Incorporating real-world evidence into the development of patient blood glucose prediction algorithms for the ICU

Oisin Fitzgerald,1,2 Oscar Perez-Concha,1 Blanca Gallego,1 Manoj K. Saxena,2 Lachlan Rudd,3 Alejandro Metke-Jimenez,4 and Louisa Jorm1

1Centre for Big Data Research in Health, UNSW Sydney, Sydney, NSW, Australia 2The George Institute for Global Health, UNSW Sydney, Sydney, NSW, Australia 3Data and Analytics, eHealth NSW, Chatswood, NSW, Australia, and 4Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia

Received 11 January 2021; Revised 10 February 2021; Editorial Decision 2 March 2021; Accepted 22 March 2021

ABSTRACT

Objective: Glycemic control is an important component of critical care. We present a data-driven method for predicting intensive care unit (ICU) patient response to glycemic control protocols while accounting for patient heterogeneity and variations in care.

Materials and Methods: Using electronic medical records (EMRs) of 18,961 ICU admissions from the MIMIC-III dataset, including 318,574 blood glucose measurements, we train and validate a gradient boosted tree machine learning (ML) algorithm to forecast patient blood glucose and a 95% prediction interval at 2-hour intervals. The model uses as inputs irregular multivariate time series data relating to recent in-patient medical history and glycemic control, including previous blood glucose, nutrition, and insulin dosing.

Results: Our forecasting model using routinely collected EMRs achieves performance comparable to previous models developed in planned research studies using continuous blood glucose monitoring. Model error, expressed as mean absolute percentage error is 16.5%–16.8%, with Clarke error grid analysis demonstrating that 97% of predictions would be clinically acceptable. The 95% prediction intervals achieve near intended coverage at 93%–94%.

Discussion: ML algorithms built on observational data sources, such as EMRs, present a promising approach for personalization and automation of glycemic control in critical care. Future research may benefit from applying a combination of methodologies and data sources to develop robust methodologies that account for the variations seen in ICU patients and difficulty in detecting the extremes of observed blood glucose values.

Conclusion: We demonstrate that EMRs can be used to train ML algorithms that may be suitable for incorporation into ICU decision support systems.

Key words: critical care, blood glucose, prediction, electronic health records, machine learning

INTRODUCTION

Data science is expected to play a major role in enabling personalized medicine in the critical care setting.1–3 Critically ill patients who enter the intensive care unit (ICU) exhibit considerable heterogeneity, with large variation in treatment response among patients with similar diagnoses.4 The widespread adoption of electronic medical record (EMR) systems and the resulting accumulation of large clinical datasets enables the development of data-driven, precision medicine approaches to addressing clinical heterogeneity. By working well with the large size and complexity commonly found in these, data machine learning (ML) has become the analytic tool of choice.5

ML can play a role in tasks important in ICU precision medicine, such as prediction and state estimation,6 subgroup discovery,3 and
Glycemic control in the ICU is complex, with previous studies using observational datasets demonstrating considerable variation in ICU patient blood glucose levels. Stress-related hyperglycemia, defined as blood glucose levels above 180 mg/dL, is common on presentation to the ICU with 1 multi-ICU hospital reporting rates of 9% and 28.6% of non-diabetic and diabetic ICU patients over nearly 20 000 ICU admissions. Hyperglycemia evolved as an adaptive response to acute trauma (the “fight or flight” response). However, in ICU patients, hyperglycemia is a marker of increased morbidity and mortality and is potentially harmful. As such, glycemic control is considered a core aspect of critical care, with insulin therapy the primary approach to controlling blood glucose levels. However, a further complication is the risk that excessive treatment may cause hypoglycemia, where blood glucose falls below 70 mg/dL, with severe hypoglycemia (< 40 mg/dL) associated with increased mortality. Hypoglycemia (< 70 mg/dL) has been reported in 2% and 6% of diabetic and non-diabetic ICU patients over nearly 20 000 ICU admissions. However, while a broad consensus has been reached that moderate glycemic control should be the clinical standard, there is less evidence on glycemic control in critical care. We use Catboost, given recent evidence of its gradient boosting outperforming other ML algorithms at forecasting in complex hierarchical time-series datasets and being relatively quick to train. Additionally, in order to investigate the presence of patient heterogeneity and its impact on predictive performance, we cluster the observed glucose trajectories into subgroups and evaluate the ML algorithms on these groups.

