NORTHWESTERN UNIVERSITY

Advancing Computational Methods to Derive Insights from Real-world Health Data

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Biomedical Informatics

By

Yu Deng

EVANSTON, ILLINOIS

June 2022

Abstract

In 2009, the Health Information Technology for Economic and Clinical Health Act (HITECH) promoted national use of electronic health records (EHR) in the US by giving incentives to providers who adopt 'meaningful use' of EHRs. As of 2017, nearly 86% of office-based physicians had adopted EHRs. EHRs have rich information including structured data like diagnosis, medication, patient encounters, laboratory tests, semi-structured data such as problem lists, and unstructured data such as patient notes. The large amount of information in EHRs presents abundant opportunities for clinical research but also challenges. EHRs can facilitate clinical research either alone or combined with other data sources including evaluating drug comparative effectiveness, facilitating patient recruitment for clinical trials, assessing gene-disease associations and many others. In this work, we explored insights that researchers can derive from using EHR data, the challenges it presents along the way, particularly, in comparison to traditional registry data and cohort data, and how to overcome these challenges from a methodology point of view. More specifically, in the second chapter, we focused on the application of EHRs on drug comparative effectiveness. In this chapter, we used structured EHR data to assess second-line type 2 diabetes medication comparative effectiveness on renal disorder to provide real-world evidence for clinical trial findings. In the third chapter, we presented data completeness/accuracy related challenges in EHR data specifically in the context of lupus subtype identification; In the fourth chapter, we proposed a natural language processing-based approach to address some of the challenges mentioned by improving the accuracy of disease phenotypes. In the last chapter, we developed a method to improve cardiovascular disease prediction in traditional cohort studies that could potentially be applied to EHR data in the future. Together, this thesis highlights the insights we can derive from EHR data, the challenges EHR

data presents when applied in clinical research and offer ways to overcome some of the challenges from a methodology point of view.

Acknowledgment

This work cannot be completed without the mentorship and support from many people during my PhD journey. First and foremost, I would like to thank my thesis advisor Dr. Abel Kho. Abel once said that the key to patient care is care and this extends to all other things in life. Abel deeply cares about his students and their personal growth. From him, I learned not only how to do good research but also how to be a good person and a good leader. Through my PhD years, there were many times when I felt I was not good enough or I was not able to finish my projects. He has always been encouraging and genuinely believed in me. I want to thank my thesis committee member, Dr. Lihui Zhao. From him, I learned the importance of being detail-orientated and how to generate quality work. The project I worked on with him involved editing over 10 times to ensure good quality of work. I want to thank my committee members Dr. Ramana Davuluri and Dr. Yuan Luo, both of whom provided value advice on my thesis work. I want to thank my close friend, Marcus Gruwell. Being an international student and studying abroad was not easy for me. Marcus and his family kindly treated me as their own and have been a huge support for both my academic and personal life. I want to thank my good friend, Susan Park. I can remember the days when we studied late in the library working on deadlines. Susan, thank you for transforming these hard times from lemons to lemonade. I want to thank my good friends Angel Bohannon, Eunie Cho, Ashley Haluck-Kangas, and my former roommates Yizhen Zhong and Shimeng Liu. Each of them have been a huge inspiration for me and I am so grateful to have grown together with them in this journey. Last but not the least, I want to thank my parents and my sister for their support and unconditional love. They made the person I am today, and I am forever thankful.

Dedication

I dedicate this thesis to my parents and sister.

Table of Contents

Abstract *3*

Acknowledgment *5*

Dedication *6*

List of Tables, Illustrations, Figures, and Graphs *9*

Chapter One: Introduction *11*

1.1 Electronic health records overview **11**

1.2 Drug comparative effectiveness using EHRs and its recent progress **13**

1.3 Predictive modeling using EHR data **14**

1.4 Existing Challenges using EHR data **15**

1.5 The development of Computational phenotyping **17**

Chapter Two: Use of real-world evidence data to evaluate the comparative effectiveness of second-line type 2 diabetes medications on chronic kidney disease *20*

Abstract **20**

Keywords **21**

Introduction **21**

Methods **22**

2.1 Study population 22

2.3 Renal outcomes 24

2.4 Covariates 24

2.5 Data cleaning 25

2.6 Statistical analysis 26

2.7 Sensitivity analysis 27

Results **27**

3.1 Risk for incident CKD, CKD hospitalization, and eGFR<45mL/min outcome in primary analysis 29

3.2 Sensitivity analysis 30

Discussion **31**

Conclusion **34**

Chapter Three: Comparison of electronic health record data with adjudicated registry data for identification of systemic lupus erythematosus subtypes using unsupervised learning *50*

Abstract **50**

Introduction **51**

Methods **54**

Results **59**

Discussion **66**

Chapter Four: Natural language processing for lupus nephritis computational phenotyping ***79***

Abstract **79**

Introduction **80**

Methods **82**

Data Source 82

Algorithm development: lupus nephritis phenotype 83

Model training and evaluation 84

External validation 84

Results **85**

Discussion **86**

Limitations **87**

Conclusion **87**

Chapter Five: Deep Neural Network Survival Model for Cardiovascular Disease Risk Prediction ***91***

Abstract **91**

Background **92**

Methods **94**

Results **100**

Discussion **102**

Limitations **104**

Conclusion **104**

Reference ***111***

Appendices ***123***

Vita ***147***

List of Tables, Illustrations, Figures, and Graphs

Table 1. Common data elements in EHR. 11
Table 2. Baseline characteristics of study cohort 36
Table 3. Hazard ratio in the fully adjusted cox regression model in primary analysis....44
Table 4. Hazard ratio in the fully adjusted cox regression model in sensitivity analysis....45
Table 5. General study cohort and cluster characteristics using CLD data vs NMEDW data for latent class analysis. 70
Table 6. Concordance table for each SLICC criteria between EDW data and CLD data. 71
Table 7. Cluster characteristics using CLD data vs EDW data on a subset of criteria having concordance > 70%. 72
Table 8. Cluster characteristics using CLD data vs EDW data on a subset of individuals with concordance across criteria > 65% 73
Table 9. Algorithm description....88
Table 10. Model performance....89
Table 11. Baseline characteristics for each race and sex group in training/internal validation dataset and external validation dataset. 75
Table 13. C-statistics plot by race-gender groups. In terms of calibration in 10x10 cross-validation 79

Figure 1. Cohort selection flowchart. 46
Figure 2. Unadjusted Kaplan Meier curve for incident CKD in different ADM groups.....47
Figure 3. Unadjusted Kaplan Meier curve for CKD hospitalization in different ADM groups. 48
Figure 4. Unadjusted Kaplan Meier curve for eGFR<45mL/min outcome in different ADM groups. 49
Figure 5. Consistency comparison between CLD clustering results vs EDW clustering results; 1a: TSNE visualization for clustering on CLD data; 1b: TSNE visualization for clustering on EDW data; 1c: concordance table for patient membership from EDW clustering vs C 74
Figure 6. TSNE plot for CLD clustering results vs EDW clustering results on the subset of criteria with concordance>70%; Left plot: clustering on CLD data; Right plot: clustering on EDW data 75
Figure 7. number of patients left using individual concordance cutoff from 0%-100%. 76
Figure 8. TSNE plot for CLD clustering results vs EDW clustering results on individuals with concordance > 65%; Left plot: clustering on CLD data; Right plot: clustering on EDW data; bottom: membership concordance table....77
Figure 9. ESRD Kaplan Meier curve grouped by different clustering groups. A.1, B.1: cluster on whole dataset; A.2, B.2: cluster on criteria that have concordance > 70%; A.3, B.3: cluster on patients that have concordance>65%....78
Figure 10. Number of patients with SLE.....90
Figure 11. Frameworks for neural network survival models.....106
Figure 12. C-statistics for PCEs, Nnet-survival, Deepsurv, Cox-nnet, and Cox PH-TWI in 10x10 cross-validation and MESA external validation....107
Figure 13. Kaplan-Meier Observed Event Rate and Predicted Event Rate for the ASCVD Outcome in the 10x10 cross-validation.....108
Figure 14. Kaplan-Meier Observed Event Rate and Predicted Event Rate for the ASCVD Outcome in the MESA Cohort. 109

Figure 15. Kaplan-Meier Observed Event Rate and Recalibrated Predicted Event Rate for the ASCVD Outcome in the MESA Cohort.. 110

Table S 1. Diagnosis codes for type 2 diabetes. 123
Table S 2. Diagnosis codes for study variables. 124
Table S 3. Change over time in HbA1c value by second-line ADM groups.....126
Table S 4. Hazard ratio for CKD incidence outcome in primary analysis. 128
Table S 5. Hazard ratio for CKD hospitalization outcome in primary analysis. 130
Table S 6. Hazard ratio for eGFR < 45 mL/min outcome in primary analysis. 132
Table S 7. Hazard ratio for CKD incidence outcome in sensitivity analysis.....134
Table S 8. Hazard ratio for CKD hospitalization outcome in sensitivity analysis. 136
Table S 9. Hazard ratio for eGFR < 45 mL/min outcome in sensitivity analysis.....138
Table S 10. Number of patients who were on only two medications in each second-line ADM group. ..140
Table S 11. Hazard ratio in the fully adjusted cox regression model among patients who took only two ADMs during the exposure period. 141
Table S 12. Regex search for Nasal/oral ulcer, arthritis, renal disorder, and lupus nephritis.....142
Table S 13. CUIs and their definition. 144
Table S 14. C-statistics for PCEs, Nnet-survival, Deepsurv, Cox-nnet, and Cox PH-TWI in 10x10 cross-validation and MESA external validation.....145

Figure S 1. BIC varying with number of clusters. 146

Chapter One: Introduction

1.1 Electronic health records overview

Electronic health record (EHR) systems are databases designed to store individual medical information recorded by health care providers during clinical care and healthcare administration (1). In 2009, the Health Information Technology for Economic and Clinical Health Act (HITECH) as part of the 2009 American Recovery and Reinvestment Act encouraged the use of EHRs by providing incentives to health care providers that met the "meaningful use" criteria. The criteria include using EHRs for relevant purposes and meeting certain technological requirements (2). As of 2017, nearly 86% of office-based physicians had adopted EHRs. Today, EHRs are in wide use and collect a broad variety of patient medical information including structured data, unstructured data, and semi-structured data. Structured data includes demographic information, diagnoses, lab results, medication use, even some environmental exposure data, such as smoking status (3). Unstructured data is primarily text such as clinical notes, while semi-structured data might be problem lists or radiology reports which, while also text, have consistent form or structure. The details of some common data elements and their description and examples in EHRs are shown in Table 1.

Table 1. Common data elements in EHRs.

Data elements	Descriptions, examples	Data structure
Demographics	Age, gender, race, sex, date of birth	Structured data
Encounters	Encounter start date, end date, encounter types,	Structured data

Clinical notes	Physician notes, discharge summaries, nursing documentation	Unstructured data
Diagnosis	ICD-9/ICD-10, SNOMED-CT	Structured data
Medication	RxNorm code, ATC	Structured data
Vital signs	Height, weight, blood pressure	Structured data
Laboratory test	LOINC	Structured data
Problem list	A list of diagnoses that allow physicians to quickly review a patient's ongoing health problems	Semi-structured data
Others	Social economics (income, occupation), environmental factors (smoking)	Structured data

Abbreviations: ICD-9-CM: International Classification of Diseases-Tenth Revision, ICD-10: International Classification of Diseases-Tenth Revision, SNOMED-CT: Systemized Nomenclature of Medicine – Clinical Terms, ATC: Anatomical Therapeutic Chemical classification, LOINC: Logical Observation Identifiers Names and Codes.

EHRs can advance healthcare research by themselves or in connection with other data sources such as registry data, genomic data, and socioeconomic status. Some active areas of research using EHRs examine drug comparative effectiveness, drug-drug interaction, drug repurposing, genome-disease association, disease progression modeling, and patient recruitment (1,4–6). Because of the scope of this work, we will focus on two areas of research: 1) Evaluating drug comparative effectiveness in disease outcome and 2) Predictive modeling of disease development.

1.2 Drug comparative effectiveness using EHRs and its recent progress

Traditionally, researchers evaluate the benefits and safeties of a drug through clinical trials starting from patient recruitment through longitudinal follow-up patients over time until adequate outcome data is collected. The whole process is often complex, costly and time intensive. In addition, one study can measure only a limited number of drugs and endpoints. In comparison, data in EHRs is already collected over time which makes it efficient and cost-effective. Further, its longitudinal nature enables both cross-sectional and longitudinal cohort studies. Many studies have now (7,8) used EHRs to evaluate treatment effectiveness in various disease areas including cardiovascular disease, cancer (9), diabetes (10,11), Alzheimer (12), and others (13–15). Several large efforts aim to conduct comparative effectiveness research using EHRs, sometimes referred to now as real-world evidence data. The Patient Centered Outcomes Research Network (PCORI) launched PCORnet, a distributed research network, to support comparative effectiveness research using nation-wide EHR data (16). The Health Care Systems Research Collaboratory supported by the National Institutes of Health (NIH) Common Fund aimed to strengthen national capacity to implement cost-effective, large-scale pragmatic clinical trials using data from routine clinical care (17). Some of the projects from the Collaboratory include evaluating Chlorhexidine versus routine bathing on multidrug-resistant organisms and all-cause bloodstream infections (18). In July 2021, the FDA approved Prograf (tacrolimus) in combination with other immunosuppressant drugs for organ rejection among patients with lung transplantation (19). This action marked a new chapter for real-world evidence studies as it demonstrated that a well-designed, observational study using real-world data (RWD) can be considered adequate under FDA regulations.

1.3 Predictive modeling using EHR data

Predictive modeling is another popular area of application of EHR data. Studies have used EHR data to develop predictive models for suicide prevention (20), mortality prediction in intensive care unit (21), cardiovascular disease prediction (22), acute kidney disease prediction (23) and many others. In recent years, more non-traditional approaches, such as machine learning methods have been developed to make use of the high dimensional data in EHRs to improve the predictive accuracy (24–26). Among these machine learning models, neural networks have had breakthrough success especially when applied to unstructured data such as image recognition and text classification (27–30). Efforts have been made to improve predictive accuracy by combining different approaches with neural networks. This includes using data from different modules (31), incorporating longitudinal information (26,32), and integrating an external knowledge base (33). Choi et al. used a list of one-hot encodings, a way of converting categorical data to a binary vector, to represents a patient's diagnosis histories. This list is then used as features for a recurrent neural network model (26) to predict heart failure. Ramsy et al. developed Med-Bert to pretrain embeddings on structured EHR data. The model took the longitudinal history of patient diagnoses to facilitate disease prediction (34). However, most of these models are applicable to binary or continuous outcomes rather than time-to-event data type. It is common in EHR studies that individuals are lost to follow-up (censored data) before the failure or event time and standard neural network models cannot train or test on these individuals, which leads to sample size reduction.

In 1995, Faraggi-Simon first combined neural network architectures with the Cox PH model to make use of censored information as well as to model non-linear features-outcome relations (35). Since then, there has been increasing interest in incorporating neural network architectures in

survival analysis. In current literature, there are two main ways of modeling time-to-event using neural networks: (i) adapting Cox PH model and using partial likelihood loss, e.g., Cox-nnet (36) and Deepsurv (37); or (ii) discretizing survival time and using a heuristic loss function, e.g., Nnet-survival (38). However, there is still limited application of these neural network models to disease prediction to make use of information from patients who are lost-to-follow-up, despite this being a common problem in EHR data and other medical data.

1.4 Existing Challenges using EHR data

Despite the unprecedented opportunities EHRs bring for clinical research, it also carries many challenges. Because EHRs were primarily designed for administrative purpose, the data itself is of varying quality, completeness, and biased based on the population coverage, selective missingness of data, or fragmentation of these data across sites (39).

Completeness

Data is missing in EHRs for several different reasons (39). Data is missing in a sense that a patient who seeks care in one hospital may go for care in another which causes care fragmentation and discontinuity. Madden et al. investigated data missingness among patients with depression and bipolar disorder by comparing the data in EHRs vs more complete data from insurance claim (40). They found that 60% and 54% of outpatient behavioral care data were missing in EHRs among depression and bipolar patients respectively. Data normally expected to be recorded could also be missing by mistake as never captured even at the time of patient registration. Culbertson et al., (41) investigated the completeness of demographic information across 13 healthcare facilities. They found that not only the completeness of different data elements varied but the completeness of the same data elements also varied greatly among

different health institutes (41). Lastly, data is missing in the sense that EHRs only capture information when a patient has a care episode. This is different from rigorously designed experiments where information is typically collected repeatedly on predefined time points.

Accuracy

Data accuracy is another issue related to EHR data. Data inaccuracy can occur because of human recording error or systematic problems due to avoidance of liability. A review by Chan et al. in 2010 examined the quality of the same clinical concepts across multiple institutions and found a great deal of variability. A study in Australia had patients review their own medical records to determine how many items were incorrect. They found that demographic details have high accuracy (94%) while allergies have low accuracy (61%) (42). In a study of evaluating data accuracy in EHR documentation, Weng et al. found major differences in symptom reporting between eye symptom questionnaire and what was documented in the EHRs (43).

Complex

EHR data is also highly complex. The data is often high dimensional, heterogeneous, and sparse. Data may also have random errors and contains a mixture of continuous variables and discrete variables. Different hospitals often have different terminology systems to define diagnoses, medications, laboratory tests and others. Some commonly used terminologies are ICD-10-CM, SNOMED-CT, ATC, LOINC codes. Just ICD-10-CM alone has more than 70,000 codes. Many efforts have been dedicated not only to define the concepts but also to the relationships and classification hierarchies among concepts. A lot of this data is also locked in clinical notes. which requires the use of sophisticated machine learning methods such as natural language processing to mine useful information from these notes. The data is also complex in the sense

that the relationship between data elements is often intertwined. For example, metformin is prescribed for type 2 diabetes patients who usually have high Hemoglobin A1C (HbA1c) values. In this case, the correlations among metformin, type 2 diabetes, and HbA1c make it harder for some statistical methods to model outcomes, especially when certain models (e.g., regression models) assume independence among covariates (45).

Bias

All the above-mentioned challenges with EHRs could be the sources of bias. Bias can occur when the selected population is not representative from the target population. In the case of EHR data, sicker patients who access the healthcare system frequently usually have more information recorded in EHRs while information is lacking on the population that are healthier (46). Therefore, identifying the types of bias in the study and on what stage they occur in the EHR system is critical to draw valid conclusions for observational or other types of studies.

1.5 The development of Computational phenotyping

Despite recognized limitations in the current use of EHRs, there are potential methodological improvements that can address some of these challenges. As an example, one method that has emerged is EHR-based computational phenotyping, which can accurately capture phenotypes from noisy observational data for clinical research. Funded by the National Human Genome Research Institute in 2007, the eMERGE network aims to combine genetic data with EHRs in support of genomic medicine. As an important step for this process, the network has developed 64 phenotypes.

Traditionally, phenotype development relies primarily on approaches that are heuristic, rule-based, and iterative which usually require domain expertise and is labor intensive. Later, more

and more studies adopt a supervised learning approach as reported in a review by Shivade et al. in 2013 (47). Many of these studies relied heavily on structured data such as ICD-9 (48), its successor ICD-10 (49), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) (50), RxNorm (51), and Logical Observation Identifiers Names and Codes (LOINC) (52). Although structured data is readily available and easy to access, a large amount of information in EHRs is locked in clinical notes. Natural language processing is a subfield of computer science that aims for machines to understand text or spoken words the way that humans can. In earlier studies, natural language processing algorithms usually work as feature extraction tools to mine useful features from clinical notes. Those features include but are not limited to key words, bag of words, term frequency – inverse document frequency (TF-IDF), and relationships between concepts. In recent years, models like Bidirectional Encoder Representations from Transformers (BERT) process whole notes at once for downstream tasks, which avoids feature extraction and information loss during the process (53,54). However, such an approach is computationally expensive, which may hinder its wide adoption in phenotyping. A review paper by Zeng et al. on computational phenotyping showed promise of using natural language processing to improve phenotyping accuracy (28).

In this thesis, we first evaluated the second-line type 2 diabetes medication comparative effectiveness on renal disorder using only structured data to define phenotypes and covariates in chapter 2 (55). We then evaluated how phenotypes defined with structured EHR data compared with phenotypes drawn from curated registry data in chapter 3. In chapter 4, we evaluated how well natural language processing-based methods for phenotyping can enhance phenotypes by combining structured EHR data and unstructured EHR data (56). In the last chapter, we focused on the value of novel machine learning methods on disease prediction. More specifically, we

explored the use of novel neural network survival methods for 10-year cardiovascular disease risk prediction using cohort data in chapter 5 (57). While not directly using EHR data, this method could potentially be applied in high dimensional EHR data in the future to improve predictive modeling. Taken together, this thesis highlights the insights we can derive from EHR data, challenges that EHR data presents when applied in clinical research, and offers ways to overcome some of the challenges from a methodological perspective.

Chapter Two: Use of real-world evidence data to evaluate the comparative effectiveness of second-line type 2 diabetes medications on chronic kidney disease

Abstract

Chronic kidney disease (CKD) is a common complication of type 2 diabetes mellitus (T2DM). Approximately one-third of patients with T2DM also have CKD. In clinical trial studies, several anti-diabetic medications (ADM) show evidence of preventing the progression of CKD. Biguanides (e.g., metformin) are widely accepted as the first line medication. However, the comparativeness effectiveness of second line ADMs on CKD outcomes in T2DM is unclear. In addition, results from clinical trials may not generalize into routine clinical practice. In this study, we aimed to investigate the association of second line ADMs with incident CKD, CKD hospitalization, and eGFR<45mL/min in T2DM patients using real-world data from electronic health records. Our study found that treatment with sodium-glucose cotransporter 2 **(**SGLT-2) inhibitors was significantly associated with a lower risk of CKD incidence in both primary analysis (hazard ratio, 0.43; 95% CI, [0.22;0.87]; p-value,0.02) (SU) as a second-line ADM. Treatment with a dipeptidyl peptidase 4 (DPP-4) inhibitor was significantly associated with lower CKD incidence (hazard ratio, 0.7; 95% CI, [0.53;0.96]; p-value, 0.03) and lower CKD hospitalization events (hazard ratio, 0.6; 95% CI, [0.37; 0.96]; p-value, 0.04) in the primary analysis. However, both associations were not significant in the sensitivity analysis. We did not observe significant association between use of glucagon-like peptide 1 receptor agonists (GLP-1RA), Thiazolidinediones (TZD), insulin and CKD incidence or hospitalization compared to use of SU as the second-line ADM.

Keywords

Type 2 diabetes, type 2 diabetes treatment, chronic kidney disease, electronic health records, second-line anti-diabetic medication, real-world data, sodium-glucose cotransporter 2, dipeptidyl peptidase 4, glucagon-like peptide 1 receptor agonists, Thiazolidinediones, Sulfonylureas, Insulin

Introduction

Type 2 diabetes is a major risk factor for chronic kidney disease (CKD) and is the leading cause of end stage renal disease (ESRD) with approximately 20% - 40% of patients developing diabetic nephropathy (1). Sulfonylureas (SU), dipeptidyl peptidase 4 (DPP-4) inhibitor, glucagon-like peptide 1 receptor agonists (GLP-1RA), sodium-glucose cotransporter 2 (SGLT-2), Thiazolidinediones (TZD), and Insulin are commonly used as second-line medication in addition to metformin. In the past decade, several randomized clinical trials have shown evidence of the newer anti-diabetic medications (ADMs) such as SGLT-2 and GLP-1RA in reducing risk for renal disease outcome compared to placebo (58–62)abl. However, there are few clinical trials (ongoing or finished) that directly compare the effectiveness of the newer AMDs to the older ones such as SU and DPP-4 inhibitor. Furthermore, patients in clinical trial studies are more compliant with therapy for a number of reasons (e.g. support from study staff), therefore, it is unclear how well results from clinical trials may apply to the general population in real clinical practice(63). Electronic health records (EHRs), a major source of real-world evidence data, can facilitate the understanding of treatment effectiveness in clinical practice using patient level data from the routine operation of the healthcare system and complement evidence on the efficacy of medications from randomized controlled trials (RCTs) (64). Recently, a study using real-world data investigated the comparative effectiveness of SGLT-2 inhibitor, GLP-1RA, DPP-4 inhibitor,

and SU in type 2 diabetics for preventing renal disease (65). This study was conducted within the Department of Veterans Affairs, which serves an older (mean age = 65.46), white (71.95%), male (94.49%) population. Therefore, additional research is required to compare the effectiveness on renal disease across the multitude of currently available drugs and drug classes. In this study, we compared the effect of commonly used second line anti-diabetic medications including SU, DPP-4 inhibitor, GLP-1RA, SGLT-2 inhibitor, TZD, and Insulin on renal outcomes using EHRs from a large integrated health delivery system.

Methods

2.1 Study population

The Northwestern Medicine Electronic Data Warehouse (NMEDW) is the primary data repository for all the medical records of patients who receive care within the Northwestern Medicine system (66). Established in 2007, the NMEDW contains records for over 3.8 million patients, with most EHR data going back to at least 2002, and with some billing claims data going back to 1998 or even earlier (66). We included patients who met the following criteria: 1) at least one prescription to an ADM 2) at least one diagnosis code for type 2 diabetes (see

Table S 1 for type 2 diabetes diagnosis codes)(10) 3) no excluded diagnoses: pregnancy, type 1 diabetes mellitus 4) at least one year of records in the database before their first ADM exposure and 5) at least three years of continuous ADM prescription records after first ADM exposure. Given that the first SGLT-2 inhibitor was approved by FDA for use in the United States in March 2013, we only included patients who had their first ADM exposure after 2013-03-01. Figure 1 shows the flow chart of the patient selection process.

We defined an ADM sequence as the chronological series of ADM prescription orders a patient received from the date of their first ADM prescription to one year later (e.g., “Metformin, Glipizide, Glipizide, Glucophage, Sitagliptin, Glipizide”). We then mapped each generic name and brand name to its respective drug class in Anatomical Therapeutic Chemical Classification (e.g., “Biguanides, SU, SU, Biguanides, DPP-4 inhibitor, SU”) (67). Repeated listings of the same medication in each sequence were understood to be refilled prescription. Therefore, in this example, only the first occurrence of a medication is kept in the final sequence (e.g., “Biguanides, SU, DPP-4 inhibitor”). During the 1-year period, only patients who were treated with biguanides as their first line of medication and had a second medication of either DPP-4 inhibitor, SGLT-2 inhibitor, GLP-1RA, SU, TZD, or Insulin were included in the study.

2.2 Exposure

The exposure in this study was the medication sequence during a patient one-year exposure period as described above. The sequences we were interested in include 'Biguanides, SU', 'Biguanides, DPP-4 inhibitor', 'Biguanides, Insulin', ‘Biguanides, GLP-1RA’, ‘Biguanides, SGLT-2 inhibitor’ and ‘Biguanides, TZD’. A patient might switch to a third medication during the one-year exposure time, we considered them in the same group as the patients who did not. For example, ‘Biguanides, SU, DPP-4’ was considered in the same second-line ADM group as ‘Biguanides, SU’. Patients who were only on Biguanides during the 1-year drug exposure period were excluded from the study. Index date was defined as one year after first the exposure to ADM.

2.3 Renal outcomes

Our primary outcome is CKD incidence identified by the first appearance of an associated International Classification of Diseases, 9th/10th revision (ICD9/10) diagnosis code (see Table S 2)[7]. Patients who had CKD, ESRD or macroalbuminuria before index date were excluded in the analysis for this outcome. In addition to the incident CKD event, we evaluated ADMs' associations with CKD hospitalization. CKD hospitalizations were identified based on ICD9/10 codes. Because diagnosis codes may be entered by clinicians in the EHRs only for more severe cases of CKD resulting in a significant delay between onset of CKD and an observable diagnosis or admission, we also performed analysis using laboratory value eGFR < 45 mL/min as an additional outcome. We used eGFR < 45 mL/min as an indicator of patients having moderate or more severe CKD and more specific than using a higher threshold of < 60mL/min. For this outcome, patients who had eGFR < 45mL/min, CKD diagnosis, or CKDs hospitalization prior to index date were excluded from the analysis for this outcome. For any of the outcomes, patients were followed up from their index date until meeting criteria of one of the above outcomes, last hospital visit date, end of 5-year follow up period or end of study date (2019-10-29) whichever comes first.

