Journal of Medical Systems (2019) 43: 185
https://doi.org/10.1007/s10916-019-1321-6

PATIENT FACING SYSTEMS

Towards more Accessible Precision Medicine: Building a more Transferable Machine Learning Model to Support Prognostic Decisions for Micro- and Macrovascular Complications of Type 2 Diabetes Mellitus

Era Kim, PhD[1,2] · Pedro J. Caraballo, MD[3,4] · M. Regina Castro, MD[5] · David S. Pieczkiewicz, PhD[1] · Gyorgy J. Simon, PhD[1,6]

Received: 23 October 2018 / Accepted: 1 May 2019 / Published online: 17 May 2019

Abstract

Although machine learning models are increasingly being developed for clinical decision support for patients with type 2 diabetes, the adoption of these models into clinical practice remains limited. Currently, machine learning (ML) models are being constructed on local healthcare systems and are validated internally with no expectation that they would validate externally and thus, are rarely transferrable to a different healthcare system. In this work, we aim to demonstrate that (1) even a complex ML model built on a national cohort can be transferred to two local healthcare systems, (2) while a model constructed on a local healthcare system's cohort is difficult to transfer; (3) we examine the impact of training cohort size on the transferability; and (4) we discuss criteria for external validity. We built a model using our previously published Multi-Task Learning-based methodology on a national cohort extracted from OptumLabs® Data Warehouse and transferred the model to two local healthcare systems (i.e., University of Minnesota Medical Center and Mayo Clinic) for external evaluation. The model remained valid when applied to the local patient populations and performed as well as locally constructed models (concordance: .73–.92), demonstrating transferability. The performance of the locally constructed models reduced substantially when applied to each other's healthcare system (concordance: .62–.90). We believe that our modeling approach, in which a model is learned from a national cohort and is externally validated, produces a transferable model, allowing patients at smaller healthcare systems to benefit from precision medicine.

Keywords Machine learning · Large national data · External validation · Transferable model · Complications of type 2 diabetes · Precision medicine

This article is part of the Topical Collection on *Patient Facing Systems*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10916-019-1321-6) contains supplementary material, which is available to authorized users.

✉ Era Kim, PhD
kimx2757@umn.edu

[1] Institute for Health Informatics, University of Minnesota, 8-100 Phillips Wangensteen Building, 516 Delaware St. SE, Minneapolis, MN 55455, USA

[2] OptumLabs Visiting Fellow, Cambridge, MA, USA

[3] Division of General Internal Medicine. Department of Medicine, Mayo Clinic, Rochester, MN, USA

[4] Center for Translational Informatics and Knowledge Management, Mayo Clinic, Rochester, MN, USA

[5] Division of Endocrinology and Metabolism, Department of Medicine, Mayo Clinic, Rochester, MN, USA

[6] Department of Medicine, University of Minnesota, Minneapolis, MN, USA

Introduction

Machine learning (ML) holds great promise for precision medicine, providing a new way to generate evidence that enhances clinical practice guidelines with more personalized recommendations [1]. One disease that would immensely benefit from this approach is Type 2 Diabetes Mellitus (T2DM), a complex chronic disease that requires multifactorial risk-reduction strategies to prevent and manage clinically significant micro- and macrovascular complications [2]. In recent years, ML models have been increasingly developed for clinical decision support for T2DM patients, yet the adoption of these models into clinical practice remains limited [3–6].

ML models are developed on a *training* cohort and are expected to be applicable to the population from which the training cohort was drawn (*training population*). *Internal validation* ensures that the model generalizes from the training cohort to the training population. If the model performs well on the training population, the model is said to *generalize* to that population; if it does not, the model is said to *overfit* the training cohort. Most existing ML models are internally validated; they generalize well to the patient population of their healthcare system. However, when these models encounter patients who are atypical for their healthcare systems but may be common in other parts of the nation, they will invariably fail. Uncertainty around whether a model works for a particular patient erodes the applicability of the model to clinical decision support, hindering the adoption of ML models into clinical practice.

In observational studies, *external validation* is considered the strongest evidence of the generalizability of a model. A model is said to be *transferable* to a different healthcare system, if the model generalizes to the patients in the target healthcare system. Thus, external validation ensures not only generalizability but also transferability. If models were transferable, they would reliably apply to a much greater variety of patients, reducing the uncertainty. Today's ML models are rarely transferable. Even, the simplest decision-tree-based models require retraining to be applied to a different healthcare system [7]. Given the expectation that complex ML models will not be transferable, they are rarely externally validated [8–10].

A great amount of effort has been made to establish interoperability standards such as OMOP, PCORnet, and i2b2 for large-scale, long-term studies, in which models are developed and tested across multiple sites [11]. These studies are predominantly epidemiological and health service studies [12–16], while ML studies are relatively few.

In this paper, we aim to demonstrate that (1) even a complex ML model built on a national cohort can be transferred to two local healthcare systems, (2) while a model constructed on a local healthcare system's cohort is difficult to transfer to a different healthcare system; (3) we examine the impact of training cohort size on the transferability; and (4) we discuss criteria for external validity.

Methods

Datasets

We used three datasets. The first dataset was a large national dataset, containing 10-year claims and EHR data (Jan 1, 2006 - Dec 31, 2015) from the OptumLabs Data Warehouse (OLDW) [17, 18], a database which included retrospective administrative claims data on commercially insured and Medicare Advantage enrollees with linked EHR data from a nationwide network of provider groups ($N = 951{,}793$ patients diagnosed with T2DM before Jan 1, 2011, an index date). The second and third datasets were from two local healthcare systems. The second dataset was 9.5-year EHR data (Jan 1, 2008 - Jun 30, 2016) from University of Minnesota Medical Center (UMMC) in Minneapolis, MN ($N = 12{,}797$ patients diagnosed with T2DM before Jun 1, 2011). The third dataset was 8-year EHR data (Jan 1, 2007 - Dec 31, 2014) from Mayo Clinic, Rochester (MCR), MN ($N = 5479$ patients diagnosed with T2DM before Jan 1, 2010). All three datasets contained patients' demographics, smoking status, diagnoses, lab results, vital signs including BMI, and prescription drug information.

