Journal of the American Medical Informatics Association, 2024, **31(8)**, 1763–1773
https://doi.org/10.1093/jamia/ocae140
Advance access publication 19 June 2024
Research and Applications

Research and Applications

A novel hyperparameter search approach for accuracy and simplicity in disease prediction risk scoring

Yajun Lu, PhD[1], Thanh Duong, BS[2,3], Zhuqi Miao, PhD[4], Thanh Thieu, PhD[3,5], Jivan Lamichhane, MD[6], Abdulaziz Ahmed, PhD[7], Dursun Delen, PhD[8,9,]*

[1]Department of Management and Marketing, Jacksonville State University, Jacksonville, AL 36265, United States, [2]Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, United States, [3]Department of Machine Learning, Moffitt Cancer Center and Research Institute, Tampa, FL 33612, United States, [4]School of Business, The State University of New York at New Paltz, New Paltz, NY 12561, United States, [5]Department of Oncological Sciences, University of South Florida Morsani College of Medicine, Tampa, FL 33612, United States, [6]The State University of New York Upstate Medical University, Syracuse, NY 13210, United States, [7]Department of Health Services Administration, School of Health Professions, The University of Alabama at Birmingham, Birmingham, AL 35233, United States, [8]Center for Health Systems Innovation, Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK 74078, United States, [9]Department of Industrial Engineering, Faculty of Engineering and Natural Sciences, Istinye University, Sariyer/Istanbul 34396, Turkey

*Corresponding author: Dursun Delen, PhD, Center for Health Systems Innovation, Department of Management Science and Information Systems, Oklahoma State University, Business Building, Stillwater, OK 74078, United States (dursun.delen@okstate.edu)

Abstract

Objective: Develop a novel technique to identify an optimal number of regression units corresponding to a single risk point, while creating risk scoring systems from logistic regression-based disease predictive models. The optimal value of this hyperparameter balances simplicity and accuracy, yielding risk scores of small scale and high accuracy for patient risk stratification.

Materials and Methods: The proposed technique applies an adapted line search across all potential hyperparameter values. Additionally, DeLong test is integrated to ensure the selected value produces an accuracy insignificantly different from the best achievable risk score accuracy. We assessed the approach through two case studies predicting diabetic retinopathy (DR) within six months and hip fracture readmissions (HFR) within 30 days, involving cohorts of 90 400 diabetic patients and 18 065 hip fracture patients.

Results: Our scores achieve accuracies insignificantly different from those obtained by existing approaches, reaching AUROCs of 0.803 and 0.645 for DR and HFR predictions, respectively. Regarding the scale, our scores ranged 0-53 for DR and 0-15 for HFR, while scores produced by existing methods frequently spanned hundreds or thousands.

Discussion: According to the assessment, our risk scores offer simple and accurate predictions for diseases. Furthermore, our new DR score provides a competitive alternative to state-of-the-art risk scores for DR, while our HFR case study presents the first risk score for this condition.

Conclusion: Our technique offers a generalizable framework for crafting precise risk scores of compact scales, addressing the demand for user-friendly and effective risk stratification tool in healthcare.

Key words: disease prediction; risk scoring system; hyperparameter search; electronic health record.

Introduction

Risk scoring systems have emerged as a favored approach to predict a range of health conditions in diverse healthcare settings. Notable examples include the Framingham Risk Scores[1,2] and SCORE[3] for foreseeing coronary heart disease, LACE[4] and HOSPITAL[5] for anticipating death or readmission after hospital discharge, IScore[6] for predicting death and disability after an acute stroke, and Mortality Risk Score[7] for estimating mortality in adults. These risk scoring systems often trace their development methodology back to the regression coefficient-based scoring principles.[8] Building upon these foundational principles, Sullivan et al[9] presented a comprehensive and systematic approach that has found significant traction in real-world healthcare scenarios and has been employed in creating well-known scoring systems such as the Framingham Risk Score and LACE.

The benefits of risk score systems are manifold. Firstly, they can provide clinicians with an easy-to-understand tool for estimating patient risk and making informed medical decisions.[2,10] By utilizing score systems, healthcare professionals can assess the likelihood of specific health outcomes or complications, aiding in treatment plans and preventive measures.[11,12] Additionally, a user-friendly risk score system also promotes patient engagement and behavior change. When patients understand their risk scores, they are more likely to comprehend potential health consequences, leading to active participation in health management and adopting beneficial lifestyle changes.[9]

Although the risk score system offers many advantages, little improvement has been made to the score derivation methodology since the earlier work performed by Sullivan et al.[9] A notable gap pertains to a hyperparameter defined as *the*

Received: November 29, 2023. **Revised:** May 7, 2024. **Editorial Decision:** May 27, 2024; **Accepted:** May 28, 2024

number of regression units in the disease prediction model to be mapped to a single point in the risk scoring system. A typical example of the "regression units" is the log-odds in the logistic regression model. For simplicity, we henceforth denoted this hyperparameter as B. Specifically, the gap is that effective approaches in determining a suitable value for B have not been adequately explored. The B value is important as it determines the granularity of the risk score. Higher granularity, corresponding to low B values, means using more risk score points to correspond to a given amount of regression-modeled risk. It results in a larger and more complex scale for the score system, which may introduce practical inconveniences in real-world implementation. On the plus side however, an intuitive benefit of the highly granular scoring system is that it captures a greater amount of information from the original regression model, consequently preserving better predictive accuracy, as measured by area under ROC curve (AUROC).[13] When a scoring system has low granularity, intuitively, it sacrifices information from the regression model, thereby compromising accuracy. Nonetheless, relatively low granularity results in a smaller scale that finds widespread adoption in real-world healthcare settings due to its simplicity. Notable examples of risk scores with narrowed ranges that have gained significant usage in practical healthcare contexts include LACE[4] and HOSPITAL.[5] Therefore, addressing the challenge posed by scoring systems with either a low granularity, resulting in reduced predictive precision, or a highly granular scale leading to practical inconveniences, necessitates the development of a scoring approach that strikes a balance between the scale simplicity and the prediction accuracy. Although grid search, a classical hyperparameter tuning technique in machine learning, may be used to handle the issue, it can be computationally intensive when dealing with a multitude of hyperparameters, each having a wide range of possible values.[14]

In order to fill the gap, we have established two main *objectives* for this study: (1) Develop a novel hyperparameter search algorithm to identify the "best" amount of regression units in a disease prediction model, which should correspond to a single point in a risk scoring system for achieving a balance between the scale and accuracy for the risk score. (2) Assess the algorithm's ability to generate compact-scale risk scores that preserve the majority of predictive accuracy from the root regression models by conducting two case studies, one on predicting diabetic retinopathy (DR) and the other on predicting hip fracture readmission (HFR).

Methods

Data source and preprocessing

In this study, we utilized the Oracle Cerner Health Facts Electronic Health Records (EHR) data warehouse as our data source. Health Facts comprises clinical data extracted from over 200 hospitals across the United States that operate on Cerner EHR systems during 2000-2018. The data encompasses a wide range of information, including patients' timestamped encounters, demographics, diagnoses, procedures, medications, laboratory results, vital signs, etc. Oracle Cerner collects and integrates the data in accordance with established procedures that adhere to the Health Insurance Portability and Accountability Act (HIPAA) laws. The Institutional Review Boards (IRB) at Oklahoma State University (OSU) exempted the study from review because the data has been completely de-identified according to HIPAA regulations. All the data collection, preprocessing, and analysis involved in this study were performed on the devices hosted at OSU.