**Materials and Methods**

**Data**

Our dataset is constructed from the publicly available MIMIC-III database (ethics approval was not required). There is variation in the recording of insulin inputs between the EMR applications (MetaVision and CareVue) used to construct MIMIC-III. In the CareVue portion, it is difficult to determine exactly when insulin was administered for some patients. As a result, the CareVue portion of the 60 000 ICU admissions available in MIMIC-III were excluded. The resulting dataset comprises deidentified health data associated with 18 691 ICU stays and 14 742 critical care patients at Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2008 and 2012.

The outcome of interest is the 2-hour ahead point of care blood glucose at each 2-hour time-step from a patient’s entry to ICU (Figure 1). We predict the value and a 95% prediction interval, a range of values in which we expect the observed value to fall 95% of the time. The dataset includes a range of available input features, fully listed in Supplementary Appendix A, including common point of care physiological measures, laboratory results, mechanical ventilation parameters, vasopressors, patient type, amount of dextrose in IV infusions, enteral nutrition, parenteral nutrition and past values.
of point of care blood glucose, laboratory glucose, and insulin inputs.

Insulin is recorded as intravenous (IV) infusion or push and bolus injection, with short-, medium-, and long-acting variants available for bolus injections and long and short for IV infusions and pushes. Insulin is coded as the total amount over the 2-hour time intervals for each of the 5 available route/type combinations. Aspects of the dataset construction, in particular the type of insulin inputs, are based on code made available by previous researchers. Nutritional composition (total calories and percentage of carbohydrates, sugar, fat, and protein) of enteral nutritional products come from the manufacturers’ websites (full details listed in Supplementary Appendix B), while parenteral nutrition composition is as recorded in the dataset or imputed based on average composition.

Missing data and irregular measurement frequencies are a common issue in analysis of EMRs. In order to deal with irregular measurements, we carry forward the last value of each input feature and calculate the elapsed time since that measurement. Both the value of the input feature (possibly carried forward) and the elapsed time are used as input features in model development. For blood glucose, insulin, and nutritional intake features, we additionally carry forward the previous 24 hours’ worth of values as new input features in order to account for potential delays in action and correlations over time (Figure 1).

Models

We develop 2 ML approaches using the Catboost gradient boosting library. These models were chosen as they present alternative approaches to predicting both a point estimate and uncertainty quantification through probabilistic forecasting. The first is a Catboost regression model with dual estimation of the expected outcome and the standard deviation of the prediction distribution, the ‘uncertainty regression’ model. This form of estimation can be performed using the class CatBoostRegressor with the argument loss_function = “RMSEWithUncertainty” in the Python version of Catboost 2.4. The second model is a combination of quantile regressions with models for quantiles of 0.025, 0.5, and 0.975, the “quantile regression” model. This form of estimation can be performed similar to the above, for example for the 0.975 quantile as CatBoostRegressor (loss_function=“Quantile: alpha = 0.95”, ...). For both the quantile and uncertainty model the glucose is modeled as the (natural) logarithm transformed blood glucose, resulting in an approximately normally distributed variable. Calculation of the Box-Cox transformation exponent suggests an optimal transform of $\lambda = 0.3$, with the logarithm (corresponding to $\lambda = 0$) being chosen to simplify back-transformation. The point estimate (mean value) for blood glucose is calculated as $\exp(\hat{\mu} + \hat{\sigma}/2)$ where $\hat{\mu}$ and $\hat{\sigma}$ are the estimates of the mean and standard deviation on the logarithmic scale. The 95% prediction interval is calculated as
CLARA clustering algorithm. This algorithm is designed to scale well when clustering the ICU admission blood glucose trajectories using the Euclidean distance as the measure of similarity between patients. The number of clusters is determined via the elbow method as the point of inflection in the curve describing the decrease within cluster variation with the number of clusters. We report the patient daily average blood glucose level for day 1 and 2 of their ICU stay, diabetic status, and 30-day mortality rate for each cluster.

