Quantifying the impact of physical activity on future glucose trends using machine learning

Nichole S. Tyler, Clara Mosquera-Lopez, Gavin M. Young, Joseph El Youssef, Jessica R. Castle, Peter G. Jacobs

PII: S2589-0042(22)00158-4
DOI: https://doi.org/10.1016/j.isci.2022.103888
Reference: ISCI 103888

To appear in: iSCIENCE

Received Date: 24 June 2021
Revised Date: 19 November 2021
Accepted Date: 4 February 2022

Please cite this article as: Tyler, N.S., Mosquera-Lopez, C., Young, G.M., El Youssef, J., Castle, J.R., Jacobs, P.G., Quantifying the impact of physical activity on future glucose trends using machine learning, iSCIENCE (2022), doi: https://doi.org/10.1016/j.isci.2022.103888.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 The Author(s).
Exercising with type 1 diabetes

Low Fitness (n = 9)  High Fitness (n = 11)

In-Clinic Exercise

Exercising with type 1 diabetes

Low Fitness

Same Intensity

High Fitness

Time

Glucose

Exercise

Machine Learning Model

Predicting glucose during exercise improves with adaptation

Predict with population model

Adapting model reduces error in subsequent sessions

Improved model accurately predicts dangerous hypoglycemia

Predicted

Actual

Time

Glucose
Quantifying the impact of physical activity on future glucose trends using machine learning

Nichole S. Tyler¹*, Clara Mosquera-Lopez¹, Gavin M. Young¹, Joseph El Youssef², Jessica R. Castle², Peter G. Jacobs¹**

Affiliations:
¹ Artificial Intelligence for Medical Systems (AIMS) Lab, Department of Biomedical Engineering; Oregon Health & Science University; Portland, OR, 97232; USA
² Harold Schnitzer Diabetes Health Center, Division of Endocrinology; Oregon Health & Science University; Portland, OR, 97239; USA

** Lead Contact: Peter G. Jacobs, Email address: jacobsp@ohsu.edu

* Corresponding Author: Nichole S. Tyler, Email address: tylern@ohsu.edu
Summary (148/150 words)

Prevention of hypoglycemia (glucose <70 mg/dL) during aerobic exercise is a major challenge in type 1 diabetes. Providing predictions of glycemic changes during and following exercise can help people with type 1 diabetes avoid hypoglycemia. A unique dataset representing 320 days, and 50,000+ time points of glycemic measurements was collected in adults with type 1 diabetes who participated in a 4-arm crossover study evaluating insulin-pump therapies, whereby each participant performed 8 identically designed in-clinic exercise studies. We demonstrate that even under highly controlled conditions, there is considerable intra- and inter-participant variability in glucose outcomes during and following exercise. Participants with higher aerobic fitness exhibited significantly lower minimum glucose and steeper glucose declines during exercise. Adaptive, personalized machine learning (ML) algorithms were designed to predict exercise-related glucose changes. These algorithms achieved high accuracy in predicting the minimum glucose and hypoglycemia during and following exercise sessions, for all fitness levels.

Keywords: Type 1 Diabetes, Diabetes, Aerobic Exercise, Decision Support, Artificial Intelligence, Machine Learning, Medicine, Endocrinology
INTRODUCTION

Physical activity has been shown to reduce cardiovascular risk factors in people with type 1 diabetes (Bohn et al., 2015) and regular physical exercise has recently been shown to result in improved time in target glucose range (70-180 mg/dL) (Riddell et al., 2020a). However, exercise is also known to cause substantial changes in glucose. These changes in glucose vary per exercise modality (Bussau et al., 2006; Colberg et al., 2016; Lascar et al., 2014; Moniotte et al., 2017; Reddy et al., 2018; Yardley et al., 2013) and are most dramatic during steady aerobic exercise (Riddell et al., 2020b). There is an increased risk of hypoglycemia during exercise that occurs due to altered muscular uptake of glucose during exercise, and delayed hypoglycemia that can occur on nights following exercise due to changes in insulin-sensitivity (Man et al., 2009; McMahon et al., 2007; Reddy et al., 2019; Wahren, 1977). These dynamic processes underlying glucose uptake are compounded by regular bouts of exercise (Boulé et al., 2005; Steenberg et al., 2019). While regular exercise can improve overall health, avoiding hypoglycemia during exercise is a known challenge for people with type 1 diabetes (Wilson et al., 2020b).

