AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

Machine learning for initial insulin estimation in hospitalized patients

Minh Nguyen (ID)[1,*], **Ivana Jankovic**[2,*], **Laurynas Kalesinskas**[1], **Michael Baiocchi**[3], **and Jonathan H. Chen**[4]

[1]Department of Biomedical Data Science, Stanford University, School of Medicine, Stanford, California, USA, [2]Division of Endocrinology, Department of Medicine, Stanford University, School of Medicine, Stanford, California, USA, [3]Department of Epidemiology and Population Health, Stanford University, Stanford, California, USA, and [4]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, California, USA

*Co-first authors.

Corresponding Author: Jonathan H. Chen, MD, PhD, Stanford University, School of Medicine, 1265 Welch Rd, Medical School Office Building X213, Stanford, CA 94305, USA; jonc101@stanford.edu

Received 11 January 2021; Revised 12 March 2021; Editorial Decision 5 May 2021; Accepted 7 May 2021

ABSTRACT

Objective: The study sought to determine whether machine learning can predict initial inpatient total daily dose (TDD) of insulin from electronic health records more accurately than existing guideline-based dosing recommendations.

Materials and Methods: Using electronic health records from a tertiary academic center between 2008 and 2020 of 16,848 inpatients receiving subcutaneous insulin who achieved target blood glucose control of 100-180 mg/dL on a calendar day, we trained an ensemble machine learning algorithm consisting of regularized regression, random forest, and gradient boosted tree models for 2-stage TDD prediction. We evaluated the ability to predict patients requiring more than 6 units TDD and their point-value TDDs to achieve target glucose control.

Results: The method achieves an area under the receiver-operating characteristic curve of 0.85 (95% confidence interval [CI], 0.84-0.87) and area under the precision-recall curve of 0.65 (95% CI, 0.64-0.67) for classifying patients who require more than 6 units TDD. For patients requiring more than 6 units TDD, the mean absolute percent error in dose prediction based on standard clinical calculators using patient weight is in the range of 136%-329%, while the regression model based on weight improves to 60% (95% CI, 57%-63%), and the full ensemble model further improves to 51% (95% CI, 48%-54%).

Discussion: Owing to the narrow therapeutic window and wide individual variability, insulin dosing requires adaptive and predictive approaches that can be supported through data-driven analytic tools.

Conclusions: Machine learning approaches based on readily available electronic medical records can discriminate which inpatients will require more than 6 units TDD and estimate individual doses more accurately than standard guidelines and practices.

Key words: machine learning, diabetes mellitus, insulin, medical informatics, clinical decision support

INTRODUCTION

Background and Significance

Poorly controlled glucose is both common and dangerous in hospitalized patients, reflecting deficiencies in common standard practices in insulin dosing. Hyperglycemia, defined as a blood glucose >140 mg/dL, occurs in 22% to 46% of non–critically ill hospitalized patients,[1] and can lead to serious complications, including infections, cardiovascular events, and increased overall mortality.[2] The increased odds of mortality among patients with a blood glucose above 145 mg/dL is 1.3 to 3 times that of patients with normal glucose (70-110 mg/dL), independent of illness severity.[3]

The treatment for hyperglycemia in inpatients is insulin, a hormone essential to enabling cells to uptake glucose from the blood for energy. However, insulin has a narrow therapeutic window when given as a medication, and overtreatment can lead to dangerous hypoglycemia causing seizure, arrhythmia, or even death. As such, predicting an accurate insulin dose is critical for clinical outcomes. The existing standard of care for estimating initial insulin dose prediction is unfortunately highly variable, as it is typically driven mainly by individual clinical judgment supplementing crude weight-based clinical calculators, often leading to ineffective glucose control.[4]

Practice guidelines for inpatient insulin dosing primarily revolve around weight-based clinical calculators that estimate the total daily dose (TDD) of insulin required to be 0.4 to 0.6 units/kg among non-elderly patients with good kidney function.[5,6] This calculator results in a range of dosing that can vary by 50%, requiring prescribers to use variable clinical experience to adjust for factors such as age, suspected insulin sensitivity, and renal function. Even in optimal conditions, existing TDD guidelines are based on a dosing schema of unclear provenance chosen in published studies[7] that have not been clinically validated. The current practice leads to significant dosing heterogeneity even within the same patient's hospitalization.[8]

For admitted patients, prescribing an initial insulin dose is often challenging, as there is limited information available at this early stage, whereas titration is often a simpler problem because a patient's insulin sensitivity can be estimated from their response to previous insulin doses. Specialized glucose management services have been established to help minimize both hyperglycemia and hypoglycemia in the inpatient setting. These consult services have improved blood glucose control and cost savings.[9,10] However, the number of patients who need consults often exceeds the capacity of these services.[11] Decision support that assists in insulin dosing could improve inpatient glycemic control at scale.

