



# Electronic Medical Record Risk Modeling of Cardiovascular Outcomes Among Patients with Type 2 Diabetes

Dongzhe Hong · Daniel Fort · Lizheng Shi · Eboni G. Price-Haywood 

Received: April 27, 2021 / Accepted: June 5, 2021 / Published online: June 18, 2021  
© The Author(s) 2021

## ABSTRACT

**Introduction:** Increased utilization of electronic health records (EHR) has enriched databases for creating risk models. We used machine learning techniques to develop an EHR-based risk model locally fitted to patients with type 2 diabetes mellitus (T2DM) for predicting cardiovascular disease.

**Methods:** This retrospective observational study was conducted within Ochsner Health, Louisiana, USA, between 2013–2017. Data analysis included 6245 patients who had two outpatient diagnoses of T2DM recorded on separate days or a diagnosis recorded during an inpatient encounter. Baseline clinical data were limited to 180 days before the index diagnosis. Cardiovascular outcomes were coronary heart

disease (CHD), heart failure and stroke. Machine learning approaches were used to select predictor variables into Cox proportional hazards models for each outcome. Locally fit equations were compared to “generalized” risk equations (RECODE, AS-CVD, QRISK3) using model discrimination and calibration.

**Results:** Among factors identified in the Ochsner ( $n = 11$ ), RECODE ( $n = 14$ ), AS-CVD ( $n = 15$ ) and QRISK3 ( $n = 23$ ), only age was common to all four risk equations. The Ochsner model had high internal discrimination for CHD (C-statistics 0.85) and better discrimination than RECODE (C-statistics 0.45), the QRISK3 (C-statistics 0.72) and AS-CVD (C-statistics 0.54).

**Conclusions:** The Ochsner model overestimated 5-year CHD risk, but had relatively higher calibration than the other models in CHD. Risk equations fitted for local populations improved cardiovascular risk stratification for patients with T2DM. Application of machine learning simplified the models compared to “generalized” risk equations.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s13300-021-01096-w>.

D. Hong · L. Shi  
Department of Health Policy and Management,  
Tulane University School of Public Health and  
Tropical Medicine, New Orleans, LA, USA

D. Fort · E. G. Price-Haywood (✉)  
Center for Outcomes and Health Services Research,  
Ochsner Health, New Orleans, LA, USA  
e-mail: eboni.pricehaywood@ochsner.org

E. G. Price-Haywood  
University of Queensland Ochsner Clinical School,  
New Orleans, LA, USA

**Keywords:** Diabetes; Machine learning; Type 2 diabetes mellitus; Cerebrovascular stroke; Cardiovascular disease; Heart failure

### Key Summary Points

Type 2 diabetes mellitus and related complications are prevalent and result in heavy economic and disease burdens both within the US healthcare system and globally.

This study developed predictive risk models for coronary heart disease, heart failure and stroke tailored to an integrated delivery health system patient population with type 2 diabetes and compared the performance of the locally fitted model to the QRisk3, RECODE and ASCVD risk equations.

The locally fitted model performed significantly better than the other three models for predicting incident cardiovascular disease in the health system population.

Use of population-specific clinical data and application of machine learning methods can transform existing general predictive models to locally fitted models that perform better in local populations.

## DIGITAL FEATURES

This article is published with digital features, including a summary slide, to facilitate understanding of the article. To view digital features for this article go to <https://doi.org/10.6084/m9.figshare.14717004>.

## INTRODUCTION

Type 2 diabetes mellitus (T2DM) is one of the most prevalent chronic diseases in the world and results in heavy economic and disease burdens both within the US healthcare system and globally [1–3]. The prevalence of type 2 diabetes (T2DM) has continuously increased over the last decade in the US from 4.21% (12.1

million) in 2002 to 9.4% in 2015, according to a recent retrospective study and the 2017 National Diabetes Statistics Report [4, 5]. Recent studies reported that the overall prevalence of diabetes in the US is projected to reach 21% in 2050. Total estimated direct medical cost of T2DM and its related complications were last reported at \$237 billion in 2017 [6, 7]. The majority of the costs associated with diabetes are attributed to the micro-/macrovascular complication events [8–10].