2.4 Covariates

We included 5 major categories of covariates: demographics, laboratory tests related to CKD, diagnoses from medical history before index date, medications that are known to affect CKD related outcome, insurance status and smoking status. Demographics includes age, gender, race

were collected at baseline (index date). Race categories included White, Black, Asian, and other races. We used White race as the reference group. Insurance status categories included Medicaid, self-pay, Medicare, commercial insurance. Commercial insurance was used as the reference group. Laboratory test included hemoglobin A1C (HbA1c), high density lipoprotein cholesterol (HDL), total cholesterol (TOTCHL), serum creatinine, body mass index (BMI). Medications include angiotensin-converting enzyme (ACE) inhibitors, aldosterone receptor antagonists (ARA), angiotensin II receptor blocker (ARB), antiplatelet drugs, beta-blockers, calcium channel blockers, diuretics, statins, and other lipid modifying drugs (loop and thiazide diuretics). Medical history includes cardiovascular disease (CVD), congestive heart failure (CHF), hypertension, vascular disease, vascular complication of diabetes (including skin ulcer), diabetic neuropathy, diabetic oculopathy, dyslipidemia, and other diabetic complications (diabetic nephropathy, nephrotic syndrome, nephritis/nephropathy, lower extremity amputations). For CKD hospitalization outcome, prior CKD hospitalization is also included as a covariate.

2.5 Data cleaning

We included the medical history and medication use covariates for any time before the index date. For lab tests, we used the measurement closest to the index date in the prior 2 years. We treated All the lab tests (HbA1c, HDL, TOTCHL, BMI, serum creatinine) as continuous variables. We excluded extreme values that were likely to be erroneous values (e.g. HDL > 100 mg/dL, BMI< 8 kg/m^2, BMI> 100 kg/m^2, HbA1c>20% (195 mmol/mol), TOTCHL > 1000 mg/dL) from the study. A patient may not have any laboratory test during our 2-year searching time window. In addition, some patients also have missing values in race and insurance status in

EHRs. To deal with the missing data, we used multiple imputation by chained equation (68). In our analysis, we used predictive mean matching as the imputation method for continuous variables and polytomous regression imputation for unordered categorical data. We created 10 multiple imputed datasets. For each imputation, we set the number of iterations as 20. We fit cox regression model on each imputed dataset. We then used Rubin's rule to pool coefficient estimates from 10 cox regression models (69).

2.6 Statistical analysis

We generated a simple statistical summary and stratified patients by second-line ADM medication. For categorical variables, we conducted Chi-square test for differences among second-line ADM classes. For continuous variables, we performed t-test to test the differences among second-line ADM classes at baseline. To evaluate the association of ADM drug classes and three different CKD-related outcomes, we developed a series of cox proportional hazard regression models. In our baseline model, we included only ADM class variables. In our basic demographic model, we included both ADM class and basic demographic information. In our demographic-medical history model, we included ADM class variables, basic demographics, and medical history. In the full model, we included all the mentioned variables in the Covariates section. Cox regressions were conducted using 'survival' package in R, version 3.6.0. Multiple data imputation was performed using 'MICE' package in R, version 3.6.0. Estimate pooling from 10 imputed datasets were performed using 'Hmisc' in R, version 3.6.0. Descriptive data statistics were generated using 'TableOne' module in python, version 3.7.3.

2.7 Sensitivity analysis

We conducted a sensitivity analysis to evaluate the robustness of our association findings. In the sensitivity analysis, we excluded patients with missing race, insurance status, HbA1c, total cholesterol, BMI, SBP, eGFR, and serum creatinine instead of imputing the missing data. The same cox regression models were performed to assess the association of ADMs with the three above mentioned outcomes.

Results

We identified 3403 patients in our final cohort who started with Biguanides (mainly metformin) as their first ADM and used either a DPP-4 inhibitor, SGLT-2 inhibitor, GLP-1RA, TZD, SU, or an insulin as their second line medication. Table 1 shows the baseline characteristics of patients stratified by second-line ADM classes. For CKD incident outcome, we further excluded patients who had CKD diagnosis before the index date, which left us 3216 patients in the final cohort for CKD incidence outcome (see Figure 1). For eGFR outcome, we further excluded patients with prior eGFR<45mL/min, which left us 2805 patients in the final cohort for eGFR<45mL/min outcome.

Among the 3403 patients, overall, the mean age of the population is 59.8 (S.D = 12.0). Among these patients, 125 (3.67%) patients have missing data in race. For these who have race information, 2408 (73.5%) were white, 458 (14.0%) black, 178 (5.4%) Asian, 234 (7.1) % classified as other races; 1853 (54.5%) patients were male, and 1550 (45.5%) were female. There are 44 (1.29%) patients with missing insurance. Among the patients with insurance, 113 (3.5%) were self-pay, 1706 (50.8%) had commercial insurance, 128 (3.8%) were Medicaid, 1407

(41.9%) were Medicare. For laboratory test, there were 391 (11.49%) patients with missing HDL-C, 129 (3.79%) patients with missing HbA1c, 18 (0.5%) patients with missing BMI, and 462 (13.6%) patients with missing TOTCHL. For those patients with available laboratory values, the mean of HDL-C is 46.0 mg/dL (S.D. = 12.3), the mean of HbA1c is 7.4% (57 mmol/mol) (S.D. = 1.5), the mean of TOTCHL is 166.5 mg/dL (S.D. = 40.4), the median of serum creatinine is 0.9 mg/dL ([Q1,Q3] = [0.7-1.0]). Among the 3403 patients, 1192 (35.0%) patients had DPP-4 inhibitors as second line ADM, 208 (6.1%) patients had GLP-1RA as second line ADM, 215 (6.3%) patients had insulin as second ADM, 1348 (39.6%) used SU as second line ADM, 355 (10.4%) patients had SGLT-2 inhibitor as second line ADM and 85 (2.5%) patients had TZD as second line ADM.

Statistical tests showed that age, race, insurance, smoking status, HDL-C, HbA1c, BMI, baseline serum creatinine, eGFR, CKD, CHF, diabetic neuropathy, other complications, vascular disease, VCD, oculopathy, CVD, use of ACE inhibitors, ARA, antiplatelet, diuretics, and statin were significantly different among the six second line ADM groups (see Table 2). In general, patients in newer medication groups (DPP-4 inhibitors, SGLT-2 inhibitors, GLP-1RA) tended to be younger, more likely to be white and have commercial insurance compared to patients in older medication groups (insulin, SU, TZD). In addition, patients in the insulin group tended to have higher HbA1c, lower eGFR and higher prevalence of use of renal related medications and comorbidities compared to other medication groups. In our regression analysis, in the full model, we adjusted for these differences by including demographics, medication history, diagnosis history, and laboratory tests as covariates.

During the 5-year follow up, 232 of 3216 patients (7.2%) had incident CKD with median length of follow-up of 3.23 years. Out of 3403 patients, 109 (3.2%) patients had CKD hospitalization

with median length of follow-up of 3.31 years. Out of 2805 patients, 277 (9.9%) patients had eGFR < 45 mL/min with median length of follow-up of 3.21 years. Figure 2, Figure 3, and Figure 4 showed the unadjusted Kaplan Meier curves for incident CKD, CKD hospitalization, and eGFR<45 min/mL outcomes, respectively. We also examined the change of HbA1c during the 5-year follow up within each second line ADM groups. Patients in the insulin group started with the highest HbA1c among all groups. For this group, HbA1c remained stable during the first 4 years and went up during year 4 to year 5. HbA1c also went up in SGLT-2 group when comparing the fifth year's value to the baseline value (first two years prior to the first ADM medication). For the GLP-1RA group, HbA1c in year five slightly decreased from baseline but remained relatively stable overall. For DPP-4, SU, TZD groups, HbA1c increased slightly during the 5-yearsfollow up. (Table S 3).

3.1 Risk for incident CKD, CKD hospitalization, and eGFR<45mL/min outcome in primary analysis

Table 2 showed the number of events and hazard ratio of all three renal outcome for each second-line ADM class using SU as reference group in a fully adjusted model in the primary analysis (see Table S 4 for hazard ratio in baseline model, basic demographics model, basic demographics/medical history model in primary analysis)[7]. For CKD incidence, both SGLT-2 inhibitor (HR, 0.43; 95% CI, [0.22;0.87]; P=0.02) and DPP-4 inhibitor (HR, 0.71; 95% CI, [0.53;0.96]; P=0.03) are significantly associated with lower CKD incidence event. GLP-1RA was associated with reduced risk for CKD incidence (HR, 0.52; 95% CI, [0.21;1.29]; P=0.16), but the P-value is not significant. For CKD hospitalization outcome, DPP-4 is significantly associated with lower risk for CKD hospitalization (HR, 0.60; 95% CI, [0.37;0.96]; P=0.03). Use

of the other ADMs also did not show significant difference in CKD hospitalization outcome compared to use of SU. For eGFR<45mL/min outcome, we observed that SGLT-2 inhibitor was associated with lower risk for eGFR<45mL/min (HR, 0.58; 95% CI, [0.32, 1.07]; P=0.08). However, this association was not statistically significant. We did not observe significant association between DPP-4 inhibitor and eGFR<45mL/min.

3.2 Sensitivity analysis

Table 4 showed hazard ratios for ADMs relative to use of SU in the sensitivity analysis in fully adjusted model (see Table S 7, Table S 8, and Table S 9 for hazard ratio in baseline model, basic demographics model, basic demographics/medical history model in sensitivity analysis)[7]. Our sensitivity analyses included only those patients with complete data. For the incident CKD event, after excluding patients with missing race, insurance status, HbA1c, insurance status, HDL, total cholesterol, BMI, baseline serum creatinine, and baseline eGFR, 2364 patients remained. Our analysis showed that in the fully adjusted model, SGLT-2 inhibitor use (HR, 0.42; 95% CI, [0.19, 0.92], P=0.03) was significantly associated with lower incidence of CKD. DPP-4 inhibitor use (HR, 0.75; 95% CI, [0.52;1.08]; P=0.12) was associated with lower CKD incidence but the P value was not statistically significant. For the CKD hospitalization outcome, after excluding patients with missing values, there were 2499 patients left. we did not observe any significant association between a particular second line ADM class relative to SU. For eGFR<45mL/min outcome, use of SGLT-2 inhibitor showed lower risk for eGFR<45mL/min compared to the reference group, though this association was not statistically significant (HR, 0.61; CI, [0.31,1.18], P=0.14).

Discussion

Randomized clinical trials and observational studies are two major ways to assess associations between medication and corresponding clinical outcomes (70). Although randomized clinical trials are rigorously designed, they have several limitations in testing the association between ADMs and renal outcomes among type 2 diabetes patients. First, the major randomized clinical trials compared ADM's effect to a placebo group. To date the ongoing Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study (GRADE) (54) is the only clinical trial (finished or ongoing) that we are aware of that evaluates the comparative effectiveness of newer ADMs vs older ADMs on glucose control (71). Further, most clinical trials treated second line ADMs as a monotherapy (59,72). Therefore, it is unclear how the effect changes when the ADMs are used as add-on therapy on top of metformin, which is the most common case in the primary care setting. Secondly, patients in clinical trials are better positioned to adhere to medication and therapy than in real clinical settings. Therefore, the results drawn from clinical trials may not apply to real clinical settings. Observational studies using EHR data like ours could potentially close these gaps by providing evidence from the real-world data of general populations.

In this study, we used EHR data to investigate the association of second-line ADMs with renal outcomes among 3403 type 2 diabetes patients who initiated treatment with metformin. We found that both the use of SGLT-2 inhibitor and DPP-4 inhibitor showed significant association with less incident CKD as compared to use of SU as a second-line medication. However, in the sensitivity analysis, only SGLT-2 inhibitor remained significantly associated. For the CKD hospitalization event, DPP-4 inhibitor appeared protective for CKD hospitalization in the

primary analysis, but the association was not statistically significant in the sensitivity analysis. For eGFR<45mL/min outcome, the association between use of SGLT-2 inhibitors and lower event rate of eGFR<45 mL/min was marginally significant (P=0.08). The different results for eGFR<45mL/min and ICD based CKD incidence outcome likely reflects the discrepancy of recording time between lab values and diagnosis code.

In each second-line ADM group, around half of the patients switched to a third medication during the one-year exposure period (see Table S 10 for details). To evaluate the impact of switching to a third or more medication might have on the result, we examined patients who took 2 medications only during the one-year exposure. We observed that the "Biguanide, DPP-4", "Biguanides, SGLT-2" group had lower hazard risk for CKD incidence compared to the reference group (see Table S 11). However, the association was not significant (P=0.19 for DPP-4 inhibitor group and P=0.13 for SGLT-2 inhibitor group). This is likely due to small size after restricting to patients taking 2 medications only. In addition, SGLT2 is a relatively new medication, the follow-up time might not be long enough to develop CKD.

The above results from primary and secondary analysis are largely in line with results from placebo-controlled cardiovascular outcomes trials, which showed empagliflozin, canagliflozin, dapagliflozin, and Linagliptin all had beneficial effects on indices of CKD (73,74). For example, the CREDENCE (Canagliflozin and Renal Events in Diabetes with Established Nephropathy Clinical Evaluation) trial showed that type 2 diabetes patients with prior albuminuria assigned canagliflozin (a SGLT-2 inhibitor) had lower risk of composite renal events compared to placebo group (73). Cooper et al. found Linagliptin (a DPP-4 inhibitor) significantly reduced the risk of kidney disease events compared to placebo among type 2 diabetes patients (74). Groop et al (75)

found linagliptin on top of renin-angiotensin-aldosterone system significantly reduced albuminuria among patients with type 2 diabetes and renal disfunction. We did not find statistically significant difference between GLP-1RA and SU, Insulin and SU, TZD and SU in either primary analysis or sensitivity analysis regarding CKD incidence and CKD hospitalization event. For GLP-1RA, previous studies showed inconclusive evidence of its effect in reduction of renal disease (76). The LEADER (76) study showed GLP-1RA had benefit in reduction of new onset albuminuria while the LIRA-RENAL study showed no improvement in urine ACR, and eGFR change among renal impaired T2DM patients compared to the placebo group (77). For TZD vs SU, Insulin vs SU, we did not find evidence of one is superior to the other in renal disease outcome in the literature.

Our study has several strengths. First, our clearly defined exposure time starting from first prescription of metformin, with additional prescription of a second line ADM within a year: this ensured comparison of patients who were in a clinically similar diabetes stage. Second, we only included patients who started ADMs after 2013-03-01 which is the time when the first SGLT-2 inhibitor was approved by FDA for use. This ensures that patients using different second-line ADMs have comparable follow-up times and avoids the immortal time bias (78). Third, we used a rigorously defined patient cohort: we excluded any patients who had type 1 diabetes diagnosis or gestational diabetes. In addition, we only included patients who had 3 continuous years of ADM prescriptions since their first ADM exposure. This makes sure that each patient has sufficient and comparable data depth. Last while there could be confounders that were not included in the study, our list of covariates is relatively comprehensive based on a review of the literature. This list contains basic demographic information, laboratory tests, and a wide range of diagnoses that might not only be associated with renal disease, but also suggest disease severity.

Our study has several limitations. Firstly, for simplicity, we could not distinguish between a switch of medication from an add-on medication. For example, patients who have a DPP-4 inhibitor as the second medication in their sequence could either: 1) have switched entirely from metformin to a DPP-4 inhibitor, or 2) began using metformin and a DPP-4 inhibitor together. Secondly, while we controlled for as many variables/confounders as we could, it is possible that there remain other covariates we did not capture, which could affect prescription bias or are still associated with the renal outcomes. Therefore, conclusions drawn from our study should be interpreted with caution. Thirdly, as EHR data is primarily designed for administrative purpose, it is well known that EHRs have missing data issues. Because this is a single site study, patients may have sought care elsewhere. Therefore, it is likely that there were data points that we were unable to capture in this single site study. Additionally, this was a single center study of an academic hospital in a major US city, limiting its generalizability to other populations. Finally, diabetes duration is an important factor that affects renal outcomes. In our study, we did not control diabetes duration as a covariate because it was unavailable in our dataset. That said, we collected medication sequences at first exposure to ADMs, suggesting a clinically recent disease onset time.

Conclusion

In conclusion, our study assessed the association of common second-line medications with renal outcomes as compared to SU's from March 2013 through October 2019 using EHRs. Similar to previous studies, our results showed that SGLT-2 inhibitors were consistently associated with lower CKD incidence in both primary and sensitivity analyses and DPP-4 inhibitor was

significantly associated with lower CKD incidence in primary analysis. We did not observe any statistical significance between Insulin and SU, TZD and SU in either primary analysis or sensitivity analysis regarding CKD incidence and CKD hospitalization event in our dataset. Unlike reported in Xie et al.'s study (65), We also did not observe significant benefit of GLP-1RA compared to SU in CKD hospitalization and CKD incidence, possibly due to small sample size. Additional research using multi-site real-world data and larger sample sizes may be needed to confirm the generalizability of our results.

Table 2. Baseline characteristics of study cohort

	Grouped by second-line medication								
	Missing	**Overall**	**DPP-4**	**GLP-1**	**Insulin**	**SGLT2**	**SU**	**TZD**	**P-Value**
n		3403	1192	208	215	355	1348	85	
Basic Demographics									
Age, mean (SD)	0	59.8 (12.0)	59.8 (11.6)	54.9 (10.8)	59.5 (12.3)	56.3 (11.4)	61.2 (12.3)	63.5 (10.9)	<0.001
Gender, n (%)		1853 (54.5)	656 (55.0)	91 (43.8)	111 (51.6)	204 (57.5)	739 (54.8)	52 (61.2)	
Race, n (%)									
Asian	125	178 (5.4)	61 (5.3)	11 (5.7)	8 (4.0)	15 (4.3)	75 (5.7)	8 (9.8)	<0.001

Black		458 (14.0)	136 (11.9)	30 (15.5)	65 (32.2)	21 (6.1)	198 (15.1)	8 (9.8)	
Other races		234 (7.1)	86 (7.5)	9 (4.7)	15 (7.4)	25 (7.2)	97 (7.4)	2 (2.4)	
White		2408 (73.5)	860 (75.2)	143 (74.1)	114 (56.4)	285 (82.4)	942 (71.8)	64 (78.0)	
Insurance, n (%)									
Self-pay	44	118 (3.5)	43 (3.6)	5 (2.5)	13 (6.2)	5 (1.4)	49 (3.7)	3 (3.6)	<0.001
Commercial		1706 (50.8)	622 (52.8)	144 (70.6)	101 (48.3)	241 (68.3)	559 (42.0)	39 (46.4)	
Medicaid		128 (3.8)	34 (2.9)	8 (3.9)	6 (2.9)	11 (3.1)	69 (5.2)		
Medicare		1407 (41.9)	480 (40.7)	47 (23.0)	89 (42.6)	96 (27.2)	653 (49.1)	42 (50.0)	

Smoking status, n (%)		475 (14.0)	167 (14.0)	31 (14.9)	47 (21.9)	36 (10.1)	186 (13.8)	8 (9.4)	0.004
Laboratory test									
HDL-C, mean (SD)	391	46.0 (12.3)	45.9 (12.2)	47.0 (13.0)	46.2 (12.4)	44.6 (12.0)	45.9 (12.2)	49.8 (16.4)	0.033
HbA1c, mean (SD)	129	7.4 (1.5)	7.3 (1.3)	7.2 (1.6)	7.9 (2.0)	7.3 (1.4)	7.5 (1.6)	7.0 (1.1)	<0.001
BMI, mean (SD)	18	34.2 (7.7)	33.4 (7.4)	37.2 (7.4)	34.3 (7.1)	35.9 (8.1)	33.9 (7.9)	33.6 (7.3)	<0.001
TOTCHL, mean (SD)	462	166.5 (40.4)	165.3 (39.7)	168.0 (35.0)	163.4 (48.4)	169.2 (38.2)	167.9 (41.4)	155.3 (31.3)	0.059
Baseline serum	166	0.9 [0.7,1]	0.9 [0.7,1]	0.9 [0.7,1]	0.9 [0.8,1.1]	0.8 [0.7,1]	0.9 [0.8,1]	0.9 [0.8,1.1]	0.005

creatinine, median [Q1,Q3]									
eGFR, mean (SD)	199	76.4 (28.8)	77.1 (28.7)	62.7 (14.9)	60.3 (22.1)	82.7 (28.8)	79.2 (30.0)	73.1 (28.4)	<0.001
Diagnosis History									
CKD, n (%)		187 (5.5)	62 (5.2)	11 (5.3)	23 (10.7)	10 (2.8)	74 (5.5)	7 (8.2)	0.004
CHF, n (%)		213 (6.3)	62 (5.2)	16 (7.7)	23 (10.7)	16 (4.5)	90 (6.7)	6 (7.1)	0.028
Hypertension, n (%)		2761 (81.1)	983 (82.5)	161 (77.4)	181 (84.2)	274 (77.2)	1089 (80.8)	73 (85.9)	0.089
Dyslipidemia, n (%)		2635 (77.4)	939 (78.8)	152 (73.1)	169 (78.6)	272 (76.6)	1039 (77.1)	64 (75.3)	0.534

Diabetic Neuropathy, n (%)		332 (9.8)	105 (8.8)	19 (9.1)	45 (20.9)	29 (8.2)	127 (9.4)	7 (8.2)	<0.001
Other Complications, n (%)		274 (8.1)	87 (7.3)	17 (8.2)	37 (17.2)	30 (8.5)	93 (6.9)	10 (11.8)	<0.001
Vascular disease, n (%)		330 (9.7)	103 (8.6)	22 (10.6)	40 (18.6)	29 (8.2)	130 (9.6)	6 (7.1)	<0.001
VCD, n (%)		466 (13.7)	147 (12.3)	19 (9.1)	37 (17.2)	58 (16.3)	185 (13.7)	20 (23.5)	0.005
oculopathy, n (%)		128 (3.8)	28 (2.3)	11 (5.3)	24 (11.2)	13 (3.7)	47 (3.5)	5 (5.9)	<0.001

CVD, n (%)		644 (18.9)	196 (16.4)	43 (20.7)	64 (29.8)	57 (16.1)	268 (19.9)	16 (18.8)	<0.001
hypoglycemia, n (%)		43 (1.3)	6 (0.5)	3 (1.4)	4 (1.9)	6 (1.7)	23 (1.7)	1 (1.2)	0.116
Medication History									
ACE inhibitors, n (%)		2162 (63.5)	765 (64.2)	128 (61.5)	151 (70.2)	204 (57.5)	849 (63.0)	65 (76.5)	0.005
ARA, n (%)		271 (8.0)	82 (6.9)	18 (8.7)	26 (12.1)	26 (7.3)	117 (8.7)	2 (2.4)	0.037
ARB, n (%)		251 (7.4)	97 (8.1)	25 (12.0)	13 (6.0)	25 (7.0)	84 (6.2)	7 (8.2)	0.054
CCBs, n (%)		817 (24.0)	294 (24.7)	47 (22.6)	61 (28.4)	71 (20.0)	321 (23.8)	23 (27.1)	0.274

antiplatelet, n (%)		1116 (32.8)	376 (31.5)	83 (39.9)	99 (46.0)	99 (27.9)	431 (32.0)	28 (32.9)	<0.001
Beta Blocker, n (%)		1098 (32.3)	374 (31.4)	62 (29.8)	74 (34.4)	99 (27.9)	459 (34.1)	30 (35.3)	0.228
Diuretics, n (%)		1368 (40.2)	466 (39.1)	91 (43.8)	108 (50.2)	124 (34.9)	544 (40.4)	35 (41.2)	0.011
Lipid modifier, n (%)		541 (15.9)	190 (15.9)	38 (18.3)	37 (17.2)	50 (14.1)	204 (15.1)	22 (25.9)	0.111
Statin, n (%)		2284 (67.1)	805 (67.5)	135 (64.9)	154 (71.6)	224 (63.1)	897 (66.5)	69 (81.2)	0.024

This table shows the baseline patient characteristics for CKD hospitalization outcome before missing data imputation. For CKD incidence outcome, patients with prior ESRD and prior CKD were excluded from the study cohort. For composite renal outcome,

patients with prior ESRD were excluded from the study cohort. Abbreviations: HDL, high density cholesterol, HbA1c, hemoglobin A1c; BMI, body mass index; TOTCHL, total cholesterol; VCD: vascular complications of diabetes; CHF, congestive heart failure; CVD, cardiovascular disease; ACE inhibitor, angiotensin-converting enzyme inhibitors; ARA, aldosterone receptor antagonists; ARB, angiotensin II receptor blocker. DPP-4, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones. Measurement units for laboratory tests are as the following: HDL-C, mg/dL; HbA1c, %; BMI, kg/m^2; TOTCHL, mg/dL; serum creatinine, mg/dL.

*Other complications include diabetic nephropathy, nephrotic syndrome, nephritis, diabetic retinopathy, cataract, lower extreme amputation

Table 3. Hazard ratio in the fully adjusted cox regression model in primary analysis

	CKD incidence (N=3216, E = 232)		CKD hospitalization (N=3403, E = 109)		eGFR < 45 mL/min (N = 2805 , E = 277)	
	HR (95% CI)	Pval	HR (95% CI)	Pval	HR (95% CI)	Pval
DPP-4	0.71, [0.53;0.96]	**0.03**	0.6,[0.37;0.96]	**0.04**	0.95, [0.72;1.26]	0.73
GLP-1RA	0.52, [0.21;1.30]	0.16	1.05,[0.37;3.02]	0.92	0.87, [0.46;1.65]	0.68
Insulin	0.93, [0.54;1.59]	0.80	0.52,[0.24;1.17]	0.11	1.18 [0.71;1.95]	0.52
SGLT-2	0.43, [0.22;0.87]	**0.02**	0.81,[0.31;2.09]	0.66	0.58, [0.32;1.07]	0.08
TZD	1.03, [0.49;2.13]	0.93	1.25,[0.44;3.70]	0.65	0.96, [0.44;2.08]	0.91

Abbreviations: HR, hazard ratio; DPP-4, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones; Pval, p-value; E, number of events.

Table 4. Hazard ratio in the fully adjusted cox regression model in sensitivity analysis

	CKD incidence (N=2364, E=159)		CKD hospitalization (N=2499, E =75)		eGFR < 45 mL/min (N =2164 , E =197)	
	HR (95% CI)	Pval	HR (95% CI)	Pval	HR (95 %)	Pval
DPP-4	0.75; [0.52;1.08]	0.12	0.75;[0.42;1.32]	0.32	0.98,[0.71;1.36]	0.90
GLP-1	0.53; [0.19;1.48]	0.23	0.94;[0.26;3.43]	0.92	0.97,[0.48;1.96]	0.93
Insulin	0.72; [0.37;1.43]	0.35	1.00; [0.39;2.57]	1.00	1.49,[0.83;2.65]	0.18
SGLT-2	0.42; [0.19;0.92]	**0.03**	1.35;[0.50;3.66]	0.56	0.61,[0.31;1.18]	0.14
TZD	1.15; [0.49;2.71]	0.75	0.9; [0.21;3.90]	0.89	0.97,[0.42;2.27]	0.95

Abbreviations: HR, hazard ratio; DPP-4, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones; Pval, p-value; E, number of events.

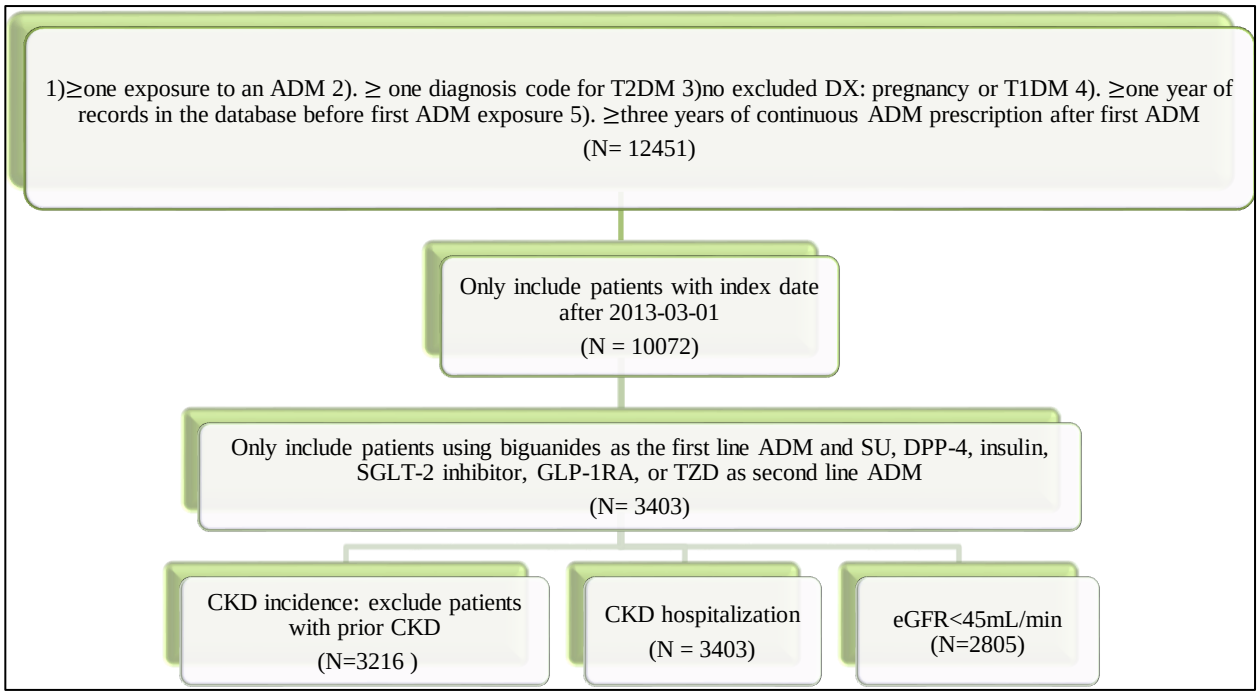


Figure 1. Cohort selection flowchart.