An alternate approach to formal consult services is a remote glucose monitoring service in which a consulting endocrinologist provides teams with insulin dosing suggestions based on chart review, without examining the patient, which has shown success in reducing the proportion of patients with hyperglycemia and reducing hypoglycemic events.[12] The success of such remote glucose control programs suggests that electronic health records may contain sufficient information for prescribing insulin, and can be leveraged using automated machine learning methods.

Prior research in the use of machine learning for diabetes-related problems has mostly focused on detecting adverse glycemic events and predicting blood glucose and insulin bolus doses using continuous glucose monitoring measurements in outpatients and has been limited to short-term predictions under 60 minutes.[13–23] Studies predicting insulin bolus doses have focused on titration, adjusting previous insulin doses and relying on manual physician calculation.[13,20–23] Although they showed promising results for outpatient type 1 dia-

betes management, these studies used either simulated data or evaluation metrics focused on glycemic control and not direct assessment of predicted insulin doses from patient data. These algorithms may not apply to hospitalized patients who are more clinically unstable than outpatients, and only have noncontinuous blood glucose checks typically no more than 4 times per day. To our knowledge, there have been no prior studies predicting actual insulin doses in inpatients, although one study predicting what dose of insulin clinicians would order yielded an error of approximately 73%.[24]

Common machine learning methods used in prior diabetes-related research include multivariate regression, support vector regression, and deep learning,[15,25–30] though no method has been consistently shown to be superior.[13] Instead of choosing a single algorithm, an ensemble machine learning approach, such as the SuperLearner that we apply here, uses a weighted combination of multiple learning algorithms to achieve better predictive performance than any single algorithm alone.[31]

OBJECTIVE

Our objective is to determine whether initial inpatient insulin requirements could be more accurately predicted from readily available electronic health record data using machine learning methods than existing weight-based guidelines. In stage I, we predicted whether a patient will require more than 6 units of TDD, ie, "low" vs "higher" insulin users, as a binary prediction. In stage II, for patients who require more than 6 units of TDD, we predicted the point-value TDD that the patients required to achieve good glucose control.

MATERIALS AND METHODS

We present our results following the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement.[32] Our study is approved by the institutional review board of the Stanford University School of Medicine.

Data source and cohort

Using electronic health record data from a tertiary academic medical center from 2008 to 2020, we retrospectively identified a cohort of unique patients who achieved target "good" glucose control during their most recent hospital encounters. Patients were considered to have good control if they had at least 3 blood glucose measurements by glucometer that were within the target range of 100 to 180 mg/dL[6,33] on a calendar day without any measurement outside this range, consistent with inpatient diabetes management guidelines[34] We excluded patients who were on total parenteral nutrition (TPN) or peripheral parenteral nutrition (PPN), tube feeds, insulin pumps, insulin infusions, or any rarely used insulin formulations (ordered fewer than 25 times in all records). Because insulin dosing is traditionally weight based, we also excluded patients with missing weights, about 2.4% of our original cohort. If patients had more than 1 good day, we selected the first good day of their most recent hospitalization.

Features

Included features were weight, height, age, sex, race, insurance status (public vs private), creatinine, diet (nothing by mouth, carb-controlled, other), counts of microbiology lab orders, and amount of glucocorticoid use within the previous 48 hours. Hemoglobin

Table 1. Summary of demographics and some important variables in the full cohort

	Mean	SD	Count	Proportion
Age, y	63.8	14.4		
Sex				
Female			7497	44.5%
Male			9351	55.5%
Weight, kg	84.1	24.0		
Height, cm	168.3	11.1		
Race				
Asian			2479	14.7%
Black			869	5.2%
Native American			71	0.4%
Pacific Islander			374	2.2%
White			9008	53.5%
Other			3562	21.1%
Unknown			485	2.9%
Insurance				
Public			10 129	60.2%
Private			6709	39.8%
Diet				
NPO			3323	19.7%
Carb controlled			3700	22.0%
Other			9825	58.3%
HbA1c, %	6.57	1.49		
Creatinine, mg/dL	1.40	1.42		
First glucose, mg/dL	148	62		
History of basal insulin use				
No			13 984	83.0%
Yes			2864	17.0%

HbA1c: hemoglobin A1c; NPO: nothing by mouth.

A1c was classified into 4 categories: missing, <5.7, between 5.7 and 9, and >9 as normal, high, and panic high defined by our reference clinical laboratory. The total amount of glucocorticoid was normalized to glucocorticoid equivalents.[35–37] Counts of major International Classification of Diseases codes by most general parent category level were also included.[38] For example, the International Classification of Diseases code E11.9, type 2 diabetes without complications, would be counted as simply category E: diagnoses for endocrine, nutritional, and metabolic diseases. Table 1 displays a summary of demographics and selected key characteristics. Supplementary Table 1 displays a list of included features.

