The challenge of predicting blood glucose concentration changes in patients with type I diabetes

Neil C Borle
University of Alberta, Canada

Edmond A Ryan
University of Alberta, Canada
Alberta Diabetes Institute, Canada

Russell Greiner
University of Alberta, Canada
Alberta Machine Intelligence Institute, Canada

Abstract
Patients with Type I Diabetes (T1D) must take insulin injections to prevent the serious long term effects of hyperglycemia. They must also be careful not to inject too much insulin because this could induce (potentially fatal) hypoglycemia. Patients therefore follow a "regimen" that determines how much insulin to inject at each time, based on various measurements. We can produce an effective regimen if we can accurately predict a patient's future blood glucose (BG) values from his/her current features. This study explores the challenges of predicting future BG by applying a number of machine learning algorithms, as well as various data preprocessing variations (corresponding to 312 \{learner, preprocessed-dataset\} combinations), to a new T1D dataset that contains 29,601 entries from 47 different patients. Our most accurate predictor, a weighted ensemble of two Gaussian Process Regression models, achieved a (cross-validation) $\text{err}_1$ loss of 2.7 mmol/L (48.65 mg/dl). This result was unexpectedly poor given that one can obtain an $\text{err}_1$ of 2.9 mmol/L (52.43 mg/dl) using the naive approach of simply predicting the patient's average BG. These results suggest that the diabetes diary data that is typically collected may be insufficient to produce accurate BG prediction models; additional data may be necessary to build accurate BG prediction models over hours.

Corresponding author:
Russell Greiner, Department of Computing Science, Athabasca Hall, University of Alberta, Edmonton, AB T6G 2E8, Canada.
Email: rgreiner@ualberta.ca

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Keywords
type 1 diabetes, machine learning, blood glucose prediction

Introduction
Individuals suffering from Type I diabetes (T1D) are unable to produce insulin, meaning their bodies cannot properly regulate their blood glucose (BG)\(^1\) – that is, cannot maintain their BG between 4 and 8 mmol/L.\(^2\) As a result, T1D is a serious long term condition that can lead to microvascular, macrovascular, neurological and metabolic complications.\(^1,2\)

To manage their diabetes, patients give themselves periodic injections of insulin as directed by their health care team. Injecting too much insulin may induce hypoglycemia (BG < 4 mmol/L, in our study), which can be dangerous, possibly causing a coma. However, injecting too little insulin can result in hyperglycemia (BG > 8 mmol/L, in our study), which may lead to chronic complications such as blindness, kidney failure, nerve damage and circulatory problems.\(^1,2\) In general, a patient’s BG will depend on many factors, including past carbohydrate intake, the amount of bolus/basal insulin injected, exercise, and stress.\(^2\)

Diabetes patients try to properly maintain their BG in a normal range. This is challenging because tight glycemic control using bolus insulin injections (whether intermittent with insulin pens or boluses using insulin pumps) is associated with an increased the risk of having hypoglycemic events.\(^1\) This challenge has led to attempts to create closed-loop systems and the use of computational techniques that assist in controlling patient’s BG levels.\(^3\) An extreme example of this is the effort to create an “artificial pancreas”, which explicitly integrates automatic monitoring with automatic administration of insulin.\(^4\)

Another perspective on fully automated diabetes management views the BG control problem as two sequential subproblems:

1. “Modeling”: Learning an accurate BG prediction model that, for example, predicts the BG level at lunch given a description of the subject up until breakfast (including perhaps her previous BG values, carbohydrate intake, etc., from earlier meals), as well as the amount of insulin injected at breakfast.
2. “Controlling”: Given the current information (at breakfast), consider the effects of injecting various possible amounts of insulin – that is, \{1 unit, 1.5 units, 2 units, \ldots\}. For each, use the learned model to predict the BG value at lunch, then inject the amount that is predicted to lead to the best lunch-time BG-value. (Of course, this assumes that decisions made at breakfast only affect lunch, then lunch decisions will only affect dinner, etc. – which does not consider the longer-range effects of actions; see Bastani.\(^3\))

This paper focuses on the first subtask: developing a BG prediction system, where in general, a model \(M\) will predict the blood glucose \(\hat{BG}_{i+1} = M(x_i, \Delta t_{i+1})\) at the next time point (\(\Delta t_{i+1}\) minutes into the future), based on information currently known about this patient, including the amount of insulin (BOLUS\(_i\)) the patient decided to inject:

\[
x_i = [\text{TIME}_i, \text{BG}_i, \text{BOLUS}_i, \text{BASAL}_i, \text{EV}_i, \text{PV}_i, \text{IOB}_i, \ldots] \tag{1}
\]

Note that this over-simplifies some issues; see Borle\(^5\) for details.

An example of this subtask might be to predict an individual’s blood glucose at lunch on Tuesday at 12:15pm, given information collected up-until 8am breakfast on Tuesday. (Note this might only include the Tuesday breakfast information, or it could include other earlier information – for
example, the ellipses in equation (1) might contain information about events from yesterday, or last week).

To be precise, the goal of this work is to determine if it is possible to accurately predict a T1D patient’s future BG, from one meal to the next, based only on the information typically recorded by the patient. To do this, a model must be able to deal with the features provided in a patient’s diabetes diary, which have varied prediction horizons.

This work is an extensive effort to learn an accurate BG prediction model, which involved exploring 312 different combinations of learner and preprocessing variants. To train and evaluate each of these variants, we used a dataset of 29,601 entries collected from 47 unique patients, where each entry included only the information typically collected, including: the time of day, the patient’s current BG, the carbohydrate about to be consumed and the anticipated exercise. Our results show that, surprisingly, this information is not sufficient to produce models that can make accurate predictions.