Longitudinal electronic health records (EHRs) including diagnoses, tests, procedures, treatments, medication administrations, biomarkers and other laboratory data have been widely implemented in clinical settings and used in health services research in the US [11–15]. EHR data have been used to develop diabetes risk models. For instance, the QRISK3 prediction algorithms were developed to estimate the 10-year risk of cardiovascular disease in women and men using general practice data in England from the QResearch database [16]. Several diabetes risk models in the US have also been used to describe disease progression and support outcomes-driven evidence-based diabetes management, including the 10-year risk equations for complications of type 2 diabetes (RECODE), and American College of Cardiology/American Heart Association atherosclerotic cardiovascular (AS-CVD) disease equations [17, 18].

However, these national models may not be useful at the health system level if the local population significantly differs from the population used to build the model. The increased availability of EHR data, combined with advances in computing and machine learning methods, makes it possible to locally derive risk prediction models. Because they are built off a local population, it is possible these locally fitted models may outperform similar risk prediction models built for other populations.

Outcomes-driven evidence-based diabetes management would become widely adopted if a good prediction model were available to provide quick assessment at the point of care among specific health system populations. Thus, this study's main objectives are to: (1) describe the development of a predictive risk

models for coronary heart disease, heart failure and stroke tailored to Ochsner Health's (Louisiana's largest integrated delivery health system) patient population with type 2 Diabetes and (2) compare the performance of the Ochsner model to the risk equations for coronary heart disease of RECODE, AS-CVD and QRISK3.

## METHODS

### Population, Setting and Study Design

This study is a secondary data analysis of EHR data acquired from the Louisiana Experiment Assessing Diabetes cohort study. The LEAD cohort includes electronic health record data which were obtained from the Research Action for Health Network (REACHnet) for the time period between January 1, 2013, and October 31, 2017 [19]. Clinical data from REACHnet conform to the National Patient-Centered Clinical Research Network (PCORnet) common data model, the specification that defines a standard organization and representation of data for the PCORnet distributed research network [20]. We conducted a retrospective observational cohort study within Ochsner Health. The study population was restricted to patients who received care within Ochsner, which was a sub-population derived from the LEAD study cohort [21–23]. The definition of T2DM in the present study was formulated according to the Surveillance Prevention and Management of Diabetes Mellitus (SUPREME-DM) definitions as follows: (1) one or more of the International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) codes and Tenth Revision, Clinical Modification (ICD-10-CM) codes for type 2 diabetes mellitus associated with inpatient encounters; (2) two or more ICD codes associated with outpatient encounters on different days within 2 years; (3) combination of two or more of the following associated with outpatient encounters on different days within 2 years: (1) ICD codes; (2) fasting glucose level  $\geq 126$  mg/dl; (3) 2-h glucose level  $\geq 200$  mg/dl; (4) random glucose  $\geq 200$  mg/dl; (5) hemoglobin A1c (HbA1c)  $\geq 6.5\%$ ; (6) prescription for an

antidiabetic medications [24]. The study and analysis plan were approved by the Ochsner Health Institutional Review Board, which granted waiver of consent for this retrospective data only study.

### Data Variables

Patients' data extracted from the PCORnet common data model for the present study included demographic characteristics, clinical biomarkers, medical histories and medication utilization. The demographic characteristics included age at diabetes diagnosis, race/ethnicity and sex. Clinical information with encounter dates, dates of diagnoses and laboratory test dates included weight, height, body mass index (BMI), blood pressure, diagnoses of various diseases, total cholesterol, triglycerides, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, glycosylated hemoglobin (HbA1c) and estimated glomerular filtration rate (eGFR). We also included health-care settings for each patient encounter and medication prescription histories such as anti-hypertensive drugs, glucose-lowering drugs and lipid-lowering drugs. Antidiabetic medications included insulin, sulfonylurea, metformin, dipeptidyl peptidase-4 inhibitors, alpha glucosidase inhibitors, amylin analogs, sodium-glucose cotransporter-2 inhibitors, glucagon-like peptide-1 receptor agonists, meglitinide, thiazolidinedione, insulin analogs and medications that increase the secretion of insulin. Antihypertensive medications included beta blockers, calcium channel blockers, ACE inhibitors, diuretic, angiotensin receptor blockers, alpha blockers, sympatholytics and vasodilators. Lipid-lowering medications included statins, niacin, bile sequestrants, PCSK9 inhibitor, fibrates, ezetimibe and fish oil. The eGFR was estimated using the Modification of Diet in Renal Disease (MDRD) [25].

