Received: 6 August 2020 | Revised: 11 January 2021 | Accepted: 15 January 2021

DOI: 10.1002/pds.5199

ORIGINAL ARTICLE

WILEY

Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data

Lin Li[1] | Chuang-Chung Lee[2] | Fang Liz Zhou[1] | Cliona Molony[2] | Zoran Doder[1] | Evgeny Zalmover[1] | Kristen Sharma[1] | Juhaeri Juhaeri[1] | Chuntao Wu[1]

[1]Sanofi U.S. LLC, Bridgewater, New Jersey

[2]Sanofi U.S. LLC, Cambridge, Massachusetts

Correspondence

Lin Li, Sanofi U.S. LLC, 55 Corporate Drive, Bridgewater, NJ 08807.
Email: lin.li@sanofi.com

Present address

Kristen Sharma, Astellas Pharmaceuticals, Inc, Northbrook, Illinois

Chuntao Wu, Alexion Pharmaceuticals, Inc, Boston, Massachusetts

Funding information

Sanofi

Abstract

Purpose: To assess the performance of different machine learning (ML) approaches in identifying risk factors for diabetic ketoacidosis (DKA) and predicting DKA.

Methods: This study applied flexible ML (XGBoost, distributed random forest [DRF] and feedforward network) and conventional ML approaches (logistic regression and least absolute shrinkage and selection operator [LASSO]) to 3400 DKA cases and 11 780 controls nested in adults with type 1 diabetes identified from Optum® de-identified Electronic Health Record dataset (2007–2018). Area under the curve (AUC), accuracy, sensitivity and specificity were computed using fivefold cross validation, and their 95% confidence intervals (CI) were established using 1000 bootstrap samples. The importance of predictors was compared across these models.

Results: In the training set, XGBoost and feedforward network yielded higher AUC values (0.89 and 0.86, respectively) than logistic regression (0.83), LASSO (0.83) and DRF (0.81). However, the AUC values were similar (0.82) among these approaches in the test set (95% CI range, 0.80–0.84). While the accuracy values >0.8 and the specificity values >0.9 for all models, the sensitivity values were only 0.4. The differences in these metrics across these models were minimal in the test set. All approaches selected some known risk factors for DKA as the top 10 features. XGBoost and DRF included more laboratory measurements or vital signs compared with conventional ML approaches, while feedforward network included more social demographics.

Conclusions: In our empirical study, all ML approaches demonstrated similar performance, and identified overlapping, but different, top 10 predictors. The difference in selected top predictors needs further research.

Lin Li and Chuang-Chung Lee have equal authorship

This work was presented as a poster at the International Conference on Pharmacoepidemiology & Therapeutic Risk Management All Access in 2020.

KEYWORDS
AUC, diabetic ketoacidosis, least absolute shrinkage and selection operator, logistic regression, machine learning, prediction model

1 | INTRODUCTION

Artificial intelligence including machine learning (ML) has been increasingly used to analyze healthcare data including electronic health records (EHR).[1] ML is a natural extension to conventional analysis approaches such as logistic regression, and has been widely used to learn complex relationships or patterns from data to make accurate predictions.[1,2] Conventional analysis approaches are commonly used to identify risk factors for a disease or outcome in clinical epidemiology. However, there is limited data on performance of different ML approaches in such studies, and the theoretical superiority of more flexible ML such as XGBoost and distributed random forest (DRF) over conventional ML approaches has not been consistently observed in real-world settings.[3]

Diabetic ketoacidosis (DKA) is an acute life-threatening but preventable complication of type 1 diabetes (T1D).[4] In the United States, DKA hospitalization had increased by 54.9% from 2009 to 2014 after a decline in 2000–2009.[5] Because DKA is caused by insulin deficiency that is often precipitated by discontinuation of insulin or inadequate insulin treatment, it is important to identify risk factors for DKA and to predict DKA for effective T1D management. However, no prediction model for DKA is developed and used in clinical practice. Given the growing use of ML in clinical prediction including pharmacoepidemiology[6] with limited model performance assessment, we conducted a study to assess the empirical performances of different ML approaches in identifying risk factors for DKA and predicting DKA in adults with T1D using an EHR database.