Abbreviations: T2DM, type 2 diabetes mellitus; ADM, antidiabetic medication; DX, diagnosis; T1D, type 1 diabetes; SU, sulfonylureas; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones.

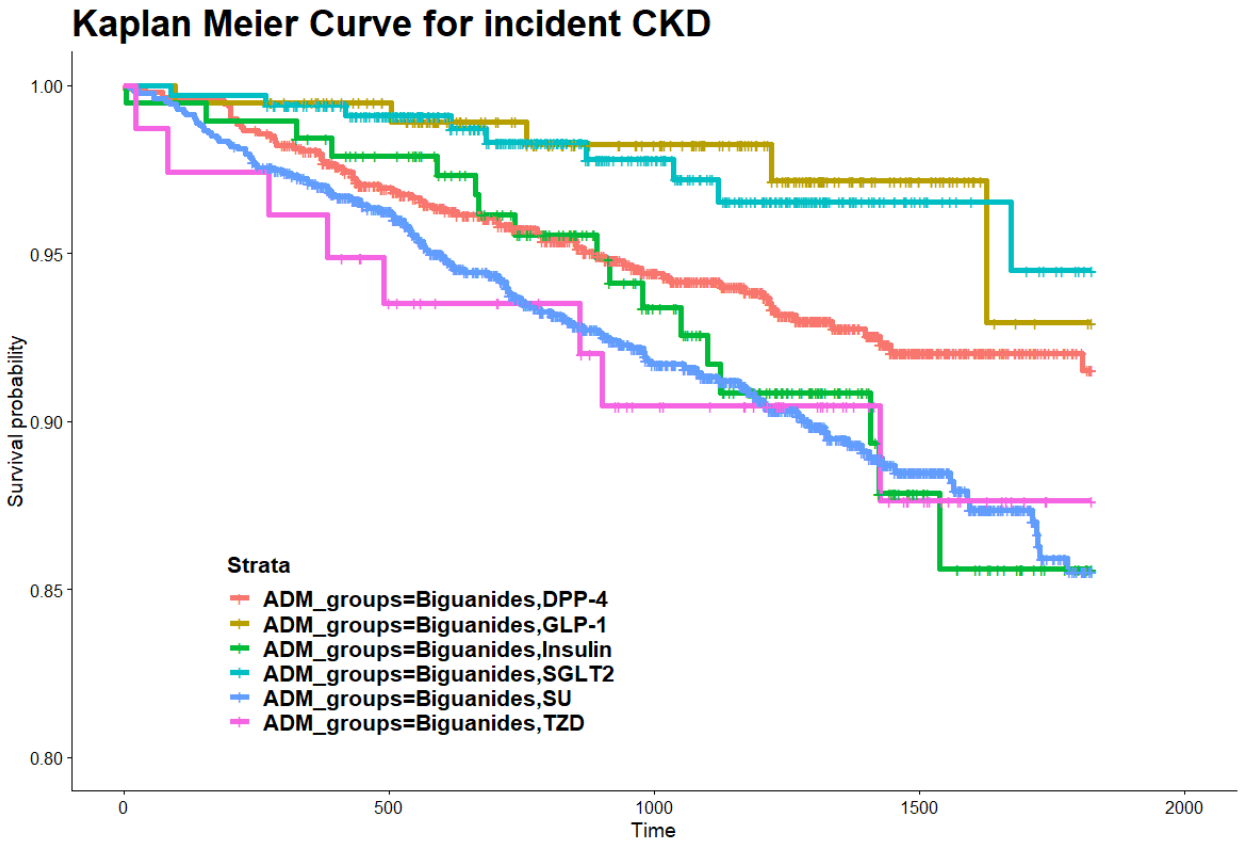


Figure 2. Unadjusted Kaplan Meier curve for incident CKD in different ADM groups.

Abbreviations: CKD, chronic kidney disease; ADM, anti-diabetic medication; DPP4, dipeptidyl peptidase 4 inhibitors; GLP-1, glucagon-like peptide receptor agonists; SGLT2 inhibitor, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones.

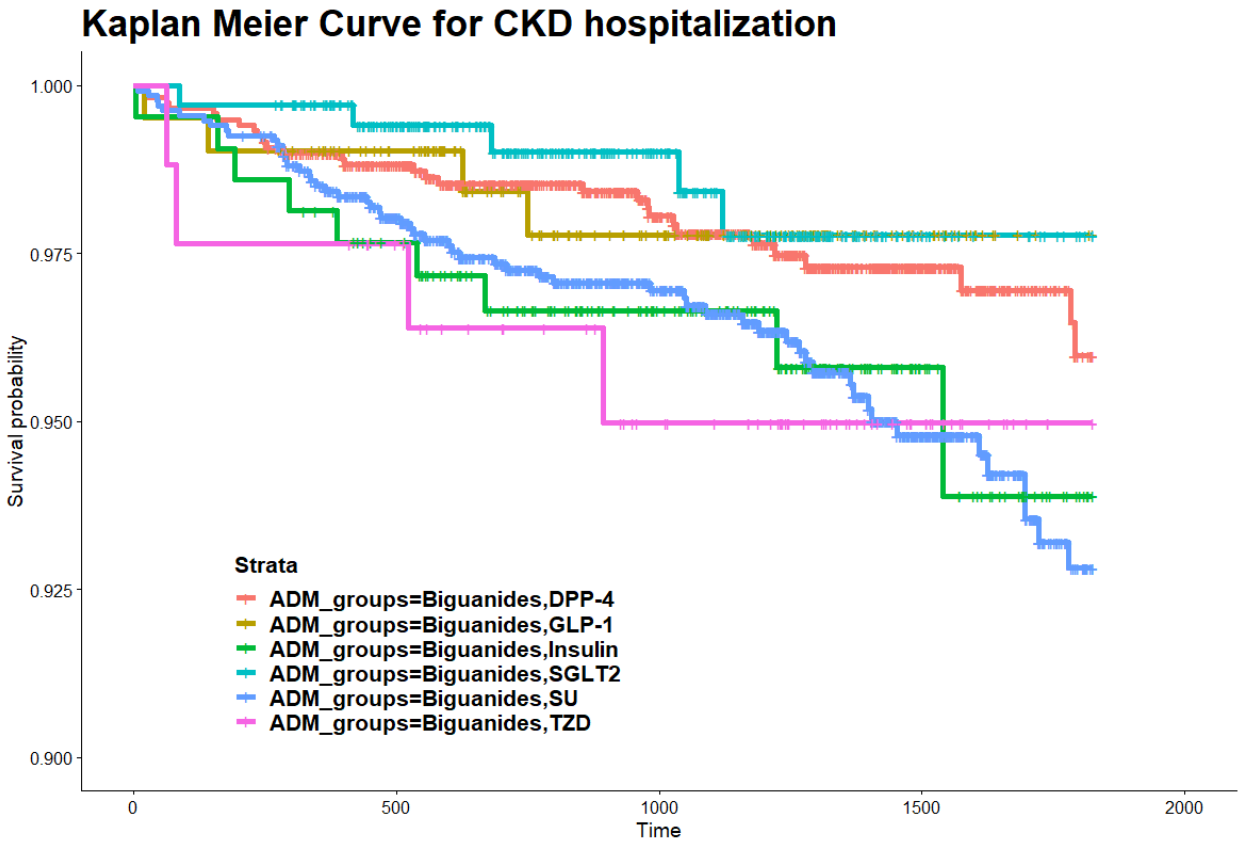


Figure 3. Unadjusted Kaplan Meier curve for CKD hospitalization in different ADM groups.

Abbreviations: CKD, chronic kidney disease; ADM, anti-diabetic medication; sDPP4, dipeptidyl peptidase 4 inhibitors; GLP1RA, glucagon-like peptide receptor agonists; SGLT2 inhibitor, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones.

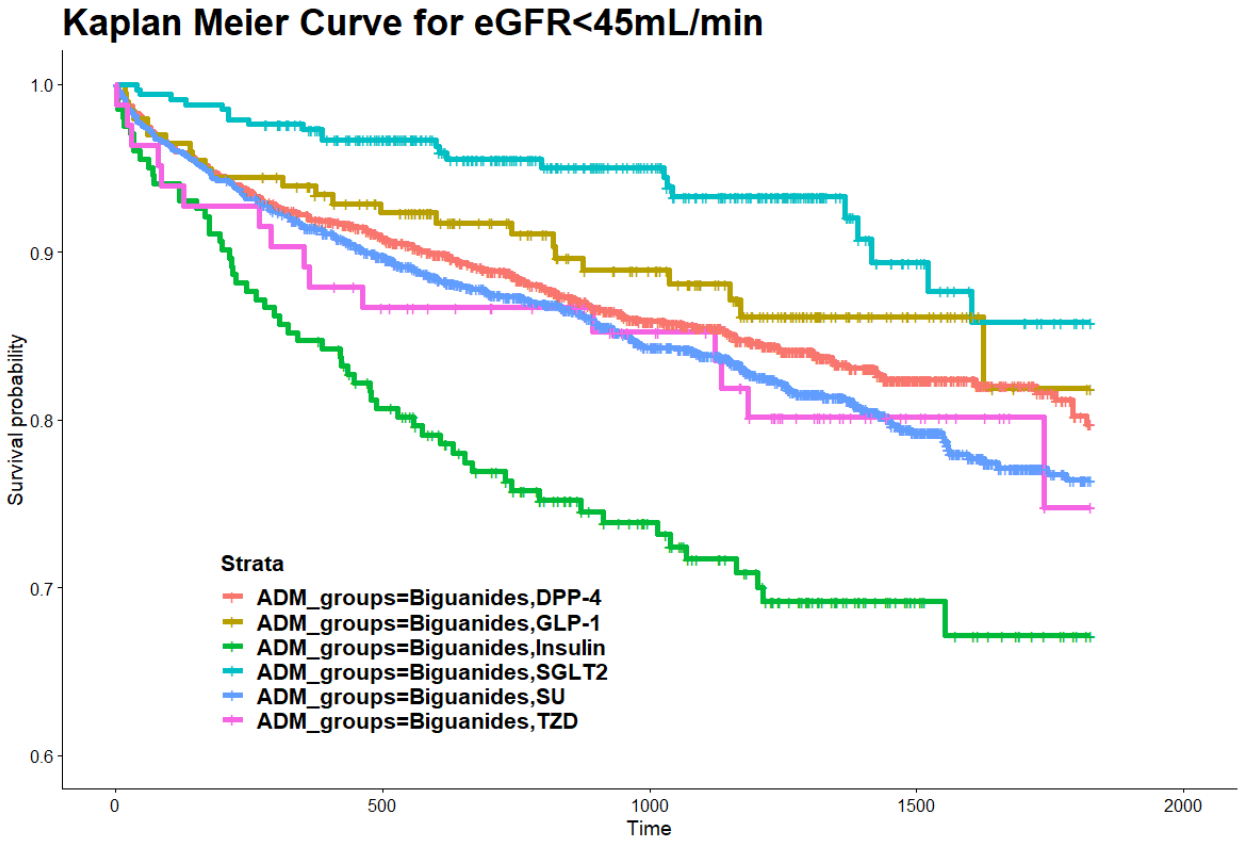


Figure 4. Unadjusted Kaplan Meier curve for eGFR<45mL/min outcome in different ADM groups.

Abbreviations: DPP4, dipeptidyl peptidase 4 inhibitors; GLP1RA, glucagon-like peptide receptor agonists; SGLT2 inhibitor, sodium-glucose cotransporter 2 inhibitor; TZD, Thiazolidinediones.

Chapter Three: Comparison of electronic health record data with adjudicated registry data for identification of systemic lupus erythematosus subtypes using unsupervised learning

Abstract

Objective: Precision medicine focuses on designing medical treatments by considering individual differences. Having rich individual patient information, electronic health record (EHR) data is an important foundation for precision medicine studies, however, EHR systems are primarily designed for administrative purposes and do not generally collect all information needed to define disease symptomology, particularly for complex systemic diseases. Systemic Lupus Erythematosus (SLE) is an autoimmune disease with diverse manifestations and a highly variable and heterogeneous presentation. In this study, we used latent class analysis to identify SLE subtypes based on disease classification criteria using either adjudicated registry data or EHR data for the same patient population, compared the subpopulations of derived from the different data sources, and explored ways to improve cluster validity when using EHR data for clustering analysis.

Materials and Methods. The Systemic Lupus International Collaborating Clinics (SLICC) group developed classification criteria for SLE that includes 11 clinical and 6 laboratory-defined manifestations. We performed latent cluster analysis using these manifestations and basic demographic information to identify distinct SLE clusters using EHR data and clinical registry data respectively. We compared the characteristics, survival curves and membership consistency of the clusters between the two datasets. To see if there are ways to improve the cluster

consistency derived from these two datasets, we further performed two sensitivity analyses by using a subset of features and a subset of patients respectively for clustering analysis. We compared the consistency of clustering results in both sensitivity analyses.

Results Our results show that EHR and registry-based data generate distinct clusters that are associated with significant differences in end-stage renal disease outcomes. A comparison of cluster membership between the EHR and registry datasets show low concordance (53%). LCA performed on a subset of patients with concordant features does not improve the subpopulation concordance. LCA performed using a subset of features that were concordant between the EHRs and registry improve the concordance to 76%.

Discussion and Conclusion Our data suggest that conclusions drawn from clustering analysis are data source-specific, and where possible, should be assessed relative to other data sources. If no other data sources are available, performing clustering analysis on a subset of features that are expected to have high concordance (e.g. lab values) may generate more reliable results.

Introduction

Precision medicine focuses on designing personalized medical treatment by considering individual attributes, including environmental, genetic, and lifestyle differences. Machine learning strategies are essential to precision medicine and help to identify subgroups within population that may have unique attributes or needs by analyzing large datasets of biological and clinical measures (79). Unsupervised learning algorithms, which do not rely on specific outcomes to help identify subgroups within datasets, can cluster patients into groups based on specified features in order to find homogeneous sub-phenotypes within a complex disease. The

techniques can be used to better understand of the etiology and pathophysiology of the disease as well as provide a the foundation for more targeted treatment (80). Disease-focused patient registries are often used for unsupervised learning. Patient registries, which are organized systems that collect uniform data based on observational study methods to evaluate specified outcomes for a specific population have long been viewed as the gold standard for precision medicine research (81). However, the development and maintenance of registries is laborious, often requiring significant effort on the part of clinical and research team members to collect and curate patient data. Electronic health record (EHR) systems are databases designed to store individual medical information recorded by health care providers during the course of clinical care and healthcare administration (1). They are in wide use and collect a broad variety of medical information including diagnoses, lab results, medication use over time as well as primary demographic information and even some environmental exposure data, such as smoking status. Several studies have demonstrated the effectiveness of using EHR data to identify homogeneous groups for more targeted treatment [2, 5, 6]. However, the information in the EHRs is recorded primarily for billing purposes and continuity of clinical care, may not be collected by the same set of people with a set of defined standards as occurs for registries, and may be incomplete (39). It remains unclear if the conclusions drawn from EHRs are comparable to those drawn from more rigorously curated data sources, such as registries, for the development precision medicine tools.

To address this question, it was essential to identify a complex disease with heterogeneous presentation and well-defined clinical and laboratory-based classification characteristics that could be used for unsupervised subpopulation analysis as well as a population of people

experiencing this disease who had information documented in both an adjudicated registry and an aligned EHR system. Systemic lupus erythematosus (SLE) is a complex autoimmune disease with heterogeneous presentation and a wide range of clinical outcomes which also has several rigorously developed classification criteria (84–86) that can be used to define disease attributes for machine learning. At the Northwestern University Feinberg School of Medicine, we have developed the Chicago Lupus Database (CLD), a registry of people with lupus with physician-adjudicated classification criteria (87) that is linked to their electronic health record data in the Northwestern Medicine Enterprise Data Warehouse (NMEDW).

Significant therapeutic breakthroughs for people with SLE have been few and far between. Only one therapy has been specifically approved or licensed for use in SLE in the past 60 years (88). The underlying heterogeneity of the disease likely contributes to the poor outcomes of clinical trials and makes effective clinical management of people with lupus challenging. Care for people with SLE and therapeutic clinical trials may be improved by identifying subpopulations of people with similar clinical manifestations using unsupervised machine learning strategies. Clinical classification criteria are developed to describe the disease manifestations in people primarily for research and clinical trials The Systemic Lupus International Collaborating Clinics (SLICC) developed a list of classification criteria for SLE based on prevalent/important SLE manifestations that includes 11 clinical criteria that are primarily defined based on clinical observations and 6 immunological criteria that are primarily defined by laboratory test results (84). For each patient documented in the CLD, a clinical expert has recorded which clinical classification criteria describe the patient. To determine whether it was possible to identify clinically relevant subpopulations of people with SLE we applied latent class analysis (LCA) to

SLE patient data in the CLD to identify subpopulations using SLE clinical classification criteria as features. LCA was chosen as our machine learning strategy because it is a non-parametric statistical model that identifies subpopulations based on multivariate categorical data (89). It assumes the independence of observations given their latent classes (i.e. subpopulation) and has been previously shown to works well with discrete data types (90–92), which are present in the CLD.

To better understand the differences between medical record data and registry data as a foundation for subpopulation identification, we focused on identifying SLE subtypes for 472 patents who had data in both the CLD and the NMEDW using the SLICC criteria and basic demographic information as a foundation for latent class analysis. We compared subpopulations identified using registry data and EHR data by comparing their cluster characteristics, assessing their relationship to significant clinical outcomes, and evaluating subpopulation membership consistency between two datasets. We further explored two methods to improve the generalizability of the clustering algorithm by performing clustering on a subset of features with high concordance across the two data sources and a subset of patients whose clinical features were highly similar between both data sources respectively.

Methods

Data sources and study population

Chicago Lupus Database (CLD): Established in 1991, the CLD is a physician-adjudicated registry of 1,052 patients with possible or definite lupus according to the revised 1982 American College of Rheumatology classification criteria (86). The CLD has laboratory data, symptoms,

and patient demographics based on each known visit. If a patient was referred, previous history information from the notes is documented. The CLD has been updated to reflect the 17 SLICC classification criteria. The 11 clinical criteria are acute cutaneous lupus, chronic cutaneous lupus, oral ulcers, nonscarring alopecia, synovitis, serositis, renal disorder, neurologic disorder, hemolytic anemia, leukopenia, thrombocytopenia and the 6 immunological criteria are: antinuclear antibody (ANA), Anti-double stranded DNA (anti-dsDNA), anti-Smith antibody (anti-Sm), antiphospholipid antibody (APA), low complement, and direct Coombs test. Patients are confirmed to have each criterion or not by physician chart review. SLE diagnosis date is the date a patient first met at least 4 ACR criteria.

Northwestern Medicine Enterprise Data Warehouse (NMEDW): Established in 2002, NMEDW is an integrated data repository that stores observations of more than 6.6 million distinct patients. Everyday, more than 2.8 billion new data elements are loaded into the database, including majority from electronic heatlh records, part from research database and others. In our study, we collected our patient EHR data from NMEDW from 2002-01-01 to 2018-08-23.

Cohort Identification: To create a cohort for our study we identified all persons in the CLD who had a diagnosis of lupus and also satisfied the SLICC classification criteria for lupus and also had medical records in the NMEDW. Medical record numbers are documented in the CLD and were used to match the same subject at NMEDW. In order to ensure sufficient and comparable data depth, we also required that participants in the CLD have at least 3 visits documented in NMEDW. Patients with missing basic demographic information and SLE diagnosis date in the CLD are excluded from the study. Because we evaluated the clustering results based on ESRD

survival outcome, patients with ESRD prior their SLE diagnosis date were also excluded from the study.

Cohort Classification: The SLICC Classification Criteria [10] includes 17 attributes, 11 clinical (acute cutaneous lupus, chronic cutaneous lupus, oral ulcers, nonscarring alopecia, synovitis, serositis, renal disorder, neurologic disorder, hemolytic anemia, leukopenia, thrombocytopenia) and 6 immunologic (antinuclear antibody (ANA), Anti-double stranded DNA (anti-dsDNA), anti-Smith antibody (anti-Sm), antiphospholipid antibody (APA), low complement, and Direct Coombs Test). Within the CLD, the presence or absence of these criteria were determined by clinician chart review. To determine presence or absence of each criterion in the NMEDW, we used diagnosis codes (ICD-9/10), procedure codes (ICD and CPT) and laboratory tests as appropriate. These algorithms are fully described in Walunas, et al (REF). Several clinical criteria were determined to have low detection sensitivity based on these structured data elements alone. Therefore, for renal disease, arthritis and oral ulcers criteria, we used a list of customized regular expression patterns designed to search for positive mentions of these three phenotypes in patient notes (93) in addition to ICD codes and laboratory tests. Patients were determined to have severe SLE complication (e.g. ESRD) based on diagnoses (ICD 9/10 codes) in the NMEDW. The project was reviewed and approved by the Northwestern University Institutional Review Boards.

Clustering analysis

We applied LCA to the NMEDW and CLD classification criteria datasets. For both datasets, we used basic demographics and SLICC criteria as features for clustering analysis. Race was self-reported from CLD data. Race was coded as Caucasian, African American (AA), Asian, and

others. Onset age was coded as young-onset (1-16), adult-onset (17 - 50), and late-onset (51-90). The Bayesian Information Criteria (BIC) was used to select from LCA-based models with a varying number of clusters (2 to 10). The model with the lowest BIC score was selected as the final model for clustering based off of each data set. We then compared differences in features among subgroups using chi-square tests, Fisher exact tests, or ANOVA based on the variable types.

To compare clinical relevance of the subgroups, we examined the relationship between the subgroups ESRD, a severe comorbidity of lupus, associated with poor outcomes, as determined by diagnosis documented in the NMEDW. We followed patients from their SLE onset date until the onset of ESRD or the end of the study date (2018-08-23) whichever came first. We plotted the Kaplan-Meier curve for each subgroup and used the log-rank test to test the significance.

To compare the consistency among clustering results, we calculated the concordance between the clustering results based on NMEDW data vs CLD data. Every cluster from NMEDW clustering was compared to its most similar cluster in the CLD data sets. If a patient i belongs to cluster j in NMEDW data and also belongs to cluster j's corresponding cluster in CLD data, then we consider it as a concordant pair. Membership concordance is calculated using the following equation:

$$\frac{\text{concordant pairs}}{\text{concordant pairs} + \text{disconcordant pairs}}$$

To explore if there are ways to improve the clustering result consistency between the EHR data and CLD data, we performed two sensitivity analyses. In the first sensitivity analysis, we

performed LCA only on a subset of criteria that have above 70% concordance across all patients. The concordance for each SLICC criterion i is calculated as the following:

$$\text{concordance(i)} = \left(N_{+,i} + N_{-,i}\right)/N,$$

Where N is the total number of patients, $N_{+,i}$ is number of patients that are positive for criteria i in both CLD and NMEDW, $N_{-,i}$ is the number of patients that are negative for criteria i in both CLD and EDW, N is the total number of patients.

In the second sensitivity analysis, we performed LCA on a subset of patients with patient level concordance above 65%. To calculate patient-level concordance, for the same patient, we compared all of his/her criteria labels in the CLD with his/her criteria labels in the NMEDW. We selected the subset of patients that having concordance above 65% and performed LCA. For example, if patient A has criteria label [+renal, +ANA, -APA, -antiSM] in the CLD and label [-renal, -ANA, +APA, -antiSM] (+ means positive for that criterion, - means negative for the criterion) in the NMEDW, the concordance for A would be 1/4 = 0.25, meaning 25% of the criteria identified for patient A are the same in both the CLD and NMEDW. A 65% threshold was used to ensure that patients in the analysis have relatively high concordance across two datasets, yet still leaves us relatively big sample size. More details of how the cutoff is selected is described in the result section.

We evaluated the clustering results of both sensitivity analyses by comparing the characteristics of subtypes and the Kaplan-Meier curves for ESRD of the subtypes. We assessed the consistency of clustering results from the CLD and the NMEDW data using concordance statistics and visualized them with t-distributed stochastic neighbor embedding (TSNE) plot (94).

Statistical software R (Version 3.6.3) were used for the data processes and statistical analysis. LCA was performed using the R poLCA package (95). Tsne plot was performed using the Rtsne package (94) .

Results

Among 856 SLE patients in the CLD with definite lupus as defined by the SLICC classification criteria, 472 were identified in the NMEDW and had at least 3 visits documented since 2002. After excluding patients who had documented ESRD before SLE diagnosis date, and patients who had missing data (3 patients did not have SLE diagnosis dates), 457 patients remained in the final cohort. The average age of SLE diagnosis is 29.7 (S.D. = 11.5), 93% are female, 51% are Caucasian, 28% are African American and 8% are Asian and 13% are from other racial groups.

Subtypes generated from the Complete CLD and NMEDW Datasets

LCA performed on the CLD-based dataset shows that when the number of clusters is 2, the BIC score is the lowest suggesting that 2 clusters give the best performance. As shown in **Error! Reference source not found.** and Figure 5, cluster 1C (C represents CLD data) has 289 patients consisting of 61% of the original population; Cluster 2C has 178 patients consisting of 39% of the original population. In general, cluster 1C has a higher percentage of young-onset patients (cluster 1C = 14% vs cluster 2C = 4%, p = 0), higher percentage of male patients (cluster 1C = 9% vs cluster 2C = 5%, p = 0.32), as well as higher percentage of non-Caucasian compared to cluster 2C (cluster 1C = 63% vs cluster 2C = 27%, p = 0). In terms of clinical and immunological

characteristics, cluster 1C has higher percentage in almost all immunological disorders including hemolytic anemia, leukopenia, thrombocytopenia, anti-ANA, anti-dsDNA, anti-Smith, anti-phospholipid, as well as a higher percentage in Coombs test, while cluster 2C has a higher percentage of acute rash, oral/nasal ulcer and arthritis. Cluster 1C also has higher percentage in alopecia. Chi-square/Fisher exact test shows that among clinical criteria, age of SLE diagnosis, race, acute cutaneous SLE, oral ulcer, alopecia, serositis, renal, hemolytic anemia, and leukopenia are significantly different between these two clusters. For immunological criteria, anti-Sm, anti-phospholipid antibody, low complement and coombs tests are significantly different between the two clusters.

LCA performed on the NMEDW-derived data shows that when the number of clusters is 3, the BIC score is the lowest suggesting that 3 clusters give the best performance. As shown in Table 5, cluster 1N (N represents NMEDW data) has 189 patients, consisting of 41% of the whole cohort; cluster 2N has 211 patients, consisting of 46% of the cohort; cluster 3N has 57 patients, consisting of 13% of the cohort. Cluster 2N has the highest percentage of late-onset SLE (8%), and Caucasian population (65%) compared to the other two clusters. Cluster 2N has low percentage of skin manifestations including acute rash, chronic rash, oral ulcer and alopecia. It also has a low percentage in every other criterion compared to cluster 1N and cluster 3N. Cluster 1N has the highest percentage in skin presentations (acute cutaneous, chronic cutaneous, ulcer, and alopecia) compared to both cluster 1 and cluster 3, while its percentage in other clinical and immunological criteria are in between that of cluster 1N and cluster 3N. Cluster 3N has low to medium percentage of skin manifestations including acute rash, chronic rash, and oral/nasal

ulcer. It has medium percentage of alopecia, while highest percentage in all other clinical and immunological criteria.

To assess the overlap between the clusters developed from the registry and EHR-based data sources, we created TSNE plots labelled to demonstrate the relationship between clusters generated from the CLD and the NMEDW. Figure 1a shows the clustering results using CLD data. The color represents clustering membership from CLD clustering: green represents patients in cluster 1C, purple represents patients in cluster 2C. The shape represents clustering membership in EDW clustering: '+' represents cluster 1N, 'circle' represents cluster 2N, and '▲' represents cluster 3N. We observe overlap between CLD cluster 1C (green symbols) and NMEDW cluster 1N while CLD cluster 2C (purple symbols) has more overlap with cluster 2N (circles). Similar observations can be drawn from assessing the data from the perspective of the NMEDW results as displayed in Figure 5b. Although patient cluster membership generated from these two datasets overlaps, there is also disagreement. Figure 5c shows that the concordance of the two clustering is 52.5%.