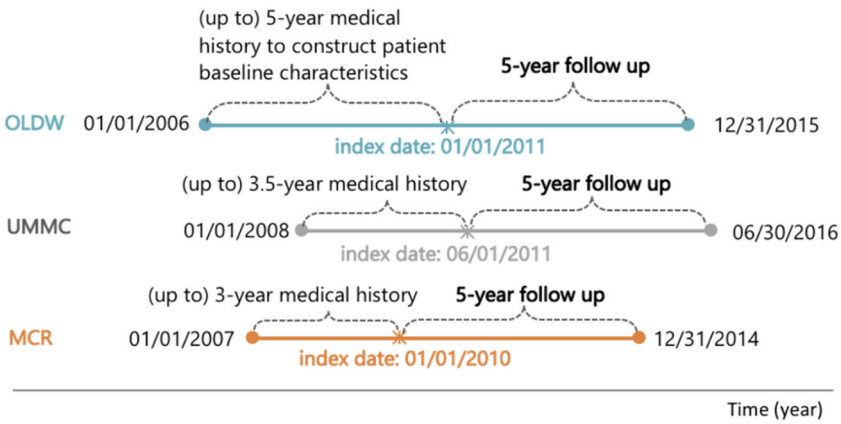

Fig. 1 Study design

Table 1 Cohort selection criteria (OLDW cohort)

Description	Patient (N)	(%)
Inclusion		
Patients w/ at least two encounters w/ diagnosis or history of T2DM during the 2-years before the index date (Jan 1, 2011)	915,739	100
Exclusion		
Patients w/ ICD-9 codes for T1DM, gestational diabetes, secondary diabetes, poisoning by hormones and synthetic substitutes, or other specified disorder of pancreatic internal secretion (source: HEDIS 2015 Diabetes Exclusions Value Set).	−208,938 = 706,801	77
Patients w/ multiple DOBs or sex.	−22,037 = 684,764	75
Patients whose age < 18 as of the index date	−178,581 = 506,183	55
Patients w/o any record since the index date.	−36,359 = 469,824	51
Patients w/o any HbA1c measurements at all before the index date[a].	−129,813 = 340,011	37
Patients w/o any SBP and DBP measurements at all before the index date[b].	−28,795 = 311,216	34
Patients w/o any pulse measurements at all before the index date.	−65,150 = 246,066	27
Patients w/o any BMI measurements at all before the index date.	−41,920 = 204,146	22
Patients w/o any known smoking status before the index date.	−30,576 = 173,570	19
Patients w/ ICD-9 codes for hyperlipidemia but no LDL, HDL, and Triglycerides measurements at all before the index date.	−8,504 = 165,066	18
Patients w/ ICD-9 codes for chronic kidney disease or chronic renal failure but no creatinine and GFR measurement at all before the index date.		
Patients w/ indication of HTN (i.e. ICD-9 codes for HTN or abnormal SBP and DBP) but no drug records at all before the index date.	−8,664 = 156,402	17
Patients w/ indication of HLD (i.e., ICD9 codes for HLD or abnormal lipids) but no drug records at all before the index date.		
Patients w/ shorter than 4-year follow-up[c].	−31,075 = 125,327	14
Patients whose minimum gap of two adjacent HbA1c measurements was larger than 3-years before the index date[d].	−44,236 = 81,091	9
Patients whose maximum gap of two adjacent HbA1c measurements was less than 1-year before the index date[d].		

Abbreviations: T2DM, Type 2 Diabetes Mellitus; T1DM, Type 1 Diabetes Mellitus; DOB, date of birth; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; BMI, Body Mass Index; LDL, Low Density Lipoprotein; HDL, High Density Lipoprotein; GFR, Glomerular Filtration Rate; HTN, Hypertension; HLD, Hyperlipidemia;

[a] If patients are diabetic, their HbA1c should be measured regularly [19]

[b] If patients are diabetic, their blood pressure should be measured at every routine visit

[c] Although survival analysis effectively handles the censoring of patients, unequal follow-up among different complications may produce biased results. Also, since complications of T2DM develop over many years, we preferred observing a patient for at least 4 years. We found that the excluded 31,075 patients had more pre-existing complications, worse lab results, and worse vital signs (long-standing T2DM and/or less controlled health conditions) than the remaining 126,327 patients

[d] In an effort to identify active primary care patients, we loosely applied recommendations for the HbA1c test [19]. Specifically, we considered a diabetic patient as a primary care patient if he had at least two HbA1c tests with a maximum gap of 3-years, and a minimum gap of 1-year during the baseline period. The excluded 44,236 patients were not different from the remaining 81,091 patients in terms of mean baseline characteristics and outcome rates. This exclusion only reduced the variance of baseline characteristics

Study design and cohort selection

This study was a retrospective cohort study (Fig. 1). We established baseline characteristics of the OLDW cohort using 5 years of medical history before the index date. From the index date, we followed them for 5 years, determining their time-to-event outcomes. When patients developed multiple complications during the follow-up period, we censored them after their first events occurred so that we focused on the most likely complication that individual patients could develop next.

With the cohort selection criteria (Table 1), we identified 81,091 (74,551 when fasting plasma glucose (FPG) was used instead of HbA1c) OLDW, 8091 UMMC, and 2247 MCR primary care patients with T2DM.

Table 2 ICD9-codes used to identify six micro- and macro complications of T2DM

Outcome	ICD-9 Code
Ischemic Heart Disease (IHD)	410.*, 411.*, 412, 413.*, 414.0x, 414.2, 414.3, 414.4, 414.8, 414.9, V45.81, V45.82
Congestive Heart Failure (CHF)	428.0, 428.1, 428.2, 428.3, 428.4, 428.9
Cerebrovascular Disease (CVD)	431, 432.9, 436, 437.0, 437.1, 437.8, 437.9, V12.54, 433.*, 434.*, 435.*, 438.*
Peripheral Vascular Disease (PVD)	440.2x, 440.4, 443.9, 445.89, 444.2x, 444.8x, 445.0x, 557.*
Chronic Kidney Disease (CKD)	585.1, 585.2, 585.3, 585.9 (CKD stage 1–3)
Chronic Renal Failure (CRF)	585.4, 585.5, 585.6, 586 (CKD stage 4–5, End stage renal disease, Renal failure)

Outcomes

We studied six time-to-event outcomes, micro- and macrovascular complications of T2DM, determined by billing diagnoses (Table 2). Although CKD and CRF are generally considered the same clinical entity, the patient could take many years to progress from the earlier to the more advanced stages of CKD, thus we defined them as separate outcomes.

Overview of model development and validation

Available variables and measurement methods could differ across healthcare systems. In our study, HbA1c was only available in UMMC, while FPG was only available in MCR. To externally validate our model on these two local datasets, we built four *model variants* based on a dataset and measurement method: $\text{variant}_{\text{OL.A1c}}$, $\text{variant}_{\text{OL.FPG}}$, $\text{variant}_{\text{UMMC}}$, and $\text{variant}_{\text{MCR}}$ (Fig. 2). Specifically, $\text{variant}_{\text{OL.A1c}}$ was built on the OLDW dataset with HbA1c included, while $\text{variant}_{\text{OL.FPG}}$ used FPG instead of HbA1c. $\text{Variant}_{\text{UMMC}}$ used HbA1c, and $\text{variant}_{\text{MCR}}$ used FPG.