KEY POINTS

- Flexible machine learning (ML) approaches do not automatically result in improved performance over conventional ML approaches
- The performances of flexible ML and conventional ML approaches were similar in predicting diabetic ketoacidosis (DKA) using an electronic health records data source
- Flexible ML and conventional ML approaches identified overlapping, but different, top 10 risk factors for DKA
- Flexible ML approaches only provided the relative importance for each predictor, while logistic regression could also estimate odds ratio for each predictor
- Interpretation of the findings of flexible ML approaches is challenging

2 | METHODS

Different ML approaches were applied in a case–control study nested in adults with T1D to identify risk factors and predict DKA. The nested case–control design can more readily and efficiently identify risk factors for DKA compared with a cohort design. Previous studies have demonstrated that this design can be used for development of clinical prediction models.[7,8]

2.1 | Data source

Optum de-identified EHR database was used in this study. It currently encompasses approximately 80 million patients from all Census regions in the United States, with at least 5 million patients from each region. Data on the full spectrum of inpatient and outpatient treatments are collected from more than 140 000 physicians at more than 600 hospitals and 6500 clinics. On average, patients contribute 4 years of medical history to the database. Information such as patient-reported symptoms and outcomes as well as treatment rationale is captured directly from providers' notes via natural language processing. Approximately 82% of patients in this database are part of an Integrated Delivery Network, which includes hospital and emergency care as well as outpatient visits. In addition, about 20% of patients can be linked with administrative claims data.

2.2 | Study patients

Patients with T1D were identified from Optum EHR database between January 1, 2007 and September 30, 2018 by adapting the Klompas algorithm.[9] The positive predictive value (PPV) of T1D using the Klompas algorithm was 89% in the original publication and 94.5% in an external validation study.[9,10] For each patient with T1D, the start of follow up for DKA event was the later of the first date on which a patient met the diabetes surveillance algorithm criteria or when a patient turned 18 years. The end of follow up was the date of first hospitalized DKA event during the follow-up, the date of last-recorded clinical activity in the database, or September 30, 2018, whichever occurred earliest.

Hospitalized DKA event was identified using ICD-9-CM (250.1) or ICD-10-CM (E1X.1) diagnosis codes which have been reported to have a PPV of 88.9%.[11] Because DKA is an emergency condition and patients with DKA rarely meet criteria to be safely discharged from emergency departments,[12] outpatient or emergency encounters without subsequent hospitalization were excluded to minimize false positive cases. The index date of a case was the date of the first DKA event occurred during the follow up. The DKA cases included in the study must had ≥1 non-emergency clinical activity encounter and insulin treatment within 365 days before the index date, had T1D for ≥365 days before the index date, and had ≥1 HbA1c measurement within 183 days before the index date. Cases who were pregnant and those who used antihyperglycemic agents indicated for type 2 diabetes only (except for metformin) within 365 days prior to or on the index date were excluded.

For each DKA case, up to 10 controls without DKA on the index date of a case were randomly selected from the T1D cohort using incidence-density sampling without replacement and matched on whether a patient's EHR data was linked with claims data. The index date for controls was assigned as the date of DKA diagnosis of their matching case. The same inclusion and exclusion criteria for cases were also applied to controls.

2.3 | Potential predictors

Seven groups of potential predictors (i.e., features in ML) and the T1D cohort entry year as well as year of the index date were explored. These groups included social demographics, lifestyle factors, health service use, treatment, chronic comorbidities, acute medical conditions as well as laboratory test results, vital signs and other common measurements (Table S1). The selection of predictors was guided by the background knowledge of DKA and data availability.

2.4 | Statistical methods

The overall structure of patient data that served as input to the analysis algorithm is shown in Figure 1. The whole study subjects were first randomly split into a training and a test data sets in a ratio of 4:1. The training set was used in the feature selection preprocessing and model building. The test set was used to assess each model performance.

2.4.1 | Feature selection preprocessing

We carried out the feature selection preprocessing from all potential predictors in the training set. First, any features with missing data ≥60% were removed. This threshold was used to keep as many as laboratory measurements which were most likely due to an absence of testing. The remaining features with missing data are listed in Table S2. The missing records of these features were each imputed with the average values of the available records of that feature or grouped as “unknown” where applicable. Otherwise, many patients' records were incomplete for models to make prediction.[13] Second, to avoid collinearity among the features which can cause unstable estimates and the sign flipping of the coefficients,[14,15] variance inflation factor (VIF) for each feature was calculated (Table S3). Features with VIF values ≥10 were flagged as highly correlated features.[16] Only one feature with higher clinical utility or lower missing data percentage was selected within a group of highly correlated features or those caused sign flipping issue.

2.4.2 | Predictive modeling algorithms

Two conventional and three flexible ML approaches were utilized to identify important risk factors for DKA from the selected features and to build prediction models.