Continuous glucose monitoring technologies (CGM) can provide real-time alerts to the occurrence of hypoglycemia (< 70 mg/dL) or hyperglycemia (> 180 mg/dL) during exercise. And while certain commercial CGM systems like the Dexcom CGM have recently been reported to achieve 13.3% mean absolute relative error (MARE) during aerobic activity (Guillot et al., 2020), use of CGM alone is not sufficient to prevent
hypoglycemia. Commercially available automated insulin delivery (AID) systems have been shown to improve time in glucose target range across real-world daily activities (Brown et al., 2019; Garg et al., 2017), but the exercise modalities of these systems are limited to user-selected modifications to basal insulin and target glucose during announced physical activity (MiniMed 670G System User Guide, Medtronic, 2017; t:slim X2 Insulin Pump with Control-IQ Technology User Guide, Tandem Diabetes Care, 2020) (Wilson et al., 2022). AID algorithms that incorporate real-time physical activity data to prevent hypoglycemia typically reduce automated insulin and, in the case of dual-hormone systems, increase glucagon in anticipation of glucose drops during aerobic exercise (Castle et al., 2018; Jacobs et al., 2016; Wilson et al., 2020a). Furthermore, adaptive AID algorithms that incorporate activity data have been developed to estimate an individual’s plasma insulin and future glucose concentrations for the purpose of personalizing insulin delivery (Hajizadeh et al., 2018a; Hajizadeh et al., 2018b). However, even these systems do not completely eliminate exercise-induced hypoglycemia. Consensus statement guidelines have been developed to help people with type 1 diabetes make decisions regarding modification of insulin dosages and carbohydrate intake prior to and during exercise (Moser et al., 2020; Riddell et al., 2017), but people with type 1 diabetes will oftentimes need to use trial-and-error approaches to learn how to avoid hypoglycemia during exercise. Both automated hormone delivery and decision support systems currently lack the ability to accurately predict exercise-induced changes in glucose. In addition, there can be significant inter and intra-person variability in glucose changes during exercise. Exercise-related glucose changes in people with type 1 diabetes have not yet been precisely quantified
in individuals and across populations when considering different insulin therapies, or baseline fitness levels.

Machine learning is a powerful tool whereby machines are designed to solve problems or perform sophisticated tasks and can even help to make medical decisions, or provide decision support, for diabetes management. Machine learning approaches have been used in disease detection (Li et al., 2018), insulin dose modification through decision support (Tyler and Jacobs, 2020; Tyler et al., 2020), and can be expanded to provide exercise decision support directly to a person living with type 1 diabetes, or to AID systems in order to adjust insulin during physical activity (Reddy et al., 2019; Wilson et al., 2020a). While algorithms that have been designed to predict future glucose exhibit relatively low root mean squared error (RMSE) during non-exercise periods (14.0 mg/dL-18.0 mg/dL) (Mosquera-Lopez and Jacobs, 2021; Pérez-Gandía et al., 2010; Zecchin et al., 2012; Zhu et al., 2020), recent studies have indicated that the accuracy of these algorithms is oftentimes far worse during exercise (46.16 mg/dL) (Hobbs et al., 2019). Machine learning models have already been developed to predict changes in glucose immediately following aerobic exercise (Ben Brahim et al., 2015; Hobbs et al., 2019; Reddy et al., 2019; Romero-Ugalde et al., 2019), and, when integrated with a decision support system, increase the minimum glucose measured during in-clinic exercise sessions (Breton et al., 2018). Still, these algorithms oftentimes have poor accuracy during real-world scenarios (Hobbs et al., 2019), demonstrate large variability in performance between individuals (Xie and Wang, 2020) and have not been evaluated across varying physical fitness levels.
Population machine learning models are trained on a group of people and are designed to provide predictions for all people. Whereas a personalized model learns an individual's unique physiology in order to improve prediction accuracy for an individual. Personalized models can be designed by training machine learning models specifically on an individual's data (Romero-Ugalde et al., 2019), by clustering a number of similar people into groups prior to model training and then training a model on that cluster (Contreras et al., 2017; Montaser et al., 2019), or by adapting a model in real-time using newly observed data in order to improve glucose predictions (Hajizadeh et al., 2018b; Hobbs et al., 2019). It is not yet clear how personalization impacts the prediction accuracy of exercise-related changes in glucose.

