESAC, AI4Life) (Figure 4.3, Table 3).

Without exception, models were more accurate on female patients than on male patients with Bayes factors greater than ten (strong evidence) for nine of the top 15 models (Figure S4). As the challenge asked participants to predict mortality status 180 days from the last visit, we examined whether there were differences in model performances based on whether the last visit was inpatient, outpatient, or an emergency room visit. Most models had lower accuracy when the last visit was an outpatient visit, with the exception of 3 models (ultramangod671, Georgetown - ESAC, AI4Life in Figure 4.4). On patients where the last visit was an emergency room visit, models showed a wide variety of accuracies. In a few cases, models that had an overall lower model accuracy had higher accuracies on patients in the emergency room (compare ProActa to PnP_India in Figure 4.4).

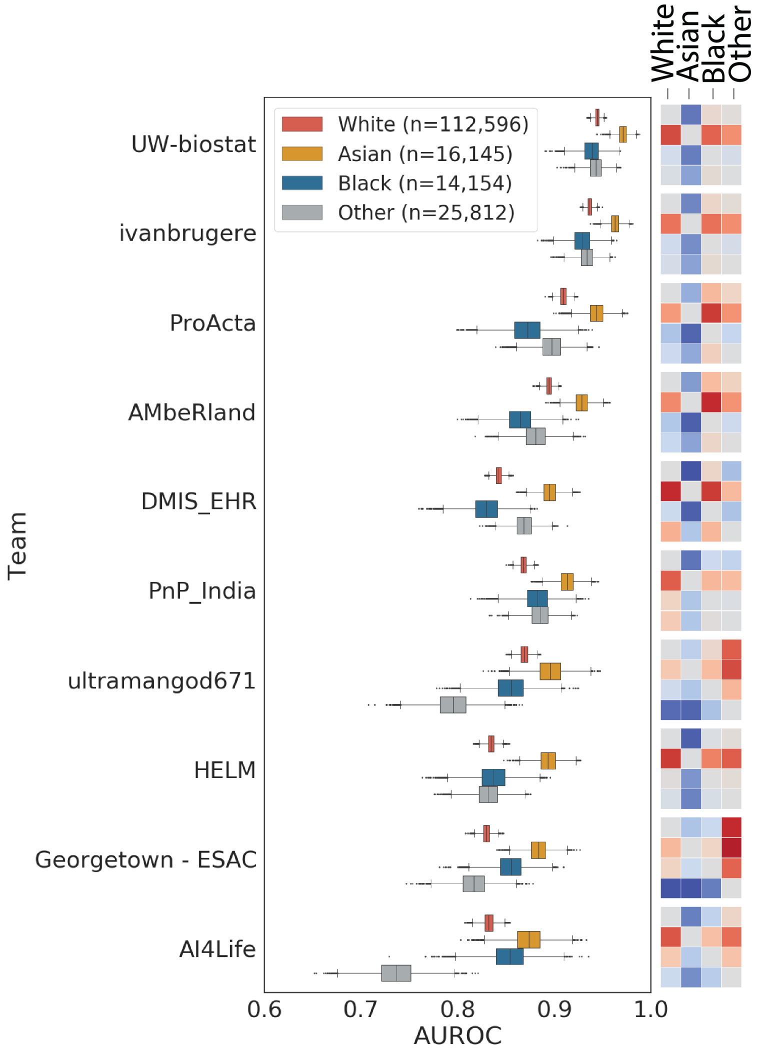


Figure 4.3: Bootstrapped distributions ($n = 10,000$) of the top 10 model AUROCs broken down by race. Model predictions were randomly sampled with replacement and scored against the benchmark gold standard. Comparisons were made between each category of race and Bayes values calculated to assess the level of evidence for the model having a higher accuracy on racial category compared to another category. The heat maps represent the log of the calculated Bayes factors when comparing racial groups within each model. The darker the red, the stronger the evidence for the racial category being higher than the comparison category. Bayes factor values range from 10000 to 0.0001. The darker the blue, the stronger the evidence for the racial category being lower than the comparison category. The color scale is normalized across all comparisons.

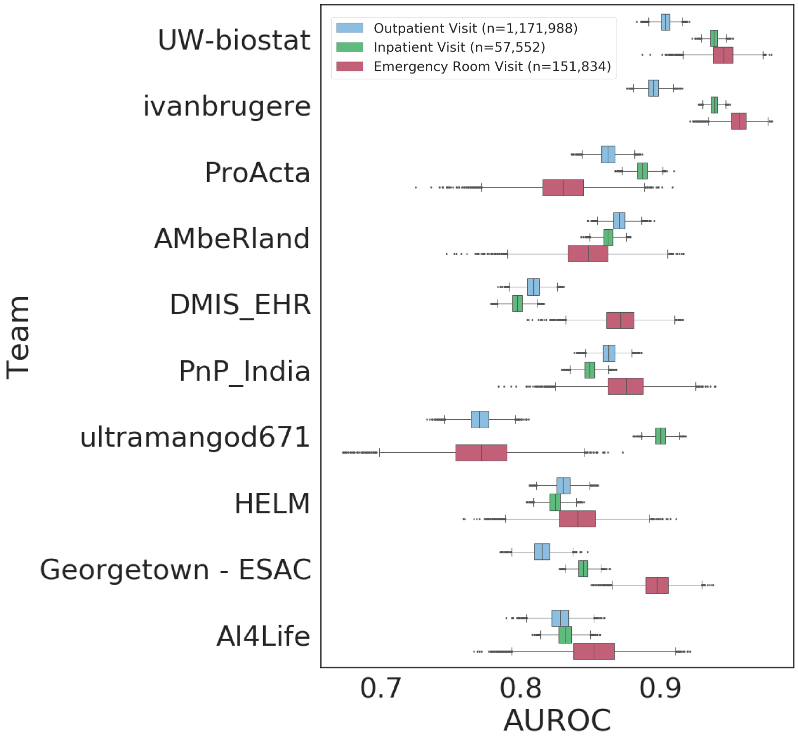


Figure 4.4: Bootstrapped distributions ($n = 10,000$) of the top 10 model AUROCs broken down by the last visit type of the patients. The visit types evaluated were Inpatient Visits, Outpatient Visits, Emergency Room visits.

4.3.3 Assessment of important features

The top five highest scoring teams were asked to adjust their dockerized models to output their trained features as a list of codes/values with associated weights from their trained models. In order to compare features across models, these teams reported which terms (SNOMED, RxNorm, LOINC, etc.) were used during any feature engineering and submitted brief descriptions of those features. All teams implemented a form of dimension reduction to operate in a far smaller modeling space than the 1.2 million unique granular concepts available in the UW dataset. Team UW-biostat derived 1405 weighted features from 1.2 million unique granular concepts; Team IvanBrugere used 1479 weighted features representing 1479 unique granular concepts; Team Proacta used 224 weighted features representing 22,934 unique granular concepts; Team AMbeRand used 35 weighted features representing 35 unique granular concepts; and Team DMIS's model used a Principal Component Analysis to generate principal components as features for their trained model. We were not able to reliably extract feature importance from their model due to the nature of their model and due to the current limitations of the "Model to data" infrastructure.

The top four teams were able to successfully extract the features and weights from their models and output them into a human readable file. Table 3 reports the top 10 features of each model, including engineered features (i.e. presence or absence of a category of diagnosis or drug) and raw concepts from the data (i.e. granular SNOMED or LOINC codes). Some of the highly weighted features included the age of the patient at their last visit, systolic and diastolic blood pressure, heart rate, and a code for Do Not Resuscitate. The impact of the highest weighted features was evaluated across all methods by either randomizing or removing features from the validation phase holdout test set. Figure 6 graphs the effect of these experiments as a percentage decrease in model performance. Age was the most important feature in most models, with the highest impact representing a 49% decrease in model performance for QiaoHezhe. While the code for “Not for resuscitation” was among the more highly weighted in the feature reporting, it was the most impactful feature in only one model, decreasing ultramangod671’s model performance by 20%, and only causing a 3% decrease in IvanBrugere’s model and a 2.3% decrease in AI4Life’s model.

4.4 Discussion

Data-driven, machine learning models are increasingly regarded as foundational to any precision medicine strategy. The assessment of model accuracy and utility in a healthcare environment is challenged by lack of resources to rigorously evaluate models, limited data availability, concerns about breach of protected health information confidentiality, and lack of infrastructure to manage model evaluations. We proposed and implemented an architecture for an unbiased and transparent assessment of methods that overcomes those limitations, and in doing so were able to improve existing methodology. We demonstrate how community challenges can provide an inclusive and rigorous environment for hosting a multisite machine learning clinical trial.

Using the ”Model to data” framework, 25 international teams were able to submit machine learning models to a private clinical dataset that otherwise would have remained in-accessible to these researchers. This was enabled by leveraging a common data model, in this case OMOP, a synthetic dataset for technical development and validation, a cloud environment hosting the synthetic data for pipeline and execution evaluation, standard containerization software, and a secure environment hosting the private clinical data.

For the first EHR DREAM Challenge, we asked participants to predict 180 day mortality status of patients in the University of Washington clinical records. We chose all-cause mortality prediction since clinical records can be linked to state death records, yielding a more complete gold standard. However, this prediction question is not immune to censoring, and is still susceptible to an open world limitation as some patients may die out of state or outside UW clinical care without the ability to map their death to UW clinical records[88]]. Another limitation, as shown both in the selected features (Table 3) and in model accuracy across last visit type (Figure 4.4) , is that all-cause mortality is not a clinically actionable question, as models trained for all-cause mortality are not specific enough for clinical action. Given that some of the highlighted features used to predict death included age, palliative care, designation for do not resuscitate, and treatments or diagnoses of cancer (neoplasms), many of the models developed in this challenge were identifying the most obvious patients in the UW population. However, the highest performing model was assigning the most weight on specific measurements and their values, not just the presence or absence of a measurement or condition, and was utilizing the hierarchical design of the available biomedical ontologies to “roll up” low level granular codes, which does point to the fact that the ”Model to data” framework can be used in concert with intelligent feature engineering and selection.

Assessing a wide variety of methods from teams allowed us to evaluate the best approaches and assess inter-method variability when holding the evaluation data constant. Interestingly, even though models were trained and evaluated on the same data, there was variance in model accuracy across different demographic groups. White, Black, and other racial groups showed differences across models, with some models scoring higher on Black patients than White patients and vice versa, while Asian patients were consistently more accurate across nearly all models. This may have to do with the cause of death, as prevalence of different causes of death may vary between different populations. Unfortunately, we did not have access to cause of death data at the time of this analysis. With the exception of the 0-17 age group, method accuracy was inversely correlated with age (Figure A.3). Younger patients who pass away in 180 days and are coming into the hospital are more likely to have extreme conditions and have a higher risk of death, while older patients are simply more likely to have diseases and health problems in general, making it more difficult to predict risk of death. Models also had varied accuracies from the last visit type, highlighting the need to develop context specific clinical prediction algorithms, although the two most accurate models were still more accurate on outpatient visits than all other less accurate models (Figure 4.4). In other cases, models were aligned in their bias, universally scoring higher on females than males (Figure (A.2)). We found no meaningful difference in model accuracy between Hispanic versus Nonhispanic ethnicity. (Figure A.1).