Conventional ML approaches included logistic regression and least absolute shrinkage and selection operator (LASSO). For logistic regression, backward selection with a cutoff *p*-value of 0.05 was used to select statistically significant predictors; adjusted odds ratio (OR) and its 95% confidence interval (CI) was estimated for each selected predictor. LASSO is a technique of parameter regularization, which was applied to a logistic regression. During regularization,

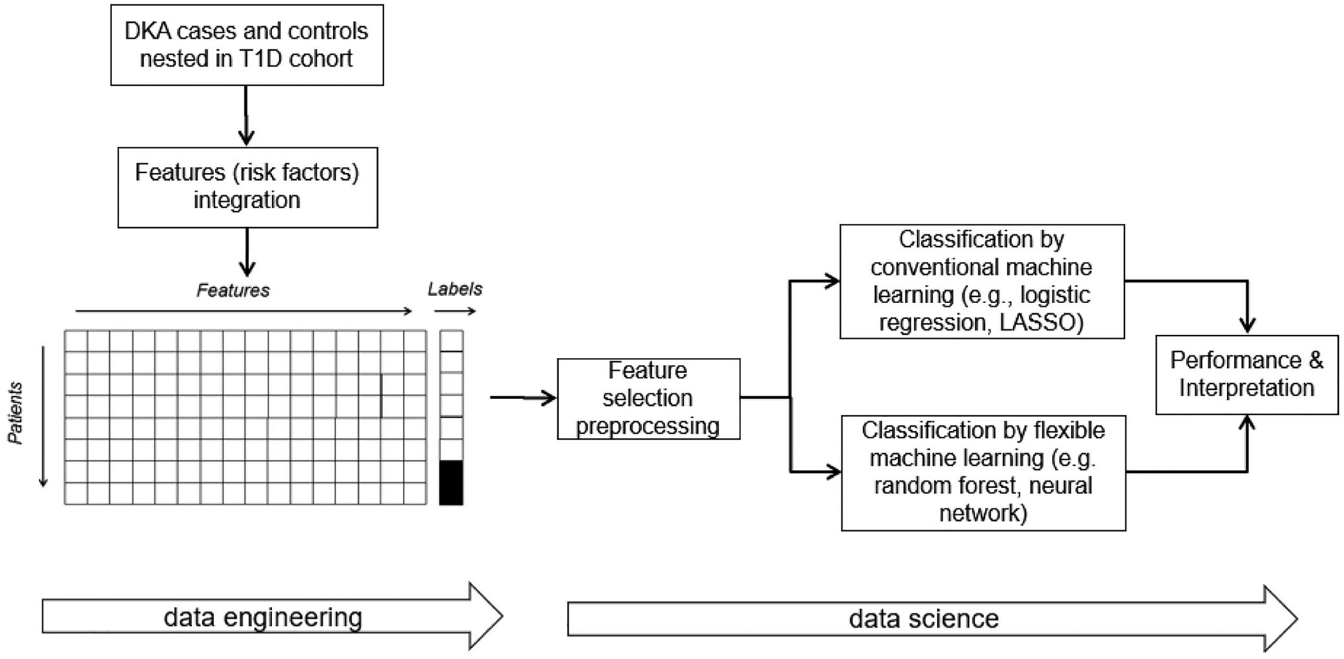
DKA cases and controls nested in T1D cohort
Features (risk factors) integration
Features
Labels
Patients
Feature selection preprocessing
Classification by conventional machine learning (e.g., logistic regression, LASSO)
Classification by flexible machine learning (e.g. random forest, neural network)
Performance & Interpretation
data engineering
data science

FIGURE 1 Data processing flowchart. DKA, diabetic ketoacidosis; LASSO, least absolute shrinkage and selection operator; T1D, type 1 diabetes

penalties are introduced to the model building process to avoid overfitting and reduce the number of covariates.[17]

Flexible ML approaches included XGBoost, DRF and feedforward network. XGBoost is a supervised learning algorithm that implements a process called boosting to yield accurate models.[18] Boosting refers to the ensemble learning technique of building many models sequentially, with each new model attempting to correct for the deficiencies in the previous one. DRF generates a forest of classification trees, rather than a single classification or regression tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees can reduce the variance. The classification process takes the average prediction over all trees to make a final prediction. Neural networks are learning algorithms inspired by the brain research. Feedforward network we used is the simplest form of neural network since its flow of information only moves in forward direction without circling back.[19] It consists of an input layer, an output layer, and several hidden layers in between. Each layer includes multiple nodes with different weights combining with input can determine the output of the network. During the learning process, these weights are updated to minimize the loss function.[20]

2.4.3 | Cross validation

During the training process, *k*-fold cross validation was conducted for each modeling. That is, the training set was further split into a training subset and a validation subset in a ratio of 4:1 to tune hyperparameters. This process was repeated five times (*k* = 5). The final model was then built by aggregating the five cross validation models and evaluated in the full training data set. Further details are provided in Appendix I.