**Background literature and its limitations**

This research deals with long-range (>2 h) predictions of Blood Glucose values, for many (>30) real (not simulated) Type I (not Type II) patients, sampled over a long time range (months to years). While there are many existing works related to modelling Blood Glucose dynamics in T1D patients, these either have limitations in their datasets or design, or do not satisfy our specific goal. A recent literature review identified 49 publications using modeling techniques for blood glucose prediction (primarily with T1D patients) of which 38 used prediction horizons that were 60 min or less. One of these publications did involve prediction horizons of 180 min but only on simulated patients, and another had 1440-min prediction horizons but used a dataset of eight patients collected over only 3 days. Several studies use data from only a single patient, often including records from fewer than 100 days, or include only data from simulated patients. Other studies include more patients but only have 3–22 days of data, or have a few years worth of data but only include three patients. Further studies analyze continuous glucose monitoring (CGM) data from larger patient sets (89 T1D patients) but are again sampled over relatively short periods of time (1 week). (There are large type-2 diabetes (T2D) datasets (163 patients over 1 year), but studies modeling T2D address an easier problem because T2D patients have less variable BG values.)

While we focus on predicting BG values many hours later, some studies instead attempt to predict the occurrence of hypoglycemic events, and only within a short window (e.g. 30–120 min). One notable study included data collected from 40 patients over the course of 3 years. However, this was a CGM study that collected bursts of data at 3-month intervals and only considered prediction horizons of up to 30 min. While this might help to protect patients from a very serious situation, it is lacking in several ways. First, such fine-grained measurements are often not practically obtainable outside of a study setting and without using a CGM device that provides measurements every 5 min. Second, these short-term predictions are not adequate for spanning the time between meals. Third, the goal of building a diabetes control system is better served with a more expressive model, as opposed to one that can only provide binary classifications – hypoglycemic or not. Note that these model provides no useful feedback for situations where patients are hyperglycemic.

In our work, we try to model blood glucose dynamics (including both hyperglycemia and hypoglycemia), using only the standard records collected at meal times. While this makes our task more challenging, we do this because it involves only the data that medical professionals most often encounter in practice.
Our study will use both many common machine learning algorithms and also some less well-known algorithms that are motivated by the existing literature. These include a model that is similar to the Gaussian Wavelet Neural Network used by Zainuddin et al., and a weighted ensemble of Gaussian Process Regression (GPR) models that are constructed in a way that is similar to Duke, who uses a GPR to learn models of individual patients that could be used to aid in cross-patient prediction.

We will evaluate the quality of our predictions in several ways. Del Favero et al. describes various measures for comparing a patient’s specific glucose reading, with a predicted value, including both standard measures (like L1, relative L1 error, and L2 losses – there called MAD, MARD and RMSE) and some “glucose-specific metrics”, such as gMAD and gRMSE. While our paper focuses on the L1 and relative L1 losses, we also include the others mentioned there.

**Main contributions**

This work has three main contributions:

1. To our knowledge, this study examines the largest multi-year dataset of diabetes diary records, collected from Type 1 diabetes patients, used for modeling future BG.
2. We provide a comprehensive study of this data, considering 312 combinations of learning algorithm, types of features, and categories of records, to determine if machine learning can create an accurate blood glucose prediction model.
3. Our results demonstrate that it is difficult for any model – whether human or machine-learned – to use this Diabetes Diary data (i.e. the information that patients typically record) to predict that patient’s BG better than a naïve baseline (in this case, predicting a patient’s average BG). This applies when considering both standard error measures, like L1 and L2 loss, and also for glucose-specific measures, such as gMAD.

This work is based on the publicly available MSc thesis, which provides additional information, including a breakdown of the individual patients in the study, more detailed results, and a comparison to a diabetologist’s performance on this prediction task.

**Materials and methods**

This section first describes our (real world) dataset (pre-processing described in Borle et al.); it then considers two ways to modify this dataset. In general, the dataset is a set of records, each from a specific patient, where each is described by a set of features – for example, in Table 3, each column is a single record, whose values are described by a set of features (corresponding to the rows). One issue is determining the set of records to include: whether the dataset includes all of the records, from all patients, versus just the subset of “Expert Predictable” records from those patients (\(D^E\) defined below). The second modification deals with how we represent each such record. We consider various “feature sets” – the original set of features (shown in Table 2), and also 12 other variants, each of which includes various new features, that are combination of those original features.

To each of these 2×13 different [record_set, feature_set] combinations, we then apply each of 12 different learning algorithms (described below) – meaning we are exploring the results of applying 12 learning algorithms to each of 2 record subsets, each record of which can be expressed using any of 13 different feature sets (original and 12 variants), corresponding to 12×2×13 = 312
experiments. Finally we describe how we evaluate the quality of each of these [learning_algorithm, record_set, feature_set], and then provide our results.

Datasets considered

This study used 47 histories from Type I diabetes patients, which were collected using the “Intelligent Diabetes Management” (IDM) software (described in Ryan et al.35). This data included patients who participated in Ryan et al.’s study, as well as additional patients who began using the IDM software after the completion of the study (up until December 2016). For further details regarding patient participation, see Ryan et al.35 Some of the participants only used the system a few times. As we wanted to focus on patients that had sufficient information to find relevant patterns, we only included patients who made at least 100 diabetes diary entries with the system – that is, produced at least 100 “sufficient” records. This led to a dataset of 16 pump users and 31 non-pump patients. Table 1 provides summary statistics for our data. The dataset used for this work differs from the one described in Borle5 in that we limit the number of patients included to those with complete data.

