Clinically Accurate Prediction of Glucose Levels in Patients with Type 1 Diabetes

Clinically accurate prediction of glucose levels in patients with type 1 diabetes

Yotam Amar[1,2]*, Smadar Shilo[1,2,3]*, Tal Oron[4,5]*, Eran Amar[1,2], Moshe Phillip[4,5]†, Eran Segal[1,2]†

Author Affiliations

1. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
2. Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel
3. Pediatric Diabetes unit, Ruth Rappaport Children's Hospital of Haifa, Rambam Healthcare Campus, Haifa, Israel
4. The Jesse Z. and Sara Lea Shafer Institute for Endocrinology and Diabetes, The National Center for Childhood Diabetes, Schneider Children's Medical Center of Israel, Petah Tikva, Israel;
5. Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel.

* Equal contribution

† Corresponding authors

Corresponding authors: Eran Segal ,eran.segal@weizmann.ac.il, Moshe Phillip, mosheph@post.tau.ac.il.

Abstract:

OBJECTIVE. Accurate prediction of blood glucose levels in patients with type 1 diabetes mellitus (T1DM) is critical both for their glycemic control and for the development of reliable closed loop systems. Here, we evaluated different computational methods for glucose prediction, assessed their performance by measures of clinical and numerical accuracy, and developed a novel computational model that optimizes these measures.

RESEARCH DESIGN AND METHODS. We utilized real-life retrospective continuous glucose monitoring (CGM) data from 141 patients with T1DM, totaling 9,083 CGM connection days (1,592,506 glucose measurements) and in silico data generated by the UVA/Padova T1DM Simulator. We compared the clinical accuracy, measured by the percentage of time in each of the Clarke Error Grid (CEG) zones, of predictions done by autoregressive models, tree- based methods, artificial neural networks, and a novel computational model that we devised and optimized for this task.

RESULTS. Our novel model, trained and tested on real-life data, achieved clinical accuracy of 99.3% and 95.8% in predicting glucose level 30 minutes and 60 minutes ahead, respectively, and reduced the percentage of glucose predictions in zones C, D and E of the CEG by 60.6% and 38.4% in these prediction horizons compared to a standard autoregressive model. The model was superior to all other prediction models across all age groups and achieved higher clinical accuracy in subgroups of patients with high glucose variability and greater time spent in hypoglycemia. Compared to real-life data, when evaluated on in silico data, the model had a higher clinical and numerical accuracy

CONCLUSIONS. A model for predicting glucose levels that optimizes for CEG zones may significantly improve clinical accuracy and clinical outcomes of treatment decisions in T1DM patients. Results obtained from simulated data may overestimate the performance of models on real-life data.

INTRODUCTION

Type 1 Diabetes Mellitus (T1DM) is one of the most common chronic diseases in children and adolescents. In the past decade, there is an alarming increase in the incidence of T1DM worldwide, which is more prominent in younger children (1,2). Management of T1DM is a challenge for both patients and caregivers. Although intensive insulin therapy can prevent microvascular complications and cardiovascular morbidity (3), it is associated with higher risk of hypoglycemia, weight gain, and an increased burden of self-management (4). Despite substantial progress in diabetes technologies over the past decades , recent studies demonstrate that clinical management of T1DM is still lacking, particularly in children and adolescents, with many patients not meeting their glycemic goals (6–8).

The development of continuous glucose monitoring (CGM) devices, which measure glucose in the interstitial fluid continuously (5) was an important milestone in diabetes monitoring technology and enabled the development of new treatment strategies such as closed loop systems, also termed "artificial pancreas"(AP) . These systems consist of three components: a CGM, an infusion pump, and a dosing algorithm (9). Personalized predictions of blood glucose (BG) levels are an essential part of AP technologies, as they are crucial for the dosing algorithms. Prediction of glucose levels is a challenging task due to a high intra- and inter-patient variability, and the influence of many factors such as food consumption, insulin dosage, physical activity, and emotional status (10). Most of the control algorithms for AP systems published to date, including Model Predictive Controllers (MPC) (11), Proportional Integral Derivative Controllers (PID) (12) and fuzzy-logic based controllers (13), use naïve, mostly linear, predictors of blood glucose levels, and rely mostly on medical logic, rather than on data-driven approaches. In addition, many of these computational models were constructed and evaluated using in silico data, generated by a T1DM Simulator (14,15), and were not evaluated using real life data.

Motivated by recent developments in the ability of machine learning algorithms to use massive amounts of data in different medical prediction tasks, we evaluated the ability of different data-driven computational models to accurately predict BG levels using both numerical and clinical performance measures. We compared linear models, non-linear tree-based and artificial neural network (ANN) models, as such non-linear models have become state of the art models in many fields in Machine Learning(16),(17). We further devised a novel ANN model optimized for clinical accuracy aiming at improving the clinical outcomes of treatment decisions based on the results of the model (18).

RESEARCH DESIGN AND METHODS

Data acquisition

The database used for training and evaluation of the computational models included retrospective CGM and insulin pumps data of patients with T1DM who visited the National Center for Childhood Diabetes at the Schneider Children's Medical Center of Israel (SCMCI) between December 2015 to December 2018. The inclusion criteria were a diagnosis of T1DM and using a CGM and an insulin pump simultaneously. Patients with insufficient data and pregnant or lactating women were excluded from the analysis. Table 1 depicts the clinical characteristics of the 141 patients included in the study. The average of CGM connection days was 64.4 ± 46.6 per patient, summing up to a total of 9,083 CGM connection days and an overall of 1,592,506 glucose measurements that were included in the analysis.

The study protocol was approved by SCMCI institutional review board. Informed consent was waived by the institutional review board as all identifying details of the patients were removed prior to the computational analysis.

Table1 Baseline characteristics of patients	
Number	141
Gender (F/M)	66/75
Weight (kg)	49.9 ± 14.6
BMI (kg/m^2)	20.5 ± 3.8
Age (years)	13.5 ± 5.2
Duration of diabetes (years)	5.2 ± 4.1
Hemoglobin A1C (%)	7.8 ± 1.0
Celiac (N/Y)	128/13
Thyroid disorder (N/Hashimoto/Graves)	126/12/4
Insulin pump	Medtronic(74), Animas(19), Omnipod(48)
Continuous glucose monitoring	Dexcom(48), Enlite(22), Guardian(24), Libre(43), Navigator(4)

Data preparation

As input for the computational models, we used 4 hours of historical CGM data and insulin dosage. Since some of the CGM devices report a glucose value every 15 minutes, we used linear interpolation to fill the missing BG values. Of note, the output metrics did not include interpolated data. Insulin dosage included both the bolus rate (which was considered as 0 if none) and the basal rate from the insulin pump records. We did not predict the time-window of one-hour post insulin bolus administration since a bolus is a very influential hidden feature that greatly affects our ability to predict correctly. In addition, in order to model how much insulin is needed per meal, a future prediction of what is the expected BG level without such a bolus is required.