2.4.4 | Model performance assessment

The model performance was assessed using different metrics below in the test set.

Area under the receiver operating characteristic (ROC) curve (AUC): It plots the true positive rate against the false positive rate,[21] and ranges from 0 to 1. The value of 1 indicates that the model predicts perfectly. Accuracy, specificity and sensitivity values were calculated based on a confusion matrix. To build a confusion matrix, a specific threshold value is required to determine whether a probability level gets assigned to a case or a control.[22] We chose accuracy as the metric to optimize to determine the threshold value. Finally, to demonstrate the variability of the model predictions, the 95% CI of AUC, accuracy, specificity, and sensitivity values were established using 1000 bootstrap samples.[23]

2.4.5 | Feature importance

For each approach, the feature importance percentage was determined by calculating the relative influence of each predictor. For conventional ML approaches, it was derived and ranked by the magnitude of standardized coefficient of each selected statistically significant predictor.[24] For XGBoost and DRF, two factors were considered to determine the relative importance of each feature: whether the variable was used to divide the decision tree node and how much prediction error has been reduced as a result of the split. That is, when split in a feature contributed to a larger decrease in the squared error, that feature was regarded as one with greater relative influence.[25] For feedforward network, the Gedeon method was used to calculate feature importance.[26] It considers the weights connecting the input features to the hidden layers.

TABLE 1 Study population attrition procession

	Patient counts (January 01, 2007–September 30, 2018)	
Individuals in Optum® de-identified Electronic Health Record database	95 823 300	
Individuals with diabetes	7 153 077	
Individuals with type 1 diabetes	169 779	
Individuals aged ≥18 years and within an Integrated Delivery Network	130 052	
Individuals with at least 1 HbA1c measurement and at least 1 year of clinical activity any time	105 816	
DKA case and control selection		
Potential candidates	Potential DKA Cases *N* = 15 454	Potential Controls[a] *N* = 105 816
After application of the study criteria on potential DKA cases before matching: • Type 1 diabetes for at least 365 days before index date • Treated with insulin within 365 days before index date • At least 1 HbA1c measurement within 183 days before index date • Without pregnancy within 365 days before index date • Without off-label use of antihyperglycemic agents indicated for type 2 diabetes only (except for metformin) within 365 days before or on index date	3400	NA
Control selection via incidence density sampling based on 1:10 matching ratio	3400	34 000
After application of the same study criteria defined above on controls	3400	11 780

Abbreviations: DKA, diabetic ketoacidosis; HbA1c, hemoglobin A1c.
[a] A control could not develop DKA before or at the matched index date but could become a case after the index date.

TABLE 2 Selected characteristics of DKA cases and controls

	DKA cases *N* = 3400	Controls *N* = 11 780	*p*-Value[a]
Calendar year of T1D cohort entry (%)			
2007	554 (16.3)	1831 (15.5)	0.111
2008	578 (17.0)	1910 (16.2)	
2009	367 (10.8)	1175 (10.0)	
2010	414 (12.2)	1469 (12.5)	
2011	376 (11.1)	1286 (10.9)	
2012	370 (10.9)	1257 (10.7)	
2013	283 (8.3)	1009 (8.6)	
2014	228 (6.7)	847 (7.2)	
2015	131 (3.9)	532 (4.5)	
2016	80 (2.4)	361 (3.1)	
2017	19 (0.6)	103 (0.9)	
Age, years, mean (SD)	42.9 (16.5)	45.4 (16.4)	<0.001
Sex (%)			
Female	1818 (53.5)	5594 (47.5)	<0.001
Male	1579 (46.4)	6184 (52.5)	
Unknown	3 (0.1)	2 (0.0)	
Race (%)			
Caucasian	2929 (86.1)	10 653 (90.4)	<0.001
African American	346 (10.2)	584 (5.0)	
Asian	12 (0.4)	71 (0.6)	
Other/Unknown	113 (3.3)	472 (4.0)	
Annual household income, $, mean (SD)	41 773 (8302)	42 935 (8938)	<0.001
Insurance type (%)			
Commercial	1306 (38.4)	6342 (53.8)	<0.001
Medicare	621 (18.3)	1519 (12.9)	
Medicaid	514 (15.1)	727 (6.2)	
Other payor type	134 (3.9)	353 (3.0)	
Uninsured	202 (5.9)	255 (2.2)	
Unknown	623 (18.3)	2584 (21.9)	
Geographic region (%)			
Midwest	2042 (60.1)	6792 (57.7)	<0.001
Northeast	315 (9.3)	1389 (11.8)	
South	646 (19.0)	2237 (19.0)	
West	269 (7.9)	1008 (8.6)	
Other/Unknown	128 (3.8)	354 (3.0)	
Lifestyle risk factors within 365 days before index date (%)			
Alcohol abuse	175 (5.1)	226 (1.9)	<0.001
Controlled substance abuse	523 (15.4)	454 (3.9)	<0.001
Health service use within 365 days before index date (%)			
Visit to endocrinologist	1789 (52.6)	7129 (60.5)	<0.001
Visit to primary care	1781 (52.4)	4959 (42.1)	<0.001
Chronic comorbidities any time between study start date and index date (%)			
Cardiovascular disease	2042 (60.1)	5927 (50.3)	<0.001
Diabetic microvascular complications	1981 (58.3)	5160 (43.8)	<0.001
Chronic liver disease	244 (7.2)	435 (3.7)	<0.001

