A Novel Hyperparameter Search Approach for Accuracy and Simplicity in Disease Prediction Risk Scoring

Yajun Lu, PhD[a], Thanh Duong, BS[b,c], Zhuqi Miao, PhD[d], Thanh Thieu, PhD[c,e], Jivan Lamichhane, MD[f], Abdulaziz Ahmed, PhD[g], Dursun Delen, PhD[h,i,*]

[a]*Department of Management and Marketing, Jacksonville State University, Jacksonville, AL, USA*
[b]*Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA*
[c]*Department of Machine Learning, Moffitt Cancer Center and Research Institute, Tampa, FL, USA*
[d]*School of Business, The State University of New York at New Paltz, New Paltz, NY, USA*
[e]*Department of Oncological Sciences, University of South Florida Morsani College of Medicine, Tampa, FL, USA*
[f]*The State University of New York Upstate Medical University, Syracuse, NY, USA*
[g]*Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, USA*
[h]*Center for Health Systems Innovation, Department of Management Science and Information Systems, Oklahoma State University, Stillwater, OK, USA*
[i]*Department of Industrial Engineering, Faculty of Engineering and Natural Sciences, Istinye University, 34396, Sariyer/Istanbul, Turkey*

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 97,876 diabetic patients and 18,065 hip fracture patients.

Results: Our scores achieve accuracies insignificantly different from those obtained by existing approaches, reaching up to AUCs of 0.814 and 0.638 for DR and HFR predictions, respectively. Regarding the scale, our most accurate scores ranged 0–25 for DR and 0–5 for HFR, while scores

*Corresponding author.
Email addresses: ylu@jsu.edu (Yajun Lu, PhD), thanh.duong@moffitt.org (Thanh Duong, BS), miaoz@newpaltz.edu (Zhuqi Miao, PhD), thanh.thieu@moffitt.org (Thanh Thieu, PhD), lamichhJ@upstate.edu (Jivan Lamichhane, MD), aahmed2@uab.edu (Abdulaziz Ahmed, PhD), dursun.delen@okstate.edu (Dursun Delen, PhD)

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.

Keywords: Disease Prediction, Risk Scoring System, Hyperparameter Search, Electronic Health Record

BACKGROUND AND SIGNIFICANCE

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 number of regression units in the disease prediction model to be mapped to a single point in the risk scoring system*, henceforth denoted as B for simplicity. Specifically, smart 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 means using more risk score points to correspond to a given amount of regression-modeled risk, resulting in a larger and more complex scale for the score system, which introduces 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. While 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.[13]

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).

MATERIAL AND METHODS

Data Source and Preprocessing

In this study, we utilized the 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' time-stamped encounters, demographics, diagnoses, procedures, medications, laboratory results, vital signs, etc. Cerner Corporation 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.[14, 15] Hip fractures (HF) significantly increase morbidity and mortality in older adults, frequently resulting in post-discharge readmissions.[16, 17] Both are significant conditions drawing extensive research attention and warranting further investigation.

- DR Data: Our DR case study replicated the logistic regression-based DR prediction model reported in a previous study by Wang et al.,[18] followed by our proposed technique to derive a risk scoring system from the model. Given the nature of replication, we employed the identical DR data, predictors, and preprocessing and modeling methods as utilized by Wang et al. The data was extracted from Health Facts®, containing 97,876 diabetes patients, with 3,749 of them having DR. The predictors include *age, creatinine, HbA1c, neuropathy, duration of diabetes, white blood cells, nephropathy, glucose, hematocrit*, and *sodium*. Wang et al. identified these 10 essential predictors out of 26 variables related to patient's demographics, duration of diabetes, complications, and laboratory results through a machine-learning-based ensemble predictor selection method.[19] Their results showed that the compact model with the 10 essential variables achieved very close accuracy to that of the full model with all 26 variables. Wang et al. applied

the complete case method to preprocess the missing values for all the predictors. Given that we adopted the identical data, no predictor had missing values in the case study. In the predictive modeling, the values of these predictors, 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. Interested readers can find detailed information about the approach in Wang et al.'s paper.[18]

- 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.[20] A flow chart detailing the process is provided in Sheet A of the Supplementary Material. In total, we included 18,065 HF patients, among whom 2,055 were readmitted to the hospital within 30 days from their HF inpatient visits. By applying the same predictor selection method utilized by Wang et al.,[18] we identified nine essential predictors from hundreds of variables related to demographics, historical visits, discharge, laboratory results, diagnoses and procedures associated with their HF inpatient visits. Specifically, the identified predictors are *blood urea nitrogen, length of stay, serum creatinine, preinp1y* (defined as the number of inpatient visits within one year before), *Charlson Comorbidity Index, platelet count, age, hematocrit*, and *hemoglobin*. To maintain consistency in preprocessing missing values across both case studies, we also applied the complete case method to these predictors. Beyond preprocessing, the predictors were utilized to predict all-cause readmissions within 30 days from the HF inpatient visits.