Evaluating models in a pseudo-prospective manner allowed us to assess how models would perform over time in the UW environment. We found that most models decreased in perfor-

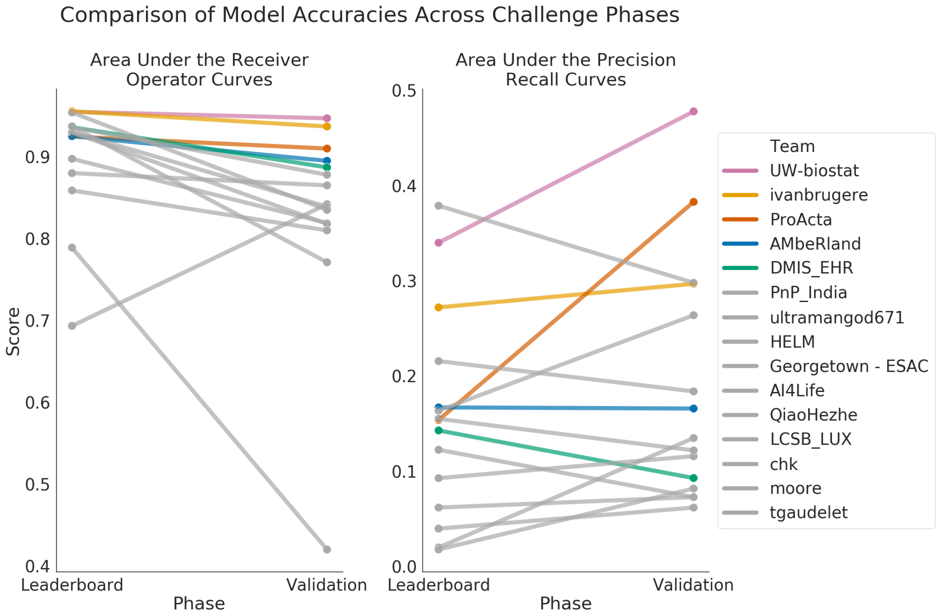


Figure 4.5: Comparison of model performance between the leaderboard phase and the validation phase. All models decreased in AUROC except for HELM while many models increased in AUPRC. The top 5 team AUROCs decreased the least between the two phases.

mance in the validation phase when compared to the leaderboard phase (Figure 4.5). This is in line with the literature, as previous studies have shown that the utility of clinical data can have a half-life of as little as three months [89]. Comparing results from the post challenge resplit data to the validation phase final results, the majority of models performed better when the training data was longitudinally closer to the test data (Table 4.3). UW-biostats explicitly down weighted older data and relied more heavily on the most recent 6-months of data in the training dataset resulting in their model having the lowest decrease in performance of any model. By combining this technique with their “roll up” feature engineering, UW-biostat’s model proved to be the most robust against longitudinal changes and concept drift. In contrast, some models dropped by a significant margin. For instance, AI4Life’s model was among the most accurate during the leaderboard phase, but dropped to tenth in the validation phase (Table 4.3). While it is difficult to completely account for their drop, one possible explanation is their overall lower accuracy on first time visiting patients (Figure A.4) combined with the increase of first time visitors in the validation data (leaderboard data - 13.8% compared to validation data - 19.7%). AI4Life's model had a high score on the resplit data, indicating that their model was susceptible to concept drift as well. Evaluating models prospectively or pseudo-prospectively evaluates models in the manner in which they will be used. Assessing models on prospective hold out data ensures that we understand how the models will perform in a live clinical setting.

4.4.1 Limitations

For this challenge, we set a submission limit of 10 hours for both the training and testing stages of model runtime. While this was implemented to limit the burden to the University of Washington secure servers, this also limited the types of models participants were able to build and excluded deep learning or more sophisticated, time consuming models. However, limiting the time also forced participants to carefully consider the efficiency of their algorithms so extraneous calculations or operations were not run.

Training and evaluating on data from a single site limited our ability to control for overfitting. While we did prospectively evaluate models on a future holdout set to try and control for overfitting, evaluation on data from one site does not properly assess model generalizability. For future assessments, we hope to partner with other hospitals to externally validate models.

4.5 Conclusion

Machine learning promises to enhance patient care and improve health outcomes; however, if not properly vetted and evaluated, risks and negative effects may be introduced. These risks include breach of privacy in the development and assessment of methods, inaccuracy or methodological bias when deployed, and the graduate loss of accuracy over time as data and business practices change. This study highlights these challenges by showing that while highly accurate methods are possible, even methods from the best of the best data scientists have considerable variability and that variability (such as differences in accuracy based on race or gender) may not be detectable from high level measures such as AUCs or accuracy. Our framework enables this assessment and also brings the community challenge culture to private datasets, in this case data that is subject to the HIPAA privacy rule. Further, machine learning methods may be able to address some causes of treatment disparities but may cause others for patients without rich longitudinal data, patients of certain races, gender or age. Based on these results, we believe that multisite standardized architecture is required to truly assess new methods and that independent oversight is required in their assessment.

Chapter 5

ENABLING PATIENT-DRIVEN RESEARCH TO PROMOTE PRECISION MEDICINE USING THE REDCAPTOWORDPRESS PLUGIN

5.1 Introduction

5.1.1 History of Direct Patient Interaction Approaches in Medical Research Study

Patient involvement in tailored research may become more important as we move toward personalized medicine. Patient-driven research is when the subjects or community in a research study take an active role in the research project design or implementation [90]. Studies have shown that patient-driven research increases the levels of participation and the amount of data gathered by the study, leading to higher success rates of the studies [91, 92]. This research method has recently gained traction in the medical research community as a way to engage patients in the research process. The major focus has been on rare diseases, where it can be difficult to recruit large cohorts for randomization, and n-of-one insights may impact medical care [93, 94, 95, 96, 97]. Many studies work with rare disease advocacy groups who have already built up a community and a pool of funds [98, 99, 100]. Patient-driven studies and patient registries rely on varying levels of patient involvement and could benefit from improved methods of communication between patients and researchers.

5.1.2 Overview of developed data capture platforms

Gathering data from patients generally requires a data capture platform, since patients are often immobile, or are spread across the country or the world. Web-based platforms have been developed to gather and manage patient generated data. The RUDY (Rare UK Diseases Study) study platform was developed as a way to handle patient enrollment, consent, and data capture for patient-driven research of rare diseases.[101] The platform also enables two way communication between the researcher and the patient. While RUDY does address the dispersed patient data capture question, the adaptability of the platform to other research questions is limited. At the time of this writing, no downloadable application was available, and, as far as we can tell, the skills necessary to adapt the platform to address new variations of patient-researcher interaction would be beyond those of the average biomedical researcher.

5.1.3 REDCap data capture platform

The Research Electronic Data Capture (REDCap) application is a popular data capture platform that, at the time of this writing, has over 600,000 users and 480,000 projects across over 2700 institutions around the world.[102] REDCap is an online research management tool designed to facilitate data capture (e.g. surveys, experimental results, etc.), data analysis, and secure data storage ensuring a HIPAA compliant environment. While not specifically designed for patient interactions, REDCap offers the flexibility to tailor the data capture tools while having an extensive API that allows external application development and integration.

REDCap is a useful tool for data management in terms of security and analysis; however, it does not lend itself to acting as a secure study patient portal. The user management system is designed for administrators of the study and can only display researcher views, not study participant views. REDCap's API has allowed it's use to be extended in projects such as the R package redcapAPI [103] and the python wrapper PyCap [104]. These extensions have been focused on researcher use and data analysis, and to date, no application that makes use of the API has used it to develop a patient portal.

5.1.4 Patient-Driven Research Requires a Patient Portal

The focus of most patient-driven research in medicine has been on rare diseases. However, many of these different projects have been more focused on enrolling patients for studies and then linking them to researchers, not enabling patients to actively participate in their own research study implementation. These projects are generally described as "registries" for different rare diseases, and are seen as a source for cohort discovery for studies and clinical trials [101, 105, 106]. Many of these registries follow the Rare Disease Registry Framework (RDRF), which implements a modular approach to data collection [107]. These solutions, however, are more researcher and research administrator focused. The patient-driven research framework offers a way for patients to take control of the research for their own conditions. While rare disease registries are useful for linking patients with researchers, these are not sufficient for returning information to the patients. Many studies will need informatics solutions that facilitate study progress updates, return of results, and patient data editing and validation.

5.1.5 The FindMyVariant Study Overview

FindMyVariant is a study helping individuals, who in the course of their medical care have had genetic variants of uncertain significance (VUS) identified, find more information about the VUS that has been identified. As clinical genetic sequencing has become more common, an increasing number of patients have had a VUS identified. When a new, family-specific variant is identified it is often classified as a VUS because there is not sufficient information to determine if it is associated with disease. Understanding the clinical importance of VUS is important to patients and families as VUS may increase patient anxiety and lead to improper medical treatment [108, 109, 110]. Family analysis, including cosegregation analysis is a strategy to find information about specific VUS. If a VUS is reclassified, this can lead to appropriate prevention and treatment [111].

Several clinical diagnostics laboratories perform family analysis for selected families [112], but building pedigree information, contacting, consenting, genotyping multiple relatives, and performing statistical analysis requires substantial research time and resources. FindMyVariant is a unique family analysis study, as it focuses on patients taking charge of pedigree building activities. Each variant analysis is patient-initiated as patients contact the study directly after they have received a clinical report with a VUS. Patients who are motivated and able to carry out the family member recruitment are enrolled, with the research team acting as a consultant, performing genetic sequencing, and conducting statistical analysis. After the patient contacts the study, the study provides education on relevant research rationale and methodology (findmyvariant.org), access to genotyping, and statistical expertise.[112]

The patient-driven framework distributes the time-consuming work of gathering information relevant to building pedigrees to those with intimate knowledge of their family dynamics. Study genetic counselors transcribe a pedigree as dictated by the patient. Patients invite family members to participate, and the study sends out genotyping kits to relatives. Genotype results are used in family cosegregation and other analyses. All results are returned directly to participants as directed by the individual participants' or relatives' consent for their own data sharing. The ultimate reclassification results are related to the patient's and their family's medical care as variant classification is likely to benefit patients and relatives. This patient-driven process generates many more tested relatives than researcher-driven research processes with the same goal (manuscript in preparation) and is well accepted by patients (manuscript submitted) and their relatives (manuscript in preparation).

5.1.6 Objective

Originally, all the data collected by the FindMyVariant patients and the findings by the researchers were shared over the phone and through email. We needed a new interface to enable direct patient data input into the research REDCap project, as well as an interface to report research progress and return of results. The objective of the study was to develop a WordPress [113] plugin that would integrate the adaptability of the REDCap data capture platform with the ease of use and intuitiveness of WordPress. The REDCapToWordPress plugin needed to enable intuitive data entry into the REDCap project, keep control of the user signups in the hands of the researcher, and ensure security and privacy of the patients.

We demonstrate the feasibility of REDCapToWordPress in the FindMyVariant study and show that the two, user-friendly and established platforms can be adapted for patient-driven research.

5.2 Methods

5.2.1 WordPress Plugin for Enabling Patient Access to REDCap

Findmyvariant.org was a static informational website developed in WordPress using the Themosis framework [114]. We developed a WordPress plugin that would display patient data from the REDCap study project and allow the participant to edit and view existing data and enter new information directly into the linked REDCap project, without having to log into REDCap as a user. We developed the plugin using HTML, CSS, JavaScript, and PHP. Our plugin manages user signup, user sessions, and will display both an administrator view for controlled participant enrollment and a participant view for reading and editing select study information. The source code for the plugin is available under an open source license in our GitHub repository (https://github.com/UWMooneyLab/REDCapToWordpress).