Our two case studies involved leveraging large-scale EHR datasets from Health Facts to predict DR and HFR. DR is a complication of diabetes that can cause vision loss or blindness over time if not diagnosed early enough and left untreated.[15,16] Hip fractures (HF) significantly increase morbidity and mortality in older adults, frequently resulting in post-discharge readmissions.[17,18] Both are significant conditions drawing extensive research attention and warranting further investigation.

- DR Data: We extracted the DR case and control cohorts using the same diagnosis codes and a similar cohort derivation method as utilized in a prior study by Wang et al.[19] Together with the cohorts, we gathered 31 variables related to patient's demographics, duration of diabetes, complications, and laboratory results, all of which have been shown in the literature to be significantly associate with DR.[20–24] In the predictive modeling, the values of these variables, during a two-year window that was 6 months preceding the first diagnosis of DR, were averaged to predict whether DR would occur within the 6-month period. This approach models the DR prediction in six months given a diabetic encounter and history in past two years.[19,25] Subsequently, we applied the complete-case preprocessing method to these variables. After preprocessing, the variables maintained distributions close to those of the raw data (the distribution plots are available in Section B of the Supplementary Material). Next, we applied a machine learning-based ensemble predictor selection method[26] to identify a reduced set of key predictors from the 31 variables. These key predictors enabled us to create a concise yet accurate risk scoring system.
- HFR Data: Regarding the selection of the patient cohort for HFR, we extracted data from Health Facts and followed a cohort derivation method similar to that used in a prior HFR study.[27] The data comprised patient demographics, historical visits, diagnoses, procedures, and seven laboratory results, constituting the initial set of variables for our analysis. To maintain methodological consistency across both case studies, we applied the same complete-case preprocessing and predictor selection methods used for DR data to this HFR dataset as well. Beyond the processing, the selected key predictors were utilized to predict all-cause readmissions within 30 days from the HF inpatient visits.

Detailed diagnosis and procedure codes for patient extraction, flow charts outlining data preprocessing, and initial sets of variables for analysis for the two case studies are provided in Section A of the Supplementary Material. For both study cohorts, we randomly partition the data into training (70%) and test subsets (30%) for predictive analysis.

Risk score derivation methods

Figure 1 shows a risk scoring framework adapted from the one established by Sullivan et al,[9] serving as the foundational pipeline for our risk score derivation. Our novel

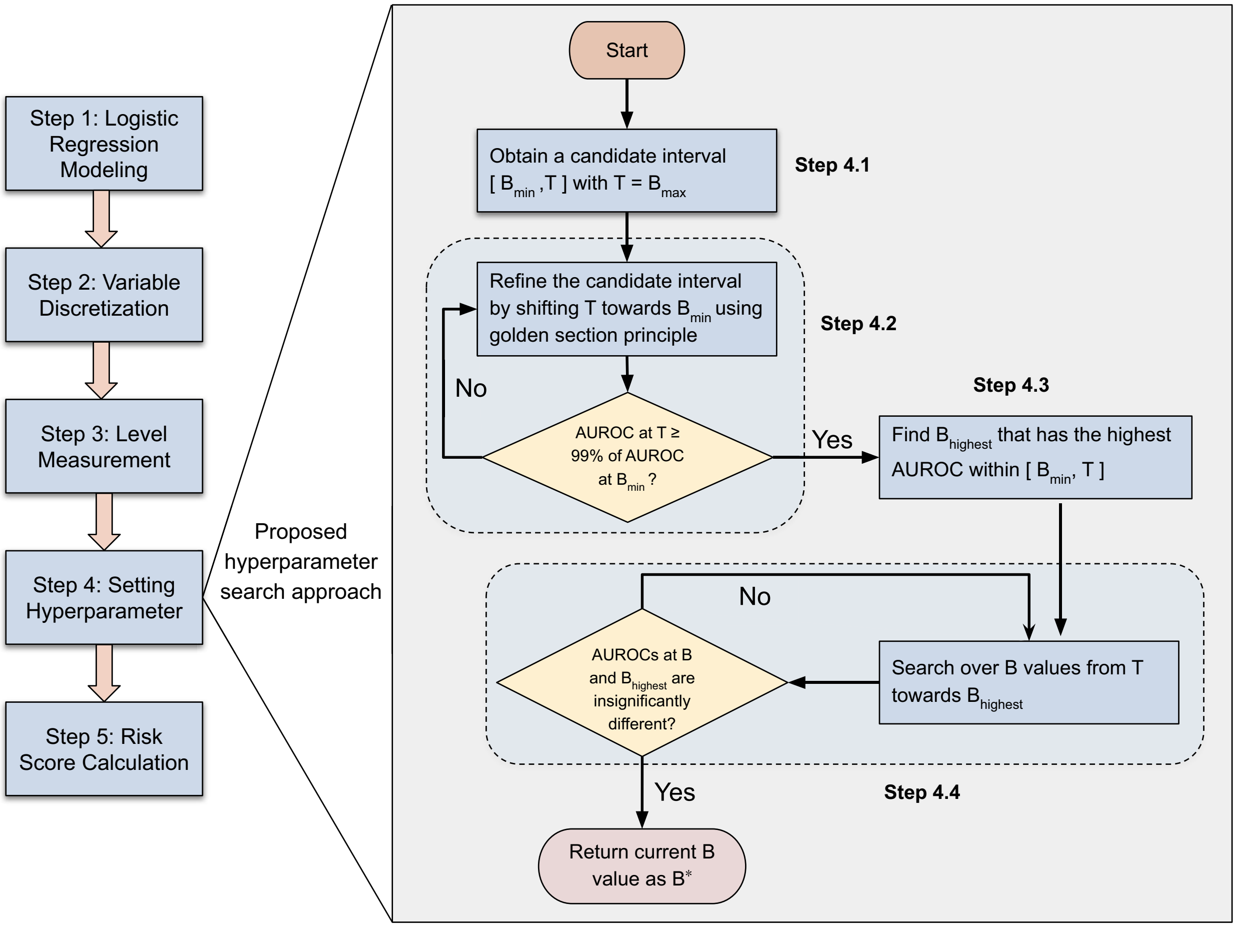
Step 1: Logistic Regression Modeling
Step 2: Variable Discretization
Step 3: Level Measurement
Step 4: Setting Hyperparameter
Step 5: Risk Score Calculation
Proposed hyperparameter search approach
Start
Obtain a candidate interval [B_{min} ,T] with T = B_{max}
Step 4.1
Refine the candidate interval by shifting T towards B_{min} using golden section principle
Step 4.2
No
AUROC at T ≥ 99% of AUROC at B_{min} ?
Yes
Step 4.3
Find $B_{highest}$ that has the highest AUROC within [B_{min}, T]
No
AUROCs at B and $B_{highest}$ are insignificantly different?
Search over B values from T towards $B_{highest}$
Yes
Step 4.4
Return current B value as B^*

Figure 1. Flowchart illustrating the risk score derivation framework and our refinements in Step 4. In this illustration, *B* denotes the number of regression units in the disease prediction model to be mapped to a single point in the risk scoring system. *T* is the right end of the candidate interval of *B* values, and updated iteratively in the algorithm.

hyperparameter search approach centers on the step of *Setting Hyperparameter B*, aiming to develop simple yet accurate risk scores, as detailed in the following.