$\text{Variant}_{\text{OL.A1c}}$ and $\text{variant}_{\text{OL.FPG}}$ were built on 80% of the national OLDW cohort, internally validated on the remaining 20% of the cohort, and externally validated on the entire UMMC and MCR cohort, respectively. $\text{Variant}_{\text{UMMC}}$ was built on 80% of the local UMMC cohort, internally validated on the remaining 20%, and externally validated on the entire MCR cohort. Likewise, $\text{variant}_{\text{MCR}}$ was built on 80% of local MCR cohort, internally validated on the remaining 20%, and externally validated on the entire UMMC cohort. We used the concordance index (C-index) [20] as a performance measure and followed [21] to convert between HbA1c and FPG.

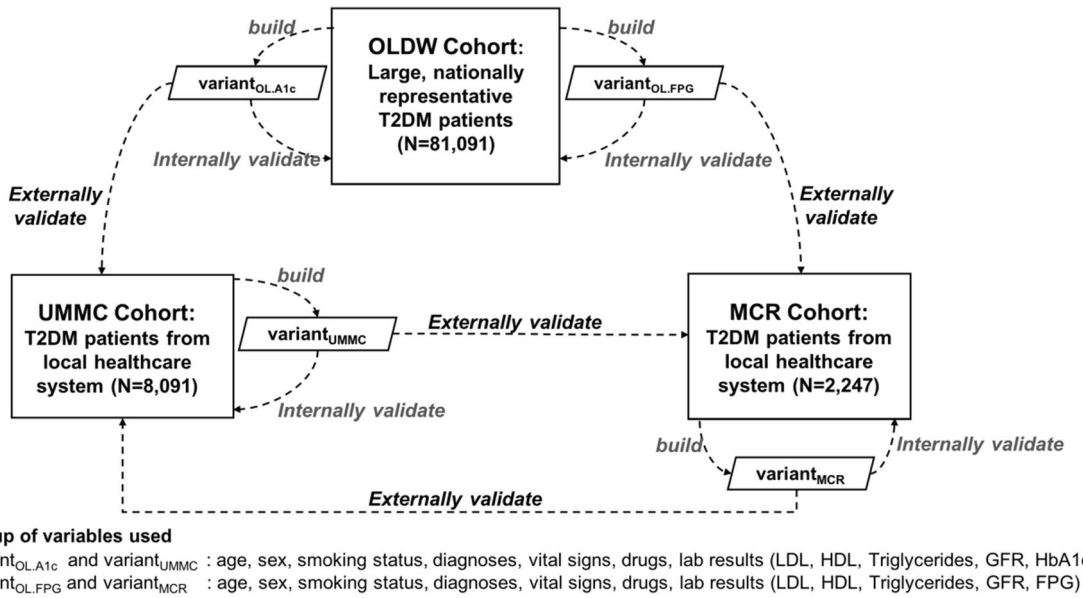

Fig. 2 Overview of model development and validation

Internal validation

In internal validation, we calculated a C-index for each model variant and outcome pair. To construct 95% confidence intervals (95CIs) of the C-index, we performed bootstrapping with 500 iterations. In each iteration, we built a model variant on a bootstrap sample and calculated a C-index for each of the outcomes on an Out-of-bag (OOB) sample. Using C-indices from the 500 OOB samples, we finally constructed 95CIs for each outcome. Internal validation was considered successful if the C-index from the internal validation fell into its 95CIs.

External validation

We examined whether a model variant was validated on a different (external) cohort. External validation was considered successful if a model variant achieved a non-random C-index (> .5), and the C-index fell into the 95CIs, explained in 2.5.

Modeling method

The model variants were developed using Multi-Task Learning (MTL)-based methodology [22], which learned six outcomes simultaneously by extracting a variance component that was common across all outcomes and isolating variance components that were specific to individual outcomes. These common and variant components referred to as *General Progression* model and *Differential Progression* models, respectively. We deliberately chose a complex model with a latent outcome (General Progression) to demonstrate that even a complex ML model could be externally validated.

Measuring transferability of a model from a large to mid-size training cohort

To understand the impact of training cohort size (N) on the transferability, we examined how the performance of $\text{variant}_{\text{OL.A1c}}$ and $\text{variant}_{\text{OL.FPG}}$ on the UMMC and MCR cohort changed as N was gradually reduced from 81,091 (OLDW) to 8091 (UMMC) and 2247 (MCR). For each N, we created 200 bootstrap samples from the entire OLDW cohort, built a $\text{variant}_{\text{OL.A1c}}$ ($\text{variant}_{\text{OL.FPG}}$) on each sample, and externally validated them on the UMMC (MCR) cohort. This gave us 200 C-indices allowing us to construct the 95CIs. To demonstrate that our findings are not specific to MTL, we also built LASSO (Least Absolute Shrinkage and Selection Operator) and GBM (Gradient Boosting Machines) models and compared our MTL with these other ML algorithms.

Results

Baseline characteristics of the OLDW, UMMC, and MCR cohorts

Table 3 shows baseline characteristics of the OLDW, UMMC, and MCR cohorts. Univariate cohort differences were tested using ANOVA for continuous variables and Chi-square tests for categorical variables. Although *p* values were adjusted by the Bonferroni correction, significant differences among the cohorts were found in most variables.

Cumulative hazard curves

Figure 3 presents the cumulative probabilities for developing each of the complications in the OLDW (median follow-up: 4.95-year), UMMC (median follow-up:4.83-year), and MCR (median follow-up:4.81) cohorts. Regarding major differences from the OLDW cohort, the UMMC cohort showed higher incidence of CKD and IHD; and the MCR cohort showed higher incidence of CKD, PVD, and CHF.

Description of the model

Table 4 presents coefficients in log hazard ratio (HR) resulting from $\text{variant}_{\text{OL.A1c}}$. The first column corresponds to the General Progression model and the remaining six columns correspond to Differential Progression models. Significant coefficients (Bonferroni corrected $P \leq .001$) are marked with an asterisk '*'. To show how to interpret them, let us consider HbA1c. higher levels of HbA1c increases the risk of progressing to any complication by a HR of 1.08(=exp.(0.0740)) (Table 4, row 4, col. 1), so it is harmful. It further increases the risk of progressing to CKD by a HR of 1.03(=exp.(0.0360)) (Table 4, row 4, col. 6), therefore we consider it to be more important in progression to CKD than the other complications. Most effect directions coincide with what is reported in the RCTs and prospective cohort studies [23–28]. A more detailed explanation about how to interpret the coefficients is found in our previous work [22].