For both study cohorts, we randomly partition the data into training (70%) and testing subsets (30%) for predictive analysis, with detailed cohort statistics summarized in Table 1.

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

Table 1: Statistics on training and test datasets for DR and HFR predictions.

	DR Dataset			
	Training		Test	
	non-DR	DR	non-DR	DR
# Patient	65,898	2,615	28,229	1,134
Age (mean (SD))	64.07 (14.07)	60.61 (13.17)	64.23 (14.09)	60.38 (13.10)
Creatinine (mean (SD))	1.07 (0.45)	1.94 (1.85)	1.07 (0.46)	1.93 (1.71)
Duration of Diabetes (mean (SD))	1.91 (1.76)	2.78 (2.05)	1.92 (1.77)	2.75 (2.04)
Glucose (mean (SD))	142.81 (46.21)	174.43 (61.91)	143.07 (46.32)	175.77 (62.06)
Hba1c (mean (SD))	7.14 (1.51)	8.35 (2.01)	7.14 (1.52)	8.40 (2.08)
Hematocrit (mean (SD))	38.99 (4.75)	36.23 (4.76)	38.90 (4.73)	36.27 (4.63)
Nephropathy = yes (%)	3173 (4.8)	731 (28.0)	1512 (5.4)	307 (27.1)
Neuropathy = yes (%)	5550 (8.4)	887 (33.9)	2427 (8.6)	355 (31.3)
Sodium (mean (SD))	138.83 (2.47)	138.50 (2.41)	138.82 (2.46)	138.46 (2.34)
White Blood Cell (mean (SD))	8.13 (2.21)	7.94 (2.18)	8.14 (2.21)	8.06 (2.39)
	HFR Dataset			
	Training		Test	
	non-HFR	HFR	non-HFR	HFR
# Patient	11,226	1,419	4,784	636
Age (mean (SD))	80.01 (9.80)	80.87 (9.22)	80.05 (9.74)	81.26 (9.04)
Length of Stay (mean (SD))	5.32 (2.79)	6.18 (3.53)	5.34 (2.78)	6.27 (3.36)
Platelet Count (mean (SD))	209.64 (81.53)	216.67 (89.54)	210.16 (82.10)	219.50 (91.35)
Blood Urea Nitrogen (mean (SD))	19.53 (10.97)	24.01 (14.37)	19.55 (11.25)	23.87 (13.29)
Hemoglobin (mean (SD))	10.11 (1.34)	10.07 (1.31)	10.14 (1.35)	10.10 (1.29)
Serum Creatinine (mean (SD))	0.97 (0.61)	1.15 (0.80)	0.97 (0.60)	1.11 (0.74)
Hematocrit (mean (SD))	30.06 (3.84)	30.09 (3.86)	30.17 (3.86)	30.12 (3.85)
Charlson Comorbidity Index (mean (SD))	1.29 (1.48)	1.72 (1.61)	1.30 (1.48)	1.67 (1.61)
preInp1Y (mean (SD))	0.35 (0.65)	0.52 (0.78)	0.35 (0.65)	0.57 (0.78)

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.