Baseline clinical data were limited to 180 days before the first recorded date of the T2DM diagnosis for consistency in the data collection period for the survival analysis described below. Documentation of clinical care after the index date was used as follow-up data

instead of baseline data. Patients with any missing value in baseline data were excluded. Among a total of 93,034 T2DM patients in the original cohort, 86,789 were excluded because of missing HbA1c data at baseline (before T2DM diagnosis documented).

## Outcomes

The main outcomes of the present study were recorded diagnosis of coronary heart disease (CHD: ICD-9 CM codes: 410-414,429.2; ICD-10 CM codes: I20-I25), heart failure (HF: ICD-9 CM codes: 402.01,402.11,402.91,428, ICD-10 code: I50) and stroke (ICD-9 CM codes: 430-436, ICD-10 CM codes: I60-I66).

## Development of Health System Tailored T2DM Risk Models

The Ochsner risk models employed Cox proportional hazards models for CHD, HF and stroke, followed by LASSO regression to select predictor variables from demographic characteristics, clinical variables, medications and biomarkers. LASSO regularization is a well-established machine learning method that can help select important variables [18, 26]. This approach fitted the Ochsner risk model via penalized maximum likelihood to minimize the risk of overfitting. In addition, the LASSO method has computational convenience and performs competitively in real examples, incorporating different penalties for different coefficients. Unimportant variables receive larger penalties than important ones, so that important variables tend to be retained in the selection process, whereas unimportant variables are more likely to be dropped [27].

## Assessment of Model Performance

All compared diabetes risk models included in this study were Cox proportional hazards models. RECODE developed multiple risk equations for T2DM complications, including cardiovascular disease, congestive heart failure, stroke and other microvascular complication outcomes [18]. However, AS-CVD and QRISK3

models only predicted the risk of developing CHD as an outcome of T2DM complication [16, 17]. Since CHD is the common outcome among all models, the Ochsner, RECODE, AS-CVD and QRISK3, model performances were compared on CHD only.

Since the study objective is to develop a specific risk prediction model for Ochsner Health System, this study used the same cohort to test the model performance for the Ochsner model as well as the other risk models. Patients were not excluded if they had missing AS-CVD or QRISK3 or RECODE data. If covariates in ASCVD, QRISK3 and RECODE were missing, the analysis used the baseline data from the respective publications.

Model discrimination was assessed by the C-statistic (area under the receiver-operating characteristic curve) [28]. The baseline survival of CHD was defined as the 5-year survival of CHD in the Ochsner cohort and was calculated by using the Kaplan-Meier survival function [29]. The baseline survival of CHD was used to conduct the performance comparison of the Ochsner model, RECODE and QRISK3. The baseline survivals of CHD in AS-CVD disease equations were used to evaluate its performance since the equations published their own gender- and race-specific baseline survivals for the US population. The patient risk of developing cardiovascular outcomes was calculated by the equation:  $P(t, x) = 1 - S(t, x)$  being the failure (event) probability, that is, the chance of an event occurring in the interval  $(0, t)$  for an individual with covariate vector  $x$ . In addition,  $S(t) = S(0)\hat{e}^{(\sum \beta \times x - \sum \beta \times \bar{x})}$ . In the equation,  $S(0)$  is the baseline survival of the three cardiovascular outcomes,  $x$  is the corresponding value of each variable in each model,  $\bar{x}$  is the corresponding mean of the cohort's characteristics for each continuous variable in each model, and  $\bar{x}$  is "0" for each categorical variable for the reference group in the model [18]. C-statistics were calculated by using  $P(t)$  as probability and event status (i.e., whether the patient had cardiovascular event).

A logistic regression model was used to assess the calibration of risk models. The outcome probability,  $P(x)$ , is a function of the prognostic

index (PI),  $\beta \times x$ , and the baseline log odds of an event,  $\beta_0 = \text{logit}\{P(0)\}$ . Assessing model calibration means comparing the observed event probabilities with those predicted by the model. The observed event probability for an individual is taken as 1 if the individual experiences an event (outcome  $Y = 1$ ) and 0 otherwise (outcome  $Y = 0$ ). We write the PI as  $\text{PI} = \beta_0 \times x\beta$ . The predicted event is  $P(x) = \text{logit}^{-1}(\text{PI}) = \{1 + e^{(-PI)}\}^{-1}$ . A logistic regression model  $\text{logit}\{\text{Pr}(Y = 1)\} = \gamma_0 + \gamma_1 \text{PI}$ , which is linear in the PI, was used to check agreement between observed and predicted probabilities[30].

