DIABETES TECHNOLOGY & THERAPEUTICS
Volume 22, Number 12, 2020

DOI: 10.1089/dia.2020.0061

DTT

ORIGINAL ARTICLE

Multi-Hour Blood Glucose Prediction in Type 1 Diabetes: A Patient-Specific Approach Using Shallow Neural Network Models

Taisa Kushner, MS,[1,2] Marc D. Breton, PhD,[3] and Sriram Sankaranarayanan, PhD[1]

Abstract

Background: Considering current insulin action profiles and the nature of glycemic responses to insulin, there is an acute need for longer term, accurate, blood glucose predictions to inform insulin dosing schedules and enable effective decision support for the treatment of type 1 diabetes (T1D). However, current methods achieve acceptable accuracy only for prediction horizons of up to 1 h, whereas typical postprandial excursions and insulin action profiles last 4–6 h. In this study, we present models for prediction horizons of 60–240 min developed by leveraging ''shallow'' *neural networks*, allowing for significantly lower complexity compared with related approaches.

Methods: Patient-specific neural network-based predictive models are developed and tested on previously collected data from a cohort of 24 subjects with T1D. Models are designed to avoid serious pitfalls through incorporating essential physiological knowledge into model structure. Patient-specific models were generated to predict glucose 60, 90, 120, 180, and 240 min ahead, and a ''transfer learning'' approach to improve accuracy for patients where data are limited. Finally, we determined subgroup characteristics that result in higher model accuracy overall.

Results: Root mean squared error was 28 ± 4, 33 ± 4, 38 ± 6, 40 ± 8, and 43 ± 12 mg/dL for 60, 90, 120, 180, and 240 min, respectively. For all prediction horizons, at least 93% of predictions were clinically acceptable by the Clarke error grid. Variance of historic continuous glucose monitor (CGM) values was a strong predictor for the need of transfer learning approaches.

Conclusions: A shallow neural network, using features extracted from past CGM data and insulin logs, can achieve multi-hour glucose predictions with satisfactory accuracy. Models are patient specific, learnt on readily available data without the need for additional tests, and improve accuracy while lowering complexity compared with related approaches, paving the way for new advisory and closed loop algorithms able to encompass most of the insulin action timeframe.

Keywords: Blood glucose prediction, Neural networks, Artificial pancreas, Type 1 diabetes.

Introduction

Type 1 diabetes mellitus (T1D) is characterized as an autoimmune condition resulting in the body's inability to produce insulin. Thus, individuals with T1D must constantly monitor blood glucose (BG) levels and adjust doses of exogenous insulin accordingly. However, determining a proper dosing schedule is complicated by the need to calculate how to control future BG levels amidst anticipated activities such as meals and exercise, using insulin doses that have action profiles lasting upwards of 4–6 h.[1] To ease patient burden, numerous advisory strategies and closed-loop systems have been created, although most rely on predictive models that forecast 15–60 min into the future.[2–4] Furthermore, we find no proposed predictive models with prediction horizons over 120 min.[5–11]

Although current predictive models still fall short of the 4–6 h insulin action profile, recent work has sought to extend

[1]Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, USA.
[2]IQ Biology, Biofrontiers Institute, University of Colorado Boulder, Boulder, Colorado, USA.
[3]Center for Diabetes Technology, University of Virginia, Charlottesville, Virginia, USA.

prediction horizons. Some notable longer term models include that of Perez-Gandıa et al. who used past BG levels and a three-layer feed-forward neural network to predict BG values 30–45 min out, with root mean squared error (RMSE) 18–27 mg/dL.[6] Pappada et al.[8] utilized a feed-forward network to predict BG 75 min out using previous continuous glucose monitor (CGM) data to achieve RMSE of 43.9 mg/dL for a vector of 15 glucose values across the 75-min horizon. In fact, the model with the longest prediction horizon identified in literature thus far is that of Georga et al.,[10,11] which predicts out to 120 min. This model utilized support vector regression and random forests methods using data from 15 patients, and reported RMSE of 7.62 mg/dL using the support vector machine[10] approach and 10.83 mg/dL using a random forest classifier.[11] However, their input sources of plasma insulin concentration, instantaneous energy expenditure, and meal-derived rate glucose rate of appearance may be difficult to obtain outside a dedicated research center. Thus, although these models show promise, a multi-hour discrepancy between predictions and insulin action profiles remains.[1]

In this study, we propose approaches to learn neural network predictive models from CGM and insulin pump logs for patients with T1D. Using shallow neural networks, our approach is able to provide useful and accurate predictions for patients with T1D over time horizons of up to 4 h. This represents a significant improvement over the current state-of-the-art in terms of the accuracies achieved over such a long-term horizon. Furthermore, our work used relatively small neural network models that significantly reduced complexity over related approaches, an important consideration for use in automated insulin delivery applications.

Research Design and Methods

Patient data

The data set was collected during the observation period of a clinical trial in individuals with T1D.[12] The study included 16 continuous subcutaneous insulin infusion (CSII) and 8 multiple daily injection (MDI) users, with a broad range of ages represented. Overall key statistics are presented in Table 1, with full details found in the original study article.[12] During the observational period for which our data are collected, individual glucose measurements were collected using a blinded Dexcom G4 Platinum CGM with Share (Dexcom, San Diego, CA). Heart rate and step count from Fitbit devices were also present, although these measurements were intermittent.