TABLE 2 (Continued)

	DKA cases *N* = 3400	Controls *N* = 11 780	*p*-Value[a]
Chronic kidney disease	859 (25.3)	1328 (11.3)	<0.001
Dementia	137 (4.0)	217 (1.8)	<0.001
Psychiatric disorder	1743 (51.3)	3580 (30.4)	<0.001
Autoimmune disorders	432 (12.7)	1577 (13.4)	0.315
Cancer	377 (11.1)	1264 (10.7)	0.575
Acute medical conditions (%)			
Infection within 7 days before index date	230 (6.8)	96 (0.8)	<0.001
Major surgery within 7 days before index date	45 (1.3)	6 (0.1)	<0.001
Non-DKA hospitalization within 30 days before index date	610 (17.9)	168 (1.4)	<0.001
Treatments (%)			
Insulin pump within 7 days before index date	154 (4.5)	153 (1.3)	<0.001
Insulin type within 7 days before index date			
Intermediate/long-acting insulin	218 (6.4)	309 (2.6)	<0.001
Rapid/short-acting insulin	285 (8.4)	512 (4.3)	<0.001
Premixed insulin	10 (0.3)	15 (0.1)	0.061
Other medications within 30 days before index date			
Systemic steroids	81 (2.4)	101 (0.9)	<0.001
Diuretics	132 (3.9)	221 (1.9)	<0.001
Antipsychotics	60 (1.8)	57 (0.5)	<0.001
Laboratory test results or vital signs within 183 days before index date, mean (SD)[b]			
HbA1c, %	9.3 (1.8)	8.0 (1.4)	<0.001
Random blood glucose level, mg/dl	194.6 (60.6)	169.9 (66.3)	<0.001
eGFR, ml/min/1.73m^2	83.1 (36.9)	96.4 (30.5)	<0.001
Total cholesterol, mg/dl	178.5 (46.7)	173.3 (38.3)	<0.001
Systolic blood pressure, mm Hg	126.8 (15.6)	124.4 (13.8)	<0.001
BMI, kg/m^2	26.4 (5.8)	27.9 (5.7)	<0.001
Height, cm	169.6 (10.2)	171.1 (10.2)	<0.001
White blood cell count, x10^3 per microliter	9.0 (3.3)	7.6 (2.7)	<0.001
Platelet count, x10^3 per microliter	268.4 (79.7)	254.5 (72.4)	<0.001
Temperature, °C	36.7 (0.3)	36.7 (0.3)	<0.001
Pulse rate, beats per minute	85.0 (12.8)	78.7 (11.9)	<0.001
Respiratory rate, breaths per minute	17.4 (2.1)	16.7 (2.1)	<0.001
Hemoglobin, g/dl	12.6 (2.1)	13.4 (1.9)	<0.001
Oxygen saturation, S_pO_2 (pulse oximetry)	97.6 (1.5)	97.6 (1.5)	0.264

Abbreviations: BMI, body mass index; DKA, diabetic ketoacidosis; eGFR, estimated glomerular filtration rate; HbA1c, hemoglobin A1c; SD, standard deviation.

[a] Based on univariate analysis.

[b] Based on non-missing values.

A model can be simplified by only including the top 10 features, and we assessed the AUC of the top 10 feature model for each approach.

The data management was conducted using Palantir Foundry system (https://www.palantir.com/palantir-foundry/) housed in Sanofi. The control selection process via incidence-density sampling was conducted using SAS 9.4 (SAS Institute, Cary, NC), and all other statistical analyses were performed using R software version 3.4 (www.r-project.org). We used H_2O R package to implement the conventional and flexible ML processes.[27]

3 | RESULTS

A total of 3400 DKA cases and 11 780 controls were selected for the final analysis (Table 1). After the feature selection preprocessing 43 features were selected to predict DKA and were described in Table 2. Compared with controls, DKA cases were younger, had lower socioeconomic status and had more comorbidities. The mean of HbA1c level was 9.3% for DKA cases and 8.0% for controls.