Patient #16 is noteworthy for having by far the most records of any patient in our dataset; it is unusual for a patient to consistently produce diabetes entries over the course of many years. Because of the large number of records, we use part of this patient’s dataset as our hyper-parameter tuning (validation) dataset, as well as for visualization. Note we then use only the remaining portion of this patient’s records in our experiments, so that there is no overlap with the portion used for tuning the parameters.

Each record $i$ corresponds to an entry in a patient’s “diabetes diary”, which includes the meal associated with the record MEAL$_i$, a time stamp (DATE$_i$ and TIME$_i$), the blood glucose value BG$_i$, the grams of carbohydrates consumed CHO$_i$, and the units of bolus (resp., basal) injected BOLUS$_i$, (resp., BASAL$_i$). The patients also entered the anticipated level of exercise using the non-numeric values {“less than normal”, “normal”, “active”, “very active”}. We converted these into numeric values (2, 4, 7 and 10 respectively) for use by standard learning algorithms.

As mentioned above, 16 of the patients in this study used insulin pumps, which each directly infuse insulin from a reservoir, via a catheter, just under a patient’s skin at a basal rate. Moreover, they also self-inject larger amounts of bolus insulin when a patient ingests carbohydrates (as a patient would with an insulin pen [http://www.diabetes.org/living-with-diabetes/treatment-and-care/medication/insulin/how-do-insulin-pumps-work.html]). Each record of each insulin pump patient includes the basal infusion rate value PV$_i$, in $\text{units} / \text{hour}$. The insulin pump settings work by partitioning the 24h clock into intervals, and setting a particular delivery rate of insulin for each interval. The PV$_i$ values for any specific record was then set to the insulin delivery rate for the interval containing the record’s time stamp. We also computed two other features: $\Delta t_i$, which is the elapsed time since the previous record (actually, $\Delta t_i$ is based on previous BOLUS$_{i-1}$ and CHO$_{i-1}$ values; see Borle et al.34) and “Insulin on Board” IOB$_i$, which captures the effect of any insulin

| # Patients | Range: # usable records | Pump users | Sex | Age | Height* | Weight |
|------------|-------------------------|------------|-----|-----|---------|--------|
| 47         | 106–4339 (30–3323)      | 16         | $9 \ddagger \ddagger / 38 \ddagger \ddagger$ | $42 \pm 13 \text{years}$ | $166 \pm 8 \text{cm}$ | $74 \pm 14 \text{kg}$ |

Values are shown as mean ± standard-deviation. See Borle$^5$ for more details about the individual patients.

$\ddagger$This is the number of records in $D^5$, defined below; see also the distribution shown in Figure 2.

$\ddagger \ddagger$We did not know the height for seven individuals; this average value was calculated using only the remaining 40 patients.
remaining in a person’s system from previous injections. This was based on Figure 1, which was computed using a simple spline to interpolate from the following pairs of elapsed time and percentage of post-injection insulin remaining: (1.66 h, 78%), (2.5 h, 48%), (3.33 h, 27%), (4.15 h, 12%), (5 h, 3%). Table 2 describes all of these features and Table 3 provides example data.

To prepare our data for the various learning algorithms, we used several different preprocessing approaches, which produced different versions of our dataset. These included several methods for handling missing values such as removing records or imputing average estimated values. For a more detailed description of this process, see Borle et al.

Subset of only “Expert Predictable” entries

As our data was collected voluntarily from patients at their own convenience, sampling intervals are not uniform, and the relevant data is not recorded for every meal. This is problematic for our predictive task as blood glucose values are more difficult to predict as more time elapses between readings. To address this issue, our clinician co-author established the following criteria of when it is reasonable to predict the next glucose value, claiming the BG is “expert predictable” (EP) at a given time if all of the following are true:

1. The preceding record is not a hypoglycemic event, as that subsequent BG reading is difficult to predict due to potential glucose counter-regulation effects and the uncertainty in BG that follows from a physiological response to hypoglycemia.
2. The blood glucose reading is present for the preceding meal. For example, to make a prediction about a patient’s blood glucose value before lunch, a record detailing his/her previous breakfast must be available.
Six of the last 8 days prior to a prediction must have records for both the current meal time and the previous meal time. For example, to predict the blood glucose before lunch, 6 of the last 8 days must have both “before lunch” and “after breakfast” entries, to help capture this “after breakfast to before lunch” transition pattern.

Figure 2 shows the number of records from each patient that qualify as EP – that is, the number of records for which our expert would feel comfortable making predictions. We will let \( D^E \) refer to the set of records that met the expert’s EP criteria; as opposed to the complete set of records, \( D^4 \).

### Feature engineering

Table 2 shows the basic features used to describe each event. Additionally, we also considered many other feature sets to see if any could lead to better performance. Some of the variants

| MEAL \(_i\) | The time of day: (Before Breakfast, After Breakfast, Before lunch, After Lunch, Before Supper, After Supper, Before Bed, During the Night) |
| DATE \(_i\) | The date as year-month-day |
| TIME \(_i\) | The time as hour:minute:second* |
| BG \(_i\) | The BG value at the current time (\( \frac{mmol}{L} \)) |
| CHO \(_i\) | The amount of carbohydrates ingested (grams) |
| BOLUS \(_i\) | The amount of insulin injected (units) |
| BASAL \(_i\) | The units of background insulin injected |
| EV \(_i\) | Numeric encoding of exercise value: \( \{2, 4, 7, 10\} \) |
| PV \(_i\) | Pump Value: The rate at which the insulin pump is infusing (\( \frac{units}{hour} \)). This is always 0 if the patient does not have a pump. |
| \( \Delta t \) | The elapsed time since last record |
| IOB \(_i\) | Insulin on Board: Estimated residual bolus insulin from the previous injection (\( \frac{mmol}{L} \)) |

See text for further description of these terms. Note this is a simplified set of features; see Borle et al.\(^{34}\) for the complete set of feature descriptions.