In silico data

In addition to the real-life data acquired, we also trained and tested the computational models on data generated by the distributed version of the UVA/Padova T1DM Simulator, which includes in silico data of 30 simulated T1DM patients (10 adults, 10 adolescents, and 10 children) (15) .This simulator was accepted in 2008 by the US Food and Drug Administration (FDA) as a substitute for preclinical trials for insulin treatments, including closed-loop algorithms for AP systems. Each virtual subject in the simulator is represented with subject-specific model parameters. For each virtual subject we generated 30 days of data for training, 7 days for validation and 7 days for testing. For children, 3 meals and 2 snacks were included for each day, while for adults, 3 meals and a single snack. Total carbohydrates consumption per day were calculated according to Dietary reference intake (DRI) recommendations (19). The input for the models was 4 hours of historical CGM data and insulin dosing, comparable to the real-life data.

Machine Learning Models

We evaluated the following machine learning models:

- Auto-Regressive model (AR) (20)
- Tree ensemble using the Random Forest Regressor (RF) implementation of Scikit-Learn (21)
- Gradient Boosting Decision Tree using LightGBM (22)
- Fully connected neural networks using R(23)
- A novel neural network architecture that we constructed, which we term Gradually connected neural network (GCN)

For each model we tried multiple hyperparameters. All the models were tested using 10-fold cross validation.

Gradually connected network

Gradually connected network (GCN) is a novel neural network architecture that we developed, composed of Gradually connected layers (GCL) with fully connected layers on top, rather than using only fully connected layers as commonly used in Artificial neural networks (Figure 1, Figure S1). GCL is similar to a fully connected layer, but with the number of connections to the output neurons gradually increasing with relation to the input order (see 'Supplementary Material'). GCN proved to be very useful for sequential data like CGM data, since it gives more expressive power to more recent data (e.g., the last 30 minutes), which are more informative and more relevant for the prediction task. The GCN models we used contain 4 GCLs with 1-2 fully connected layers on top, summing up to a total of 4-5 hidden layers.

Models optimization and evaluation

We optimized and evaluated the performance of the methods using two performance measures: Numerical accuracy, measured using the root means square error (RMSE) and Clinical accuracy, measured by the percentage of prediction in zones C-E on the Clarke error grid (CEG) (18,24). CEG is a method that quantifies clinical accuracy of predicted BG compared to reference BG, by classifying each pair of predicted and reference BG into five zones (Figure 1, panel B). Predictions that fall in the diagonal of the graph correspond to perfect agreement between the predicted BG and the reference BG, whereas points below and above the line indicate overestimation or underestimation of the actual BG values. While zones A and B indicate sufficiently accurate or acceptable errors in glucose value, zones C-E indicate unacceptable or potentially risky errors that may result in inappropriate treatment and undesired hypo- or hyperglycemia. We therefore defined predictions in zones C-E, as "Clinically hazard zones" (CHZ). When training a neural network model, the use of different loss functions can greatly affect the results of the model. The standard loss functions we used were the mean square error (MSE) and mean absolute percentage error (MAPE). In addition, we designed new loss functions with the goal of minimizing the number of predictions in the CHZ of the CEG and optimizing our model to gain maximal clinical accuracy (defined as predictions in zones A-B). For each prediction, we calculated a standard loss (MSE or MAPE) and multiplied it by a weight according to the CEG zone it corresponded to before running the optimization algorithm. The weight for each zone was deterministically chosen and we examined multiple options of different weights.

To assess statistical significance, we performed a t-test for comparisons between the models, and ANOVA between the performance of the same model on different subgroups.

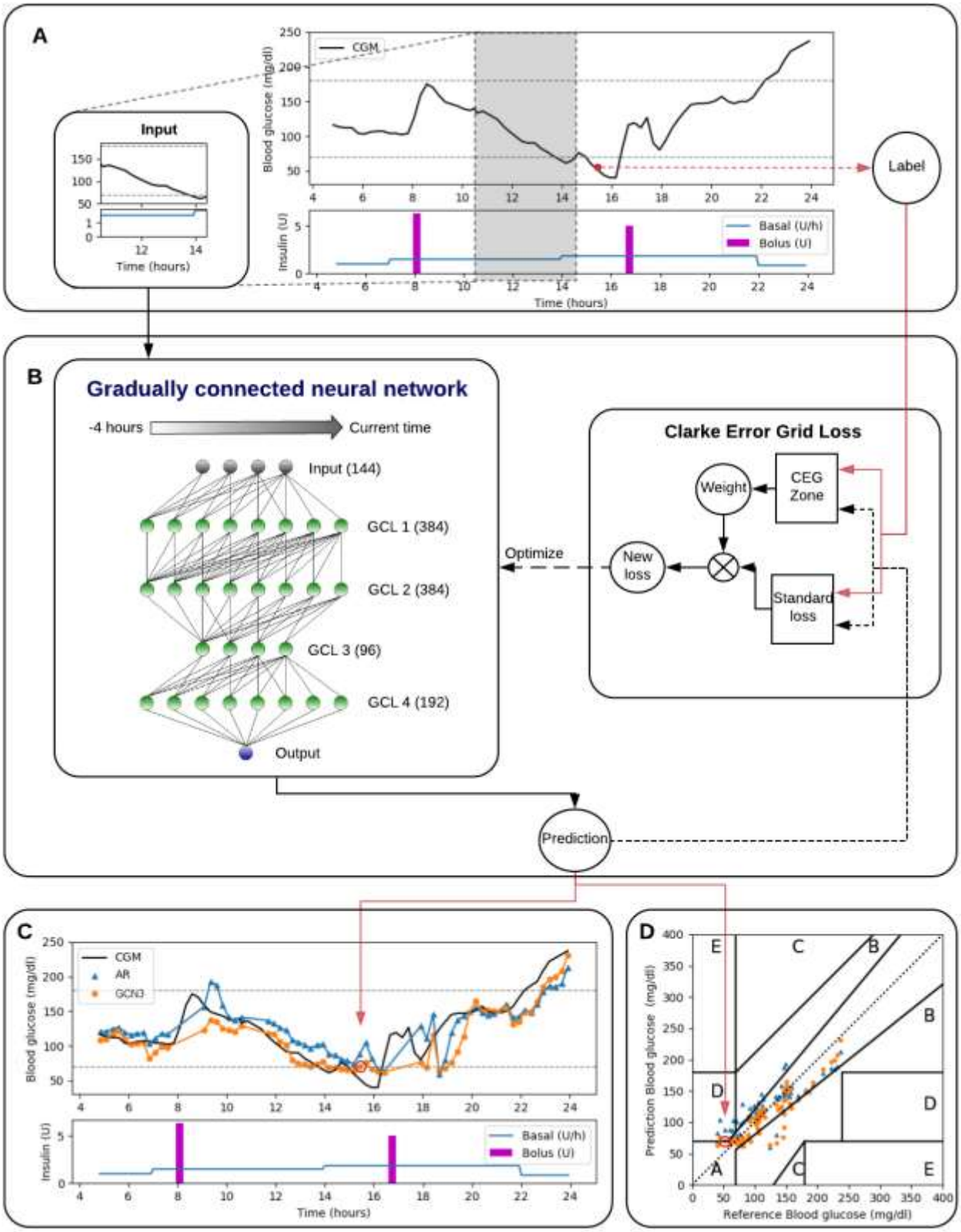

Figure 1: Illustration of a Gradually connected neural network optimized by the clarke error grid analysis, taking as input CGM data of one patient from the cohort and predicting his/her future glucose levels.