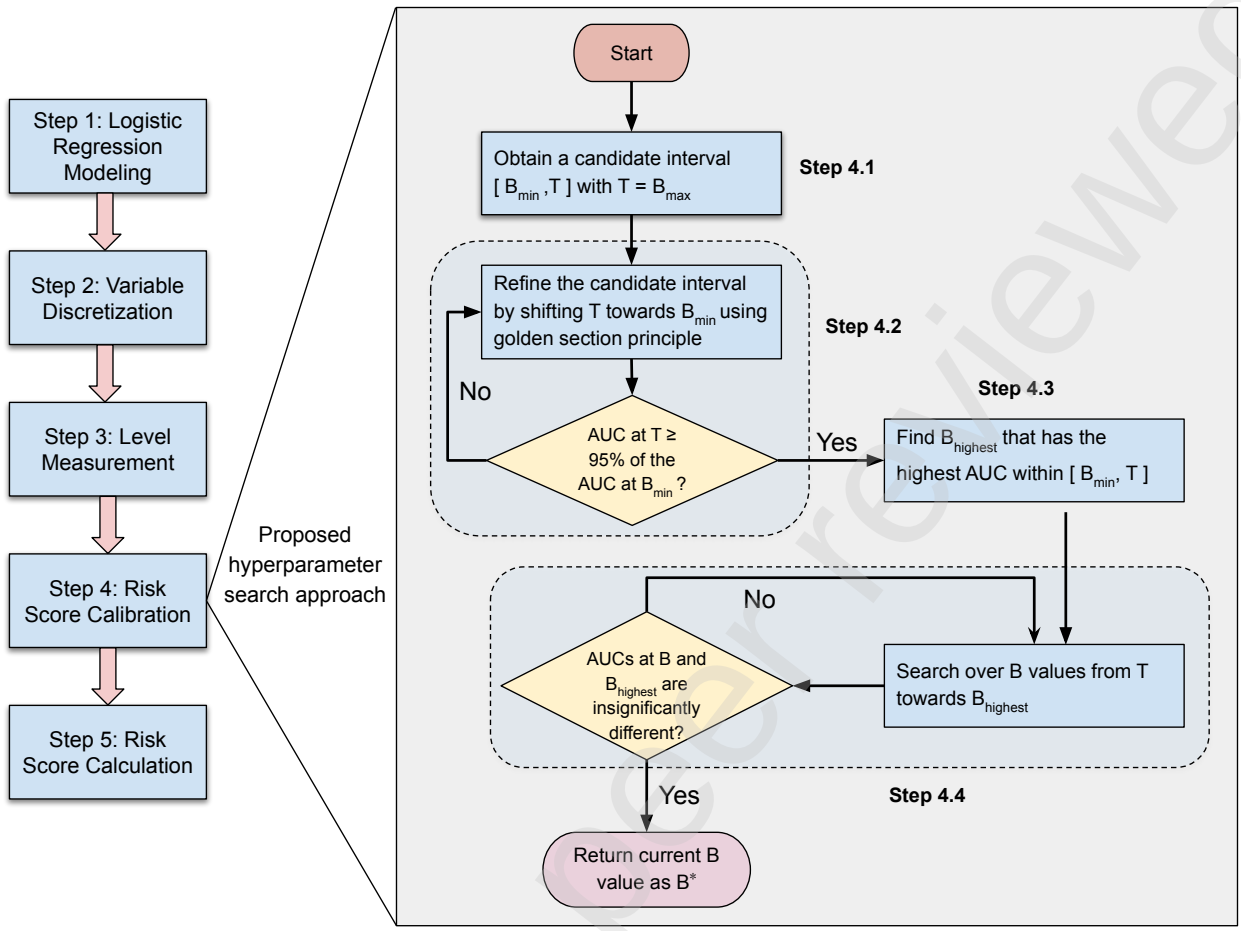


Figure 1: Flowchart illustrating the risk score derivation framework and our refinements in Step 4. In this illustration, B denotes the number of regression units in the disease prediction model to be mapped to a single point in the risk scoring system. T is the right end of the candidate interval of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

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. In our implementation, we explore four types of variable discretization cutoffs, which are respectively based on domain expertise, even sample, even length, and specific percentiles [12] (in particular, the percentiles are 5%, 20%, 80%, and 95%).

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 (e.g., 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, potentially leading to suboptimal B values. 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 Background and Significance 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] while some other studies employed the smallest coefficient.[5, 12] 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.[13] 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 95% of the accuracy at B_{min}. In our implementation, we measure accuracy using AUC.[21] The endpoint adjustment adheres to the golden section principle.[22] In other words, for each iteration, the new value of T is updated as $T' - 0.382 \times (T' - B_{\min})$, where T' 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 AUC, 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 AUC is insignificantly different from that of $B_{highest}$ according to DeLong test.[23] At last, return the found B value, denoted as B^*.