5.2.2 REDCap survey as a data collection tool

We use a developed, HIPAA compliant data storage system (REDCap), and designed the patient portal to interact with REDCap's Application Programming Interface (API) using the built in functionality of the REDCap survey to facilitate data input by the study participant. REDCap surveys are generated on top of existing REDCap project records and will take participant answers and directly input them into the REDCap project. Surveys can be customized to hide certain variables from the end user (e.g. Genetic test results of relatives), allowing the researcher to ensure private data protection. Normally, surveys are for one time use and are not designed to facilitate repeat returns to the form to edit information. Options are available to enable returns, but a unique link to the survey and a return code must be stored to make this happen. Generally, the onus is on the end user to keep track of this information; however, in building our patient portal, we designed the system to facilitate easy returns to the survey forms, doing the heavy lifting of tracking and storing pertinent return information by linking the information to the secure patient portal accounts in the WordPress site.

5.2.3 Linking the user to the REDCap record

In order to the link the patient user to their REDCap record, we stored the user's email in a custom two-column database table that links the unique email to the patient's REDCap record ID number. By default, WordPress needs the user email to make accounts, so using the user email allows seamless integration into WordPress. When the patient logs in, the email associated to their account is used to collect their study information from REDCap.

5.2.4 Information Flow and Security

Using the REDCap API, we were able to retrieve the patient record study variables that were entered by the researcher or the patient, and display them through the patient portal.

In order to pull data from a REDCap project, a unique token is used to identify users who are requesting access through the API. The disadvantage of this system is that anyone who has this token can access all data from the project and is not limited to individual records; thus, protecting this token is paramount to ensure security.

To address this problem, we handled all the API calls through a “middleman” server that acted as a risk limiting mediator, running a python Flask API application. The REDCapToWordPress plugin queries the middleman with the current user’s information. The middleman uses the user info to narrow down which records and variables are appropriate for the user to view, removing all other extraneous data, then returns a JSON object of the viewable data to the website. The token to access the REDCap project is only stored on the middleman server and the middleman server is only accessible from the website. (Figure 5.1) By limiting the flow of data to the bare minimum number of variables, we have designed the system to minimize the risk of data compromise to individual records rather than the whole project.

5.2.5 Application for University of Washington IRB approval

We applied for IRB approval to share FindMyVariant study data with the participants. This data sharing would include both their own data as well as their family information. We were allowed to share relative information with the main participant as long as the relative gave written consent. The IRB study request was approved by the University of Washington IRB.

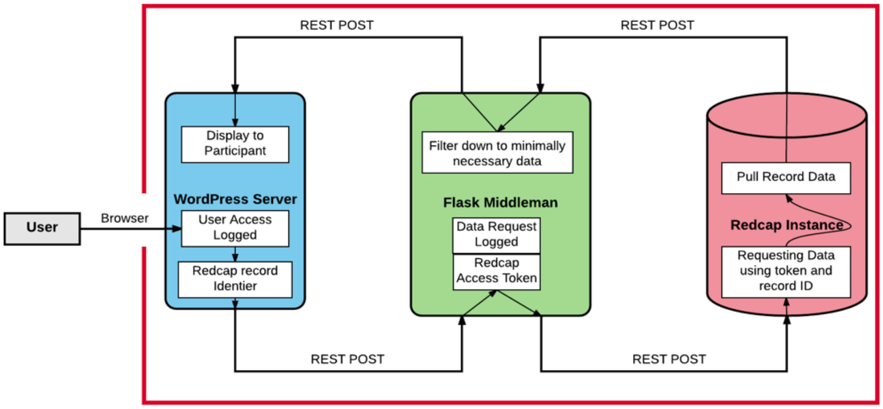


Figure 5.1: Diagram of the security protocols using the middleman server. Using the middleman as a secure token storage mechanism enables mitigating the risk of a bad actor gaining access to the REDCap project. The Flask API also allows all non-essential variables to be removed from the JSON object that is returned from the REDCap project. ITHS is the Institute of Translational Health Sciences at the University of Washington and manages IT for research with medical data at the University of Washington.

5.3 Results

The WordPress plugin we developed is currently being used by the FindMyVariant cosegregation analysis study to enroll and manage patients and their families.

5.3.1 Administrator Workflow

While the plugin has the ability to allow study participants to sign up for the research, the FindMyVariant study has very specific criteria for who can sign up for the research. Since administrator controlled signup is essential, our implementation controls access to the signup page by limiting viewing permissions to users with admin access. The sign up page allows registration of an individual using their email, first name and last name. (Figure

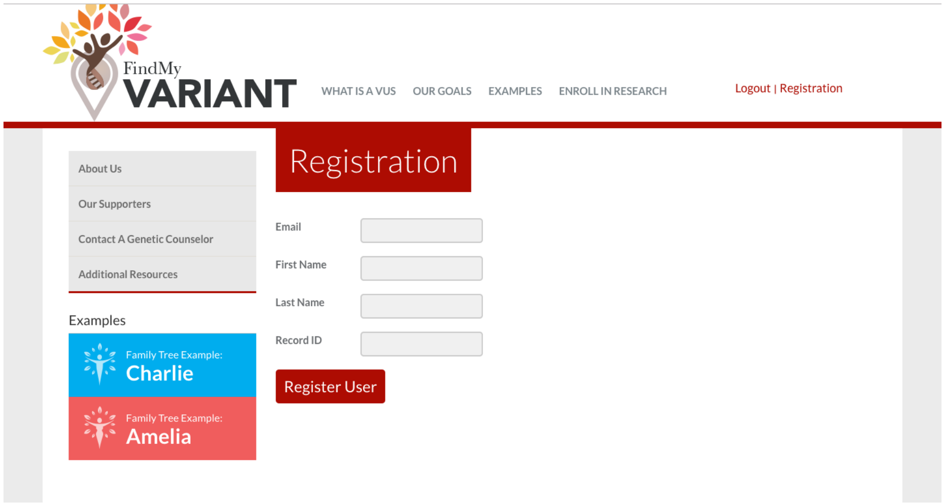
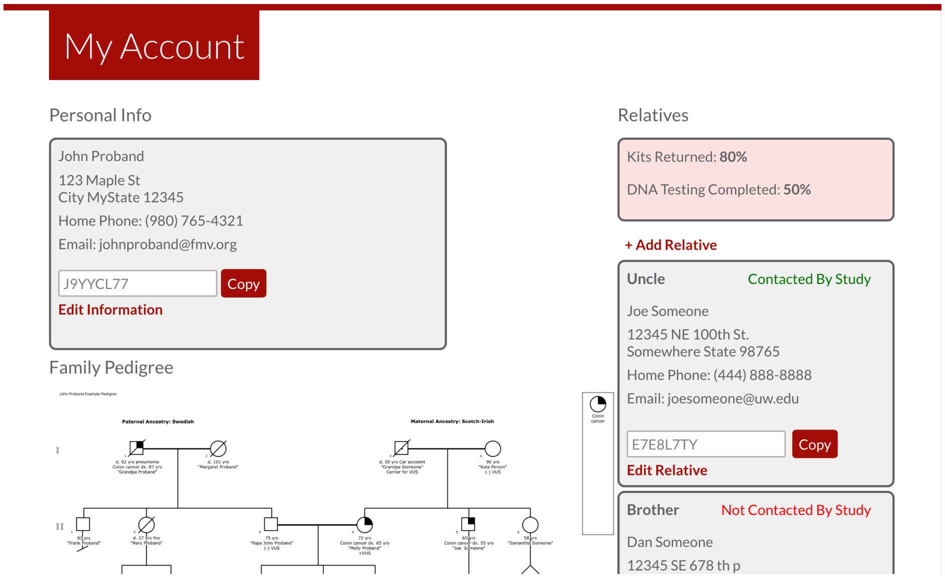


Figure 5.2: Administrator view of the FindMyVariant study website. Registration of new participants is made easier in this setup by automatically creating a record in REDCap and entering the tracking information into the local database. The new users receive an email with a link directing them to reset their password.

5.2) The RecordID field is optional depending on whether the administrator has created a REDCap record for the family. Registering participants on this page will create a new record in REDCap for this participant (or link to an existing record if RecordID is filled in). The new participant will receive an email with a link to create their password and to access their new account. The plugin stores the email and record id on the web server as a link to the newly created REDCap participant record.

5.3.2 Participant Workflow

After the new participant changes their password, they will be able to access their account which will import their data from their record in REDCap. The WordPress plugin will display the participants' personal information, relative's information (if they or the coordinator have



My Account

Personal Info

John Proband
123 Maple St
City MyState 12345
Home Phone: (980) 765-4321
Email: johnproband@fmv.org

J9YYCL77 Copy

Edit Information

Family Pedigree

Relatives

Kits Returned: **80%**

DNA Testing Completed: **50%**

+ Add Relative

Uncle Contacted By Study

Joe Someone
12345 NE 100th St.
Somewhere State 98765
Home Phone: (444) 888-8888
Email: joesomeone@uw.edu

E7E8L7TY Copy

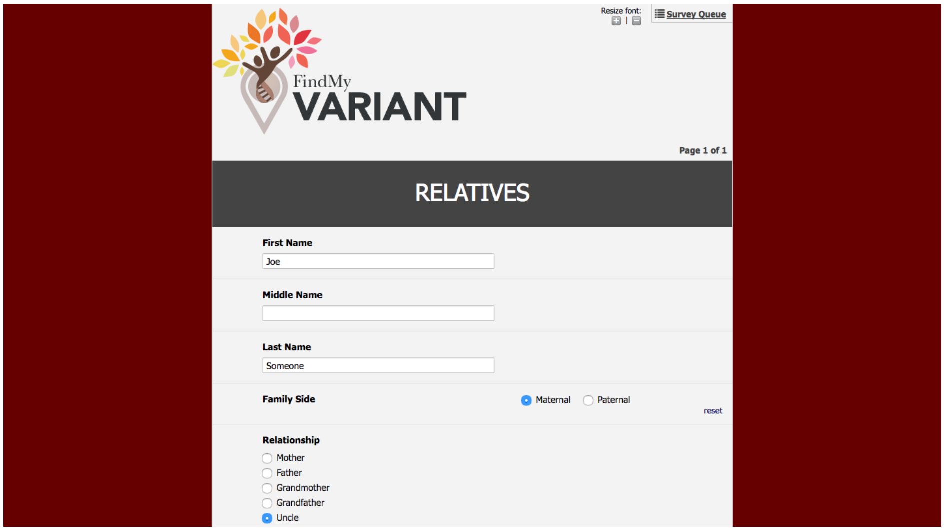
Edit Relative

Brother Not Contacted By Study

Dan Someone
12345 SE 678 th p

Figure 5.3: Patient portal interface. This interface retrieves data from the REDCap record linked to the user email and displays participant's information, relative's information, what percentage of the spit kits that were sent out have been returned, what percentage of the spit kits received have been tested, which relatives have been contacted about participating in the study and the pedigree of the family with genetic results. Links to edit the relative information take the user to a REDCap survey.

entered it), the status of the sample kits and the final pedigree once the study is complete (Figure 5.3). The user has the option to add new relatives or edit existing ones. The link to add new relatives will create a new form in the participant associated REDCap record and take the user to a REDCap survey where they can enter the relative information (Figure 5.4). The links to edit existing relatives take the user to the previously mentioned survey where they can edit the information.