*Only the hour and minute were captured for each record so the seconds field is always 00.

### Table 3. Example of data, over a single day, from patient 16.

| Index \(_i\) | 27 | 28 | 29 | 30 | 31 | 32 |
| MEAL \(_i\) | Before breakfast | After breakfast | Before lunch | After lunch | Before dinner | After dinner |
| DATE \(_i\) | 2015-11-25 | 2015-11-25 | 2015-11-25 | 2015-11-25 | 2015-11-25 | 2015-11-25 |
| TIME \(_i\) | 08:36:00 | 10:19:00 | 12:19:00 | 15:35:00 | 18:42:00 | 20:11:00 |
| BG \(_i\) | 16.2 | 14.7 | 5.6 | 6.8 | 10.5 | 3.0 |
| CHO \(_i\) | 30.0 | 0 | 30.0 | 0 | 15.0 | 0 |
| BOLUS \(_i\) | 10.4 | 0 | 3.0 | 0 | 3.8 | 0 |
| BASAL \(_i\) | 0 | 0 | 0 | 0 | 0 | 0 |
| EV \(_i\) | 4 | 4 | 4 | 4 | 4 | 4 |
| PV \(_i\) | 0.50 | 0.50 | 0.63 | 0.45 | 0.90 | 0.90 |
| \( \Delta t \) | 540 | 103 | 120 | 196 | 187 | 89 |
| IOB \(_i\) | 0.00 | 7.90 | 3.61 | 0.89 | 0.81 | 3.35 |

Note this is a simplified version of the data; Borle et al.\(^{34}\) provides the general, complete set of features.

3. Six of the last 8 days prior to a prediction must have records for both the current meal time and the previous meal time. For example, to predict the blood glucose before lunch, 6 of the last 8 days must have both “before lunch” and “after breakfast” entries, to help capture this “after breakfast to before lunch” transition pattern.

Figure 2 shows the number of records from each patient that qualify as EP – that is, the number of records for which our expert would feel comfortable making predictions. We will let \( D^E \) refer to the set of records that met the expert’s EP criteria; as opposed to the complete set of records, \( D^4 \).

**Feature engineering**

Table 2 shows the basic features used to describe each event. Additionally, we also considered many other feature sets to see if any could lead to better performance. Some of the variants
completed records that were missing entries for carbohydrates or bolus insulin, which others removed those deficit records. Some added in the day of the week as an integer feature or as a one-hot encoded feature (http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html), while others removed the “basal insulin” feature. A few variants included non-temporal patient characteristics: age, gender, height and weight. Some replaced the set of features with just the first four principal components (obtained by principal component analysis, PCA).

Our “Kok Features” variant uses computed features similar to Kok, and subsequently used by Baghdadi and Nasrabadi and Zainuddin et al. Unlike Kok’s data, however, we do not have stress level values in our data and were therefore unable to incorporate that feature.

For any given dataset variant, some ensembled learners have components that train on different subsets of the data. In addition, we also use models that include components that are trained on the data from all patients other than the current test patient, as well as learners that involve sub-models that are each trained only on data from one meal type (e.g. before breakfast).

Note we compute each of these 13 feature set variants for the original dataset $D^A$ that included all records; this leads to the datasets $\{D_1^A, D_2^A, \ldots, D_{13}^A\}$. We also compute all 13 of these feature sets for $D^E$, which includes only the “EP-filtered” subset of records; this leads to $\{D_1^E, D_2^E, \ldots, D_{13}^E\}$. Note that this reduces the predictions that our system attempts. For example, we do not attempt to predict a before-lunch BG value if there is no preceding after-breakfast reading. For each variant, we only considered the subset of the records that belonged to that variant, both for producing the model and also for estimating the quality of that model – in particular, each of the $13 \times 12$ “Expert-Predictable”-based models were trained and tested on the $D^E$ set of records.

The complete set of preprocessing variants are described in Borle et al.

**Machine learning algorithms**

As the diabetes diary data for each patient is a temporal sequence of entries, we initially considered modeling the data as a time series, using something like an HMM (https://hmmlearn.readthedocs.io/
However, we realized this would be problematic due to the irregular sampling of the data, and so decided to view our task as a set of single step predictions; this allows us to use standard supervised machine learning systems to produce regression models. Note that these predictions can use, as input, data describing various earlier time points – for example, some of our models for predicting the BG at dinner time can use information about the patient at lunch time, and also at breakfast time, and perhaps also other previous dinner times, etc.

This work considers twelve different supervised machine learners, based on seven base learning algorithms: K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Wavelet Neural Networks (WNN), Ridge Regression (RR), Random Forest Regression (RFR) and Gaussian Process Regression (GPR). We created five other, more complex, learners, by considering weighted ensembles of GPRs and by using model stacking, where the base models are produced using different subsets of the training data. For selecting hyper-parameters, we used a portion of patients #16’s data. (Note we did not use that in the final evaluation.) For more details regarding these models, including hyper-parameter selections, please refer to Borle et al.34

Model evaluation

To assess the performance of our models in general, we use various evaluation functions to measure the quality of model predictions for various instances, with respect to known true outcomes. In this paper we report our results in terms of “L_1-loss” ($err_{L_1}$), “relative L_1-loss” ($err_{rL_1}$) and “Root Mean Squared Error” (RMSE).