(A) Four hours of historical CGM data and insulin dosage are taken from the patient CGM and insulin pump data as input for the model. The dashed line represents a desired glucose range of 70 mg/dl to 180 mg/dl. **(B)** GCN models with 4 GCLs receiving the data as input. The models are optimized using the CEG loss function in which a weight is deterministically chosen for each zone. The layer is described as the layer name, GCL1-4 and the number of neurons in the layer.
(C) The output of the models, glucose predictions for 60 minutes prediction horizon, are presented for both AR and CGN models. The red arrow indicates the point of prediction generated by the GCN model using the 4 hours historical window presented in panel A as input.
(D) Predictions generated by AR and CGN models are presented on the Clarke Error Grid. The red arrow indicates the point of prediction generated by the GCN model using 4 hours of historical data presented in panel A as input.
AR - Autoregressive model, CEG-Clarke Error Grid, CGM - Continuous glucose monitoring, GCL- Gradually connected layer, GCN- Gradually connected neural network.

RESULTS:

Comparing different computational models:

Table 2
Performance of the computational models for 30 and 60 minutes glucose prediction horizons

PH (minutes)	Model	RMSE	C-E (%)	Relative change*
30	AR	23.16 (3.80)	1.87 (1.26)	0.00%
	GCN3	24.27 (3.73)	**0.74 (0.56)**	**-60.64%**
	GCN2	22.82 (3.81)	1.20 (0.83)	-35.64%
	GCN1	**22.06 (3.87)**	2.61 (1.75)	39.29%
	FC	22.85 (3.73)	2.00 (1.32)	6.72%
	LightGBM	22.26 (3.60)	2.97 (1.92)	58.48%
	RF	22.75 (3.72)	2.95 (1.98)	57.55%
60	AR	39.94 (7.24)	6.78 (5.14)	0.00%
	GCN3	42.87 (8.12)	**4.18 (4.26)**	**-38.38%**
	GCN2	39.69 (7.83)	5.46 (4.58)	-19.51%
	GCN1	37.50 (6.94)	7.28 (5.55)	7.26%
	FC	39.16 (6.91)	6.66 (5.06)	-1.80%
	LightGBM	**37.34 (6.53)**	7.64 (5.81)	12.68%
	RF	38.97 (7.04)	7.86 (5.82)	15.83%

*Relative change was calculated by dividing the percentage of prediction in zones C-E on the CEG [C-E (%)] of the model, to the percentage of prediction in these zones when using a baseline AR model.
AR - Autoregression model, CEG -Clarke Error Grid, FC- Fully connected neural network model, GCN- Gradually connected network, GCN1-3 are ensembles of several GCN models that were trained using

different zone weights and standard loss function, LightGBM - LightGBM, Gradient Boosting Decision Tree, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor

---

For each computational method, we trained multiple models using different hyperparameters for both the 30- and 60-minute prediction horizons (PH), and tested our results using 10-fold cross validation. We calculated the RMSE and percentage of time in the CHZ of the CEG for each method and every patient. Table 2 presents the mean and standard deviation of each of these measurements using the best model of each method. We compared the percentage of glucose predictions in CHZ using each of the models compared to the baseline model (expanded results are presented in table S1).

Of note, tree-based methods, RF model and a Gradient Boosting Decision Tree model had better numerical accuracy than AR models. However, both models had lower clinical accuracy than AR, motivating us to try additional computational methods that might be better suited for optimization of clinical accuracy.

We next investigated the Fully connected neural network model (FC) with 2 hidden layers and width 50, which did not improve the numerical accuracy compared to RF and LightGBM but did improve clinical accuracy, leading us to devise the Gradually connected neural network architecture.

The Gradually connected network models (GCN1-3) are ensembles of several GCN models that were trained using different zone weights and standard loss functions. The first model (GCN1) is an ensemble of 2 GCN models, both using only MSE as the loss function (all zone weights are 1). This model improved the RMSE result significantly compared to FC and achieved the lowest RMSE for 30 minutes PH and very similar RMSE result to the LightGBM for the 60 minutes PH. The second model (GCN2) is aimed at optimizing both numerical and clinical accuracy, and is an ensemble of 24 models with 6 unique zone weights, some of which focus on zones C-E of the CEG while others are more balanced and also give a higher weight to zones A-B. This model resulted in an improvement of both clinical and numerical accuracy compared to AR.

The third model (GCN3) is optimized mostly for decreasing the prediction in the CHZ of the CEG, is an ensemble of 6 models which had weights for the different zones of the CEG, and uses both MSE and MAPE for the standard loss. This model resulted in a relatively large improvement in clinical accuracy compared to AR, reducing the average percentage of predictions in zones C-E of the CEG relative to AR by 60% for 30 minutes PH and by 38% for 60 minutes PH. Of note, while GCN3 managed to significantly increase the clinical accuracy, it had the lowest numerical accuracy compared to all the other computational methods in both prediction horizons.

Comparison of our model and the Autoregressive model

To further investigate the performance of our model, we analysed several performance measures for 60 minutes glucose PH on different subgroups of patients. These analyses are presented in Figure 2 (expended results are presented in Table S4 and S5). An overall analysis showed that our model decreased the percentage of predictions in CHZ of the CEG significantly ($p < 10^{-11}$). Examining the individual level, the decrease was apparent in the majority of the cohort (123 of 141, 87%; (Fig. 2 A). Notably, the percentage of predictions in these zones was not significantly different between different age groups (p= 0.45 in GCN3 and p=0.58 for AR model) (Fig. 2 B).

To examine if our model performs differently on patients with a different degree of glycemic control, we used data on HbA1c and created 3 subgroups consisting of patients with good glycemic control, (defined by an HbA1C< 7%), moderate glycemic control (7%<HbA1C<8%) , and poor glycemic control

(HbA1C> 8%) (25)(Fig. 2 C). The clinical accuracy of our novel model, GCN3, was not affected by HbA1C level (p=0.69) , while the clinical accuracy of the AR model significantly decreased for patients with lower HbA1C level, reflected by an increase percentage of predictions in the CHZ zones of the CEG (p= 0.006) (Fig. 2 C).