Sensitivity analysis: criteria with high concordance

To assess the impact of the concordance of individual variables between the gold standard data in the CLD and the real-world clinical care data in the NMEDW we assessed concordance between the CLD and NMEDW datasets for each classification criteria attribute. The results of our concordance analysis are shown in Table 6. Concordance ranged from 52%-94%. Notably, we observed higher concordance among immunologic criteria (range 75%-94%) than clinical

criteria (range 52%-88%). We selected all criteria with 70% concordance or higher (Coombs test, hemolytic anemia, thrombocytopenia, anti-dsDNA, leukopenia, anti-Sm, APA, arthritis, alopecia, low complement, neurologic, ANA) for a secondary cluster analysis. We used 70% cutoff because it ensures criteria with relatively high concordance yet leaves us a large enough feature set to differentiate our population.

LCA on this subset of criteria and basic demographics and subsequent BIC analysis identified the optimal number of clusters for both CLD and EDW data was 2. The characteristics of the clusters are shown in Table 7. Clusters generated from CLD and EDW have similar profiles. Both cluster 1 in the NMEDW and the CLD have more young onset patients, higher percentage of non-Caucasian, and higher percentage in almost every clinical and immunological disorders except for arthritis. In CLD data, arthritis has a lower percentage in cluster 1C, although its p value is not significant.

We used TSNE plots (Figure 6) to assess overlap between clusters from the CLD and NMEDW data sets. Figure 6a shows the clustering results using CLD data. The color represents clustering membership from CLD clustering: green represents patients in cluster 1C, purple represents patients in cluster 2C. The shape represents clustering membership in EDW clustering: '+' represents cluster 1N, 'circle' represents cluster 2N. We observe overlap between cluster 1C (green symbols) and cluster 1N (+) while cluster 2C (purple symbols) has more overlap with cluster 2N (circles). Similar observations can be drawn from assessing the data from the perspective of the NMEDW results as displayed in Figure 6b. Although patient cluster membership generated from these two datasets overlaps, there is also disagreement. Figure 6c shows that the concordance of the two clustering is 75.5%.

Sensitivity analysis: patients with high criteria concordance

To assess the impact of criteria concordance for a given patient between data derived from registry and real-world clinical data on our subpopulation analysis, we first examined the impact of the level of concordance on cohort size. The number of patients in the cohort after applying a range of individual concordance cutoffs is shown in Figure 7. As concordance increases, the cohort size available for performing the clustering analysis decreases. Given the diversity of phenotypic presentations in SLE, effective clustering requires a large sample size, thus we selected a patient criteria concordance of 65% which provides relatively high concordance between the datasets at the patient level while still retaining a large enough cohort (353) on which to perform the analysis.

The characteristics of the two subtypes are shown in Table 8. When subpopulations are identified based on CLD data, cluster 1C has more younger onset patients, male, non-Caucasian compared to cluster 2C. Cluster 1C also includes persons with a who are less likely to have acute cutaneous or chronic cutaneous lupus, and oral ulcers, but have a higher likelihood of an immunological criteria. Similar to CLD data, cluster 1N derived from the NMEDW data also has higher percentage of younger onset patients and higher percent of non-Caucasian. It has lower percentage of male, though the difference is not significant ($p = 0.12$). Compared to cluster 1C from the CLD, cluster 1N has higher percentage in skin manifestations including acute and chronic cutaneous lupus, oral ulcers and alopecia.

We used TSNE plots (Figure 8) to assess overlap between clusters from the CLD and NMEDW data sets. Figure 8a shows the clustering results using CLD data. The color represents clustering membership from CLD clustering: green represents patients in cluster 1C, purple represents patients in cluster 2C. The shape represents clustering membership in EDW clustering: '+' represents cluster 1N, 'circle' represents cluster 2N. We observe '+' spreads across both green color and purple color which shows low consistency between EDW clustering result and CLD cluster result. Similar observations can be drawn from assessing the data from the perspective of the NMEDW results as displayed in Figure 8b. Although patient cluster membership generated from these two datasets overlaps, there is also disagreement. Figure 8c shows that the concordance of the two clustering membership is 44.9%.

Kaplan Meier curves for compare survival outcome

Our previous assessments of subpopulations focused on statistical comparisons of cluster composition but did not explore the clinical relevance of the subpopulations derived from registry or real-world clinical data. To explore clinical relevance in more detail, Kaplan Meier comparing rate of occurrence of end stage renal disease, a serious and life-threatening outcome for people with SLE, were generated based on the subpopulations identified in each of our three clustering experiments. The results of the Kaplan-Meier analysis are shown in Figure 9.

When using the whole dataset, cluster 2C (higher prevalence in anti-dsDNA and renal disorder) generated from CLD data have a significantly worse outcome compared to cluster 1C ($p <$

0.0001). While in NMEDW clustering, cluster 1N and cluster 2N have similar ESRD outcomes, cluster 3N has a significant worse outcome. Cluster 1N has the highest percentage of skin manifestations, low-medium percentage of renal disorder and low-medium percentage of immunological disorder, while cluster 3N has the lowest prevalence in skin manifestations, and highest percentage in renal disorder and immunological disorder. In the past studies, it is reported that renal disorder and immunological disorders including anti-dsDNA are associated with worse lupus nephritis outcome (96). This is consistent with our study. When clustering on demographics plus criteria with concordance > 70%, cluster 1 (immunological manifestation prevalent cluster) has significantly worse outcomes compared to cluster 2 (acute cutaneous/chronic cutaneous/oral ulcer/arthritis prevalent cluster) in both NMEDW- and CLD-derived data. Finally, when assessing rate of ESRD development on clusters of individuals with concordance > 65%, cluster 1C has significantly worse outcome compared to cluster 2C , while no significant different is observed in NMEDW data.

Discussion

Complex diseases, such as systemic lupus erythematosus, are often recalcitrant to the identification of therapies that make significant impact on disease outcomes. It has been speculated that this may be due to disease heterogeneity and that identification of sub-populations of patients with more similar disease characteristics would provide a foundation for research into targeted therapies and support precision medicine approaches. Physician adjudicated disease-specific registries have long been the gold standard for sub-population development. However, development and maintenance of these registries is laborious and costly. Within the last 10 years there has been significant expansion of the use of electronic health records within the United States, providing an alternative source of information about patients that is collected in the context of their care. In this study, we set out to understand similarities and differences in subpopulations of people with lupus derived from either registry or EHR data for the same cohort of people using unsupervised machine learning strategies.

Using latent class analysis on the SLICC classification criteria as described in a registry and EHR-data for a cohort of people with SLE, we saw significantly different sub-populations. This discordance likely arises from the known data quality issues of EHR data such as missingness, misclassification and measurement errors. In our dataset, a relative high percentage of serositis (33%), oral ulcer (37%) and acute cutaneous (43%) recorded in the registry were not observed in the EHRs, which is consistent with missingness that results from certain clinical attributes not being included in diagnosis for billing or being regularly documented in notes. Interestingly, most classification criteria that were identified using laboratory results, such as the direct Coombs test,

hemolytic anemia, thrombocytopenia, anti-dsDNA, APA have high concordance between EHR and registry data.

To better understand the differences in our initial results we evaluated two possible approaches to improve sub-population identification consistency: 1) clustering on a subset of patients that have high criteria concordance between the two data sources, and 2) clustering on a subset of features having high concordance between the two data sources. In our results, clustering on high concordant features was able to improve the cluster consistency in patient membership (from 52.5% to 75.5%), cluster characteristics. Considering the fact that most laboratory tests have high concordance in our dataset, the strategy of relying more on laboratory features may be adaptable for other disease type in the future study if the purpose of clustering on EHR data is to understand the true clusters.

Finally, to better assess the clinical relevance of our sub-population analysis, we assessed the relationship of the subpopulations identified in our primary and sensitivity analyses to the development of end stage renal disease, a severe complication of SLE. While sub-populations identified using registry data consistently demonstrated segmentation where hallmarks of severe disease were associated with higher likelihood of ESRD, similar results were only seen in the EHR data with high concordance of criteria detection in EHR data.

Several previous studies have applied unsupervised learning to identify sub-groups of SLE **(97–101)**. These studies showed different number of sub-populations ranging from 2 to 5. There are also variations in patient race profile, unsupervised learning algorithms applied, and types of features used. Among the studies, To et al. and Lanata et al. used a similar feature set (based on the American College of Rheumatology 1997 classification criteria for SLE) as our study

(101,102). Both of their studies identified two clusters: a skin/arthritis manifestation dominated cluster (acute cutaneous, chronic cutaneous, oral ulcer, arthritis) and a renal disorder/ds-DNA dominated cluster, which is similar to what we observed using CLD data. Nevertheless, this similar pattern is also presented in the EHR data. We identified 3 clusters when performing LCA on the whole dataset using EHR data. Cluster 1N has high prevalence in skin disorder (acute cutaneous, chronic cutaneous) and low-medium prevalence in immunological disorder and renal disorder, while cluster 3N has low-medium prevalence in skin disorder and high prevalence in immunological disorder as well as renal disorder. This shows that strong pattern may still emerge regardless of the noise from the data source (103). However, researchers should be extremely careful when assigning new subjects to subgroups based on algorithms developed using EHR data considering the membership inconsistency between the two datasets, which has a stronger emphasis on clinical attributes.

Our study has several limitations. Firstly, we chose SLE as a study case to explore clustering results from EHR and CLD data. Although SLE has a spectrum of manifestations and it's an ideal disease for clustering analysis, the results from SLE may not generalize to other disease type. Secondly, a big part of EHR information is stored in clinical notes. In our study, we incorporated regular expression to capture more information on features (oral/nasal ulcer, arthritis, renal disorder) that are expected to have relatively high missingness from structured data. We were not able to perform more advanced NLP (natural language processing) methods or apply them to the other features due to limited resource. We expect that other features will benefit from regular expression or more advanced NLP methods as well. Thirdly, our proposed technique of focusing on a subset of

features that have high concordance is at the cost of feature size reduction. If certain features are important to a study but also have low concordance, alternative approaches should be considered.

In conclusion, our study shows that both NMEDW and CLD data generate distinct clusters with significant relationships to ESRD. However, despite using the same patient cohort, subpopulations identified from EHR data are different from those derived from registry data. By performing LCA on a subset of criteria that have concordance above 70%, we are able to reduce the differences including improving the cluster consistency in patient membership, cluster characteristics and ESRD outcome. Our proposed method of using features with relative high concordance across data sources can potentially improve the algorithm generalizability and help researchers to generate more reliable clusters when using noisy data source such as EHRs.

Table 5. General study cohort and cluster characteristics using CLD data vs NMEDW data for latent class analysis.

	CLD data				NMEDW data				
Features	**General**	**Clust**	**Cluster**	**Pv**	**Genera**	**Clust**	**Cluster**	**Clust**	**Pva**
	N=457	N=280	N=		N=4	N=188(	N=207(	N=	
Age of									
<17	46(10%)	7(4%)	39(1	0	46(10	28(15%)	12(6%)	6(9%	0
17-50	389(85%	159(90	230(		389(8	156(83	178(86	55(89	
>50	22(5%)	11(6%)	11(4		22(5	4(2%)	17(8%)	1(2%	
Sex									
Male	34(7%)	9(5%)	25(9	0.32	33(7	8(4%)	10(5%)	16(25	0
Female	423(93%	168(95	255(		424(9	180(96	197(95	47(75	
Race									
Caucasian	233(51%	129(73	104(	0	233(5	77(41%)	135(65	21(34	
African	127(28%	27(15%)	101(		129(2	70(37%)	39(19%	20(32	
Asian	37(8%)	9(5%)	28(1		38(10	13(7%)	17(8%)	8(13	
Other races	60(13%	12(7%)	48(1		58(13	28(15%)	17(8%)	13(21	
Clinical									
Acute	395(86%	177(100	218(	0	227(4	184(98	29(14%	14(22	0
Chronic	114(25%	44(25%)	70(2	0.88	185(4	177(94	6(3%)	2(3%	0
Oral/Nasal	260(57%	126(71	134(	0	113(5	51(27%)	50(24%	12(20	0.59
Alopecia	95(21%)	25(14%)	70(2	0	43(10	32(17%)	6(3%)	5(8%	0
Arthritis*	430(94%	170(96	260(	0.1	372(7	158(84	157(76	56(91	0.05
Serositis	197(44%	65(37%)	132(	0.05	75(17	38(20%)	19(9%)	19(30	0
Renal*	172(38%	21(12%)	151(	<2.2	238(4	115(61	70(34%	53(85	0
Neurologic	100(22%	30(17%)	70(2	0.06	99(22	53(28%)	29(14%	17(28	0
Hemolytic	42(9%)	0(0%)	42(1	0	23(5	9(5%)	2(1%)	11(18	0
Leukopeni	417(91%	149(84	269(	0	359(7	156(83	143(69	60(97	0
Thrombocy	56(12%)	14(8%)	42(1	0.12	86(19	36(19%)	25(12%	25(41	0
Immunolo									
ANA	393(86%	152(86	241(	0.62	381(8	180(96	139(67	62(10	0
Anti-	320(70%	62(35%)	258(	<2.2	326(7	150(80	118(57	58(93	0
Anti-Sm	102(23%	12(7%)	90(3	0	83(18	47(25%)	10(5%)	25(41	0
APA	124(27%	28(16%)	95(3	0	128(2	66(35%)	25(12%	38(61	0
Low	315(69%	74(42%)	241(	<2.2	339(7	162(86	118(57	59(95	0
Coombs	13(3%)	2(1%)	11(4	0.03	17(4	9(5%)	0(0%)	7(12	0

*Used both regular expression on clinical notes and ICD to identify criterion

Table 6. Concordance table for each SLICC criteria between EDW data and CLD data.

SLICC Criteria	Concordance	Positive % only in	Positive % only in
Coombs Test	94% (430)	2% (9)	3% (14)
Hemolytic	88% (402)	8% (37)	4% (18)
Thrombocytopenia	86% (393)	4% (18)	10% (46)
Anti-dsDNA	84% (384)	7% (32)	9% (41)
Leukopenia	82% (375)	15% (69)	3% (14)
Anti-Sm	81% (370)	12% (55)	7% (32)
APA	80% (366)	10% (46)	10% (46)
Arthritis*	78% (356)	17% (78)	4% (18)
Alopecia	76% (347)	18% (82)	6% (27)
Low Complement	76% (347)	9% (41)	14% (64)
Neurologic	75% (343)	12% (55)	12% (55)
ANA	75% (343)	14% (64)	12% (55)
Renal*	73% (334)	7% (32)	20% (91)
Chronic	63% (288)	11% (50)	26% (119)
Serositis	60% (274)	33% (151)	6% (27)
Oral/Nasal	58% (265)	37% (169)	5% (23)
Acute Cutaneous	52% (238)	43% (197)	5% (23)

*Used both regular expression on clinical notes and ICD to identify criterion

Table 7. Cluster characteristics using CLD data vs EDW data on a subset of criteria having concordance > 70%.

	CLD				EDW			
	General	**Cluster 1**	**Cluster 2**	**Pval**	**General**	**Cluster 1**	**Cluster 2**	**Pval**
	N= 457	**N = 274**	**N = 183**		**N= 457**	**N = 238**	**N = 219**	
Age of								
<17	48(11%)	41(15%)	7(4%)	0.00	47(10%)	37(16%)	10(4%)	0.00
17-50	388(85%	225(82%	163(89%		390(85%	198(83%	193(88%	
>50	21(5%)	8(3%)	13(7%)		21(5%)	4(2%)	17(8%)	
SEX								
Male	34(7%)	19(7%)	15(8%)	0.44	32(7%)	21(9%)	11(5%)	0.11
Female	423(93%	255(93%	168(92%		425(93%	217(91%	208(95%	
Race								
Caucasia	232(51%	104(38%	128(70%	0.00	232(51%	83(35%)	149(68%	0.00
AA	129(28%	96(35%)	33(18%)		128(28%	93(39%)	35(16%)	
Asian	39(9%)	30(11%)	9(5%)		38(8%)	23(10%)	15(7%)	
Other	60(13%)	47(17%)	13(7%)		59(13%)	40(17%)	19(9%)	
Clinical								
Alopecia	95(21%)	69(25%)	26(14%)	0.01	44(10%)	36(15%)	9(4%)	0.00
Hemolyti	44(10%)	44(16%)	0(0%)	0.00	24(5%)	24(10%)	0(0%)	0.00
Leukope	418(91%	266(97%	152(83%	0.00	359(78%	212(89%	147(67%	0.00
Arthritis*	431(94%	255(93%	176(96%	0.09	369(81%	207(87%	162(74%	0.00
Renal*	175(38%	148(54%	27(15%)	< 2.2e-16	236(52%	164(69%	72(33%)	< 2.2e-16
Neurolog	100(22%	69(25%)	31(17%)	0.01	99(22%)	60(25%)	39(18%)	0.27
Thrombo	56(12%)	41(15%)	15(8%)	0.01	86(19%)	67(28%)	20(9%)	0.00
Immunol						0(0%)	0(0%)	
ANA	392(86%	233(85%	159(87%	0.26	382(83%	226(95%	155(71%	0.00
Anti-	316(69%	252(92%	64(35%)	< 2.2e-16	324(71%	214(90%	110(50%	< 2.2e-16
Anti-Sm	103(23%	90(33%)	13(7%)	0.00	83(18%)	74(31%)	9(4%)	0.00
APA	125(27%	96(35%)	29(16%)	0.00	129(28%	112(47%	18(8%)	< 2.2e-16
Low	316(69%	241(88%	75(41%)	< 2.2e-16	337(74%	219(92%	118(54%	< 2.2e-
Coombs	11(2%)	11(4%)	0(0%)	0.01139	17(4%)	17(7%)	0(0%)	0.00

*Used both regular expression on clinical notes and ICD to identify criterion

Table 8. Cluster characteristics using CLD data vs EDW data on a subset of individuals with concordance across criteria > 65%

	CLD				EDW			
	General N=353	Cluster 1 N = 213	Cluster 2 N = 140 (40%)	Pval	General N = 353	Cluster 1 N = 160	Cluster 2 N = 193	Pval
Age of SLE								
<17	33(9%)	28(13%)	5(4%)	0.01	33(9%)	21(13%)	12(6%)	0
17-50	301(85%	173(81%	128(91%)		301(85%	136(85%	165(86%	
>50	19(5%)	12(5%)	7(5%)		19(5%)	3(2%)	16(8%)	
SEX								
Male	30(9%)	24(11%)	6(4%)	0.03	30(8%)	9(6%)	21(11%)	0.12
Female	324(92%	190(89%	134(96%)		322(91%	150(94%	172(89%	
RACE								
Caucasian	183(52%	82(39%)	101(72%)	0	183(52%	66(41%)	117(61%	0
African American	103(29%	83(39%)	20(15%)		103(29%	60(37%)	43(22%)	
Asian	26(7%)	17(8%)	9(7%)		26(7%)	10(6%)	16(8%)	
Other races	41(12%)	32(15%)	9(7%)		41(12%)	24(15%)	17(9%)	
Clinical criteria								
Acute Cutaneous	301(85%	161(76%	140(100%	0	181(51%	154(96%	27(14%)	<2.2e
Chronic	90(25%)	51(24%)	39(28%)	0.28	142(40%	140(87%	2(1%)	<2.2e
Oral/Nasal	189(54%	91(43%)	98(70%)	0	94(27%)	45(28%)	49(26%)	0.6
Alopecia	59(17%)	46(22%)	13(9%)	0	32(9%)	26(17%)	6(3%)	0
Arthritis*	337(95%	202(95%	135(96%)	0.63	309(88%	142(89%	167(86%	0.67
Serositis	132(37%	87(41%)	45(32%)	0.05	66(19%)	38(24%)	28(15%)	0.02
Renal*	133(38%	113(53%	20(14%)	0	178(50%	94(59%)	84(44%)	0
Neurologic	70(20%)	55(26%)	15(11%)	0	71(20%)	41(25%)	30(16%)	0.02
Hemolytic	23(7%)	23(11%)	0(0%)	0	9(3%)	3(2%)	6(3%)	0.71
Leukopenia	322(91%	206(97%	116(83%)	0	285(81%	138(86%	147(76%	0.03
Thrombocytopeni	40(11%)	30(14%)	10(7%)	0.22	56(16%)	28(17%)	28(15%)	0.34
Immunological								
ANA	308(87%	187(88%	121(86%)	0.9	309(88%	156(97%	153(79%	0
Anti-dsDNA	247(70%	197(92%	51(36%)	<2.2e	259(73%	126(79%	133(69%	0.03
Anti-Sm	70(20%)	65(30%)	5(4%)	0	53(15%)	37(23%)	16(9%)	0
APA	95(27%)	75(35%)	20(14%)	0	95(27%)	58(36%)	37(19%)	0
Low Complement	243(69%	181(85%	62(44%)	<2.2e	265(75%	137(86%	128(66%	0
Coombs Test	7(2%)	7(3%)	0(0%)	0.07	6(2%)	3(2%)	3(1%)	1

*Used both regular expression on clinical notes and ICD to identify criterion

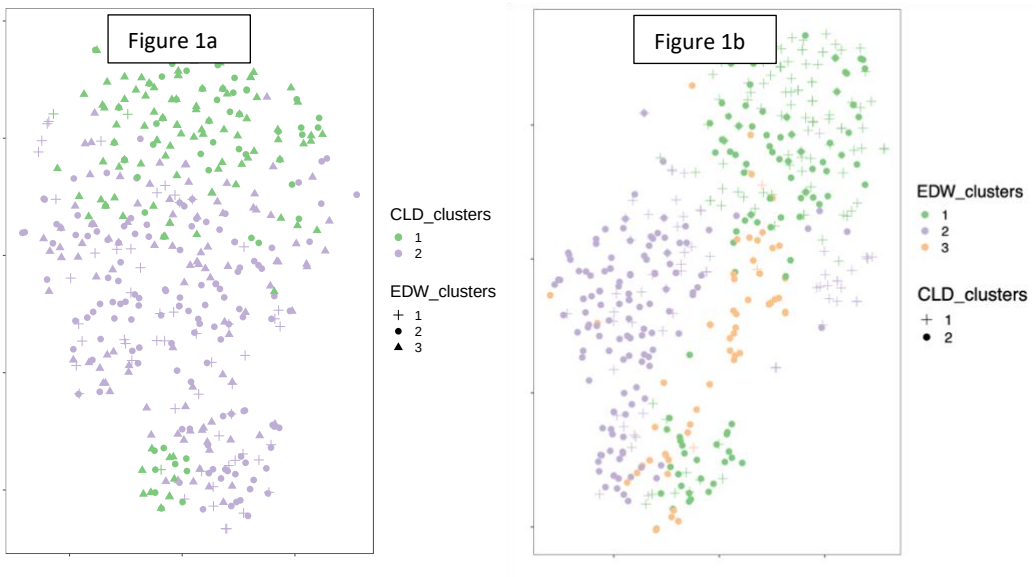


Figure 1c	EDW: Cluster 1	EDW: Cluster 2	EDW: Cluster 3
CLD: Cluster 1	127	98	54
CLD: Cluster 2	62	113	3

Figure 5. Consistency comparison between CLD clustering results vs EDW clustering results; 1a: TSNE visualization for clustering on CLD data; 1b: TSNE visualization for clustering on EDW data; 1c: concordance table for patient membership from EDW clustering vs CLD clustering.

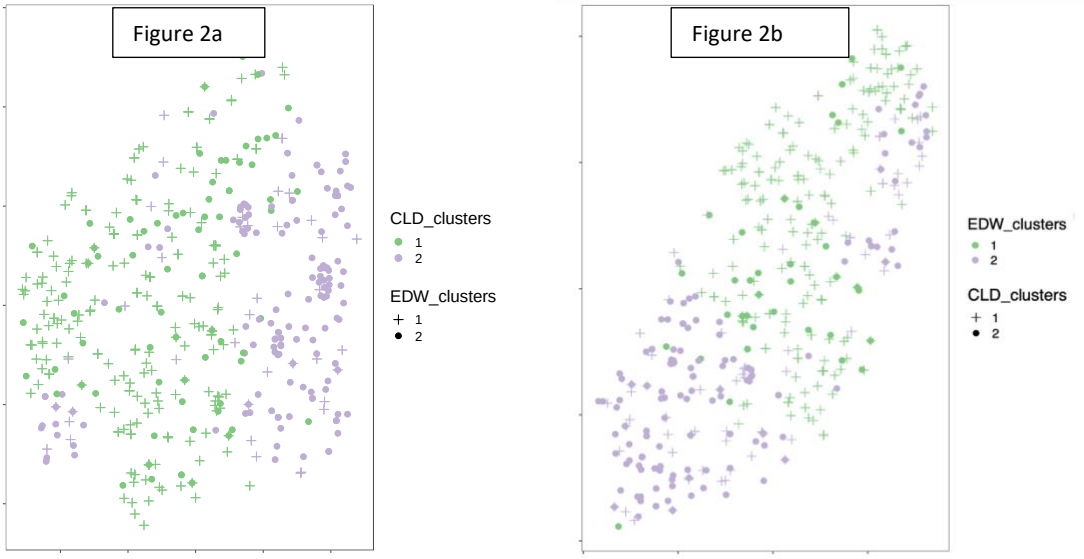


Figure 2c	EDW: Cluster 1	EDW: Cluster
CLD: Cluster 1	204	73
CLD: Cluster 2	39	141

Figure 6. TSNE plot for CLD clustering results vs EDW clustering results on the subset of criteria with concordance>70%; Left plot: clustering on CLD data; Right plot: clustering on EDW data

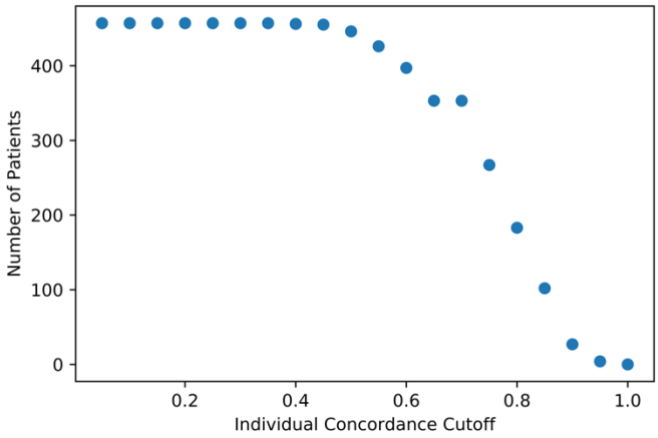


Figure 7. number of patients left using individual concordance cutoff from 0%-100%.