Page 1 of 1

RELATIVES

First Name

Joe

Middle Name

Last Name

Someone

Family Side ● Maternal ○ Paternal

reset

Relationship

○ Mother
○ Father
○ Grandmother
○ Grandfather
● Uncle

Figure 5.4: Snapshot of the survey through which patients can edit or enter their relative's contact information. This survey form enters the information directly into the patient's associated REDCap record and also allows them to return later and edit any of the information.

5.3.3 Return of Results

We use the patient portal as an interface to return study results to the patients. Once all the information has been collected from the patients and the relatives, certain fields in the REDCap project are returned as study results to the patients. Designated fields in REDCap allow the researchers to deposit the results of the cosegregation analysis; the deposited results are shown to the patient in the patient dashboard. The patients take the results returned from the study and share them with their families.

5.3.4 Patient Use

Since the implementation of the patient portal, 32 new families have joined the FindMyVariant study, with 13 (41%) of them choosing to use the patient portal for data entry and receiving study results. As of this writing, none of the families who have used the patient portal have had their VUS re-classified, since patient-driven VUS reclassification can take anywhere between 6-12 months to complete; however, these patients have used the portal to input family data with little to no complications or complaints. Once testing of the participant samples begins, it is difficult to track how patients use the patient portal. From May 15-July 16, 2018, there were 92 views of the account page on FindMyVariant.org. Using simple tracking methods, we also know that 3 unique patients were actively involved in editing their information between July 12th and July 16th.

5.4 Discussion

We developed a WordPress plugin that enables researchers to easily build and link a WordPress page that integrates REDCap research data into a patient portal. Other platform options are limited in adaptability and scope. The REDCapToWordPress plugin allows an established, well documented, and adaptable web development platform to be integrated with a widely used research data capture platform. The skills needed to customize the patient portal are limited to a basic knowledge of HTML and CSS which is more manageable than needing an understanding of systems and database engineering. Now more researchers can harness the power of patient-driven research, and unlock the potential of their research subjects.

5.4.1 IRB Concerns and Return of Results

In the FindMyVariant study, the new classification of the VUS is returned to the patient, who can choose to share the VUS with the family. While FindMyVariant does not give medical advice on how to treat and deal with the new VUS classification, the new information is normally taken by the patient to a genetic counselor or physician who works with the patient to adjust their care as appropriate. Every variant that is reclassified has clinical actionability. This is important since previous research has found that patients and communities are positively impacted by the return of genomic test results if the findings are clinically actionable [115, 116], and while heavily debated, it has been recommended that results from genetic research studies be returned only when they are clinically actionable to the patient [117, 118].

During the IRB application process for the FindMyVariant patient portal, a concern was raised over this issue of return of results to patients. What if a relative did not want the knowledge that they had the VUS returned to other family members? FindMyVariant solved this issue by allowing relatives to choose to give their written consent for one of three options: they can have their results returned to the family member who invited them to participate regardless of the results, they can choose to view the results first before they are returned, or they can choose to not have the results returned at all. The individuals maintain control of their genetic information. The information available to the family coordinator is entirely managed by the research team who has complete control over which data is entered into the patient visible fields in REDCap. For some data elements, new REDCap record fields were created for confidential information to make the REDCap database compatible with differing relative preferences for data sharing. From the patient perspective looking at the patient portal, they are not able to tell whether someone has opted out of the study, or how many sample collection kits have been sent out or returned. We reported the percentages of the study progress for the purposes of protecting relatives choices if they choose to opt out of the study. It may be possible for a patient to deduce this information based on prior knowledge or communication with relatives, but it is not shared by the study.

5.4.2 The Importance of the Middleman

The middleman is an important feature in the plugin implementation. REDCap gives API security tokens to users of REDCap that give access to the entire REDCap project. Anyone with the REDCap API token can request access to the data within the entire REDCap project, making protection of the token paramount. The middleman creates a safe place to store the token, limiting its interaction with the WordPress site. The middleman also acts as a filter of unnecessary or restricted data. By creating a Flask API on the middleman server with data specific endpoints, we can control the flow of information much like established methods for software application interfacing with databases. By having set endpoints that call specific data fields from REDCap, bad actors won't be able to directly query the REDCap project but would have to go through the middleman API, limiting their access to the data. Since API requests with the token can be used to pull extra data, the middleman will filter out any data not explicitly requested by the functions in the plugin. If the REDCap API calls were done directly from the WordPress site, a bad actor could potentially pull the entire participant record as opposed to only the components that are meant to be available to the participant. Since only limited data is returned to the WordPress site, the breach would be limited to the explicitly defined variables in the middleman API.

5.4.3 The Potential of Patient-Driven Research

Patient interactive research does not have to be limited to family studies. There is growing use of many types of social networks in research. Patients may be used to recruit friends and acquaintances and to contribute data that is more detailed and accurate than researcher driven data collection efforts. With the increasing adoption of EHR interoperability standards like Fast Healthcare Interoperability Resources (FHIR), enabling patients to enter their data directly into EHRs could be the next step in patient-driven research. With EHRs being seen more widely as sources of research data and with the push to use the EHR as an application platform, empowering patients to validate and enter their own data could open new opportunities for medical research. Recently, studies have looked at the effect of patients being given access to their clinical notes, where they are able to read the doctors notes and enter their own notes [17, 18, 19]. So far, these studies have shown that patients responded positively to having increased access to their data. All of these innovations can lead to more opportunities to engage patients in driving medically actionable discoveries about themselves and other similar patients.

5.4.4 Using REDCapToWordPress for Patient-Driven Research

The REDCapToWordPress plugin will allow researchers of all stripes to use the potential of the patient-driven research framework. With limited programming knowledge, a researcher can quickly develop a patient portal, linking their customized REDCap project with a WordPress site. With the lowered threshold of installing a patient portal and the promise of larger, more accurate datasets from patient-driven research, we see the patient-driven framework as a way to drive precision medicine in a cheaper, more efficient manner.

5.5 Conclusion

We develop a WordPress plugin that links a patient portal embedded in WordPress to a research study's REDCap project, allowing patients to directly enter their research relevant data into the REDCap project. The plugin is currently being used in the FindMyVariant study, with 14 of the 32 families opting to use the portal as their main data entry mechanism. Patient-driven research is an effective way to increase research subject engagement, improve research data validity, and increase cohort sizes, all of which drive the success rate of the study. REDCapToWordPress creates a patient portal that is easy to customize and adapt to research projects, lowering the threshold for using the patient-driven research framework.

Chapter 6
CONCLUSION

In conclusion, I summarize the contributions from each aim by reviewing the research contributions (Section 6.1) and acknowledging the limitations and opportunities for future work (Section 6.2).

6.1 Summary of Contributions

The nearly ubiquitous adoption of EHRs by healthcare systems is driving excitement and research around data driven healthcare, including research from poputhe population health to machine learning. Concerns of security, data integrity, and access to EHRs limit the realization of this vision.

In this work, I address these issues of security for access and research with health data as well as concerns of data accuracy as it pertains to population health. The four aims of this thesis contribute to building data driven healthcare system that is secure, trusted, and accessible.

6.1.1 Aim 1 summary

In Aim 1, I collected EHR data from the University of Washington enterprise data warehouse. Using a binomial test, I detected anomalous events where the number and type of diagnosis codes significantly deviated from the expected baseline. Enrichment of ICD codes were evaluated, from both a seasonal and daily perspective, to detect possible events in the data that reflected real events from the local population. Each day that had anomalous codes were considered "events" with scores associated with each event to designate the significance or magnitude of the event. For each seasonal enrichment detected and for the top 15 days with the most enriched codes, I externally validated these trends in the literature and news sites.

The binomial test detected patterns of seasonal enrichment consistent with our expectations about seasonal behavior such as accidents from snow sports such as skiing and snowboarding during the winter, and accidents related to outdoor activities in warm weather such as bites and stings from bugs, firework accidents, bicycle accidents, and water transport accidents during the summer. Annual events such as July 4th and Jan 1st were highlighted by a significant increase in the number of burns and firework related accidents. One time rare events like the Hannukah eve windstorm and the Nisqually earthquake were marked by a unique pattern of carbon monoxide related poisoning and ICD-9 codes for earthquake related injuries.

In support of Aim 1, my contributions in this study are to evaluate EHRs for their potential to serve as generalizable population health platforms, looking specifically at how "events" in the local population are reflected in the EHR. My contributions are:

1. The comparison of statistically significant code enrichments from the University of Washington EHR data to the literature and news sources finding that these enrichments

reflect local population events.

2. Reinforcement of the growing body of literature that suggests EHRs can be used as a viable generalizable population health analysis platform, even when the data analyzed was not originally designed for such analysis.

6.1.2 Aim 2 summary

In Aim 2, we carry out 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) to their on-premises secure computing infrastructure. A researcher (referred to as the model developer), who did not have direct access to the UW Medicine EHR database, piloted the system by designing and building a 180-day mortality prediction model using the available clinical records.

The model developer built three mortality prediction models using three different sets of features: demographics, demographics and binary indicators for 5 common chronic diseases, and demographics and the top 1000 most common conditions in the clinical record. The model developer was able to build models using both engineered and data driven features and achieved an Area Under the Recall Curve of 0.921 with the top 1000 most common conditions. During the course of this pilot we discovered the essential elements necessary for a researcher to fully harness the "Model to data" framework, namely a synthetic dataset that closely resembles the real data in form and size, an environment with the synthetic data allowing models to run on the synthetic dataset for debugging and runtime assessment, and a data dictionary with the more commonly used concept codes from the EHR data to enable smarter feature engineering.

My contributions in support of Aim 2:

1. Successfully enabled a model developer, with no access to the EHR data, to build accurate mortality prediction models using both engineered and data driven features from the hidden EHR data.

2. Characterizing the essential elements needed for researchers to successfully utililize the "Model to data" framework for developing predictive models.

3. Showcasing the "Model to data" approach as an alternative to other data sharing mechanisms for prediction model development.

6.1.3 Aim 3 summary

In Aim 3, we expanded the work of Aim 2 to scale the "Model to Data" framework, enabling more than one researcher to build mortality prediction models using the University of Washington clinical records without having access to the data. We hosted a community challenge, asking partipants to predict the 180-day mortality status of patients in the UW EHR from the time of their last visit. We improved upon the synthetic data from Aim 1, making that available for both download and for model debugging in a cloud environment, and created a data dictionary, making it available to partipants to aid in their model development. Participants were able to train and test their models in the University of Washington system, using

EHR as their data. We evaluated submitted models on a held out, prospectively collected EHR dataset to assign final accuracy metrics to particapants.

We had 345 registered participants, coalescing into 25 independent teams, spread over 3 continents and 10 countries. The top performing team achieved a final area under the receiver operator curve of 0.947 (95% CI 0.942, 0.951) and an area under the precision-recall curve of 0.487 (95% CI 0.458, 0.499) on the prospectively collected patient records. Top features used in the highest performing models included age of the patient at their last visit and measurements of systolic blood pressure, diastolic blood pressure, and heart rate. Models were evaluated for their accuracy across sensitive demographic strata as well as different clinical contexts, with the results indicating that different models vary in their accuracy across these groups when compared to other models, indicating that it may not solely be a data bias issue.