Consistency of the effect directions of significant differential markers among the variants

We examined whether the effect directions of modifiable clinical variables were consistent among all the variants. Since differential effects were of our major interest, we visualized only coefficients in Differential Progression models (i.e., Differential Markers) (Fig. 4). Significant coefficients (Bonferroni corrected $P \leq .001$) are in colored cells where the orange and blue color indicates a significant positive or negative coefficient, respectively. We

Table 3 Baseline characteristics of the OLDW, UMMC, and MCR cohorts

Variable	OLDW ($N = 81,091$)	UMMC ($N = 8091$)	MCR ($N = 2247$)	P value
Age (year)	60.4 ± 9.7[a]	62.6 ± 11.3	61.9 ± 12.9	<.001*
Male (%)	48	49	51	.006*
Census Region (%)				
South (reference)	43			N/A
Midwest	41	100	100	N/A
Northeast	6			N/A
West	8			N/A
Unknown	2			N/A
Smoking Status (%)				
Never Smoker	22	43	12	<.001*
Former Smoker	61	25	71	<.001*
Current Smoker	17	32	17	<.001*
Lab Results				
HbA_{1c} (%)	7.1 ± 1.2	6.9 ± 1.1	N/A	1.000*
FPG (mg/dl)	116.34 ± 43.0	N/A	131.9 ± 19.0	<.001*
LDL (mg/dl)	96.4 ± 29.2	96.6 ± 28.7	103.9 ± 23.9	0.145*
HDL (mg/dl)	44.7 ± 12.4	42.9 ± 12.3	43.9 ± 10.6	<.001*
Triglycerides (mg/dl)	173.7 ± 118.2	178.7 ± 120.5	199.3 ± 98.5	<.001*
Missing LDL	0.6	1.4	0.0	<.001*
Missing HDL	0.6	1.5	0.0	<.001*
Missing Triglycerides	0.6	1.3	0.0	<.001*
GFR (ml/min/1.73m^2)	93.1 ± 20.6	74.1 ± 16.5	54.8 ± 14.3	<.001*
Normal GFR (%)[b]	0.4	0.4	4	<.001*
Vital Signs				
BMI (kg/m^2)	34.7 ± 7.5	34.0 ± 7.0	34.1 ± 7.0	<.001*
SBP (mmHg)	130.1 ± 11.3	127.1 ± 10.3	133.4 ± 12.4	<.001*
DBP (mmHg)	75.8 ± 7.2	73.8 ± 6.8	73.4 ± 8.6	<.001*
Pulse (bpm)	76.7 ± 9.0	75.7 ± 8.8	77.7 ± 9.7	<.001*
Pre-Existing Complications[c]				
IHD	22	22	37	<.001*
CHF	7	8	5	<.001*
CVD	11	10	14	<.001*
PVD	8	7	20	<.001*
CKD	9	15	18	<.001*
CRF	3	2	8	<.001*
Severity of HTN, HLD, and DM (%)				
HTN				
No indication of HTN (No Dx, normal SBP and DBP, and no HTN drug)	7	8	8	.008*
Untreated HTN (presence of Dx, or abnormal SBP or abnormal DBP but HTN drug)	5	5	5	.655*
At most two HTN drugs	49	44	58	<.001*
At least three HTN drugs and controlled HTN (normal SBP and DBP)	28	36	17	<.001*
At least three HTN drugs but uncontrolled HTN (abnormal SBP or DBP)	11	7	12	<.001*
HLD				
Cholesterol drug use	78	87	83	<.001*
DM				
Metformin only	33	41	25	<.001*
Monotherapy except Metformin or combination therapy without insulin	37	32	35	<.001*
Insulin	30	26	40	<.001*
Drug (%)				
Aspirin	29	38	NA	
HTN Drug				
ACEI or ARB	76	76	79	.009*
Diuretic	55	57	13	<.001*
Beta Blocker	43	53	54	<.001*
Calcium Channel Blocker	29	29	31	.062*
Other	6	5	3	<.001*
Cholesterol Drug				
Statin	76	86	79	<.001*
Fibrate	15	18	17	<.001*
Other	7	24	15	<.001*

Table 3 (continued)

Variable	OLDW ($N = 81,091$)	UMMC ($N = 8091$)	MCR ($N = 2247$)	P value
Diabetes Drug				
Metformin	72	69	71	<.001*
Alpha-Glucosidase Inhibitor	0.4	0.4	0	.408*
Amylin	0.4	0.3	0	.750*
Meglitinide	1	2	2	.001*
Incretin	7	7	4	<.001*
Sulfonylurea	41	41	52	<.001*
Thiazolidinedione	25	26	21	<.001*
DDP4 Inhibitor	14	5	5	<.001*
Insulin	30	26	39	<.001*
Other	2	1	1	<.001*
Drug Missingness (%)				
No HTN, Cholesterol and DM drug info	0.1	0.1	0.1	.062*

Abbreviations: FPG, Fasting Plasma Glucose; LDL, Low Density Lipoprotein; HDL, High Density Lipoprotein; GFR, Glomerular Filtration Rate; BMI, Body Mass Index; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; IHD, Ischemic Heart Disease; CHF, Congestive Heart Failure; CVD, Cerebrovascular Disease; PVD, Peripheral Vascular Disease Chronic Kidney Disease; CRF, Chronic Renal Failure; HTN, Hypertension; HLD, Hyperlipidemia; DM, Diabetes; Dx, Diagnosis; ACEI, Angiotensin-Converting Enzyme Inhibitor; ARB, Angiotensin Receptor Blocker; DDP4, Dipeptidyl Peptidase 4

Significant coefficients (Bonferroni corrected $P \leq .001$) are marked with an asterisk '*'

[a] Data are presented as mean ± standard deviation unless otherwise indicated

[b] N/A indicates not applicable

[c] This indicates % patients with no Dx of CKD, creatinine or GFR measurements at all prior to index date. We considered them to have normal kidney function. We imputed zero for these patients' GFR

[d] When patients had already presented with outcome complication(s) on the index date, we called them pre-existing complications

found the effect directions of significant Differential Markers to be consistent all the variants except the BMI and CRF pair.

Performance evaluation

Table 5 shows performance of the model variants. Rows 1–4 present C-indices from internal validation with 95CIs in parentheses, and rows 5–8 present C-indices from external validation.

In internal validation, national and local model variants showed very similar predictive performance (.73 to .92) (Table 5, row 1–4). All the C-indices fell into their 95CIs, indicating overfitting did not occur. That is, all the variants were successful in internal validation. National model variants had much narrower 95CIs, which is expected given the large training cohort size.

In external validation, all variants achieved non-random performance. National model variants still performed well on the local cohorts (.73–.92) (Table 5, row 5,7); in fact, they performed as well as local model variants on their own populations (Table 5, row 3,4). In contrast, the performance of local model variants reduced substantially when they were applied to each other's healthcare system (.62–.90) (Table 5, row 6,8).