Next, we investigated subgroups of patients with different percentage of time spent in hypoglycemia, defined as a glucose level less than 70 mg/dl. We divided our cohort into 3 groups, in which percentage of time spent in hypoglycemia is less than 3%, between 3%-7% and above 7%. In both models there was a higher average percentage of predictions in the CHZ of the CEG when the percentage of time spent in hypoglycemia increased. However, in the high hypoglycemic risk group of patients, our model significantly decreased the average percentage of predictions in the CHZ of the CEG compared to the AR mode by 66.4% for 30 minutes PH (2.91% using AR vs. 0.98% using GCN3) and by 46% for 60 minutes PH (10.95% using AR vs. 5.91% using GCN3 ($p < 10^{-12}$ for both), thus demonstrating higher clinical accuracy for our method in this high risk population. Of note, both models had a small decrease (0.7) in the average RMSE in patients that spend a large percentage of time in hypoglycemia (above 7%). (Fig. 2 E and Fig. 2 G, Table S4).

To study the effect of glucose variability on the performance of our model, we calculated the Coefficient of variation (CV) (26,27) from the CGM measurements of each patient. This value measures blood glucose variability corrected for the mean blood glucose per patient. Next, we divided our cohort into 3 subgroups according to their CV: below 37%, 37%-45 and above 45%. As expected, the average percentage of predictions in the CHZ of the CEG and average RMSE monotonically increased with an increase in the CV using both models, reflecting the challenge of BG prediction in patients with high BG variability. Of note, even in the group of patients with the highest CV (larger than 45%), our method decreased the average percentage of predictions in the CHZ of the CEG by 66.6% for 30 minutes PH (2.82% using AR vs. 0.94% using GCN3) and by 40.9% for 60 minutes PH (11.63% using AR vs. 6.87% using GCN3) ($p < 10^{-8}$ for both), reinforcing the significantly better clinical accuracy of our method in patients with high blood glucose variability (Fig. 2 D and Fig. 2 F).

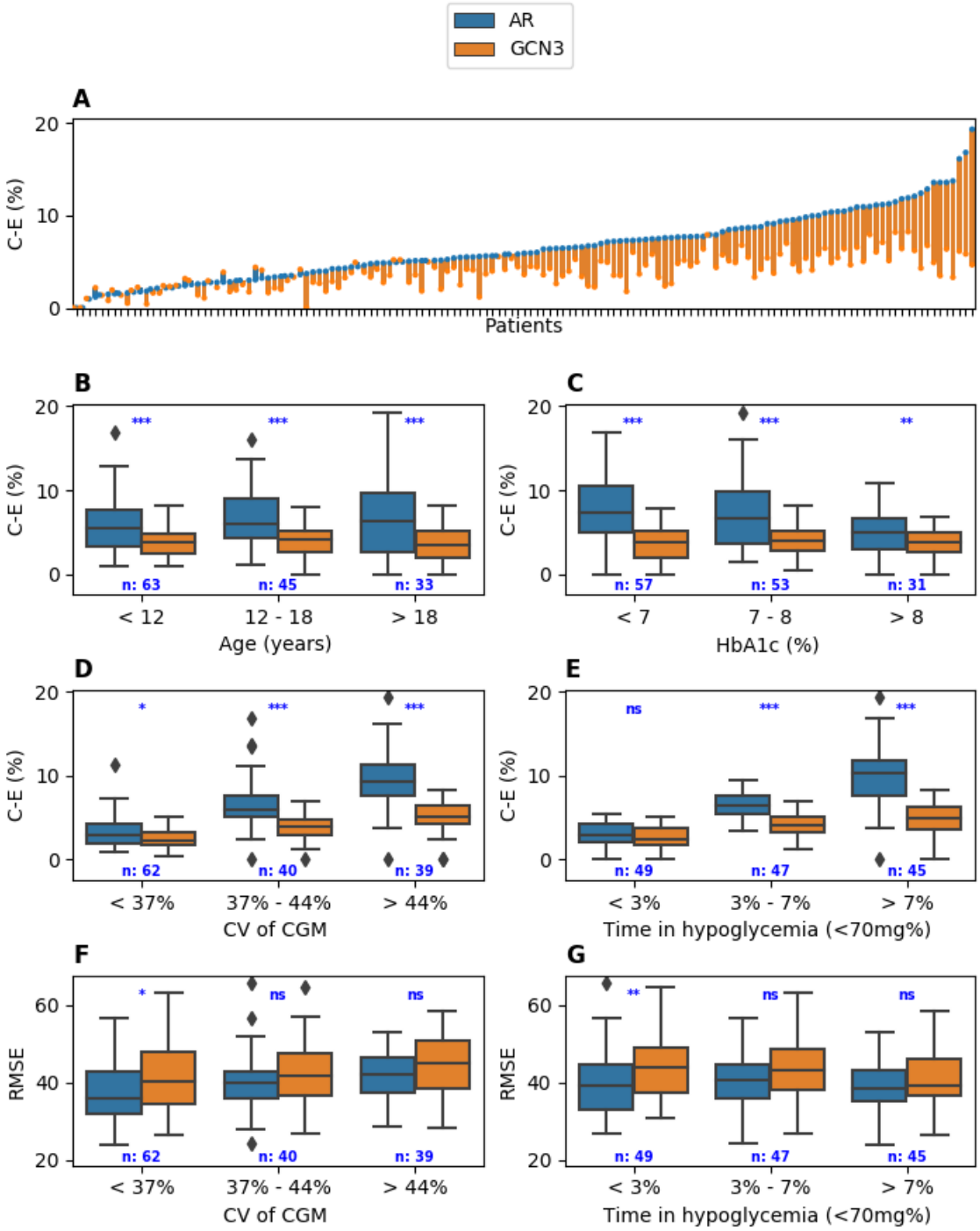

Figure 2: Analysis of the performances of our model (GCN3) compare to a baseline model (AR) on different groups of T1DM patients.

A - The percentage of predictions in zones C-E of the CEG for every patient in our cohort, the patients are sorted by AR C-E (%).

B-E - The percentage of predictions in zones C-E of the CEG for: B. Patients in different age groups. C- Different HbA1c% values. D- Different Coefficient of variation of glucose values. E- Different percent of time in the hypoglycemic range (< 70 mg%).

Significance was calculated using t-test, comparing the two different methods on each subgroup of patients. ns= Non significant, p> 0.05, * p<0.05, **p<0.01, ***p<0.001
F-G - Root means square error for patients with: F- Different Coefficient of variation of glucose values.G- Different percent of time in the hypoglycemic range (< 70 mg%).
AR- Autoregressive model, CV-Coefficient of variation , CEG-Clarke Error Grid, T1DM- Type 1 diabetes mellitus, RMSE-Root means square error

---

Analysis of the models using the Type 1 Diabetes Simulator

Similar to the analysis done on real-life data, we next trained the computational models for two glucose PH, 30 and 60 minutes, using different hyperparameters, on data generated by the UVA/Padova Type 1 Diabetes Simulator (see 'Methods'). We tested our results on 7 generated days for each patient and calculated the RMSE and percentage of time in the CHZ of the CEG for each method per patient. In order to compare the results on simulated data versus real-life data, we trained all the computational models on our real-life dataset of 141 T1DM patients, reserving the last 4 days of each patient for testing (see 'Supplementary').

