Development and validation of a machine learning model for classification of next glucose measurement in hospitalized patients

Andrew D. Zale,[a] Mohammed S. Abusamaan,[a] John McGready,[b] and Nestoras Mathioudakis[a]*

[a]Associate Professor of Medicine, Division of Endocrinology, Diabetes & Metabolism, Department of Medicine, Johns Hopkins University School of Medicine, 1830 E. Monument Street Suite 333, Baltimore, MD 21287, USA
[b]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Summary

Background Inpatient glucose management can be challenging due to evolving factors that influence a patient's blood glucose (BG) throughout hospital admission. The purpose of our study was to predict the category of a patient's next BG measurement based on electronic medical record (EMR) data.

Methods EMR data from 184,361 admissions containing 4,538,418 BG measurements from five hospitals in the Johns Hopkins Health System were collected from patients who were discharged between January 1, 2015 and May 31, 2019. Index BGs used for prediction included the 5th to penultimate BG measurements ($N = 2,740,539$). The outcome was category of next BG measurement: hypoglycemic (BG $\leq$ 70 mg/dl), controlled (BG 71−180 mg/dl), or hyperglycemic (BG > 180 mg/dl). A random forest algorithm that included a broad range of clinical covariates predicted the outcome and was validated internally and externally.

Findings In our internal validation test set, 72·8%, 25·7%, and 1·5% of BG measurements occurring after the index BG were controlled, hyperglycemic, and hypoglycemic respectively. The sensitivity/specificity for prediction of controlled, hyperglycemic, and hypoglycemic were 0·77/0·81, 0·77/0·89, and 0·73/0·91, respectively. On external validation in four hospitals, the ranges of sensitivity/specificity for prediction of controlled, hyperglycemic, and hypoglycemic were 0·64−0·70/0·80−0·87, 0·75−0·80/0·82−0·84, and 0·76−0·78/0·87−0·90, respectively.

Interpretation A machine learning algorithm using EMR data can accurately predict the category of a hospitalized patient's next BG measurement. Further studies should determine the effectiveness of integration of this model into the EMR in reducing rates of hypoglycemia and hyperglycemia.

Introduction

Diabetes mellitus (DM) is highly prevalent within the general U.S. population, as over 10% of Americans carry this diagnosis.[1] Within the hospital, DM is even more prevalent, as a diagnosis of DM increases the rate of hospitalization by two-to-six fold, and nearly one in four hospitalized patients has diabetes.[2−4] Glycemic control may be especially hard to maintain in the

hospital setting owing to the presence of multiple evolving clinical parameters that exert varying influences on glucose homeostasis. For example, *nil per os* (NPO) status, use of steroid tapers, surgical procedures, varying antihyperglycemic regimens and doses, underlying infection, and changing renal function throughout the course of a patient's hospital stay can each influence the direction of blood glucose, and it can be difficult for a clinician to assess the combined impact of multiple factors in predicting whether the next glucose measurement will be in the desired range.[5−8]

Given the challenges in reconciling multiple clinical factors when making daily insulin dose adjustments in the hospital, it is not surprising that rates of hypoglycemia and hyperglycemia remain high in this setting.[9] Hypoglycemia during a hospital admission has been linked to an increased risk of mortality, longer length-of-stay, increased complications, and increased costs.[10,11] While there are commercially available

Abbreviations: BG, blood glucose; EMR, electronic medical record; NPO, nil per os; ICU, intensive care unit; BMI, body mass index; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; AUC, area under receiver operating curve; PPV, positive predictive value; NPV, negative predictive value; PLR, positive likelihood ratio; NLR, negative likelihood ratio; CGM, continuous glucose monitor; ICD, International Classification of Diseases.

*Corresponding author.

  E-mail address: nmathio1@jhmi.edu (N. Mathioudakis).

Research in context

Evidence before this study

We searched PubMed and Google Scholar for any study using the search terms "machine learning" or "prediction" or "artificial intelligence" and "blood glucose" or "hospital diabetes" or "hypoglycemia" between January 1, 1997 and September 1, 2021. We found that there has been research on clinical decision support tools that attempt to predict when a patient's blood glucose (BG) will be hypoglycemic within a certain time interval. To our knowledge, no study has validated a model that predicts a patient's next blood glucose reading in a general hospital population using three-level classification of hyperglycemic, controlled, and hypoglycemic.

Added value of this study

We derived a random forest prediction model to identify *both* short-term hypoglycemia and hyperglycemia in a general hospitalized population that was internally and externally validated. We used a dataset that included demographic variables, diabetes diagnosis, diabetes-related medications, glucose measurements, diet, in-hospital insulin and steroid regimen, labs, and vital signs. The model has higher detection of hyperglycemia and hypoglycemia compared to a model that attempts to predict a patient's next blood glucose reading based solely on the patient's current glucose measurement ("null model").

Implications of all available evidence

Our random forest model had high sensitivity and specificity in detecting dysglycemia based on a broad range of clinical predictors. This algorithm has the potential to give provider's warning about a patient's blood glucose trend to prevent dysglycemia in the hospital setting.

glycemic management software for use in the hospital (Glucommander®, EndoTool®, Glucotab®, etc.), these tools are costly, less accurate in the hypoglycemic range, have not been evaluated in critically ill patients, cannot be used to predict glucose values several hours later, and are not used by most health systems.[12−14]

Considering the lack of published prediction models for glucose in the hospital setting and the negative implications of dysglycemia, there is a compelling need for prediction models that can help clinicians gage the trajectory of glucose for a given patient in a short-term horizon. Since insulin dose adjustments usually occur daily in the hospital,[15] prediction of the next glucose classification needs to occur at the time of each glucose observation to be most useful to clinicians. Previously published models have focused predominantly on prediction of hypoglycemia, rather than a more clinically-relevant classification of hypoglycemia, controlled, and hyperglycemia.[16] Thus, we sought to develop and

validate a prediction model that predicted the class of next glucose observation, which could be potentially embedded in the EMR to increase usability and improve outcomes.

Methods

Dataset

This was a retrospective cohort study derived from EMR data obtained from five hospitals within the Johns Hopkins Health System in Maryland and the District of Columbia. Across the five hospitals, there were 4538,510 blood glucose (BG) measurements for 118,734 hospitalized patients discharged between January 1, 2015 and May 31, 2019 who received at least one unit of insulin (either subcutaneous or intravenous) and had at least four BG measurements during the admission. Details surrounding the data extraction and data processing for our dataset have been previously described.[17] We found no change in model performance when restricting our study population to patients with a documented diagnosis of diabetes mellitus, so we decided to include patients regardless of diabetes diagnosis as long as they fulfilled the inclusion criteria. Our population included both non-ICU and ICU patients, as we anticipate that translation of a prediction model for glucose in the hospital setting using EMR-based clinical decision support would have greater adoption if implemented broadly in the hospital rather than limited to certain units.