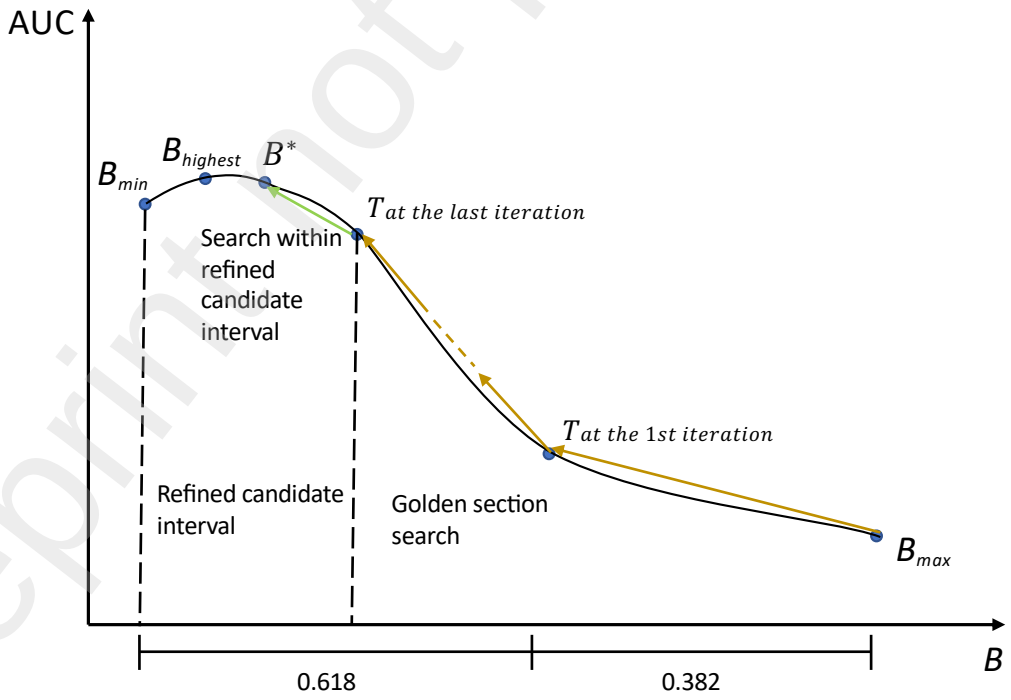


Figure 2: An illustration of the search trajectory of the developed hyperparameter search algorithm

The benefit of this search strategy is that we can leverage the intuition that with the increase of B, the AUC 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 developed algorithm 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.

RESULTS

Trend between AUC and B

Figure 3 (A) and (C) depict the relationships between AUC and B values for DR and HFR predictions, respectively when domain-expertise-based cutoffs were used for variable discretization (plots based on other variable discretization cutoffs are provided in Sheets B and C of the Supplementary Material). All 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 3 (B) and (D) provide zoomed-in views of the refined candidate intervals, showing that $B_{highest}$ does not necessarily coincide with the smallest B. Furthermore, many AUCs 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 for the hyperparam-