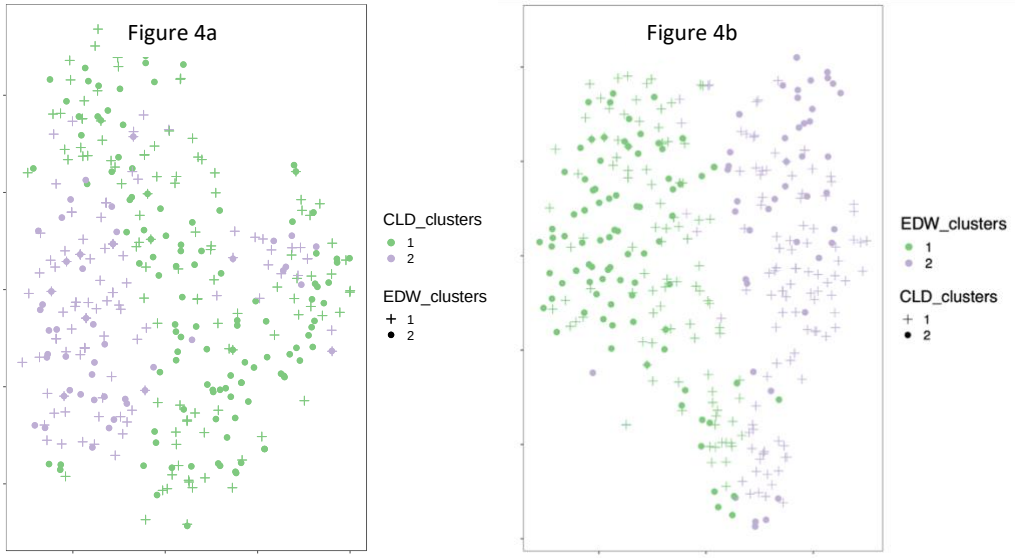


Figure 4c	EDW: Cluster1	EDW: Cluster2
CLD: Cluster1	111	100
CLD: Cluster2	48	94

Figure 8. TSNE plot for CLD clustering results vs EDW clustering results on individuals with concordance > 65%; Left plot: clustering on CLD data; Right plot: clustering on EDW data; bottom: membership concordance table

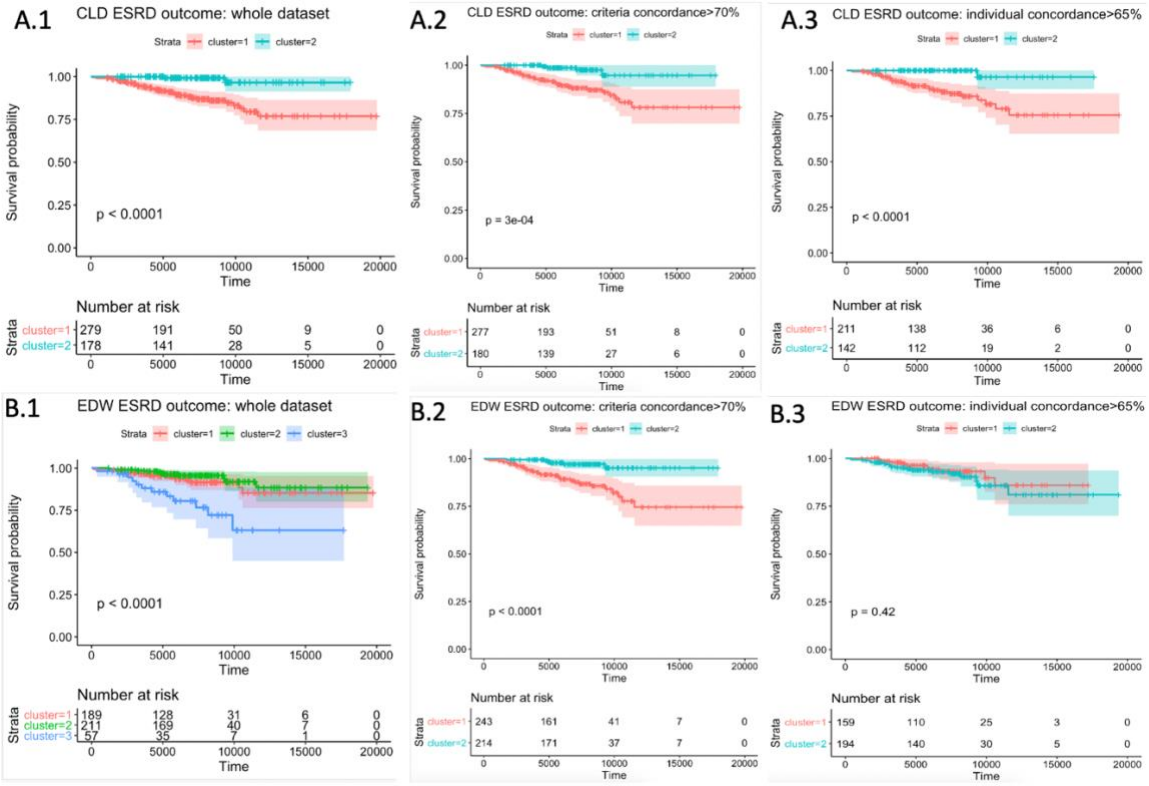


Figure 9. ESRD Kaplan Meier curve grouped by different clustering groups. A.1, B.1: cluster on whole dataset; A.2, B.2: cluster on criteria that have concordance > 70%; A.3, B.3: cluster on patients that have concordance>65%

Chapter Four: Natural language processing for lupus nephritis computational phenotyping

Abstract

SLE (systemic lupus erythematosus) is a rare autoimmune disorder with a repeated relapsing-remitting course and diverse manifestations. Lupus nephritis is one of the major risk factors for severe outcomes and mortality and is a key component for modern lupus classification criteria. Thus, accurately identifying lupus nephritis in electronic health records (EHRs) would benefit large cohort clinical studies and clinical trials where characterization of the patient population is critical for study design and analysis. While lupus nephritis can be identified through procedures billed for in EHR data, a large number of information related to lupus nephritis such as kidney biopsy are generally present in histology reports and prior medical history narratives which require sophisticated text processing to mine information out of clinical notes. In this study, we developed algorithms to identify lupus nephritis with and without natural language processing (NLP) using EHR data from the Northwestern Medicine Enterprise Data Warehouse (NMEDW). We developed four algorithms: a rule-based algorithm using only structured data only and three algorithms using different NLP models. These algorithms were validated on a dataset from Vanderbilt University Medical Center. Our best performing NLP model improved F measure in both the NMEDW dataset (0.41 vs 0.79) and the Vanderbilt dataset (0.62 vs 0.96) compared to the baseline lupus nephritis algorithm.

Introduction

Systemic Lupus Erythematosus (SLE) is an autoimmune disease that has diverse manifestations, resulting in significant morbidity and mortality (84,104). While many autoimmune diseases, such as rheumatoid arthritis, have benefitted from new classes of medications, SLE has seen few advancements in therapy in the last 50 years (105). It has been hypothesized that the heterogeneity of SLE presentations may make it challenging to understand therapeutic responses across the full scope of SLE presentations and that research studies and clinical trials would benefit from targeting subpopulations with similar disease presentations (106). Classification criteria for SLE describe a broad range of evidence-based clinical and laboratory descriptors for SLE. There are three criteria currently in use: 1) the set developed in 1983 and revised in 1997 by the American College of Rheumatology (ACR) (107) , 2) the set developed by the System Lupus International Collaborating Clinics in 2012 (SLICC) (84), and 3) the newly established European League Against Rheumatism / American College of Rheumatology (EULAR/ACR) criteria set (108). Lupus nephritis is one of the most common and severe sequelae of SLE: approximately 40% SLE patients develop lupus nephritis(109), and it is represented in all three classification criteria. Both the SLICC and EULAR/ACR criteria define “definite lupus” as having a positive anti-nuclear antibody/anti-dsDNA screen in the presence of renal biopsy proven lupus nephritis (84,108). Thus, it is a critical attribute to describe for clinical and research applications and the development of SLE subpopulations, but often it requires time consuming chart adjudication to identify patients who satisfy this criterion. Electronic health records (EHRs) are a readily available data source for describing persons with SLE that includes a record of clinical care and procedures performed, diagnoses, laboratory test result values,

medication orders, and clinical notes. However, the EHRs are primarily designed for administrative and clinical purposes; thus, data can be biased and incomplete (39), and important data for describing SLE, such as histology notes for kidney biopsies, is generally only located in text based notes that are challenging to extract information from using simple rules-based identification algorithms and text string searches. Several prior studies developed algorithms to identify lupus nephritis using administrative or claim data (110). Chibnik et al. identified lupus nephritis in claim data and reached a positive predictive value (PPV) of 88% but sensitivity and specificity were not mentioned (111). Li et al. used various combinations of International Classification of Diseases (ICD) codes to identify lupus nephritis (112). The algorithm achieved good sensitivity and specificity but a low PPV of 63.4%. Most of these studies only used structured data (i.e. ICD codes, laboratory test value), and the algorithms were often not validated in an external dataset (111,112).Thus, correctly identifying lupus nephritis from EHRs for large cohort studies, in addition to identifying critical procedures, diagnoses and lab results, also requires the development of natural language processing (NLP) tools that can read histology reports and clinical notes, and previous studies with other lab-based concepts have demonstrated that NLP can significantly improve rate of identification (93). In this study, we compared algorithms for the identification of lupus nephritis based on structured data alone with those that included three different NLP models to determine whether NLP could improve identification of lupus nephritis.

We trained and evaluated the performance of all four algorithms in a dataset from Northwestern Medicine Electronic Data Warehouse (NMEDW) and then further validated the performance in an external dataset from Vanderbilt University (VU) Medical Center.

Methods

Data Source

The Chicago Lupus Database (CLD), established in 1991, is a physician validated registry of 1,052 patients with possible or definite lupus according to the revised 1982 American College of Rheumatology classification criteria [4][5].The patients in the CLD were chart adjudicated for the ACR 1997 classification criteria and SLICC criteria . Among the 1052 patients, 878 patients had definite lupus according to the Systemic Lupus International Collaborating Clinics (SLICC) classification criteria (84). Among these patients, 178 have lupus nephritis. The presence or absence of lupus nephritis in patients in the CLD is verified by the physician chart review.

The Northwestern Medicine Electronic Data Warehouse (NMEDW) is the primary data repository for all the medical records of patients who receive care within the Northwestern Medicine system (113). Established in 2007, the NMEDW contains records for over 3.8 million patients, with most EHR data going back to at least 2002, and with some billing claims data going back to 1998 or even earlier. By linking patients in the CLD to patient records in the NMEDW through their Medical Record Numbers, we identified 818 definite SLE patients who were both in the CLD and the NMEDW. To ensure our patient cohort has sufficient depth of data in both data sources, we excluded any patients who had less than four clinical encounters documented in the NMEDW, reducing the final case cohort size to 472. Our SLE study population selection process is described in Figure 10. All inpatient and outpatient notes from transplant, nephrology, and rheumatology departments were retrieved without any provider type

restriction. The retrieved clinical narratives included pathology reports, progress notes, consult notes, and discharge notes.

Algorithm development: lupus nephritis phenotype

Lupus nephritis is defined as "having a urine protein/creatinine ratio (or 24-hour urine protein collection) equivalent to 500 mg of protein per 24 hour period, or red blood cell casts in the urine" based on the SLICC classification criteria (84). We developed 4 algorithms (see Table 9 for the details of the four algorithms) to identify lupus nephritis in the EHR data of SLE patients including a baseline algorithm that included only structured data from the EHRs and three NLP models that used structured data and clinical notes. In the baseline algorithm, a patient is predicted as lupus nephritis based on ICD9/10 diagnosis codes and laboratory test results. For the NLP models, we implemented an L2-regularized logistic regression classifier. We extracted concept unique identifier (CUI) features and regular expression features from the notes. For the CUI features, we first preprocessed the notes by removing duplicated records and tokenizing sentences. We then applied MetaMap to annotate medical concepts in each sentence (114). MetaMap is an NLP application that maps biomedical text to the Unified Medical Language System (UMLS) Metathesaurus (115) . It assigns a concept unique identifier to each word or term. CUIs that were tagged as negation by NegEx in MetaMap were excluded. For the regular expression (regex) pattern components, we used regex to search for text related to nephritis class II, nephritis class III, nephritis class IV, nephritis class V, and proteinuria (see Table S 12 for the details of regex patterns). For the NLP models, we explored three sets of features. In the first NLP model, the full MetaMap (binary) model, all MetaMap CUIs were used as binary type features. In the second NLP model, the full MetaMap (count) model, the number of occurrences

for every mapped CUIs were used as features. In the third NLP model, the MetaMap mixed model, we used a mixture of lupus nephritis related CUIs, structured data, and regex patterns as features. The CUIs include C0024143, C0268757, C0268758, C4053955, C4053958, C4053959, C4054543 (see Table S 13 for each CUI definition). There were 13 variables in total for the MetaMap mixed model including 7 features from CUIs, 5 features from regex patterns, and 1 feature from structured data.

Model training and evaluation

We split the data from NMEDW into training and testing datasets with a size ratio of 3:1. In the training dataset, we used 5-fold cross-validation to find the optimal solver and the L2 ratio. Parameter C was selected by a grid search with C ranging from 1e-5 to 1e5 with interval spacing equal to 10. We selected sag method as our optimizer to find the best parameters (116). We set the class weight as balanced to adjust for disproportionate class frequencies. Parameters that generated the best accuracy were retained. We evaluated our model in the testing set based on sensitivity, specificity, PPV, and negative predicted value (NPV). L2-regularized logistic regression was conducted using 'scikit-learn' library in Python, version 3.7.3. Regular expression was performed using 're' package in Python, version 3.7.3.

External validation

We further validated both the baseline algorithm, which included only structured data elements, and the best performing NLP model (based on results from the NU site), in an external validation dataset at Vanderbilt University. The Vanderbilt University Medical Center is a regional, tertiary care center (117). The VU data warehouse contains over 2.5 million subjects with de-identified clinical records from the EHRs collected across the past several decades. We first did a simple

SLE phenotyping algorithm based on SLE ICD9/10 codes to get a cohort on which to run our lupus nephritis algorithm. We then randomly selected 50 patients on which to evaluate our lupus nephritis algorithm. A rheumatologist manually reviewed the chart for these 50 patients. There were 16 patients with definite lupus, 1 with possible SLE, and 33 with no SLE. We evaluated the sensitivity, specificity, PPV, and NPV for the lupus nephritis baseline algorithm, and the lupus nephritis NLP model with the highest F measure based on the results from the NU dataset.

Results

Among the 472 SLE patients, there are 178 patients who developed lupus nephritis, consisting of 37.7% of the cohort. The average number of notes is 68.58 (SD = 59.37). The distribution of number of notes for the patient cohort is shown in Figure 10.

The performance for the four algorithms is shown in Table 10. All three NLP models have higher sensitivity, specificity, PPV, and NPV compared to the baseline algorithm with structured data alone. The full MetaMap (binary) model is better in sensitivity (0.63 vs 0.6), NPV (0.81 vs 0.8), and F measure (0.72 vs 0.41) compared to the full MetaMap (count) model. The MetaMap mixed model has higher sensitivity (0.74) and NPV (0.86) as well as F measure (0.79) compared to the other two models. Therefore, we selected the MetaMap mixed model as the final model to be validated at VU. In the VU dataset, which included 50 patients with SLE, the MetaMap mixed model has higher sensitivity, specificity, PPV, and NPV compared to the baseline algorithm. The F measure improved from 0.79 to 0.96 as shown in Table 10.

Discussion

In this study, we developed four algorithms to identify the lupus nephritis, a baseline algorithm using structured data only, a full MetaMap model with binary features, a full MetaMap model with count features, and a filtered MetaMap/regex/ICD mixed model. In the NU dataset, the MetaMap mixed model outperformed (F measures = 0.79) both baseline algorithm (0.41) and the other two NLP models. In the VU validation dataset, the MetaMap mixed model also improved the F measure (0.96) greatly compared to the baseline algorithm (0.62).

Error analysis

In the MetaMap mixed model, we investigated 10 SLE patients in the training set that were wrongly classified by L2-regularized logistic regression. One patient was wrongly predicted as negative for lupus nephritis with a 0.49 probability of having lupus nephritis. In the feature set the algorithm identified, the patient was positive for CUI C002413 (glomerulonephritis in the context of systemic lupus erythematosus) and was negative for all the other features. It was mentioned in the notes that the patient had ‘stage 2 LN’. Lupus nephritis class II is one of the features used in our algorithm. However, our regex did not include this specific variation of wording for lupus nephritis class II. This pattern could be incorporated in the NLP in the future to improve algorithm performance.

In another example, a 26-year-old female was wrongly predicted as positive for lupus nephritis with a probability of 0.53 of having lupus nephritis. In the feature set the algorithm identified, the patient was positive for C0024143 and proteinuria features both of which were positively associated with lupus nephritis. Our algorithm showed that patient had match for ‘proteinuria>0.5’ in the notes which was in the context of ‘negative renal disorder: either

persistent proteinuria (>0.5g/day or +++) or cellular casts'. Our regex pattern was not able to capture the negation at the beginning of the sentence. Therefore, it falsely predicted the patient as positive for lupus nephritis.

All NLP models outperformed the baseline algorithm in the NU dataset. The baseline algorithm relies solely on ICD 9/10 diagnosis and laboratory test results. In the baseline rule-based algorithm, laboratory tests missing from the EHRs greatly influenced the performance. In the NLP mixed model, by utilizing features from multiple data modalities (regex, CUI, and laboratory tests) with a penalized logistic regression model, the risk of having mislabeled data is shared among several features which may improve the generalizability of the model.

Limitations

Our study has certain limitations. We only had 50 patients in the VU validation dataset. This is due to limited resources for chart review. The small sample size may increase the chance of sample bias, which might explain the big improvement of F-measure in the external validation dataset. In addition, our algorithms were developed based on the SLE population, which significantly increases the prevalence of lupus nephritis. Our algorithms might not generalize well in a broader hospital population.

Conclusion

In conclusion, we developed four algorithms, a structured data only algorithm and three NLP models, to identify lupus nephritis phenotypes. We evaluated the algorithms in an internal and an external validation dataset. The three NLP models outperformed the baseline algorithm in both the NU dataset and the VU dataset. Our NLP algorithms can serve as powerful tools to accurately identify lupus nephritis phenotype in EHRs for clinical research.

Table 9. Algorithm description

Baseline algorithm	Rule-based	A patient is confirmed to have lupus nephritis if he/she has proteinuria>0.5mg in laboratory test or has ICD 9/10 diagnosis code for lupus nephritis.
Full MetaMap model (binary)	L2-regularized logistic regression	Features are all the non-negative mention of MetaMap CUIs. We treated CUIs as binary variables and fitted L2-regularized logistic regression to predict lupus nephritis.
Full MetaMap model (count)	L2-regularized logistic regression	The same as the full MetaMap model (binary) except that MetaMap CUIs are treated as numeric variables representing the count of instances each concept is mentioned in the clinical text.
MetaMap/regex/ICD mixed model	L2-regularized logistic regression	There are 13 features in this model including 7 CUI features, 5 RegEx features, and 1 feature from structured data.

Table 10. Model performance

Dataset	Algorithm	Sensitivity	Specificity	PPV	NPV	F Measure
NU	Baseline	0.43	0.6	0.39	0.64	0.41
NU	Full MetaMap (binary)	0.63	0.93	0.85	0.81	0.72
NU	Full MetaMap (counts)	0.6	0.95	0.88	0.8	0.71
NU	MetaMap mixed	0.74	0.92	0.84	0.86	0.79
VU	baseline	0.92	0.61	0.46	0.96	0.62
VU	MetaMap mixed	1	0.97	0.93	1	0.96

Abbreviations: SLE, systemic lupus erythematosus; NU, Northwestern University; VU, Vanderbilt university; NLP: natural language processing; PPV, positive predictive value; NPV, negative predicted value.

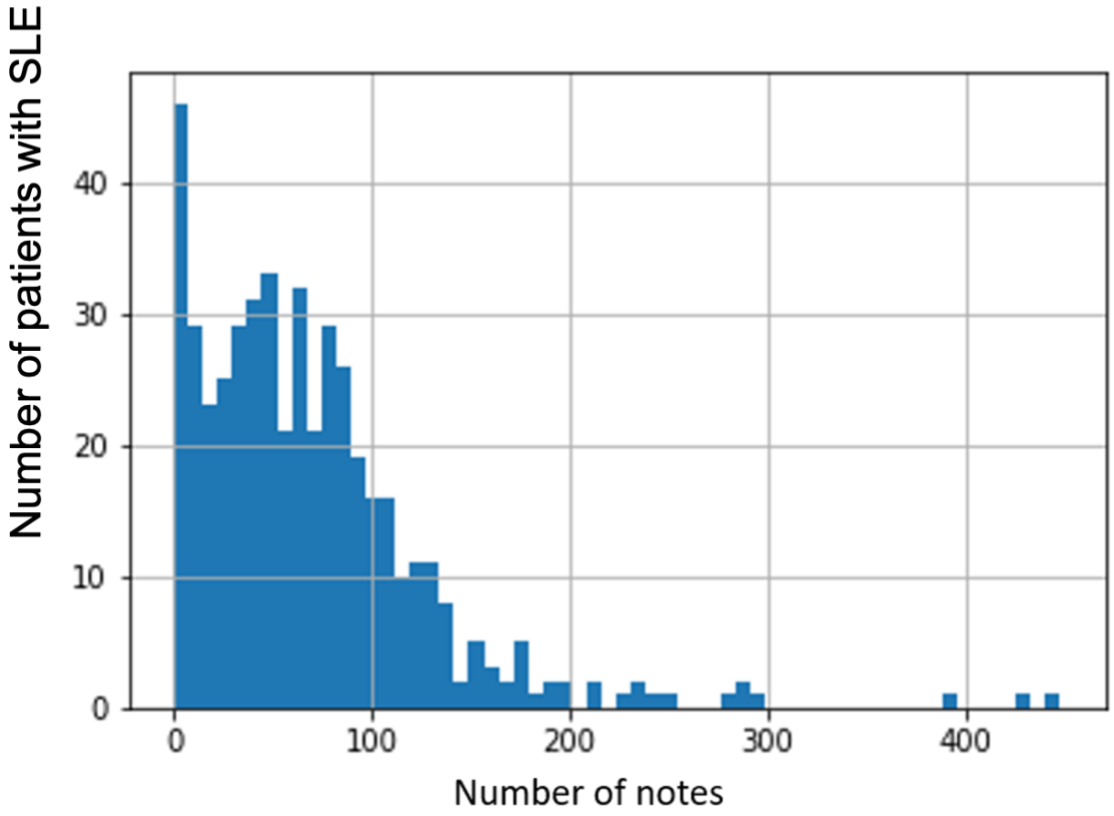


Figure 10. Number of patients with SLE.

Chapter Five: Deep Neural Network Survival Model for Cardiovascular Disease Risk Prediction

Abstract

Background: The Pooled Cohort Equations (PCEs) are race- and sex-specific Cox PH-based models used for 10-year atherosclerotic cardiovascular disease (ASCVD) risk prediction with acceptable discrimination. In recent years, neural network models have gained increasing popularity with their success in image recognition and text classification. Various survival neural network models have been proposed by combining survival analysis and neural network architecture to take advantage of the strengths from both. However, the performance of these survival neural network models compared to each other and to PCEs in ASCVD prediction is unknown.

Methods: In this study, we used 6 cohorts from the Lifetime Risk Pooling Project and compared the performance of the PCEs in 10-year ASCVD risk prediction with an all two-way interactions Cox PH model (Cox PH-TWI) and three state-of-the-art neural network survival models including Nnet-survival, Deepsurv, and Cox-nnet. For all the models, we used the same 7 covariates as used in the PCEs. We fitted each of the aforementioned models in white females, white males, black females, and black males, respectively. We evaluated models' internal and external discrimination power and calibration.

Results: The training/internal validation sample comprised 23216 individuals. The average age at baseline was 57.8 years old (SD = 9.6); 16% developed ASCVD during average follow-up of 10.50 (SD = 3.02) years. Based on 10x10 cross-validation, the method that had the highest C-statistics was Deepsurv (0.7371) for white males, Deepsurv and Cox PH-TWI (0.7972) for white females, PCE (0.6981) for black males, and Deepsurv (0.7886) for black females. In the external validation dataset, Deepsurv (0.7032), Cox-nnet (0.7282), PCE (0.6811), and Deepsurv (0.7316) had the highest C-statistics for white male,

white female, black male, and black female population, respectively. Calibration plots showed that in 10x10 validation, PCE had good calibration in white male and black male but was outperformed by neural network models in white female and black female. In external validation, all models overestimated the risk for 10-year ASCVD.

Conclusions

We demonstrated the use of the state-of-the-art neural network survival models in ASCVD risk prediction. Neural network survival models and PCEs have generally comparable discrimination and calibration.

Background

Cox Proportional Hazards (Cox PH) model is widely used to quantify the effect of covariates in relation to time-to-event outcomes or to predict the survival time for a new individual (118). Cox PH is a semi-parametric model, which consists of two main components: baseline hazard and hazard ratio. The estimates of its coefficients are obtained through optimization of the partial likelihood function, which depends on both censored and uncensored individuals.

With the availability of large datasets and high-speed computational power, neural network algorithms have become increasingly popular. Neural networks have been successful when applied to unstructured data such as image recognition and text classification (27–30). Compared to Cox PH, standard neural network architectures focus on predicting outcomes as a binary classification problem at a specific follow-up point. However, it is common in medical studies that individuals are lost to follow-up (censored data) before the failure or event time. Standard neural network models cannot train or test on these individuals, which leads to sample size reduction. In 1995, Faraggi-Simon first combined neural

network architectures with the Cox PH model to make use of censored information as well as to model non-linear features-outcome relations (35). Since then, there has been increasing interest in incorporating neural network architectures in survival analysis. In current literature, there are two main ways of modeling time-to-event using neural networks: (i) adapting Cox PH model and using partial likelihood loss, e.g., Cox-nnet (36) and Deepsurv (37); or (ii) discretizing survival time and using a heuristic loss function, e.g., Nnet-survival (38).

Atherosclerotic cardiovascular disease (ASCVD) is the leading cause of death globally (119). Currently, some commonly used prediction models for ASCVD are based on Cox PH, such as the Framingham CHD risk score and its derivatives (120). In recent years, the American College of Cardiology (ACC)/American Heart Association (AHA) guidelines developed new equations, i.e., the Pooled Cohort Equations (PCEs), to estimate 10-year ASCVD risk in non-Hispanic whites and African Americans (121). The equations are developed based on datasets from several community-based epidemiology cohort studies. The PCEs are four race-, sex-specific and Cox PH based models. It is unclear whether neural network survival models can outperform PCEs for 10-year ASCVD risk prediction. In addition, it is unclear how different architectures of neural network survival models perform compared to each other. In this study, we compared the four race- and sex-specific PCEs with race- and sex-specific state-of-the-art neural network survival models: Nnet-survival, Deepsurv, and Cox-nnet in primary ASCVD risk prediction. For fair comparison, we also included Cox PH models with all significant two-way interactions since this enables Cox PH to capture more complex relationships. For all models, we used the same seven predictors as in the PCEs. Our study is the first study to compare the state-of-the-art neural network survival models with PCEs in incident ASCVD prediction.

Methods

Model I, II: Pooled Cohort Equations, all two-way interaction Cox PH

PCEs are four Cox PH based models, each of which is for a specific race and sex group (white male, white female, black male, black female). Cox PH models the probability an individual experiences the event during a small-time interval given the individual is free of an event at the beginning of the time interval (118), which is also known as hazard rate. Specifically, the hazard function can be expressed as the follows:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) = \lambda_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}), \quad (1)$$

where t is the survival time, $\lambda_0(t)$ is the baseline hazard risk at time t, $\boldsymbol{X}_i$ is the covariates for individual i, $\boldsymbol{\beta}$ is the regression coefficient vector. The hazard function consists of two parts: baseline hazard $\lambda_0(t)$ and a hazard ratio or risk function $\exp\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)$. Cox PH assumes that the relative risk for each covariate ($\boldsymbol{\beta}$ in the equation) is constant over time. The estimate of $\boldsymbol{\beta}$ is obtained by optimizing the Cox partial likelihood function as defined below:

$$l(\beta) = \sum_{i:\Delta_i=1}\left(\boldsymbol{X}_i^T\boldsymbol{\beta} - log\sum_{j:Y_j\geq Y_i}\exp\left(\boldsymbol{X}_j^T\boldsymbol{\beta}\right)\right) \quad (2)$$

where Δ_i is the indicator for the occurrence of event, Y_j is follow-up time for individual j.

In the PCEs, seven predictors were selected based on demonstrated statistical utility using prespecified criteria. These predictors include age at baseline, systolic blood pressure (SBP), diabetes medical history,

treatment for hypertension, current smoker, high density cholesterol and total cholesterol. The interactions between age at baseline and the other predictors were tested based on p-values. Only interactions that had significant p-values ($p<0.05$) were kept in the model. The PCEs demonstrated acceptable performance in derivation samples, with C-statistics for 10-year risk prediction of 0.80 in white females, 0.76 in white males, 0.81 in black females, and 0.70 in black males in 10x10 cross-validation (121).

To capture more complex non-linear relationships between predictors and ASCVD outcome, in the Cox PH-TWI model, we included all the two-way interactions of the 7 predictors in the model for each race and sex. We then retained only the interaction terms that had significant p-values for each race and sex.