For each of these three metrics, we also consider a “glucose-specific” variant, that uses a “Clark Error Grid inspired penalty function” to re-weight the relative costs of different mis-predictions; see Del Favero et al.32 for descriptions and definitions (which refers to $err_{L_1}$ and $err_{rL_1}$ as MAD and MARD, respectively). Therefore in total, we consider the performance of our models across six different evaluation functions. To help calibrate the quality of our results, we also consider a na"ive model $M_{ave}$ that only predicts a patient’s average BG, regardless of any other information. (This is like predicting that the temperature is Edmonton is 4.6°C, on any day, independent of the temperature yesterday, or whether it is summer or winter, etc.) See Borle et al.34 for more details.

10-fold cross validation.

Each of our learners will use the entire dataset to produce a model. The next challenge is evaluating the predictive quality of each learner. Here, we use 10-fold cross validation (CV), with respect to each patient. We first partition time series history of a patient, denoted $X_i$, into 10 disjoint subsets, each corresponding to a contiguous time interval, $X_i^j$; here $X_i = \bigcup_{j=1..10} X_i^j$.

We then use 9 of the 10 segments for training in each CV round and use the remaining 1 segment for testing – so the first split would be train $X_i^{1..9}$ and test $X_i^{10}$. While the testing partition always consists of contiguous data, the training partition will not always be completely contiguous. Figure 3 provides a visualization of what it means to partition time series data into contiguous segments for the purposes of cross validation (each row is a fold) – for simplicity, here we show “5-fold CV” rather than 10, and suppress the subscript $i$.

Results

Cross validation results

For each of the 47 patient histories (excluding the portion used for hyperparameter selection; see above), for each of the $2 \times 13 \times 12$ [record_set, feature_set, learner] models, we perform 10-fold CV using that specific learner, on that specific [record_set, feature_set] dataset, to determine
its effectiveness and how well it compares to the baseline model $M_{ave}$. For details regarding the models, heat-maps describing the relative performance of different models on different dataset variants (in terms of $err_{L1}$ and $err_{rL1}$, and other measures), see Borle et al.\textsuperscript{34}

For each $[\text{record\_set}, \text{feature\_set}, \text{learner}]$ situation, we compute the average-$err_{L1}$ error as a micro-average over all the records of all 47 patients. When considering the $D^A$ data, the best result was $[\text{record\_set} = D^A, \text{feature\_set} = \text{“Original”}, \text{learner} = M_{\text{gpr}^w}]$ – that is, running the learner $M_{\text{gpr}^w}$ on $D^A$ type of data, which yielded an average loss $err_{L1} = 2.8 \text{ mmol/L}$. Towards evaluating how good this result is, we then ran the simplistic $M_{ave}$ model, on those same $D^A$ instances (recall the feature\_set does not matter here). Its average $err_{L1}$ was 3.0 mmol/L – that is, our exploration, over 156 different variants, produced a model that was only 6.14\% better than the baseline!

Even worse: note also that our “2.8 mmol/L error” claim is actually optimistic, as this was the high-water mark over many learners, which is technically not valid. If we ran a correct learner (that used internal cross-validation to identify the appropriate learner, and the best feature\_set), we anticipate the result would worse. Our main point is: even when we were “cheating”, we were still not much better than baseline!

We next considered the “easier” subset of records, from $D^E$. Here, the best combination was $[\text{record\_set} = D^E, \text{feature\_set} = \text{“Original”}, \text{learner} = M_{\text{gpr}^w}]$ – again running the $M_{\text{gpr}^w}$ learner, but here on $D^E$ data, with error $err_{L1} = 2.7 \text{ mmol/L}$. This was only 7.1\% better than the baseline (running $M_{ave}$ learner on these same $D^E$ data), of $err_{L1}$ of 2.9 mmol/L.

To help understand why our best result was not better, Figure 4 shows the predictions of $M_{\text{gpr}^w}$ for the processed entries from patient#16 that were used for selecting hyper-parameters (using $D^A$ data). Here, we can see that even the best model is unable to account for the high amount of variance present in the BG records for this patient.

Figure 5 plots the variance in each patient’s BG history and the corresponding patient’s $err_{L1}$ loss that $M_{\text{gpr}^w}$ was able to achieve (using $D^A$ data). This figure shows that the variance of a patient’s blood glucose was highly correlated with the $err_{L1}$ test loss (0.93 Pearson Correlation).

One obvious question is whether these disappointing results were due to the sample size – perhaps we would get better results if we had more records for each patient? We therefore explored whether a patient’s $err_{L1}$ loss varied with that patient’s number of records. However, a scatter plot of $err_{L1}$ loss versus number of records, for each patient in the dataset (Figure 6), suggests that there is no such relationship – here the Pearson Correlation is –0.14.
Other evaluation measures

We then considered the relative $L1$-loss, $err_{L1}$, and found that the best value (on $D^4$) was obtained using the same learner $M_{gpr}^w$, but on the $D_{11}^4$ feature set. This is the feature set that includes a “day of the week” feature and a basal insulin feature, replaces missing carbs values with imputed mean values, and excludes records without bolus values. Its $err_{L1}$ loss of 0.356 was 15.88% better than the loss of $M_{ave}$: 0.422. Similarly, on the $D^5$ records, the best result was again $M_{gpr}^w$ on $D_{11}^4$, achieving 0.348, which was 18.97% better than $M_{ave}$’s score of 0.430.

In addition to our $err_{L1}$ and $err_{rl1}$ metrics, we also computed RMSE and the glucose-specific versions of these metrics; see Table 4. We were surprised to find that the best model, across all six

![Figure 4. Model $M_{gpr}^w$: our GPR ensemble’s predictions on data from patient 16.](image1)

![Figure 5. Model $M_{gpr}^w$: average $err_{L1}$ as a function of BG variance, for all patient histories.](image2)
Figure 6. Model $M_{w}^{gr}$: average $err_{L1}$ as a function of the # of diabetes diary entries for a patient, for all patient histories.