For both PH, all of the computational models had a much better performance of clinical and numerical accuracy,as reflected by both a lower RMSE and percentage of time in CHZ of the CEG when trained and tested on the simulated data. For example, when trained and tested on real-life data, the highest percentage of prediction in zone A of the CEG for PH 60 was 66.8% on real-life data, and 82.3%, on the simulated data (Table S2 and Table S3).

DISCUSSION:

In this study, we developed a novel computational model using only real-time BG measurements and insulin dosage, that significantly improves blood glucose predictions at 30 and 60 minutes ahead compared to the commonly used AR model. Our model, optimized for clinical accuracy, reduces prediction errors that are considered hazardous and may lead to inappropriate clinical decisions. We trained and tested the models using a real-life dataset of 141 heterogeneous T1DM patients with over 1,592,506 glucose measurements, making it the largest dataset studied to date for this task. Our model is based on BG measurements and insulin dosage which are available for many T1DM patients and can therefore be easily and efficiently implemented in clinical practice. It can run on devices such as smartphones for everyday use, or as a part of a diabetes advisor. Most importantly, our method can be implemented in closed loop systems and replace existing glucose prediction algorithms in order to maximize clinical accuracy.

The computational methods that we analyzed included autoregressive models, tree-based methods, and ANNs. ANNs have certain properties that are ideal for approximating glucose levels, including an unknown non-linear function with multidimensional inputs that varies over time and contains a certain amount of noise. In the last decade, several other studies have tried to harness the power of ANNs for predicting glucose levels in T1DM subjects (28–32). These studies used relatively small data sets and were not optimized for clinical accuracy. In this study, we employed a large data set of T1DM patients with variable clinical characteristics, enabling us to construct robust neural networks that generalize well to unseen data. Our models optimized for clinical accuracy were superior to the more conventional computational models as reflected by an increase in the percentage of predictions in the CHZ of the CEG. Based on our analysis of the clinical and numerical performances of the different computational methods (table 2), we concluded that there is no clear trade-off between achieving a higher percentage of

predictions in the CHZ of the CEG and maximizing the RMSE, leading us to use an ensemble of methods with different optimization for both measures. We therefore created three GCN models, with different optimizations for clinical and numerical accuracies. Our GCN3 model optimized for maximal clinical accuracy, using the zones of the CEG in the loss function, achieved clinical accuracy of 99.3% and 95.8% in predicting glucose level 30 minutes and 60 minutes ahead respectively, and reduced the percentage of glucose predictions in the CHZ zones of the CEG by 60.6% and 38.4% in these prediction horizons compared to the commonly used AR model. Importantly, the clinical accuracy of our models was pronounced in specific clinical settings, in which the current models perform poorly, such as patients at high risk for hypoglycemia or those with increased glucose variability.

Our analysis also demonstrated the necessity of using real-life data for the construction and evaluation of computational models for glucose prediction. By comparing the performance of our model on real-life data to in-silico data, we revealed that achieving high clinical and numerical accuracy is much more challenging when real-life data is used. It is plausible that the simulators are limited in their ability to capture the complexity of glucose fluctuation in T1DM and that their database does not contain unexpected events that typically appear in real-life data. However, as in real life data, our prediction model on the in- silico data performed better than all other methods.

The strengths of our study are our large dataset, which allowed us to use advanced neural networks that cannot be constructed using small datasets, the clinical heterogeneity of our cohort, and the use of real-world data. However, our model has several limitations. It is plausible that incorporation of more clinical inputs effecting glucose homeostasis such as meal content and physical activity, would have resulted in enhanced predictive abilities. Our model is based on retrospective CGM and insulin pump data. As such, we lack information that may add to the accuracy of the model. Using data from CGM devices for prediction models is challenging as they suffer from the existence of a time lag between changes in blood glucose and interstitial fluid (33,34). We believe that measuring blood glucose directly, whether invasive or not, will improve the prediction of the model. Finally, our model was validated using 10-fold cross validation and was not validated on an additional external database. Further prospective studies are needed to evaluate the clinical utility of the model

In conclusion, we present a novel ANN model for the prediction of glucose levels in T1D patients. Our model, optimized for clinical accuracy, was able to predict glucose levels better than the currently used algorithms. Accurate prediction of glucose levels is likely to improve treatment decisions and pave the way for improved outcomes in T1DM patients.

Supplementary Material

Gradually Connected layer

Gradually Connected Layer (GCL) have 2 parameters "output_rows" and "step_size", for input $x \in M_{rows \times columns}$ the output will be y$\in M_{rows \times \frac{columns}{step_size}}$ where each cell in column i of the output is a linear combination of the first $i \cdot step\_size$ columns of the input.

In a fully connected layer for input vector $x = (x_0, x_1 \dots x_n)$ and output vector $y = (y_0, y_1 \dots y_m)$ will have weights matrix $W \in M_{n \times m}$ where: $y = x \cdot W$ whereas in a Gradually Connected Layer (GCL) we force W to be a upper triangular block matrix (Figure S1).

In practice, we implemented it by flattening the input, so the input is a vector of size $(rows \times columns)$, and then multiplying it with a weights matrix is a block upper triangular Matrix $W \in M_{rows \cdot columns \times output\_rows \cdot \frac{columns}{step\_size}}$ where the height of the blocks is $output\_rows$ and the width of the blocks is $(step\_size \cdot columns)$, so $\forall c \in columns, W_{i<c \cdot rows, j<output\_rows \cdot \frac{c}{step\_size}} = 0$.