Model interpretation
For each feature we calculate the associated change in the loss function that would occur with removal of the feature to assess its importance in forecasting blood glucose.

RESULTS
Cohort description
Our dataset includes 14 742 patients. Accounting for readmission, the total number of ICU admissions is 18 691 with an average stay of 3.9 days. Most admissions are nondiabetic (76.7%), and the cohort is evenly split between medical (42%) and surgical (46%) ICU patients, with a small number of cardiac ICU admissions (11%). There are 318 574 glucose measurement events, an average of 4.3 per patient per day with 22.5% of these above 170 mg/dL and 1.3% below 70 mg/dL. Insulin is administered fairly frequently during ICU admissions (37.0% of admissions), with the most common route of administration being bolus injection (31.3% of admissions), followed by IV infusion (19.9% of admissions). Nutritional intake information, in the form of enteral and parenteral nutrition, is present for 18.0% and 3.5% of admissions. Further details on the cohort are available in Table 1 including a comparison of the training and test splits. As with all EMR systems, absence of a record cannot be taken as absence of an event.

Model performance
The overall performance of the models is outlined in Table 2. The 95% prediction intervals for the quantile regression and uncertainty regression models include 94% and 93%, respectively, of the measured postadmission blood glucose values. There was no clear difference between the models in terms of point prediction error (MAPE and RMSE), with the quantile regression model marginally better. Both models perform best in the most frequently observed glycemic range (100–200 mg/dL), followed by the hyperglycemic (≥ 200 mg/dL) and then hypoglycemic ranges (< 70 mg/dL), with the uncertainty regression model achieving MAPEs of 12.4%, 22.5%, and 96.3% for these ranges. Additionally, there is variation in model performance by ICU admission type, with the uncertainty regression model performance highest for surgical (including cardiac surgery) (MAPE: 15.1%), followed by cardiac (MAPE: 17.9%) and medical admissions (MAPE: 19.0%). We show further breakdown of the results below (Figures 1–3) for the uncertainty regression model. These illustrate that the majority (97.1%) of model predictions lead to clinically acceptable decisions (Figure 2) and that the performance of the model varies by observed blood glucose value with better performance in terms of point prediction accuracy and prediction interval coverage for blood glucose values in the range 100–200 mg/dL (Figure 3).

Model interpretation
For each feature we calculate the associated change in the loss function that would occur with removal of the feature to assess its importance in forecasting blood glucose.

Performance by cluster
To assess the performance of the models across patient subgroups, we cluster the ICU admission blood glucose trajectories using the CLARA clustering algorithm. This algorithm is designed to scale on large datasets by clustering on subsamples and then aggregating the results across the whole dataset. The Euclidean distance is chosen as the measure of similarity between patients. The number of clusters is determined via the elbow method as the point of inflection in the curve describing the decrease within cluster variation with the number of clusters. We report the patient daily average blood glucose level for day 1 and 2 of their ICU stay, diabetic status, and 30-day mortality rate for each cluster.

Model interpretation
For each feature we calculate the associated change in the loss function that would occur with removal of the feature to assess its importance in forecasting blood glucose.

Performance by cluster
To assess the performance of the models across patient subgroups, we cluster the ICU admission blood glucose trajectories using the CLARA clustering algorithm. This algorithm is designed to scale on large datasets by clustering on subsamples and then aggregating the results across the whole dataset. The Euclidean distance is chosen as the measure of similarity between patients. The number of clusters is determined via the elbow method as the point of inflection in the curve describing the decrease within cluster variation with the number of clusters. We report the patient daily average blood glucose level for day 1 and 2 of their ICU stay, diabetic status, and 30-day mortality rate for each cluster.
Performance by cluster
Based on the elbow method the blood glucose trajectories were clustered into 3 clusters. Cluster descriptive statistics and performance metrics are displayed in Table 3. Cluster 1 contained the largest percentage of diabetics and had the highest blood glucose values and 30-day mortality rates, with clusters 2 and 3 exhibiting increasingly fewer diabetic admissions and lower mortality and blood glucose values, respectively. Based on the MAPE and MAD, the model performs best for patients in Cluster 2, those patients with the greatest stability in their blood glucose measurements. Examples of observed blood glucose values, forecasts, and the cluster means for 4 patients are shown in Figure 4.