Table 4. Performance of the best models across all metrics on $DA$ and $DE$ datasets.

| Metric       | Naive model error | Best model error | Percent improvement | Best model | Best dataset variant |
|--------------|-------------------|------------------|---------------------|------------|----------------------|
| All records $DA$ |                   |                  |                     |            |                      |
| $err_{L1}$ (MAD) | 2.96              | 2.78             | 6.14                | $M_{w}^{gr}$ | $D_{1}^{A}$           |
| gMAD         | 5.45              | 5.14             | 5.70                | $M_{w}^{gr}$ | $D_{1}^{A}$           |
| $err_{L1}$ (MARD) | 0.422             | 0.355            | 15.88               | $M_{w}^{gr}$ | $D_{11}^{A}$          |
| gMARD        | 0.772             | 0.660            | 14.50               | $M_{w}^{gr}$ | $D_{1}^{A}$           |
| RMSE         | 3.67              | 3.58             | 2.49                | $M_{w}^{gr}$ | $D_{1}^{A}$           |
| gRMSE        | 5.16              | 5.02             | 2.70                | $M_{w}^{gr}$ | $D_{1}^{A}$           |
| Expert predictable records $DE$ |                   |                  |                     |            |                      |
| $err_{L1}$ (MAD) | 2.91              | 2.70             | 7.12                | $M_{w}^{gr}$ | $D_{1}^{E}$           |
| gMAD         | 5.31              | 4.98             | 6.28                | $M_{w}^{gr}$ | $D_{1}^{E}$           |
| $err_{L1}$ (MARD) | 0.430             | 0.348            | 18.97               | $M_{w}^{gr}$ | $D_{11}^{E}$          |
| gMARD        | 0.783             | 0.648            | 17.28               | $M_{w}^{gr}$ | $D_{1}^{E}$           |
| RMSE         | 3.58              | 3.47             | 2.98                | $M_{w}^{gr}$ | $D_{1}^{E}$           |
| gRMSE        | 5.01              | 4.86             | 2.95                | $M_{w}^{gr}$ | $D_{1}^{E}$           |

Note that these error values are micro-averages.

metrics was $M_{w}^{gr}$, and the best feature set was either the original data ($D_{1}^{A}$ or $D_{1}^{E}$) for for $err_{L1}$ (MAD) or RMSE and their glucose-specific variants gMAD and gRMSE, and was the 11th dataset ($D_{11}^{A}$ or $D_{11}^{E}$) for $err_{L1}$ (MARD) and its glucose-specific variant gMARD.
Discussion

Our results show that our best learning algorithm is more accurate than a naive baseline – but only slightly – but that, even in the best situation or relatively easy instances (on the Expert Predictable records $D^E$) it can achieve an average $err_{L1}$-loss of only 2.7 mmol/L. This loss means that, if the patient’s blood glucose was normal (e.g. 6 mmol/L), the learned model may incorrectly identify the patient as either hypoglycemic (as $6 - 2.7 < 4$ mmol/L) or hyperglycemic ($6 + 2.7 > 8$ mmol/L).

Together with the strong relationship between glucose variance and prediction error, this highlights how challenging it is to create models that produce fine-grained blood glucose predictions when only using diabetes diary entries – that is, using only the information that is commonly available to medical practitioners. We also see that the $D^E$-learning task was slightly easier than the $D^A$-learning task, as our results showing the best model on $D^E$ had an error that was 2.88% better than on $D^A$ – even though it was trained on fewer training instances.

Having tried 312 different combinations of learners, record_set and feature_set variants, and observing minimal differences in their predictive performance, it seems unlikely that another [learner, record_set, feature_set] combination would be better. Note that these approaches include models that use data from multiple patients, which suggest that simply including more patients in the study is not likely to improve model performance. Moreover, since the model accuracy did not seem to improve as the number of records increased, we suspect that simply collecting more of these entries for each individual will not improve model performance.

There are many possible reasons why modeling T1D glucose levels based on this standard type of diabetes diary data is so challenging. It is possibly just an artifact of our study: perhaps many of the patients who volunteered, did so because their diabetes was difficult to manage. (Although, of course, this was not an inclusion criterion.) Another reason could be that inaccuracies and omissions of variables in data prevent the model from producing accurate predictions. These omissions could possibly include: not knowing the site where the bolus insulin was injected, how much scar tissue was present at the injection site, skin temperature, how accurately the carbohydrate value was recorded, the accuracy of the recorded insulin dose, the levels of different hormones, whether the patient was menstruating, accuracy of recording exertion and/or stress, insulin age or storage conditions, amount of blood flow at the injection site and likely yet other factors. Given our belief that training more accurate models will require additional relevant variables, future research might incorporate more confounding variables, such as injection location, glucon levels, meal protein/fat content, amount of fiber (complex carbohydrates), influence of food order, and relationships between hyperglycemia and gastric motility/emptying. However, it is not clear which, if any, of such variables are sufficient to explain the response, nor whether they can be practically captured in a clinical setting. Another approach might be to avoid such a long prediction time – as mentioned above, many systems have much better prediction accuracy by making short term forecasts – on the order of tens of minutes, instead of hours. Of course, these modifications change our task, as they involve data, or timing, that does not correspond to the typical diabetes diary. So while they would probably lead to a more accurate model, that does not change the message of this paper: that the information collected in standard diabetes diaries is not sufficient to accurately estimate the patient’s next-meal BG.