Data partitioning

Because of the high correlation between subsequent BG and insulin values, and the neural network's ability to overfit the training data, the standard method of selecting a random 20% subsection of data to be used as testing data to evaluate the model's accuracy, and the remaining 80% to be used as training data to learn the model is prone to yield falsely high accuracy results (Fig. 1A).[13] Thus, rather than performing a purely randomized 80/20 split, we utilized a sectioned window randomization method to ensure correlations are minimized. For each individual, we subdivided our time series data into continuous *windows* of data sampled at 5-min intervals. Within these windows, we obtained input–output data vector pairs using a sliding window. Next, rather than randomizing the individual data vectors, we randomized the *windows* into training and testing data utilizing an 80:20 ratio. This allowed us to ensure a minimum gap of 4 h existed between end time of each training data vector, and start time of each testing data vector, limiting correlations between the two (Fig. 1B). For prediction time horizons <240 min, we ensured that every hour of the day is represented in both the training and test data. Because of data limitations when utilizing a large window, this was not achievable for the 240-min prediction time horizon. In this case, care was taken to ensure a wide range of times of day and night are represented in both training and test data. Two independent randomizations of windows into 80/20 splits were performed, models were trained and validated on each split independently, and reported results were averages from both splits. No significant differences were found between the runs.

Table 1. Individual Demographics and Data Overview

Characteristic	*Mean ± standard deviation (range)*
Total number	24
Age, years	44.4 ± 11.8 (21–62)
Weight, kg	42 ± 16 (53–130)
T1D duration, years	21 ± 11 (1–45)
HbA1C, %	7.2 ± 1.1 (5.3–9.7)
Continuous days of data, days	37.8 ± 14.0 (22.9–75.0)
CGM measurements, number	11,534 ± 4377 (6247–22,493)
SMBG measurements, number	209.0 ± 156.6 (20–691)
Insulin boluses, number	253.6 ± 132.0 (112–627)
Meals and snacks, number	229.1 ± 144.6 (66–565)

Statistics for individual demographics and overview of collected measurements. Values are given as mean and standard deviation, along with ranges. Units reported as (number) indicate the total number of measurements.

CGM, continuous glucose monitor; T1D, type 1 diabetes.

Statistical analysis of data quality

Without sufficiently varied data, neural network models were unable to generalize.[14] To evaluate whether we have sufficiently varied data from an individual to train accurate models, we considered the distribution of historic CGM values. By testing goodness of fit of various distributions onto aggregate CGM data, we determined gamma distributions to be the best fit. Gamma distributions are two-parameter, continuous distributions that are widely used to model continuous variables that are always positive and have skewed distributions. In the case of BG values, distributions tend to be skewed toward higher values rather than lower values. Various factors influenced this trend, in particular, the unequal risk of hyper- versus hypoglycemia. Fitting distributions to available data grants us the ability to easily describe data based on the shape and scale parameters, (α,β), of the distribution fit.

We fit gamma distributions, $\Gamma(\alpha,\beta)$, to individual patient's historic data using maximum likelihood estimation in MATLAB.[15] The shape parameter, α, denotes a stretching or shrinking of the distribution and the scale parameter, β,