Additionally, we included counts of relevant lab results in quantiles to handle sparsity. We quantized lab value distributions into decile bins, assigned values to the bins, and then counted the frequencies of bin membership. This method naturally deals with missingness by yielding count vectors of zeros over all bins if a particular numerical lab is not available. The included labs were albumin, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, anion gap, total protein, total bilirubin, troponin, blood urea nitrogen, calcium, potassium, sodium, lactate, blood gas, hemoglobin, white blood cell count, platelet count, eosinophil, and absolute neutrophil count. For blood glucose, we used the exact value of the first available (admission) measurement and summary statistics of all measurements (mean, median, minimum, maximum, SD, and total number of measurements) prior to prediction time. Although we included patient history of basal insulin use as a feature, no insulin doses the patients received during the present admission were included to reflect the aim of predicting initial insulin dosing (as opposed to titration). We restricted the feature set to only include data
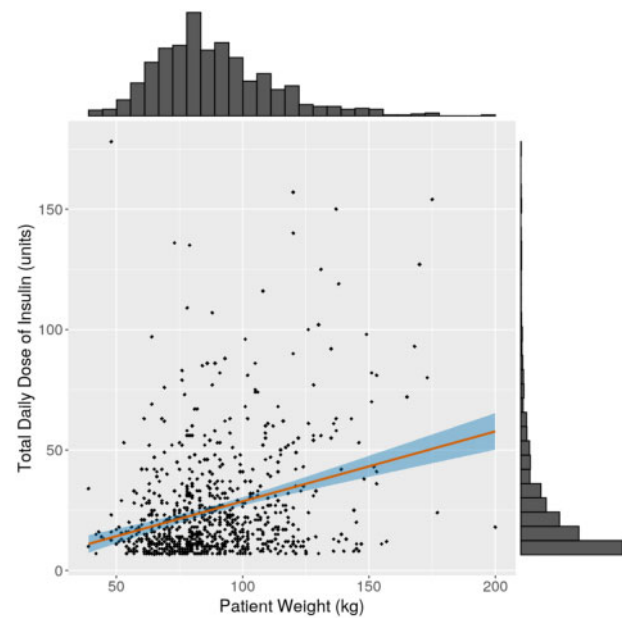
Figure 1. Plot of weight vs total daily dose with regression line and its confidence interval.

available before the days that patients achieved good glucose control (prediction time).

Outcomes and 2-stage prediction framework

In the first stage of modeling, we undertook a binary prediction of whether a patient received more than 6 units of TDD ("higher" insulin users, positive class) or ≤6 units ("low" insulin users, negative class). We chose this threshold based on our data distribution, as about 75% of our cohort required a TDD of 6 units or below. Additionally, 6 units is the minimum dose that could be split into basal and prandial insulin without an insulin pump. Therefore, maintaining glycemic control for "low" insulin users could reasonably require little more than monitoring and sliding scale insulin, whereas "higher" insulin users may require a basal-bolus regimen. Our baseline model is a univariate logistic regression model with weight as the predictor.

In the second stage of modeling, we aimed to predict the point-value TDD in "higher" users. Among the "higher" insulin users, the range of insulin needs varied broadly (Figure 1). We log transformed TDD for modeling purposes, given the right skew of the outcome variable. Our primary baseline model was a univariate regression model with weight, approximating existing guidelines using weight as the predictor. A secondary baseline was the estimated TDD using a clinical calculator often considered in clinical practice: $TDD = c * patient\text{-}weight$ (kg), where c is a factor from 0.4 to 0.6.[33]

Statistical methods

We randomly split our data by patients into 80% for training and 20% for testing. For both of the prediction stages, we chose 3 supervised machine learning algorithms, including regularized regression, random forest, and gradient boosted trees. We trained an ensemble SuperLearner algorithm, a more generalized "regression" procedure, to integrate these base algorithms to achieve higher performance and stability, in which each base algorithm contributed a weight to the ensemble algorithm.[31] We used 10-fold cross-validation for training with SuperLearner and tested on the independent test set.

To evaluate our models for the 2 stages on the test set, we used several different metrics. For the stage I binary prediction, we used the area under the receiver-operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). The receiver operating characteristics curve summarizes the trade-off between true positive rate (sensitivity) and true negative rate (specificity) for a predictive model using different probability thresholds. However, using only AUROC can be misleading, especially in highly imbalanced data.[39] In such cases, the AUPRC offers additional information. The precision-recall curve distinguishes the trade-off between the true positive rate (recall or sensitivity) and the positive predictive value (precision). The positive predictive value (precision) is a more informative measurement than the true negative rate (specificity) as it is not overshadowed by the large number of true negatives (low insulin users) and is much more sensitive to true positives. Higher values for AUROC and AUPRC indicate better model performance. An AUROC of 0.5 indicates a random classifier, whereas an AUPRC lowest value equals the fraction of positives.[39]