Models III and IV: Deepsurv and Cox-nnet

Deepsurv and Cox-nnet are both adaptations of the standard Cox PH (37). Instead of assuming the linear relationship between covariates and log-hazard, the Deepsurv and Cox-nnet models can automatically learn the non-linear relationship between risk factors and an individual's risk of failure by its linear (i.e., multi-layer perceptron) and non-linear (activation functions) transformation. Specifically, the log-risk function $\boldsymbol{X}_i^T\boldsymbol{\beta}$ in the Cox equation as shown in Eq. (1) is replaced by the output from neural network $h_{w,\boldsymbol{\beta}'}(\boldsymbol{X}_i)$, where $\boldsymbol{\beta}'$ is the weight for the last hidden layer and w is the weight for other hidden layers for neural network (see Figure 11A).

$$h_{w,\boldsymbol{\beta}'}(\boldsymbol{X}_i) = \boldsymbol{G}\left(\boldsymbol{W}\boldsymbol{X}_j + \boldsymbol{b}\right)^T\boldsymbol{\beta}'$$

The neural network optimizes the log-partial likelihood function similar to the standard Cox model by tuning parameters $\boldsymbol{W}, \boldsymbol{\beta}'$:

$$l(\boldsymbol{W}, \boldsymbol{\beta}') = \sum_{i:\Delta_i=1}\left(h_{w,\boldsymbol{\beta}'}(\boldsymbol{X}_i) - log\sum_{j:Y_j\geq Y_i}(\exp\left(h_{w,\boldsymbol{\beta}'}(\boldsymbol{X}_j)\right)\right) .$$

Cox-nnet was proposed to deal with high dimensional features especially in genomic studies. To avoid overfitting, Cox-nnet introduces a ridge regularization term and subsequently, the partial log likelihood in Eq. (2) is extended as the following:

$$l(\boldsymbol{W}, \boldsymbol{\beta'}) = \sum_{i:\Delta_i=1} \left(\boldsymbol{G}(\boldsymbol{W}\boldsymbol{X}_i + \boldsymbol{b})^T \boldsymbol{\beta'} - log \sum_{j:Y_j \geq Y_i} \exp\left(\boldsymbol{G}\left(\boldsymbol{W}\boldsymbol{X}_j + \boldsymbol{b}\right)^T \boldsymbol{\beta'} \right) \right) + \lambda(\| \boldsymbol{\beta'} \|_2 + \| \boldsymbol{W} \|_2),$$

In addition to L_2-regularizer, Cox-nnet also allows drop-out for regularization to avoid overfitting. The model is based on Theano framework, therefore, Cox-nnet can be run on a Graphics Processing Unit or multiple threads.

The Deepsurv model also allows the above-mentioned regularization techniques to avoid overfitting. In addition to that, Deepsurv adapted modern techniques to improve the training of the network such as introducing scaled Exponential Linear Units (SELU) as the activation function (37).

Because the structure of $\lambda_i(t) = \lambda_0(t) \exp(\theta)$ is still used in the Cox-nnet and Deepsurv, proportional hazard assumption still stands in the sense that the relative risk between any individual i and j is constant over time.

Model V: Nnet-survival

Nnet-survival is a fully parametric survival model that discretizes survival time. Nnet-survival is proposed to improve two main aspects of the neural network model that are adapted from Cox model: computational speed and the violation of the proportional hazard assumption. Neural network survival models that adapt from Cox model (e.g., Deepsurv, Cox-nnet) use partial likelihood function as the loss function to optimize. The partial likelihood function is calculated based on not only the current individual

but also all the individuals that are at risk at the time point. This makes it difficult to use stochastic gradient descent or mini-batch gradient descent, both of which use a small subset of the whole dataset. Therefore, both Deepsurv and Cox-nnet may have slow convergence and cannot be applied to large datasets that runs out of memory [9]. Nnet-survival was proposed to discretize time, which transforms the model into a fully parametric model and avoids the use of partial likelihood as the loss function. In Nnet-survival models, follow-up time is discretized to n intervals. Hazard h_j is defined as the conditional probability of surviving time interval j given the individual is alive at the beginning of interval j. Survival probability at the end of interval j can be then calculated as the following:

$$S_j = \prod_{i=1}^{j}(1 - h_i).$$

The loss function is defined as the following:

$$L = h_j \prod_{i=1}^{j-1}\left(1 - h_{(i)}\right),$$

for individuals who failed at interval j, and

$$L = \prod_{i=1}^{j-1}\left(1 - h_{(i)}\right),$$

for individuals who are censored at the second half of interval $j - 1$ or the first half of interval j.

There are two main architectures of Nnet-survival: a flexible version and a proportional hazards version. In the flexible version, output layers have m neurons, where m is the number of intervals and each output neuron represents the survival probability at the specific time interval given an individual is alive at the beginning of the time interval. In the proportional hazard version, the final layer only has a single neuron representing $\boldsymbol{X}_i^T\boldsymbol{\beta}$:

$$h_\beta(\boldsymbol{X_i}) = \boldsymbol{X}_i^T \cdot \boldsymbol{\beta}\ ,$$

In our study, the flexible version is used, with its architecture of the flexible version shown in FigurB.

Statistical analysis

In this study, we used the harmonized, individual-level data from 6 cohorts in the Lifetime Risk Pooling Project, including Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study (CHS), Framingham Offsprinig study, Coronary Artery Risk Development in Young Adults (CARDIA) study, the Framingham Original study, and the Multi-Ethnic Study of Atherosclerosis (MESA). The first 5 cohort data were used for model development and internal validation, and the MESA data was used for external validation. We included individuals that meet the following criteria: (i) age between 40 to 79; and (ii) free of a previous history of myocardial infarction, stroke, congestive heart failure, or atrial fibrillation. ASCVD was defined as nonfatal myocardial infarction or coronary heart disease death, or fatal or nonfatal stroke (see (121) for details of selection criteria). All study individuals were free of ASCVD at the beginning of the study and were followed up until the first ASCVD event, loss to follow up, or death, whichever came first. We fit PCE, Cox PH with all two-way interactions (Cox PH-TWI), Nnet-survival, Deepsurv, and Cox-nnet models in white male, white female, black male, and black female participants. For comparison purposes, for all the models, we included the same predictors as used in the PCEs: age at baseline, systolic blood pressure (SBP), diabetes medical history, treatment for hypertension, current smoker, high density cholesterol (HDL-C) and total cholesterol. Individuals who

had missing data at baseline were excluded from the study. Individuals who were lost to follow-up were censored.

To obtain high performance neural network survival models, we manually tuned various hyper-parameters including learning rate, number of layers, number of neurons, number of epochs, batch size, momentum, optimizer, learning rate decay, batch normalization, L_2 regularization, and dropout. After selecting the optimal hyper-parameters, we evaluated model performance through internal validation with 10x10 cross validation and external validation with the MESA data. To perform 10x10 cross-validation, we randomly partitioned the pooled cohort data into 10 equal-sized subsamples. Of the 10 subsamples, 9 subsamples were used as training data and the remaining single subsample was retained as the validation data for testing the model. Each of the subsamples is used in turn as the validation data. We repeated this process 10 times, during which 100 models were built. The average C-statistics and calibration plot of the 100 models were used as the final 10x10 cross-validation result. In the calibration plots, the observed and predicted events were shown in deciles (38). For the external validation, we trained the model in the whole harmonized dataset (not including MESA cohort), and evaluated the model discrimination and calibration in the external MESA cohort (122). To compare if the differences among C-statistics were significant in neural network models vs. PCE models, we performed significant test using method proposed by Mogensen et al (123). MESA is a more contemporary cohort that had lower CVD event rate compared to the earlier cohorts (121). This difference could cause models have poor calibration in MESA. To overcome this, we performed recalibration on all models using the method proposed by Pennells et al. (124). Briefly, we first calculated rescaling factors that were needed to bring predicted risks in line with observed risks using regression model in MESA dataset. We then applied the rescaling factors to the original predicted risk and got recalibrated risk estimates for all participants.

Nnet-survival, Deepsurv, and Cox-nnet were implemented in python, version 3.7.3. Cox PH model was conducted using the "survival" package in R, version 3.6.0. C-statistics was calculated using the "survC1"

package in R, version 3.6.0 (125). Significant test was done using the "pec" package in R, version 3.6.0 (123). Regression model for recalibration was performed using "scikit-learn" module in python, version 3.7.3 (126).

All data were de-identified, and all study protocols and procedures were approved by the Institutional Review Board at Northwestern University with a waiver for informed consent. All methods were performed in accordance with the relevant guidelines and regulations.

Results

Overall, there were 23216 participants, including 8644 white male, 1354 black male, 10719 white female, 2499 black female individuals. The average age at baseline was 57.8 years old (SD = 9.6). Among these individuals, 16.0% developed ASCVD with average follow-up of 10.50 (SD = 3.02) years. The mean SBP value was 127.1 mmHg (SD = 21.0), the mean HDL-C value was 51.6 mg/dL (SD = 16.4), total cholesterol was 217.8 mg/dL (SD = 43.0). For binary predictors, 4.6% individuals had a history of diabetes, 26.0% individuals were current smokers, 31.6 % individuals had treatment for hypertension. The descriptive statistics for each race and sex group were shown in Table 1.

In the MESA external validation dataset, there were 4259 individuals in total. The average age at baseline was 61.6 years old (SD = 9.6). Among the 4259 individuals, 331 (7.77%) developed ASCVD with average follow-up years of 10.97 years old (SD = 2.48). Among these individuals, there were 1194 white male, 799 black male, 1284 white female, and 982 black female. Baseline characteristics of the study sample were shown in Table 1, stratified by sex and race group.

In 10x10 cross validation, in the white male population (see Figure 12), Deepsurv achieved the highest C-statistics (0.7371) among all the models. In the white female population, Deepsurv had the highest C-statistics (0.7972) comparable to Cox PH-TWI (0.7972). In the black male population, PCE had the highest C-statistics (0.6981). In the black female population, Deepsurv had the highest C-statistics (0.7886). The details of C-statistics for each model and race sex group were shown in Supplemental Table 1. In the external validation dataset, in white male population, Deepsurv had the highest C-statistics (0.7032). In white female population, Cox-nnet had the highest C-statistics (0.7282). In black male population, PCE had the highest C-statistics (0.6811). In black female population, Deepsurv (0.7316) had the highest C-statistics and this difference was statistically significant compared to PCE (p=0.004, see Supplemental Table 1). In general, Deepsurv had the highest C-statistics for 4 times followed by PCE which had the highest C-statistics for 2 times followed by Cox-nnet and Nnet-survival. However, the difference between all neural network models vs. PCE are not significant except for Deepsurv in black female (see Supplemental Table 1). Overall, Deepsurv had the best C-statistics for 4 times followed by PCE which had the best C-statistics for 2 times.

In terms of calibration in 10x10 cross-validation (see Figure 13), the calibration plot showed that all five models had similar calibration compared to PCE in white male population. In white female and black female population, neural network models outperformed PCE. In black male population, PCE had the best calibration compared to all other models. In the MESA external dataset, calibration plot showed that all five models overestimated the event rate among all race gender groups. In the white male population, all five models had similar overestimation with predicted event rate ranging from 0 to 0.57 compared to 0-0.2 in the observed event rate (see Figure 14). In the white female population, PCE, Cox-nnet, and Nnet-survival had better calibration compared to the other groups. In the black male population, Deepsurv and Cox PH-TWI were closer to the Kaplan Meier estimation compared to other models. In the black female population, Cox-nnet had better calibration compared to the other models. After recalibrating the model

by fitting linear regression models, over-estimation of event risk was greatly reduced in all models among all race gender groups Figure 15. Overall, models in female groups had better calibration than in male groups. Nnet-survival had relatively good performance among all race gender groups.

Discussion

In this study, we implemented state-of-the-art neural network survival models in predicting 10-year risk for a first ASCVD event. Our results showed that overall, when using the same predictors as in the PCEs, neural network survival models and PCE had comparable performance. Neural network survival models outperformed PCE in white male, white female, and black female population by slim margin. However, the difference is not statistically significant expect for Deepsurv in black female. In terms of calibration, in internal validation dataset, PCEs had good calibration in white male and black male population. In white and black female population, neural network models outperformed PCEs. In external validation dataset, all models over-estimated the event rate in all four race-sex groups. Recalibration largely reduced the overestimation.

Theoretically, among the different neural network survival models, Nnet-survival is faster in training than Deepsurv and Cox-nnet models. Nnet-survival's loss function only relies on individuals in the current minibatch which allows mini-batch gradient descent while both Deepsurv and Cox-nnet require the entire dataset for each gradient descent update. On the other hand, the discretization of time-to-event in Nnet-survival leads to a less smooth predicted survival curve compared to Deepsurv and Cox-nnet.

In prior studies, Gensheimer et al applied Cox PH, Nnet-survival, Deepsurv, and Cox-nnet in life expectancy prediction using the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) dataset (38). The dataset consisted of 9105 individuals and 39 predictors. The four neural network survival models generated similar C-statistics compared to the Cox PH model, which

was consistent with our findings in ASCVD prediction. Both the SUPPORT dataset and our dataset had low dimension number of predictors. Several studies explored other machine learning methods for CVD prediction. Joo et al (127) applied logistic regression, deep neural networks, random forests, and LightGBM to predict CVD as a binary outcome using the Korean National Health Insurance Service–National Health Sample Cohort dataset. The authors found that deep neural network had better performance (C-statistics = 0.7446) compared to the PCE (C-index = 0.7381) in that cohort. However, the ML models used more predictors (hemoglobin level, diastolic blood pressure, presence of proteinuria, serum aspartate aminotransferase, serum alanine aminotransferase, and total cholesterol) compared to the PCE. In another study, Dimopoulos et al implemented KNN, random forest, and decision tree to predict CVD compared to the HellenicSCORE, a Cox regression based model (128). Their results showed that ML models have comparable performance compared to the HellenicSCORE (129) using 5 and 13 same predictors respectively but were not able to outperform the baseline model.

Similar to other machine learning models, neural network models often show advantage in modeling non-linear complex relationships between predictors and outcome. The explanatory variables used in our study are all well-studied predictors of cardiovascular disease. The biologic basis for many of these variables are understood, and they are known to predict independently and often linearly. In this situation, simpler models can be the best since they can accurately capture a linear biologic relation without sacrificing interpretation. Similar conclusions were reached in data comparing three machine learning methods to a simpler logistic regression model for predicting death after acute myocardial infarction. In the study, two of the 3 machine learning algorithms improved discrimination by a slim margin (130). In the follow up editorial by Engelhard et al. (131), they mentioned that machine learning has been most impactful with complex data (e.g., high dimensional, highly structured, and difficult to summarize without substantial loss of information). Our findings further support this hypothesis. In the future, we expect neural network

based models to be a powerful tool in CVD prediction when more abundant data and more complex repeated measures data become available (e.g. electronic health records data).

Limitations

Our study has several limitations. First, the cohorts we used from the Lifetime Risk Pooling Project were the same cohorts used in the derivation of the PCEs. This may have led to some optimism in the performance of the PCEs. Second, the participants of our external validation cohort, MESA, were perhaps healthier than the general population. More importantly, they received intensive screening for subclinical CVD, which influenced health behaviors and preventive interventions including use of effective drug therapies; this may result in the lower event rate in MESA participants than what would have been predicted because of the use of effective preventive therapies selectively in higher-risk individuals.

Conclusion

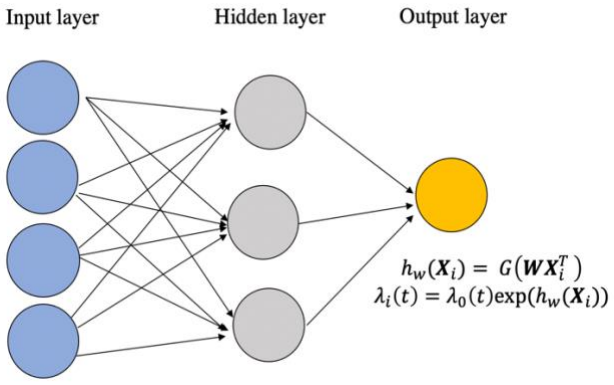
Neural network survival models can achieve comparable discrimination if not superior performance compared to the PCEs in 10-year time-to-ASCVD prediction in the white female, white male, black female, and black male population in our dataset. In future studies, high dimensional features or longitudinal data should be considered to fully explore the benefits of neural network survival models.

Table 1. Baseline characteristics for each race and sex group in training/internal validation dataset and external validation dataset.

	Overall	White male	Black male	White female	Black female
Training/internal validation dataset					
N	23216	8644	1354	10719	2499
Number of Events, n (%)	3705 (16.0)	1788 (20.7)	300 (22.2)	1217 (11.4)	400 (16.0)
Age (year), mean (SD)	57.8 (9.6)	58.0 (9.6)	57.3 (9.5)	58.0 (9.7)	56.4 (9.3)
SBP (mm Hg), mean (SD)	127.1 (21.0)	127.2 (19.5)	131.8 (21.5)	125.5 (21.6)	130.8 (22.7)
HDL-C (mg/dL), mean (SD)	51.6 (16.4)	43.8 (12.7)	49.7 (16.0)	56.8 (16.5)	57.5 (16.4)
TOTCHL (mg/dL), mean (SD)	217.8 (43.0)	212.1 (39.9)	208.6 (44.4)	223.9 (43.8)	216.0 (45.4)
HXDIAB, n (%)	1069 (4.6)	295 (3.4)	175 (12.9)	294 (2.7)	305 (12.2)
Smoker, n (%)	6035 (26.0)	2294 (26.5)	441 (32.6)	2723 (25.4)	577 (23.1)
RXHYP, n (%)	7326 (31.6)	2226 (25.8)	707 (52.2)	2951 (27.5)	1442 (57.7)
MESA external validation dataset					
N	4259	1194	799	1284	982
Number of Events, n (%)	331 (7.8)	104 (8.7)	85 (10.6)	79 (6.2)	63 (6.4)
Age (year), mean (SD)	61.6 (9.6)	61.9 (9.6)	61.5 (9.6)	61.5 (9.6)	61.3 (9.4)
SBP (mmHg), mean (SD)	126.3 (21.0)	123.7 (18.3)	130.0 (19.2)	121.8 (21.4)	132.4 (22.8)
HDL-C (mm/dL), mean (SD)	52.2 (15.5)	45.2 (12.1)	46.5 (12.5)	58.8 (15.8)	56.9 (15.6)
TOTCHL (mm/dL), mean (SD)	193.3 (35.7)	189.2 (34.4)	182.0 (34.6)	202.3 (34.4)	195.7 (36.5)
HXDIAB, n (%)	354 (8.3)	57 (4.8)	117 (14.6)	51 (4.0)	129 (13.1)
Smoker, n (%)	628 (14.7)	137 (11.5)	166 (20.8)	160 (12.5)	165 (16.8)
RXHYP, n (%)	1676 (39.4)	389 (32.6)	370 (46.3)	402 (31.3)	515 (52.4)

Abbreviations: SBP, systolic blood pressure; HDL-C, high density cholesterol; TOTCHL, total cholesterol; HXDIAB, history of diabetes; RXHYP, history of hypertension.

A. Deepsurv/Cox-nnet



B. Nnet-survival

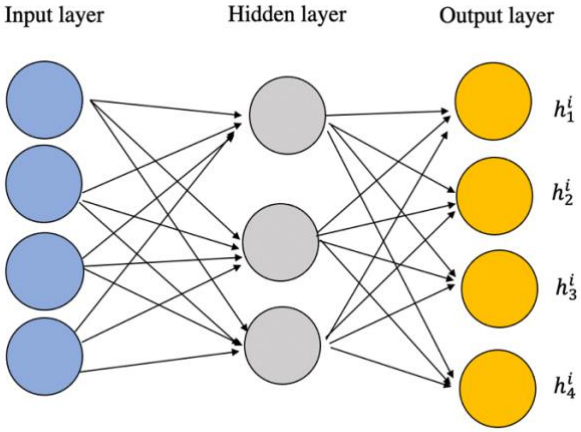


Figure 11. Frameworks for neural network survival models.

The framework of Deepsurv/Cox-nnet is shown in panel A. The framework of Nnet-survival is shown in panel B. In Figure 1A, the Deepsurv/Cox-nnet model outputs $h_w(\boldsymbol{X}_i)$ which is used to replace the log risk $\boldsymbol{X}_i^T\boldsymbol{\beta}$ in the Cox model. In Figure 1B, in the Nnet-survival model, the output layers generate h_j^i which is the hazard for individual i at time j.

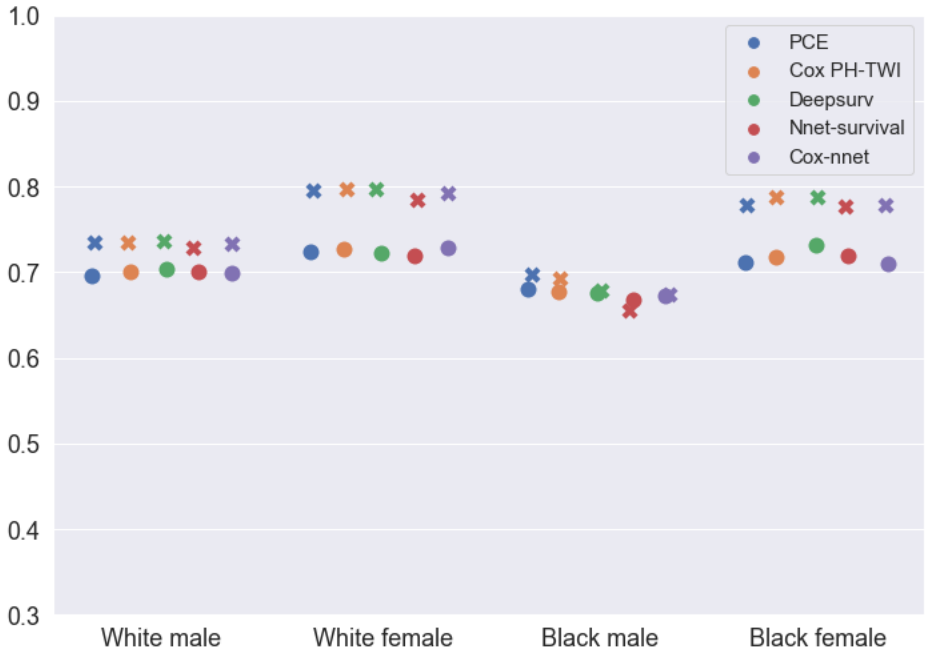


Figure 12. C-statistics for PCEs, Nnet-survival, Deepsurv, Cox-nnet, and Cox PH-TWI in 10x10 cross-validation and MESA external validation.

The ‘x’ markers represent C-statistics in 10x10 cross-validation, the ‘o’ markers represent C-statistics in MESA external validation.

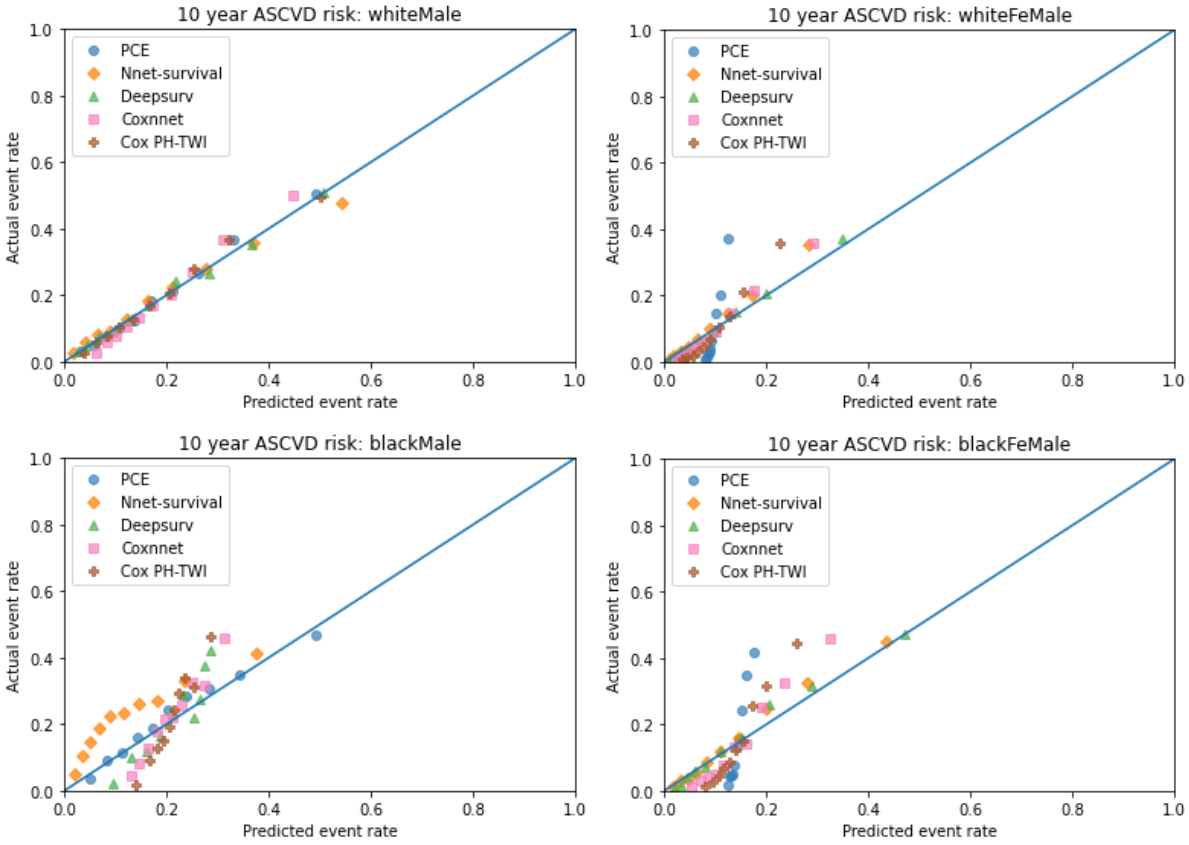


Figure 13. Kaplan-Meier Observed Event Rate and Predicted Event Rate for the ASCVD Outcome in the 10x10 cross-validation.

For each model, we divided participants into 10 group (decile) based on their sorted predicted event probability. Then, for each decile, mean observed event rate (Kaplan-Meier method) was plotted against mean predicted event rate. In a perfectly calibrated model, the predicted event rate would be the same as the observed event rate in each decile. This means that all points would be clustered around the blue identity line.

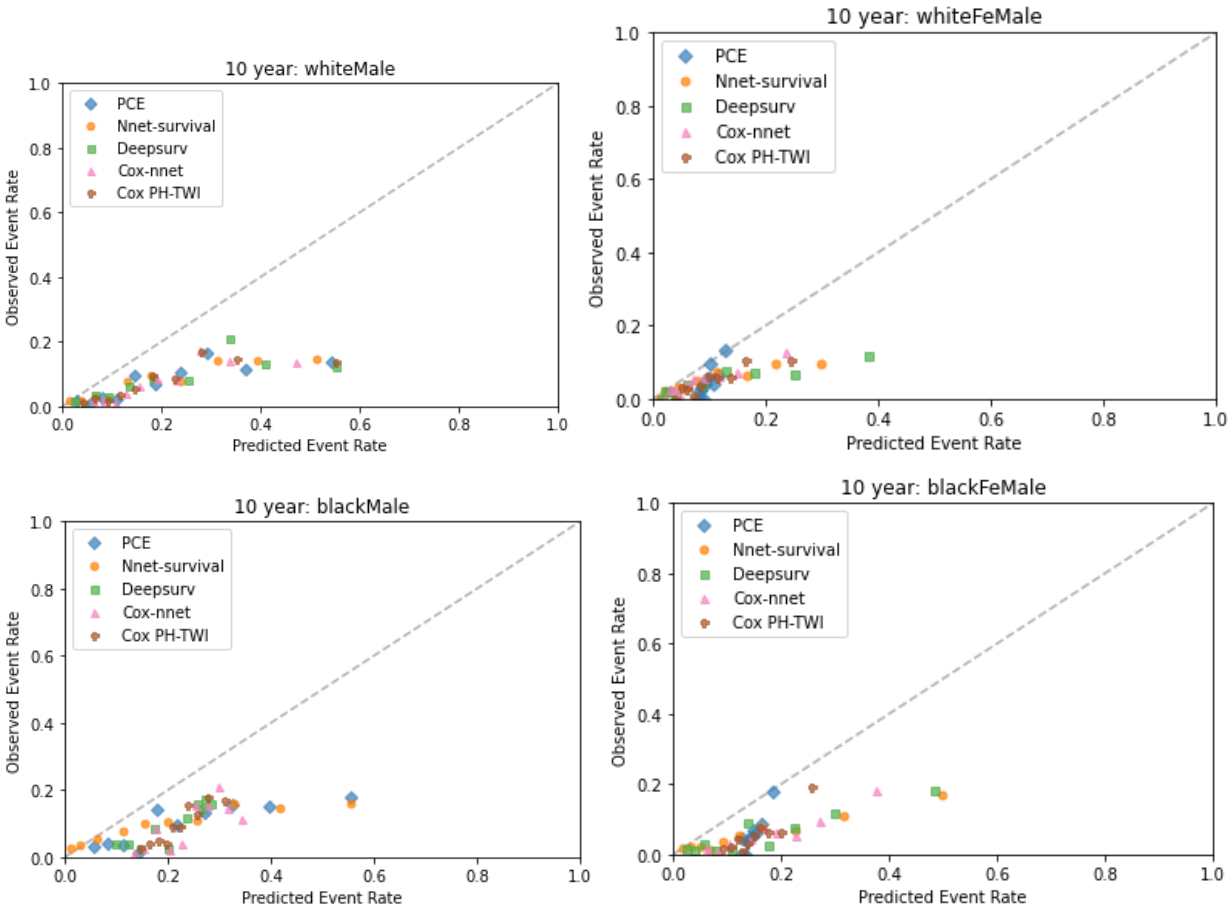


Figure 14. Kaplan-Meier Observed Event Rate and Predicted Event Rate for the ASCVD Outcome in the MESA Cohort. For each model, we divided participants into 10 group (decile) based on their sorted predicted event rate. Then, for each decile, mean observed event rate (Kaplan-Meier method) was plotted against mean predicted event rate. In a perfectly calibrated model, the predicted event rate would be the same as the observed event rate in each decile. This means that all points would be clustered around the dotted identity line.