Conclusion

This work explored the challenge of accurately predicting future blood glucose values in Type I diabetes patients. Our extensive explorations – involving 12 different learning algorithms, over 13 different feature sets, and both the original set of records, as well as the subset of “easier to predict”
records (312 combinations) – found that, on average, the model with the lowest expected $err_{L1}$ was a confidence weighted Gaussian process regression model ($M_{gpr}^w$). Using 10-fold cross validation on 29,601 blood glucose records from 47 different patients, this $M_{gpr}^w$ model performed only 7.1% better than the naïve “mean predicting” model ($M_{avg}$). Anecdotally, a diabetologist also attempted to do this task – predicting the BG for the next meal. We found that our learned model (insignificantly) outperformed the diabetologist’s, in terms of a simple unbiased loss function, but that the diabetologist performed (insignificantly) better when the evaluation was biased toward predicting hypoglycemic events; see Borle.⁵

These results showed that our model could achieve an expected absolute error of 2.7 mmol/L (48.65 mg/dl), which is disconcertingly large given that this is based on the type of data that is frequently collected and used for clinical practice – that is, records collected at meal times by the patients themselves. These results strongly suggest that the standard data collected by T1D patients, while apparently sufficient for clinical treatment of T1D, is not sufficient for accurately predicting blood glucose levels. We conjecture that using patient data that is sampled more frequently (perhaps using a device like FreeStyle Libre⁴⁶) and that includes additional features would improve both the ability of professionals and machine learning practitioners to more accurately predict patient’s blood glucose levels, but there is a practical trade-off between patient convenience and highly detailed record keeping.

Acknowledgements

The authors gratefully acknowledge help from Haipeng (Paul) Li, as well as our visiting summer interns Abhinav Agrawalla, Prachi Agrawal and Pranjal Daga from the Indian Institute of Technology, Kharagpur.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded in part by a pilot project grant from the Alberta Diabetes Institute (ADI). RG gratefully acknowledges funding from the Alberta Machine Intelligence Institute (Amii), and NSERC; E.A.R. from ADI; NB from NSERC.

ORCID iD

Neil C Borle https://orcid.org/0000-0002-5960-1236

References

1. Daneman D. Type 1 diabetes. *Lancet* 2006; 367(9513): 847–858.
2. Kok P. Predicting blood glucose levels of diabetics using artificial neural networks. Research Assignment for Master of Science, Delf University of Technology, 2004.
3. Bastani M. Model-free intelligent diabetes management using machine learning. Master’s Thesis, Department of Computing Science, University of Alberta, 2014.
4. Lunze K, Singh T, Walter M, et al. Blood glucose control algorithms for type 1 diabetic patients: a methodological review. *Biomed Signal Process Control* 2013; 8(2): 107–119.
5. Borle N. The challenge of predicting future blood glucose for patients with type I diabetes. Master’s Thesis, Department of Computing Science, University of Alberta, 2017.
6. Contreras I and Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res* 2018; 20(5): e10775.

7. Tresp V, Briegel T and Moody J. Neural-network models for the blood glucose metabolism of a diabetic. *IEEE Trans Neural Netw* 1999; 10(5): 1204–1213.

8. Baghdadi G and Nasrabadi AM. Controlling blood glucose levels in diabetics by neural network predictor. In: *Engineering in medicine and biology society, 2007. EMBS 2007. 29th annual international conference of the IEEE*, Lyon, France, 22 August 2007, pp. 3216–3219. Piscataway, NJ: IEEE.

9. Zainuddin Z, Pauline O and Ardl C. A neural network approach in predicting the blood glucose level for diabetic patients. *Int J Comput Intell* 2009; 5: 72–79.

10. Asad M, Qamar U, Zeb B, et al. Blood glucose level prediction with minimal inputs using feedforward neural network for diabetic type 1 patients. In: *Proceedings of the 2019 11th international conference on machine learning and computing*, Zhuhai, China, 22 February 2019, pp.182–185. New York, NY: ACM.

11. Aiello EM, Toffanin C, Messori M, et al. Postprandial glucose regulation via kNN meal classification in type 1 diabetes. *IEEE Control Syst Lett* 2018; 3(2): 230–235.

12. Sun X, Yu X, Liu J, et al. Glucose prediction for type 1 diabetes using kLMS algorithm. In: *2017 36th Chinese control conference (CCC)*, Da Lian, China, 26 July 2017, pp. 1124–1128. Piscataway, NJ: IEEE.

13. Georga EI, Protopappas VC, Polyzos D, et al. Online prediction of glucose concentration in type 1 diabetes using extreme learning machines. In: *Engineering in medicine and biology society (EMBC), 2015 37th annual international conference of the IEEE*, Milan, Italy, pp. 3262–3265. Piscataway, New Jersey: IEEE.

14. Andreassen S, Benn JJ, Hovorka R, et al. A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. *Comput Meth Prog Biomed* 1994; 41(3–4): 153–165.

15. Pappada SM, Cameron BD and Rosman PM. Development of a neural network for prediction of glucose concentration in type 1 diabetes patients. *J Diabetes Sci Technol* 2008; 2(5): 792–801.

16. Valletta JJ, Chipperfield AJ and Byrne CD. Gaussian process modelling of blood glucose response to free-living physical activity data in people with type 1 diabetes. In: *Engineering in medicine and biology society. 2009. EMBC 2009. Annual international conference of the IEEE*, Minneapolis, MN, 3 September 2009, pp. 4913–4916. Piscataway, N.J: IEEE.

17. Duke DL. *Intelligent diabetes assistant: a telemedicine system for modeling and managing blood glucose*. Pittsburgh: Carnegie Mellon University, 2010.

18. Zarkogianni K, Mitsis K, Litsa E, et al. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. *Med Biol Eng Comput* 2015; 53(12): 1333–1343.