TABLE 3 The performance of study models with full set of features in the test data set

Models	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Logistic regression	0.821 (0.804–0.837)	0.827 (0.814–0.839)	0.409 (0.321–0.497)	0.947 (0.925–0.969)
LASSO	0.821 (0.805–0.838)	0.827 (0.814–0.839)	0.407 (0.318–0.496)	0.948 (0.925–0.970)
XGBoost	0.819 (0.802–0.836)	0.825 (0.813–0.837)	0.414 (0.311–0.518)	0.944 (0.916–0.971)
DRF	0.817 (0.799–0.834)	0.827 (0.815–0.839)	0.420 (0.319–0.522)	0.944 (0.917–0.971)
Feedforward network	0.817 (0.799–0.834)	0.825 (0.812–0.837)	0.400 (0.291–0.508)	0.947 (0.920–0.975)

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; DRF, distributed random forest; LASSO, least absolute shrinkage and selection operator.

TABLE 4 Top 10 features by each study model

	Conventional machine learning		Flexible machine learning		
Rank	Logistic regression	LASSO	XGBoost	DRF	Feedforward network
1	Insurance type – uninsured	HbA1c	HbA1c	HbA1c	Race – Asian
2	HbA1c	Non-DKA hospitalization	Non-DKA hospitalization	White blood cell count	Insurance type – uninsured
3	Non-DKA hospitalization	Insurance type - uninsured	White blood cell count	Non-DKA hospitalization	Geographic region – Northeast
4	BMI	BMI	Hemoglobin	Hemoglobin	Race – African American
5	Pulse rate	Pulse rate	Pulse rate	Pulse rate	Geographic region – West
6	Psychiatric disorder	Psychiatric disorder	BMI	Random glucose level	Platelet count
7	Age	Age	Oxygen saturation	Respiratory rate	Gender - female
8	Calendar year of diabetes cohort entry	Calendar year of diabetes cohort entry	Random glucose level	Platelet count	HbA1c
9	White blood cell count	White blood cell count	Platelet count	eGFR	Non-DKA hospitalization
10	Acute infection	Acute infection	eGFR	BMI	White blood cell count

Abbreviations: BMI, body mass index; DKA, diabetic ketoacidosis; DRF, distributed random forest; eGFR, estimated glomerular filtration rate; HbA1c, hemoglobin A1c; LASSO, least absolute shrinkage and selection operator.

3.1 | Model performance with full set of features

In the training set, XGBoost outperformed the other 4 approaches with an AUC of 0.887, followed by feedforward network (AUC = 0.859), LASSO and logistic regression (AUC = 0.829 for each), and DRF (AUC = 0.808). In the test set, the AUC values ranged from 0.817 to 0.821 among these models and the difference decreased to 0.004 only (Table 3). The 95% CI of accuracy values ranged between 0.812 and 0.839. While the specificity values were higher than 0.9 for all models, the sensitivity values were only as high as 0.4. Consistent with the AUC findings, the differences in accuracy, sensitivity and specificity between flexible and conventional ML approaches were all minimal in the test set (Table 3). The confusion matrices are provided in Table S4.

3.2 | Feature importance

HbA1c level, non-DKA hospitalization, and white blood cell count were identified as one of top 10 features across all 5 models (Table 4 and Figure S1). Logistic regression and LASSO consistently identified the same top 10 features with slightly different ranks and most of them are well-established risk factors for DKA. XGBoost and DRF also identified almost the same top 10 features and eight were laboratory test results or vital signs, while feedforward network selected a very different set of top 10 features and six were social demographics. Compared with the conventional ML approaches, XGBoost and DRF identified the same five features in their top 10 features, while feedforward network identified the same four features.

In the logistic regression model, there were 16 positive predictors of DKA (i.e., with increased risk) such as higher HbA1c, non-DKA hospitalization, and so on, and seven negative predictors (i.e., with decreased risk) such as older age, higher annual household income, and so on (Table S5).

For each approach, the AUC values of the top 10 feature model were close to that of the full model in the test set ranging from 0.785 to 0.802 (Figure S2).