All BG measurements were included in calculations of summary measures used in the lookback windows for our prediction model. However, we limited the BG measurements used for prediction (i.e. index BG measurements) to the 5th to penultimate BG measurements throughout admission (Figure 2). Furthermore, we limited index BG measurements to those which were followed by another BG measurement within our defined prediction horizon of 5 min to 10 h. The rationale for these criteria for our index unit of observation is as follows: (a) to provide sufficient data for prediction, we did not begin predictions until at least 4 past BG measurements were available (typical number of BG measurements per day in non-critically ill patients); (b) we excluded BG measurements that were followed by a BG in 5 min as repeated values in this short window could reflect either clinical concern for a spurious BG reading or may reflect the same hypoglycemic or hyperglycemic episode[18]; (c) we selected 10 h as the upper limit of our prediction horizon, since this is typically the longest window of time between BG checks in hospitalized patients on medical/surgical floors (interval between bedtime and morning BG check); (d) the penultimate BG measurement was included as this was the last BG measurement that was proceeded by another BG measurement for prediction. In the process of including
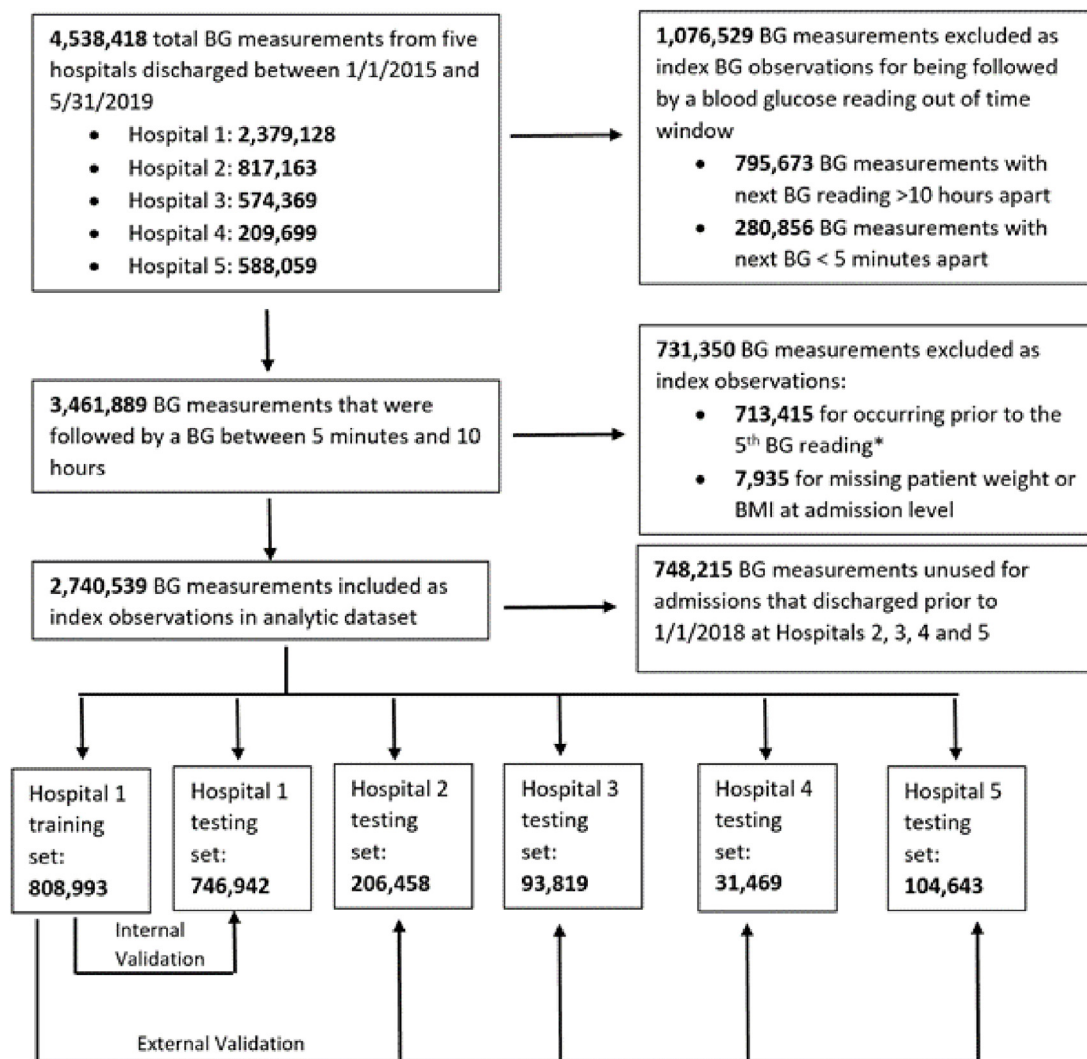
Figure 1. Study flowchart. *These BG readings were included as historical glucose data in the admission, but not as index observations.

only BG measurements with the next BG measurement between five minutes and ten hours, the final BG measurement was automatically excluded (as this BG measurement had no time to next BG measurement). Of note, although the first four BG measurements of the admission were not included as index BGs for prediction, these measurements *were* included in summary measures that fell within in the lookback windows for other exposure variables (e.g. average glucose over previous 24 h or average glucose since admission) relative to index BG measurements. Finally, since glucose and insulin dosing are associated with weight, we excluded 1391 admissions (1·2% of all admissions) in which weight or body mass index (BMI) data were not available.

Based on the above criteria, 2740,539 BG measurements from 48,370 patients across all five hospitals

were included as index observations in our prediction model. Figure 1 describes how the analytical dataset was created from the raw EMR data. The study protocol was approved by the Institutional Review Board of the Johns Hopkins School of Medicine with a waiver of informed consent.

Outcome

The outcome of interest was a three-level category of the next BG (i.e. following each index BG): hypoglycemic (≤70 mg/dL), controlled (71−180 mg/dL), or hyperglycemic (>180 mg/dL). We selected the cut-off of 70 mg/dL for hypoglycemia to align with our hospitals' hypoglycemia treatment policies. While we did consider the outcome of clinically significant hypoglycemia (BG <54 mg/dL), the event rate for this outcome was

exceedingly low and makes prediction challenging. We believe that using a higher threshold for hypoglycemia prediction would allow hospital-based clinicians more time to react and adjust. We considered treating BG as a continuous variable, but our modeling resulted in right-skewed error distributions with mean average errors in the 20−40 mg/dL ranges, causing poor predictive performance when attempting to further classify by threshold to treat for hyperglycemia and hypoglycemia.

Predictors

Predictors were selected based on clinical knowledge and findings from previous studies that support the inclusion of demographic variables, diabetes diagnosis, diabetes-related medications, BG measurements, diet, in-hospital insulin and steroid regimen, labs, and vital signs.[19−21] Supplemental Table 1 gives definitions of variables included in the model. We included both static (e.g. age) and time-varying predictors (e.g. lab results, vital signs). While hemoglobin A1C would be an obvious candidate for a predictor of glucose, since a majority of patients (∼65%) did not have an A1C obtained during admission, this variable was not included as a predictor. Time-varying predictors were defined in time frames that were relevant in relation to the index BG observation. For example, insulin doses and steroid doses on board at the time of the index BG observation were defined based on the pharmacologic duration of action of the given medication. Supplemental Table 2 summarizes each predictor variable by hospital.

Missing data and imputation

Most of the missing data were in the laboratory measurements and vital signs. We used the following approach to impute missing values: if the value was missing, we imputed the laboratory or vital sign value as equal to the mean value for that patient's admission; if the mean admission value could not be calculated due to absent data at the admission level, we imputed the laboratory or vital sign value as equal to the median of that hospital's value. Supplemental Table 3 lists each hospital's median for each laboratory and vital sign value and the proportion of patients in the full dataset (prior to any exclusion) with a missing value. To determine if missing data were missing at random, we compared the sample characteristics of samples in which a given laboratory value or vital sign was present in the EMR to those that required imputation for that predictor (Supplemental Table 4).

Model selection and development

We screened 14 different machine learning methods using five-fold cross validation on a subsample of 14,000 observations to determine which models to pursue further based on overall model accuracy and

computational efficiency. A summary of our machine learning screening results are reported in Supplemental Table 5 by Cohen's kappa (level of agreement).[22] Among the screened machine learning methods, we found that the LogitBoost algorithm, random forest, stochastic gradient boosting, and C5.0 had the highest level of agreement. Despite the relatively high model performance, they were relatively insensitive to the rarest event (hypoglycemia). We tested the re-sampling techniques of up-sampling, down-sampling, and SMOTE on a subset of our dataset in the LogitBoost, random forest, stochastic gradient boosting, and C5.0 methods. These sampling techniques mildly improved the predictive accuracy of hypoglycemia.