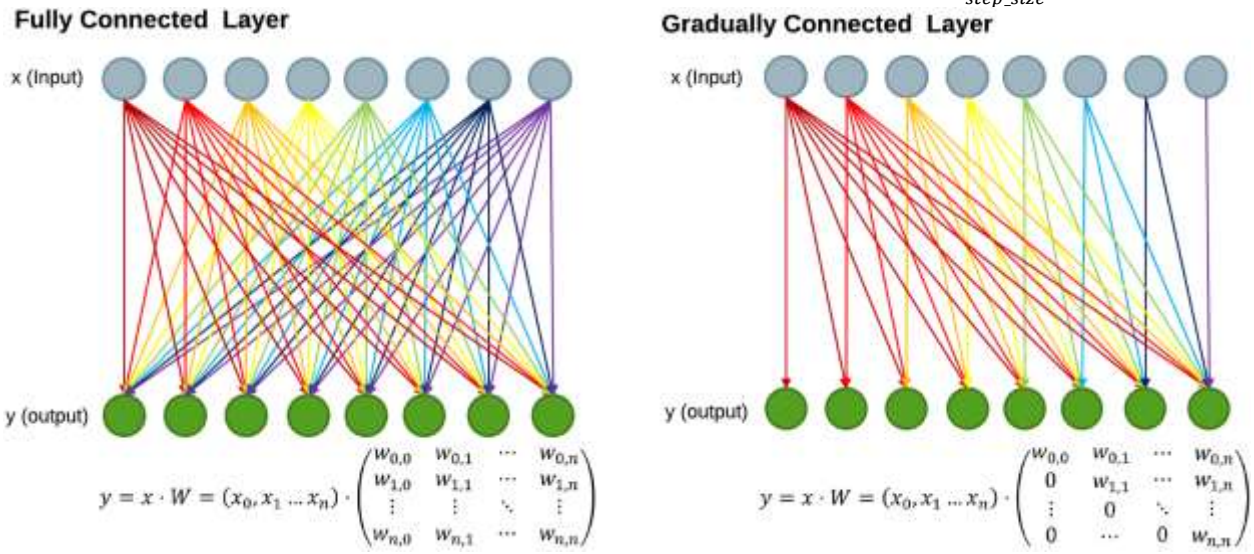

Figure S1: Illustration of a Fully Connected Layer with input and output of size n and the weights matrix $W \in M_{n \times n}$ that defines it (left). Illustration of a Gradually Connected Layer with input of size $n \times 1$, $output\_rows = 1$, $step\_size = 1$ and the upper triangular block weights matrix $W \in M_{n \times n}$ that defines it (right).

Expanded results

Table S1 is an extension of table 2, with the addition of the results of the mean absolute percentage error (MAPE) and the percentages of the predictions in each of the clarke error grid zones using the different computational methods

Table S1
Clinical and numerical accuracy of the computational model (10- fold cross validation)

PH (minutes)	Model	RMSE	MAPE	C-E (%)	Relative change*	A	B	C	D	E
30	AR	23.16 ±3.80	12.01 ±3.80	1.87 ±1.26	0.00%	83.64%	14.48%	0.07%	1.80%	0.01%
	gcn3	24.27 ±3.73	12.05 ±3.73	**0.74 ±0.56**	**-60.64%**	82.60%	16.66%	0.03%	0.69%	0.02%
	gcn2	22.82 ±3.81	**11.24 ±3.81**	1.20 ±0.83	-35.64%	**84.73%**	14.07%	0.04%	1.16%	0.01%
	gcn1	**22.06 ±3.87**	11.57 ±3.87	2.61 ±1.75	39.29%	84.51%	12.88%	0.05%	2.55%	0.01%
	FC	22.85 ±3.73	11.93 ±3.73	2.00 ±1.32	6.72%	83.72%	14.29%	0.05%	1.94%	0.01%
	LightGBM	22.26 ±3.60	11.84 ±3.60	2.97 ±1.92	58.48%	84.07%	12.97%	0.03%	2.93%	0%
	RF	22.75 ±3.72	12.05 ±3.72	2.95 ±1.98	57.55%	83.66%	13.39%	0.05%	2.89%	0%
60	AR	39.94 ±7.24	22.12 ±7.24	6.78 ±5.14	0.00%	63.24%	29.98%	0.61%	5.98%	0.19%
	gcn3	42.87 ±8.12	20.39 ±8.12	**4.18 ±4.26**	**-38.38%**	62.61%	33.21%	0.16%	3.77%	0.26%
	gcn2	39.69 ±7.83	**19.37 ±7.83**	5.46 ±4.58	-19.51%	65.55%	28.99%	0.19%	5.15%	0.11%
	gcn1	37.50 ±6.94	20.88 ±6.94	7.28 ±5.55	7.26%	**65.56%**	27.17%	0.34%	6.87%	0.07%
	FC	39.16 ±6.91	21.78 ±6.91	6.66 ±5.06	-1.80%	63.48%	29.85%	0.45%	6.05%	0.16%
	LightGBM	**37.34 ±6.53**	21.26 ±6.53	7.64 ±5.81	12.68%	64.81%	27.54%	0.32%	7.26%	0.06%
	RF	38.97 ±7.04	22.01 ±7.04	7.86 ±5.82	15.83%	63.66%	28.48%	0.52%	7.25%	0.09%

*Relative change was calculated by dividing the percentage of prediction in zones C-E on the CEG [C-E (%)] of the model, to the percentage of prediction in these zones when using a baseline AR model.
AR - Autoregression model, CEG -Clarke Error Grid, FC- Fully connected neural network model, GCN- Gradually connected network, GCN1-3 are ensembles of several GCN models that were trained using different zone weights and standard loss function, LightGBM - LightGBM, Gradient Boosting Decision Tree, MAPE - mean absolute percentage error, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor

Table S2 presents the results on a models trained using all the real- life data of T1DM patients and tested on the last 4 days of each patient (that were not included in the training).

Table S2
Clinical and numerical accuracy of the computational models (tested on the last 4 days of each patient)

PH (minutes)	Model	RMSE	MAPE	C-E (%)	Relative change*	A	B	C	D	E
30	AR	22.72 ±6.31	11.83 ±6.31	1.97 ±2.06	0.00%	83.86%	14.18%	0.10%	1.85%	0.01%
	gcn3	23.69 ±7.23	11.77 ±7.23	**0.77 ±1.09**	**-61.00%**	84.06%	15.17%	0.05%	0.70%	0.02%
	gcn2	22.29 ±6.82	**11.06 ±6.82**	1.27 ±1.58	-35.60%	**85.38%**	13.35%	0.04%	1.20%	0.02%
	gcn1	**21.49 ±6.44**	11.17 ±6.44	2.30 ±2.63	17.26%	85.23%	12.47%	0.05%	2.24%	0.01%
	FC	22.68 ±6.32	11.50 ±6.32	1.75 ±1.86	-11.12%	84.72%	13.54%	0.04%	1.69%	0.02%
	LightGBM	22.09 ±5.85	11.75 ±5.85	3.05 ±3.00	55.26%	83.77%	13.18%	0.04%	3.00%	0%
	RF	22.20 ±5.98	11.77 ±5.98	2.93 ±3.02	49.32%	83.77%	13.30%	0.07%	2.86%	0%
60	AR	39.13 ±10.43	22.08 ±10.43	7.21 ±9.65	0.00%	62.80%	29.99%	0.68%	6.37%	0.16%
	gcn3	41.31 ±13.14	19.72 ±13.14	**4.06 ±4.00**	**-43.65%**	64.95%	30.99%	0.23%	3.61%	0.23%
	gcn2	38.17 ±11.89	**18.97 ±11.89**	5.71 ±9.18	-20.77%	66.39%	27.90%	0.22%	5.39%	0.10%
	gcn1	**36.23 ±10.82**	20.11 ±10.82	7.48 ±10.08	3.78%	**66.86%**	25.66%	0.35%	7.05%	0.09%
	FC	38.60 ±10.71	21.03 ±10.71	6.68 ±9.37	-7.41%	64.41%	28.91%	0.40%	6.10%	0.17%
	LightGBM	37.03 ±9.64	21.47 ±9.64	8.16 ±10.31	13.19%	63.99%	27.85%	0.38%	7.71%	0.07%
	RF	37.95 ±9.92	21.86 ±9.92	8.24 ±10.29	14.32%	63.31%	28.45%	0.57%	7.62%	0.05%

*Relative change was calculated by dividing the percentage of prediction in zones C-E on the CEG [C-E (%)] of the model, to the percentage of prediction in these zones when using a baseline AR model.
AR - Autoregression model, CEG -Clarke Error Grid, FC- Fully connected neural network model, GCN- Gradually connected network, GCN1-3 are ensembles of several GCN models that were trained using different zone weights and standard loss function, LightGBM - LightGBM, Gradient Boosting Decision Tree, MAPE - mean absolute percentage error, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor

Table S3 presents the results of the computational models on In- silico data generated by the data generated by the distributed version of the UVA/Padova Type 1 Diabetes Simulator, For each virtual subject we generated 30 days of data for training, 7 days for validation and 7 days for testing (see methods). When compared to the results obtained using real-life data (Table S2), these results are much better, especially on 60 min PH, both in percentage of glucose prediction in CHZ and in percentage of glucose prediction in zone A of the CEG.