4 | DISCUSSION

We evaluated the performance of different ML approaches in a nested case–control study that used an EHR database to identify risk factors for DKA in adults with T1D. We found that prediction of DKA is achievable with either conventional or flexible ML approaches and that the differences in performance were minimal among these approaches. All approaches could consistently identify the known risk factors for DKA including HbA1c and non-DKA hospitalization. XGBoost and DRF included more laboratory test results or vital signs in their top 10 features, while feedforward included more social demographics. The flexible ML approaches only provided the relative importance for each predictor, while logistic regression could also estimate OR for each predictor. Therefore, interpreting the findings by the flexible ML approaches is challenging.

In this study, we found that flexible ML approaches offered very limited improvement over conventional ones in predicting DKA using an EHR database. In a systematic review of performance comparison of logistic regression with ML for clinical prediction modeling, the AUC of logistic regression and ML models were similar when comparisons were restricted to studies with low risk of bias.[3] Other studies also reported comparable AUCs between flexible and conventional ML models.[28,29] Although the accuracy and specificity were above 0.8 in this study, sensitivity was only 0.4. The main reason for the low sensitivity is the probability threshold we used in constructing the confusion matrix. For simplicity, we chose overall accuracy to optimize when determining the threshold. As a result, the threshold was set relatively high, which led to low sensitivity and high specificity.

In general, the DKA risk factors that were most often selected by different models are clinically sensible as a triggering factor for insulin deficiency or discontinuation of insulin. Known risk factors for DKA include high HbA1c level, infection, surgery/trauma, younger age, female sex, low BMI, low socioeconomic status, and so on,[4,30] and either conventional or flexible ML approaches included some of these factors in their top 10 features. However, XGBoost and DRF included more laboratory test results or vital sign measurements compared with conventional models, while feedforward network included more social demographics. Because each model uses different theories and algorithms to determine the feature importance, direct comparison cannot be made to explain the different features selected by various models. Despite this, one possible explanation for the difference is that some of the identified laboratory test results or vital signs reflect underlying causes of DKA, for example increased white blood cell count and pulse rate with infection, or elevated hemoglobin with dehydration. Another possible explanation is that some test results and social demographics are largely interrelated, e.g., high random blood glucose level and uninsured status may both be associated with suboptimal diabetes management.

This study has several limitations that should be considered. First, misclassification of DKA was possible, because we could not retrieve medical records to validate the outcome. However, the PPV for the proposed approach of DKA identification was 89%.[11] Second, unlike administrative claims data, there is no patient enrollment information in EHR data. To minimize the possibility that the medical encounters in EHR data are incomplete, the study patients were selected among those with non-emergency clinical activities recorded within 1 year prior to the index date, assuming all medical encounters were captured in the defined time window. In addition, we applied other inclusion and exclusion criteria which may limit the implementation of these predictive algorithms in a read-world setting. Third, several laboratory measurements including white blood cell count had missing data close to 60% and the impact on model performance from the use of imputation for these features is unknown. However, the high percentages of these missing values were driven by controls, suggesting they were most likely to lack laboratory testing because controls had fewer medical conditions which could trigger laboratory testing than cases. Fourth, the predictor selection is based on the knowledge of DKA and data availability. This priori feature selection may limit the learning potential performance of some of the more complex ML algorithms and affect the interpretation of feature importance results. Fifth, each model uses different theories and algorithms to determine the feature importance. Therefore, principled comparison cannot be made across models to explain the differences in feature selection. Last, the model performance was assessed based on one typical nested case–control study using an EHR database. This needs to be considered when interpreting the generalizability of these results.

Overall, the flexible ML approaches offered very limited performance improvement over conventional ones in predicting DKA using structured data recorded in EHR data source in this study. Both conventional and flexible ML approaches identified overlapping, but different top 10 risk factors for DKA. Further research is needed to determine the conditions under which the flexible ML approaches would outperform the conventional ones and vice versa and to better understand the reasons for differences in feature importance ranking among these approaches.

ACKNOWLEDGMENTS

The study was funded by Sanofi U.S. LLC. This study was designed as an analysis based on de-identified Electronic Health Record data, and there was no active enrollment or active follow-up of study subjects, and no data were collected directly from individuals.

CONFLICT OF INTEREST

Lin Li, Chuang-Chung Lee, Fang Liz Zhou, Cliona Molony, Zoran Doder, Evgeny Zalmover, and Juhaeri Juhaeri are employees of Sanofi.

Kristen Sharma was an employee of Sanofi at the time of this study, and is a current employee of Astellas Pharmaceuticals, Inc. Chuntao Wu was an employee of Sanofi at the time of this study, and is a current employee of Alexion Pharmaceuticals, Inc.