If a model is well calibrated, the estimates of  $\gamma_0$  and  $\gamma_1$  are identically 0 and 1, respectively. The model calibration was assessed with three tests at time  $t$ : (1) intercept test, (2) slope test and (3) joint test [30].

All analyses were conducted using R, version 4.0.3 [31]. All methods were carried out in accordance with relevant guidelines and regulations. This study was funded by the Ochsner Health Clinical Research and Innovation Support Program (CRISP). The data analyzed in this study were not identifiable. This study was deemed exempt from Institutional Review Boards from Tulane University and Ochsner Health. The permission to access the data was granted as the project was funded and led by the data owner (Ochsner Health). The content of this publication is solely the responsibility of the author(s) and does not necessarily represent the views of the sponsoring health system. The authors do not have any conflicts of interest to disclose.

## RESULTS

A total of 6245 patients were included in the present study. Table 1 shows the baseline characteristics for the T2DM cohort in Ochsner Health System. The mean (standard deviation [SD]) of age was 61.0 (11.7) years old. Most of the study population was female (51.5%) and White (59.2%). The mean (SD) hemoglobin A1c (HbA1c) was 7.4 (1.7) mmol/mol, and 1284 (20.6%) of the patients had HbA1c > 8%. In addition, the percentage of the population who

**Table 1** Characteristics of the T2DM cohort in Ochsner Health (2013–2017) documented within 180 days prior to first notation of diagnosis in EHR

	<i>N</i> = 6245
<b>Demographics</b>	
Age, years (mean, SD)	61.0 (11.7)
Sex ( <i>n</i> , %)	
Woman	3225 (51.6)
Man	3020 (48.4)
Race ( <i>n</i> , %)	
White	3694 (59.2)
Black	2456 (39.3)
Asian	65 (1.0)
Other	30 (0.5)
<b>Clinical or biomarkers (mean, SD)</b>	
BMI, kg/m <sup>2</sup>	33.9 (8.0)
HbA1c, %	7.4 (1.7)
HbA1c > 8%, <i>n</i> (%)	1284 (20.6)
<b>Blood pressure, mmHg</b>	
Systolic	133.0 (18.0)
Diastolic	78.2 (10.9)
LDL cholesterol, mg/dl	109.4 (36.8)
HDL cholesterol, mg/dl	44.1 (12.4)
Total cholesterol, mg/dl	182.6 (44.8)
Triglycerides, mg/dl	145.6 (87.2)
Estimated GFR, ml/min/1.73 m <sup>2</sup>	51.5 (36.6)
<b>Medical history (<i>n</i>, %)</b>	
Hypertension history	4895 (78.4)
CHD history	1037 (16.6)
Heart failure history	540 (8.7)
Stroke history	597 (9.6)
<b>Medication prescription history (<i>n</i>, %)</b>	
Hypolipidemic drugs	2139 (34.3)
Statin	2010 (32.2)
Antidiabetic drugs	1360 (21.8)

**Table 1** continued

	<i>N</i> = 6245
Insulin	250 (4.0)
Metformin	1044 (16.7)
Sulfonylurea	219 (3.5)
SGLT2 inhibitor	6 (0.1)
GLP1 receptor agonist	28 (0.5)
Antihypertensive drugs	3540 (56.7)

*T2DM* type 2 diabetes mellitus; *BMI* body mass index; *DL* low-density lipoprotein; *HDL* high-density lipoprotein; *GFR* glomerular filtration rate; *GLP-1* glucagon-like peptide-1; *SGLT2* sodium-glucose cotransporter 2

had hypertension history, CHD history, HF history and stroke history was 78.4%, 16.6%, 8.7% and 9.6%, respectively. Lastly, the percentages of using hypolipidemic, antidiabetic and antihypertensive drugs at baseline were 34.3%, 21.8% and 56.7%, respectively. During the follow-up period, 413 (6.6%) patients had CHD, 295 (4.7%) had HF, and 105 (1.7%) had stroke.