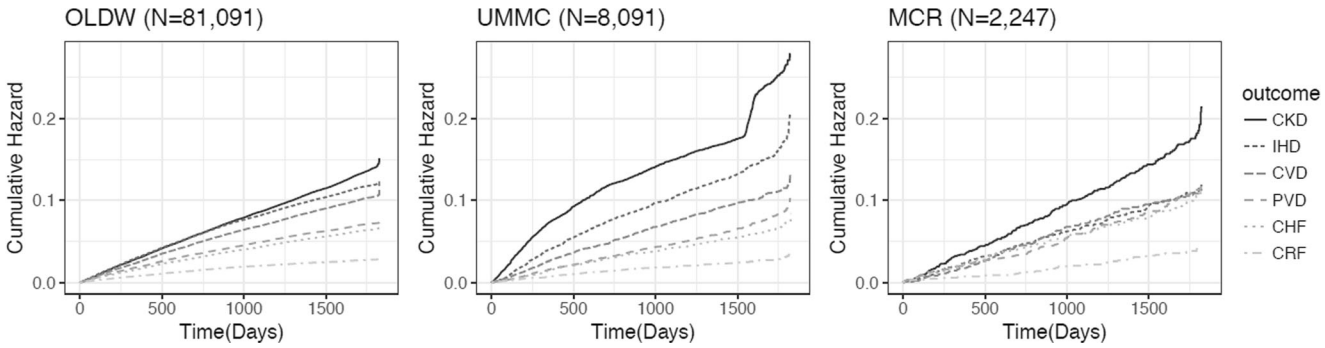

Fig. 3 Kaplan-Meier curves of the Nelson-Aalen Estimator for the OLDW, UMMC, and MCR cohort

Table 4 Coefficients in log hazard ratio from variant$_{OL.A1c}$

Variable	General and Differential Progression models						
	Gen	IHD	CHF	CVD	PVD	CKD	CRF
Low-Density Lipoprotein (LDL)	0.0010*	0.0020*	0.0000	0.0020*	0.0000	−0.0030*	0.0000
High-Density Lipoprotein (HDL)	−0.0060*	−0.0090*	−0.0010	−0.0010	0.0000	0.0070*	0.0060*
Triglycerides (Trigl)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HbA1c	0.0740*	−0.0200	0.0100	−0.0540*	0.0220	0.0360*	0.0090
Glomerular Filtration Rate (GFR)	−0.0200*	0.0170*	0.0090*	0.0170*	0.0160*	−0.0270*	−0.0280*
Normal GFR	−2.0910*	1.8460*	0.9550*	1.4120*	1.1840*	−2.8310*	−1.6150*
Body Mass Index (BMI)	−0.0070*	0.0060*	0.0490*	−0.0110*	0.0060*	0.0070*	−0.0020
Systolic Blood Pressure (SBP)	0.0100*	−0.0030	0.0000	0.0010	0.0030	0.0000	0.0060*
Diastolic Blood Pressure (DBP)	−0.0140*	0.0030	0.0000	0.0080*	−0.0110*	0.0000	−0.0130*
Pulse	0.0050*	−0.0060*	0.0110*	−0.0020	0.0040*	−0.0010	0.0070*
Former Smoker	0.1630*	0.1210*	0.0000	−0.0270	0.0330	−0.0310	−0.0180
Current Smoker	0.2290*	−0.1190*	0.1480*	0.0000	0.3040*	−0.1220*	−0.0100
Age	0.0270*	0.0080*	0.0210*	0.0190*	0.0040	−0.0190*	−0.0440*
Male	0.2430*	0.0410	−0.0380	−0.3720*	−0.0790	0.1760*	−0.0440
Midwest	−0.0750*	−0.1310*	0.0780*	−0.1050*	−0.2240*	0.2490*	−0.0730*
Northeast	−0.0980*	−0.0380	−0.0890	−0.0020	−0.2430*	0.1130	0.1570*
West	−0.2770*	−0.5000*	0.1370*	−0.2400*	−0.4060*	0.6550*	−0.2000*
Unknown Region	−0.0550	−0.0800	0.1180	−0.0840	−0.0760	0.0000	−0.1730
Aspirin Use	0.1190*	0.1170*	0.0360	−0.0050	−0.0250	−0.0670*	−0.0250
Untreated Hypertension (HTN)	0.2630*	−0.0310	0.0770	−0.1560	0.1400	0.1370	0.0000
HTN Mono Therapy	0.1000*	−0.0810	0.0490	0.0000	−0.1880*	0.1250*	0.5070*
HTN Combination Therapy, Controlled HTN	0.2100*	−0.0570	0.2000*	−0.0530	−0.0390	0.0010	−0.0420
HTN Combination Therapy, Uncontrolled HTN	0.0620*	−0.0640	−0.0050	−0.0370	0.0000	0.0690	0.1560*
Cholesterol (Chol) Medication Use	−0.0210*	0.0920*	−0.1850*	0.0000	−0.0370	0.0280	−0.1660*
Diabetes Combination Therapy	0.1030*	0.0010	−0.0320	0.0000	−0.0360	0.0130	0.0110
Diabetes Insulin Therapy	0.1610*	0.0040	0.1330*	0.0000	0.0650	−0.1270*	0.0000
Normal LDL,No Hyperlipidemia (HLD) Dx,No Chol Drug	0.0000	−0.0850	−0.5610	0.0000	0.2860	−0.2640	−0.4220
Normal HDL,No HLD Dx,No Chol Drug	0.0000	−0.2170	0.7690	0.0000	0.0000	0.0000	0.5870
Normal Trigl,No HLD Dx,No Chol Drug	0.1710	0.1890*	0.2080	0.0250	−0.3140	0.0000	−0.2260
Pre-existing Ischemic Heart Disease	−1.6650*		2.5150*	2.1850*	2.1880*	1.7850*	1.6730*
Pre-existing Congestive Heart Failure	6.2490*	−4.9330*		−5.9320*	−5.8640*	−5.8230*	−5.6370*
Pre-existing Cerebrovascular Disease	−0.0390	0.3770*	0.2680*		0.6360*	0.1250	0.2110*
Pre-existing Peripheral Vascular Disease	0.0460	0.3610*	0.3080*	0.3580*		0.1800*	0.2400*
Pre-existing Chronic Kidney Disease	−0.5160*	0.4260*	0.5450*	0.5820*	0.5230*		1.3870*
Pre-existing Chronic Renal Failure	−0.9730*	1.0880*	1.2410*	0.9840*	1.2700*	−3.8070*	

Significant coefficients (Bonferroni corrected $P \leq .001$) are marked with an asterisk '*'