Thus, we decided to derive a machine learning algorithm in which we could tune the predictions based on the probability outputs. Previous research by Elbaz et al. demonstrated that stratification by class probability derived from *P*-values from a logistic regression output could improve model performance in predicting hypoglycemia.[23] The highest performing model was a random forest classification model with 35 random variables in each node. We used the *cutpointr* package in R to maximize the sum of sensitivity and specificity for each class to determine probability cutoffs.[24]

We conducted sensitivity analyses in which we limited the study population to patients with type 1 diabetes (T1DM) or type 2 diabetes who had basal insulin on board (T2DM) at the time of the BG reading and shortened the prediction horizon to five minutes to two hours. We used all exclusion and inclusion criteria as described above for the training and internal validation set, created the model, and used *cutpointr* to determine cut-points for each class in each model.

Model validation

Hospital 1 is the largest hospital in our health system, so data from this hospital were used to train the random forest model. We chronologically split observations from Hospital 1 into a training and validation set. Patients discharged prior to January 1, 2018 had all observations included in the training set and those discharged on or after January 1, 2018 had all observations included in the validation set (52%/48% training/validation split in Hospital 1). We used a chronological approach for validation to simulate how a developed machine learning method will perform when applied to future observations and minimize the influence of secular trends on predictive accuracy.

Internal validation was performed by using the model from the training set for Hospital 1 for prediction in the validation set for Hospital 1. External validation was performed by using the Hospital 1 training for prediction in Hospitals 2, 3, 4, and 5 separately. Although our five hospitals are part of a health system, the hospital size, demographics, and inpatient glucose management protocols
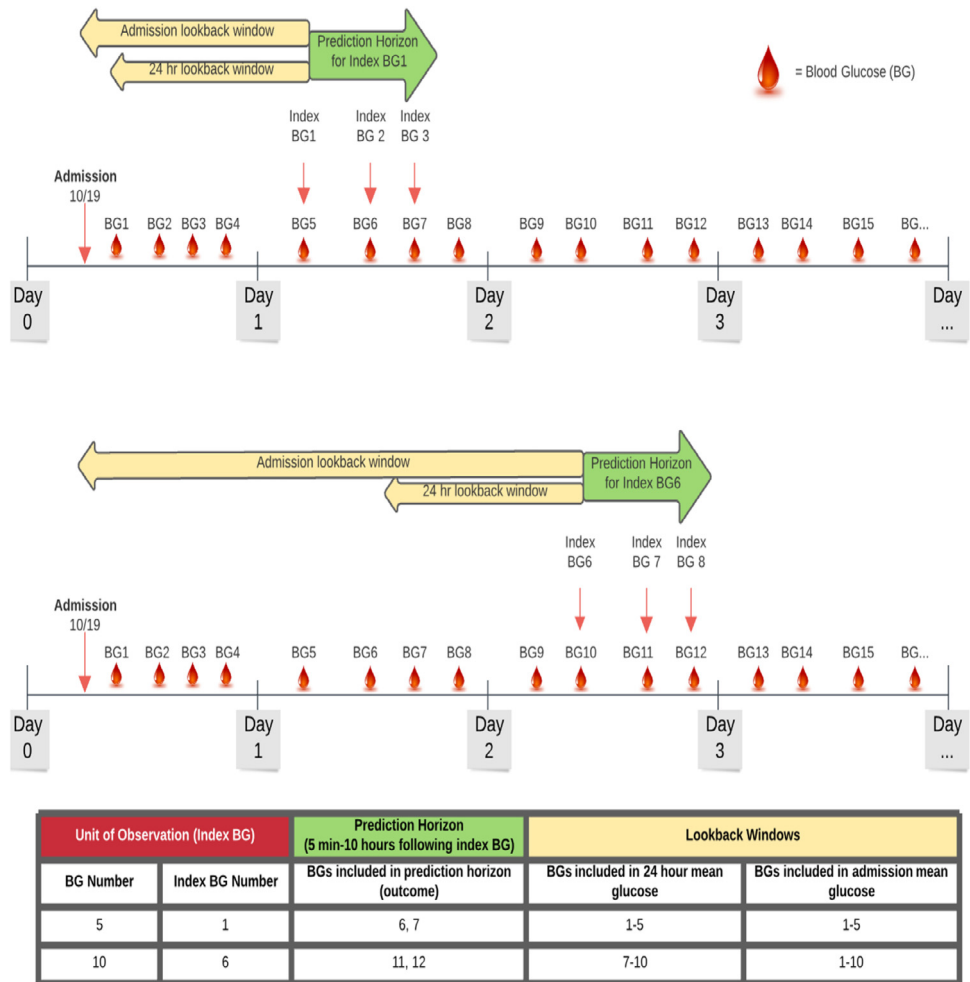
Figure 2. Lookback window, index observation, and prediction horizon. *Top:* Starting with the 5th blood glucose (BG) measurement, our machine learning algorithm begins to make predictions about a patient's next BG measurement in an expected prediction horizon of five minutes-ten hours. Data preceding the 5th BG measurement are included in summary statistics over the previous 24 h and admission as applicable.

Bottom: As a patient's admission becomes longer, BG measurements that were earlier in the admission may no longer be included in the 24 h summary statistics but will continue to be included in the admission summary statistics. The five minute to ten hour prediction horizon continues to roll with each new BG measurement.

differ to such an extent that we consider validation of data from Hospital 1 to other hospitals in our system to reflect external validation. Observations in the external validation sets were limited to patients who were discharged on or after January 1, 2018 for the same rationale as described above. We also evaluated the model by predicting the class of next BG reading based only on the index BG reading ("null model").

Statistical analysis

All statistical analyses were performed using R statistical software version 3.6.2 (R Foundation for Statistical Computing)[25] and Stata software, version 15.1 (StataCorp LLC).[26] Machine learning algorithms were developed using the 'Caret' R package.[27] The random forest algorithm originated from the 'parRF' function in 'Caret'.

Descriptive statistics were used to summarize the patient population at the admission level and at the index BG observation level by hospital. For continuous measures, normality was assessed using tests of skewness. All continuous variables were non-normally distributed and thus reported as medians and interquartile ranges. Categorical variables were reported as counts and frequencies. Differences in continuous measures across the five hospitals were evaluated using the Kruskal-Wallis test, and categorical measures were compared with the chi-square test. Differences between two datasets (internal or external validation sets) were evaluated using the Wilcoxon rank-sum test and chi-square test.