For the stage II point-value TDD prediction in "higher" insulin users, we used the mean absolute error (MAE) and mean absolute percent error (MAPE) as the 2 evaluation metrics for interpretability. MAE reflects the average magnitude of error by comparing the predicted vs observed values. MAPE is the magnitude of error normalized over the observed values, reflecting how far the predictions are off from the truth as a percentage. MAPE is similar in concept to the mean absolute relative difference used to evaluate blood glucose readings for continuous glucose monitors[40]; in both cases, errors at low values are more clinically significant than the same absolute error at a high value. The lower the MAE and MAPE are, the better the predictions are. Using bootstrap sampling, we obtained 95% confidence intervals (CIs) to better compare different results within each prediction stage. We also did a manual chart review of a sample of high error cases.

RESULTS

The final cohort had 16 868 unique patients and 87 features. Additionally, we created a subset of this cohort, excluding patients who were in intensive care units (ICUs), yielding a non–critically ill cohort of 13 037 unique patients.

Stage I: Predicting low insulin need—binary prediction

For the baseline univariate logistic regression weight-based model, the AUROC and AUPRC estimates were 0.57 (95% CI, 0.55-0.60) and 0.29 (95% CI, 0.22-0.35), respectively. Our SuperLearner algorithm achieved an AUROC of 0.85 (95% CI, 0.84-0.87) and an AUPRC of 0.65 (95% CI, 0.64-0.67). All base algorithms contributed roughly equal weights to the ensemble SuperLearner model. As it is clinically more dangerous to classify low insulin users (negative class) as high insulin users (positive class), which may lead to hypoglycemia, a higher true negative rate (specificity) is more desirable. The higher the specificity is, the lower the sensitivity (recall) is and the higher the positive predictive value (precision) is. Choosing a relatively high and conservative prediction probability threshold of 0.4 yielded 90% specificity with 56% sensitivity (recall) and a positive predictive value of 64% that a patient will fall in the "higher" TDD group. Figure 2 shows the stage I binary prediction's calibration plot of observed probability vs predicted probability of a patient requiring more than 6 units of insulin.
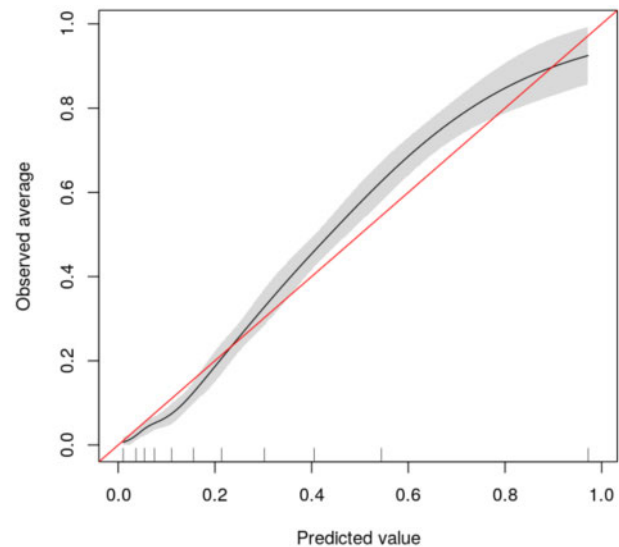

Figure 2. Calibration plot for binary prediction of "low" vs "higher" insulin users.

Stage II: Point-value TDD prediction

After the initial stage I binary prediction, we followed with a stage II point-value TDD prediction for "higher" insulin users who received more than 6 units of TDD. The prediction task was challenging because the range of TDD was wide, the distribution of TDD was still heavily right skewed, and the sample size was reduced to 3970 patients. On average, our data showed that it took about 2.2 days (SD of 4.4 days) (Supplementary Figure 1) from admission for physicians to titrate insulin to achieve good control, ie, blood glucose in the range of 100 to 180 mg/dL.

Using the standard clinical TDD formula with $c = (0.4\text{-}0.6)$, the MAE was 25.4 units (19.5-32.2 units) and the MAPE was 186% (136%-329%). Here, 25.4 units and 186% were the errors for $c = 0.5$. The lower bounds (19.5 units or 136%) and upper bounds (32.2 units and 329%) correspond to $c = 0.4$ and $c = 0.6$. A larger factor c yielded larger errors. For the baseline univariate linear regression weight-based model, the estimates for MAE and MAPE were 14 units (95% CI, 12.8-14.5) and 60% (95% CI, 57%-63%), respectively. This baseline model yielded an intercept of 10.5 units, and for every 1-kg increase in weight, TDD increases by about 0.66% (Figure 1).