Scoring framework

Step 1. Logistic Regression Modeling: Construct a logistic regression model to predict the presence of a health condition (modeled as a binary target variable y) based on n predictors, denoted by $x_1, x_2, \ldots, x_n$. The model can be represented as Equation (1):

$$\ln \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \tag{1}$$

where p represents the probability that $y = 1$, indicating that patients developed DR or were readmitted respectively in our two case studies. The value $\ln \frac{p}{1-p}$, known as "log-odds," is used to model patient risk of having the health condition. While β_0 is the intercept and β_i represents the coefficient for the predictor x_i.

Step 2. Variable Discretization: Convert continuous predictors to categorical variables by discretizing them into multiple intervals (aka the levels of the resulted ordinal categorical variables) using meaningful cutoffs based on medical expertise. Statistical methods, such as percentile-based cutoffs, are frequently employed to discretize continuous variables in the literature.[12] Our comparison (provided in Section C of the Supplementary Material) revealed minor differences in AUROC between the medically meaningful and statistical cutoffs. Hence, in this article, we focus on reporting the results based on medically meaningful discretization due to the clinical relevance and interoperability of this approach.

Step 3. Regression Unit Measurement for Levels: This step involves measuring the regression units, specifically log-odds in our case, for every level of each categorical variable. The measurement follows the subsequent procedure: Given any variable x_i, we first determine a reference value for each level, which is the mid-value for intervals and a modeled value in logistic regression for categorical variables (eg 0 for modeling female and 1 for modeling male). Then, the level with the lowest reference value corresponds to the lowest-risk level if the coefficient is positive; otherwise, it is the level with the highest reference value. Denote the reference value of the lowest-risk level as

W_{min}, then for a level of x_i with reference value of W, the log-odds assigned to the level are expressed as $\beta_i(W - W_{min})$.

Step 4. Setting Hyperparameter B: The hyperparameter, in this case, is the number of log-odds corresponding to a single risk score point. It can be determined by multiplying the coefficient of a selected base variable by a factor, such as $\beta_{age} \times 5$. However, the approaches for selecting the base variable and the factor in current literature lack a delicate design to generate risk scores that achieve both high simplicity and accuracy. Our novel hyperparameter search algorithm, elaborated in the next subsection on "Hyperparameter Searching," addresses this gap, constituting the primary innovation of this study.

Step 5. Risk Score Calculation: Once B is determined, the associated risk score for each level of a predictor can be calculated using the formula $\beta_i(W - W_{min})/B$ and round it to the nearest integer. The overall risk score of a patient will be the sum of the risk scores corresponding to each variable's measurement of the patient.

Hyperparameter searching

As discussed in the Introduction section, a smaller value for the hyperparameter B leads to a more granular risk score, preserving greater predictive power from the regression model. However, it may result in an unnecessarily large scale, posing inconvenience for clinical applications. On the other hand, a larger B value yields a simpler scale but incurs a loss of accuracy. Hence, a clever choice of the B value is crucial for simplifying the risk score system without compromising accuracy. Many scores used a multiple of β_{age},[4,6,7,9] to account for increasing risk associated with aging, while some other studies employed the smallest coefficient[5,12] to ensure all scores to be larger than one. However, none of the approaches adequately considered both the scale and accuracy of the risk score. Grid-search-based enumeration across all predictors and all potential factor values for each predictor is an intuitive approach to tackle the issue, but it can be computationally expensive and time-consuming.[14] Our new approach, rather than engaging in a two-dimensional search across variables and factors, executes a uni-dimensional search directly over all feasible B values. The flow diagram is illustrated in Figure 1, with steps explained below:

Step 4.1 Obtain all possible B values by multiplying the coefficient of each variable by all potential factor values (we used $1, 2, \ldots, 10$ in our implementation). Then, sort the resulting B values in an ascending order and define a candidate interval $[B_{min}, T]$ with $T = B_{max}$ initially to cover the entire range of B values.

Step 4.2 Iteratively refine the candidate interval by adjusting the right endpoint T from B_{max} towards B_{min} until the accuracy at T reaches at least 99% of the accuracy at B_{min}. In our implementation, we measure accuracy using AUROC. The endpoint adjustment adheres to the golden section principle.[28] In other words, for each iteration, the new value of T is updated as $T\prime - 0.382 \times (T\prime - B_{\min})$, where $T\prime$ represents the previous value of T.

Step 4.3 Within the refined candidate interval $[B_{min}, T]$, identify the B value associated with the highest AUROC, denoted as $B_{highest}$.

Step 4.4 Search from the right endpoint of the refined candidate interval T towards $B_{highest}$ to find the first B value whose AUROC is insignificantly different from that of $B_{highest}$ *via* DeLong test at a 0.01 significance level.[29] Finally, return the found B value, denoted as B^*.

The benefit of this search strategy is that we can leverage the intuition that with the increase of B, the AUROC demonstrates an overall declining trend as larger B tends to yield less granularity in the risk score. The trend enables us to perform directional search to find a suitable B value sooner. Specifically, *Step* 4.2 enables us to quickly skip B values close to the right end of the trend that are associated with low accuracies, as illustrated in Figure 2. Furthermore, once the refined candidate interval is determined, the search from T to $B_{highest}$, as described in *Step* 4.4, saves effort of performing DeLong test exhaustively for the B values less than $B_{highest}$.

All the data cleaning, analysis and algorithm development presented in this article were implemented using Python 3.10. The logistic regression models used in this study were created and executed using the "*glm*()" function from the Python *statsmodels* 0.14.0 module. Our code is publicly available on GitHub at https://github.com/yajun668/RiskScoring.

Results

Descriptive statistics of case study cohorts

After preprocessing, our DR cohort consisted of 90 400 diabetic patients, among whom 3380 were diagnosed with DR. The HFR cohort included 18 065 HF patients, among whom 2055 were readmitted to the hospital within 30 days from their HF inpatient visits. The selected key predictors and their detailed statistics within the training and test datasets of the two cohorts are presented in Table 1, showing insignificant difference across all variables except for creatinine between the training and test datasets.

Trend between AUROC and *B*

Figure 3A and C depicts the relationships between AUROC and B values for DR and HFR predictions, respectively. Both plots demonstrate a consistent downward trend, aligning with the intuitive expectation that higher B values lead to lower granularity of the risk scoring system, ultimately compromising its accuracy. Figure 3B and D provides zoomed-in views of the refined candidate intervals, showing that $B_{highest}$ does not necessarily coincide with the smallest B. Furthermore, many AUROCs in the interval appear very close, indicating that towards the right-hand side of the interval, there are competitive B values that could result in narrower scales of risk scores, with statistically insignificant differences in accuracy compared to that at $B_{highest}$. All the observations favorably support the design of our proposed hyperparameter search algorithm.