Model interpretation
The variable importance measures for the 30 most important features are shown in Table 4.

DISCUSSION
We report a predictive ML algorithm using routinely collected EMR data that takes as input a patient’s medical history during their ICU stay, and at baseline, and forecasts blood glucose at 2-hour ahead time-steps from ICU entry. The model achieves performance that is comparable to previous models developed in planned research stud-
ies using continuous blood glucose monitoring devices and, like these, performs better for some patient groups and blood glucose values with relatively poor performance for predicting hypoglycemia.

Our proposed ML model demonstrates a high degree of accuracy in forecasting blood glucose in the range 70–200 mg/dL. The overall and surgical patient MAPEs of 16.5% and 15.1% for 2-hour forecasts of serum blood glucose compares with 15.9% (for a 135-min forecast) achieved in a planned research study on surgical ICU patients using continuous glucose monitoring.35 While other research has achieved lower error margins, this was through shorter forecasting timeframes (30–75 min) reducing comparability.36–38 The current approach aimed to have the potential to assist with glycemic control as practiced in the ICUs studied, with the forecast timeframe a lower bound on the frequency of observed blood glucose measurement.3 As assessed using CEGA, the majority of predictions (97.2%) will lead to a clinically appropriate response. The use of prediction intervals is a strength, with quantification of uncertainty important for glycemic prediction due to potentially complex dynamics decreasing point prediction performance35 and the presence of measurement error.33

One strength of data-driven approaches to glucose prediction in the ICU is the ability to incorporate a wide range of predictor variables, as seen in previous research.35–38 This contrasts with mathematical modeling approaches that generally rely on previous blood glucose measures along with insulin and glucose intake information.24 However, as noted in previous research,35 a risk of ML methods is their “black-box” nature and the potential learning of spurious relationships local to the dataset. Indeed, an approach to comprehensive interrogation (from a physiological standpoint) of the data-driven approaches used in the past research is missing from the literature. In the current research, the variable importance measures are unsurprising, with previous measures of blood glucose, dextrose/carbohydrate intake, diabetes, insulin, and patient weight being of high predictive importance. The presence of ICU unit type as a relatively important feature may account for variation in patient characteristics and treatment patterns unaccounted for in the modeled input features.

Our model performs best in cardiac surgical ICU patients. Generally, these patients are seen as the least critically ill51 and have been a focus of several previous studies using data-driven methods to forecast blood glucose in the ICU.35,36,38,52 However, no previous articles have compared predictive performance between medical and surgical patients for large numbers of patients. The greater stability in their metabolic state may ease forecasting through less shift in their blood glucose distribution over time or with treatment (eg, insulin). Blood glucose in metabolically unstable groups, such as patients with diabetic ketoacidosis (MAPE: 26.5%), or emergency compared to elective admissions in general (MAPE: 17.3% vs

![Figure 2](image-url)

**Figure 2.** Clarke error grid analysis using forecast and observed blood glucose values from the test dataset. The percentage of points in each region are A: 71.4%; B: 25.6%; C: 0.2%; D: 2.8%; E: < 0.1, with the regions ranked (A–E) in terms of increasingly dangerous clinical consequences for the prediction error.

![Figure 3](image-url)

**Figure 3.** (A) Prediction interval coverage by measured blood glucose value using forecast and observed blood glucose values from the test dataset. The overall coverage is 93.2%. (B) Mean absolute percentage error by measured blood glucose value. We see a pattern of poor performance in prediction of hypoglycemic and, to a lesser extent, hyperglycemic events.
15.0%), is more difficult to predict. Previous theoretical research using dynamic models has shown that time dependent variations in insulin resistance and feeding patterns common to ICU patients can result in complex glucose responses.32,50,53 This further highlights the importance of quantifying uncertainty.