Table 2 provides the coefficients of the Ochsner risk equations for each of cardiovascular outcomes. The LASSO regularization method revealed that common variables in Ochsner models include age, BMI, systolic blood pressure, HbA1c and eGFR. The other significant predictors were medical histories, such as CHD, HF and hypertension, followed by medication prescription histories and race.

Among factors identified as statistically significant in the Ochsner ( $n = 11$ ), RECODE ( $n = 14$ ), AS-CVD ( $n = 15$ ) and QRISK3 ( $n = 23$ ), only age was common to all four risk equations (Supplementary Material Table S1). Three significant predictors of CHD were common between the Ochsner and RECODE models, including age, HbA1c (%) and HDL cholesterol (mg/dl). Only two significant predictors of CHD were common between the Ochsner model and QRISK3 equations, including age and BMI (kg/

m<sup>2</sup>). Five significant predictors of CHD were common between the Ochsner and AS-CVD models, which included age, sex, race, HbA1c and HDL cholesterol.

Table 3 presents the comparisons of model discrimination with alternative risk equations among the Ochsner T2DM cohort. The Ochsner model equations had high internal discrimination with C-statistics of 0.85 for CHD. The Ochsner model equations had better discrimination than RECODE with C-statistics 0.46, AS-CVD disease equations with C-statistics 0.54 and the QRISK3 with C-statistics of 0.72 for CHD.

Table 4 shows the logistic regression results of the prognostic index on having CHD in the Ochsner T2DM cohort. The estimate of the intercept in the Ochsner model suggested that the predicted risk of having CHD at 5 years is about  $\exp^{-3.829} = 0.021$  higher than a perfect calibration, indicating that the Ochsner model overestimated the 5-year CHD risk. Along with the joint test results, a miscalibration for the Ochsner model was common with all the other models (Supplementary Material Table S2). Among the four models, the Ochsner model equations had a relatively high internal calibration.

## DISCUSSION

In an era of learning health systems during which health policy changes are driving population health management to improve the quality, cost and experience of healthcare, health systems need reliable, reproducible predictive analytic tools that account for the diverse characteristics of populations they serve and allow for better patient care. Our results show a significantly better performance in the locally fitted Ochsner model than the other three models for predicting incident cardiovascular disease in the Ochsner population. These findings suggest locally fitted models may provide more useful predictive analytics compared to existing broader models. Although we only compared the number of significant predictors of CHD among the four models, this study found that the Ochsner model required fewer

**Table 2** Coefficients of the Ochsner models for calculating 5-year risk of CHD, HF and stroke

	CHD	HF	Stroke
<b>Demographics</b>			
Age, years	0.02143	0.03372	0.01705
BMI, kg/m <sup>2</sup>	− 0.00959	0.01053	− 0.02817
<b>Sex</b>			
Man	0.13083		
<b>Race</b>			
Black	0.22707	0.29277	0.03210
Asian	− 0.70858	− 0.91104	− 0.21774
Other	− 0.87601		− 0.06151
<b>Clinical or biomarkers</b>			
Systolic blood pressure, mmHg	0.00213	0.00455	0.00990
HbA1C, %	0.05607	0.03734	0.09647
LDL cholesterol, mg/dl	− 0.00078	− 0.00374	
HDL cholesterol, mg/dl	− 0.00847		
Triglycerides, mg/dl			0.00067
Estimated GFR, ml/min/1.73 m <sup>2</sup>	0.00292	− 0.00108	0.00111
<b>Medical history</b>			
Hypertension history	0.55760	0.59334	0.23297
CHD history	1.84915	0.37780	0.36549
Heart failure history	0.46568	1.88251	0.58397
Stroke history	− 0.14675	0.18941	0.65132
<b>Medication prescription history</b>			
Hypolipidemic drugs	0.34079		
Antidiabetic drugs	0.05846		
Antihypertensive drugs	− 0.20982	− 0.17883	