	Training		Validation			
		Internal	External			
	Hospital 1	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5
Number of Admissions	38,470	35,257	14,874	9296	4905	9630
Age, years, median (IQR)	59·0 (46·0, 69·0)	59·0 (46·0, 69·0)	63·0 (52·0, 74·0)***	71·0 (59·0, 81·0)***	71·0 (58·0, 82·0)***	74·0 (62·0, 84·0)***
Sex						
Female	18,785 (48·8%)	17,159 (48·7%)	7472 (50·2%)**	4775 (51·4%)***	2661 (54·3%)***	4834 (50·2%)*
Male	19,685 (51·2%)	18,098 (51·3%)	7402 (49·8%)	4521 (48·6%)	2244 (45·7%)	4796 (49·8%)
Race						
Black	14,581 (37·9%)	13,343 (37·8%)***	4361 (29·3%)***	2723 (29·3%)***	1277 (26·0%)***	1742 (18·1%)***
White	20,521 (53·3%)	18,419 (52·2%)	9599 (64·5%)	5278 (56·8%)	3114 (63·5%)	6105 (63·4%)
Other	3368 (8·8%)	3495 (9·9%)	914 (6·1%)	1295 (13·9%)	514 (10·5%)	1783 (18·5%)
BMI, kg/m, median (IQR)	26·8 (22·7, 31·9)	26·8 (22·8, 31·9)	28·1 (23·4, 34·0)***	27·4 (23·2, 32·8)***	25·8 (22·3, 30·5)***	26·3 (22·7, 30·7)***
Weight, pounds, median (IQR)	171·0 (141·0, 206·0)	171·0 (141·0, 206·0)	177·0 (145·0, 216·0)***	173·0 (142·0, 211·0)***	164·0 (136·0, 197·0)***	164·0 (137·0, 197·0)***
Diabetes diagnosis						
None	30,793 (80·0%)	27,714 (78·6%)***	10,354 (69·6%)***	6557 (70·5%)***	4044 (82·4%)***	6872 (71·4%)***
T1DM	558 (1·5%)	528 (1·5%)	320 (2·2%)	254 (2·7%)	62 (1·3%)	229 (2·4%)
T2DM	6788 (17·6%)	6636 (18·8%)	4137 (27·8%)	2430 (26·1%)	768 (15·7%)	2453 (25·5%)
Other DM	331 (0·9%)	379 (1·1%)	63 (0·4%)	55 (0·6%)	31 (0·6%)	76 (0·8%)
Home insulin	2914 (7·6%)	2350 (6·7%)***	1241 (8·3%)**	1348 (14·5%)***	431 (8·8%)**	1115 (11·6%)***
Home steroid	2156 (5·6%)	1801 (5·1%)**	466 (3·1%)***	555 (6·0%)	339 (6·9%)***	656 (6·8%)***
Average BG, median (IQR)	122·2 (106·6, 148·7)	121·4 (106·2, 148·0)**	119·7 (102·9, 158·4)***	129·0 (109·0, 167·5)***	114·8 (101·6, 143·2)***	124·2 (106·8, 157·5)***
Number of BG measurements, median (IQR)	15·0 (7·0, 35·0)	15·0 (7·0, 36·0)*	12·0 (6·0, 26·0)***	11·0 (6·0, 23·0)***	7·0 (5·0, 16·0)***	10·0 (5·0, 22·0)***
Length of stay, days, median (IQR)	5·5 (3·4, 9·4)	5·8 (3·7, 9·9)***	4·9 (3·1, 8·0)***	4·9 (3·2, 7·5)***	5·1 (3·4, 8·0)***	4·9 (3·3, 7·4)***

Table 1: **Baseline Characteristics of Study Population at Admission Level Comparing Each Validation Set to the Training Set.**

IQR = interquartile range; BMI = body mass index; T1DM = type 1 diabetes; T2DM = type 2 diabetes; BG = blood glucose.

  *   *P*-value < 0·05.

  **   *P*-value < 0·01.
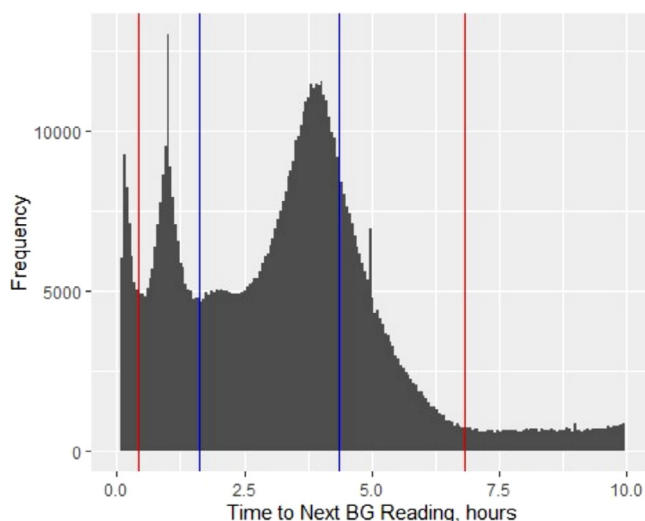
  ***   *P*-value < 0·001.

Figure 3. Time to next blood glucose reading in test set of Hospital 1. Red lines mark the middle 90% (5th percentile-95th percentile) and blue lines mark the middle 50% (25th percentile−75th percentile).

To report performance measures in a three-level classification problem, each class is considered individually and the summary statistics of all three classes are reported. Model performance measures were calculated by creating two-by-two contingency tables for the outcome of interest vs. all others (e.g. controlled vs. hyperglycemia or hypoglycemia; hyperglycemic vs. controlled or hypoglycemic; hypoglycemic vs. controlled or hyperglycemic). These contingency tables were used to calculate area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive and negative predictive value, and positive and negative likelihood ratio.

Role of the funding source

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The corresponding author (NM), ADZ, MSA, and JM had full access to all data in the study and they took the decision to submit the manuscript for publication.

Results

Cohort characteristics

Table 1 shows the baseline characteristics of the study population, comparing each hospital's validation set to the training set. There were statistically significant differences with respect to nearly all covariates for individual comparisons of each hospital to the training set. The training set population was younger (median age 59 years), majority male (19,685 admissions; 51·2%; *P*-value < 0·001 for all comparisons), and had a lower

proportion of White patients (20,521 admissions; 53·3%; *P*-value < 0·001 for all comparisons). The external validation sets had a median age of 63 to 74 years, were predominantly female (7472/14,874 [50·2%] to 2661/4905 [54·3%] admissions) and had a higher proportion of White patients (5278/9296 [56·8%] to 9599/14,874 [64·5%] admissions;). A lower proportion of training set patients were on insulin at home (2914 admissions; 7·6%; *P*-value < 0·01 for all comparisons) than the external validation set patients (1241/14,874 [8·3%] to 1348/9296 [14·5%] admissions; median range:). The 25th−75th and 5th−95th percentiles for time to next BG reading were 1·63−4·37 h and 0·42−6·83 h, respectively, in the test set of Hospital 1 (Figure 3).

When comparing the hospital 1 validation set (internal validation) to the training set, there were demographic differences with respect to age and race, but not sex. The groups had similar weight and BMI; however, the populations differed in regard to diabetes diagnosis, home insulin use, home steroid use, average admission BG, number of BG measurements, and length of stay. Of note, global comparison showed that the hospitals differed statistically with respect to all characteristics, both at the admission level (Supplemental Table 6) and the index BG level (Supplemental Table 2).