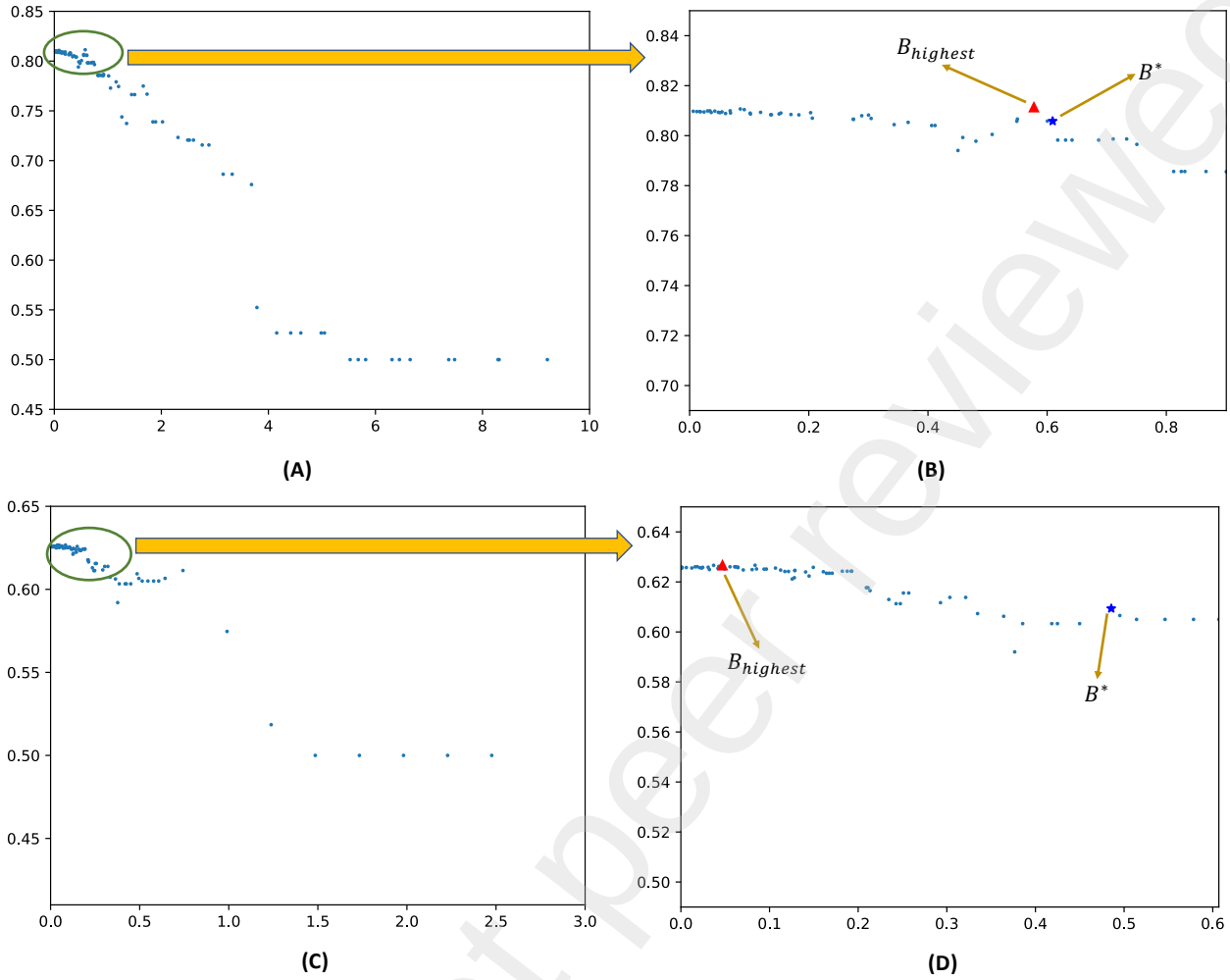


Figure 3: The AUC and B relationships for DR (A and B) and HFR (C and D) when domain-expertise-based cutoffs were used for variable discretization. (A) and (C) provide overall trends over all B values considered. (B) and (D) provide zoomed-in views of the AUC-B relationship for DR and HFR, respectively.

eter B in literature. For each B value, the four predictor discretization cutoffs, based on domain expertise, even sample, even length, and specific percentiles, are evaluated for comparison. Table 2 summarizes the AUCs and scales of the risk scores across different combinations of B values and predictor discretization methods. Upon observation, our algorithm consistently generates risk scores closely aligned with those obtained using other B values in terms of AUC for each discretization method. It is worth noting that, though the AUCs associated with B^* are slightly lower than those associated with other B values, the difference is statistically insignificant according to DeLong test. Remarkably, the scales of the risk scores derived through our approach exhibit much simpler ranges, with the highest scores spanning from 24 to 42, in contrast to 85 to 269 for $5\beta_{\text{age}}$ and 1,210 to 3,789 for B_{min} in the context of DR prediction. Similarly, for HFR prediction, the