All risk factors listed in Table 1 were considered for inclusion; the ones listed in Table 2 were also selected for inclusion. The 5-year risk of an outcome can be computed as  $1 - S(0)\hat{e}(\sum \beta \times x - \sum \beta \times \bar{x})$ . Here  $\beta$  is the coefficient, and  $x$  is the covariate vector for an individual patient within the Ochsner cohort.  $S(0)$  for the CHD, HF and stroke are 0.898, 0.928 and 0.975, respectively. For example, a 65-year-old white man with BMI 34.9 kg/m<sup>2</sup>, systolic blood pressure 143 mmHg, HbA1c 8.4%, LDL cholesterol 110 mg/dl, HDL cholesterol 45 mg/dl, triglycerides 145 mg/dl and estimated GFR 52 ml/min/1.73 m<sup>2</sup> and with hypertension history, without CHD/HF/stroke history, and currently using antidiabetic and antihypertensive drugs, would have a 5-year CHD risk of  $1 - 0.898 \times \exp(0.02143 \times 65 - 0.00959 \times 34.9 + 0.13083 \times 1 + 0.00213 \times 143 + 0.05607 \times 8.4 - 0.00078 \times 110 - 0.00847 \times 45 + 0.00292 \times 52 + 0.55760 \times 1 + 0.05846 \times 1 - 0.20982 \times 1 - 1.37) = 0.192$  or 19.2% 5-year risk, where 1.37 is the mean  $\sum \beta \times \bar{x}$

CHD coronary heart disease; HF heart failure; BMI body mass index; LDL low-density lipoprotein; HDL high-density lipoprotein; GFR glomerular filtration rate

**Table 3** C-statistics of risk model performance for each model in the Ochsner T2DM cohort

Models	C-statistics for CHD
Ochsner model	0.85 (0.83, 0.87)
RECODE	0.46 (0.43, 0.49)
QRISK3	0.72 (0.70, 0.74)
AS-CVD	0.54 (0.51, 0.57)

T2DM type 2 Diabetes Mellitus; CHD coronary heart disease

predictors, implying future efficiencies in data extraction and mapping. We did not compare the number of significant predictors of HF and stroke since neither QRISK3 nor AS-CVD predicted the incident risk of developing HF or stroke. This study also found that the Ochsner model showed the best discrimination of predicting cardiovascular risk in the Ochsner T2DM cohort. The experience from the learning health system will be disseminated to other health systems in the state and other regions.

While the discrimination of the Ochsner model was significantly better than the other models for CHD, none of the four models performed well on any of the calibration tests. Failure of calibration tests where risk is overestimated on the high end, as seen in the Ochsner model, is associated with overfitting in models with rare events [32]. One of the benefits of the penalized regression methods undertaken in this investigation is the prevention of exactly

that outcome, suggesting overfitting may not be the simple root of the observed problem. Regardless, as the calibration error is in the direction of overestimating risk, it might be argued the result would still serve a purpose in successfully identifying individuals for preventive measures.

While the better performance of a locally fitted model may appear logical, the cross-model comparison on model performance is very preliminary and should be interpreted with caution. For example, the RECODE model derived its risk equations from ACCORD (2001–2009) clinical trial data [18]. However, some of the required variables for this model were not available for each patient in our cohort. Although the RECODE model states its risk equations are tolerant of missing data, its performance may have been significantly hampered in our investigation. On the other hand, if a commonly generalized risk prediction model requires data quality identical to a randomized controlled trial for success, then it has already introduced significant barriers to local implementation.

A systematic review of prediction models for cardiovascular disease risk in the general population argued that the predictive performance of most models for predicting CVD risk is heterogeneous, and the usefulness of most models remains unclear [33]. This systematic review also concluded that it is impossible to recommend which specific model should be used in which setting or location, which was broadly

**Table 4** Logistic regression results of PI on having CHD in the Ochsner T2DM cohort

Model	Coefficients	Coefficient	S.E	<i>p</i> value	95% Confidence interval	
Ochsner	PI	0.882	0.045	0.000	0.794	0.970
	Intercept	− 3.829	0.096	0.000	− 4.018	− 3.641
RECODE	PI	1.288	0.081	0.000	1.129	1.447
	Intercept	− 2.845	0.060	0.000	− 2.962	− 2.727
QRISK3	PI	0.065	0.143	0.651	− 0.216	0.346
	Intercept	− 2.744	0.220	0.000	− 3.175	− 2.313
AS-CVD	PI	0.473	0.060	0.000	0.355	0.591
	Intercept	− 3.058	0.096	0.000	− 3.246	− 2.869



supported by the poor results of the nationally generalizable models in our local population.