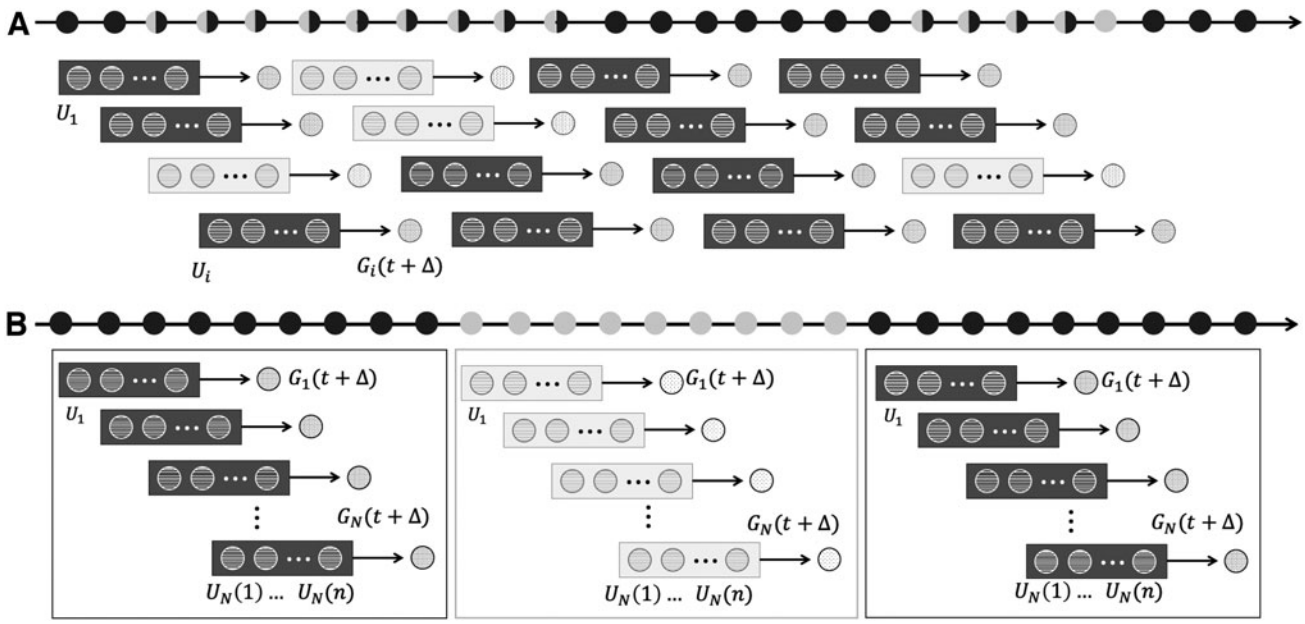

FIG. 1. Graphics depicting the standard random data partitioning method (top) and the window method utilized in this paper (bottom *Ui*). In both graphics, the black trajectory line represents the original time series from which input is obtained and the points are color coded into training (dark) and testing (light) points. Top graphic **(A)** depicts how a random selection of training versus testing results in significant overlap between points used in both training and testing (blue/yellow split color). Bottom graphic **(B)** depicts the window randomization method utilized in this article, which minimizes correlations between training and testing data that may result in falsely high testing accuracy. Rather than randomizing individual input–output vector pairs (U,G), we randomize windows, depicted as outlined boxes. The color-coded boxes depict how points are grouped to fall into larger windows, boxed, and these windows are partitioned using an 80/20 split to be used as either training (dark) or test (light) sets. Within each larger window, we show how input (boxed striped points, U_i) and prediction [single point, dotted, $G_i(t + \Delta)$] points used to train and test the network are contained fully within a box, with no overlap between testing and training. Furthermore, we ensure there is at least a 4 h time gap between the last prediction point within each box, $U_N(n)$, and the first input point of the following box, $U_1(1)$.

denotes a spreading out of the distribution. Specifically, $\alpha = \left(\frac{\mu}{\sigma}\right)^2$, where μ is the mean and σ is the standard deviation of the data. Geometrically, the larger the shape parameter α is, the lower the peak of the distribution is and the further away from zero it is. For the scale parameter we have $\beta = \frac{\sigma^2}{\mu}$, the larger β is, the more spread out the distribution. Metrics on goodness of fit of distributions for all individuals are provided in Supplementary Table S1.

Neural network model structure

To predict BG values, feedforward (artificial) neural networks were utilized. A feedforward neural network model is a black box, *connectionist* modeling approach, loosely based on biological neural networks. These models seek to translate input data to an output prediction through a series of interconnected neurons.[14] Each neuron is associated with a *nonlinear activation function*, and the connections between neurons have associated weights and biases that determine how a given neuron weighs and propagates incoming information. The presented models used the common rectified linear unit function defined as $\sigma(z) = \max(z,0)$ for all neurons.

Although neural network models are well suited for learning predictive models from data, the process of inferring these models can lead to networks learning incorrect causal relationships from correlations in data. Figure 2 illustrates one such pitfall as related to models learned on CGM and insulin pump data: the prevalence of insulin boluses before a meal and the subsequent rise in BG level is potentially mistaken by a naive approach as a causal relationship, thus leading to the outright wrong inference that insulin causes BG levels to increase.

As demonstrated by Narasimhamurthy et al.,[16] resultant models might predict with high accuracy; however, they can be dangerous if used for instance to treat low BG levels. To avoid such pitfalls, we developed a novel, *physiologically informed* neural network design that makes it easier to incorporate background biological facts such as insulin-lowering effect on BG levels. This is performed by (1) constraining training weights to match physiological knowledge (e.g., insulin contributes negatively) and (2) partitioning the first layer into insulin and noninsulin inputs, to mimic learning an ''insulin on board'' term (Fig. 3). Neural network weights are restricted such that insulin has a negative effect, and other inputs provide positive influence, the extent to which is determined in the training phase. A detailed analysis of this type of network structure, in particular how it can be utilized to improve model conformance to known physiology such as ''increased insulin decreases blood glucose values'' can be found in the recent work of Kushner et al.[17]

Transfer learning

For a subset of patients who have limited data, as characterized by the parameters of the gamma distributions fit to historic CGM data, we utilized an adaptive training scheme known as transfer learning to improve accuracy. Transfer

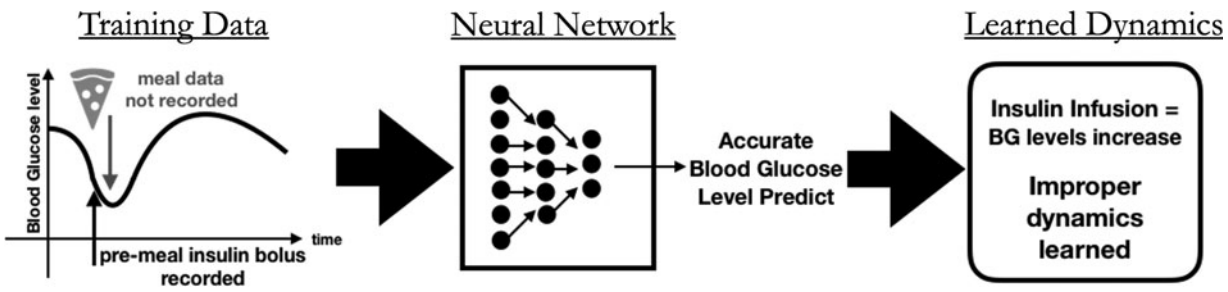

FIG. 2. A common problem with machine learning—neural network models falsely attribute increases in blood glucose levels to previous large doses of insulin, rather than meals, owing to the unavailability of meal data. This results in highly accurate predictive models that are incredibly dangerous to use in advisory systems owing to models learning improper dynamics and taking dangerous dosing decisions: adding insulin when blood glucose is low.

learning is a broad term referring to the technique by which a learned model is reused as a starting point to learn a model for a different, but related, task or is fine tuned to improve accuracy on a subset of the original model. This technique is often used in cases where data are limited, or training from scratch would take a very long time, often leading to increased generalization in the model.[18] In our case, we considered the application of model ''fine-tuning.'' In this transfer learning protocol, the biases and weights are trained on aggregate data from all patients, and then these are used as starting seeds for training the patient-specific model. In this retraining step, only the last layer of weights and biases are adjusted using backpropagation with the individual patient's data. The idea behind this is that the first layers of the network pick out general patterns in BG trends, and the final layer ''fine tunes'' the model for each patient.