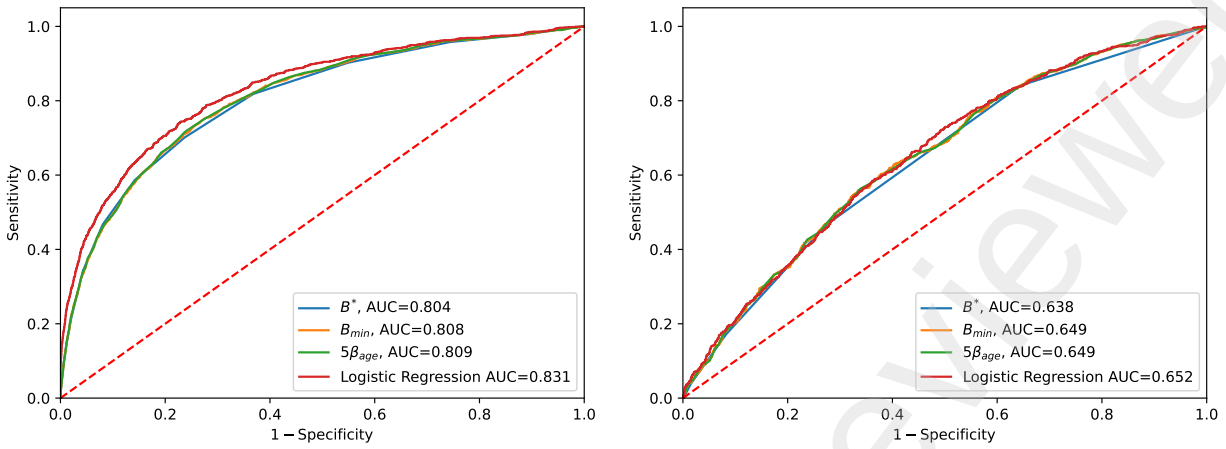
highest scores in our scales spans from only 5 to 10, whereas $5\beta_{\text{age}}$ and $B_{\min}$ show ranges with the highest score as high as 61 to 128 and 13,835 to 28,214, respectively.

Table 2: AUCs and scales of the risk scores across B values and predictor discretization methods for DR and HFR.

Case		B	Expert	Even Sample	Even Length	Specific Percentiles
DR	AUC	B^*	0.804	0.814	0.790	0.807
		$B_{\min}$	0.808	0.816	0.803	0.813
		$5\beta_{age}$	0.809	0.815	0.803	0.813
	Scale	B^*	0–24	0–25	0–42	0–26
		$B_{\min}$	0–2,510	0–1,210	0–3,789	0–1,835
		$5\beta_{age}$	0–178	0–85	0–269	0–131
HFR	AUC	B^*	0.638	0.612	0.633	0.610
		$B_{\min}$	0.649	0.625	0.653	0.618
		$5\beta_{age}$	0.649	0.625	0.653	0.618
	Scale	B^*	0–5	0–6	0–10	0–10
		$B_{\min}$	0–19,372	0–13,835	0–28,214	0–21,984
		$5\beta_{age}$	0–87	0–61	0–128	0–100

We additionally compared the risk scores with the corresponding logistic regression models—the root model from which the scores are derived—in terms of AUC. The AUC plots when domain-expertise-based cutoffs were used are displayed in Figure 4, with the plots for other types of cutoffs provided in Sheets D and E of the Supplementary Material. All the plots illustrate a very marginal difference, up to 0.041 (as observed in the even length plots for DR, available in Sheet D of the Supplementary Material), between the predictive accuracy achieved by the risk scores and that of logistic regressions. This aligns with what has been reported in the literature,[18] reiterating the effectiveness of the entire risk scoring framework in maintaining strong predictive capacity from logistic regressions.

Furthermore, we report the new risk score systems for DR and HFR (with domain-expertise-based cutoffs) in Table 3 (scoring systems utilizing other types of cutoffs can be found in Sheets F and G of the Supplementary Material). Note that the risk score derived using $5\beta_{age}$ for DR, presented in the table is essentially equivalent to the score system proposed by Wang et al.[18] 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 <0.5 and 0.5–1 for *creatinine* and <4, 4–6, and 6–8 for *white blood cell*. These levels can be combined because they share the



(A) (B)

Figure 4: Comparison between AUCs of risk scores and AUC of Logistic Regression: for (A) DR prediction and (B) HFR prediction.

identical risk points. A similar finding can be observed for the HFR risk score as well. Many levels can be combined, for instance <5 and 5–7 for *length of stay* and 4–6 and >6 for *Charlson comorbidity index*. More interestingly, our new scoring system requires a even more concise set of predictors, specifically *blood urea nitrogen, length of stay, preInp1Y*, and *Charlson comorbidity index*, rather than the original 9 variables, because the other predictors exhibit 0 point 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.[24, 25, 26, 27] Machine learning and deep learning models are often employed to tackle the challenges because of their high predictive accuracy.[28, 29, 30] However, the inherent black-box nature of machine/deep learning often poses challenges in interpreting the results for clinicians.[31] Additionally, many existing EHR systems in hospitals lack support for complex machine-learning models.[32]