Table S3
Clinical and numerical accuracy of the computational models on In silico data

PH (minutes)	Model	RMSE	MAPE	C-E (%)	Relative change*	A	B	C	D	E
30	AR	21.42	11.83	1.28	0.00	85.46%	13.26%	0.02%	1.26%	0.01%
	FC	15.68	8.51	0.94	-26.56	93.45%	5.61%	0%	0.93%	0%
	GCN	15.85	8.28	**0.74**	**-42.19**	93.79%	5.47%	0.01%	0.73%	0%
	LightGBM	**15.53**	**7.94**	1.02	-20.31	**94.02%**	4.96%	0.01%	1.01%	0%
	RF	17.64	9.45	1.38	7.81	91.60%	7.01%	0.01%	1.37%	0%
60	AR	37.02	16.59	3.35	0.00	73.55%	23.10%	0.04%	3.26%	0.06%
	FC	**30.09**	14.21	2.88	-14.03	79.77%	17.35%	0.06%	2.79%	0.03%
	GCN	30.46	**13.21**	**2.44**	**-27.16**	**82.31%**	15.25%	0.12%	2.26%	0.06%
	LightGBM	30.39	14.07	3.33	-0.60	80.03%	16.63%	0.06%	3.23%	0.04%
	RF	32.64	15.6	3.8	13.43	76.39%	19.80%	0.04%	3.73%	0.04%

*Relative change was calculated by dividing the percentage of prediction in zones C-E on the CEG [C-E (%)] of the model, to the percentage of prediction in these zones when using a baseline AR model.
AR - Autoregression model, CEG -Clarke Error Grid, FC- Fully connected neural network model, GCN- Single gradually connected network, optimized using MSE loss, LightGBM - LightGBM, Gradient Boosting Decision Tree, MAPE - mean absolute percentage error, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor

Tables S4 and S5 presents the analysis of the performance of the models on different subgroups of patients. Table S4 presents the mean value of each performance measure for each computational model. Significance between the performance of the two models on each subgroup was calculated using t-test.

.

Table S4
Comparison between the computational models on subgroups of patients

PH (minutes)		Subgroups	C-E (%)		Signficance	RMSE		Signficance
			AR	GCN3		AR	GCN3	
30	Age	< 12 y	1.82	0.84	<0.0001	24.35	25.53	0.0701
		12-18 y	1.95	0.71	<0.0001	22.37	23.53	0.0936
		> 18 y	1.86	0.66	<0.0001	22.96	24.05	0.2944
	HbA1c	< 7%	2.34	0.8	<0.0001	23.4	24.48	0.4032
		7%-8%	1.91	0.75	<0.0001	22.94	24.07	0.0352
		> 8%	1.59	0.69	<0.0001	23.21	24.42	0.1115
	CV of CGM	< 37%	0.96	0.46	0.0005	22.18	23.76	0.0581
		37%-45%	1.97	0.78	<0.0001	23.43	24.23	0.2552
		> 45%	2.69	0.96	<0.0001	23.67	24.94	0.1283
	Percentage of time spent in hypoglycemia	< 3%	0.77	0.42	<0.0001	23.58	24.95	0.1066
		3%-7%	2	0.83	<0.0001	23.25	24.3	0.137
		> 7%	2.97	1	<0.0001	22.55	23.56	0.1958
60	Age	< 12 y	5.99	3.89	0.0002	40.38	43.09	0.0375
		12-18 y	6.74	4.01	<0.0001	40.6	43.72	0.0459
		> 18 y	6.63	3.52	0.0005	37.8	40.84	0.1032
	HbA1c	< 7%	7.56	3.66	<0.0001	36.85	37.94	0.6186
		7%-8%	7.11	3.99	<0.0001	39.03	41.92	0.0194
		> 8%	5.17	3.82	0.0012	42.55	46.71	0.0029
	CV of CGM	< 37%	3.32	2.55	0.0399	37.34	41.96	0.0211
		37%-45%	6.62	3.91	<0.0001	40.11	42.43	0.0788
		> 45%	9.49	5.11	<0.0001	42.09	44.41	0.1349
	Percentage of time spent in hypoglycemia	< 3%	3.06	2.63	0.1134	39.91	44.47	0.0099
		3%-7%	6.59	4.1	<0.0001	40.5	43.36	0.0501
		> 7%	10.08	4.93	<0.0001	39.18	40.54	0.3745

Hypoglycemia was defined as glucose level < 70 mg/dl.
AR - Autoregression model, CV - Coefficient of variation, CGM -Continuous glucose monitoring, CEG -Clarke Error Grid, GCN3- Gradually connected network 3, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor.

Table S5 presents the comparison of mean values of each performance measure of the same computational model across different subgroups of patients. Significance was calculated using ANOVA.

Table S5
Comparison of the performance of the same computational model across different subgroups of patients

PH (minutes)		Subgroups	C-E (%) Signficance		RMSE Signficance	
			AR	GCN3	AR	GCN3
30	Age	< 12 y, 12-18 y, > 18 y	0.5464	0.4469	0.1672	0.2526
	HbA1c	< 7%, 7%-8%, > 8%	0.0028	0.7106	0.0009	<0.0001
	CV of CGM	< 37%, 37%-45%, > 45%	<0.0001	<0.0001	0.0121	0.3569
	Percentage of time spent in hypoglycemia	< 3%, 3%-7%, > 7%	<0.0001	<0.0001	0.6834	0.0557
60	Age	< 12 y, 12-18 y, > 18 y	0.8588	0.345	0.0262	0.0204
	HbA1c	< 7%, 7%-8%, > 8%	0.0285	0.6743	0.8491	0.8479
	CV of CGM	< 37%, 37%-45%, > 45%	<0.0001	0.0001	0.1624	0.3717
	Percentage of time spent in hypoglycemia	< 3%, 3%-7%, > 7%	<0.0001	<0.0001	0.4178	0.1962

Hypoglycemia was defined as glucose level < 70 mg/dl.
AR - Autoregression model, CV - Coefficient of variation, CGM -Continuous glucose monitoring, CEG -Clarke Error Grid, GCN3- Gradually connected network 3, PH - prediction horizon , RMSE - Root means square error, RF- Random Forest Regressor.