My contributions in support of Aim 3:

1. Further showcasing the "Model to data" framework as a viable and scalable solution to expanding access to patient records for predictive model development, without risking patient privacy violations.

2. Evaluating 15 different mortality prediction models developed by teams from around the world on the same data set to objectively assess and compare model accuracy to establish the "state of the art" in mortality prediction models.

3. Highlighting the need for further research into machine learning bias in medicine, es-

pecially in relation to sensitive demographic strata.

6.1.4 Aim 4 summary

In Aim 4, I developed a patient portal for the FindMyVariant research project to link research subjects with their study results and enabling the patient driven research paradigm used by FindMyVariant. This tool was a Wordpress plugin that securely connects a patient portal webpage in Wordpress to a REDCap project hosting the research data.

Through the web portal, study admins could easily onboard new research participants and those participants could see their family tree information, track the progress of their study, and receive the results of their family genetics study. During the course of the project, 32 new families joined, and of those, 14 used the patient portal to coordinate their family studies.

My contributions in support of Aim 4:

1. Developing an open source Wordpress plugin that securely links patients with the research results of their family genetics study through a patient portal.

2. Bringing an informatics solution to the patient-driven research domain, enabling patients to better coordinate the research studies that impact their health.

6.2 Limitations and Future Work

6.2.1 Aim 1 limitations and future work

As with any study of EHRs, we cannot rule out biases due to site-specific coding practices or changes in practitioner knowledge of the health record system that may effect the prevalence or presence of codes that would otherwise show an event. Future work should focus on further validation studies to evaluate the representation of the UWMC data in the Seattle Region. Another future solution would be to run our method at more sites across Washington, feeding the live statistics into an aggregation mechanism for a more robust population view.

Our methods also run the risk of identifying false positive events in the EHR data that are simply spurious coding events that are not reflective of real events in the population. Although we use bonferroni correction to minimize this risk, this method may be an overly conservative approach, and future studies need to research methods for controlling the false positive rate without losing true positives.

Finally, our method only uses diagnoses codes to find events in the clinical records. This limits our ability to find all possible events that are being captured. Future efforts could use Natural Language Processing on clinical note texts to better inform detected trends and find "enriched" keywords on the detected days to add context to the detected events in a data driven automated manner.

6.2.2 Aim 2 and 3 limitations and future work

Model interpretation, such as gathering feature importance scores, is difficult under a "Model to data" paradigm. Currently model developers need to output their features as an output file from their docker containers. However, if their model is not a simple set of coeffients or parameters, this may not be possible under the MTD framework. Future work should focus on developing reliable feature collection methods that don't rely on the model developers to output their features, but can extract feature importance by either simulations or perturbations of the underlying data.

During the community challenge, we set a submission limit of 10 hours for both the training and testing stages of model runtime. While this was implemented to limit the burden to the University of Washington secure servers, this also limited the types of models participants were able to build and excluded deep learning or more sophisticated, time consuming models. However, limiting the time also forced participants to carefully consider the efficiency of their algorithms so extraneous calculations or operations were not run. For future efforts, finding either cloud based solutions or increasing the amount of on-site computational resources could improve the types of models participants would be able to build.

The remote nature of the MTD framework limits the opportunities for manual hyperparameter tuning which usually requires direct interaction with data. However, automated methods to tune the hyperparameters work with the proposed pipeline. The emergence of AutoML, as well as other algorithms including grid and random search, reinforcement learning, evolutionary algorithms, and Bayesian optimization, allows hyperparameter optimization to be automated and efficient. Future work could research methods for giving participants more insight into their models final parameters, while still protecting patient privacy.

While we found that in many cases, models differed in their accuracy across sensitive demographic strata, the methods used to evaluate these differences may be susceptible to the differences in demographic strata prevalences within the dataset. Future efforts should focus on developing more robust statistical methods to confirm these findings of potential bias across different models.

6.2.3 Aim 4 limitations and future work

While the patient portal developed for this aim was useful to FindMyVariant, in order to adapt the patient portal to another research project, the variables that interact with the REDCap API would need to be redefined and recoded, making the current version of this plugin limited in its use case scenarios. The same holds for the customizability of the patient portal display. Some coding skills are necessary to customize the layout of the patient portal, limiting this plugin's generalizability to other REDCap based projects. Future work on this project should focus on making the code between the patient portal and the REDCap API more adaptable to other REDCap projects, removing the need to manually code to allow the plugin to interact with new REDCap projects. Future efforts should focus on building a backend user interface that would allow the manager of the site to customize the layout of the patient portal via a graphical user interface.

6.2.4 Conclusion Overview

This work serves to advance the knowledge about methods and tools for interacting with health data from the prespective of both the researcher and patient in a secure and ethical manner as well as advancing the body of literature enhancing trust in the EHR as a generalizable population health platform. In Aim 1, I evaluated an EHR for its potential to serve as a generalizable population health platform. In Aim 2, I piloted a Model-to-data framework as a mechanism to deliver predictive methods to sensitive health data. In Aim 3, I scaled the Model-to-data framework and hosted a mortality prediction community challenge. In Aim 4, I built a patient portal allowing patients to edit and view their health information and enabling patient driven research. These four aims serve to bring the vision of the data driven health system closer to fruition, and the contributions of this work will aid in the development of a system that is private, accessable, and accurate in its delivery of care.

BIBLIOGRAPHY

[1] Dustin Charles, Meghan Gabriel, Talisha Searcy, and Others. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2014. *ONC data brief*, 23(4), 2015.

[2] Dawn Heisey-Grove and Vaishali Patel. Physician motivations for adoption of electronic health records. *Washington, DC: Office of the National Coordinator for Health Information Technology*, 2014.

[3] Guthrie S Birkhead, Michael Klompas, and Nirav R Shah. Uses of electronic health records for public health surveillance to advance public health. *Annu. Rev. Public Health*, 36(1):345–359, March 2015.

[4] Spencer S Jones, Robert S Rudin, Tanja Perry, and Paul G Shekelle. Health information technology: an updated systematic review with a focus on meaningful use. *Ann. Intern. Med.*, 160(1):48–54, January 2014.

[5] Beau Norgeot, Benjamin S Glicksberg, and Atul J Butte. A call for deep-learning healthcare. *Nat. Med.*, 25(1):14–15, January 2019.

[6] Jianying Hu, Adam Perer, and Fei Wang. Data driven analytics for personalized healthcare. In Charlotte A Weaver, Marion J Ball, George R Kim, and Joan M Kiel, editors, *Healthcare Information Management Systems: Cases, Strategies, and Solutions*, pages 529–554. Springer International Publishing, Cham, 2016.

[7] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.*, 25(10):1419–1428, October 2018.

[8] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.*, 19(6):1236–1246, November 2018.

[9] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PLoS One*, 14(2):e0211057, February 2019.

[10] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.*, 18(Suppl 4):122, December 2018.

[11] Michael Klompas, Michael Murphy, Julie Lankiewicz, Jason McVetta, Ross Lazarus, Emma Eggleston, Patricia Daly, Paul Oppedisano, Brianne Beagan, Chaim Kirby, and Richard Platt. Harnessing electronic health records for public health surveillance. *Online J. Public Health Inform.*, 3(3), December 2011.

[12] Michael Klompas, Emma Eggleston, Jason McVetta, Ross Lazarus, Lingling Li, and Richard Platt. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care*, 36(4):914–921, April 2013.

[13] Michael S Calderwood, Richard Platt, Xuanlin Hou, Jessica Malenfant, Gillian Haney, Benjamin Kruskal, Ross Lazarus, and Michael Klompas. Real-time surveillance for tuberculosis using electronic health record data from an ambulatory practice in eastern massachusetts. *Public Health Rep.*, 125(6):843–850, November 2010.

[14] Amanda F Elliott, Arthur Davidson, Flora Lum, Michael F Chiang, Jinan B Saaddine, Xinzhi Zhang, John E Crews, and Chiu-Fang Chou. Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions in the united states. *Am. J. Ophthalmol.*, 154(6 Suppl):S63–70, December 2012.

[15] Michael Klompas, Gillian Haney, Daniel Church, Ross Lazarus, Xuanlin Hou, and Richard Platt. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *PLoS One*, 3(7):e2626, July 2008.

[16] Sharon E Perlman, R Charon Gwynn, Carolyn M Greene, Amy Freeman, Claudia Chernov, and Lorna E Thorpe. NYC HANES 2013-14 and reflections on future population health surveillance. *J. Urban Health*, July 2018.

[17] Susan S Woods, Erin Schwartz, Anais Tuepker, Nancy A Press, Kim M Nazi, Carolyn L Turvey, and W Paul Nichol. Patient experiences with full electronic access to health records and clinical notes through the my HealtheVet personal health record pilot: qualitative study. *J. Med. Internet Res.*, 15(3):e65, March 2013.

[18] Kim M Nazi, Carolyn L Turvey, Dawn M Klein, Timothy P Hogan, and Susan S Woods. VA OpenNotes: exploring the experiences of early patient adopters with access to clinical notes. *J. Am. Med. Inform. Assoc.*, 22(2):380–389, March 2015.

[19] Catherine T Fant and Deborah S Adelman. Sharing clinical notes with patients. *Nurse Pract.*, 42(10):1, October 2017.

[20] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. Big healthcare data: preserving security and privacy. *Journal of Big Data*, 5(1):1, January 2018.

[21] Carol Pierannunzi, Shaohua Sean Hu, and Lina Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (BRFSS), 2004-2011. *BMC Med. Res. Methodol.*, 13:49, March 2013.

[22] Michael Klompas, Jason McVetta, Ross Lazarus, Emma Eggleston, Gillian Haney, Benjamin A Kruskal, W Katherine Yih, Patricia Daly, Paul Oppedisano, Brianne Beagan, Michael Lee, Chaim Kirby, Dawn Heisey-Grove, Alfred DeMaria, Jr, and Richard Platt. Integrating clinical practice and public health surveillance using electronic medical record systems. *Am. J. Public Health*, 102 Suppl 3:S325–32, June 2012.

[23] Mary Ann Cooney, Michael F Iademarco, Monica Huang, William R MacKenzie, and Arthur J Davidson. The public health community platform, electronic case reporting, and the digital bridge. *J. Public Health Manag. Pract.*, 24(2):185–189, 2018.

[24] Rachel D Melamed, Hossein Khiabanian, and Raul Rabadan. Data-driven discovery of seasonally linked diseases from an electronic health records system. *BMC Bioinformatics*, 15 Suppl 6:S3, May 2014.

[25] Barbara Michiels, Van Kinh Nguyen, Samuel Coenen, Philippe Ryckebosch, Nathalie Bossuyt, and Niel Hens. Influenza epidemic surveillance and prediction based on electronic health record data from an out-of-hours general practitioner cooperative: model development and validation on 2003–2015 data. *BMC Infect. Dis.*, 17(1):84, 2017.

[26] Aurora O Amoah, Sonia Y Angell, Hannah Byrnes-Enoch, Sam Amirfar, Phoenix Maa, and Jason J Wang. Bridging the gap between clinical practice and public health: Using EHR data to assess trends in the seasonality of blood-pressure control. *Prev Med Rep*, 6:369–375, June 2017.