Score system comparison

To assess the effectiveness of our proposed approach, we compared the risk scores developed using B^* with those derived based on $5\beta_{\text{age}}$ and $B_{\min}$—two commonly used values

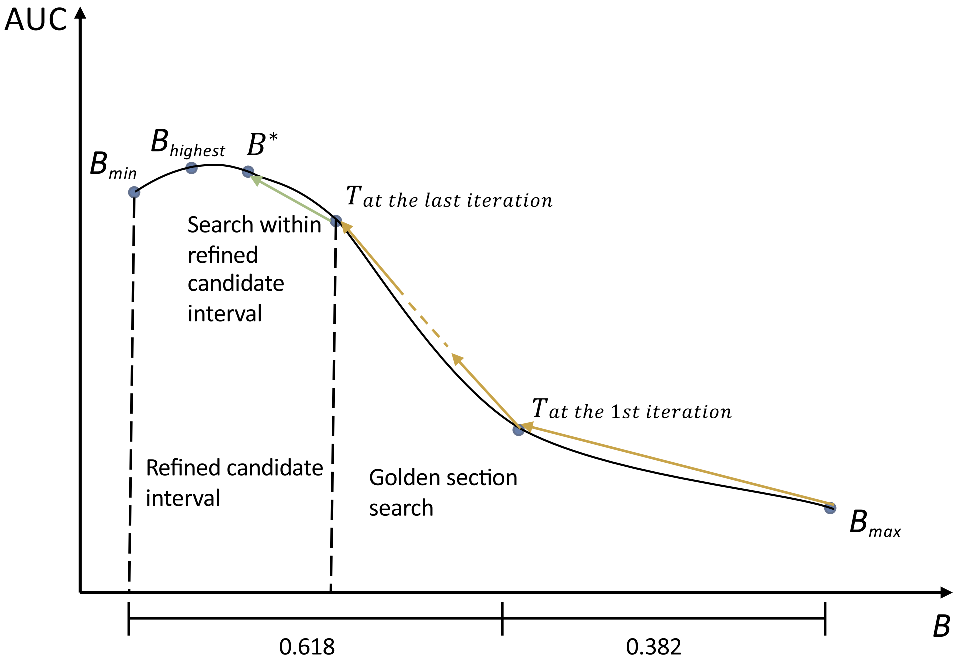
AUC
B_{min}
$B_{highest}$
B^*
$T_{at\ the\ last\ iteration}$
Search within refined candidate interval
$T_{at\ the\ 1st\ iteration}$
Refined candidate interval
Golden section search
B_{max}
B
0.618
0.382

Figure 2. An illustration of the search trajectory of the developed hyperparameter search algorithm.

for the hyperparameter B in the literature. Table 2 summarizes the AUROCs and scales of the risk scores across different B values. Upon observation, our algorithm consistently generates risk scores closely aligned with those obtained using other B values in terms of AUROC. It is worth noting that, though the AUROC associated with B^* is slightly lower than those associated with other B values for HFR, the difference is statistically insignificant according to DeLong test. Remarkably, the scales of the risk scores derived through our approach exhibit much simpler ranges. In the context of DR prediction, our risk score ranged up to only 53, in contrast to 191 for $5\beta_{\mathrm{age}}$ and 11 018 for B_{min}. Similarly, for HFR prediction, our highest score is only 15, whereas $5\beta_{\mathrm{age}}$ and B_{min} have ranges with the highest score as high as 93 and 18 728, respectively.

We additionally compared the risk scores with the corresponding logistic regression models—the root model from which the scores are derived—in terms of AUROC. The AUROC plots are displayed in Figure 4, illustrating marginal differences, up to 0.029 (as observed for $5\beta_{age}$ for DR), between the predictive accuracy achieved by the risk scores and that of logistic regressions. This aligns with what has been reported in the literature,[19] reiterating the effectiveness of the entire risk scoring framework in maintaining strong predictive capacity from logistic regressions. Figure 4 also shows comparable area under the precision-recall curve (AUPRC)[30] across evaluated risk scores for each condition. They all outperformed random models, suggesting their ability in differentiating patients by disease risk. DR risk scores outperformed the random model to a greater extent than HFR risk scores, which aligns with AUROC findings. However, the class imbalance in our data likely limited AUPRC performance. Integrating techniques for handling imbalanced data into regression models has the potential to improve AUPRC for risk scores.

Furthermore, we report the new risk score systems for DR and HFR in Table 3. Note that the risk score derived using $5\beta_{age}$ for DR, presented in the table, is essentially a variation of the score system proposed by Wang et al[19] with two additional predictors and slightly adjusted scores for certain levels. Compared to it, the risk score derived using our approach, B^*, significantly simplified the system, by aggregating many levels across a multitude of predictors, such as <60 and $[60, 80)$ for glucose, as well as "African American" and "Other" for race. These levels can be combined because they share the identical risk points. A similar finding can be observed for the HFR risk score as well. Many levels can be combined, for instance, <65 and $[65, 75)$ for age, and <9 and $[9, 9.7)$ for hemoglobin. More interestingly, our new scoring system requires an even more concise set of nine predictors, specifically BUN, hemoglobin, hematocrit, length of stay, preInp1Y, preER1Y, Charlson comorbidity index, age, and platelet count, rather than the 12 variables chosen by feature selection, because the other predictors exhibit 0 risk points across all levels, resulting in no effect on final risk score.

Discussion

The widespread deployment of EHR systems has made a tremendous volume of digitized clinical data available. Coupled with advancements in medical informatics and analytics, it has provided valuable and actionable insights for addressing a wide range of healthcare challenges, including the high-cost patients identification, disease prediction, patient triaging, and treatment plan optimization, among others.[31–34] Machine learning and deep learning models are often employed to tackle the challenges because of their high

Table 1. Descriptive statistics on training and test datasets for DR and HFR predictions.

	DR dataset					
	Training		Test		*P*-value[a]	
	Non-DR	DR	Non-DR	DR	Non-DR	DR
# Patient (%)	60 936 (96.3)	2344 (3.7)	26 084 (96.2)	1036 (3.8)	–	–
Creatinine, mean (SD)	1.06 (0.45)	1.96 (1.88)	1.06 (0.45)	1.80 (1.60)	0.355	0.018
HbA1c, mean (SD)	7.13 (1.50)	8.35 (2.03)	7.14 (1.51)	8.39 (1.97)	0.173	0.605
Diabetes duration, mean (SD)	1.92 (1.76)	2.75 (2.02)	1.92 (1.77)	2.76 (2.05)	0.999	0.929
White blood cell, mean (SD)	8.11 (2.19)	7.97 (2.21)	8.12 (2.19)	8.01 (2.31)	0.547	0.619
Glucose, mean (SD)	142.42 (46.17)	173.96 (61.61)	142.61 (46.24)	174.75 (62.36)	0.575	0.733
Age, mean (SD)	64.16 (14.08)	60.47 (13.37)	64.05 (14.07)	60.82 (12.77)	0.306	0.475
Hematocrit, mean (SD)	38.99 (4.71)	36.23 (4.72)	39.04 (4.69)	36.42 (4.71)	0.233	0.294
Sodium, mean (SD)	138.87 (2.46)	138.59 (2.37)	138.86 (2.45)	138.48 (2.39)	0.537	0.249
BUN, mean (SD)	19.66 (9.45)	27.45 (14.78)	19.67 (9.57)	26.82 (14.51)	0.846	0.245
Anion gap, mean (SD)	9.47 (2.55)	9.52 (2.71)	9.45 (2.55)	9.36 (2.62)	0.419	0.131
Nephropathy = yes (%)	3030 (5.0)	656 (28.0)	1281 (4.9)	278 (26.8)	0.715	0.516
Neuropathy = yes (%)	5197 (8.5)	782 (33.4)	2190 (8.4)	349 (33.7)	0.529	0.884
Race (%)						
African American	10 867 (17.8)	897 (38.3)	4669 (17.9)	377 (36.4)	0.417	0.435
Caucasian	45 355 (74.4)	1270 (54.2)	19 332 (74.1)	586 (56.6)		
Other	4714 (7.7)	177 (7.6)	2083 (8.0)	73 (7.0)		