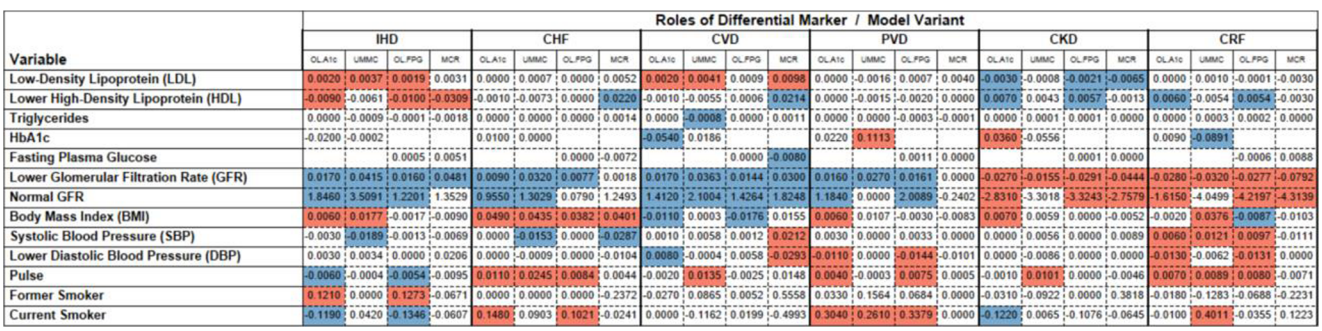

Fig. 4 Consistency of the effect directions of significant Differential Markers among the variants

Table 5 Performance of national ($variant_{OL.A1c}$, $variant_{OL.FPG}$) and local ($variant_{UMMC}$, $variant_{MCR}$) model variants

No	Validation type	Validation cohort	Model variant	IHD	CHF	CVD	PVD	CKD	CRF
1	Internal	OLDW	$variant_{OL.A1c}$	.74 (.73–.87)	.81 (.73–.87)	.74 (.73–87)	.75 (.73–87)	.80 (.73–87)	.86 (.73–.87)
2			$variant_{OL.FPG}$	.75 (.73–.87)	.81 (.73–.87)	.74 (.73–.87)	.74 (.73–.87)	.81 (.73–.87)	.87 (.73–.87)
3		UMMC	$variant_{UMMC}$	.73 (.71–.94)	.80 (.71–.94)	.75 (.71–.94)	.73 (.71–.94)	.80 (.71–.94)	.92 (.71–.94)
4		MCR	$variant_{MCR}$	.78 (.70–.94)	.82 (.68–.95)	.73 (.68–.94)	.75 (.70–.94)	.76 (.70–.94)	.90 (.70–.95)
5	External	UMMC	$variant_{OL.A1c}$	.73	.76	.75	.75	.79	.92
6			$variant_{MCR}$	.62	.73	.67	.70	.74	.90
7		MCR	$variant_{OL.FPG}$	.73	.76	.73	.74	.75	.92
8			$variant_{UMMC}$	.69	.72	.67	.69	.67	.90

Evaluation of training cohort size on transferability

Table 6 presents transferability from a small to a large training cohort size (N) of the three ML models. When N was 2247, the incidence of an outcome in a bootstrap sample was occasionally too low, resulting in training failure. With this smallest N, we were not able to evaluate the transferability of the GBM model; we were barely able to train MTL and LASSO models without CRF, which had the lowest incidence among the outcomes (2.3%).

Generally, transferability increases as N increases. This trend is not specific to MTL but repeats with LASSO and GBM models as well. Regardless of types of models, transferability reaches a steady state at the size of 30 K. Plots for the transferability are given in Online Resource 1(validation cohort: UMMC) and Online Resource 2 (validation cohort: MCR).

Table 7 (essentially a copy of appropriate rows from Tables 5 and 6) compares the performance of $variant_{OL.A1c.2247}$ and $variant_{MCR}$, constructed on the same sample size (N = 2247) and externally evaluated on the UMMC cohort; and analogously, $variant_{OL.FPG.8091}$ and $variant_{UMMC}$, constructed on the same sample size (N = 8091) and externally evaluated on the MCR cohort. Each variant pair used the same set of variables and the same number of training instances. The only difference is that one was trained on the national OLDW cohort and the other was trained on the locale-specific UMMC or MCR cohort.

Discussion

In healthcare, the generalizability of study results has always been emphasized as evidenced by various reporting standards for clinical trials [29], observational studies [30], diagnostic and prognostic studies [31, 32], and meta-analyses [33, 34]. Detailed and transparent reporting is helpful to objectively evaluate the generalizability. However, simply following the standards does not guarantee a generalizable (or transferable) model. Although we followed the TRIPOD statement [32], the external validation of local model variants failed. In this section, we discuss the importance of external validity since we believe it facilitates the adoption of ML models into clinical practice.

External validity The concept of external validity is more complicated than it seems to be [35, 36]. Especially, in order to judge external validity, it is important to define a target population on which external validity is dependent, but the definition is often omitted in most studies, limiting the transferability of a model. Additionally, the criteria for external validity is undefined. Performance is a predominant criterion for external validity in the vast majority of literature [37, 38]; however, external validity could, in theory, include clinical findings (e.g., prognostic factors). Going beyond risk prediction, in order for ML models to be useful for patients' prognosis or treatment development, the use of a compound criterion including predictive performance and clinical findings will become increasingly important. Lastly, institutional policies pose great challenges to external validation because they can influence the available variables for study. For example, health disparities are highly predictive of T2DM and its complications [39–41] but are not commonly collected in routine practice. Models using such variables will eventually be impractical in many other healthcare systems. Therefore, it is crucial to use a set of variables that are commonly observable and able to explain the outcome variable sufficiently so that the model achieves reasonable predictive power.

Dangers of ignoring external validity One may argue that as long as a model is only applied to its training population (a model is never transferred to a different healthcare system), internal validity is sufficient. This is not necessarily true. As previously mentioned, a model can encounter patients atypical for its training

Table 6 Transferability (Mean C-index and 95CIs) of at various OLDW training cohort size