Network training

Networks are trained using backpropagation in TensorFlow software.[19] The Adam (adaptive moment estimation) optimizer, which is an extension to stochastic gradient descent that keeps separate learning rates for each weight, is used.[20] Models are trained with 80,000 epochs for all networks. Retraining, which is carried out when adapting a generalized model to a specific individual, is performed to 50,000 epochs. During the training phase, a randomly selected subset of training data is labeled ''internal test set.'' The size of this subset is 15%–20% of the training data, based on training data set size. After every 1000 epochs, the accuracy is tested on this internal test set. The internal test set is used to measure convergence of the training process in terms of prediction accuracy over this internal test set. We note this ''internal test set'' is a subset of training data, rather than selected from the true hold out set that is used for final model validation. The number of epochs was selected to be the minimum number of steps, rounded to the nearest 1000, which consistently resulted in convergence in training accuracy as measured by stabilization of training error for this internal test set.

Accuracy assessment and features selection

Model prediction accuracy is assessed using RMSE and the Clark Error Grid[21] on a separate test dataset, prediction accuracy of the models is compared with a zero hold model, $G(t+\Delta)=G(t)$. RMSE along with heteroskedasticity in error

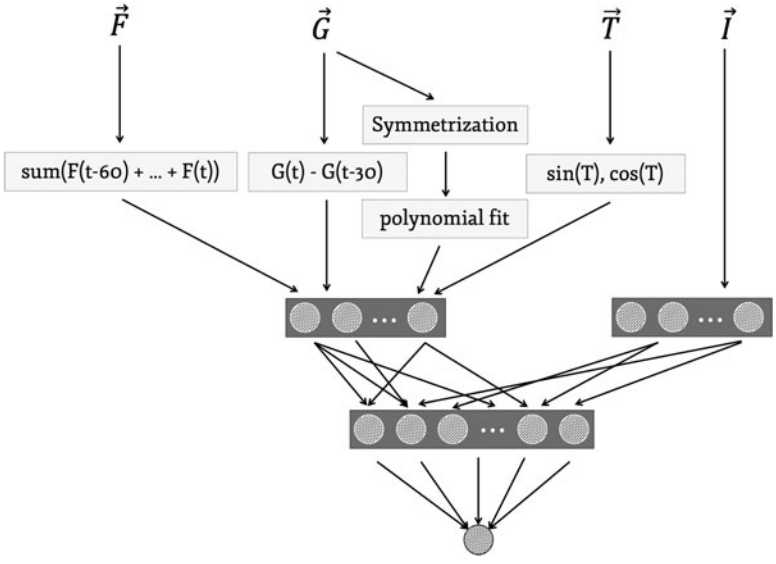

FIG. 3. Schematic illustrating the network structure. Data preprocessing steps are depicted in light gray boxes, and neural network nodes in dark gray. Network input vectors are as follows: $\vec{F}=$ carbs consumed (g), $\vec{G}=$ glucose input from CGM (mg/dL), $\vec{T}=$ time of day (h), $\vec{I}=$ insulin input (units/5 min). The polynomial ax^3+bx^2+cx+d is fit to symmetrized CGM data over the past 60 min, and coefficients a,b,c,d are used as network inputs. Symmetrization is based on the work of Kovatchev et al.[22] Note the first network layer separates insulin inputs from all other inputs, which the second is fully connected. CGM, continuous glucose monitor.

is used to inform features selected as inputs to the network. These inputs are identified systematically by finding patterns in input data that resulted in decreased prediction accuracy, and adding features based on these patterns into the model inputs with the goal of decreasing RMSE and heteroskedasticity. The process is started with the simplest input of raw CGM and insulin historical data and built to the final inputs presented in the Results section.

Results

Average training time for a network was 8 min 6 s and for retraining in the transfer learning approach, average time was 2 min 39 s on a Macbook Pro laptop with 16 GB RAM.

Final network structure

The final network is selected to both minimize error on test data and be parsimonious in the number of neurons and layers. The networks are structured as given in Figure 3, with 16 neurons in the first layer, 16 in the second, and 1 output neuron. The number of hidden layers was selected based on previous work showing a two-layer feedforward network sufficient for use in the BG prediction task.[9] Number of neurons per layer was determined akin to,[9] by training range sizes and selecting that which minimized error when tested on the internal test set during trainings. The neurons that receive input from insulin pump data are fully connected to the second hidden layer, as are the neurons that receive input from noninsulin pump sources. However, the inputs receiving insulin pump data are not connected to neurons receiving noninsulin pump data. It has previously been shown that this structure allows for smaller networks with high prediction accuracy and permits us to appropriately constrain weights for network conformance.[9]

Feature identification for networks

Final inputs to the network were identified systematically, starting with the simplest input of raw CGM and insulin historical data, and building to the final inputs defined hereunder and given in Figure 4:

- a,b,c,d: coefficients from a third order polynomial, $ax^3 + bx^2 + cx + d$, fit to the past 1 h of BG (CGM) data. CGM values are scaled according to the method developed by Kovatchev et al.[22] before fitting. Fit is identified using the polynomial curve fitting function, *polyfit*(),in MATLAB.[15]
- $\boldsymbol{G}(t) - \boldsymbol{G}(t - 30\,\text{min})$: difference in CGM values over the past half hour, in mg/dL.
- $\boldsymbol{h(t)}$: hour of the day input as $\sin\left(\frac{2\pi \text{HR}}{24}\right)$, $\cos\left(\frac{2\pi \text{HR}}{24}\right)$, where HR is the hour of day (0–23) at time t.
- $\text{sum}(\boldsymbol{F}(t - 60\,\text{min}{:}\underline{t}))$: sum of carbohydrate intake for the past 1 h.
- $\boldsymbol{I}(\boldsymbol{t} - 180\,\text{min}),\ldots,\boldsymbol{I(t)}$: raw insulin pump values (units delivered over 5 min), input at 5-min intervals over the past 180 min.