[27] M R Boland. A Systems-Level approach to understand the seasonal factors of early development with clinical and pharmacological applications. 2017.

[28] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, January 2009.

[29] L A Fingerhut and M Warner. The ICD-10 injury mortality diagnosis matrix. *Inj. Prev.*, 12(1):24–29, February 2006.

[30] Recommended framework for presenting injury mortality data. `https://www.cdc.gov/mmwr/preview/mmwrhtml/00049162.htm`, August 1997. Accessed: 2018-10-18.

[31] 2015-ICD-10-CM-and-GEMs. September 2014.

[32] CNN.com - major U.S. quakes - february 28, 2001. *CNN*, February 2001.

[33] Largest recorded earthquake in WA was 17 years ago. `https://www.king5.com/article/news/local/largest-recorded-earthquake-in-wa-was-17-years-ago/281-67102021`, January 2018. Accessed: 2018-11-21.

[34] Local news — carbon-monoxide poisoning kills burien man — seattle times newspaper. `http://community.seattletimes.nwsource.com/archive/?date=20070124&slug=dige24m`. Accessed: 2018-11-13.

[35] Hanukkah eve wind storm ravages western washington beginning on december 14, 2006. `http://www.historylink.org/File/8042`. Accessed: 2018-11-13.

[36] Reena K Gulati, Tao Kwan-Gett, Neil B Hampson, Atar Baer, Dennis Shusterman, Jamie R Shandro, and Jeffrey S Duchin. Carbon monoxide epidemic among immigrant populations: King county, washington, 2006. *Am. J. Public Health*, 99(9):1687–1692, September 2009.

[37] CNN.com - monorail train catches fire in seattle - may 31, 2004. *CNN*, June 2004.

[38] F/V yardarm knot Fire/Chlorine release — IncidentNews — NOAA. `https://incidentnews.noaa.gov/incident/7054`. Accessed: 2018-12-18.

[39] Rick Ruddell, Matthew O Thomas, and Lori Beth Way. Breaking the chain: Confronting issueless college town disturbances and riots. *J. Crim. Justice*, 33(6):549–560, November 2005.

[40] Kent E Glindemann, Douglas M Wiegand, and E Scott Geller. Celebratory drinking and intoxication: A contextual influence on alcohol consumption. *Environ. Behav.*, 39(3):352–366, May 2007.

[41] John A Staples and Donald A Redelmeier. The april 20 cannabis celebration and fatal traffic crashes in the united states. *JAMA Intern. Med.*, 178(4):569–572, April 2018.

[42] Cameron Bell, Jessica Slim, Hanna K Flaten, Gordon Lindberg, Wiktor Arek, and Andrew A Monte. Butane hash oil burns associated with marijuana liberalization in colorado. *J. Med. Toxicol.*, 11(4):422–425, December 2015.

[43] Gang Luo, Peter Tarczy-Hornoch, Adam B Wilcox, and E Sally Lee. Identifying patients who are likely to receive most of their care from a specific health care system: Demonstration via secondary analysis. *JMIR Med Inform*, 6(4):e12241, November 2018.

[44] Claudia Allen, Terrisca R Des Jardins, Arvela Heider, Kristin A Lyman, Lee McWilliams, Alison L Rein, Abigail A Schachter, Ranjit Singh, Barbara Sorondo, Joan Topper, and Scott A Turske. Data governance and data sharing agreements for Community-Wide health information exchange: Lessons from the beacon communities, 2014.

[45] Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.*, 25(8):969–975, August 2018.

[46] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan van der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.*, 216:574–578, 2015.

[47] Jeffrey G Klann, Matthew A H Joss, Kevin Embree, and Shawn N Murphy. Data model harmonization for the all of us research program: Transforming i2b2 data into the OMOP common data model. *PLoS One*, 14(2):e0212463, February 2019.

[48] Simson L Garfinkel. De-identification of personal information. Technical report, National Institute of Standards and Technology, October 2015.

[49] Bradley Malin, Latanya Sweeney, and Elaine Newton. Trail re-identification: learning who you are from where you have been. *Proc. LIDAP-WP12*, 2003.

[50] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, David J Wales, and Ritankar Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform*, 4(3):e28, September 2016.

[51] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. March 2017.

[52] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, May 2016.

[53] Randi Foraker, Douglas L Mann, and Philip R O Payne. Are synthetic data derivatives the future of translational medicine? *JACC Basic Transl Sci*, 3(5):716–718, October 2018.

[54] Richard E Murray, Patrick B Ryan, and Stephanie J Reisinger. Design and validation of a data simulation model for longitudinal healthcare data. *AMIA Annu. Symp. Proc.*, 2011:1176–1185, October 2011.

[55] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.*, August 2017.

[56] Justin Guinney and Julio Saez-Rodriguez. Alternative models for sharing confidential biomedical data. *Nat. Biotechnol.*, 36(5):391–392, May 2018.

[57] Enterprise container platform — docker. `https://www.docker.com/`. Accessed: 2019-8-9.

[58] Singularity. `https://sylabs.io/`. Accessed: 2019-11-18.

[59] Kyle Ellrott, Alex Buchanan, Allison Creason, Michael Mason, Thomas Schaffter, Bruce Hoff, James Eddy, John M Chilton, Thomas Yu, Joshua M Stuart, Julio Saez-Rodriguez, Gustavo Stolovitzky, Paul C Boutros, and Justin Guinney. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol.*, 20(1):195, September 2019.

[60] Wendong Ge, Jin-Won Huh, Yu Rang Park, Jae-Ho Lee, Young-Hak Kim, and Alexander Turchin. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. *AMIA Annu. Symp. Proc.*, 2018:460–469, December 2018.

[61] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.*, 24(1):198–208, January 2017.

[62] Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.*, 17(8):470–486, July 2016.

[63] Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang, and Adam A Margolin. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. *Nat. Genet.*, 45(10):1121–1126, October 2013.

[64] Stephen F Weng, Luis Vaz, Nadeem Qureshi, and Joe Kai. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One*, 14(3):e0214365, March 2019.

[65] Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.*, 21(4):578–582, July 2014.

[66] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.*, 17(2):124–130, March 2010.

[67] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M Yunes, Ameet S Talwalkar, Susanna Repo, Michael L Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W A Buchan, Kevin Bryson, David T Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio

Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N Wass, Michael J E Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A I Kourmpetis, Aalt D J van Dijk, Cajo J F ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C Babbitt, Steven E Brenner, Christine Orengo, Burkhard Rost, Sean D Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nat. Methods*, 10(3):221–227, March 2013.

[68] Yuxiang Jiang, Tal Ronnen Oron, Wyatt T Clark, Asma R Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S Funk, Indika Kahanda, Karin M Verspoor, Asa Ben-Hur, Da Chen Emily Koo, Duncan Penfold-Brown, Dennis Shasha, Noah Youngs, Richard Bonneau, Alexandra Lin, Sayed M E Sahraeian, Pier Luigi Martelli, Giuseppe Profiti, Rita Casadio, Renzhi Cao, Zhaolong Zhong, Jianlin Cheng, Adrian Altenhoff, Nives Skunca, Christophe Dessimoz, Tunca Dogan, Kai Hakala, Suwisa Kaewphan, Farrokh Mehryary, Tapio Salakoski, Filip Ginter, Hai Fang, Ben Smithers, Matt Oates, Julian Gough, Petri Törönen, Patrik Koskinen, Liisa Holm, Ching-Tai Chen, Wen-Lian Hsu, Kevin Bryson, Domenico Cozzetto, Federico Minneci, David T Jones, Samuel Chapman, Dukka Bkc, Ishita K Khan, Daisuke Kihara, Dan Ofer, Nadav Rappoport, Amos Stern, Elena Cibrian-Uhalte, Paul Denny, Rebecca E Foulger, Reija Hieta, Duncan Legge, Ruth C Lovering, Michele Magrane, Anna N Melidoni, Prudence Mutowo-Meullenet, Klemens Pichler, Aleksandra Shypitsyna, Biao Li, Pooya Zakeri, Sarah ElShal, Léon-Charles Tranchevent, Sayoni Das, Natalie L Dawson, David Lee, Jonathan G Lees, Ian Sillitoe, Prajwal Bhat, Tamás Nepusz, Alfonso E Romero, Rajkumar Sasidharan, Haixuan Yang, Alberto Paccanaro, Jesse Gillis, Adriana E Sedeño-Cortés, Paul Pavlidis, Shou Feng, Juan M Cejuela, Tatyana Goldberg, Tobias Hamp, Lothar Richter, Asaf Salamov, Toni Gabaldon, Marina Marcet-Houben, Fran Supek, Qingtian Gong, Wei Ning, Yuanpeng Zhou, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Stefano Toppo, Carlo Ferrari, Manuel Giollo, Damiano Piovesan, Silvio C E Tosatto, Angela Del Pozo, José M Fernández, Paolo Maietta, Alfonso Valencia, Michael L Tress, Alfredo Benso, Stefano Di Carlo, Gianfranco Politano, Alessandro Savino, Hafeez Ur Rehman, Matteo Re, Marco Mesiti, Giorgio Valentini, Joachim W Bargsten, Aalt D J van Dijk, Branislava Gemovic, Sanja Glisic, Vladmir Perovic, Veljko Veljkovic, Nevena Veljkovic, Danillo C Almeida-E-Silva, Ricardo Z N Vencio, Malvika Sharan, Jörg Vogel, Lakesh Kansakar, Shanshan Zhang, Slobodan Vucetic, Zheng Wang, Michael J E Sternberg, Mark N Wass, Rachael P Huntley, Maria J Martin, Claire O'Donovan, Peter N Robinson, Yves Moreau, Anna Tramontano, Patricia C Babbitt, Steven E Brenner, Michal Linial, Christine A Orengo, Burkhard Rost, Casey S Greene, Sean D Mooney, Iddo Friedberg, and Predrag Radivojac. An expanded evaluation of protein function prediction methods shows an im-

provement in accuracy. *Genome Biol.*, 17(1):184, September 2016.

[69] Binghuang Cai, Biao Li, Nikki Kiga, Janita Thusberg, Timothy Bergquist, Yun-Ching Chen, Noushin Niknafs, Hannah Carter, Collin Tokheim, Violeta Beleva-Guthrie, Christopher Douville, Rohit Bhattacharya, Hui Ting Grace Yeo, Jean Fan, Sohini Sengupta, Dewey Kim, Melissa Cline, Tychele Turner, Mark Diekhans, Jan Zaucha, Lipika R Pal, Chen Cao, Chen-Hsin Yu, Yizhou Yin, Marco Carraro, Manuel Giollo, Carlo Ferrari, Emanuela Leonardi, Silvio C E Tosatto, Jason Bobe, Madeleine Ball, Roger A Hoskins, Susanna Repo, George Church, Steven E Brenner, John Moult, Julian Gough, Mario Stanke, Rachel Karchin, and Sean D Mooney. Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. *Hum. Mutat.*, 38(9):1266–1276, September 2017.

[70] John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15(3):285–289, June 2005.

[71] Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A survey of the State-of-the-Art. August 2019.

[72] Stefanie Jauk, Diether Kramer, Birgit Großauer, Susanne Rienmüller, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study. *J. Am. Med. Inform. Assoc.*, 27(9):1383–1392, July 2020.