Model specification

The most important variables in the random forest model were index BG value, 24 h minimum BG, 24 h average BG, 24 h peak BG, and previous BG. Notably, insulin doses and home medications were *not* among the top 20 most important variables (Figure 4). Figure 5 shows an algorithm to convert an observation's random forest probabilities for each glucose outcome to achieve a final class prediction. If a BG value had a probability
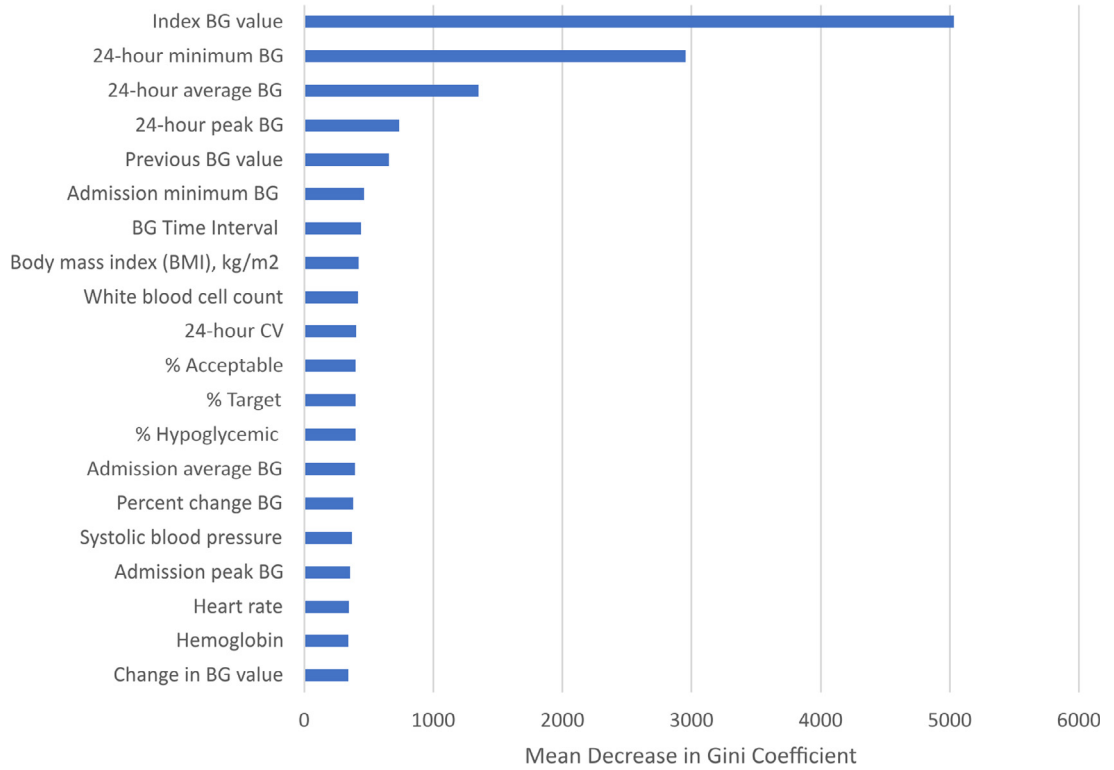
Figure 4. Variable Importance Plot of Top 20 Predictors. Variable importance plot based on the mean decrease in Gini, which is a measure of how much heterogeneity (i.e. misclassification) is lost when a predictor is used in a random forest node.
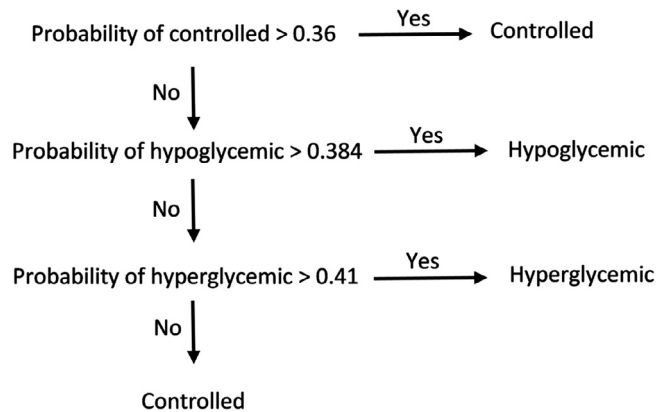

Figure 5. Probability cutpoints to determine the predicted class of a patient's next BG reading. This shows how probabilities for each BG category are used in an algorithmic fashion to determine the final class of BG. Cutpoints were selected that maximized the sum of sensitivity and specificity for each class.

of controlled > 0·36, that observation was predicted controlled; if a BG value had a probability of controlled ≤ 0·36 and a probability of hypoglycemic > 0·384, that observation was predicted hypoglycemic. If neither of these first two conditions were met and the observation had a probability of hyperglycemic > 0·41, that observation was predicted hyperglycemic. If all three probability thresholds were not reached, then the observation defaulted to the controlled classification.

Model performance
The training set prevalence for controlled, hyperglycemic, and hypoglycemic were 0·74, 0·25 and 0·01, respectively. In the Hospital 1 test set, 0·73, 0·26 and

	Null Model			Random Forest Model		
	Controlled	**Hyper**	**Hypo**	**Controlled**	**Hyper**	**Hypo**
Prevalence	0·73	0·26	0·01	0·73	0·26	0·01
AUROC	–	–	–	0·87 (0·87, 0·87)	0·91 (0·91, 0·91)	0·90 (0·90, 0·91)
Sensitivity	0·87 (0·87, 0·87)	0·70 (0·70, 0·70)	0·25 (0·25, 0·26)	0·77 (0·77, 0·77)	0·77 (0·77, 0·77)	0·73 (0·72, 0·74)
Specificity	0·68 (0·68, 0·68)	0·91 (0·90, 0·91)	0·99 (0·99, 0·99)	0·81 (0·80, 0·81)	0·89 (0·89, 0·89)	0·91 (0·91, 0·91)
PPV	0·86 (0·86, 0·87)	0·72 (0·72, 0·72)	0·25 (0·25, 0·26)	0·91 (0·91, 0·92)	0·70 (0·70, 0·70)	0·10 (0·10, 0·11)
NPV	0·70 (0·70, 0·70)	0·90 (0·90, 0·90)	0·99 (0·99, 0·99)	0·57 (0·57, 0·57)	0·92 (0·92, 0·92)	1·00 (1·00, 1·00)
PLR	2·75 (2·73, 2·76)	7·42 (7·36, 7·48)	22·73 (21·87, 23·62)	3·99 (3·95, 4·02)	6·84 (6·79, 6·90)	7·79 (7·69, 7·90)
NLR	0·19 (0·18, 0·19)	0·33 (0·33, 0·33)	0·75 (0·75, 0·76)	0·28 (0·28, 0·28)	0·26 (0·26, 0·26)	0·30 (0·29, 0·30)

Table 2: **Null model and random forest model performance and 95% confidence intervals on internal chronologic validation.**
AUC = area under the receiver operating characteristic curve; PPV = positive predictive value; NPV = negative predictive value; PLR = positive likelihood ratio; NLR = negative likelihood ratio; hyper = hyperglycemia; hypo = hypoglycemia. 95% CI shown in parentheses.
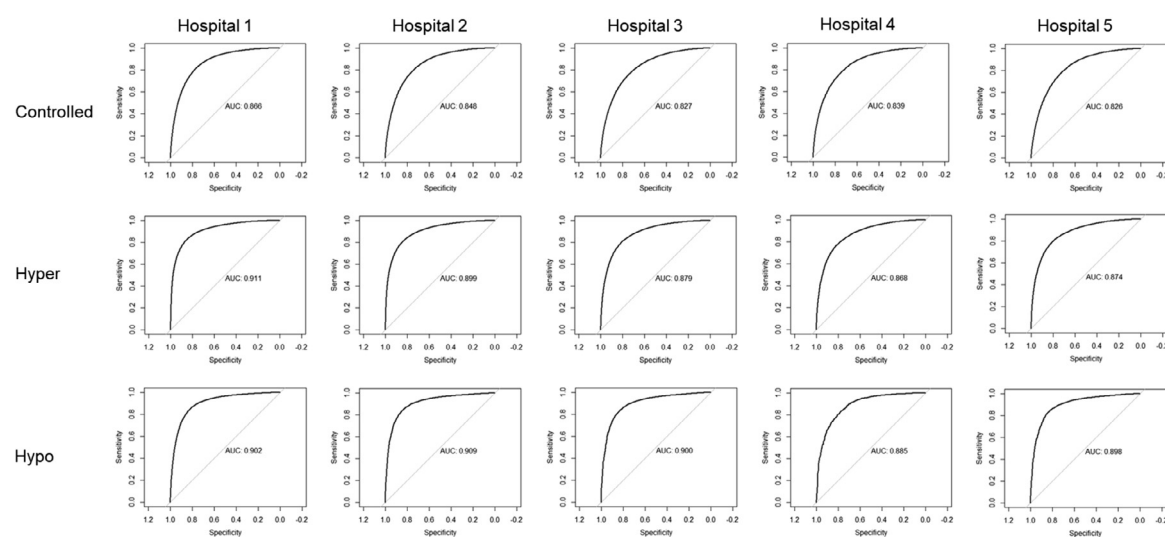

Figure 6. AUC performance of each class of prediction by hospital. AUC curves were plotted for each class (controlled, hyperglycemia, and hypoglycemia) by comparing the sensitivity and specificity at different cutpoints for each class individually (controlled vs. not controlled, hyperglycemic vs. not hyperglycemic, and hypoglycemic vs. not hypoglycemic). The AUC curves of Hospital 1 are the class-specific model performance in the test set and the AUC curves of Hospitals 2−5 are the class-specific model performances in the external validation sets.