Herein we characterize the impact of aerobic exercise on glucose changes using a unique dataset collected during highly controlled, aerobic exercise sessions in adults with type 1 diabetes. Glucose variations are characterized per participant, insulin therapy, and are further explored with respect to baseline physical fitness. Personalized machine learning models were then designed to estimate the minimum glucose during aerobic exercise and four hours following the start of exercise, and to quantify the impact of personalization on model accuracy. We considered three machine learning algorithms, including a multivariate adaptive regression spline (MARS) model (Friedman, 1991), a previously described logistic regression model (Breton et al., 2018), and an autoregressive (AR) model based on a previously described autoregressive model with exogenous inputs (ARX) (Romero-Ugalde et al., 2019). The dataset used to
train and benchmark the approach was collected in a previously published study whereby aerobic exercise was performed 8 times per study participant under identical exercise intensity and duration, meal content and timing conditions, and across multiple diabetes management strategies including automated insulin delivery, automated insulin and glucagon delivery, insulin pump therapy with predictive low-glucose suspend, and standard insulin pump therapy (Castle et al., 2018). The findings obtained from this unique dataset can serve as a benchmark for comparison with other adaptive prediction algorithms, since we anticipate that the repeatability of the changes in glucose will be substantially reduced under free-living exercise conditions compared with these controlled conditions.

RESULTS

Variations in blood glucose dynamics during identically designed exercise scenarios

To evaluate the repeatability of exercise-related glucose changes, participant glucose outcomes were obtained from 20 adults with type 1 diabetes who each performed 8 identically-designed aerobic exercise sessions at 70% VO₂max for 43.2 minutes on average (N = 160 observations). To control for additional variability in glucose trends that can impact exercise-related glucose changes, the in-clinic exercise sessions were designed such that participants consumed a self-selected breakfast at 8 am, daily activities at 10 am, lunch at 12 pm, and performed exercise at 2 pm. Meals of identical nutritional content were consumed at the same time, and aerobic treadmill exercise was performed at the same time for each of the 8 in-clinic visits. Figure 1 shows the variability in the changes in blood glucose during exercise for each participant across
the entire study (Figure 1A) and also organized by insulin therapy (Figure 1B-E). The difference in exercise-related blood glucose changes measured during highly controlled exercise sessions (Figure 1B-E, connecting dashed and solid lines) are reported as the difference averaged across all study arms, per participant in Table 1. Glucose dropped during exercise for nearly every exercise session, and glucose dropped further in the 4 hour period after exercise was concluded for some subjects (Figure 1B-E, circles).

Despite highly repeatable exercise conditions, food intake, and glucose management strategies, there was still substantial intra-participant variability of the change in glucose during exercise across all 8 identical exercise scenarios, ranging across participants from 23.1 mg/dL (participant 13) to 56.4 mg/dL (participant 9) (Table 1). While variability is smaller for some participants when looking at the two exercise sessions performed under a given diabetes management strategy, substantial variability in glucose changes during exercise is still observable for other study participants (Figure 1B-E). The average change in blood glucose during exercise and variability in this change is reported per therapy arm and per participant in Table 1.