AUTHOR CONTRIBUTIONS

Chuntao Wu, Cliona Molony, Lin Li, Chuang-Chung Lee: Study concept and design. **Chuang-Chung Lee:** Data acquisition and statistical analysis. **Lin Li, Chuang-Chung Lee, Chuntao Wu, Fang Liz Zhou, Cliona Molony, Zoran Doder, Evgeny Zalmover, Kristen Sharma, Juhaeri Juhaeri:** Interpretation of data. **Lin Li:** Drafting of the manuscript. **Lin Li, Chuang-Chung Lee, Chuntao Wu, Fang Liz Zhou, Cliona Molony, Zoran Doder, Evgeny Zalmover, Kristen Sharma, Juhaeri Juhaeri:** Critical revision of the manuscript for important intellectual content. **Chuntao Wu, Juhaeri Juhaeri:** Study supervision. **Lin Li, Chuang-Chung Lee:** Accountable for accuracy and integrity.

ORCID

Lin Li https://orcid.org/0000-0002-3972-735X

REFERENCES

1. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318.
2. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012;16(5):933-951.
3. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22.
4. Misra S, Oliver NS. Diabetic ketoacidosis in adults. *BMJ*. 2015;351:h5660.
5. Benoit SR, Zhang Y, Geiss LS, Gregg EW, Albright A. Trends in diabetic ketoacidosis hospitalizations and in-hospital mortality - United States, 2000-2014. *MMWR Morb Mortal Wkly Rep*. 2018;67(12):362-365.
6. Sessa M, Khan AR, Liang D, Andersen M, Kulahci M. Artificial intelligence in Pharmacoepidemiology: a systematic review. Part 1-overview of knowledge discovery techniques in artificial intelligence. *Front Pharmacol*. 2020;11:1028.
7. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol*. 2008;8:48.
8. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. *Am J Epidemiol*. 2012;175(7):715-724.
9. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care*. 2013;36(4):914-921.
10. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA. Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health record data. *Pharmacoepidemiol Drug Saf*. 2018;27(10):1053-1059.
11. Bobo WV, Cooper WO, Epstein RA Jr, Arbogast PG, Mounsey J, Ray WA. Positive predictive value of automated database records for diabetic ketoacidosis (DKA) in children and youth exposed to antipsychotic drugs or control medications: a Tennessee Medicaid study. *BMC Med Res Methodol*. 2011;11:157.
12. Ford W, Self WH, Slovis C, McNaughton CD. Diabetes in the emergency department and hospital: acute care of diabetes patients. *Curr Emerg Hosp Med Rep*. 2013;1(1):1-9.
13. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)*. 2013;1(3):1035.
14. Mullet GM. Why regression coefficients have the wrong sign. *J Qual Technol*. 1976;8(3):121-126.
15. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. *J Interdiscip Math*. 2010;13(3):253-267.
16. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36(1):27-46.
17. Nykodym T, Kraljevic T, Hussami N, Rao A, Wang A. *Generalized Linear Modeling with H2O*. Mountain View, CA: H2O.ai; 2017; http://h2o.ai/resources/, Accessed June 26, 2019.
18. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Paper presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13–17, 2016, 2016; San Francisco, CA, USA.
19. Montana DJ, Davis L. Training feedforward neural networks using genetic algorithms. Paper presented at: the 11th international joint conference on artificial intelligence1989; Barcelona, Catalonia. *Spain*. 1989;1:762–767.
20. Candel A, Parmar V, LeDell E, Arora A. *Deep Learning with H_2O*. Mountain View, CA: H2O.ai; 2016.
21. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861-874.
22. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*. 2011;2(1):37-63.
23. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
24. Nathans LL, Oswald FL, Nimon K. Interpreting multiple linear regression: a guidebook of variable importance. *Pract Assess Res Eval*. 2012;17:1–19.
25. Cook D. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*. Sebastopol, CA: O'Reilly Media; 2016.
26. Gedeon TD. Data mining of inputs: analysing magnitude and functional measures. *Int J Neural Syst*. 1997;8(2):209-218.
27. *H_2O: R Interface for H2O* [computer program]. 2018.
28. Allam A, Nagy M, Thoma G, Krauthammer M. Neural networks versus logistic regression for 30 days all-cause readmission prediction. *Sci Rep*. 2019;9(1):9277.
29. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open*. 2020;3(1):e1918962.
30. Danne T, Garg S, Peters AL, et al. International consensus on risk management of diabetic ketoacidosis in patients with type 1 diabetes treated with sodium-glucose cotransporter (SGLT) inhibitors. *Diabetes Care*. 2019;42:1147-1154.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Li L, Lee C-C, Zhou FL, et al. Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data. *Pharmacoepidemiol Drug Saf*. 2021;30:610–618. https://doi.org/10.1002/pds.5199