0·01 of BG measurements were controlled, hyperglycemic, and hypoglycemic respectively (Table 2). The random forest AUC for the classes of controlled, hyperglycemic and hypoglycemic were 0·87, 0·91, and 0·90, respectively (Figure 6). The sensitivity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·77, 0·77, and 0·73, respectively. The specificity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·81, 0·89, and 0·91, respectively. On external validation (Table 3), the random forest AUC for the classes of controlled, hyperglycemic, and hypoglycemic ranged between 0·83−0·91, 0·87−0·90, and 0·85−0·90, respectively. The external validation sensitivity for prediction of controlled, hyperglycemic, and hypoglycemic were 0·64−0·70, 0·75−0·80, and 0·76−0·78, respectively. The specificity for prediction of controlled,

hyperglycemic, and hypoglycemic were 0·80−0·87, 0·82−0·84, and 0·87−0·90, respectively.

The presence of hyperglycemia and hypoglycemia in the internal validation set patients with T1DM was 0·50 and 0·03 (Supplemental Table 7). In a parallel random forest model trained and internally validated on patients with only T1DM, sensitivity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·74, 0·65, and 0·53, respectively. The specificity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·71, 0·88, and 0·90, respectively. The presence of hyperglycemia and hypoglycemia in the internal validation set patients with T2D was 0·51 and 0·01. In a parallel random forest model trained and internally validated on patients with only T2DM, sensitivity for a prediction of controlled, hyperglycemic, and

	Hospital 2			Hospital 3			Hospital 4			Hospital 5		
	Controlled	Hyper	Hypo	Controlled	Hyper	Hypo	Controlled	Hyper	Hypo	Controlled	Hyper	Hypo
Prevalence	0.65	0.33	0.02	0.63	0.35	0.03	0.65	0.33	0.02	0.68	0.30	0.02
AUROC	0.85 (0.85, 0.85)	0.90 (0.90, 0.90)	0.85 (0.85, 0.85)	0.91 (0.90, 0.91)	0.88 (0.88, 0.88)	0.90 (0.89, 0.91)	0.84 (0.84, 0.84)	0.87 (0.86, 0.87)	0.89 (0.87, 0.90)	0.83 (0.82, 0.83)	0.87 (0.87, 0.88)	0.90 (0.89, 0.90)
Sensitivity	0.69 (0.69, 0.69)	0.80 (0.80, 0.81)	0.78 (0.77, 0.80)	0.64 (0.64, 0.65)	0.79 (0.79, 0.80)	0.78 (0.76, 0.80)	0.70 (0.69, 0.71)	0.76 (0.75, 0.77)	0.69 (0.65, 0.73)	0.69 (0.69, 0.70)	0.75 (0.74, 0.75)	0.76 (0.75, 0.78)
Specificity	0.84 (0.84, 0.85)	0.84 (0.84, 0.85)	0.89 (0.88, 0.89)	0.84 (0.84, 0.84)	0.82 (0.82, 0.83)	0.87 (0.87, 0.87)	0.81 (0.81, 0.82)	0.82 (0.82, 0.83)	0.90 (0.90, 0.91)	0.80 (0.80, 0.81)	0.84 (0.84, 0.84)	0.88 (0.88, 0.88)
PPV	0.89 (0.89, 0.89)	0.72 (0.72, 0.72)	0.12 (0.12, 0.12)	0.87 (0.87, 0.87)	0.71 (0.70, 0.71)	0.13 (0.13, 0.14)	0.88 (0.87, 0.88)	0.68 (0.67, 0.69)	0.11 (0.10, 0.12)	0.88 (0.88, 0.89)	0.67 (0.66, 0.67)	0.12 (0.12, 0.13)
NPV	0.60 (0.59, 0.60)	0.90 (0.89, 0.90)	1.00 (0.99, 1.00)	0.59 (0.58, 0.59)	0.88 (0.88, 0.88)	0.99 (0.99, 0.99)	0.59 (0.58, 0.60)	0.87 (0.87, 0.88)	0.99 (0.99, 1.00)	0.55 (0.55, 0.56)	0.89 (0.88, 0.89)	0.99 (0.99, 0.99)
PLR	4.41 (4.33, 4.48)	5.16 (5.10, 5.23)	6.85 (6.71, 6.99)	4.02 (3.92, 4.12)	4.45 (4.37, 4.53)	6.08 (5.91, 6.25)	3.77 (3.62, 3.93)	4.25 (4.12, 4.38)	7.17 (6.70, 7.67)	3.55 (3.47, 3.63)	4.73 (4.65, 4.82)	6.45 (6.27, 6.63)
NLR	0.37 (0.37, 0.37)	0.23 (0.23, 0.24)	0.24 (0.23, 0.26)	0.42 (0.42, 0.43)	0.25 (0.25, 0.26)	0.25 (0.23, 0.27)	0.37 (0.36, 0.38)	0.29 (0.28, 0.30)	0.35 (0.30, 0.39)	0.38 (0.38, 0.39)	0.30 (0.29, 0.31)	0.27 (0.25, 0.29)

Table 3: **Random forest model performance and 95% confidence intervals upon external validation.**

AUC = area under the receiver operating characteristic curve; PPV = positive predictive value; NPV = negative predictive value; PLR = positive likelihood ratio; NLR = negative likelihood ratio; hyper = hyperglycemia; hypo = hypoglycemia. 95% CI shown in parentheses.

hypoglycemic were 0·80, 0·64, and 0·35, respectively. The specificity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·67, 0·85, and 0·95, respectively.

When limiting the prediction horizon to BG readings between five minutes and two hours, the prevalence among the internal validation set of controlled, hyperglycemic, and hypoglycemic BG readings were 0·71, 0·27, and 0·02, respectively (Supplemental Table 8). In a parallel random forest model trained and internally validated on patients with BG readings in this prediction horizon, the sensitivity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·82, 0·82, and 0·82, respectively. The specificity for a prediction of controlled, hyperglycemic, and hypoglycemic were 0·83, 0·93, and 0·92, respectively.

Discussion

Exploring a broad panel of machine learning algorithms, we identified the random forest function as a modeling technique capable of predicting the next class of a hospitalized patient's BG measurement with a high degree of detection for the two minority classes of hyperglycemia and hypoglycemia. To our knowledge, our prediction model is the first inpatient-based model to use the short-term prediction horizon of *next* BG measurement without the use of continuous glucose monitoring. Unlike other inpatient-based glucose prediction models, which tend to focus only on the outcome of hypoglycemia and limit the prediction horizon to any time during the admission, the short-term prediction horizon and three-factor level classification of the present model could be translated into a clinical decision support tool to influence glycemic management in near real-time.