Physical fitness impacts changes in glucose observed during physical activity

Baseline aerobic fitness was assessed by VO$_2$max norms for men and women using a rating scale from the American College of Sports Medicine (American College of Sports Medicine’s Complete Guide to Fitness & Health by Barbara Bushman, 2017) that ranks individuals on a scale of very poor, poor, fair, good, excellent, and superior. We found that participants with higher aerobic fitness (rated as good, excellent, and superior VO$_2$max) exhibited significantly lower minimum glucose during aerobic exercise than
those with lower aerobic fitness (rated as very poor, poor, and fair VO\textsubscript{2}max) (average minimum glucose 75.9 mg/dL vs 103.1 mg/dL, p <0.001). Participants with higher aerobic fitness also exhibited lower CGM-measured minimum glucose compared with participants with lower aerobic fitness in the 4-hours following the start of exercise (70.4 mg/dL vs 85.4 mg/dL, p <0.001). And, the higher aerobic fitness participants had significantly steeper glucose drops during exercise (-2.2 mg/dL/min vs -1.8 mg/dL/min, p <0.05) (Figure 2A-C). Participants with higher aerobic fitness exhibited lower glucose values across the in-clinic study days (Figure 2D-E), with significantly lower glucose during activities of daily living when they were physically active (p < 0.05), during the aerobic exercise, and in the overnight period following in-clinic aerobic exercise.

*Population model predictions achieve good prediction accuracy*

Three types of population machine learning models were designed: a MARS model to predict minimum glucose following exercise, a logistic regression model to predict hypoglycemia following exercise, and an AR model to predict CGM values at the end of exercise. Features used to model minimum glucose during and after exercise were extracted from the data collected during each of the in-clinic exercise sessions (N = 160 exercise sessions) and are defined in Table S1. Leave-one-participant-out cross-validation was used during algorithm training to develop generalizable predictive models (Figure S1). Accuracy of the three machine learning models to predict minimum blood glucose at the end of exercise and also CGM-measured minimum glucose during the 4 hours following the start of exercise are reported in Table 2. The population MARS model estimated minimum glucose during exercise with an MAE of 20.0 mg/dL; a
sensitivity of 63%, and an accuracy of 67% to predict hypoglycemia when cross-validated across all 20 participants with each participant left out during the training. The population logistic regression model achieved a sensitivity of 64% and accuracy of 61% in predicting hypoglycemia during exercise when cross-validated on all 20 participants. The population AR model exhibited worse MAE than the MARS model, 23.8 mg/dL, and achieved the highest sensitivity (71%) and accuracy (81%) to predict CGM-measured glucose < 70 mg/dL 40 minutes after the start of exercise, when cross-validated across all 20 participants.

For longer prediction horizons of 4 hours after the start of exercise, the population MARS model exhibited a MAE of 20.1 mg/dL, and a sensitivity of 62% and an accuracy of 56% to detect CGM-measured hypoglycemia when cross-validated across all 20 participants. The results of the logistic regression model to predict hypoglycemia during exercise and 4 hours following the start of exercise were similar both during exercise and 4-hours after exercise. The logistic regression model achieved a sensitivity of 63% and accuracy of 58% to detect CGM-measured hypoglycemia when cross-validated across all 20 participants. The AR model was not designed for the 4-hour predictive window and therefore results are not shown.

Prior exercise-related changes in glucose help to predict future nadir glucose

The benefit of personalization was evaluated by first considering whether the inclusion of participant exercise history, or data collected during previous exercise sessions, can improve accuracy to predict the minimum glucose during exercise. To do this, a second
MARS model was designed that also incorporates participant exercise history features (Table S2). Exercise data features that were found to be predictive of future glucose trends included (1) the participant’s average metabolic expenditure measured during other aerobic exercise sessions, and (2) the average change in glucose measured during other aerobic exercise sessions by the participant. When evaluated on the holdout set, the MARS model that included exercise history reduced MAE by 39%, from 23.4 mg/dL to 14.3 mg/dL, improved sensitivity to predict hypoglycemia during exercise from 50% to 73%, and improved accuracy from 75% to 81% (Table 2). Cross-validation across all 20 participants showed that the inclusion of participants’ exercise history into the MARS model reduced MAE from 20.0 mg/dL to 17.6 mg/dL, improved sensitivity from 63% to 66% to detect hypoglycemia, and improved accuracy from 67% to 70%.