Our ensemble model with full features improved the regression with MAE of 12.2 units (95% CI, 11.0-13.2 units) and MAPE of 51% (95% CI, 48%-54%). The best base algorithm was lasso regression, which contributed about 49% of the weight to the ensemble model. Random forest and gradient boosted trees each contributed about 35% and 16%, respectively. The lasso regression alone had an MAE of 12.3 units (95% CI, 11.5-13.2 units) and MAPE of 53% (95% CI, 51%-57%), comparable to the ensemble model's result but with substantially less computational complexity. Table 2 summarizes the results from both prediction stages.

Figure 3 visualizes the observed values of TDD vs the predicted values using 3 different approaches: TDD calculators, weight only, and full feature machine learning models.

Non-ICU cohort

A manual review of labs for patients with high prediction errors suggested that the acuity levels of these patients were high. We

Table 2. Results from 2-stage predictions

Stage I: Binary prediction for "low" vs "higher" insulin users	AUROC (95% CI)	AUPRC (95% CI)
Ensemble model with full features	0.85 (0.84-0.87)	0.65 (0.64-0.67)
Logistic regression with weight only	0.57 (0.55-0.60)	0.29 (0.22-0.35)
Stage II: Point-value TDD prediction among "higher" insulin users	**MAE (95% CI)**	**MAPE (95% CI)**
Ensemble model with full features	12 (11.0 -13.2)	51% (48%-54%)
Regression model with weight only	14 (12.8-14.5)	60% (57%-63%)
TDD = 0.4 * patient-weight (kg)	20	136%
TDD = 0.5 * patient-weight (kg)	25	186%
TDD = 0.6 * patient-weight (kg)	32	329%

Performance metrics for stage I and stage II predictions. Full feature ensemble models included other features besides patient weight as described in the Materials and Methods. For the MAE and MAPE, results were compared with the estimated TDD using clinical calculators defined by $c*patient\text{-}weight$, where c is a constant of 0.4, 0.5, or 0.6. CIs were not included due to the deterministic nature of the calculation.

AUPRC: area under the precision-recall curve; AUROC: area under the receiver-operating characteristic curve; CI: confidence interval; MAE: mean absolute error; MAPE: mean absolute percent error; TDD: total daily dose.
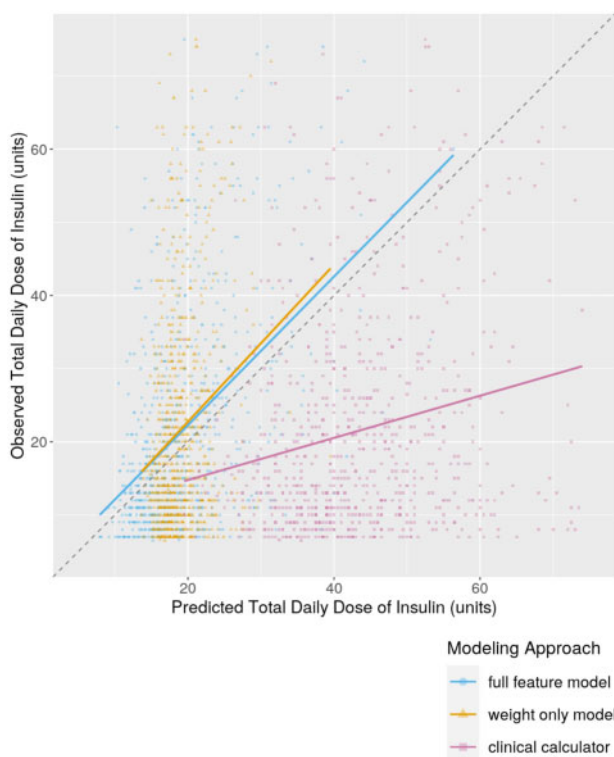

Figure 3. Plot of observed vs predicted total daily dose for all 3 modeling approaches.

hypothesized that repeating the predictions excluding critically ill patients would decrease error. Excluding ICU patients from our original cohort, we performed the same predictions to compare the results with the full feature SuperLearner model. The stage I binary prediction results improved slightly, with an AUROC of 0.86 (95% CI, 0.85-0.88) and an AUPRC of 0.69 (95% CI, 0.67-0.71). Unexpectedly, for stage II point-value TDD prediction, the performance decreased. The MAE was 13.0 units (95% CI, 11.6-14.5 units) and the MAPE was 57% (95% CI, 51%-59%).

Variable importance

For variable importance, we looked at 2 measures produced by the random forest model (Supplementary Figure 2). The first was per-cent increase in node purity (%IncNodePurity), which measures the increase in tree node homogeneity that results from splits of a given variable, averaged over all the trees. A node is purer if there are fewer splits. The second metric was the percent increase in mean squared error (%IncMSE), which is calculated when a variable is not included in the model. This metric is considered more robust and informative because it uses out-of-bag samples in which values of a variable are randomly permuted to compute prediction accuracy.[41] The most important variable in both metrics was patient weight (kg). Other important variables were the following: summary statistics of all glucose measurements by glucometers, the first available blood glucose, counts of serum glucose measurements in deciles, diet, hemoglobin A1c category, creatine, history of basal insulin use, and the count of historical diagnosis codes E (endocrine related) and Z (factors influencing health status).