Unlike previous models published in this area, we found that few clinical predictors besides glycemic measures were informative. These findings support the notion described by Kovatchev et al. that there are measurable disturbances in blood glucose in the 48 h preceding a hypoglycemic event.[28–30] Since changes in insulin doses are ultimately reflected by changes in glucose trends, it is not surprising that glycemic summary measures confer the most information about the next glucose measurement. From an implementation standpoint, the fact that glycemic data are readily available in EMR systems and easily extractable means that a model that is primarily derived from glycemic measures could be more easily adopted by outside hospital systems. Moreover, despite our broader inclusion criteria, our model performed as well as models based on more restrictive criteria, which may increase the generalizability and clinical application of the model. A meta-analysis of the performance of hypoglycemic prediction tools that did not use CGM data calculated pooled estimates for sensitivity, specificity, PLR and NLR of 0·76 (0·66

−0·84), 0·92 (0·88−0·95), 10·14 (6·13−16·77), and 0·26 (0·17−0·38), respectively.[31] Furthermore, our model performs well with high discrimination of a given class based on the class-specific AUCs.

Among 19 published hypoglycemia prediction models, only six were developed using blood glucose data rather than CGM data.[31] Mueller et al. used a hypothesis free, Bayesian machine learning analytics platform to create ensembles of linear models.[32] Their AUC in a hold-out model for hypoglycemia in the next 12 months was 0·77. Ruan et al. analyzed how accurately 18 different machine learning methods predicted whether an admission would have at least one event of hypoglycemia.[19] Their XGBoost model had an AUC of 0·96 with a precision of 0·88 and recall of 0·70. Jensen et al. predicted nocturnal hypoglycemia with an AUC of 0·79, sensitivity of 0·75, and specificity of 0·70; however, two of the four components in their linear discriminant analysis involved glucose measurements from CGM. Sudharsan et al. utilized self-monitored blood glucose checked once or twice daily to predict hypoglycemia in the next 24 h.[33] Their sensitivity was 0·92 and specificity was 0·70, and their model's specificity rose to 0·90 when they added medication information. Compared to these previously published models, our random forest model was able to discern hypoglycemia and hyperglycemia from controlled in a broader inpatient population within a narrower prediction horizon.

We recently published findings using a stochastic gradient boosting machine learning model for prediction of iatrogenic hypoglycemia using a similar EMR dataset;[17] while our previous model had greater sensitivity for prediction of hypoglycemia, the present random forest model had a higher PLR and positive predictive value for hypoglycemia, which may be more impactful in modifying end user behavior by minimizing false positives. Of note, in our previous model, the prediction horizon for hypoglycemia was 24 h from the index BG measurement, whereas in the present study, it was the *next* BG measurement (variable time interval). In addition, the three-level outcome prediction of the present study has broader clinical applicability than one focused solely on hypoglycemia.

Artificial intelligence tools have been linked to improved diagnostic confidence.[34] Our machine learning model can support or refute a physician's belief about what the patient's glucose status will be at the next BG measurement. PLRs and NLRs express how much the probability of a diagnosis increases or decreases from the pre-test probability based on the model's prediction[16,19,35] Likelihood ratios between 5 and 10 and 0·1 to 0·2 represent moderate shifts in pretest to posttest probability; likelihood ratios between 2 and 5 and 0·5 to 0·2 generate small, but sometimes important, changes in probability.[36] In this study, the relatively small PLR for controlled glucose means that being predicted as controlled should not change a provider's expectations because the pre-test probability of controlled would be relatively high in most cases. On the other hand, for rare outcomes, like hypoglycemia, the PLR is higher. Thus, if our model alerts the clinician to hypoglycemia, providers should have a low threshold to modify the antihyperglycemic regimen to avoid an adverse outcome.

Our sensitivity analyses highlight the importance of key considerations when selecting exclusion criteria for the development of a machine learning model. When comparing the general patient population model to patients with known diagnoses of T1DM or T2DM, the general patient population model has a similar PLR to that of the T2DM model; the T1DM has the lowest PLR for the prediction of hypoglycemia. The sensitivity in the general patient population model is much higher, compared to a higher specificity in the T1DM and T2DM models. Interestingly, choosing a more homogenous patient population did not increase sensitivity for hypoglycemia. When using the shorter prediction horizon of 5 min to 2 h, there was a global improvement in model performance measures, especially in the PLR of hyperglycemia and hypoglycemia. This result is not unsurprising because limiting the prediction horizons allows for less time between BG readings. Therefore, we suspect that the index BG reading has a higher correlation to the next BG reading when limited to a smaller prediction horizon. In terms of clinical utility, however, the generalizability of this model is limited since most patients in the hospital do not receive BG measurements every two hours.

Our study has several strengths. Notably, we predicted the class of a patient's next BG measurement, which is a much narrower time frame than previously published studies and allows for potential integration in the EMR as a decision support tool. We also included as many glycemic parameters as possible. We took a naÿve approach to variable and model selection so that we could select the most accurate model regardless of technique and included variables. The sample size and number of predictor variables were large. Furthermore, we demonstrated that our external validation sets were each different from our training set and externally validated our model on four other hospitals in our health system. Our external validation sets performed well in detecting hypoglycemia and hyperglycemia in hospitalized patients. Of note, while there are commercially available software for glycemic prediction/insulin dosing support that have been evaluated in multiple healthcare settings (e.g. Glucommander), the model specifications for these algorithms have not been published due to proprietary considerations.

There are some limitations to the present study. We were unable to extract all pertinent information from a patient's admission such as dextrose dose, insulin doses from total parenteral nutrition formulations, and carbohydrates actually consumed during meals. Furthermore, some measures like A1C were excluded since

most patients did not have this lab test during their admission. We limited our model to start making predictions on the fifth BG measurement during the admission to guarantee a minimum amount of historical glucose data from which to make predictions. Likewise, we limited our training and validation sets to include BG observations that were followed by a BG observation within a five minute to ten-hour window. Since our goal was to create a general machine learning model that is agnostic to the time between BG observations (which cannot be necessarily predicted in a hospital setting), we believe that this prediction horizon allows for a clinically useful prediction horizon. Our missing data analysis showed that for most patient characteristics, there was a statistically significant difference between the BG readings that had laboratory values and vital signs compared to those BG readings that required imputation for that given predictor. However, most of these statistically significant differences (expected given large dataset) are small in magnitude and we are unsure about how to interpret the clinical significance of the non-randomness in missing laboratory values or vital signs. Nevertheless, these findings highlight a limitation to using EMR data that is incomplete based on the variable data ordered by different providers.

Additionally, when training our model, we discovered that our modeling techniques outperformed in the general hospitalized patient population compared to patients with documented diagnoses of T1DM or T2DM with basal insulin on board. We do recognize that this is an important clinical limitation since these patients are considered at high risk for dysglycemia. We hypothesize that some of this is related to coding practices in our EMR, as we are entirely reliant on ICD-10 codes for diabetes diagnosis. Moreover, we found that only 1.3% of readings in patients with T2DM were hypoglycemic, compared to 3.1% of readings in patients with T1DM and 1.4% of patients with no known diagnosis of diabetes. We believe that since our population included any patient who received insulin, this insulin exposure poses the greatest risk for hypoglycemia, even without a known diagnosis of diabetes. Our hypoglycemic model suffers from PPV of 0·10, driven by a specificity of the controlled (majority) class of 0·81. This limitation is similar to previous findings that demonstrated a high false alarm rates despite high sensitivity and specificity due to class imbalance.[37] In previous research, the high hypoglycemic false alarm rate was reduced by increasing specificity by excluding any hypoglycemic event that was transient. Since we did not use CGM data, we were unable to analyze the effects of transient hypoglycemic events on model performance. Finally, our model performance was lower among patients with coded type 1 and type 2 diabetes compared to the overall cohort; since these patients are often the most challenging to manage in the hospital, further research is needed to identify additional predictor variables for this population.

To our knowledge, this is the first machine learning model that can predict the 3-level outcome of glycemic control for the next glucose measurement throughout a patient's admission. Our model does not rely on data from continuous glucose monitors and can be applied to any hospitalized patient, regardless of insulin use or ICU status. Our next step will be to adapt and integrate our model into the EMR. Further studies will be needed to evaluate whether information about the predicted level of glycemic control based on this machine learning model has the potential to improve glycemic control and related clinical sequalae.