Article Information

Acknowledgments. We thank members of the Segal lab for useful discussions.

Funding. The study was funded through an EFSD award supported by EFSD/Novo Nordisk. The funder had no influence on the results of the study. E.S. is supported by the Crown Human Genome Center; D. L. Schwarz; J. N. Halpern; L. Steinberg; and grants funded by the European Research Council and the Israel Science Foundation

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions.

Y.A., S.S. and T.O. conceived the project, designed and conducted the analyses, interpreted the results, wrote the manuscript and are listed in random order. E.A. contributed to the analysis M.P. and . E.S. conceived and directed the project and analyses, designed the analyses, interpreted the results and wrote the manuscript. E.S. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References:

1. Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G, EURODIAB Study Group. Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study. *Lancet*. 2009 Jun 13;373(9680):2027–33.
2. Quinn M, Fleischman A, Rosner B, Nigrin DJ, Wolfsdorf JI. Characteristics at diagnosis of type 1 diabetes in children younger than 6 years. *J Pediatr*. 2006 Mar;148(3):366–71.
3. Diabetes Control and Complications Trial Research Group, Nathan DM, Genuth S, Lachin J, Cleary P, Crofford O, et al. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med.* 1993 Sep 30;329(14):977–86.
4. Hypoglycemia in the diabetes control and complications trial. the diabetes control and complications trial research group. *Diabetes*. 1997 Feb;46(2):271–86.
5. Facchinetti A. Continuous glucose monitoring sensors: past, present and future algorithmic challenges. *Sensors Basel Sensors*. 2016 Dec 9;16(12).
6. Wood JR, Miller KM, Maahs DM, Beck RW, DiMeglio LA, Libman IM, et al. Most youth with type 1 diabetes in the T1D Exchange Clinic Registry do not meet American Diabetes Association or International Society for Pediatric and Adolescent Diabetes clinical guidelines. *Diabetes Care*. 2013 Jul;36(7):2035–7.
7. Petitti DB, Klingensmith GJ, Bell RA, Andrews JS, Dabelea D, Imperatore G, et al. Glycemic control in youth with diabetes: the SEARCH for diabetes in Youth Study. *J Pediatr*. 2009 Nov;155(5):668-72.
8. Miller KM, Foster NC, Beck RW, Bergenstal RM, DuBose SN, DiMeglio LA, et al. Current state of type 1 diabetes treatment in the U.S.: updated data from the T1D Exchange clinic registry. *Diabetes Care*. 2015 Jun;38(6):971–8.
9. Hovorka R. Closed-loop insulin delivery: from bench to clinical practice. *Nat Rev Endocrinol*. 2011 Feb 22;7(7):385–95.
10. Cobelli C, Renard E, Kovatchev B. Artificial pancreas: past, present, future. *Diabetes*. 2011 Nov;60(11):2672–82.
11. Magni L, Raimondo DM, Bossi L, Man CD, De Nicolao G, Kovatchev B, et al. Model predictive control of type 1 diabetes: an in silico trial. *J Diabetes Sci Technol*. 2007 Nov;1(6):804–12.
12. Doyle FJ, Huyett LM, Lee JB, Zisser HC, Dassau E. Closed-loop artificial pancreas systems: engineering the algorithms. *Diabetes Care*. 2014;37(5):1191–7.
13. Atlas E, Nimri R, Miller S, Grunberg EA, Phillip M. MD-logic artificial pancreas system: a pilot study in adults with type 1 diabetes. *Diabetes Care*. 2010 May;33(5):1072–6.
14. Man CD, Micheletto F, Lv D, Breton M, Kovatchev B, Cobelli C. The UVA/PADOVA type 1 diabetes simulator: new features. *J Diabetes Sci Technol*. 2014 Jan 1;8(1):26–34.
15. Kovatchev BP, Breton M, Man CD, Cobelli C. In silico preclinical trials: a proof of concept in

closed-loop control of type 1 diabetes. J Diabetes Sci Technol. 2009 Jan;3(1):44–55.

16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2012 May 24;60(6):84–90.

17. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*. 2012 Nov 1

18. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care.* 1987 Oct;10(5):622–8.

19. Nutrient Recommendations : Dietary Reference Intakes (DRI) [Internet]. [cited 2019 Mar 14]. Available from: https://ods.od.nih.gov/Health_Information/Dietary_Reference_Intakes.aspx

20. Leal Y, Garcia-Gabin W, Bondia J, Esteve E, Ricart W, Fernández-Real J-M, et al. Real-time glucose estimation algorithm for continuous glucose monitoring using autoregressive models. *J Diabetes Sci Technol*. 2010 Mar 1;4(2):391–403.

21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011;

22. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017;

23. Chollet F. Keras., GitHub. https://github.com/fchollet/keras. 2015;

24. Kovatchev B, Anderson S, Heinemann L, Clarke W. Comparison of the numerical and clinical accuracy of four continuous glucose monitors. *Diabetes Care.* 2008 Jun;31(6):1160–4.

25. American Diabetes Association. 6. Glycemic Targets: Standards of Medical Care in Diabetes-2018. *Diabetes Care.* 2018;41(Suppl 1):S55–64.

26. Suh S, Kim JH. Glycemic variability: how do we measure it and why is it important? *Diabetes Metab J.* 2015 Aug;39(4):273–82.

27. DeVries JH. Glucose variability: where it is important and how to measure it. *Diabetes.* 2013 May;62(5):1405–8.

28. Zecchin C, Facchinetti A, Sparacino G, Cobelli C. Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information. *Comput Methods Programs Biomed.* 2014;113(1):144–52.

29. Pérez-Gandía C, Facchinetti A, Sparacino G, Cobelli C, Gómez EJ, Rigla M, et al. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. *Diabetes Technol Ther. 2010 Jan;12(1):81–8.*

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

31. Pappada SM, Cameron BD, Rosman PM, Bourey RE, Papadimos TJ, Olorunto W, et al. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes Technol Ther.* 2011 Feb;13(2):135–41.

32. Mougiakakou SG, Nikita KS. A neural network approach for insulin regime and dose adjustment in type 1 diabetes. *Diabetes Technol Ther.* 2000;2(3):381–9.

33. Andelin M, Kropff J, Matuleviciene V, Joseph JI, Attvall S, Theodorsson E, et al. Assessing the accuracy of continuous glucose monitoring (CGM) calibrated with capillary values using capillary or venous glucose levels as a reference. *J Diabetes Sci Technol*. 2016 Jul;10(4):876–84.

34. Luijf YM, Mader JK, Doll W, Pieber T, Farret A, Place J, et al. Accuracy and reliability of continuous glucose monitoring systems: a head-to-head comparison. *Diabetes Technol Ther*. 2013 Aug;15(8):722–7.