To symmetrize BG values, CGM values were scaled using the method of Kovatchev et al.[22] The third degree polynomial fit to past glucose data was tested as a method for determining key features of past glucose traces, and was found to improve model prediction accuracy over simply using BG

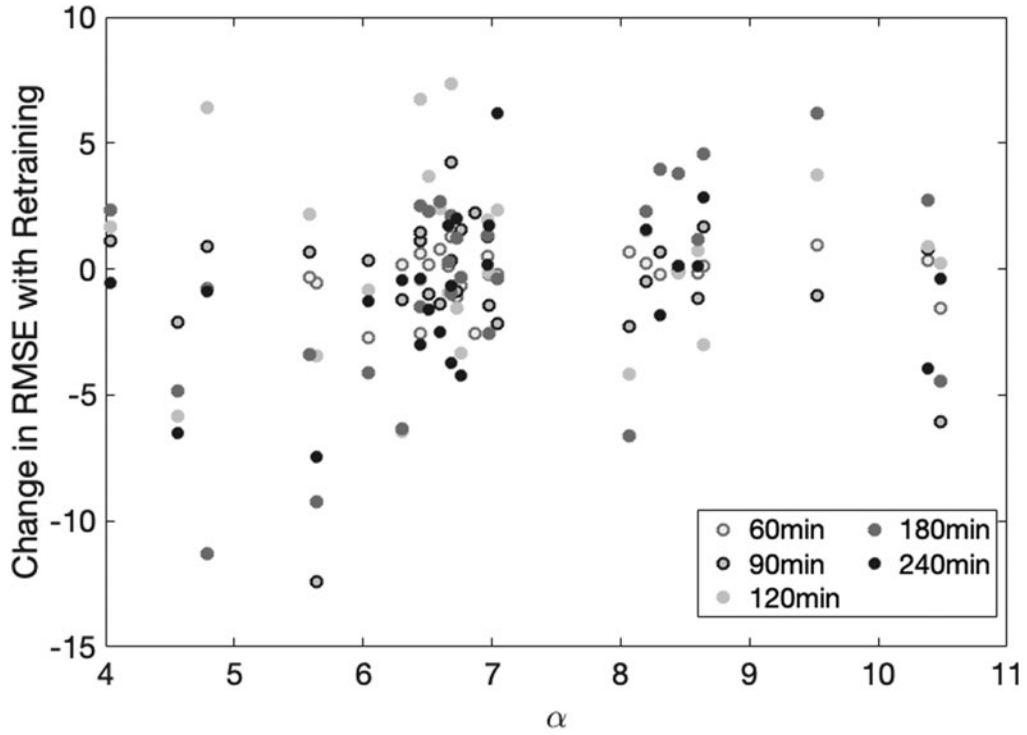

FIG. 4. Change in RMSE by α parameter, for each individual, at each time horizon. We find transfer learning yields on average a decrease in RMSE when $\alpha < 7.6$ while values $\alpha > 7.6$ show an average increase in RMSE, although $\alpha = 6.03$ and $\alpha = 10.08$ significant fluctuations among individuals exist. RMSE, root mean squared error.

Table 2. Accuracy Metrics from Purely Individualized Models

Horizon	RMSE, mg/dL	RMSE, zero hold model, mg/dL	MARD, %	Clark error grid analysis zones				
				Zone A, %	Zone B, %	Zone C, %	Zone D, %	Zone E, %
60	32 ± 6	37 ± 7	17 ± 3	72	24	0	3	0
90	39 ± 7	47 ± 10	21 ± 5	63	31.5	0.5	5	0
120	45 ± 9	53 ± 13	27 ± 7	55	38	1	6	0
180	49 ± 13	65 ± 16	28 ± 7	52	40	1	7	0
240	49 ± 16	69 ± 15	30 ± 9	50	43	2	5	0

Accuracy metrics and percentages of predictions within Clarke error zones of clinical correctness. Both zones A and B are clinically acceptable errors with zone A corresponding to deviations of <20%, or predictions in the hypoglycemic range and zone B corresponding to benign errors. Zones C–E are potentially dangerous, with increasing degree of inaccuracy. The zero hold model comparison is the model $G(t+\Delta) = G(t)$. Results are presented when transfer learning is not included.
RMSE, root mean squared error.

values. Similarly, other features such as the hour of the day, difference in CGM values, and sum of food intake over the past hour were considered subsequently. To determine a feature, we checked correlations in prediction accuracy with features in glucose, insulin, and meal inputs. This method worked particularly well for meal inputs.

As an illustrative example, we show how the carbohydrate term was added after finding that RMSE was notably higher if an individual had consumed carbs over the past hour (with), versus if they had not (without):

RMSE with: 31 mg/dL (zero hold model RMSE 45 mg/dL)
RMSE without: 24 mg/dL (zero hold model RMSE 34 mg/dL)

Building from this discrepancy, we alter the model to include a meal term, sum($\boldsymbol{F}(t - 60\,\text{min}:\underline{t})$), as input. After this term is included, we observed improved mean model accuracy across all individuals when a meal had occurred (with), and found no loss in accuracy when no meal occurred (without):

RMSE with: 27 mg/dL (zero hold model RMSE 45 mg/dL)
RMSE without: 24 mg/dL (zero hold model RMSE 34 mg/dL)

As demonstrated in this meal term example, our input identification method improved robustness of prediction accuracy and also decreased heteroskedasticity in residuals.