[73] Raquel Norel, John Jeremy Rice, and Gustavo Stolovitzky. The self-assessment trap: can we all be better than average? *Mol. Syst. Biol.*, 7:537, October 2011.

[74] Noah Hammarlund. Racial treatment disparities after machine learning surgical Risk-Adjustment. 2019.

[75] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019.

[76] Timothy Bergquist, Yao Yan, Thomas Schaffter, Thomas Yu, Vikas Pejaver, Noah Hammarlund, Justin Prosser, Justin Guinney, and Sean Mooney. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. *J. Am. Med. Inform. Assoc.*, July 2020.

[77] Marion Fahey, Anthony Rudd, Yannick Béjot, Charles Wolfe, and Abdel Douiri. Development and validation of clinical prediction models for mortality, functional outcome and cognitive impairment after stroke: a study protocol. *BMJ Open*, 7(8):e014607, August 2017.

[78] B Smolin, Y Levy, E Sabbach-Cohen, L Levi, and T Mashiach. Predicting mortality of elderly patients acutely admitted to the department of internal medicine. *Int. J. Clin. Pract.*, 69(4):501–508, April 2015.

[79] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, May 2018.

[80] Roxana Daneshjou, Yanran Wang, Yana Bromberg, Samuele Bovo, Pier L Martelli, Giulia Babbi, Pietro Di Lena, Rita Casadio, Matthew Edwards, David Gifford, David T Jones, Laksshman Sundaram, Rajendra Rana Bhat, Xiaolin Li, Lipika R Pal, Kunal Kundu, Yizhou Yin, John Moult, Yuxiang Jiang, Vikas Pejaver, Kymberleigh A Pagel, Biao Li, Sean D Mooney, Predrag Radivojac, Sohela Shah, Marco Carraro, Alessandra Gasparini, Emanuela Leonardi, Manuel Giollo, Carlo Ferrari, Silvio C E Tosatto, Eran Bachar, Johnathan R Azaria, Yanay Ofran, Ron Unger, Abhishek Niroula, Mauno Vihinen, Billy Chang, Maggie H Wang, Andre Franke, Britt-Sabina Petersen, Mehdi Pirooznia, Peter Zandi, Richard McCombie, James B Potash, Russ B Altman, Teri E Klein, Roger A Hoskins, Susanna Repo, Steven E Brenner, and Alexander A Morgan. Working toward precision medicine: Predicting phenotypes from exomes in the critical assessment of genome interpretation (CAGI) challenges. *Hum. Mutat.*, 38(9):1182–1192, September 2017.

[81] Christophe G Lambert, Amritansh, and Praveen Kumar. Transforming the 2.33m-patient medicare synthetic public use files to the OMOP CDMv5: ETL-CMS software and processed data available and feature-complete. September 2016.

[82] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

[83] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery.

[84] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6638–6648. Curran Associates, Inc., 2018.

[85] Greg Ridgeway. Generalized boosted models: A guide to the gbm package. *Update*, 1(1):2007, 2007.

[86] Jonathan H Chen, Muthuraman Alagappan, Mary K Goldstein, Steven M Asch, and Russ B Altman. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Inform.*, 102:71–79, June 2017.

[87] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc*, 2020:191–200, May 2020.

[88] Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, and Michael E Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Inform. Assoc.*, 24(6):1052–1061, November 2017.

[89] Sharon E Davis, Thomas A Lasko, Guanhua Chen, and Michael E Matheny. Calibration drift among regression and machine learning models for hospital mortality. *AMIA Annu. Symp. Proc.*, 2017:625–634, 2017.

[90] Andrea Cornwall and Rachel Jewkes. What is participatory research? *Soc. Sci. Med.*, 41(12):1667–1676, 1995.

[91] Mark L Fuerst. Patient-Driven model may accelerate breast cancer research. *Oncology Times*, 38(13):10, July 2016.

[92] Liam Ennis and Til Wykes. Impact of patient involvement in mental health research: longitudinal study. *Br. J. Psychiatry*, 203(5):381–386, November 2013.

[93] G H Guyatt, J L Keller, R Jaeschke, D Rosenbloom, J D Adachi, and M T Newhouse. The n-of-1 randomized controlled trial: clinical usefulness. our three-year experience. *Ann. Intern. Med.*, 112(4):293–299, February 1990.

[94] Paul A Scuffham, Jane Nikles, Geoffrey K Mitchell, Michael J Yelland, Norma Vine, Christopher J Poulos, Peter I Pillans, Guy Bashford, Chris del Mar, Philip J Schluter, and Paul Glasziou. Using n-of-1 trials to improve patient management and save costs. *J. Gen. Intern. Med.*, 25(9):906–913, September 2010.

[95] Naihua Duan, Richard L Kravitz, and Christopher H Schmid. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J. Clin. Epidemiol.*, 66(8 Suppl):S21–8, August 2013.

[96] Jonathan A Shaffer, Louis Falzon, Ken Cheung, and Karina W Davidson. N-of-1 randomized trials for psychological and health behavior outcomes: a systematic review protocol. *Syst. Rev.*, 4:87, June 2015.

[97] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per. Med.*, 8(2):161–173, March 2011.

[98] Deirdre Pinto, Dominique Martin, and Richard Chenhall. The involvement of patient organisations in rare disease research: a mixed methods study in australia. *Orphanet J. Rare Dis.*, 11:2, January 2016.

[99] Jenny Leese, Sheila Kerr, Annette McKinnon, Erin Carruthers, Catherine Backman, Linda Li, and Anne Townsend. Evolving Patient-Researcher collaboration: An illustrative case study of a Patient-Led knowledge translation event. *J. Particip. Med.*, 9(1):e13, 2017.

[100] Lucia Monaco and Lucia Faccio. Patient-driven search for rare disease therapies: the fondazione telethon success story and the strategy leading to strimvelis. *EMBO Mol. Med.*, 9(3):289–292, March 2017.

[101] M K Javaid, L Forestier-Zhang, L Watts, A Turner, C Ponte, H Teare, D Gray, N Gray, R Popert, J Hogg, J Barrett, R Pinedo-Villanueva, C Cooper, R Eastell, N Bishop, R Luqmani, P Wordsworth, and J Kaye. The RUDY study platform - a novel approach to patient driven research in rare musculoskeletal diseases. *Orphanet J. Rare Dis.*, 11(1):150, November 2016.

[102] Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G Conde. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.*, 42(2):377–381, April 2009.

[103] Benjamin. redcapAPI.

[104] Scott Burns and Aaron Browne. Version 1.0, May 2014.

[105] Kathryn R Napier, Megan Tones, Chloe Simons, Helen Heussler, Adam A Hunter, Meagan Cross, and Matthew I Bellgard. A web-based, patient driven registry for

angelman syndrome: the global angelman syndrome registry. *Orphanet J. Rare Dis.*, 12(1):134, August 2017.

[106] Matthew I Bellgard, Kathryn R Napier, Alan H Bittles, Jeffrey Szer, Sue Fletcher, Nikolajs Zeps, Adam A Hunter, and Jack Goldblatt. Design of a framework for the deployment of collaborative independent rare disease-centric registries: Gaucher disease registry model. *Blood Cells Mol. Dis.*, 68:232–238, February 2018.

[107] Matthew I Bellgard, Lee Render, Maciej Radochonski, and Adam Hunter. Second generation registry framework. *Source Code Biol. Med.*, 9(1):14, June 2014.

[108] J O Culver, C D Brinkerhoff, J Clague, K Yang, K E Singh, S R Sand, and J N Weitzel. Variants of uncertain significance in BRCA testing: evaluation of surgical decisions, risk perception, and cancer distress. *Clin. Genet.*, 84(5):464–472, November 2013.

[109] Suzanne C O'Neill, Christine Rini, Rachel E Goldsmith, Heiddis Valdimarsdottir, Lawrence H Cohen, and Marc D Schwartz. Distress among women receiving uninformative BRCA1/2 results: 12-month outcomes. *Psychooncology*, 18(10):1088–1096, October 2009.

[110] Sandra van Dijk, Christi J van Asperen, Catharina E Jacobi, Geraldine R Vink, Aad Tibben, Martijn H Breuning, and Wilma Otten. Variants of uncertain clinical significance as a result of BRCA1/2 testing: impact of an ambiguous breast cancer risk message. *Genet. Test.*, 8(3):235–239, 2004.

[111] Mitzi L Murray, Felecia Cerrato, Robin L Bennett, and Gail P Jarvik. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. *Genet. Med.*, 13(12):998–1005, December 2011.

[112] Lauren T Garrett, Nathan Hickman, Angela Jacobson, Robin L Bennett, Laura M Amendola, Elisabeth A Rosenthal, and Brian H Shirts. Family studies for classification of variants of uncertain classification: Current laboratory clinical practice and a new Web-Based educational tool. *J. Genet. Couns.*, 25(6):1146–1156, December 2016.

[113] WordPress.com: Create a free website or blog. `https://wordpress.com/`. Accessed: 2018-2-2.

[114] Kristofer @specik, Nicolas Joly @dotmastaz, and Mighty Horst @mightyhorst. The themosis framework. `https://framework.themosis.com`. Accessed: 2017-9-25.

[115] Joon-Ho Yu, Julia Crouch, Seema M Jamal, Holly K Tabor, and Michael J Bamshad. Attitudes of african americans toward return of results from exome and whole genome sequencing. *Am. J. Med. Genet. A*, 161A(5):1064–1072, May 2013.

[116] Susan Brown Trinidad, Evette J Ludman, Scarlett Hopkins, Rosalina D James, Theresa J Hoeft, Annie Kinegak, Henry Lupie, Ralph Kinegak, Bert B Boyer, and Wylie Burke. Community dissemination and genetic research: moving beyond results reporting. *Am. J. Med. Genet. A*, 167(7):1542–1550, July 2015.

[117] Gail P Jarvik, Laura M Amendola, Jonathan S Berg, Kyle Brothers, Ellen W Clayton, Wendy Chung, Barbara J Evans, James P Evans, Stephanie M Fullerton, Carlos J Gallego, Nanibaa' A Garrison, Stacy W Gray, Ingrid A Holm, Iftikhar J Kullo, Lisa Soleymani Lehmann, Cathy McCarty, Cynthia A Prows, Heidi L Rehm, Richard R Sharp, Joseph Salama, Saskia Sanderson, Sara L Van Driest, Marc S Williams, Susan M Wolf, Wendy A Wolf, eMERGE Act-ROR Committee and CERC Committee, CSER Act-ROR Working Group, and Wylie Burke. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. *Am. J. Hum. Genet.*, 94(6):818–826, June 2014.

[118] Wylie Burke, Barbara J Evans, and Gail P Jarvik. Return of results: ethical and legal distinctions between research and clinical care. *Am. J. Med. Genet. C Semin. Med. Genet.*, 166C(1):105–111, March 2014.