	HFR dataset					
	Training		Test		*P*-value[a]	
	Non-HFR	HFR	Non-HFR	HFR	Non-HFR	HFR
# Patient	11 226 (88.78)	1419 (11.22)	4784 (88.27)	636 (11.73)	–	–
Age, mean (SD)	80.01 (9.80)	80.87 (9.22)	80.05 (9.74)	81.26 (9.04)	0.823	0.373
Length of stay, mean (SD)	5.32 (2.79)	6.18 (3.53)	5.34 (2.78)	6.27 (3.36)	0.819	0.607
Platelet count, mean (SD)	209.64 (81.53)	216.67 (89.54)	210.16 (82.10)	219.50 (91.35)	0.711	0.511
BUN, mean (SD)	19.53 (10.97)	24.01 (14.37)	19.55 (11.25)	23.87 (13.29)	0.904	0.835
Hemoglobin, mean (SD)	10.11 (1.34)	10.07 (1.31)	10.14 (1.35)	10.10 (1.29)	0.115	0.640
Creatinine, mean (SD)	0.97 (0.61)	1.15 (0.80)	0.97 (0.60)	1.11 (0.74)	0.748	0.307
Hematocrit, mean (SD)	30.06 (3.84)	30.09 (3.86)	30.17 (3.86)	30.12 (3.85)	0.093	0.860
CCI, mean (SD)	1.29 (1.48)	1.72 (1.61)	1.30 (1.48)	1.67 (1.61)	0.595	0.541
Potassium	4.02 (0.44)	4.06 (0.46)	4.01 (0.43)	4.04 (0.47)	0.200	0.452
Sodium	137.09 (3.74)	137.23 (4.03)	137.06 (3.77)	137.20 (3.81)	0.617	0.852
preInp1Y[b] (%)						
0	8335 (74.2)	926 (65.3)	3571 (74.6)	389 (61.2)	0.684	0.132
1	1809 (16.1)	244 (17.2)	745 (15.6)	131 (20.6)		
2	1082 (9.6)	249 (17.5)	468 (9.8)	116 (18.2)		
preER1Y[c] (%)						
0	7230 (64.4)	792 (55.8)	3068 (64.1)	321 (50.5)	0.905	0.063
1	2127 (18.9)	288 (20.3)	906 (18.9)	152 (23.9)		
2	1869 (16.6)	339 (23.9)	810 (16.9)	163 (25.6)		

Abbreviation: CCI = Charlson Comorbidity Index.

[a] The *P*-values are associated with the statistical tests comparing variable differences between the training and test datasets.

[b] Number of inpatient visits within 1 year before.

[c] Number of emergency department visits within 1 year before.

predictive accuracy.[35–37] However, the inherent black-box nature of machine/deep learning often poses challenges in interpreting the results for clinicians.[38] Additionally, many existing EHR systems in hospitals lack support for complex machine-learning models.[39]

In contrast, risk scores are easy to interpret, understand, and implement in healthcare settings, contributing to their considerable attention and real-world applications. The novel hyperparameter search algorithm developed in this study enable the creation of simple yet accurate risk scores, which can support medical decision making in various aspects of patient care. Firstly, risk scoring systems developed using comprehensive socioeconomic and clinical determinants enable healthcare professionals to compute patients' risk of developing specific conditions in the future. A high risk score can serve as early-warning tool, prompting timely intervention for effective care management. Additionally, the risk score's interpretability, along with insights into how each