Accuracy metrics

Table 2 presents accuracy metrics from purely individualized models forecasting BG values to prediction horizons 60–240 min. For all time horizons, we find at least 93% of prediction to be clinically acceptable for determining treatment strategy by the Clarke error grid (Table 2).[21]

Data characterization and prediction accuracy

We have identified a direct relationship between prediction accuracy and the shape parameter, α, of the distribution in the training data set. This relationship is strongest for prediction horizons of 120 min, and tapers for both longer and shorter term predictions (Table 3). For horizons of 90–180 min, the fit regression is statistically significant with $P < 0.05$ versus the null hypothesis of a constant model (Table 3). For the edge cases of 60 and 240 min, the relationship is slightly weaker ($P = 0.0634$ and $P = 0.133$, respectively). For all practical purposes, we find the significance strong enough such that partitioning based on the shape parameter is beneficial in most cases for these prediction horizons. We note that drop off at the 240-min horizon is likely owing to overall limited data when training these networks.

We have also identified an inverse relationship between the scale parameter, β, of the distribution and prediction accuracy. Similar to the relation with the shape parameter, the strongest correlation is found for prediction horizons of 120 min. However, this relation is weaker than that between the shape parameter and prediction accuracy, and hence the β parameter was not utilized in selection criteria for transfer learning (Table 3).

Overall, this translates to models learned from individuals who have lower average BG values being less generalizable when testing on unseen data. We present further analysis on this finding in the Discussion and Conclusion sections.

Leveraging aggregate data to improve accuracy through transfer learning

By grouping individuals based on parameters of the gamma distributions versus prediction accuracy, we find that on average, individuals with small α parameters, $\alpha < 7.6$, fared worse than those with larger values with most

Table 3. Prediction Accuracy Versus Gamma Distribution Parameters and Improvements in Root Mean Squared Error with Retraining, by Time Horizon

Horizon	Slope (α)	P (α)	Slope (β)	P (β)
60	1.439	0.0634	0.019112	0.95
90	1.651	0.0126	−0.143	0.473
120	2.2559	0.00421	−0.448	0.0609
180	0.9523	0.0269	−0.18339	0.151
240	1.06	0.133	−0.256	0.19

P-values and slopes for regression fit to either shape (α) or scale (β) parameters of gamma distributions on historic CGM values versus percent of predictions within 10%. We find the α parameter, rather than the β parameter to have strongest correlation.

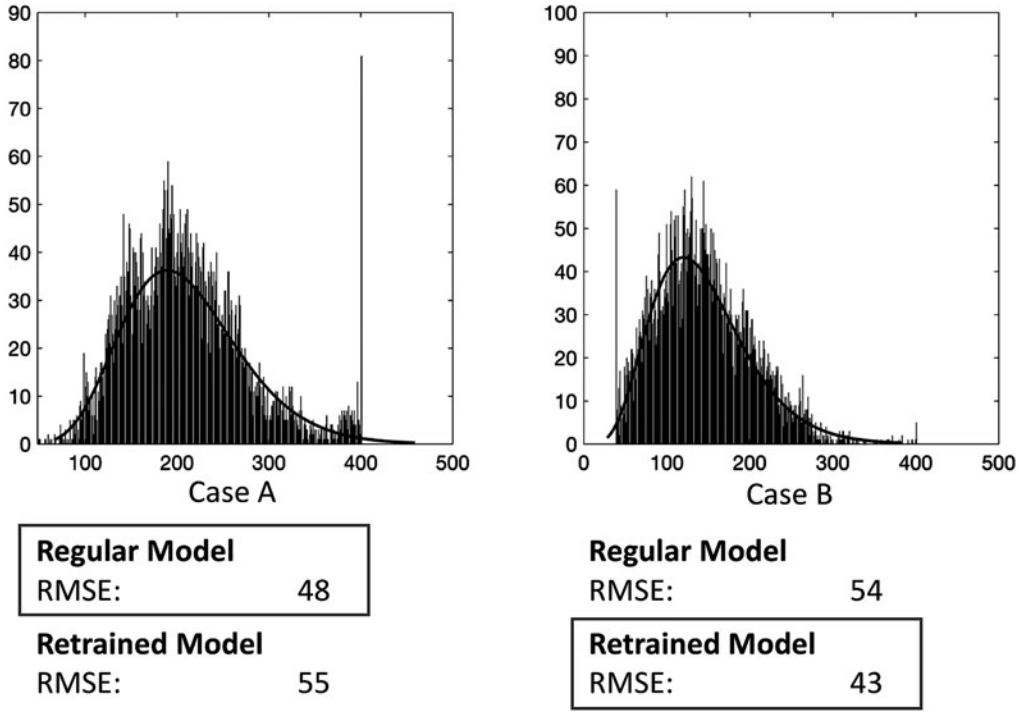

FIG. 5. Case A: example individual who did not benefit from retraining, historic CGM data reflects a wide distribution with $\alpha = 10.08$. Case B: example individual who benefited from retraining, historic CGM values reflect a tighter distribution with $\alpha = 6.03$. We note the spike in values at 40 and 400 are owing to the bounds on CGM output and may not reflect correct blood glucose values.

individuals benefiting with an average reduction in RMSE of across all prediction horizons of −2.6 mg/dL (range: −12.4 to 7.3 mg/dL) compared with an *increase* in RMSE of +0.27 mg/dL (range: −6.6 to 6.2 mg/dL) for $\alpha < 7.6$ (Fig. 4). Whereas the extent to which an individual benefits varies, we find the relationship on the population level strong enough to conclude that it is beneficial to have a larger spread in BG values when training a predictive model.