This study has several limitations. First, the Ochsner model assesses a 5-year risk compared to the 10-year risk prediction of the comparative models. It is possible that required predictors from the other models which failed in our analysis would become significant in predicting longer term risk in years 5–10 after diagnosis. In addition, it may underestimate the performance of AS-CVD disease equations because our analysis used the gender- and race-specific baseline 10-year survival estimates derived from the AS-CVD study [17]. We cannot directly apply the baseline 5-year survival estimates from the Ochsner T2DM cohort. We have to accumulate more years of data for a cohort with 10 years' follow-up from the current cohort of a maximum 5-year follow-up (EHR records from 2013 to 2017). Additionally, the study used apparent validation which employed the same cohort as the training sample. Therefore, the Ochsner model's predictive performance estimates could be more optimistic than other validation methods, such as split-sample validation and bootstrap validation [27]. It is also a predictable result because the applied machine learning method was fitted to the local population. Furthermore, the performance of the Ochsner model may be overestimated when applying to patients with missing data since patients with baseline missing data were excluded from model development. Imputation of missing data was not conducted because of the large proportion of patients with missing information (> 25%). The analysis handled missing data by minimizing the number of predictors using the machine learning approach. The LASSO regularization can keep relatively important clinical factors in real-world clinical settings. The Ochsner model had fewer covariates in the risk prediction models than other models. Notably, in real-world settings, patients with missing data within an EHR often have missing risk assessments until the required data are captured and model calculations are subsequently updated. Lastly, the incidence of cardiovascular outcomes may be overestimated among the Ochsner T2DM cohort since we only required a 180-day

baseline prior to first recorded diagnosis as an "all-comer" approach for health system's population health management needs. Thus, the comparison of model performance may also be biased because of the potential bias of the cardiovascular outcome ascertainment.

## CONCLUSIONS

Use of population-specific clinical data and application of machine learning methods can transform existing general predictive models to locally fitted models that perform better in local populations. Predictive analytics are increasingly incorporated into population health management strategies for risk profiling patients, evaluating the comparative effectiveness of different therapeutic plans and estimating long-term outcomes for different treatment goals. "Generalized" risk prediction models do not necessarily have to be re-built for the local population; however, researchers and clinicians should be cautious about the results of these models when applying them to local populations as the risks may be over- or underestimated. Locally fitted models may provide better support for achieving population-specific strategies.

## ACKNOWLEDGEMENTS

**Compliance with Ethics Guidelines.** The data analyzed in this study were not identifiable. This study was deemed exempt from Institutional Review Boards from Tulane University and Ochsner Health. The permission to access the data was granted as the project was funded and led by the data owner (Ochsner Health).

**Funding.** This study was funded by the Ochsner Health Clinical Research and Innovation Support Program (CRISP). The journal's rapid service fee was funded by the authors.

**Prior Publication.** This work was presented as an abstract at the International Society for

Pharmacoeconomics and Outcomes Research (ISPOR) Annual Meeting in May 18–22, 2019.

**Authorship.** All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

**Author Contributions.** All authors (Dongzhe Hong, Daniel Fort, Eboni G. Price-Haywood, Lizheng Shi) take full responsibility for the work, including the study design (Lizheng Shi, Eboni G. Price-Haywood), data collection (Lizheng Shi, Eboni G. Price-Haywood), data analysis (Dongzhe Hong, Daniel Fort, Lizheng Shi), and the decision to draft, submit and publish the manuscript (Dongzhe Hong, Daniel Fort, Eboni G. Price-Haywood, Lizheng Shi).

**Disclosures.** The authors, Dongzhe Hong, Daniel Fort, Eboni G. Price-Haywood and Lizheng Shi, have nothing to disclose.