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

Variable	Category	Risk Score for DR		
		B^*	$5\beta_{age}$	$B_{\min}$
Neuropathy	No	0	0	0
	Yes	2	11	153
Nephropathy	No	0	0	0
	Yes	1	7	105
Creatinine	<0.5	0	0	0
	0.5–1	0	3	47
	1–1.5	1	8	116
	1.5–2	2	13	185
	>2	3	22	313
Hba1c	<6	0	0	0
	6–8	1	7	96
	8–10	2	14	192
	10–12	3	20	287
	>12	4	31	431
Duration of Diabetes	<1	0	0	0
	1–2	0	2	25
	2–3	0	4	50
	3–4	1	5	75
	>4	2	15	217
White Blood Cell	<4	2	18	247
	4–6	2	16	222
	6–8	2	13	189
	8–12	1	10	138
	>12	0	0	0
Glucose	<60	0	0	0
	60–80	0	1	17
	80–100	0	3	37
	100–200	1	7	97
	>200	3	22	311
Age	<35	2	12	173
	35–50	1	9	127
	51–65	1	6	85
	66–75	0	3	49
	76–85	0	2	21
	>85	0	0	0
Hematocrit	<30	3	24	333
	30–35	3	18	256
	35–40	2	14	199
	40–50	1	8	114
	>50	0	0	0
Sodium	<136	0	0	0
	136–144	1	9	129
	>144	2	16	228

Variable	Category	Risk Score for HFR		
		B^*	$5\beta_{age}$	$B_{\min}$
Blood Urea Nitrogen	<11	0	0	0
	11–15	0	3	765
	15–19	0	6	1,377
	19–27	1	10	2,295
	>27	1	18	3,978
Length of Stay	<5	0	0	0
	5–7	0	4	871
	7–14	1	13	2,829
	>14	2	24	5,223
Serum Creatinine	<0.6	0	1	110
	0.6–0.8	0	0	90
	0.8–0.9	0	0	72
	0.9–1.2	0	0	47
	>1.2	0	0	0
preInp1Y[a]	<1	0	0	0
	1–2	1	12	2,663
	>2	1	16	3,551
Charlson Comorbidity Index	<4	0	0	0
	4–6	1	8	1,843
	>6	1	10	2,304
Platelet Count	<143	0	0	0
	143–177	0	0	39
	177–213	0	0	74
	213–268	0	1	120
	>268	0	1	200
Age	<65	0	0	0
	65–75	0	2	485
	76–80	0	4	817
	81–85	0	5	1,037
	86–90	0	6	1,258
	>90	0	6	1,368
Hematocrit	<26.9	0	0	0
	26.9–28.8	0	2	395
	28.8–30.7	0	3	715
	30.7–33.1	0	5	1,076
	>33.1	0	7	1,649
Hemoglobin	<9	0	4	990
	9–9.7	0	3	735
	9.7–10.3	0	2	540
	10.3–11.1	0	1	330
	>11.1	0	0	0

[a] Number of inpatient visits within one year before.

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 be applied to support healthcare professionals in various medical decision making processes. Firstly, it empowers healthcare professionals to compute patients' risk scores, which not

only enables the evaluation of patients' condition severity but also serves as a practical early-warning tool for physicians to identify high-risk patients, thereby optimizing the allocation of medical resources to those in immediate need. Additionally, the risk score's interpretability, along with insights into how each feature contributes to the total risk score, empowers patients to better grasp the factors that pose health risks. As a result, patients are more likely to modify unhealthy behaviors to reduce their risk score. Hence, our approach has the potential to facilitate positive behavior change among patients by empowering them to recognize and address the factors that may negatively impact their health.

Compared to the state-of-the-art DR risk score,[18] 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.

An additional finding from our exploration of predictor discretization's impact on risk scoring is that statistical cutoffs, such as even sample, even length, or specific percentiles can support generating equivalent (see the *Even Length* vs. *Expert* columns for HFR in Table 2) or, sometime, even more accurate (compare the *Even Sample* vs. *Expert* columns for DR in Table 2) risk scores compared to those derived using domain-expertise-based cutoffs. Given that statistical cutoffs are easily automated, this finding provides a favorable evidence supporting the endeavor to automate the entire risk scoring process, as exemplified in existing literature.[12]

Limitations: Our proposed technique enables the risk score system to closely mirror the predictive accuracy of regression models. Numerous factors throughout the stages of preprocessing, modeling, and deployment have the potential to impact the actual accuracy of regression models in real-world prediction. Examples include handling missing values, addressing data imbalance, and the geographical and care setting differences between modeling and deployment. Any of these factors may create a chain effect on the resulting risk score through their influence on the root re-

gression model. Given that the objective of this study is to develop and assess a hyperparameter searching approach to generate risk scores of a compact scale that retain as much accuracy as possible from the root regression models, the indirect influence of these factors on risk scores is beyond the scope of this work, thus deferred for future research.