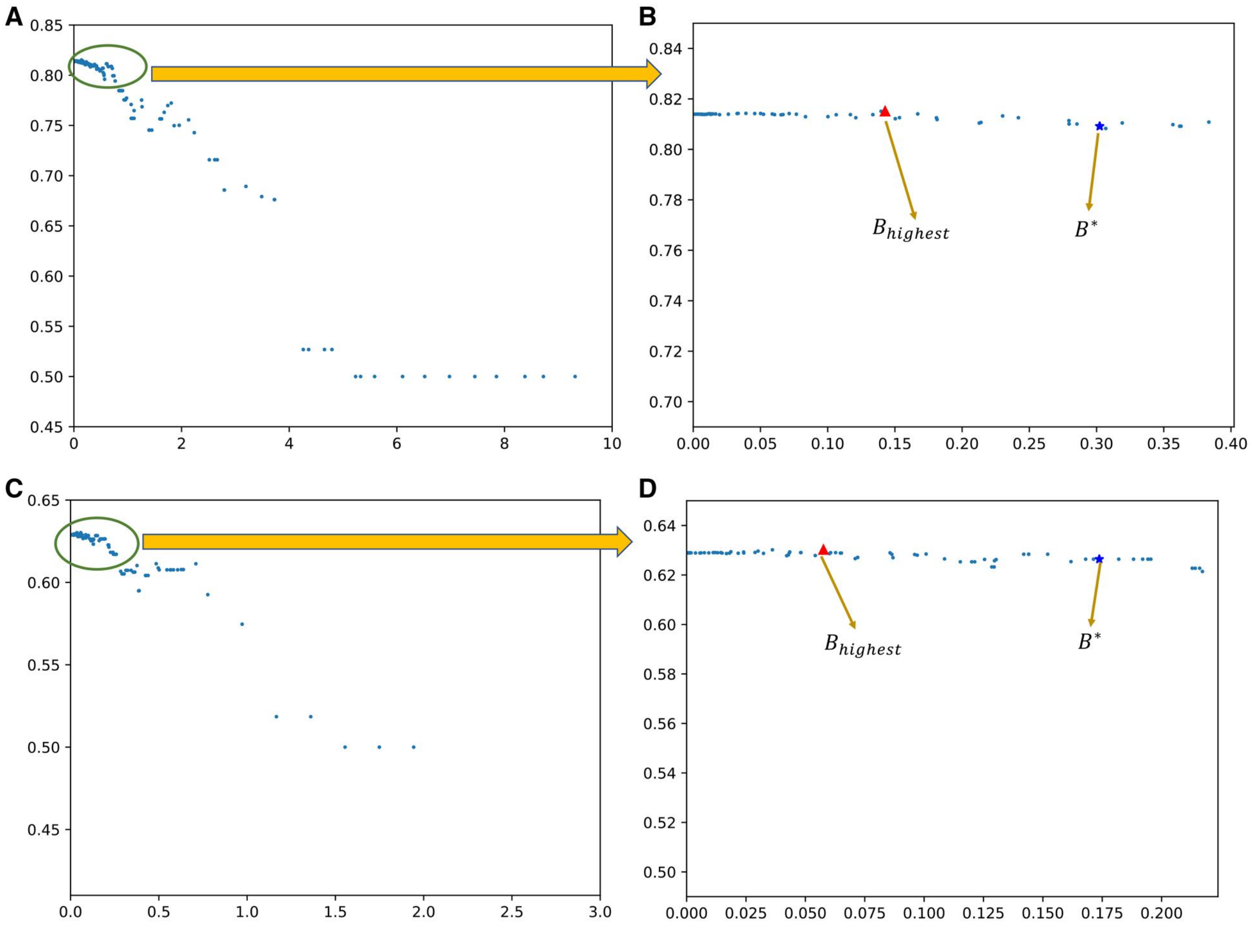

Figure 3. The AUROC and *B* relationships for DR (A and B) and HFR (C and D). (A) and (C) provide overall trends over all *B* values considered. (B) and (D) provide zoomed-in views of the AUROC-*B* relationship for DR and HFR, respectively.

Table 2. Comparison of AUROCs and risk score scales between the proposed method (B^*) and existing approaches ($B_{\min}$ and $5\beta_{age}$) for DR and HFR.

	AUROC			Score scale		
	B^*	$5\beta_{age}$	$B_{\min}$	B^*	$5\beta_{age}$	$B_{\min}$
DR	0.803	0.802	0.804	0-53	0-191	0-11 018
HFR	0.645	0.651	0.652	0-15	0-93	0-18 728

feature contributes to the total risk score, empowers patients to better grasp the factors that pose health risks. With the knowledge, patients are more likely to take action to address the factors that negatively impact their health.[40,41]

Compared to the state-of-the-art DR risk score,[19] our new DR risk score system, generated using the algorithm proposed in this study, exhibits equivalently high accuracy with a significantly simpler scale. As for the risk score for HFR, to the best of our knowledge, this is the first study in developing a risk score system for this condition. The two new risk score systems not only demonstrate the effectiveness of our proposed approach but also offer highly potential alternatives, once externally validated, for the prediction and risk stratification for DR and HFR respectively. Furthermore, while our case studies concentrated solely on two conditions, DR and HFR, our approach can serve as a general framework for developing risk scores for other health conditions as well.

Recommendations for accurate risk scoring

Our technique enables the creation of concise risk scores closely mirroring the accuracy of regression models. Therefore, robust regression models are the cornerstone for accurate risk scoring. Various factors spanning data collection, preprocessing, modeling, and deployment influence regression modeling, subsequently the accuracy of risk score, in real-world disease prediction applications. Key considerations encompass data representativeness, incorporation of comprehensive socioeconomic and clinical variables, handling missing values, addressing data imbalance, and the geographical and care setting differences between modeling and deployment. Analysts should select data from sources aligned with the geographic and care settings where the model will be deployed and thoroughly evaluate the data quality before modeling to ensure the relevance and robustness of their models in the target settings.[42,43] Various methods exist for handling missingness and imbalance within health data, yet there is a lack of widespread consensus and acceptance within the scientific community regarding the most effective methodology.[44,45] Distinct methodologies yield different models and predictive outcomes.[46] Analysts should carefully consider

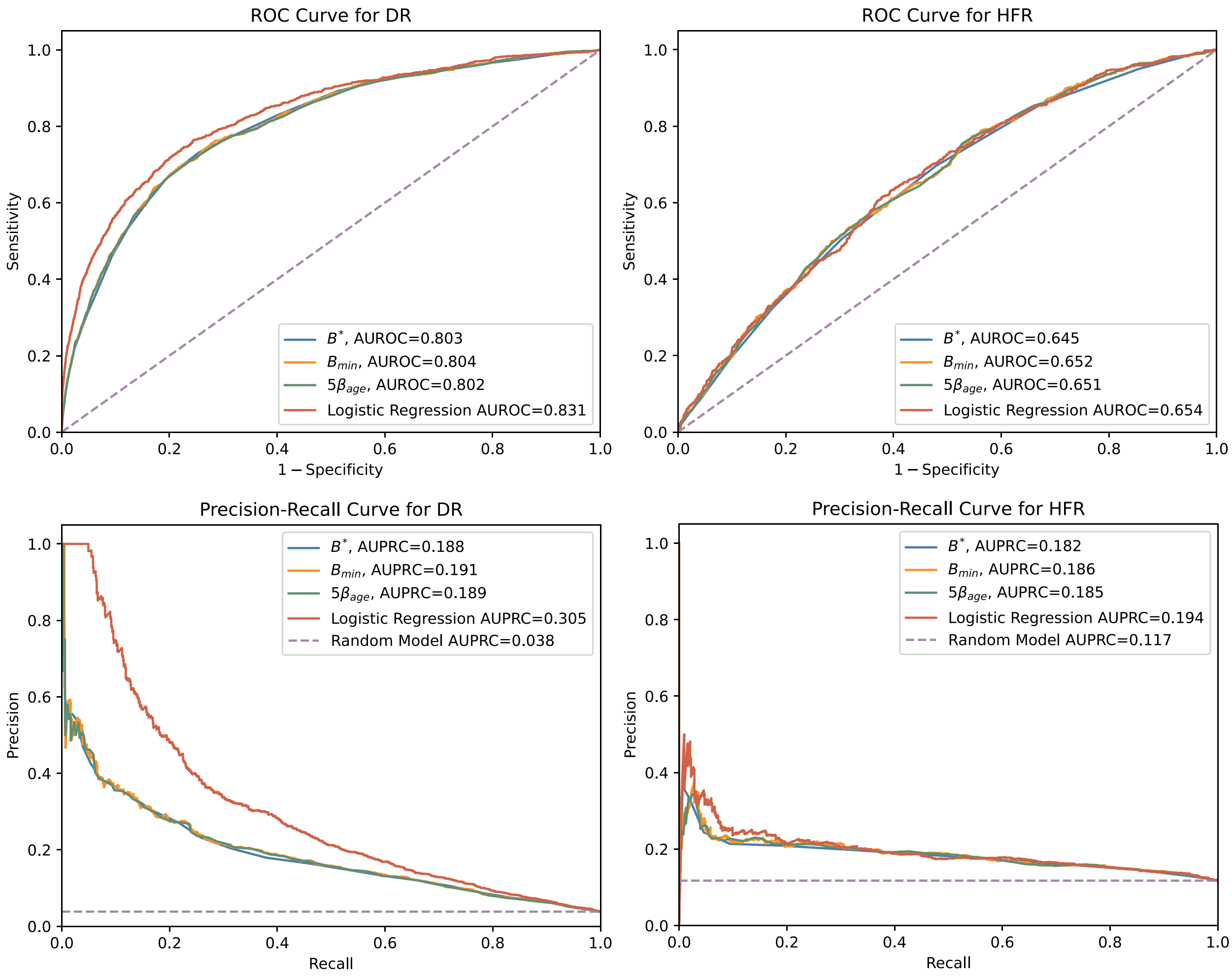

Figure 4. Comparison among risk scores and logistic regressions in AUROC and AUPRC.

various aspects, such as clinical relevance, interpretability, generalizability, and accuracy, for optimal model selection. When incorporating variables, analysts should thoroughly examine available metrics concerning both health and socioeconomics, then leverage clinical and biological expertise, along with feature selection techniques,[47] to identify essential predictors for accurate modeling. EHR data is rich in clinical data but often lacks socioeconomic variables, which are crucial for understanding and addressing many health conditions.[48–52] By integrating socioeconomic factors into EHR,[53,54] the endeavor of crafting more comprehensive and accurate risk scores can be significantly bolstered.