To improve prediction accuracy for patients in the $\alpha < 7.6$ group, we utilize the transfer learning protocol described in the Methods section. In general, we find this method to be particularly helpful for patients with small α parameters, an example presented in Figure 5. For individuals with larger alpha values, this method more often results in no improvements in prediction accuracy, with a higher likelihood of worsening the prediction accuracy by a few points (Fig. 5). The cutoff point was determined by iteratively testing cutoff points from $5 < \alpha < 10$ to determine which provided highest improvement in RMSE. To have a single value cutoff, 7.6 was selected as the midpoint of 7.0 and 8.1, which performed equally well because of a lack of individuals in the range $7.1 < \alpha < 8.0$. We note that the amount of testing and training data decreases as prediction horizon increases, which could contribute to the discrepancy between benefits gained by various prediction horizons (Table 4).

Discussion

We found that shallow neural network models, with inputs selected based on feature identification, are able to predict future BG values up to 240 min with clinically valuable

Table 4. Accuracy Metrics for Final Models

Horizon	*RMSE, mg/dL*	*RMSE, zero hold model, mg/dL*	*MARD, %*	*Clark error grid analysis zones*				
				Zone A, %	*Zone B, %*	*Zone C, %*	*Zone D, %*	*Zone E, %*
60	28 ± 4	37 ± 7	15 ± 3	75	22	0	3	0
90	33 ± 4	47 ± 10	19 ± 4	68	27	0	5	0
120	38 ± 6	53 ± 13	24 ± 6	58	36	0	6	0
180	40 ± 8	65 ± 16	25 ± 6	53	40	0	7	0
240	43 ± 12	69 ± 15	26 ± 8	52	41	0	7	0

Accuracy metrics and percentages of predictions within Clarke error zones of clinical correctness. Both zones A and B are clinically acceptable errors with zone A corresponding to deviations of <20%, or predictions in the hypoglycemic range and zone B corresponding to benign errors. Zones C–E are potentially dangerous, with increasing degree of inaccuracy. The zero hold model comparison is the model $G(t + \Delta) = G(t)$. Results are presented when transfer learning is included for individuals with $\alpha < 7.6$.

accuracy. In particular we find that in all cases at least 93% of our predictions fall into Clarke error grid zones A and B, meaning that a treatment strategy based on these predictions would be safe (Table 4). Furthermore, when compared with the most recently proposed neural network models for BG prediction at 60-min horizons,[23] we found our models improved both RMSE (28 mg/dL vs. 43 mg/dL) and percent of predictions in zones A and B of the Clarke error grid (97% vs. 96%) while utilizing a significantly less complex model.

On an individual level, we find that model accuracy and best training approach depends heavily on the distribution of an individual's available CGM data.

By fitting gamma distributions to each individual's distribution of historic CGM values, we find a two-group divide based on the shape parameter of the distribution: group (A) individuals with a tighter distribution of CGM values, $\alpha < 7.6$, and group (B) individuals with a more varied distribution, $\alpha > 7.6$. By training two models for each individual, one with a standard training protocol and one with a retraining protocol that leverages aggregate patient data, we find that those in group (A) are more likely to benefit from retraining with added data, whereas those in group (B) are more likely to not benefit. We note a narrower distribution does not imply to be well controlled, as the distribution can center around any value. Although this cutoff point was validated with multiple independent partitions of our original dataset and models trained, and shown to balance improvements in RMSE across all prediction horizons, we note our patient dataset did not include individuals with $7.1 < \alpha < 8.0$, and this methodology run on a different dataset may find adjustments in cutoff point are needed.

We propose two explanations for this trend: an overinfluence of a narrow range of CGM values during training, and a high penalty from outliers during testing. First, the process of training neural networks relies on minimizing error over thousands of subsamples of the training dataset. For individuals with $\alpha < 7.6$, the range of CGM values most often encountered is narrow and centered around the mean leading the model to become very accurate in the narrow range while sacrificing accuracy on outliers. These outliers in turn present large errors during the test period. By first training on aggregate data, the model encounters more outliers and weighs the mean less, resulting in more balanced training and improved accuracy for rarer events, such as hypoglycemic values.

On the contrary, we postulate that for individuals who already have a wide distribution of CGM values, they likely do not benefit from the retraining approach as the added data do not widen the distribution of CGM values seen in the training process and simply serves to reduce the patient specificity of the model.

Finally, we note that while heart rate and step data from activity tracking devices (e.g., Fitbit) was available and tested as potential input, it did not have a significant affect or prediction accuracy. However, despite this finding, we do not rule out the importance of these inputs. Rather, it is hypothesized that this lack of improvement is likely a result of insufficient breadth of training data containing activity tracking information.

Conclusions

Using a physiologically motivated network structure, we were able to utilize shallow neural networks to develop multihour glucose predictive models. These models improve RMSE over previously published work,[5–9,23] and double previously studied prediction horizons[10,11] from 120 to 240 min while removing the dependence on measurements of plasma insulin concentration, instantaneous energy expenditure, and meal-derived rate of glucose appearance, which may be difficult to obtain outside a dedicated research center.

By use of gamma distributions fit to the collected CGM data for an individual, we provide a simple method for quantifying if data of sufficient quality exist for training a patient-specific model. Using this metric, we show that individuals with tighter BG control and lower mean BG values ($\alpha < 7.6$) do not appear to generate optimal data for model development and we developed a transfer learning approach to improve accuracy for these individuals. We find the transfer learning approach to be better suited as a method for model accuracy versus simply boosting or normalizing an individual's training data by duplicating training points that fall in the category of ''rare events,'' as it not only increases the number of rare events, but also increases the variety of dynamics the network is exposed to. Although seemingly counterintuitive, we note this is a result of large errors for ''rare events,'' such as hypo and hyperglycemia, which are not highly represented in training data for such individuals but may exist in test data.