DISCUSSION

Insulin has a very narrow therapeutic range,[42] and the consequences of overprescribing are more immediately critical than underprescribing. Though inadequate glycemic control is common in hospitalized patients, we found that the amount of insulin required to achieve good control was heavily skewed toward small doses. Thus, in stage I we first predicted TDD insulin usage at a threshold of 6 units. A benchmark weight-based-only classifier was not much better than random (AUROC 0.57, AURPC 0.29), but when using the full set of variables readily available in the electronic medical record data, we demonstrate that a machine learning classification approach can perform substantially better (AUROC 0.85, AUPRC 0.65). These findings demonstrate the ability to discriminate between "low" insulin users vs "higher" insulin users. Our stage I binary prediction could be clinically useful when initiating insulin dose for hospitalized patients to immediately risk-stratify patients based on how likely they are to need significant doses of insulin at all, potentially shortening the insulin titration time to achieve good glucose control while decreasing hypoglycemia risk.

Precise point-value TDD prediction is more relevant for the "higher" insulin users. However, it is more challenging than the binary prediction, as there is a wide variability of patient responses to insulin. Additionally, there is likely a range of TDDs with which patients' blood glucoses would stay in the range of good control, though this is not knowable from retrospective data. For this task, the generalized linear regression model did best compare to random forest and gradient boosted trees. It is worth noting that the point-

value TDD estimates from the standard clinical calculator performed much worse than both univariate and multivariate machine learning prediction models (Figure 3). In clinical practice, the weight-based calculator is only a starting point to anchor clinical decision making; our stage II point-value TDD predictor provides a more accurate anchoring value. Additionally, from the univariate weight-only regression model coefficients, we can interpret that for patients who are classified as "higher" insulin users, their baseline insulin dose is about 10 units, and an increase in 1 kg of patient weight is associated with 0.66% increase in predicted TDD. Although the weight-only regression model performed well, the full feature model showed improved performance for TDD above approximately 30 units (Figure 3). Patient weight is the most important predictor of TDD, more so than any other features, though common TDD calculators do not accurately capture its contribution.

Limitations of this study include that the TDD values were heavily skewed so there was more heterogeneity and fewer data in the higher range of prescribed insulin. Because we aimed to predict the initial insulin dose, as this is the more pressing clinical need than titration, we intentionally did not include any titration insulin doses and their corresponding exact glucose values prior to prediction time, limiting informative data for prediction. Moreover, erroneous, missing, and unavailable data are common in electronic health record data. For example, details of patient diet intake and home doses of diabetes medications, including insulin, are important for TDD prediction but are not reliably available. We used features that are most commonly available and applied quantization for labs to handle major missingness. Even given extensive electronic health record data and robust machine learning methods, we found patient insulin requirements to be highly variable, indicating that predicted dosing recommendations would still need to provide a range of values for a clinician to consider. Finally, this is a retrospective observational study, and the algorithms were trained on a cohort of patients who by definition were able to achieve good blood glucose control.

These approaches illustrate the capability to substantially estimate better insulin dosing than existing standards of care. It is important to note that clinicians struggle with many of the same missing data and patient variability issues in practice as well, and here we offer a tool that will provide clinicians with a more accurate estimate upon which to base their clinical decision making. We envision this work to be the foundation of an electronic health record–integrated decision support tool at the time of initial subcutaneous insulin ordering in patients with hyperglycemia like our cohort by inclusion and exclusion criteria. First, the binary prediction could be used to identify patients who could likely be monitored on sliding scale insulin alone. For patients who are predicted to require "higher" insulin doses, the second stage of prediction could suggest an anchoring daily dose upon which to apply clinical reasoning, similar to but more accurate than the current function of the clinical calculator. Among the predicted "higher" insulin users, there are potentially several visions for implementation, including targeting patients for endocrinology consults or contrasting the predicted dose with the estimated dose from current practice. With time and further validation, clinicians may feel more confident in prescribing insulin algorithmically while avoiding hypoglycemia, thus achieving glycemic control faster and potentiating the benefits to patients and the healthcare system.[43]

Our modeling framework will need prospective validation before implementation to address covariate shifts and temporal changes. Specifically, quantization of lab values will need to be recomputed on an updated training dataset. A new test set that includes more recent data could be necessary to assess model performance and the decaying relevance of clinical data over time. Furthermore, although we aimed to predict initial insulin dosing, more clinically meaningful prospective evaluation metrics such as inpatient hypoglycemia and hyperglycemia, or time to good control, could be used postimplementation. It is also important to check the calibration of our model based on race and evaluate performance differences among racial groups if deployed in populations with significantly different demographics. Future prospective studies are necessary to assess how well the algorithms may generalize to all inpatients requiring insulin, as these populations may differ from the retrospective cohort used here.