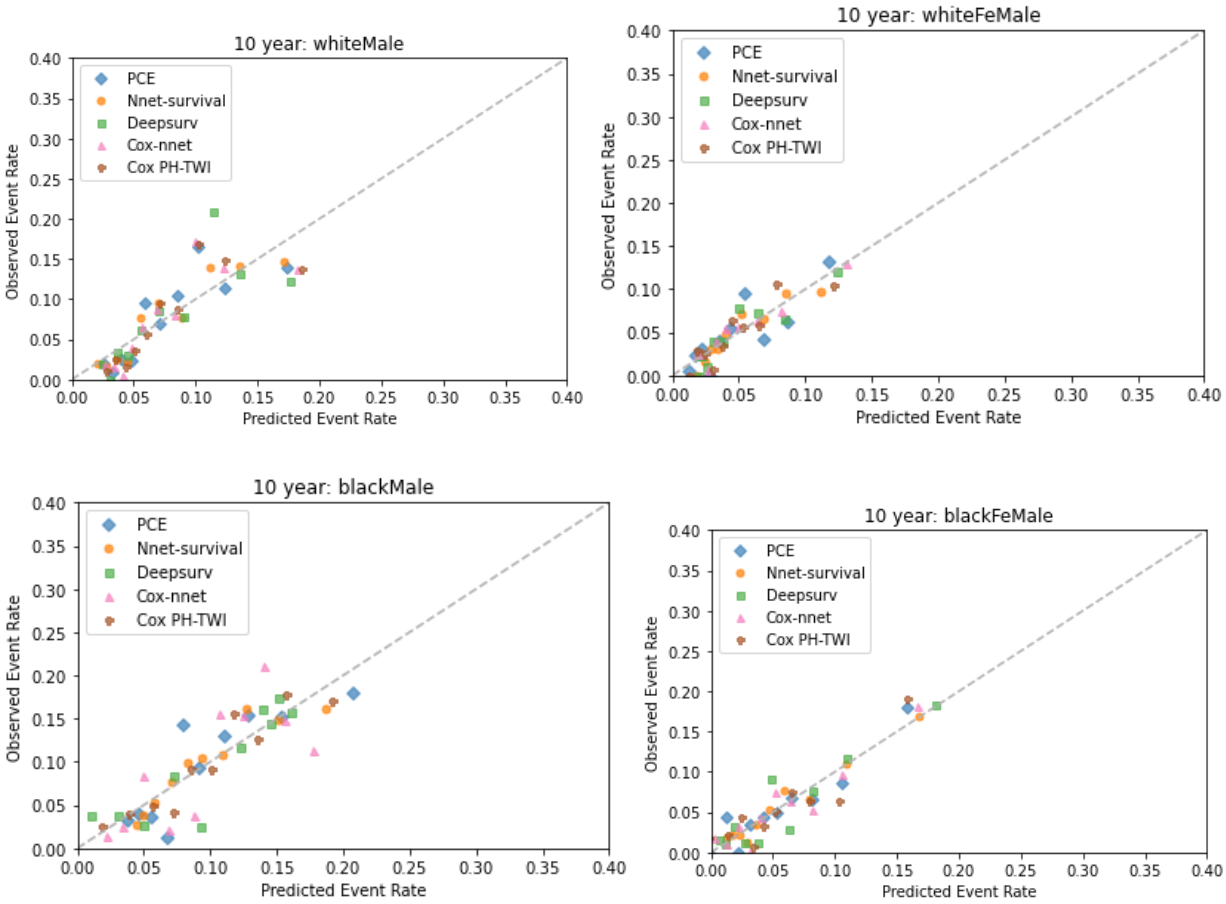


Figure 15. Kaplan-Meier Observed Event Rate and Recalibrated Predicted Event Rate for the ASCVD Outcome in the MESA Cohort. For each model, we divided participants into 10 group (decile) based on their sorted predicted event rate. Then, for each decile, mean observed event rate (Kaplan-Meier method) was plotted against mean predicted event rate. In a perfectly calibrated model, the predicted event rate would be the same as the observed event rate in each decile.

Reference

1. Kruse CS, Stein A, Thomas H, Kaur H. The use of Electronic Health Records to Support Population Health: A Systematic Review of the Literature. J Med Syst [Internet]. 2018 [cited 2020 Jun 23];42(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6182727/

2. Blumenthal D, Tavenner M. The "Meaningful Use" Regulation for Electronic Health Records. New England Journal of Medicine. 2010 Aug 5;363(6):501–4.

3. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining Data From Electronic Health Records [Internet]. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2 [Internet]. Agency for Healthcare Research and Quality (US); 2019 [cited 2022 Mar 10]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551878/

4. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learn Health Syst. 2019 Jan 16;3(1):e10076.

5. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Medical Research Methodology. 2021 Oct 27;21(1):234.

6. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017 Jan 1;106(1):1–9.

7. Ramsey SD, Adamson BJ, Wang X, Bargo D, Baxi SS, Ghosh S, et al. Using electronic health record data to identify comparator populations for comparative effectiveness research. J Med Econ. 2020 Dec;23(12):1618–22.

8. Building electronic data infrastructure for comparative effectiveness research: accomplishments, lessons learned and future steps | Journal of Comparative Effectiveness Research [Internet]. [cited 2022 Mar 14]. Available from: https://www.futuremedicine.com/doi/abs/10.2217/cer.14.73

9. Manion FJ, Harris MR, Buyuktur AG, Clark PM, An LC, Hanauer DA. Leveraging EHR Data for Outcomes and Comparative Effectiveness Research in Oncology. Curr Oncol Rep. 2012 Dec;14(6):494–501.

10. Deng Y, Ghamsari F, Lu A, Yu J, Zhao L, Kho A. Comparative Effectiveness of Second Line Anti-Diabetic Medication on Incidence of Chronic Kidney Disease [Internet]. Available from: 10.6084/m9.figshare.13426469

11. Comparative effectiveness of incident oral antidiabetic drugs on kidney function - PubMed [Internet]. [cited 2020 Aug 20]. Available from: https://pubmed.ncbi.nlm.nih.gov/22258320/

12. Dacks PA, Armstrong JJ, Brannan SK, Carman AJ, Green AM, Kirkman MS, et al. A call for comparative effectiveness research to learn whether routine clinical care decisions can protect from dementia and cognitive decline. Alz Res Therapy. 2016 Aug 20;8(1):33.

13. Fiks AG, Grundmeier RW, Margolis B, Bell LM, Steffes J, Massey J, et al. Comparative Effectiveness Research Using the Electronic Medical Record: An Emerging Area of Investigation in Pediatric Primary Care. J Pediatr. 2012 May;160(5):719–24.

14. Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib | Journal of Comparative Effectiveness Research [Internet]. [cited 2022 Mar 14]. Available from: https://www.futuremedicine.com/doi/full/10.2217/cer-2018-0032

15. Comparative Effectiveness Study of Osteoporosis Medications Using Electronic Medical Records [Internet]. [cited 2022 Mar 14]. Available from: https://dash.harvard.edu/handle/1/42061455

16. PCORI Awards $93.5 Million to Develop National Network to Support More Efficient Patient-Centered Research [Internet]. PCORI Awards $93.5 Million to Develop National Network to Support More Efficient Patient-Centered Research | PCORI. 2013 [cited 2022 Mar 14]. Available from: https://www.pcori.org/news-release/pcori-awards-935-million-develop-national-network-support-more-efficient-patient

17. HCS Research Collaboratory [Internet]. 2013 [cited 2022 Mar 14]. Available from: https://commonfund.nih.gov/hcscollaboratory

18. Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Heim L, et al. Chlorhexidine versus routine bathing to prevent multidrug-resistant organisms and all-cause bloodstream infections in general medical and surgical units (ABATE Infection trial): a cluster-randomised trial. Lancet. 2019 Mar 23;393(10177):1205–15.

19. Research C for DE and. FDA approves new use of transplant drug based on real-world evidence. FDA [Internet]. 2021 Sep 30 [cited 2022 Mar 14]; Available from: https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-new-use-transplant-drug-based-real-world-evidence

20. Kessler RC, Hwang I, Hoffmire CA, McCarthy JF, Petukhova MV, Rosellini AJ, et al. Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. Int J Methods Psychiatr Res. 2017 Sep;26(3).

21. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H, et al. Using electronic health record collected clinical variables to predict medical intensive care unit mortality. Ann Med Surg (Lond). 2016 Sep 6;11:52–7.

22. Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, et al. Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation. JMIR Med Inform. 2020 Jul 6;8(7):e17257.

23. Wang Y, Wei Y, Yang H, Li J, Zhou Y, Wu Q. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Medical Informatics and Decision Making. 2020 Sep 21;20(1):238.

24. Tomašev N, Harris N, Baur S, Mottram A, Glorot X, Rae JW, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nat Protoc. 2021 Jun;16(6):2765–87.

25. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Med. 2018 May 8;1(1):1–10.

26. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data. Circulation: Cardiovascular Quality and Outcomes. 2019 Oct;12(10):e005114.

27. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2012 [cited 2021 Aug 17]. Available from: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

28. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019 Jan;16(1):139–53.

29. Zhao Y, Hong Q, Zhang X, Deng Y, Wang Y, Petzold L. BERTSurv: BERT-Based Survival Models for Predicting Outcomes of Trauma Patients. arXiv:210310928 [cs] [Internet]. 2021 Mar 19 [cited 2021 Aug 17]; Available from: http://arxiv.org/abs/2103.10928

30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2021 Aug 17]. Available from: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

31. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks - ScienceDirect [Internet]. [cited 2022 Mar 14]. Available from: https://www.sciencedirect.com/science/article/pii/S1532046419302813

32. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017 Mar 1;24(2):361–70.

33. Shang J, Ma T, Xiao C, Sun J. Pre-training of Graph Augmented Transformers for Medication Recommendation. arXiv:190600346 [cs] [Internet]. 2019 Nov 26 [cited 2022 Mar 14]; Available from: http://arxiv.org/abs/1906.00346

34. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit Med. 2021 May 20;4(1):1–13.

35. Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995;14(1):73–82.

36. Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLOS Computational Biology. 2018 Apr 10;14(4):e1006076.

37. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018 Feb 26;18(1):24.

38. Gensheimer MF, Narasimhan B. A Scalable Discrete-Time Survival Model for Neural Networks. PeerJ. 2019 Jan 25;7:e6257.

39. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117–21.

40. Missing clinical and behavioral health data in a large electronic health record (EHR) system | Journal of the American Medical Informatics Association | Oxford Academic [Internet]. [cited 2022 Mar 12]. Available from: https://academic.oup.com/jamia/article/23/6/1143/2399287?login=true

41. Culbertson A, Goel S, Madden MB, Safaeinili N, Jackson KL, Carton T, et al. The Building Blocks of Interoperability. A Multisite Analysis of Patient Demographic Attributes Available for Matching. Appl Clin Inform. 2017 Apr 5;8(2):322–36.

42. Tse J, You W. How accurate is the electronic health record? - a pilot study evaluating information accuracy in a primary care setting. Stud Health Technol Inform. 2011;168:158–64.

43. Data Accuracy in Electronic Medical Record Documentation | Electronic Health Records | JAMA Ophthalmology | JAMA Network [Internet]. [cited 2022 Mar 13]. Available from: https://jamanetwork.com/journals/jamaophthalmology/article-abstract/2598417

44. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. Journal of Biomedical Informatics. 2014 Dec 1;52:199–211.

45. Bower JK, Patel S, Rudy JE, Felix AS. Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. Curr Epidemiol Rep. 2017 Dec;4(4):346–52.

46. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014 Apr;21(2):221–30.

47. ICD - ICD-9 - International Classification of Diseases, Ninth Revision [Internet]. 2022 [cited 2022 Mar 14]. Available from: https://www.cdc.gov/nchs/icd/icd9.htm

48. The Web's Free 2022 ICD-10-CM/PCS Medical Coding Reference [Internet]. [cited 2022 Mar 14]. Available from: https://www.icd10data.com/

49. SNOMED Home page [Internet]. SNOMED. [cited 2022 Mar 14]. Available from: https://www.snomed.org/

50. RxNorm [Internet]. U.S. National Library of Medicine; [cited 2022 Mar 14]. Available from: https://www.nlm.nih.gov/research/umls/rxnorm/index.html

51. Home [Internet]. LOINC. [cited 2022 Mar 14]. Available from: https://loinc.org/

52. Zhou S, Wang L, Wang N, Liu H, Zhang R. CancerBERT: a BERT model for Extracting Breast Cancer Phenotypes from Electronic Health Records. arXiv:210811303 [cs] [Internet]. 2022 Mar 9 [cited 2022 Mar 14]; Available from: http://arxiv.org/abs/2108.11303

53. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] [Internet]. 2019 May 24 [cited 2022 Mar 14]; Available from: http://arxiv.org/abs/1810.04805

54. Deng Y, Ghamsari F, Lu A, Yu J, Zhao L, Kho AN. Use of real-world evidence data to evaluate the comparative effectiveness of second-line type 2 diabetes medications on chronic kidney disease. medRxiv. 2021 Jan 1;2021.06.15.21258963.

55. Deng Y, Pacheco JA, Chung A, Mao C, Smith JC, Zhao J, et al. Natural language processing to identify lupus nephritis phenotype in electronic health records. arXiv:211210821 [cs] [Internet]. 2021 Dec 20 [cited 2022 Mar 14]; Available from: http://arxiv.org/abs/2112.10821

56. Deng Y, Liu L, Jiang H, Peng Y, wei Y, Zhong Y, et al. Comparison of State-Of-The-Art Neural Network Survival Models With The Pooled Cohort Equations for Cardiovascular

Disease Risk Prediction [Internet]. In Review; 2021 Oct [cited 2022 Mar 14]. Available from: https://www.researchsquare.com/article/rs-958135/v1

57. Alicic RZ, Rooney MT, Tuttle KR. Diabetic Kidney Disease: Challenges, Progress, and Possibilities. CJASN [Internet]. 2017 May 18 [cited 2020 Aug 20]; Available from: https://cjasn.asnjournals.org/content/early/2017/07/12/CJN.11491116

58. Wilkinson SV, Tomlinson LA, Iwagami M, Stirnadel-Farrant HA, Smeeth L, Douglas I. A systematic review comparing the evidence for kidney function outcomes between oral antidiabetic drugs for type 2 diabetes. Wellcome Open Res. 2018;3:74.

59. Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, et al. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. New England Journal of Medicine. 2017 Aug 17;377(7):644–57.

60. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes | NEJM [Internet]. [cited 2021 Apr 27]. Available from: https://www.nejm.org/doi/full/10.1056/NEJMoa1504720

61. Wiviott SD, Raz I, Bonaca MP, Mosenzon O, Kato ET, Cahn A, et al. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. New England Journal of Medicine. 2019 Jan 24;380(4):347–57.

62. Heerspink HJL, Stefánsson BV, Correa-Rotter R, Chertow GM, Greene T, Hou F-F, et al. Dapagliflozin in Patients with Chronic Kidney Disease. New England Journal of Medicine [Internet]. 2020 Sep 24 [cited 2020 Oct 12]; Available from: https://www.nejm.org/doi/10.1056/NEJMoa2024816

63. Understanding the Gap Between Efficacy in Randomized Controlled Trials and Effectiveness in Real-World Use of GLP-1 RA and DPP-4 Therapies in Patients With Type 2 Diabetes | Diabetes Care [Internet]. [cited 2020 Jul 15]. Available from: https://care.diabetesjournals.org/content/40/11/1469

64. Schneeweiss S, Patorno E. Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments. Endocrine Reviews [Internet]. 2021 Mar 12 [cited 2021 Jun 11];(bnab007). Available from: https://doi.org/10.1210/endrev/bnab007

65. Xie Y, Bowe B, Gibson AK, McGill JB, Maddukuri G, Yan Y, et al. Comparative Effectiveness of SGLT2 Inhibitors, GLP-1 Receptor Agonists, DPP-4 Inhibitors, and Sulfonylureas on Risk of Kidney Outcomes: Emulation of a Target Trial Using Health Care Databases. Diabetes Care. 2020 Nov 1;43(11):2859–69.

66. Northwestern Medicine Enterprise Data Warehouse (NMEDW): Research: Feinberg School of Medicine: Northwestern University [Internet]. [cited 2021 Apr 27]. Available from: https://www.feinberg.northwestern.edu/research/cores/units/edw.html

67. WHOCC - ATC/DDD Index [Internet]. [cited 2020 Oct 9]. Available from: https://www.whocc.no/atc_ddd_index/

68. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011 Mar;20(1):40.

69. Barnard J, Rubin DB. Small-Sample Degrees of Freedom with Multiple Imputation. Biometrika. 1999;86(4):948–55.

70. Ross JS. Randomized Clinical Trials and Observational Studies Are More Often Alike Than Unlike. JAMA Intern Med. 2014 Oct 1;174(10):1557–1557.

71. GRADE Study Group. Glycemia Reduction Approaches in Diabetes: A Comparative Effectiveness Study [Internet]. clinicaltrials.gov; 2021 Feb [cited 2021 Apr 26]. Report No.: NCT01794143. Available from: https://clinicaltrials.gov/ct2/show/NCT01794143

72. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JFE, Nauck MA, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. New England Journal of Medicine. 2016 Jul 28;375(4):311–22.

73. Effects of Linagliptin on Cardiovascular and Kidney Outcomes in People With Normal and Reduced Kidney Function: Secondary Analysis of the CARMELINA Randomized Trial | Diabetes Care [Internet]. [cited 2020 Sep 9]. Available from: https://care.diabetesjournals.org/content/early/2020/05/20/dc20-0279

74. Cooper ME, Perkovic V, McGill JB, Groop P-H, Wanner C, Rosenstock J, et al. Kidney Disease End Points in a Pooled Analysis of Individual Patient–Level Data From a Large Clinical Trials Program of the Dipeptidyl Peptidase 4 Inhibitor Linagliptin in Type 2 Diabetes. American Journal of Kidney Diseases. 2015 Sep 1;66(3):441–9.

75. Linagliptin Lowers Albuminuria on Top of Recommended Standard Treatment in Patients With Type 2 Diabetes and Renal Dysfunction | Diabetes Care [Internet]. [cited 2020 Sep 8]. Available from: https://care.diabetesjournals.org/content/36/11/3460?ijkey=d7fbf7b1ce210bede909e4a7587050d9cadd67a0&keytype2=tf_ipsecsha

76. Mann JFE, Ørsted DD, Brown-Frandsen K, Marso SP, Poulter NR, Rasmussen S, et al. Liraglutide and Renal Outcomes in Type 2 Diabetes. New England Journal of Medicine. 2017 Aug 31;377(9):839–48.

77. Davies MJ, Bain SC, Atkin SL, Rossing P, Scott D, Shamkhalova MS, et al. Efficacy and Safety of Liraglutide Versus Placebo as Add-on to Glucose-Lowering Therapy in Patients With Type 2 Diabetes and Moderate Renal Impairment (LIRA-RENAL): A Randomized Clinical Trial. Diabetes Care. 2016 Feb;39(2):222–30.

78. Suissa S. Lower Risk of Death With SGLT2 Inhibitors in Observational Studies: Real or Bias? Diabetes Care. 2018 Jan 1;41(1):6–10.

79. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database (Oxford) [Internet]. 2020 Mar 17 [cited 2020 Jun 23];2020. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7078068/

80. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation. 2015 Jan 20;131(3):269–79.

81. Workman TA. Defining Patient Registries and Research Networks [Internet]. Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks [Internet]. Agency for Healthcare Research and Quality (US); 2013 [cited 2020 Jun 23]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK164514/

82. Use of cluster analysis to define COPD phenotypes | European Respiratory Society [Internet]. [cited 2020 Jun 23]. Available from: https://erj.ersjournals.com/content/36/3/472

83. Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine. 2015 Oct 28;7(311):311ra174-311ra174.

84. Petri M, Orbai A-M, Alarcón GS, Gordon C, Merrill JT, Fortin PR, et al. Derivation and Validation of Systemic Lupus International Collaborating Clinics Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheum. 2012 Aug;64(8):2677–86.

85. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, et al. 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis & Rheumatology. 2019;71(9):1400–12.

86. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis & Rheumatism. 1997;1725–1725.

87. Chicago Lupus Database: Systemic Lupus Research Studies: Feinberg School of Medicine: Northwestern University [Internet]. [cited 2020 Jun 23]. Available from: https://www.lupus.northwestern.edu/research/cld.html

88. Mahieu MA, Strand V, Simon LS, Lipsky PE, Ramsey-Goldman R. A critical review of clinical trials in systemic lupus erythematosus. Lupus. 2016 Sep;25(10):1122–40.

89. Andersen R, Hagenaars JA, McCutcheon AL. Applied Latent Class Analysis. Canadian Journal of Sociology / Cahiers canadiens de sociologie. 2003;28(4):584.

90. Kwan MY, Arbour-Nicitopoulos KP, Duku E, Faulkner G. Patterns of multiple health risk–behaviours in university students and their association with mental health: application of latent class analysis. Health Promot Chronic Dis Prev Can. 2016 Aug;36(8):163–70.

91. Whitson HE, Johnson KS, Sloane R, Cigolle CT, Pieper CF, Landerman L, et al. Identifying Patterns of Multimorbidity in Older Americans: Application of Latent Class Analysis. Journal of the American Geriatrics Society. 2016;64(8):1668–73.

92. Empirically Derived Subgroups of Self-Injurious Thoughts and Behavior: Application of Latent Class Analysis - Dhingra - 2016 - Suicide and Life-Threatening Behavior - Wiley Online Library [Internet]. [cited 2020 Jun 23]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/sltb.12232?casa_token=z63KM1TseMYAAAAA%3AJIBvMGU162x-mrpr96e_R1GaHifTrvFyv3XNfLBRBcooK2lJB6ck0piwkdp5kl-trIEzLjKTTm7RaA8

93. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. arXiv:180604820 [cs] [Internet]. 2018 Jun 14 [cited 2020 Jul 20]; Available from: http://arxiv.org/abs/1806.04820

94. Maaten L van der, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008 Nov;2579–605.

95. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis. Journal of Statistical Software.

96. Mechanisms of Kidney Injury in Lupus Nephritis – the Role of Anti-dsDNA Antibodies [Internet]. [cited 2020 Aug 10]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4569852/

97. Artim-Esen B, Çene E, Şahinkaya Y, Ertan S, Pehlivan Ö, Kamali S, et al. Cluster analysis of autoantibodies in 852 patients with systemic lupus erythematosus from a single center. J Rheumatol. 2014 Jul;41(7):1304–10.

98. Pego-Reigosa JM, Lois-Iglesias A, Rúa-Figueroa Í, Galindo M, Calvo-Alén J, de Uña-Álvarez J, et al. Relationship between damage clustering and mortality in systemic lupus erythematosus in early and late stages of the disease: cluster analyses in a large cohort from the Spanish Society of Rheumatology Lupus Registry. Rheumatology (Oxford). 2016 Jul 1;55(7):1243–50.

99. Li PH, Wong WHS, Lee TL, Lau CS, Chan TM, Leung AMH, et al. Relationship between autoantibody clustering and clinical subsets in SLE: cluster and association analyses in Hong Kong Chinese. Rheumatology (Oxford). 2013 Feb 1;52(2):337–45.

100. Tang X, Huang Y, Deng W, Tang L, Weng W, Zhang X. Clinical and Serologic Correlations and Autoantibody Clusters in Systemic Lupus Erythematosus: A Retrospective Review of 917 Patients in South China. Medicine. 2010 Jan;89(1):62–7.

101. Lanata CM, Paranjpe I, Nititham J, Taylor KE, Gianfrancesco M, Paranjpe M, et al. A phenotypic and genomics approach in a multi-ethnic cohort to subtype systemic lupus erythematosus. Nature Communications. 2019 Aug 29;10(1):3902.

102. To C. Prognostically distinct clinical patterns of systemic lupus erythematosus identified by cluster analysis. Lupus [Internet]. 2009 [cited 2020 Aug 7]; Available from: https://journals.sagepub.com/doi/pdf/10.1177/0961203309345767

103. A phenotypic and genomics approach in a multi-ethnic cohort to subtype systemic lupus erythematosus | Nature Communications [Internet]. [cited 2020 Aug 4]. Available from: https://www.nature.com/articles/s41467-019-11845-y

104. Update on Lupus Nephritis | American Society of Nephrology [Internet]. [cited 2020 Jul 8]. Available from: https://cjasn.asnjournals.org/content/12/5/825

105. Murphy G, Isenberg DA. New therapies for systemic lupus erythematosus — past imperfect, future tense. Nature Reviews Rheumatology. 2019 Jul;15(7):403–12.

106. Dörner T, Furie R. Novel paradigms in systemic lupus erythematosus. The Lancet. 2019 Jun 8;393(10188):2344–58.

107. Criteria [Internet]. [cited 2021 Feb 15]. Available from: https://www.rheumatology.org/Practice-Quality/Clinical-Support/Criteria

108. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Annals of the Rheumatic Diseases. 2019 Sep 1;78(9):1151–9.

109. Hoover PJ, Costenbader KH. Insights into the Epidemiology and Management of Lupus Nephritis from the U.S. Rheumatologist's Perspective. Kidney Int. 2016 Sep;90(3):487–92.

110. Moores KG, Sathe NA. A systematic review of validated methods for identifying systemic lupus erythematosus (SLE) using administrative or claims data. Vaccine. 2013 Dec 30;31:K62–73.

111. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. Lupus. 2010 May;19(6):741–3.

112. Li T, Lee I, Jayakumar D, Huang X, Xie Y, Eisen S, et al. Development and validation of lupus nephritis case definitions using United States veterans affairs electronic health records. Lupus. 2020 Nov 11;0961203320973267.

113. Enterprise Data Warehouse [Internet]. [cited 2020 Aug 25]. Available from: https://www.nucats.northwestern.edu/resources/data-science-and-informatics/nmedw/index.html

114. MetaMap - A Tool For Recognizing UMLS Concepts in Text [Internet]. [cited 2020 Aug 23]. Available from: https://metamap.nlm.nih.gov/

115. Unified Medical Language System (UMLS) [Internet]. U.S. National Library of Medicine; [cited 2020 Dec 8]. Available from: https://www.nlm.nih.gov/research/umls/index.html

116. sklearn.linear_model.Ridge — scikit-learn 0.23.2 documentation [Internet]. [cited 2020 Dec 11]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

117. Vanderbilt University Medical Center | [Internet]. [cited 2021 Feb 16]. Available from: https://www.vumc.org/main/home

118. Cox DR. Analysis of survival data. Vol. 21. CRC Press; 1984.

119. Cardiovascular diseases [Internet]. [cited 2021 Aug 17]. Available from: https://www.who.int/westernpacific/health-topics/cardiovascular-diseases

120. D'Agostino RB Sr, Grundy S, Sullivan LM, Wilson P, for the CHD Risk Prediction Group. Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. JAMA. 2001 Jul 11;286(2):180–7.

121. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. Circulation. 2014 Jun 24;129(25_suppl_2):S49–73.

122. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. Stat Med. 2011 May 10;30(10):1105–17.

123. Mogensen UB, Ishwaran H, Gerds T. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software [Internet]. 2012; Available from: http://www.jstatsoft.org/v50/i11/

124. Pennells L, Kaptoge S, Wood A, Sweeting M, Zhao X, White I, et al. Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. Eur Heart J. 2019 Feb 14;40(7):621–31.

125. survC1-package: C-Statistics for Risk Prediction Models with Censored... in survC1: C-Statistics for Risk Prediction Models with Censored Survival Data [Internet]. [cited 2021 Aug 17]. Available from: https://rdrr.io/cran/survC1/man/survC1-package.html

126. scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation [Internet]. [cited 2022 Mar 2]. Available from: https://scikit-learn.org/stable/

127. Joo G, Song Y, Im H, Park J. Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access. 2020;8:157643–53.

128. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Medical Research Methodology. 2018 Dec 29;18(1):179.

129. Panagiotakos DB, Fitzgerald AP, Pitsavos C, Pipilis A, Graham I, Stefanadis C. Statistical modelling of 10-year fatal cardiovascular disease risk in Greece: the HellenicSCORE (a calibration of the ESC SCORE project). Hellenic J Cardiol. 2007 Apr;48(2):55–63.

130. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. JAMA Cardiol. 2021 Jun 1;6(6):633–41.

131. Engelhard MM, Navar AM, Pencina MJ. Incremental Benefits of Machine Learning—When Do We Need a Better Mousetrap? JAMA Cardiology. 2021 Jun 1;6(6):621–3.

Appendices

Table S 1. Diagnosis codes for type 2 diabetes.

ICD 9 codes	**250.20, 250.30, 250.40, 250.50, 250.60, 250.70, 250.80, 250.90, 250.22, 250.32, 250.42, 250.52, 250.62, 250.72, 250.82, 250.92**
ICD 10 codes	**E11.9, E13.01, E13.9, E11.00, E11.01, E13.10, E13.00**
SNOMED codes	**44054006, 197763012, 474213016, 200951011, 78158011, 48EB2F20-59A4-4676-A1C0-40880362224F, 359642000, 81531005, 719216001, 237599002, 199230006, 237627000, 9859006, 190331003, 703138006, 314903002, 190390000, 314902007, 190389009, 313436004**

Table S 2. Diagnosis codes for study variables.

Covariates	**ICD 9 codes**	**ICD 10 codes**
CHF	402.x1, 404.x1, 404.x3, 428.xx	I11.0, I13.0, I13.2, I50.x, I09.81
Hypertension	401.x-405.x	I10-13, I15-I16
Hypoglycemic Events	250.30, 250.32, 251.0, 251.1, 251.2	E11.64x, E16.0, E16.1, E16.2
Other diabetic complications	Diabetic Nephropathy: 250.40, 250.42 Nephrotic Syndrome: 581.x Nephritis: 583.xx Diabetic oculopathy: 250.50, 250.52, 362.0x, 366.41 Diabetic Cataracts: 366.41 Lower Extremity Amputations: V49.7x Diabetic Retinopathy: 250.50, 250.52, 362.0x	Diabetic Nephropathy: E11.2x Nephrotic Syndrome: N04.x Nephritis: N05.x, N08 Diabetic oculopathy: E11.3x Diabetic Cataracts: E11.36 Lower Extremity Amputations: Z89.4xx, Z89.5xx, Z89.6xx Diabetic Retinopathy: E11.31 – E11.35
Diabetic neuropathy	250.60, 250.62, 357.2, 362.01-362.06	E11.4x
Dyslipidemia	272.0, 272.1, 272.2, 272.3, 272.4	E78.0 – E78.5, E78.00
Tobacco Use	305.1, V15.82	F17.2x, Z87.891

CVD	Stroke: 430, 431, 432.x, 433.x1, 434.x1, 435.x PAD: 440.x, 441.x, 443.2x, 444.x, 445.x IHD: 410.x, 411.x, 414.12	Stroke: I63.x, I60.x, I61.x, I62.x, I63.xxx, G45.0-G45.2, G45.8, G45.9 PAD: I70.x, I71.x, I74.x, I75.x, I77.7x IHD: I20.x, I21.x, I22.x, I23.x, I24.x, I25.42
Vascular Disease	440.x, 441.x, 442.x, 443.2x, 443.9, 444.x, 445.x	I70.x, I71.x, I72.x, I73.9, I74.x, I75.x, I77.7x
VCD	250.70, 250.72, 607.84, 707.1x, 707.8, 707.9	E11.5x, E11.62x, L97.xxx
Renal outcomes		
End stage renal disorder	585.6	N18.6
Chronic kidney disease	285.21, 403.x, 404.x, 582.x, 585.x, 586.x	D63.1, E11.22, I12.x, I13.x, N03.x, N18.x

Abbreviations: CHF, congestive heart failure; CVD, cardiovascular disease; VCD: vascular complications of diabetes.

Table S 3. Change over time in HbA1c value by second-line ADM groups.

Medication groups	Baseline	Year_1 (N=3130)	Year_2 (N = 3011)	Year_3 (N=2349)	Year_4 (N = 1781)	Year_5 (N = 986)
DPP-4i	7.3 (1.3)	7.39(1.32)	7.40(1.36)	7.37(1.34)	7.49 (1.49)	7.43 (1.43)
GLP-1RA	7.2 (1.6)	7.28 (1.5)	7.35 (1.69)	7.18 (1.43)	7.27(1.51)	7.17(1.51)
Insulin	7.9 (2.0)	7.93 (1.93)	7.85(1.80)	7.85(1.70)	7.80(1.74)	8.23(1.94)
SGLT-2i	7.3 (1.4)	7.47(1.39)	7.49(1.38)	7.43(1.41)	7.32(1.13)	7.62(1.77)
SU	7.5 (1.6)	7.56(1.45)	7.63(1.54)	7.70(1.56)	7.63(1.52)	7.62(1.44)
TZD	7.0 (1.1)	7.18(1.16)	7.17(1.47)	7.49(1.51)	7.33(1.23)	7.14(1.25)

This table shows the mean and standard deviation of HbA1c value by year and second-line ADM group from two years prior two index date till the end of the fifth year follow up. Baseline: first two years prior to index date. Abbreviations: ADM, anti-diabetic medication; HR, hazard ratio; CI, confidence interval; DPP-4i, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like

peptide receptor agonists; SGLT-2i, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 4. Hazard ratio for CKD incidence outcome in primary analysis.

Sequence group	HR	HR 95% CI	P-value
Baseline model			
Biguanides, DPP-4 inhibitor	0.64	[0.48;0.86]	0.00
Biguanides, GLP-1RA	0.28	[0.11;0.67]	0.00
Biguanides, Insulin	0.91	[0.55;1.50]	0.70
Biguanides, SGLT-2 inhibitor	0.31	[0.16;0.60]	<0.001
Biguanides, TZD	1.00	[0.49;2.04]	0.99
Basic demographics model			
Biguanides, DPP-4 inhibitor	0.73	[0.54;0.98]	0.03
Biguanides, GLP-1RA	0.45	[0.18;1.10]	0.08
Biguanides, Insulin	1	[0.60;1.68]	0.99
Biguanides, SGLT-2 inhibitor	0.43	[0.22;0.86]	0.02
Biguanides, TZD	1.06	[0.52;2.17]	0.87
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.70	[0.52;0.95]	0.02
Biguanides, GLP-1RA	0.40	[0.16;0.98]	0.05
Biguanides, Insulin	0.81	[0.48;1.37]	0.44
Biguanides, SGLT-2 inhibitor	0.43	[0.22;0.86]	0.02
Biguanides, TZD	0.98	[0.47;2.04]	0.96
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.71	[0.53;0.96]	0.03
Biguanides, GLP-1RA	0.52	[0.21;1.30]	0.16

Biguanides, Insulin	0.93	[0.55;1.59]	0.80
Biguanides, SGLT-2 inhibitor	0.43	[0.22;0.87]	0.02
Biguanides, TZD	1.03	[0.50;2.15]	0.93

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 5. Hazard ratio for CKD hospitalization outcome in primary analysis.

Variable	HR	HR 95% CI	Pval
Baseline model			
Biguanides, DPP-4 inhibitor	0.57	[0.36;0.89]	0.01
Biguanides, GLP-1RA	0.49	[0.18;1.34]	0.16
Biguanides, Insulin	0.97	[0.48;1.97]	0.94
Biguanides, SGLT-2 inhibitor	0.38	[0.15;0.95]	0.04
Biguanides, TZD	1.02	[0.37;2.82]	0.96
Basic demographics model			
Biguanides, DPP-4 inhibitor	0.69	[0.44;1.07]	0.10
Biguanides, GLP-1RA	0.99	[0.35;2.77]	0.99
Biguanides, Insulin	1.05	[0.52;2.15]	0.88
Biguanides, SGLT-2 inhibitor	0.70	[0.28;1.77]	0.46
Biguanides, TZD	1.06	[0.38;2.94]	0.91
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.56	[0.35;0.91]	0.02
Biguanides, GLP-1RA	0.96	[0.34;2.71]	0.93
Biguanides, Insulin	0.52	[0.24;1.14]	0.10
Biguanides, SGLT-2 inhibitor	0.85	[0.33;2.16]	0.73
Biguanides, TZD	1.27	[0.45;3.61]	0.66
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.60	[0.37;0.96]	0.03
Biguanides, GLP-1RA	1.05	[0.37;3.02]	0.92

Biguanides, Insulin	0.52	[0.24;1.17]	0.11
Biguanides, SGLT-2 inhibitor	0.81	[0.31;2.09]	0.66
Biguanides, TZD	1.25	[0.44;3.70]	0.65

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 6. Hazard ratio for eGFR < 45 mL/min outcome in primary analysis.

Variable	HR	HR 95% CI	Pval
Baseline model			
Biguanides, DPP-4 inhibitor	0.82	[0.62;1.09]	0.17
Biguanides, GLP-1RA	0.65	[0.35;1.20]	0.16
Biguanides, Insulin	1.22	[0.75;1.99]	0.41
Biguanides, SGLT-2 inhibitor	0.47	[0.26;0.83]	0.01
Biguanides, TZD	0.91	[0.42;1.94]	0.80
Basic demographics model			
Biguanides, DPP-4 inhibitor	0.92	[0.70;1.22]	0.57
Biguanides, GLP-1RA	1.00	[0.53;1.87]	0.99
Biguanides, Insulin	1.42	[0.87;2.31]	0.17
Biguanides, SGLT-2 inhibitor	0.66	[0.37;1.19]	0.17
Biguanides, TZD	0.97	[0.45;2.07]	0.93
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.92	[0.70;1.22]	0.58
Biguanides, GLP-1RA	0.89	[0.47;1.68]	0.72
Biguanides, Insulin	1.19	[0.72;1.97]	0.49
Biguanides, SGLT-2 inhibitor	0.64	[0.36;1.15]	0.14
Biguanides, TZD	0.95	[0.44;2.07]	0.90
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.95	[0.72;1.26]	0.73
Biguanides, GLP-1RA	0.87	[0.46;1.65]	0.68

Biguanides, Insulin	1.18	[0.71;1.95]	0.52
Biguanides, SGLT-2 inhibitor	0.58	[0.32;1.07]	0.08
Biguanides, TZD	0.96	[0.44;2.08]	0.91

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 7. Hazard ratio for CKD incidence outcome in sensitivity analysis.

Variable	HR	HR 95% CI	P-value
Baseline model			
Biguanides, DPP-4 inhibitor	0.68	[0.48;0.97]	0.03
Biguanides, GLP-1RA	0.35	[0.13;0.97]	0.04
Biguanides, Insulin	0.95	[0.51;1.78]	0.87
Biguanides, SGLT-2 inhibitor	0.35	[0.16;0.75]	0.01
Biguanides, TZD	1.19	[0.52;2.73]	0.68
Basic demographics			
Biguanides, DPP-4 inhibitor	0.76	[0.54;1.09]	0.14
Biguanides, GLP-1RA	0.53	[0.19;1.47]	0.22
Biguanides, Insulin	1.07	[0.57;2.04]	0.83
Biguanides, SGLT-2 inhibitor	0.46	[0.21;1.00]	0.05
Biguanides, TZD	1.13	[0.49;2.61]	0.77
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.73	[0.51;1.04]	0.09

Biguanides, GLP-1RA	0.47	[0.17;1.29]	0.14
Biguanides, Insulin	0.85	[0.44;1.64]	0.62
Biguanides, SGLT-2 inhibitor	0.46	[0.21;1.00]	0.05
Biguanides, TZD	1.05	[0.45;2.44]	0.92
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.75	[0.52;1.08]	0.12
Biguanides, GLP-1RA	0.53	[0.19;1.48]	0.23
Biguanides, Insulin	0.72	[0.37;1.43]	0.35
Biguanides, SGLT-2 inhibitor	0.42	[0.19;0.92]	0.03
Biguanides, TZD	1.15	[0.49;2.71]	0.75

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 8. Hazard ratio for CKD hospitalization outcome in sensitivity analysis.

Variable	HR	HR 95% CI	P-value
Baseline model			
Biguanides, DPP-4 inhibitor	0.69	[0.41;1.16]	0.16
Biguanides, GLP-1RA	0.63	[0.20;2.06]	0.45
Biguanides, Insulin	1.16	[0.49;2.74]	0.74
Biguanides, SGLT-2 inhibitor	0.59	[0.23;1.51]	0.28
Biguanides, TZD	0.83	[0.20;3.45]	0.80
Basic demographics			
Biguanides, DPP-4 inhibitor	0.85	[0.50;1.44]	0.55
Biguanides, GLP-1RA	1.17	[0.35;3.90]	0.80
Biguanides, Insulin	1.21	[0.50;2.93]	0.67
Biguanides, SGLT-2 inhibitor	1.10	[0.43;2.85]	0.84
Biguanides, TZD	0.79	[0.19;3.28]	0.74
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.65	[0.37;1.15]	0.14

Biguanides, GLP-1RA	1.03	[0.30;3.57]	0.96
Biguanides, Insulin	0.84	[0.33;2.14]	0.71
Biguanides, SGLT-2 inhibitor	1.30	[0.49;3.48]	0.60
Biguanides, TZD	0.92	[0.21;3.98]	0.92
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.75	[0.42;1.32]	0.32
Biguanides, GLP-1RA	0.94	[0.26;3.43]	0.92
Biguanides, Insulin	1.00	[0.39;2.57]	1.00
Biguanides, SGLT-2 inhibitor	1.35	[0.50;3.66]	0.56
Biguanides, TZD	0.90	[0.21;3.90]	0.89

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 9. Hazard ratio for eGFR < 45 mL/min outcome in sensitivity analysis.

Variable	HR	HR 95% CI	P-value
Baseline model			
Biguanides, DPP-4 inhibitor	0.79	[0.57;1.08]	0.14
Biguanides, GLP-1RA	0.76	[0.38;1.50]	0.43
Biguanides, Insulin	1.39	[0.81;2.41]	0.24
Biguanides, SGLT-2 inhibitor	0.48	[0.26;0.90]	0.02
Biguanides, TZD	1.05	[0.46;2.41]	0.90
Basic demographics			
Biguanides, DPP-4 inhibitor	0.89	[0.64;1.22]	0.46
Biguanides, GLP-1RA	1.09	[0.54;2.19]	0.81
Biguanides, Insulin	1.60	[0.92;2.79]	0.10
Biguanides, SGLT-2 inhibitor	0.70	[0.37;1.32]	0.27
Biguanides, TZD	1.02	[0.45;2.34]	0.96
Basic demographics/medical history model			
Biguanides, DPP-4 inhibitor	0.95	[0.69;1.31]	0.75

Biguanides, GLP-1RA	0.98	[0.48;1.98]	0.96
Biguanides, Insulin	1.44	[0.81;2.57]	0.21
Biguanides, SGLT-2 inhibitor	0.68	[0.36;1.28]	0.23
Biguanides, TZD	0.96	[0.41;2.23]	0.93
Fully adjusted model			
Biguanides, DPP-4 inhibitor	0.98	[0.71;1.36]	. 0.90
Biguanides, GLP-1RA	0.97	[0.48;1.96]	0.93
Biguanides, Insulin	1.49	[0.83;2.65]	0.18
Biguanides, SGLT-2 inhibitor	0.61	[0.31;1.18]	0.14
Biguanides, TZD	0.97	[0.42;2.27]	0.95

Abbreviations: HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 10. Number of patients who were on only two medications in each second-line ADM group.

Medication Groups	Overall number	Switched to 3rd	Did not switch
Biguanides, DPP-4i	1192	553	639
Biguanides, GLP-1RA	208	106	102
Biguanides, insulin	215	102	113
Biguanides, SGLT-2i	355	119	236
Biguanides, SU	1348	503	845
Biguanides, TZD	85	40	45

Abbreviations: DPP-4i, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2i, sodium-glucose cotransporter 2 inhibitor; Thiazolidinediones, TZD

Table S 11. Hazard ratio in the fully adjusted cox regression model among patients who took only two ADMs during the exposure period.

	CKD incidence (N=1869, E = 38)		CKD hospitalization (N=1913, E = 67)	
	HR (95% CI)	Pval	HR (95% CI)	Pval
DPP-4	0.75, [0.49;1.15]	0.19	0.69, [0.36;1.35]	0.28
GLP-1RA	1.08, [0.32;3.64]	0.90	1.50, [0.42;5.40]	0.53
Insulin	1.19, [0.56;2.51]	0.65	0.34, [0.10;1.10]	0.07
SGLT-2 inhibitor	0.49, [0.20;1.23]	0.13	0.91, [0.29;2.81]	0.86
TZD	1.00, [0.35;2.83]	0.99	0.90, [0.25;3.23]	0.87

In this analysis, missing data in covariates were imputed. Abbreviations: ADMs, anti-diabetic medications; HR, hazard ratio; CI, confidence interval; DPP-4 inhibitor, dipeptidyl peptidase 4 inhibitors; GLP-1RA, glucagon-like peptide receptor agonists; SGLT-2 inhibitor, sodium-glucose 2 inhibitor; Thiazolidinediones, TZD

Table S 12. Regex search for Nasal/oral ulcer, arthritis, renal disorder, and lupus nephritis.

SLICC criteria	**Concept**	**keywords**
Nasal/Oral ulcer	Nasal/Oral ulcer	Oral ulcer, nasal ulcer
Arthritis	Arthritis	Arthritis, synovitis
Proteinuria/red cell cast (renal disorder)	Proteinuria > 0.5 mg	Proteinuria > 0.5 mg
	Urine/creatinine ratio > 0.5 mg/mg	Urine/creatinine ratio > 0.5 mg/mg
	24-hour urine protein >0.5gm	24-hour urine protein > 0.5gm
	Red cell cast	Red cell cast
lupus nephritis (renal disorder)	Nephritis class II	Nephritis class II
	Nephritis class III	Nephritis class III
	Nephritis class IV	Nephritis class IV, mesangial proliferative GN
	Nephritis class V	Nephritis class V, membranous nephritis

Table S 13. CUIs and their definition.

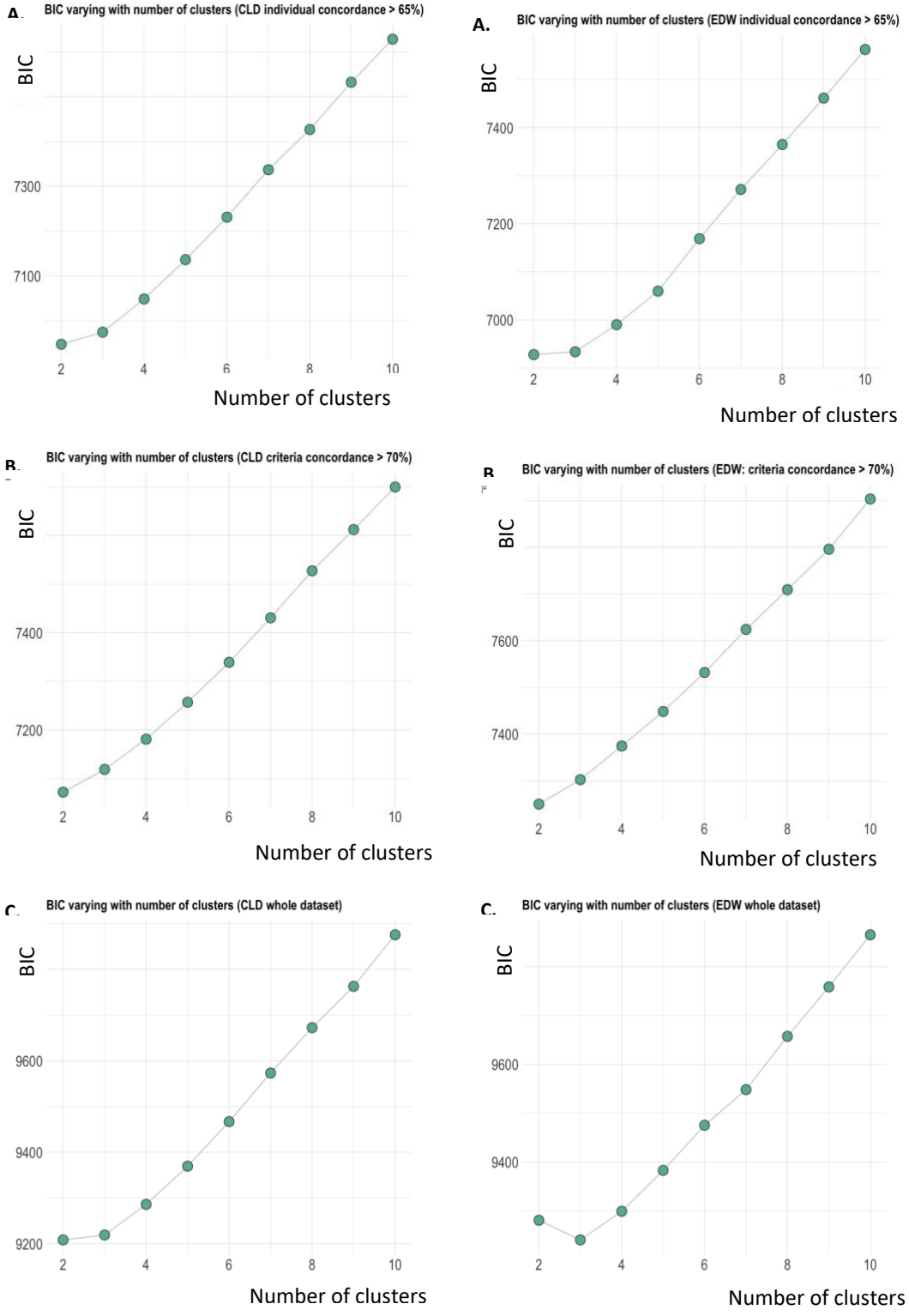
CUIs	Definition
C0024143	Glomerulonephritis in the context of systemic lupus erythematosus.
C0268757	Lupus nephritis - WHO Class IV
C0268758	Lupus nephritis - WHO Class V
C4053955	Systemic lupus erythematosus nephritis, with active or inactive diffuse, segmental or global endo- or extracapillary glomerulonephritis involving greater than or equal to 50% of all glomeruli, typically with diffuse subendothelial immune deposits, with or without mesangial alterations.
C4053958	Systemic lupus erythematosus nephritis exhibiting mesangial hypercellularity or mesangial expansion by light microscopy, with mesangial immune deposits. Isolated subepithelial or subendothelial deposits may be visible by immunofluorescence or electron microscopy, but not by light microscopy
C4053959	Systemic lupus erythematosus nephritis with active of inactive focal, segmental or global endo- or extracapillary
C4054543	Membranous nephritis associated with systemic lupus

Table S 14. C-statistics for PCEs, Nnet-survival, Deepsurv, Cox-nnet, and Cox PH-TWI in 10x10 cross-validation and MESA external validation.

		PCE	Cox PH-TWI	Deepsurv	Nnet-survival	Cox-nnet
White male	10x10 CV	0.7349	0.7349	**0.7371**	0.7281	0.7334
	MESA	0.6959 (-)	0.7004 (0.31)	**0.7032** (0.16)	0.7007 (0.53)	0.6990 (0.59)
White female	10x10 CV	0.7963	**0.7972**	**0.7972**	0.7849	0.7926
	MESA	0.7238 (-)	0.7276 (0.26)	0.7225 (0.67)	0.7187 (0.31)	**0.7282** (0.49)
Black male	10x10 CV	**0.6981**	0.6925	0.6790	0.6554	0.6745
	MESA	**0.6811** (-)	0.6772 (0.66)	0.6759 (0.79)	0.6672 (0.43)	0.6731 (0.53)
Black female	10x10 CV	0.7787	0.7884	**0.7886**	0.7774	0.7782
	MESA	0.7112 (-)	0.7173 (0.64)	**0.7316** (0.00)	0.7188 (0.21)	0.7100 (0.91)

C-statistics for all models and p-value (in the parentheses) for the difference of PCE models vs. other models. The highest C-statistics for each race and sex group are bolded. Abbreviations: PCE, Pooled Cohort Equation; Cox PH-TWI, all two way interaction Cox Proportional Hazard Model; CV, cross-validation; Cox PH, Cox Proportional Hazards model; MESA: Multi-Ethnic Study of Atherosclerosis.

Figure S 1. BIC varying with number of clusters.



Vita

Yu Deng

CONTACT INFORMATION
Feinberg School of Medicine, 303 E Superior street, Chicago, IL.
Email:yudeng2015@u.northwestern.edu
Mobil: 3124839916

RESEARCH INTEREST

I am a PhD student in biomedical informatics. My focus is on applying statistical learning methods for clinical research. I have applied various machine learning techniques to find disease subtypes and novel risk factors. I have used deep learning/joint modeling for disease prediction on time series data as well as using Natural Language Processing (NLP) to support computational phenotyping and chatbot.

EDUCATION

Northwestern University, Chicago, IL, USA
- PhD Candidate, biomedical informatics, Feinberg School of Medicine, expected graduation Feb 2022
- Advisor: Abel Kho
- Coursework: Programming for Big Data, Advanced Biostatistics, Statistical Theory and Method

INTERNSHIP

Applied Scientist Intern, Amazon, USA, 2021
- Alexa Conversation, Natural language understanding, Chatbot, sequence to sequence model.
- Developed data postprocessor to support complex user utterances in multiturn conversation.

Data Science Intern, National Institute of Health (NIH), Bethesda, MD, USA, 2019
- Analyzed full text chemical corpus for downstream analysis such as name entity recognition.
- Developed customized CNN architectures utilizing longitudinal data for cardiovascular disease prediction.

SKILLS

Programming Skills
- Python, R, SQL

Machine Learning Algorithms
- **Deep Learning** (MLP, CNN, RNN), Classical & Penalized Regression Methods (LASSO, Ridge), SVM, Random Forest, KNN, Adaboosting, Cox regression; K-means, hierarchical clustering, non-Negative Tensor Factorization

Statistical Analysis
- A/B testing, missing data imputation, survival analysis

SELECTED PROJECTS

Developed Deep Learning-survival Model for Cardiovascular Disease (CVD) Prediction, 2020
- Developed CNN based models to forecast cardiovascular disease using time series data. Our model tackled the problems of patient irregular visit patterns, subject drop-out, and correlation within repeated measurements.
- Designed 3 convolution layers with convolution kernels convolved with the layer input over the temporal dimension.
- Explore different parameter settings include learning rate, number of hidden layers, epoch, activation function, early stop.

Use of Clinical Phenotypes and Non-negative Tensor Factorization (NTF) for Heart Failure (HF) Prediction, 2017
- Performed NTF on large scale, sparse medical record data; Generated latent clusters.
- Performed dimension reduction including NTF, PCA on medical record data to get important features.
- Used the output of different feature reduction techniques as the input features of logistic regression.

Use of NLP to Improve Systemic Lupus Erythematosus Criteria Identification in Electronic Health Records, 2019
- Developed SQL queries to extract large scale clinical notes from hospital data warehouse.
- Used Metamap to extract name entities, used l1/l2 penalized logistic regression used for feature selection and prediction.
- Evaluated model performance using sensitivity and specificity. NLP based algorithm improved renal sensitivity from 0.4 to 0.7.

SELECTED PUBLICATIONS/CONFERENCES

Comparison of the State-of-Art Neural Network Survival Models vs Cox PH on Cardiovascular Disease Prediction
Deng Y., Peng Y., Wei Y., Lu Z., Zhao L. (2021). Manuscript under revision, BMC medical research methodology

Natural Language Processing to Identify Lupus Nephritis Phenotype in Electronic Health Records
Deng Y., Pacheco J., Walunas T., Luo Y. (2021). Manuscript under revision, BMC biomedical informatics and decision making

Association of Second-line Type 2 Diabetes Medication and Chronic Kidney Disease
Deng Y., Ghamsari F., Lu A., Yu J., Kho A. (2021). medRxiv 2021.06.15.21258963; doi: https://doi.org/10.1101/2021.06.15.212589 63

BERTSurv: BERT-Based Survival Models for Predicting Outcomes of Trauma Patients
Zhao Y., Hong Q., **Deng Y**., Wang Y., Petzold. (2021). IEEE international conference on Data Mining (ICDM).

Natural Language Processing for EHR-Based Computational Phenotyping
Zeng, Z., **Deng Y**., Li X., Naumann T., & Luo Y. (2018). Natura l Language Processing for EHR -Based Computational Phenotyping. IEEE/ACM transactions on computational biology and bioinformatics.

HONOURS/ DISTINCTIONS

Travel award, International Conference on Intelligent Biology and Medicine (ICIBM), 2021
Invited presentation, International Chinese Statistical Association (ICSA) Midwest Meeting, 2021
Honorable Mention in Student Poster Competition, International Chinese Statistical Association (ICSA), 2018
First Prize in Student Poster Competition, Northwestern Biomedical Informatics Day, 2017