Contributors
NM and MA contributed to study concept and design, data extraction and assessment, verified the underlying data, statistical synthesis, interpretation of data, and drafted the manuscript.

AZ contributed to study concept and design, statistical analysis, interpretation of data, and drafted the manuscript. JM contributed to study design and concept, statistical analysis, and critical revision of the manuscript. All authors reviewed and edited the manuscript

Data availability statement
The data underlying this article will be shared on reasonable request to the corresponding author.

Supplementary materials
Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2022.101290.

References

1 Prevention CfDCa. *National Diabetes Statistics Report, 2020*. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020:12–15.
2 Bo S, Ciccone G, Grassi G, et al. Patients with type 2 diabetes had higher rates of hospitalization than the general population. *J Clin Epidemiol*. 2004;57(11):1196–1201.
3 De Berardis G, D'Ettorre A, Graziano G, et al. The burden of hospitalization related to diabetes mellitus: a population-based study. *Nutr Metab Cardiovasc Dis*. 2012;22(7):605–612.
4 Kufeldt J, Kovarova M, Adolph M, et al. Prevalence and distribution of diabetes mellitus in a maximum care hospital: urgent need for HbA1c-screening. *Exp Clin Endocrinol Diabetes*. 2018;126(2):123–129.
5 Lemieux I, Houde I, Pascot A, et al. Effects of prednisone withdrawal on the new metabolic triad in cyclosporine-treated kidney transplant patients. *Kidney Int*. 2002;62(5):1839–1847.
6 Hricik DE, Bartucci MR, Moir EJ, Mayes JT, Schulak JA. Effects of steroid withdrawal on posttransplant diabetes mellitus in cyclosporine-treated renal transplant recipients. *Transplantation*. 1991;51(2):374–377.
7 Yokoyama H, Fukumoto S, Koyama H, Emoto M, Kitagawa Y, Nishizawa Y. Insulin allergy; desensitization with crystalline zinc-insulin and steroid tapering. *Diabetes Res Clin Pract*. 2003;61(3):161–166.
8 Gosmanov AR, Umpierrez GE. Management of hyperglycemia during enteral and parenteral nutrition therapy. *Curr Diabetes Rep*. 2013;13(1):155–162.
9 Schnipper JL, Barsky EE, Shaykevich S, Fitzmaurice G, Pendergrass ML. Inpatient management of diabetes and hyperglycemia among general medicine patients at a large teaching hospital. *J Hosp Med*. 2006;1(3):145–150.
10 Brodovicz KG, Mehta V, Zhang Q, et al. Association between hypoglycemia and inpatient mortality and length of hospital stay in hospitalized, insulin-treated patients. *Curr Med Res Opin*. 2013;29(2):101–107.
11 Varlamov EV, Kulaga ME, Khosla A, Prime DL, Rennert NJ. Hypoglycemia in the hospital: systems-based approach to recognition, treatment, and prevention. *Hosp Pract (1995)*. 2014;42(4):163–172.
12 Klonoff DC. The need for separate performance goals for glucose sensors in the hypoglycemic, normoglycemic, and hyperglycemic ranges. *Diabetes Care*. 2004;27(3):834–836.
13 Klonoff DC. Continuous glucose monitoring: roadmap for 21st century diabetes therapy. *Diabetes Care*. 2005;28(5):1231–1239.
14 Dassau E, Cameron F, Lee H, et al. Real-Time hypoglycemia prediction suite using continuous glucose monitoring: a safety net for the artificial pancreas. *Diabetes Care*. 2010;33(6):1249–1254.
15 Inzucchi SE. Clinical practice. Management of hyperglycemia in the hospital setting. *N Engl J Med*. 2006;355(18):1903–1911.
16 Mathioudakis NN, Everett E, Routh S, et al. Development and validation of a prediction model for insulin-associated hypoglycemia in non-critically ill hospitalized adults. *BMJ Open Diabetes Res Care*. 2018;6(1):e000499.
17 Mathioudakis N, Abusamaan M, Shakarchi A, et al. Development and validation of a machine learning model to predict near-term risk of iatrogenic hypoglycemia in hospitalized patients. *JAMA Netw Open*. 2021;4(1):e2030913. https://doi.org/10.1001/jamanetworkopen.2020.30913.
18 Weinberg ME, Bacchetti P, Rushakoff RJ. Frequently repeated glucose measurements overestimate the incidence of inpatient hypoglycemia and severe hyperglycemia. *J Diabetes Sci Technol*. 2010;4(3):577–582.
19 Ruan Y, Bellot A, Moysova Z, et al. Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. *Diabetes Care*. 2020;43(7):1504–1511.
20 Winterstein AG, Jeon N, Staley B, Xu D, Henriksen C, Lipori GP. Development and validation of an automated algorithm for identifying patients at high risk for drug-induced hypoglycemia. *Am J Health Syst Pharm*. 2018;75(21):1714–1728.
21 Kyi M, Gorelik A, Reid J, et al. Clinical prediction tool to identify adults with type 2 diabetes at risk for persistent adverse glycemia in hospital. *Can J Diabetes*. 2021;45(2):114–121. e3.
22 Ben-David A. About the relationship between ROC curves and Cohen's kappa. *Eng Appl Artif Intell*. 2008;21(6):874–882.
23 Elbaz M, Nashashibi J, Kushnir S, Leibovici L. Predicting hypoglycemia in hospitalized patients with diabetes: a derivation and validation study. *Diabetes Res Clin Pract*. 2021;171:108611.
24 Thiele C, Hirschfeld G, Cutpointr. Improved Estimation and Validation of Optimal Cutpoints in R. *J Stat Softw*. 2021;98(11):1–27.
25 R Core Team. *R: A language and Environment For Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017.
26 StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: stataCorp LLC; 2017.
27 Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1–26.
28 Palerm CC, Bequette BW. Hypoglycemia detection and prediction using continuous glucose monitoring-a study on hypoglycemic clamp data. *J Diabetes Sci Technol*. 2007;1(5):624–629.
29 Kovatchev BP, Cox DJ, Farhy LS, Straume M, Gonder-Frederick L, Clarke WL. Episodes of severe hypoglycemia in type 1 diabetes are preceded and followed within 48 h by measurable disturbances in blood glucose. *J Clin Endocrinol Metab*. 2000;85(11):4287–4292.
30 Kovatchev BP, Cox DJ, Gonder-Frederick LA, Young-Hyman D, Schlundt D, Clarke W. Assessment of risk for severe hypoglycemia among adults with IDDM: validation of the low blood glucose index. *Diabetes Care*. 1998;21(11):1870–1875.
31 Kodama S, Fujihara K, Shiozaki H, et al. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: meta-analysis. *JMIR Diabetes*. 2021;6(1):e22458.
32 Mueller L, Berhanu P, Bouchard J, et al. Application of machine learning models to evaluate hypoglycemia risk in type 2 diabetes. *Diabetes Ther*. 2020;11(3):681–699.
33 Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J Diabetes Sci Technol*. 2015;9(1):86–90.
34 Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am J Med*. 2018;131(2):129–133.
35 Parikh R, Parikh S, Arun E, Thomas R. Likelihood ratios: clinical application in day-to-day practice. *Indian J Ophthalmol*. 2009;57(3):217–221.
36 Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The evidence-based medicine working group. *JAMA*. 1994;271(9):703–707.
37 Dave D, Erraguntla M, Lawley M, et al. Improved low-glucose predictive alerts based on sustained hypoglycemia: model development and validation study. *JMIR Diabetes*. 2021;6(2):e26909.