EMRs are a messy data source not designed for medical research. However, the current research has highlighted their practical value in the development of predictive models. A major strength of the current ML based approach is the easy incorporation of large numbers of input features which can be modeled without manual feature engineering. Further, the model easily accounted for the irregularly recorded and missing data common in EMRs. With up to 60% of blood glucose measurements not observed at every 2-hour time-step, this was of vital importance. Indeed, the model achieved a level of performance comparable to previous planned research studies based on extensive measurement of a small number of patients. This demonstrates that using sparser data from many patients should be a key aspect of future research on glycemic prediction or control algorithms. The current research has limitations. As noted, performance was worse for more extreme values of blood glucose. In order to have clinical utility, it is important that the model can detect hyperglycemia and hypoglycemia. Detection of hyperglycemia was only slightly worse than values in the ICU normal blood glucose range. However, similar to previous research, our point estimates were unable to detect hypoglycemia at 2-hour forecasts.35 However, by forecasting an interval, we increase the potential to flag circumstances in which hypoglycemia is a risk, with 41% of hypoglycemic events captured within the prediction intervals. A postanalysis investigation

| Table 3. Cluster descriptive statistics and model performance metrics for the uncertainty regression model |
|---------------------------------------------------------|-----------------|-------------------|----------------|
| **Cluster descriptive statistics**                       | Cluster 1       | Cluster 2         | Cluster 3      |
| % of admissions                                          | 26.8            | 51.8              | 21.4           |
| Mean (SD) blood glucose                                  |                 |                   |                |
| Day 1                                                   | 200.6 (80.0)    | 131.8 (96.0)      | 104.5 (30.4)   |
| Day 2                                                   | 170.2 (69.5)    | 133.6 (42.5)      | 120.5 (42.5)   |
| Day 3                                                   | 163.5 (64.5)    | 133.6 (43.7)      | 124.4 (43.1)   |
| Diabetic (%)                                             | 44.4            | 22.9              | 11.0           |
| Died within 30 days of ICU stay                          | 17.8            | 15.5              | 9.4            |
| **Model performance metrics**                            |                 |                   |                |
| RMSE (mg/dL)                                            | 47.7            | 58.7              | 30.2           |
| MAPE (%)                                                | 18.6            | 15.4              | 16.7           |
| MAD (mg/dL)                                             | 21.1            | 14.2              | 13.9           |
| Coverage (nominal 95%)                                   | 92.5            | 93.8              | 93.2           |

**Figure 4.** Example glucose and prediction trajectories for 4 patients. The model forecasts blood glucose in 2 hours, and these graphs unroll the predictions over a 48-hour period. (A) is a patient from Cluster 1, (B) and (C) are patients from Cluster 2, (D) is a patient from Cluster 3. The red dashed lines indicate the relevant cluster mean trajectories.
of a 2-hour ahead hypoglycemic prediction model showed poor results with the model unable to do more than predict the base rate of ~1.3% (Supplementary Appendix D1). This difficulty of forecasting hypoglycemic clearly warrants further research and analysis. An issue is ensuring the accuracy of a hypoglycemic reading, with the potential that bedside capillary measurements are not an accurate enough tool in critically ill patients.54 Indeed, previous research has identified poor perfusion as a cause of inaccurate hypoglycemia measurements.55 In the current study patients, the average diastolic blood pressure was 56.8 mmHg during hypoglycemic events compared with an average measure of 60.4 mmHg. Alternatively, oscillation in blood glucose, as seen in theoretical ultradian insulin–glucose dynamic model studied in previous research53 may account for transient hypoglycemia.

It is also important to consider the extent to which these results are dataset specific. As the data are from a single institution, the degree to which institutional factors impact the results cannot be quantified. There exists evidence of variation in glycemic control protocols between institutions26 and future research would be required to investigate the degree of model retraining required to adapt to local protocols. The dataset also constrained the timeframe of predictions. Previous research using continuous blood glucose monitoring devices could be trained on and, hence, predict blood glucose every 5 min for up to 2 hours. Using a sparsely measured EMR to predict over such short time frames would likely be an unjustifiable extrapolation given evidence of nonlinear glucose dynamics.53 Both of these limitations point a way forward for future research. Through pooling EMRs from several institutions, along with data from continuous blood glucose monitoring devices (for potentially a subset of a patients), we may be able to develop robust predictive models.