Limitations

There are several limitations with this study. (1) Our risk scoring algorithm is essentially a statistical approach revealing associations rather than establishing causality. The generated risk scores should be viewed as decision support tools for healthcare professionals, with application and interpretation contingent upon clinical expertise. (2) In the case studies, we optimized risk scores for high AUROCs. Other accuracy measures were not necessarily preserved to the same degree. (3) Socioeconomic factors were unavailable within the EHRs used, thus omitted from the case studies. (4) The risk scores for DR and HFR created in the case studies require validation using external data.

Conclusion

In this study, we introduce a novel hyperparameter search algorithm intended to automatically determine an optimal amount of log-odds that should be calibrated to a single score in a risk scoring system to achieve a balance between accuracy and simplicity within the risk scoring system. The implications of our proposed approach in healthcare settings are substantial as it delivers simple yet accurate risk scores that support healthcare professionals and decision makers in patient stratification, treatment planning, and various medical decision-making processes. Additionally, on the patient side, the risk score encourages them to adopt healthier behaviors, undergo early screenings, and prioritize preventive measures before conditions deteriorate. Our future research will focus on evaluating the developed approach across a broader spectrum of health conditions and conducting external validations for our new DR and HFR risk score systems. In addition, exploring improved algorithms that can balance

Table 3. Risk scoring systems derived using B^*, $5\beta_{age}$, and $B_{\min}$ for DR and HFR.

Variable	Level	Risk score for DR		
		B^*	$5\beta_{age}$	$B_{\min}$
Neuropathy	No	0	0	0
	Yes	3	11	638
Nephropathy	No	0	0	0
	Yes	2	6	365
Creatinine	<0.5	0	0	0
	[0.5, 1)	1	4	203
	[1, 1.5)	2	9	502
	[1.5, 2)	4	14	801
	≥2	7	24	1357
HbA1c	<6	0	0	0
	[6, 8)	2	7	383
	[8, 10)	4	13	767
	[10, 12)	6	20	1150
	≥12	8	30	1725
Diabetes duration	<1	0	0	0
	[1, 2)	0	2	96
	[2, 3)	1	3	192
	[3, 4)	1	5	287
	≥4	4	15	833
White blood cell	<4	4	14	773
	[4, 6)	3	12	694
	[6, 8)	3	10	589
	[8, 12)	2	8	431
	≥12	0	0	0
Glucose	<60	0	0	0
	[60, 80)	0	1	76
	[80, 100)	1	3	166
	[100, 200)	2	8	434
	≥200	7	24	1393
Age	<35	3	12	704
	[35, 50)	2	9	515
	[50, 65)	2	6	343
	[65, 75)	1	4	200
	[75, 85)	0	2	86
	≥85	0	0	0
Hematocrit	<30	6	21	1215
	[30, 35)	5	16	933
	[35, 40)	4	13	725
	[40, 50)	2	7	415
	≥50	0	0	0
Sodium	<136	0	0	0
	[136, 144)	3	11	620
	≥144	5	19	1094
BUN	<11	0	0	0
	[11, 15)	0	0	3
	[15, 19)	0	0	7
	[19, 27)	0	0	14
	≥27	0	0	24
Anion gap	<5	3	11	648
	[5, 7)	3	10	575
	[7, 10)	2	8	453
	[10, 12)	2	6	330
	[12, 17)	1	3	159
	≥17	0	0	0
Race	African American	1	4	248
	Other	1	2	124
	Caucasian	0	0	0

Variable	Level	Risk score for HFR		
		B^*	$5\beta_{age}$	$B_{\min}$
BUN	<11	0	0	0
	[11, 15)	1	4	727
	[15, 19)	1	6	1308
	[19, 27)	2	11	2180
	≥27	3	19	3778
Hemoglobin	<9	1	5	945
	[9, 9.7)	1	3	702
	[9.7, 10.3)	0	3	516
	[10.3, 11.1)	0	2	315
	≥11.1	0	0	0
Hematocrit	<26.9	0	0	0
	[26.9, 28.8)	0	2	377
	[28.8, 30.7)	1	3	682
	[30.7, 33.1)	1	5	1028
	≥33.1	1	8	1573
Length of stay	<5	0	0	0
	[5, 7)	1	4	811
	[7, 14)	2	13	2637
	≥14	4	24	4869
preInp1Y[a]	0	0	0	0
	1	1	6	1300
	≥2	2	13	2600
preER1Y[b]	0	0	0	0
	1	0	2	475
	≥2	1	5	949
Charlson Comorbidity Index	<4	0	0	0
	[4, 6)	1	8	1678
	≥6	2	10	2098
Age	<65	0	0	0
	[65, 75)	0	2	443
	[75, 80)	1	4	745
	[80, 85)	1	5	946
	[85, 90)	1	6	1147
	≥90	1	6	1248
Platelet count	<143	0	0	0
	[143, 177)	0	0	39
	[177, 213)	0	0	74
	[213, 268)	1	1	120
	≥268	1	1	200
Potassium	<3.6	0	1	260
	[3.6, 3.9)	0	1	184
	[3.9, 4.1)	0	1	130
	[4.1, 4.4)	0	0	76
	≥4.4	0	0	0
Sodium	<134	0	1	112
	[134, 136)	0	0	75
	[136, 138)	0	0	50
	[138, 140)	0	0	25
	≥140	0	0	0
Creatinine	<0.6	0	0	92
	[0.6, 0.8)	0	0	75
	[0.8, 0.9)	0	0	61
	[0.9, 1.2)	0	0	39
	≥1.2	0	0	0

[a] Number of inpatient visits within 1 year before.
[b] Number of emergency department visits within 1 year before.

multiple accuracy measures while streamlining the risk score scale presents an intriguing avenue for future work.

Acknowledgments

The authors gratefully acknowledge Oracle Cerner for sharing Health Facts EHR database to support this research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Oracle Cerner.

Author contributions

Yajun Lu (Conceptualization, Data curation, Methodology, Coding, Formal analysis, Writing—original draft, Writing—review & editing), Thanh Duong (Data curation, Coding, Formal analysis), Zhuqi Miao (Conceptualization, Data curation, Methodology, Coding, Formal analysis, Writing—original draft, Writing—review & editing), Thanh Thieu (Conceptualization, Methodology, Formal analysis), Jivan Lamichhane (Writing—review & editing, Validation), Abdulaziz Ahmed (Writing—review & editing, Validation), and Dursun Delen (Conceptualization, Writing—review & editing).

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest

None declared.

Data availability

The data used in this study were obtained from Oracle Cerner. Interested researchers may request the associated data directly from Oracle Cerner.