Validation cohort	Training Cohort Size	Model	IHD	CHF	CVD	PVD	CKD	CRF
UMMC	2,247	MTL ($variant_{OL.A1c.2247}$)	.62 (.53–.66)	.69 (.57–.73)	.66 (.57–.71)	.68 (.54–.72)	.69 (.63–.73)	NA
		$LASSO_{OL.A1c.2247}$	.58 (.50–.64)	.65 (.53–.72)	.63 (.53–.71)	.62 (.51–.70)	.70 (.60–.75)	NA
	8,091	MTL ($variant_{OL.A1c.8091}$)	.70 (.67–.72)	.72 (.66–.75)	.70 (.65–.73)	.69 (.65–.73)	.76 (.72–.78)	.84 (.70–.90)
		$LASSO_{OL.A1c.8091}$	.69 (.64–.72)	.71 (.64–.75)	.70 (.65–.73)	.68 (.60–.72)	.76 (.72–.78)	.85 (.71–.91)
		$GBM_{OL.A1c.8091}$	.60 (.56–.63)	.70 (.65–.74)	.67 (.61–.71)	.67 (.62–.71)	.68 (.64–.72)	.83 (.72–.89)
	30,000	MTL ($variant_{OL.A1c.30K}$)	.72 (.71–.73)	.75 (.73–.77)	.73 (.71–.75)	.73 (.72–.75)	.78 (.77–.79)	.89 (.86–.91)
		$LASSO_{OL.A1c.30K}$	.72 (.70–.73)	.76 (.74–.77)	.74 (.71–.75)	.73 (.71–.75)	.78 (.77–.79)	.90 (.86–.91)
		$GBM_{OL.A1c.30K}$	.63 (.60–.65)	.73 (.71–.75)	.70 (.67–.72)	.71 (.68–.73)	.73 (.69–.75)	.90 (.87–.91)
	81,091	MTL ($variant_{OL.A1c.81091}$)	.74 (.73–.74)	.78 (.77–.78)	.75 (.75–.75)	.75 (.75–.75)	.79 (.79–.79)	.92 (.92–.92)
		$LASSO_{OL.A1c.81091}$	.73 (.72–.74)	.77 (.76–.78)	.74 (.73–.75)	.74 (.73–.75)	.79 (.78–.79)	.91 (.89–.92)
		$GBM_{OL.A1c.81091}$	.65 (.63–.66)	.74 (.73–.75)	.72 (.71–.73)	.72 (.72–.73)	.74 (.72–.75)	.91 (.90–.92)
MCR	2,247	MTL ($variant_{OL.FPG.2247}$)	.64 (.52–.71)	.70 (.53–.80)	.64 (.53–.70)	.64 (.52–.71)	.67 (.54–.74)	NA
		$LASSO_{OL.FPG.2247}$	.60 (.50–.69)	.69 (.48–.79)	.62 (.49–.69)	.61 (.48–.68)	.65 (.51–.73)	NA
	8,091	MTL ($variant_{OL.FPG.8091}$)	.76 (.70–.79)	.74 (.57–.80)	.68 (.59–.72)	.70 (.63–.75)	.72 (.62–.77)	.84 (.66–.90)
		$LASSO_{OL.FPG.8091}$	.76 (.69–.80)	.75 (.64–.80)	.69 (.61.73)	.69 (.58–.74)	.71 (.63–.77)	.82 (.67–.91)
		$GBM_{OL.FPG.8091}$	.61 (.54–.66)	.73 (.66–.78)	.64 (.58–.69)	.66 (.58–.70)	.62 (.54–.69)	.67 (.50–.80)
	30,000	MTL ($variant_{OL.FPG.30K}$)	.79 (.76–.81)	.76 (.71–.81)	.71 (.67–.74)	.74 (.71–.77)	.72 (.67–.77)	.89 (.83–.92)
		$LASSO_{OL.FPG.30K}$	.79 (.77–.81)	.78 (.73–.82)	.72 (.69–.74)	.75 (.72–.76)	.71 (.67–.77)	.89 (.83–.93)
		$GBM_{OL.FPG.30K}$	.66 (.61–.70)	.79 (.74–.81)	.69 (.66–.70)	.69 (.66–.72)	.66 (.58–.71)	.88 (.80–.92)
	74,551	MTL ($variant_{OL.FPG.74551}$)	.81 (.80–.81)	.79 (.77–.80)	.73 (.73–.74)	.77 (.77–.77)	.72 (.72–.73)	.91 (.91–.92)
		$LASSO_{OL.FPG.74551}$	.80 (.79–.81)	.79 (.75–.82)	.73 (.71–.74)	.76 (.74–.77)	.71 (.69–.74)	.90 (.88–.93)
		$GBM_{OL.FPG.74551}$	.69 (.67–.72)	.80 (.78–.82)	.70 (.68–.71)	.71 (.69–.72)	.70 (.63–.74)	.91 (.87–.93)

population. Also, institutional policies vary among healthcare systems and could change at anytime. Most models for clinical applications are based on patient characteristics and disease characteristics, but we must be aware that policies can influence outcomes in a way that is not determined by these characteristics [35]. Consequently, models implicitly incorporate the effects of the policies into the effects of physiological factors. Thus, external validation is helpful at least to detect the presence of institutional policies.

Sample size and a large national training set We found that a larger sample size led to increased transferability, however, transferability reached a steady state at a sample size of 30 K. While this sample size was the same for all three modeling algorithms we tried, it is possible that other models, such

Table 7 Performance comparison: national representative training cohort vs. locale-specific training cohort

Validation cohort	Training Cohort Size	Model	IHD	CHF	CVD	PVD	CKD	CRF
UMMC	2,247	$variant_{OL.A1c.2247}$	.62 (.53–.66)	.69 (.57–.73)	.66 (.57–.71)	.68 (.54–.72)	.69 (.63–.73)	NA
		$variant_{MCR}$	.62	.73	.67	.70	.74	.90
MCR	8,091	$variant_{OL.FPG.8091}$	.76 (.70–.79)	.74 (.57–.80)	.68 (.59–.72)	.70 (.63–.75)	.72 (.62–.77)	.84 (.66–.90)
		$variant_{UMMC}$	.69	.72	.67	.69	.67	.90

as deep learning, would saturate at a different sample size. We found that $variant_{OL.A1c.2247}$ performed worse than $variant_{MCR}$, indicating a lack of representativeness. A representative training set should be an unbiased reflection of a population such that predictions of a model should not be dependable but consistent across various subpopulations. With a sample size of 2247, $variant_{OL.A1c.2247}$ was not able to learn enough detail about the national OLDW cohort, therefore it performed inadequately on patients in the local healthcare system. More importantly, $variant_{OL.FPG.8091}$ built on the national OLDW training set outperformed $variant_{UMMC}$ in external validation even when the sample size and the set of variables are kept identical, which demonstrates a national training set yields a more transferable ML model.

Limitations Although many patients were excluded due to missing lab results and vital signs measurements, we did not impute them. Diabetic patients are supposed to receive routine check-ups and see their primary care physicians regularly. For patients without any HbA1c or blood pressure measurements *at all* during the baseline period, it was uncertain whether we could establish their baseline characteristics correctly.

Summary

We demonstrated that even a complex ML model can be successfully externally validated when it is constructed on a large national dataset collected from multiple providers. Having a large national dataset allows the model to capture patients who may be atypical in certain parts of the nation but common in others. It can also help to marginalize institutional policies that could suddenly change so render a trained model inaccurate. Conversely, we have also demonstrated that models built on a local healthcare system's data did not transfer to a different healthcare system. Given the current status quo of ML models being built on local healthcare systems and not being externally validated, their robustness can be questioned. We advocate for external validation for ML models to make them more robust, which will undoubtedly help in their adoption for clinical decision support.

Conclusions

Our modeling approach, in which a model is learned from a national cohort and is externally validated, can facilitate the production of a more transferable model, allowing for precision medicine to be more accessible and to become a reality.