CONCLUSION

In conclusion, we demonstrate that EMRs can be used to develop blood glucose prediction models that achieve a high degree of accuracy. There is strong potential for developing EMR-based ICU decision support systems for glycemic control and other key ICU physiological parameters. The uncertainty in blood glucose dynamics necessitates the quantification of uncertainty in any forecasts which we demonstrate can be achieved using ML tools. There remain challenges, such as detection of hypoglycemia and assessing the impact of variation of institutional glycemic control policies on model performance. Future research on glycemic control will benefit from greater emphasis on incorporation of real-world data, such as those from EMRs.

FUNDING

This work was supported by the Commonwealth Industrial and Scientific Research Organisation (CSIRO), Health NSW, and the Australian government through an Australian Government Research Training Program scholarship and CSIRO top up scholarship.

AUTHOR CONTRIBUTIONS

OF conceived the idea for the study. OF and OPC conceived the initial design of the study. All authors provided critical methodological feedback and contributed to the design of the study. OF prepared the dataset for the study and performed the analysis. OF and LJ drafted the article. All authors contributed to critical revision of the manuscript.

SUPPLEMENTARY MATERIAL

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

DATA AVAILABILITY STATEMENT

The data underlying this article are available at the MIMIC-III open source database at https://mimic.physionet.org/, and can be accessed following completion of required training as detailed at https://mimic.physionet.org/getting-started/access/.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Vincent JL. The coming era of precision medicine for intensive care. Crit Care 2017; 21 (Suppl 3): 314.
2. Maslove DM, Lamontagne F, Marshall JC, Heyland DK. A path to precision in the ICU. Crit Care 2017; 21 (1): 1–9.
3. Seymour CW, Gomez H, Chang C-CH, et al. Precision medicine for all? Challenges and opportunities for a precision medicine approach to critical illness. Crit Care 2017; 21 (1): 257.
4. Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. Lancet 2018; 392 (10141): 75–87.

Table 4. Feature importance values for the uncertainty regression model

| Feature                                      | Loss function (RMSE reduction) |
|----------------------------------------------|---------------------------------|
| Blood glucose (most recent)                  | 30.74                           |
| Time to blood glucose measurement (most recent) | 6.22                           |
| Time of day (hours)                          | 5.66                           |
| Insulin infusion amount over past 2 hours    | 4.02                           |
| Blood glucose (second most recent)           | 3.21                           |
| Blood glucose (fourth most recent)           | 2.05                           |
| Dextrose input in ongoing IV fluid           | 1.86                           |
| Blood glucose (third most recent)            | 1.39                           |
| Patient weight                               | 1.23                           |
| Cardiac ICU admission                        | 1.14                           |
| Insulin injection (short-acting) amount over past 10 hours | 1.02                           |
| Diabetic patient                             | 1.00                           |
| Insulin injection (short-acting) amount over past 20 hours | 0.91                           |
| Insulin injection (short-acting) amount over 2–4 hours ago | 0.83                           |
| Insulin injection (short-acting) amount over past 4 hours | 0.81                           |
| Insulin injection (short-acting) amount over 4–6 hours ago | 0.81                           |
| Blood glucose (fifth most recent)            | 0.77                           |
| Insulin injection (short-acting) amount over past 4 hours | 0.62                           |
| Volume of ongoing enteral nutrition          | 0.55                           |
| Most recent blood lactate value              | 0.55                           |
| Insulin injection (short-acting) amount over past 8 hours | 0.54                           |
| Insulin injection (short-acting) amount over past 18 hours | 0.48                           |
| Insulin infusion amount over past 2 hours    | 0.48                           |
| Insulin injection (short-acting) amount over past 22 hours | 0.45                           |
| Most recent heart rate measurement           | 0.44                           |
| Time to blood glucose measurement (second most recent) | 0.44                           |
| Most recent creatinine result                | 0.44                           |
| Most recent systolic blood pressure          | 0.43                           |
| Sugar content in IV fluid over past 4 hours  | 0.41                           |
| Most recent sodium result                    | 0.37                           |