For longer prediction horizons of 4 hours, the MARS model that included exercise history performed similarly to the MARS model that was designed without exercise history, when cross-validated across all 20 participants (Table 2).

Adaptive personalization improves the accuracy of predictive models

The benefit of personalization was also investigated through adaptation of the machine learning models to better predict individual participants’ exercise-related glucose changes. Stochastic gradient descent (An overview of gradient descent optimization algorithms, Ruder, 2016) was used to incorporate the exercise information obtained from a participants exercise session (e.g., data collected during their first study visit) in order to update the population model parameters. The adapted model was then used to
predict the same participant’s outcomes for a separate, held-out exercise session (e.g., their second study visit). This adaptation procedure was repeated for each held-out exercise session, enabling the machine learning model parameters to adapt to an individual’s data over time as more exercise sessions were observed. Personalization of the model coefficients through stochastic gradient descent adaptation improved the accuracy of all of the predictive algorithms (see Table 2) to estimate glucose during exercise and 4 hours after the start of exercise. The improvement from adaptation was not influenced by the order of the observed exercise sessions, and we report the results from the original order prior to shuffling. Gradient descent adaptation of model coefficients reduced the predictive error of the MARS model from an MAE of 20.0 mg/dL to an MAE of 18.1 mg/dL, reduced sensitivity from 63% to 61%, and significantly improved the 20-fold cross-validation accuracy of the MARS model in predicting hypoglycemia during exercise from 67% to 78% (p<0.05). The predictive error per-participant can be seen in Table 1. Adaptation of the logistic regression parameters improved the sensitivity to predict hypoglycemia during exercise from 64% to 68%, and significantly improved the accuracy from 61% to 70% (p<0.05), when cross-validated across all 20 participants. Adaptation of the AR model improved the cross-validation MAE from 23.8 mg/dL to 22.0 mg/dL, and improved the sensitivity to detect hypoglycemia during exercise from 71% to 76% and accuracy from 81% to 83%.

For longer prediction horizons of 4 hours following the start of exercise, adaptation reduced the predictive error and improved the accuracy of all of the models. The personalization through adaptation of the MARS model coefficients significantly reduced
the MAE from 20.1 mg/dL to 18.3 mg/dL, reduced sensitivity from 62% to 56%, and significantly increased the accuracy to predict CGM-measured hypoglycemia 4 hours following exercise from 56% to 68% (p<0.05). The adaptation of the MARS model designed to include prior exercise session metrics reduced the MAE from 21.1 mg/dL to 18.2 mg/dL, reduced sensitivity from 74% to 57% and increased the accuracy to detect CGM-measured hypoglycemia 4 hours following exercise from 57% to 69% when cross-validated across all 20 participants. Adaptation of the logistic regression model increased sensitivity from 63% to 64%, and significantly improved the accuracy from 58% to 70% (p<0.05) to predict CGM-measured hypoglycemia in the 4 hours following exercise when cross-validated across all 20 participants.

Figure 3 shows the Parkes consensus grid of the MARS model cross-validation across all 20 participants in predicting glucose at the end of exercise. Personalization of the population MARS model increased the number of observations in the consensus error grid region A from 110 observations to 115 observations, with no changes in regions C, D, or E. When exercise history was included in the design of the MARS model, adaptation increased the values in region A to 118 observations, with no observations in regions D and E and 99.4% of observations in the combined A + B regions (Figure 3C).

Physical fitness impacts predictive performance

The MARS models performed equivalently for higher fitness vs. lower fitness study participants in terms of mean absolute relative error (Table 3). The AR performed worse for the higher fitness participants than the lower fitness participants. The
accuracy to detect hypoglycemia during exercise, and in the 4 hours following start of exercise, was nominally lower in all machine learning models when evaluated on participants with higher aerobic fitness. Adaptation improved the accuracy to predict hypoglycemia for participants with higher and lower aerobic fitness, and across both prediction horizons (Table 3).