As an individual's behavior and BG control patterns can change over time, in future work we aim to extend our modeling approach to address the question ''how to best select historic data for training.'' This would require a longer data set, on the order of 6+ months, which we did not have for this work.

Overall, our modeling method present, to our knowledge, the most accurate BG forecasts out to 240 min, with at least 93% of predictions considered clinically acceptable by the Clarke error grid analysis, enabling this approach to pave the way for new advisory and closed loop algorithms that would be able to encompass most of the insulin action timeframe.

Acknowledgment

All opinions expressed are those of the authors and not necessarily of the NSF or JDRF.

Author Disclosure Statement

M.D.B. reports personal fees and honoraria from Dexcom, Tandem, Hillo, and Air Liquide; research grants from Sanofi, Tandem, Dexcom, and NovoNordisk; and nonfinancial research support from Novo Nordisk, Dexcom, and Tandem. T.K. has consulted for Tandem. S.S. reports no competing financial interests.

Funding Information

This work has been supported by the U.S. National Science Foundation (NSF) under awards number 1815983 and 1932189, and by the JDRF under award number 1-SRA-2019-818-S-B.

Supplementary Material

Supplementary Table S1

References

1. Swan KL, Dziura JD, Steil GM, et al.: Effect of age of infusion site and type of rapid-acting analog on pharmacodynamic parameters of insulin boluses in youth with type 1 diabetes receiving insulin pump therapy. Diabetes Care 2009;32:240–244.
2. Steil GM, Rebrin K, Darwin C, et al.: Feasibility of automating insulin delivery for the treatment of type 1 diabetes. Diabetes 2006;55:3344–3350.
3. Garg SK, Weinzimer SA, Tamborlane WV, et al.: Glucose outcomes with the in-home use of a hybrid closed-loop insulin delivery system in adolescents and adults with type 1 diabetes. Diabetes Technol Ther 2017;19:1–9.
4. Gondhalekar R, Dassau E, Doyle FJ III: Periodic zone-mpc with asymmetric costs for outpatient- ready safety of an artificial pancreas to treat type 1 diabetes. Automatica 2016;71:237–246.
5. Mhaskar HN, Pereverzyev SV, van der Walt MD: A deep learning approach to diabetic blood glucose prediction. Front Appl Math Stat 2017;3:1–11.
6. Perez-Gandıa C, Facchinetti A, Sparacino G, et al.: Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes Technol Ther 2010;12:81–88.
7. Mougiakakou SG, Prountzou A, Iliopoulou D, et al.: Neural network-based glucose-insulin metabolism models for children with type 1 diabetes. Conf Proc IEEE Eng Med Biol Soc 2006;2006:3545–3548.
8. 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;13:135–141.
9. Dutta S, Kushner T, Sankaranarayanan S: Robust data-driven control of artificial pancreas systems using neural networks. In: Computational Methods in Systems Biology, Proceedings in Lecture Notes in Computer Science, vol. 11095. Switzerland: Springer International Publishing, 2018, pp. 183–202.
10. Georga EI, Protopappas VC, Ardigo D, et al.: Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J Biomed Health Inform 2013;17:71–81.
11. Georga EI, Protopappas VC, Polyzos D, et al.: A predictive model of subcutaneous glucose concentration in type 1 diabetes based on Random Forests. Conf Proc IEEE Eng Med Biol Soc 2012;2012:2889–2892.
12. Breton MD, Patek SD, Lv D, et al.: Continuous glucose monitoring and insulin informed advisory system with automated titration and dosing of insulin reduces glucose variability in type 1 diabetes mellitus. Diabetes Technol Ther 2018;20:531–540.
13. Rolnick D, Veit A, Belongie S, et al.: Deep learning is robust to massive label noise. arXiv Preprint 2017; arXiv:1705.10694.
14. Goodfellow I, Bengio Y, Courville A, et al.: Deep Learning. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, 2016.
15. Matlab Version 2018b. Natick, MA: The MathWorks, Inc., 2018.
16. Narasimhamurthy M, Kushner T, Dutta S, et al.: Verifying conformance of neural network models: invited paper. In: 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). Colorado: ACM/IEEE, 2019. pp. 1–8.
17. Kushner T, Sankaranarayanan S, Breton M: Conformance verification of neural network models for glucose-insulin dynamics. In: Proceedings of the 23rd ACM International Conference on Hybrid Systems: Computation and Control (HSCC). Sydney, Australia: ACM, 2020. pp. 1–12.
18. Pan SJ, Yang Q: A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–1359.
19. Abadi M, Agarwal A, Barham P, et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv Preprint 2016; arxiv:1603.04467.
20. Kingma DP, Ba J: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR). San Diego: ICLR, 2015.
21. Clarke WL: The original Clarke error grid Analysis (EGA). Diabetes Technol Ther 2005;7:776–779.
22. Kovatchev BP, Cox DJ, Gonder-Frederick LA, et al.: Symmetrization of the blood glucose measurement scale and its applications. Diabetes Care 1997;20:1655–1658.
23. Amar Y, Shilo S, Oron T, et al.: Clinically accurate prediction of glucose levels in patients with type 1 diabetes. Diabetes Technol Ther 2020 [Epub ahead of print]. doi: 10.1089/dia.2019.0435.

Address correspondence to:
Taisa Kushner, MS
Department of Computer Science
University of Colorado Boulder
Campus Box 596 UCB
Boulder, CO 80309
USA

E-mail: taisa.kushner@colorado.edu