CONCLUSION

A machine learning approach can predict higher vs low insulin requirements among hospitalized patients with better discrimination than standard weight-based methods. Prediction of initial daily insulin dosing with an ensemble learning method is more accurate compared with current practice recommendations. Challenges remain due to the wide variability of patient response and narrow therapeutic window of insulin, but more accurate initial point-value TDD estimation can provide an improved dosing anchor for clinical decision making to improve inpatient glucose control.

AUTHOR CONTRIBUTIONS

IJ and JHC conceived the study. IJ, LK, and MN queried the data. MN implemented the algorithm, performed statistical analyses, and drafted the initial manuscript. All authors contributed to the study and analysis design, critically revising and reviewing the final manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

The electronic health records data underlying this article was provided by STARR (STAnford medicine Research data Repository) (https://med.stanford.

edu/starr-tools.html). The data can be accessed for research purposes after Institutional Review Board approval via the Stanford Research Informatics Center.

CONFLICT OF INTEREST STATEMENT

REFERENCES

1. Umpierrez GE, Hellman R, Korytkowski MT, *et al.*; Endocrine Society. Management of hyperglycemia in hospitalized patients in non-critical care setting: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab* 2012; 97 (1): 16–38.
2. Umpierrez GE, Pasquel FJ. Management of inpatient hyperglycemia and diabetes in older adults. *Diabetes Care* 2017; 40 (4): 509–17.
3. Falciglia M, Freyberg RW, Almenoff PL, *et al.* Hyperglycemia-related mortality in critically ill patients varies with admission diagnosis. *Crit Care Med* 2009; 37 (12): 3001–9.
4. Jankovic I, Chen J. 1235-P: Identifying trends in the management of inpatient diabetes at a University Teaching Hospital, 2008-2018. *Diabetes* 2020; 69 (Supplement 1). doi:10.2337/db20-: 1235-P.
5. American Diabetes Association. 15. Diabetes care in the hospital: standards of medical care in diabetes—2020. *Diabetes Care* 2020; 43 (Supplement 1): S193–S202.
6. Inzucchi S. Management of diabetes mellitus in hospitalized patients. UpToDate. https://www.uptodate.com/contents/management-of-diabetes-mellitus-in-hospitalized-patients?search=Management%20of%20Diabetes%20Mellitus%20in%20Hospitalized%20Patients&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1. Accessed 3 December 2020.
7. Umpierrez GE, Smiley D, Zisman A, *et al.* Randomized study of basal-bolus insulin therapy in the inpatient management of patients with type 2 diabetes (RABBIT 2 Trial). *Diabetes Care* 2007; 30 (9): 2181–6.
8. Kodner C, Anderson L, Pohlgeers K. Glucose management in hospitalized patients. *Am Fam Physician* 2017; 96 (10): 648–54.
9. Mandel SR, Langan S, Mathioudakis NN, *et al.* Retrospective study of inpatient diabetes management service, length of stay and 30-day readmission rate of patients with diabetes at a community hospital. *J Community Hosp Intern Med Perspect* 2019; 9 (2): 64–73.
10. Pietras SM, Hanrahan P, Arnold LM, *et al.* State-of-the-art inpatient diabetes care: the evolution of an academic hospital. *Endocr Pract* 2010; 16 (3): 512–21.
11. Ross AJ, Anderson JE, Kodate N, Thompson K, *et al.* Inpatient diabetes care: complexity, resilience and quality of care. *Cogn Tech Work* 2014; 16 (1): 91–102.
12. Rushakoff RJ, Sullivan MM, MacMaster HW, *et al.* Association between a virtual glucose management service and glycemic control in hospitalized adult patients: an observational study. *Ann Intern Med* 2017; 166 (9): 621–7.
13. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res* 2018; 20 (5): e10775.
14. Vu L, Kefayati S, Idé T, *et al.* Predicting nocturnal hypoglycemia from continuous glucose monitoring data with extended prediction horizon. *AMIA Annu Symp Proc* 2019; 2019: 874–82.
15. Zhu T, Li K, Chen J, *et al.* Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J Healthc Inform Res* 2020; 4 (3): 308–24.
16. Albers DJ, Levine M, Gluckman B, *et al.* Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS Comput Biol* 2017; 13 (4): e1005232.
17. Pesl P, Herrero P, Reddy M, *et al.* An advanced bolus calculator for type 1 diabetes: system architecture and usability results. *IEEE J Biomed Health Inform* 2016; 20 (1): 11–7.