**Data Availability.** The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

**Open Access.** This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

## REFERENCES

- Hong D, Si L, Jiang M, Shao H, Ming W-K, Zhao Y, et al. Cost effectiveness of sodium-glucose cotransporter-2 (SGLT2) inhibitors, glucagon-like peptide-1 (GLP-1) receptor agonists, and dipeptidyl peptidase-4 (DPP-4) inhibitors: a systematic review. *Pharmacoeconomics*. 2019;37(6):777–818.
- Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. *Phys Ther*. 2008;88(11):1254–64.
- Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. *Lancet*. 2017;389(10085):2239–51.
- Association AD. Economic costs of diabetes in the US in 2007. *Diabetes Care*. 2008;31(3):596–615.
- Centers for Disease Control and Prevention. National diabetes statistics report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services. 2017.
- Association AD. Economic costs of diabetes in the US in 2017. *Diabetes Care*. 2018;41(5):917–28.
- Boyle JP, Thompson TJ, Gregg EW, Barker LE, Williamson DF. Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Popul Health Metr*. 2010;8(1):29.
- Uusitupa M, Siitonen O, Aro A, Pyörälä K. Prevalence of coronary heart disease, left ventricular failure and hypertension in middle-aged, newly diagnosed type 2 (non-insulin-dependent) diabetic subjects. *Diabetologia*. 1985;28(1):22–7.
- Harris MI. Undiagnosed NIDDM: clinical and public health issues. *Diabetes Care*. 1993;16(4):642–52.
- Partanen J, Niskanen L, Lehtinen J, Mervaala E, Siitonen O, Uusitupa M. Natural history of peripheral neuropathy in patients with non-insulin-dependent diabetes mellitus. *N Engl J Med*. 1995;333(2):89–94.
- Evans JA, inventor; Azron Inc, assignee. Electronic medical records system. United States patent; US 5924074A. 1999 Jul 13. 2021. Available from: <https://patents.google.com/patent/US5924074A/>.
- Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database

- Analysis Task Force Report—Part II. Value in Health. 2009;12(8):1053–61.
13. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inf.* 2008;77(5):291–304.
  14. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. *Med Care.* 2012;50:S30.
  15. Navaneethan SD, Jolly SE, Schold JD, Arrigain S, Saupe W, Sharp J, et al. Development and validation of an electronic health record-based chronic kidney disease registry. *Clin J Am Soc Nephrol.* 2011;6(1):40–9.
  16. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ.* 2017;357:j2099.
  17. 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: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol.* 2014;63(25):2935–59.
  18. Basu S, Sussman JB, Berkowitz SA, Hayward RA, Yudkin JS. Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODE) using individual participant data from randomised trials. *Lancet Diabetes Endocrinol.* 2017;5(10):788–98.
  19. Bazzano AN, Wharton MK, Monnette A, Nauman E, Price-Haywood E, Glover C, et al. Barriers and facilitators in implementing non-face-to-face chronic care management in an elderly population with diabetes: a qualitative study of physician and health system perspectives. *J Clin Med.* 2018;7(11):451.
  20. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc.* 2014;21(4):578–82.
  21. Hong D, Shen Y, Monnette A, Liu S, Shao H, Nauman E, et al. Does the encounter type matter when defining diabetes complications in electronic health records? *Med Care.* 2020;58:S53–9.
  22. Shen Y, Shi L, Nauman E, Katzmarzyk PT, Price-Haywood EG, Bazzano AN, et al. Inverse association between HDL (high-density lipoprotein) cholesterol and stroke risk among patients with Type 2 Diabetes Mellitus. *Stroke.* 2019;50(2):291–7.
  23. Shen Y, Shi L, Nauman E, Katzmarzyk PT, Price-Haywood EG, Yin P, et al. Race and sex differences in rates of diabetic complications. *J Diabetes.* 2019;11(6):449–56.
  24. Nichols GA, Desai J, Lafata JE, Lawrence JM, O'Connor PJ, Pathak RD, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. *Prev Chronic Dis.* 2012. <https://doi.org/10.5888/pcd9.110311>.
  25. Wang Y, Katzmarzyk P, Horswell R, Zhao W, Johnson J, Hu G. Comparison of the heart failure risk stratification performance of the CKD-EPI equation and the MDRD equation for estimated glomerular filtration rate in patients with Type 2 diabetes. *Diabet Med.* 2016;33(5):609–20.
  26. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika.* 2007;94(3):691–703.
  27. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245.
  28. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* 2000;56(2):337–44.
  29. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J Am Stat Assoc.* 1988;83(402):414–25.
  30. Royston P. Tools for checking calibration of a Cox model in external validation: approach based on individual event probabilities. *Stand Genomic Sci.* 2014;14(4):738–55.
  31. R Development Core Team. The R Project for Statistical Computing. 2021. Available from: <https://www.r-project.org/>.
  32. Pavlou M, Ambler G, Seaman SR, Guttman O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ.* 2015;351:h868.
  33. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ.* 2016;353:i2416.