Code availability

The code supporting the findings of this study is shared publicly on GitHub at https://github.com/yajun668/RiskScoring.

Ethics information

The Institutional Review Boards (IRB) at Oklahoma State University exempted the study from review because the data have been completely de-identified according to HIPAA regulations. The entire process of data collection and analysis took place on devices associated with Oklahoma State University.

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837-1847.
2. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743-753.
3. Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*. 2003;24(11):987-1003.
4. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can Med Assoc J*. 2010;182(6):551-557.
5. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern Med*. 2013;173(8):632-638.
6. Saposnik G, Kapral MK, Liu Y, et al. IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation*. 2011;123(7):739-749.
7. Austin PC, van Walraven C. The Mortality Risk Score and the ADG Score: two points-based scoring systems for the Johns Hopkins Aggregated Diagnosis Groups (ADGs) to predict mortality in a general adult population cohort in Ontario, Canada. *Med Care*. 2011;49(10):940-947.
8. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? *J Clin Epidemiol*. 2002;55(10):1054-1055.
9. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med*. 2004;23(10):1631-1660.
10. Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. *Stat Med*. 2016;35(22):4056-4072.
11. Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet*. 2009;373(9665):739-745.
12. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. Autoscore: a machine learning-based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Med Inform*. 2020;8(10):e21798.
13. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159.
14. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13(2):281-305.
15. Yau JWY, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35(3):556-564.
16. Hatfield M, Nguyen TH, Chapman R, et al. Identifying the mechanism of missingness for unspecified diabetic retinopathy disease severity in the electronic health record: an IRIS® Registry analysis. *J Am Med Inform Assoc*. 2023;30(6):1199-1204.
17. Tarazona-Santabalbina FJ, Belenguer-Varea Á, Rovira-Daudi E, et al. Early interdisciplinary hospital intervention for elderly patients with hip fractures: functional outcome and mortality. *Clinics*. 2012;67(6):547-555.
18. Zhang J, Yang M, Ge Y, Ivers R, Webster R, Tian M. The role of digital health for post-surgery care of older patients with hip fracture: a scoping review. *Int J Med Inform*. 2022;160:104709.
19. Wang R, Miao Z, Liu T, et al. Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records. *J Clin Med*. 2021;10(7):1473.
20. Chang YC, Wu WC. Dyslipidemia and diabetic retinopathy. *Rev Diabet Stud*. 2013;10(2-3):121-132.
21. Ding J, Wong TY. Current epidemiology of diabetic retinopathy and diabetic macular edema. *Curr Diab Rep*. 2012;12(4):346-354.
22. Cheng YJ, Gregg EW, Geiss LS, et al. Association of A1C and fasting plasma glucose levels with diabetic retinopathy prevalence in the US population: implications for diabetes diagnostic thresholds. *Diabetes Care*. 2009;32(11):2027-2032.
23. Irace C, Scarinci F, Scorcia V, et al. Association among low whole blood viscosity, haematocrit, haemoglobin and diabetic retinopathy in subjects with type 2 diabetes. *Br J Ophthalmol*. 2011;95(1):94-98.
24. Davis MD, Fisher MR, Gangnon RE, et al. Risk factors for high-risk proliferative diabetic retinopathy and severe visual loss: Early Treatment Diabetic Retinopathy Study Report# 18. *Invest Ophthalmol Vis Sci*. 1998;39(2):233-252.
25. Ng K, Steinhubl SR, DeFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circ: Cardiovasc Qual Outcomes*. 2016;9(6):649-658.

26. Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inform Assoc*. 2019;26(3):242-253.
27. Checketts JX, Dai Q, Zhu L, Miao Z, Shepherd S, Norris BL. Readmission rates after hip fracture: are there prefracture warning signs for patients most at risk of readmission? *J Am Acad Orthop Surg*. 2020;28(24):1017-1026.
28. Kiefer J. Sequential minimax search for a maximum. *Proc Am Math Soc*. 1953;4(3):502-506.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
30. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
31. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395-405.
32. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20(1):117-121.
33. Dixon BE, Jabour AM, Phillips EO, Marrero DG. An informatics approach to medication adherence assessment and improvement using clinical, billing, and patient-entered data. *J Am Med Inform Assoc*. 2014;21(3):517-521.
34. Wang M, Sushil M, Miao BY, Butte AJ. Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data. *J Am Med Inform Assoc*. 2023;30(7):1323-1332.
35. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;25(10):1419-1428.
36. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.
37. Dong X, Deng J, Rashidian S, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. *J Am Med Inform Assoc*. 2021;28(8):1683-1693.
38. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev: Data Min Knowl Discov*. 2019;9(4):e1312.
39. O'Brien C, Goldstein BA, Shen Y, et al. Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. *MDM Policy Pract*. 2020;5(1):2381468319899663.
40. Schmälzle R, Renner B, Schupp HT. Health risk perception and risk communication. *Policy Insights Behav Brain Sci*. 2017;4(2):163-169.
41. Ferrer RA, Klein WM. Risk perceptions and health behavior. *Curr Opin Psychol*. 2015;5:85-89.
42. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151.
43. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50 Suppl(0):S21-9.
44. Tsiampalis T, Panagiotakos D. Methodological issues of the electronic health records' use in the context of epidemiological investigations, in light of missing data: a review of the recent literature. *BMC Med Res Methodol*. 2023;23(1):180.
45. Sarwar T, Seifollahi S, Chan J, et al. The secondary use of electronic health records for data mining: data characteristics and challenges. *ACM Comput Surv*. 2022;55(2):1-40.
46. Miao Z, Sealey MD, Sathyanarayanan S, Delen D, Zhu L, Shepherd S. A data preparation framework for cleaning electronic health records and assessing cleaning outcomes for secondary analysis. *Inform Syst*. 2023;111:102130.
47. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Comput Biol Med*. 2019;112:103375.
48. Kamalapathy PN, Dunne PJ, Yarboro S. National evaluation of social determinants of health in orthopedic fracture care: decreased social determinants of health is associated with increased adverse complications after surgery. *J Orthop Trauma*. 2022;36(7):e278-82-e282.
49. Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: a scientific review. *Diabetes Care*. 2021;44(1):258-279.
50. White-Williams C, Rossi LP, Bittner VA, et al. Addressing social determinants of health in the care of patients with heart failure: a scientific statement from the American Heart Association. *Circulation*. 2020;141(22):e841-e863.
51. Northwood M, Ploeg J, Markle-Reid M, Sherifali D. Integrative review of the social determinants of health in older adults with multimorbidity. *J Adv Nurs*. 2018;74(1):45-60.
52. Marmot M, Wilkinson R. *Social Determinants of Health*. Oup Oxford; 2005.
53. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *J Am Med Inform Assoc*. 2020;27(11):1764-1773.
54. Weir RC, Proser M, Jester M, Li V, Hood-Ronick CM, Gurewich D. Collecting social determinants of health data in the clinical setting: findings from national PRAPARE implementation. *J Health Care Poor Underserved*. 2020;31(2):1018-1035.

Journal of the American Medical Informatics Association, 2024, 31, 1763–1773
https://doi.org/10.1093/jamia/ocae140
Research and Applications