19. Liu C, Vehi J, Oliver N, et al. Enhancing blood glucose prediction with meal absorption and physical exercise information. *arXiv preprint arXiv:190107467*, 2018.

20. Ali JB, Hamdi T, Fnaiech N, et al. Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network. *Biocybern Biomed Eng* 2018; 38(4): 828–840.

21. Magni P and Bellazzi R. A stochastic model to assess the variability of blood glucose time series in diabetic patients self-monitoring. *IEEE Trans Biomed Eng* 2006; 53(6): 977–985.

22. Gadaleta M, Facchinetti A, Grisan E, et al. Prediction of adverse glycemic events from continuous glucose monitoring signal. *IEEE J Biomed Health Inform* 2018; 23(2): 650–659.

23. Quinn CC, Shardell MD, Terrin ML, et al. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. *Diabetes Care* 2011; 34(9): 1934–1942.

24. Chemlal S, Colberg S, Satin-Smith M, et al. Blood glucose individualized prediction for type 2 diabetes using iphone application. In: *Bioengineering conference (NEBEC), 2011 IEEE 37th annual Northeast*, Troy, NY, 1–3 April 2011, pp.1–2. Piscataway, NJ: IEEE.

25. Sudharsan B, Peeples M and Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J Diabetes Sci Technol* 2014; 9(1): 86–90.
26. Bunescu R, Struble N, Marling C, et al. Blood glucose level prediction using physiological models and support vector regression. In: *Machine learning and applications (ICMLA), 2013 12th international conference on*, Miami, FL, 4–7 December 2013, vol. 1, pp. 135–140. Washington, DC: IEEE.

27. Eren-Oruklu M, Cinar A and Quinn L. Hypoglycemia prediction with subject-specific recursive time-series models. *J Diabetes Sci Technol* 2010; 4(1): 25–33.

28. Pappada SM, Cameron BD, Rosman PM, et al. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes Technol Ther* 2011; 13(2): 135–141.

29. Plis K, Bunescu R, Marling C, et al. A machine learning approach to predicting blood glucose levels for diabetes management. In: *Modern artificial intelligence for health analytics papers from the AAAI-14*, 2014. Quebec City, 27-31 July. Palo Alto, CA: The AAAI Press.

30. Doike T, Hayashi K, Arata S, et al. A blood glucose level prediction system using machine learning based on recurrent neural network for hypoglycemia prevention. In: *2018 16th IEEE international new circuits and systems conference (NEWCAS)*, Montreal, 24 June 2018, pp. 291–295. Piscataway, NJ: IEEE.

31. Fox I, Ang L, Jaiswal M, et al. Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, London, UK, 19 July 2018, pp. 1387–1395. New York, NY: ACM.

32. Del Favero S, Facchinetti A and Cobelli C. A glucose-specific metric to assess predictors and identify models. *IEEE Trans Biomed Eng* 2012; 59(5): 1281–1290.

33. Clarke WL. The original clarke error grid analysis (ega). *Diabetes Technol Ther* 2005; 7(5): 776–779.

34. Borle NC, Ryan EA and Greiner R. The challenge of predicting meal-to-meal blood glucose concentrations for patients with type i diabetes. *arXiv preprint arXiv:190312347*, 2019.

35. Ryan EA, Holland J, Stroulia E, et al. Improved a1c levels in type 1 diabetes with smartphone app use. *Can J Diabetes* 2017; 41(1): 33–40.

36. Al-Taee AM, Al-Taee MA, Al-Nuaimy W, et al. Smart bolus estimation taking into account the amount of insulin on board. In: *Computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing (CIT/IUCC/DASC/PICOM), 2015 IEEE international conference on*, 26 October 2015, pp. 1051–1056. Piscataway, NJ: IEEE.

37. Mudaliar SR, Lindberg FA, Joyce M, et al. Insulin aspart (b28 asp-insulin): a fast-acting analog of human insulin: absorption kinetics and action profile compared with regular human insulin in healthy nondiabetic subjects. *Diabetes Care* 1999; 22(9): 1501–1506.

38. Gerich JE. Glucose counterregulation and its impact on diabetes mellitus. *Diabetes* 1988; 37(12): 1608–1617.

39. Weisang G and Awazu Y. Vagaries of the euro: an introduction to arima modeling. *Case Stud Bus Ind Govt Stat* 2008; 2(1): 45–55.

40. Koivisto VA and Felig P. Alterations in insulin absorption and in blood glucose control associated with varying insulin injection sites in diabetic patients. *Ann Intern Med* 1980; 92(1): 59–61.

41. Unger RH and Cherrington AD. Glucagonocentric restructuring of diabetes: a pathophysiologic and therapeutic makeover. *J Clin Investig* 2012; 122(1): 4.

42. Paterson M, Smart C, Lopez P, et al. Increasing the protein quantity in a meal results in dose-dependent effects on postprandial glucose levels in individuals with type 1 diabetes mellitus. *Diabet Med* 2017; 34(6): 851–854.

43. Ahola A, Harjutsalo V, Forsblom C, et al. Associations of dietary macronutrient and fibre intake with glycaemia in individuals with type 1 diabetes. *Diabet Med* 2019; 36(11): 1391–1398.

44. Faber EM, van Kampen PM, Clement-de Boers A, et al. The influence of food order on postprandial glucose levels in children with type 1 diabetes. *Pediatr Diabetes* 2018; 19(4): 809–815.

45. De Boer S, Masclee A and Lamers C. Effect of hyperglycemia on gastrointestinal and gallbladder motility. *Scand J Gastroenterol* 1992; 27(sup194): 13–18.

46. Blum A. Freestyle libre glucose monitoring system. *Clin Diabetes* 2018; 36(2): 203–204.