DISCUSSION
Herein we demonstrate that there is substantial variability in glucose changes during aerobic exercise in people with T1D even under highly repeatable food intake and exercise conditions, and that these changes are impacted by baseline physical fitness levels. We also present adaptive glucose-forecasting algorithms and demonstrate how personalization and prior history can improve the accuracy to predict minimum glucose during and following aerobic exercise. To our knowledge, this is the first analysis of exercise-related glucose changes and prediction strategies using an ideal dataset of highly regimented, identical study exercise visits and across multiple insulin therapies.

In the published clinical study data set used to train the proposed predictive algorithms (Castle et al., 2018), the specific variations in glucose during exercise were not presented and a demonstration of differences between individuals with varying aerobic fitness was not presented. The data demonstrate that individuals living with type 1 diabetes will experience considerable variability during exercise, even when exercise occurs in the context of identical meals, exercise intensity and duration, insulin therapy, and scheduled daily activities. For some participants, the magnitude of this variability was diminished when examined within the context of an individual insulin therapy. From
a clinical perspective, this highlights the challenge and uncertainty that individuals face
during aerobic exercise; even if someone could undertake the exact same daily
activities, meals, and exercise practices, there will be differences in their glucose
outcomes during exercise. Part of this variability is explained by insulin therapy and
insulin-on-board, but there are many other factors such as activity level in the days
preceding exercise, and stressors such as sleep quality, illness, or timing of menstrual
cycle that affect insulin sensitivity and glycaemia following exercise. And, baseline
physical fitness can also have a significant impact on glycemic outcomes during
exercise. The high intra- and inter-participant variability in glucose trends during
exercise presents an opportunity for adaptive machine learning approaches to help
people with type 1 diabetes avoid acute and long-term complications related to
hypoglycemia.

The impact of exercise on glucose trends during exercise, and across participants with
varying physical fitness levels, is still an open question (Moser et al., 2020; Yardley and
Sigal, 2021). While an inverse relationship has previously been observed between the
regularity of exercise and the rate of severe hypoglycemia (Bohn et al., 2015), it has also
been reported that participants with higher aerobic fitness exhibit a greater risk of
hypoglycemia (Al Khalifah et al., 2016). We contribute definitive findings that
participants with higher aerobic fitness exhibit significantly steeper glucose trends
during exercise, experienced significantly lower glucose at the end of exercise, and
exhibit nominally lower variability in their glycemic outcomes. This may be due to
physiologic differences; regular exercise impacts muscle fiber content (Yan et al., 2010),
and a single bout of exercise can prime muscle for future glucose uptake (Steenberg et al., 2019). Behavioral differences are also a factor, as participants with higher aerobic fitness may sustain physical activity and metabolic expenditure longer and more consistently than participants with lower aerobic fitness. And although participants with varying aerobic fitness exhibited significantly different glucose outcomes following exercise, personalized metrics such as VO$_2$max and fitness ranking require in-clinic evaluation and are not yet feasible features for incorporation into the design of accessible predictive algorithms. It was also observed that participants with higher aerobic fitness were shown to have significantly lower CGM across the entirety of the 4-arm clinical study; sensor readings for these participants were significantly lower during activities of daily living, exercise, and in the nighttime and 48-hrs following aerobic exercise. This precise knowledge can help to inform new strategies to help people of different fitness levels avoid exercise-related hypoglycemia.