Funding This work was supported by NIH award R01 LM011972, NSF awards IIS 1602198. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Compliance with ethical standard

Conflict of interest The access to the claims and EHR data from the OLDW was made possible through use of an OptumLabs research credit. Author Era Kim owns stock in UnitedHealth Group.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Obermeyer, Z., and Emanuel, E. J., Predicting the future - big data, machine learning, and clinical medicine. The New England journal of medicine 375(13):1216–1219, 2016.
2. Florez, J. C., Precision medicine in diabetes: Is it time? Diabetes Care 39(7):1085–1088, 2016.
3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I., Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15:104–116, 2017.
4. Perveen, S. et al., A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression. Comput. Struct. Biotechnol. J. 13(December):1445–1454, 2017, 2016.
5. Lagani, V. et al., Development and validation of risk assessment models for diabetes-related complications based on the DCCT/EDIC data. J. Diabetes Complications 29(4):479–487, 2015.
6. Cichosz, S. L., Johansen, M. D., and Hejlesen, O., Toward big data analytics. Review of Predictive Models in Management of Diabetes and Its Complications, 2016.
7. Bengio, Y., Delalleau, O., and Simard, C., Decision trees do not Generaliza to new variations. Comput. Intell. 26(4):449–467, 2010.
8. Lisboa, P. J., and Taktak, A. F. G., The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks 19(4):408–415, 2006.
9. Cruz, J. A., and Wishart, D. S., Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2:59–77, 2006.

10. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I., Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13:8–17, 2015.
11. Weeks, J., and Pardee, R., Learning to share health care data : A brief timeline of influential common data models and distributed health data networks in U.S. health care research. Gener. Evid. Methods to Improv. patient outcomes 7(1):1–7, 2019.
12. Arterburn, D. et al., Comparative effectiveness and safety of bariatric procedures for weight loss. Ann. Intern. Med. 169(11):741–750, 2018.
13. Inge, T. H. et al., Comparative effectiveness of bariatric procedures among adolescents : The PCORnet bariatric study. Surg. Obes. Relat. Dis. 14(9):1374–1386, 2018.
14. C. L. Roumie *et al.*, "Performance of a computable phenotype for identification of patients with diabetes within PCORnet : The Patient - Centered Clinical Research Network," no. December 2018, pp. 1–8, 2019.
15. Chubak, J. et al., The Cancer research network : A platform for epidemiologic and health services research on cancer prevention, care, and outcomes in large, stable populations. Cancer Causes Control 27(11):1315–1323, 2016.
16. Hripcsak, G., Ryan, P. B., Duke, J. D., and Shah, N. H., R. Woong, and V. Huser, "Characterizing treatment pathways at scale using the OHDSI network," 113(27):7329–7336, 2016.
17. Wallace, P. J., Shah, N. D., Dennen, T., Bleicher, P. A., and Crown, W. H., Optum labs: Building a novel node in the learning health care system. Health Aff. 33(7):1187–1194, 2014.
18. OptumLabs, "OptumLabs and OptumLabs Data Warehouse (OLDW) Descriptions and Citation," *Cambridge, MA: n.p.,* PDF, Reproduced with permission from OptumLabs, 2018.
19. American Diabetes Association (ADA), "Standards of Medical Care in Diabetes - 2017," *Diabetes Care*, vol. 40 (sup 1), no. January, pp. s4–s128, 2017.
20. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J., On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat. Med. 30(10): 1105–1117, 2011.
21. Nathan, D. M., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D., and Heine, R. J., Translating the A1C assay into estimated average glucose values. Diabetes Care 31(8):1473–1478, 2008.
22. E. Kim, D. S. Pieczkiewicz, M. R. Castro, P. J. Caraballo, and G. J. Simon, "Multi-Task Learning to Identify Outcome-Specific Risk Factors that Distinguish Individual Micro and Macrovascular Complications of Type 2 Diabetes," *AMIA 2018 Informatics Summit Proc.*, 2018.
23. Deedwania, P. C. et al., Differing predictive relationships between baseline LDL-C, systolic blood pressure, and cardiovascular outcomes. Int. J. Cardiol. 222:548–556, 2016.
24. Despres, J. P., Lemieux, I., Dagenais, G. R., Cantin, B., and Lamarche, B., HDL-cholesterol as a marker of coronary heart disease risk: The Quebec cardiovascular study. Atherosclerosis 153: 263–272, 2000.
25. Retnakaran, R., Cull, C. A., Thorne, K. I., Adler, A. I., and Holman, R. R., Risk factors for renal dysfunction in type 2 diabetes. Diabetes 55(6):1832–1839, 2006.
26. Franklin, S. et al., Does the relation of blood pressure to coronary heart disease risk change with aging?: The Framingham heart study. Circulation 103(9):1245–1249, 2001.
27. Evans, G. W. et al., Effects of intensive blood-pressure control in type 2 diabetes mellitus. N. Engl. J. Med. 362(17):1575–1585, 2010.
28. Li, W. et al., Body mass index and heart failure among patients with type 2 diabetes mellitus. Circ. Hear. Fail. 8(3):455–463, 2015.
29. Schulz, K. F. et al., CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. BMC Med. 8(1):18, 2010.
30. von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., and Vandenbroucke, J. P., The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Int. J. Surg. 12(12): 1495–1499, 2014.
31. Bossuyt, P. M. et al., RESEARCH METHODS & REPORTING STARD 2015 : An updated list of essential items for. Radiographies 277(3):1–9, 2015.
32. Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Eur. Urol. 67(6):1142–1151, 2015.
33. Ahmed, I., Debray, T. P. A., Moons, K. G. M., and Riley, R. D., Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med. Res. Methodol. 14(1):3, 2014.
34. Abo-Zaid, G., Sauerbrei, W., and Riley, R. D., Individual participant data meta-analysis of prognostic factor studies: State of the art? BMC Med. Res. Methodol. 12:56, 2012.
35. Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A., and Vandenbroucke, J. P., How to assess the external validity of therapeutic trials: A conceptual approach. Int. J. Epidemiol. 39(1):89–94, 2010.
36. Van Soest, J. et al., Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility. Med. Phys. 44(9), 2017.
37. Huang, J., and Ling, C. X., Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. 17(3):299–310, 2005.
38. Sokolova, M., and Lapalme, G., A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45(4): 427–437, 2009.
39. Osborn, C. Y., Groot, M., and Wagner, J. A., Racial and ethnic disparities in diabetes complications in the northeastern United States: The role of socioeconomic status. J. Natl. Med. Assoc. 105(1):51–58, Jan. 2013.
40. Maier, W. et al., The impact of regional deprivation and individual socio-economic status on the prevalence of type 2 diabetes in Germany. A pooled analysis of five population-based studies. Diabet. Med. 30(3):e78–e86, Mar. 2013.
41. Hu, R., Shi, L., Rane, S., Zhu, J., and Chen, C. C., Insurance, racial/ethnic, SES-related disparities in quality of care among US adults with diabetes. J. Immigr. Minor. Heal. 16(4):565–575, 2014.