18. Sangi M, Win KT, Shirvani F, *et al.* Applying a novel combination of techniques to develop a predictive model for diabetes complications. *PloS One* 2015; 10 (4): e0121569.
19. Dagliati A, Malovini A, Decata P, *et al.* Hierarchical Bayesian logistic Regression to forecast metabolic control in type 2 DM patients. *AMIA Annu Symp Proc* 2016; 2016: 470–9.
20. Cappon G, Marturano F, Vettoretti M, *et al.* In silico assessment of literature insulin bolus calculation methods accounting for glucose rate of change. *J Diabetes Sci Technol* 2019; 13 (1): 103–10.
21. Cappon G, Vettoretti M, Marturano F, *et al.* A neural-network-based approach to personalize insulin bolus calculation using continuous glucose monitoring. *J Diabetes Sci Technol* 2018; 12 (2): 265–72.
22. Noaro G, Cappon G, Vettoretti M, *et al.* Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy. *IEEE Trans Biomed Eng* 2021; 68 (1): 247–55.
23. Guzman Gómez GE, Burbano Agredo LE, Martínez V, *et al.* Application of artificial intelligence techniques for the estimation of basal insulin in patients with type I diabetes. *Int J Endocrinol.* 2020; 2020: 7326073. doi:https://doi.org/10.1155/2020/7326073.
24. Liu X, Jankovic I, Chen JH. Predicting inpatient glucose levels and insulin dosing by machine learning on electronic health records. *medRxiv*, doi: 10.1101/2020.03.02.20029017, 5 Mar 2020, preprint: not peer reviewed.
25. Li K, Daniels J, Liu C, *et al.* Convolutional recurrent neural networks for glucose prediction. *IEEE J Biomed Health Inform* 2020; 24 (2): 603–13.
26. Sun Q, Jankovic MV, Bally L, *et al.* Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network. arXiv, doi: https://arxiv.org/abs/1809.03817, 11 Sep 2018, preprint: not peer reviewed.
27. Zhu T, Li K, Herrero P, *et al.* A deep learning algorithm for personalized blood glucose prediction. *In: Proceedings of the 27th International Joint Conference on Artificial Intelligence*; 2018: 74–8; Stockholm, Schweden.
28. Oviedo S, Vehí J, Calm R, *et al.* A review of personalized blood glucose prediction strategies for T1DM patients. *Int J Numer Method Biomed Eng* 2017; 33 (6): e2833–54. doi:10.1002/cnm.2833.
29. Mhaskar HN, Pereverzyev SV, van der Walt MD. A deep learning approach to diabetic blood glucose prediction. *Front Appl Math Stat* 2017; 3: 14. doi:10.3389/fams.2017.00014
30. Li K, Liu C, Zhu T, *et al.* GluNet: a deep learning framework for accurate glucose forecasting. *IEEE J Biomed Health Inform* 2020; 24 (2): 414–23.
31. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; 6: Article25.
32. Heus P, Damen JAAG, Pajouheshnia R, *et al.* Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open* 2019; 9 (4): e025611.
33. Inzucchi SE. *Diabetes Facts and Guidelines.* New Haven, CT: Yale Diabetes Center; 2011. https://medicine.yale.edu/intmed/drc/diabetescenter/living/50135_Yale%20National%20F_102165_284_13584_v1.pdf. Accessed November 26, 2020.
34. Sawin G, Shaughnessy AF. Glucose control in hospitalized patients. *Am Fam Physician* 2010; 81 (9): 1121–4.
35. Meikle AW, Tyler FH. Potency and duration of action of glucocorticoids. Effects of hydrocortisone, prednisone and dexamethasone on human pituitary-adrenal function. *Am J Med* 1977; 63 (2): 200–7.
36. Singer M, Webb A. *Oxford Handbook of Critical Care.* London: Oxford University Press https://oxfordmedicine.com/view/10.1093/med/9780199235339.001.0001/med-9780199235339. Accessed December 17, 2020.
37. Czock D, Keller F, Rasche FM, *et al.* Pharmacokinetics and pharmacodynamics of systemically administered glucocorticoids. *Clin Pharmacokinet* 2005; 44 (1): 61–98.
38. *ICD-10 Version. 2019.* https://icd.who.int/browse10/2019/En. Accessed November 26, 2020.
39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10 (3): e0118432.

40. Reiterer F, Polterauer P, Schoemaker M, *et al.* Significance and reliability of MARD for the accuracy of CGM systems. *J Diabetes Sci Technol* 2017; 11 (1): 59–67.

41. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2009: 593–4.

42. High-alert' medications and patient safety. *Int J Qual Health Care* 2001; 13 (4): 339–40.

43. Haque WZ, Demidowich AP, Sidhaye A, *et al.* The financial impact of an inpatient diabetes management service. *Curr Diab Rep* 2021; 21 (2): 1–9. doi:10.1007/s11892-020-01374-0