Other groups have presented various methods to predict glucose during exercise. Reddy et al. (Reddy et al., 2019) developed a hypoglycemia prediction algorithm during exercise using a decision tree and random forest algorithm. This random forest model utilized data within first 10 minutes of aerobic exercise to form predictions, and achieved an 86% sensitivity and 87% specificity to hypoglycemia. This approach does not describe adaptation or personalization of models or utilize exercise history. It was also limited in that it required data during the first 10 minutes of exercise to estimate hypoglycemia which makes it impossible for the algorithm to provide automated hormone dosing or decision support prior to the start of exercise. The algorithms
proposed in this manuscript do not use data during the exercise event. The proposed
algorithms were designed for use prior to the start of exercise, for the purpose of
modifying hormone doses and/or carbohydrate intake. The AR model that we evaluated
in this paper was presented originally in Romero-Ugalde et al. as an ARX model, where
the model was designed to predict CGM values at 30 minutes following aerobic stair-
step exercise, and achieved an RMSE of 7.75 mg/dL (Romero-Ugalde et al., 2019). We
repeated the methods described in Romero-Ugalde et al., and while we discovered this
method achieves fair accuracy to predict CGM < 70 mg/dL, we were unable to achieve
the performance that was previously reported. While the AR model, based on the ARX
model described by Romero-Ugalde et al., did not achieve the same predictive error as
the MARS model, the adaptation methods presented herein improved the accuracy of
the AR model to predict CGM < 70 mg/dL and reduced the RMSE. Since the AR model
only included the 0, 10, and 20-minute CGM data points as feature inputs, we explored
whether including the 5 and 15-minute CGM data points would improve the accuracy of
the AR model. However, we found that when including these data points, there was no
statistically significant improvement in the accuracy. This was likely because the CGM
data points at 0, 10, and 20 minutes were smoothed, and so they included information
from the 5 and 15 minute data points. Breton et al. developed a hypoglycemia
prediction algorithm utilizing the contextual physical activity predictors identified by Ben
Brahim et al. (Ben Brahim et al., 2015). The accuracy of this model was not reported
and does not describe personalization (Breton et al., 2018). In the current paper, we
used identical features described by Breton et al. and demonstrated the performance of
the model. We additionally showed that adaptation can significantly improve the
performance in predicting hypoglycemia during exercise. Each of the prior publications
as well as our findings identified the importance of CGM or SMBG measurements at the
start of exercise as a critical predictive feature. The current manuscript extends the
work done previously by emphasizing the importance of personalization and physical
fitness considerations when designing glucose forecasting algorithms during exercise.

Personalization of the population-based machine learning models was shown to
improve the accuracy in almost every model-framework, across both short-term and
long-term prediction horizons, and across all validation scenarios. Adaptation of model
parameters using stochastic gradient descent was shown to significantly improve the
accuracy of detecting hypoglycemia during exercise for the MARS and logistic
regression models. And adaptation of the MARS and AR models improved overall
accuracy of predictions in terms of MAE. Personalization of the MARS framework that
included exercise history as an input feature significantly improved predictive accuracy
to detect hypoglycemia during exercise. The personalized MARS models exhibited
similar RMSE values for both short-term and long-term prediction horizons. This is
likely due to the study design whereby participants were most active during exercise,
and were instructed to rest until dinner. And, for some participants, the nadir glucose
occurred during exercise and was equivalent for both prediction horizons. In real-world
scenarios, predictive RMSE may be higher when people do activities that introduce
variability in glucose in the 4-hour period. Taken together across all of the models and
validation strategies presented in Table 2, personalization resulted in an average
reduction in minimum glucose error estimations by 12.9%, and an average increase in
hypoglycemia prediction accuracy of 21.0%. A strength of the personalization methods presented in this manuscript is the simplicity of the gradient descent approach, which is computationally inexpensive and can be implemented easily in other predictive frameworks with just a few lines of code.

In summary, individuals on insulin pump therapy who perform aerobic exercise under highly regimented, nearly identical conditions and intensities will experience day-to-day variations in exercise-related glucose changes during and following exercise. Baseline physical fitness significantly impacts changes in glucose during exercise. Under these controlled conditions, glucose data at the start of exercise, as well as data from prior exercise sessions are informative of anticipated changes in glucose during future exercise sessions across participants of varying physical fitness levels. And while machine learning models can predict the expected changes in glucose during exercise and can be personalized to provide more accurate predictions, further work is needed to accurately predict hypoglycemia in participants with higher baseline physical fitness. Further studies are forthcoming to determine the performance of our adaptation strategy on at-home exercise session data across participants with varying physical fitness. The scientific community is invited to apply this benchmarking dataset in their research by contacting the lead author for access to the data.