Appendices

Appendix A

APPENDIX

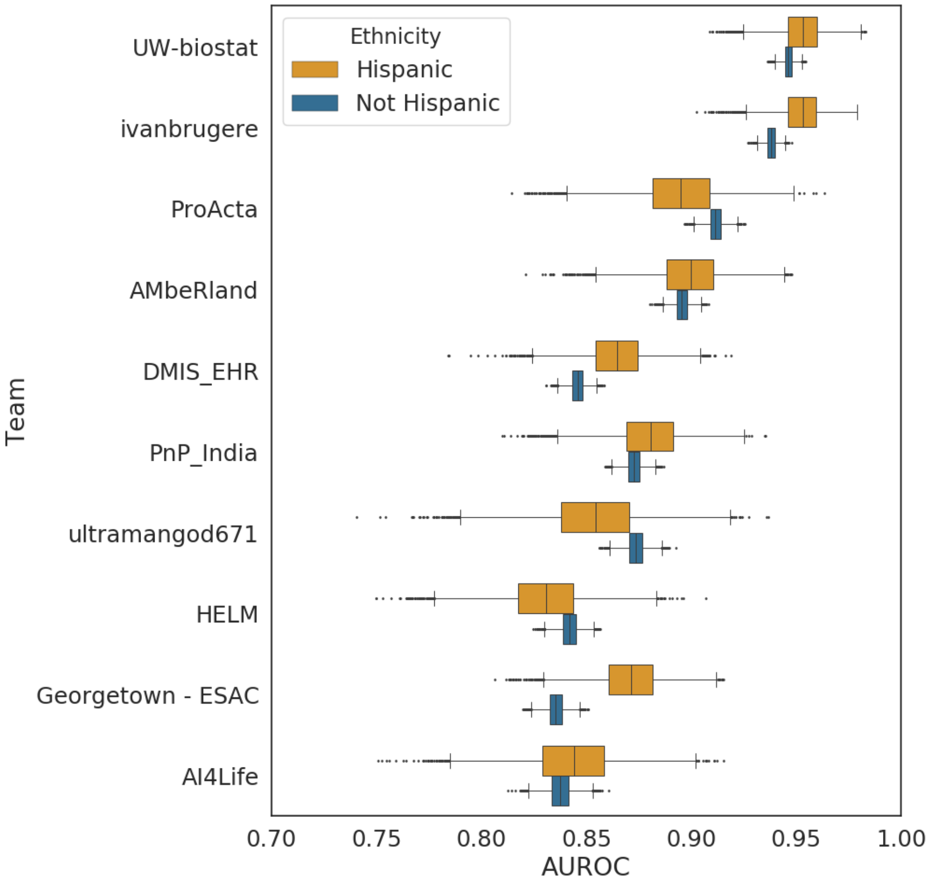


Figure A.1: Bootstrapped distributions (n=10,000) of the top 10 model AUROCs across the ethnicity demographic. Model predictions were randomly sampled with replacement and scored against the benchmark gold standard.

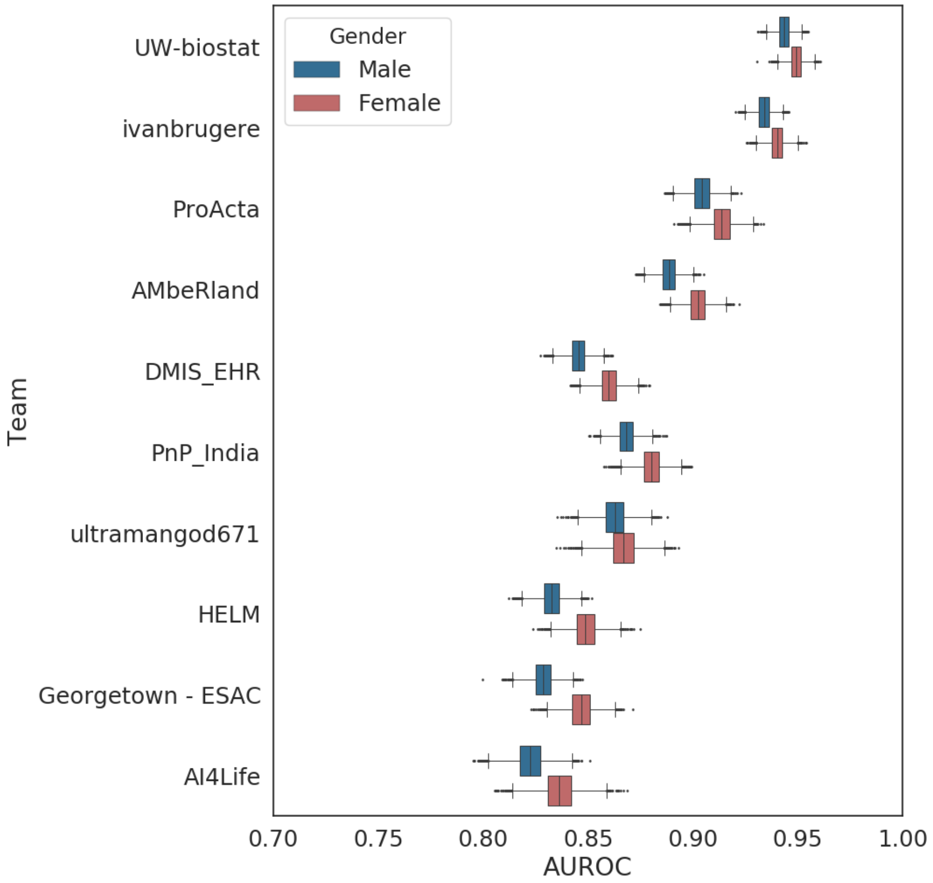


Figure A.2: Bootstrapped distributions (n=10,000) of the top 10 model AUROCs across the gender demographic. Model predictions were randomly sampled with replacement and scored against the benchmark gold standard.

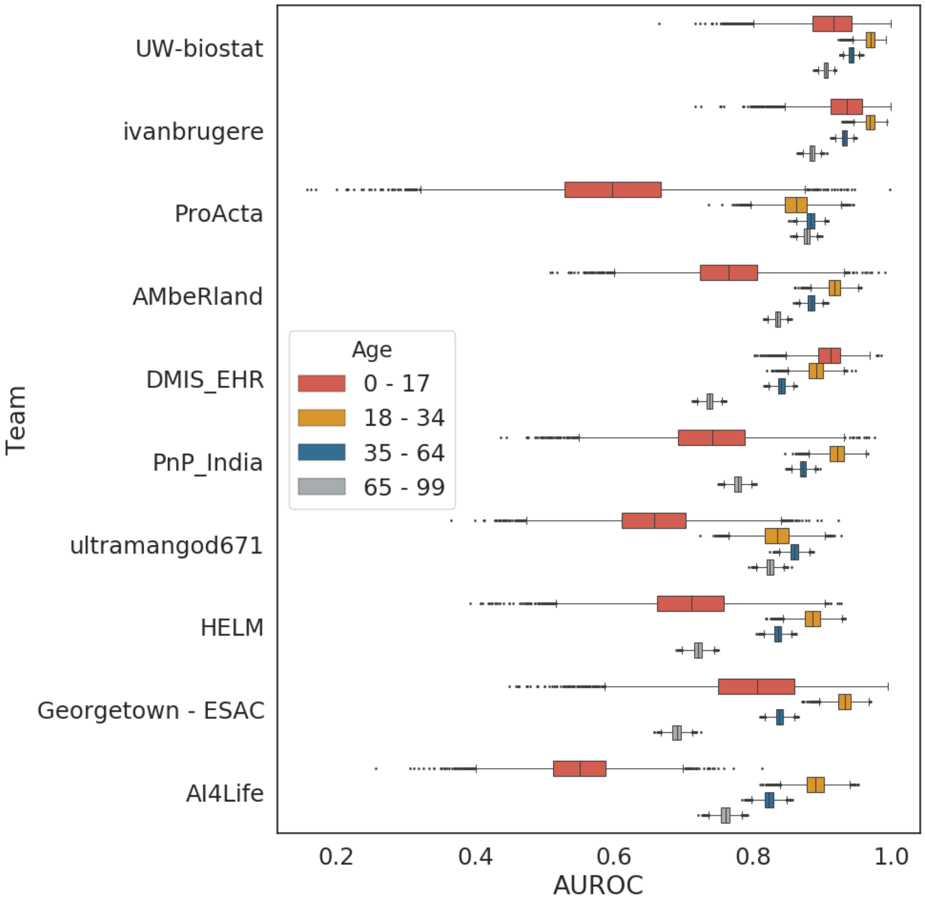


Figure A.3: Bootstrapped distributions (n=10,000) of model performance across age groups. The small number of patients age 0-17 results in high variance in model performance. For the most part, model performance is inversely correlated with age where models tend to me more accurate on 18-34 year olds and less accurate on 65-99 year olds.

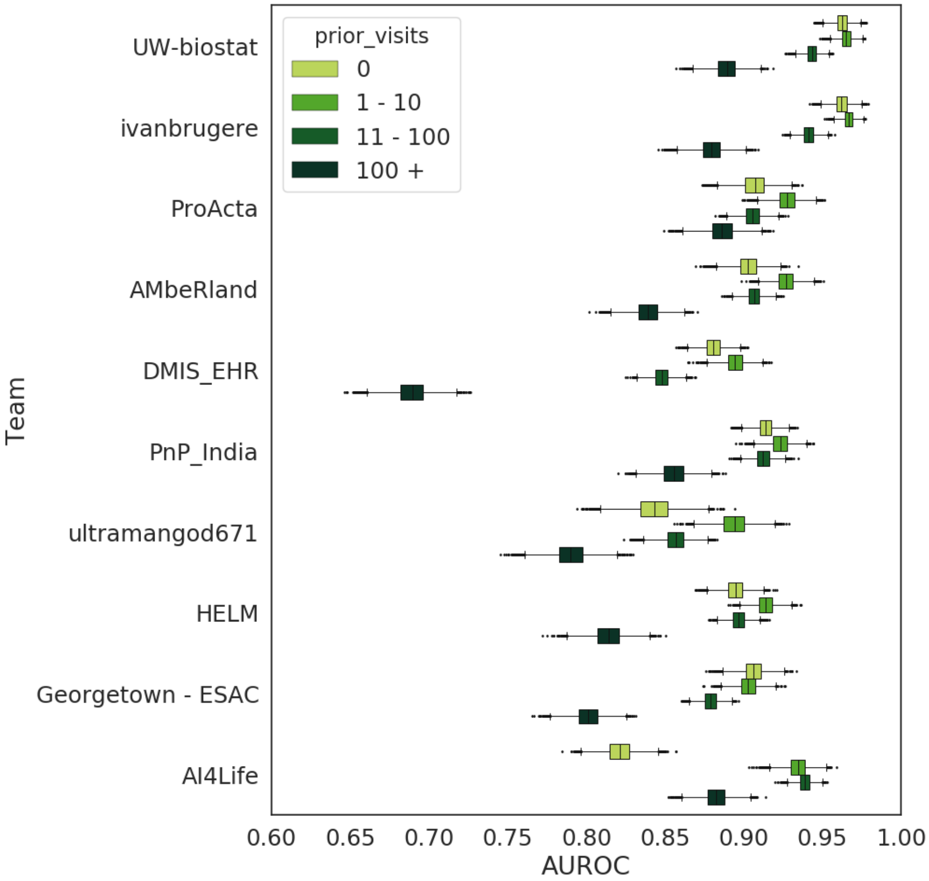


Figure A.4: Bootstrapped distributions (n=10,000) of model performance across the number of records available in each patient's clinical history. Most models perform the best on 1-10 records, with a decrease in accuracy on patient records with no history or 11-100 records, with a large decrease in performance on records of 100 or more.