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.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Cerner Corporation 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 author(s) and do not necessarily reflect the views of the Cerner Corporation.

FUNDING STATEMENT

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

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

CONTRIBUTORSHIP STATEMENT

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

DATA AVAILABILITY STATEMENT

The Cerner EHR data used in this study can be requested through the Oklahoma State University Center for Health Systems Innovation (CHSI) at `https://business.okstate.edu/chsi`.

CODE AVAILABILITY STATEMENT

The code supporting the findings of this study is currently accessible from the corresponding author upon request. Upon manuscript acceptance, it will be made publicly available on GitHub.

ETHICS INFORMATION

The Institutional Review Boards (IRB) at Oklahoma State University exempted the study from review because the data has 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):1837-47.

[2] D'Agostino Sr RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.

[3] Conroy RM, Pyörälä K, Fitzgerald Ae, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European Heart Journal. 2003;24(11):987-1003.

[4] Van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Canadian Medical Association Journal. 2010;182(6):551-7.

[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 Internal Medicine. 2013;173(8):632-8.

[6] Saposnik G, Kapral MK, Liu Y, Hall R, O'Donnell M, Raptis S, et al. IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. Circulation. 2011;123(7):739-49.

[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. Medical Care. 2011;49(10):940.

[8] Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? Journal of Clinical Epidemiology. 2002;55(10):1054-5.

[9] Sullivan LM, Massaro JM, D'Agostino Sr RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Statistics in Medicine. 2004;23(10):1631-60.

[10] Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. Statistics in Medicine. 2016;35(22):4056-72.

[11] Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. The Lancet. 2009;373(9665):739-45.

[12] Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N, et al. Autoscore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Medical Informatics. 2020;8(10):e21798.

[13] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012;13(2).

[14] Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35(3):556-64.

[15] Hatfield M, Nguyen TH, Chapman R, Myrick AC, Leng T, Mbagwu M, et al. Identifying the mechanism of missingness for unspecified diabetic retinopathy disease severity in the electronic health record: an IRIS® Registry analysis. Journal of the American Medical Informatics Association. 2023;30(6):1199-204.

[16] Tarazona-Santabalbina FJ, Belenguer-Varea A, Rovira-Daudi E, Salcedo-Mahiques E, Cuesta-Peredo D, Domenech-Pascual JR, et al. Early interdisciplinary hospital intervention for elderly patients with hip fractures: functional outcome and mortality. Clinics. 2012;67:547-56.

[17] 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. International Journal of Medical Informatics. 2022;160:104709.

[18] Wang R, Miao Z, Liu T, Liu M, Grdinovac K, Song X, et al. Derivation and Validation of Essential Predictors and Risk Index for Early Detection of Diabetic Retinopathy Using Electronic Health Records. Journal of Clinical Medicine. 2021;10(7):1473.

[19] Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. Journal of the American Medical Informatics Association. 2019;26(3):242-53.

[20] 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? Journal of the American Academy of Orthopaedic Surgeons. 2020;28(24):1017-26.

[21] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145-59.

[22] Kiefer J. Sequential minimax search for a maximum. Proceedings of the American Mathematical Society. 1953;4(3):502-6.

[23] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837-45.

[24] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012;13(6):395-405.

[25] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. 2013;20(1):117-21.

[26] Dixon BE, Jabour AM, Phillips EO, Marrero DG. An informatics approach to medication adherence assessment and improvement using clinical, billing, and patient-entered data. Journal of the American Medical Informatics Association. 2014;21(3):517-21.

[27] 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. Journal of the American Medical Informatics Association. 2023;30(7):1323-32.

[28] Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2018;25(10):1419-28.

[29] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-58.

[30] Dong X, Deng J, Rashidian S, Abell-Hart K, Hou W, Rosenthal RN, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. Journal of the American Medical Informatics Association. 2021;28(8):1683-93.

[31] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019;9(4):e1312.

[32] O'Brien C, Goldstein BA, Shen Y, Phelan M, Lambert C, Bedoya AD, et al. Development, implementation, and evaluation of an in-hospital optimized early warning score for patient deterioration. MDM Policy & Practice. 2020;5(1):2381468319899663.