LIMITATIONS

As a limitation, the candidate model structures described here must be compatible with gradient-based optimization procedures, and further evaluation is required before being
implemented in other non-linear model frameworks such as neural networks or decision tree structures. The models in this paper were designed and evaluated on in-clinic exercise data; future studies examining at-home exercise sessions will be required to develop algorithms for real-world use. Our analysis utilized data from 20 participants, and accounts for 320 cumulative days of real-world data, 160 days of which represent in-clinic exercise data, with over 50,000 data time points. While the sample size is small, we propose that this analysis reflects an ideal scenario, and that these results reflect the upper bound of adaptation performance and glucose variability.
ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number 1DP3DK101044-01, F31DK121436 and 1 R01DK120367-01, R01DK1225833-01, R01DK120367-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

N.S.T. has nothing to disclose. P.G.J. and J. R. C. have a financial interest in Pacific Diabetes Technologies, Inc. a company that may have a commercial interest in this type of research. No other potential conflicts of interest relevant to the article were reported.

INCLUSION AND DIVERSITY

This analysis utilized previously published human subjects data. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. We worked to
ensure gender balance in the recruitment of human subjects. The final participant population demographics reflected the population of the city of Portland, OR. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper received support from a program designed to increase minority representation in science.

PRIOR PUBLICATION

Parts of this study were presented as a poster at the American Diabetes Association 78th Scientific Sessions, Orlando, FL, June 22-26 2018
Figure 1. Change in blood glucose measured during identical aerobic exercise sessions.

(A) The change in glucose measured during 8 identical exercise sessions across a 4-arm clinical study. The box plot represents the median and interquartile range of the change in glucose measured during exercise, cross symbols represent outlier values and each whisker extends to the most extreme data point that is not an outlier (n = 158 observations of SMBG data from 20 participants, whereby each participant is represented by 8 SMBG observations, and participant 18 is represented by 6 SMBG observations).

(B-E) the change in glucose measured during aerobic exercise within a given insulin therapy. The black x symbol represents the change in glucose measured during an exercise session, and there are two x symbols per participant per study arm. The line drawn between two black x symbols represents the difference in glucose outcomes measured between the two identically-designed exercise sessions (n = 158 observations of SMBG data from 20 participants across 4 study arms, whereby participants are represented by 2 observations per study arm, and data is not available for participant 18 in the standard of care study arm). The open black circle represents the change glucose measured from the start of exercise, to the minimum glucose measured 4 hours after exercise, and these outcomes are connected by a dotted black line (n = 160 observations of CGM data from 20 participants across 4 study arms, whereby participants are each represented by two observations per study arm).
Figure 2. Differences in glycemic response across baseline physical fitness. Box plots represent the median and interquartile range of the data, cross symbols represent outlier values and each whisker extends to the most extreme data point that is not an outlier. * represents significant differences p<0.05 between boxplot groups as determined by an independent t-test. ** represents significant differences p<0.05 between boxplot groups as determined by Wilcoxon rank-sum test. ○ represents significant differences p<0.05 between sensor glucose as determined by a Wilcoxon rank-sum test.

(A) The slope of glucose during aerobic exercise is significantly steeper in participants with higher aerobic fitness (n = 88 observations collected from 11 participants) than participants with lower aerobic fitness (n = 70 observations collected from 9 participants) (average trend -2.2 mg/dL/min vs -1.8 mg/dL/min, p = 0.03).

(B) The minimum glucose measured during aerobic exercise is significantly lower in participants with higher aerobic fitness (n = 88 observations collected from 11 participants) than in participants with lower aerobic fitness (n = 70 observations collected from 9 participants) (average minimum glucose 75.9 mg/dL vs 103.1 mg/dL, p = 4.7E-9)

(C) The minimum glucose measured by CGM in the 4-hrs following the start of aerobic exercise is significantly lower in participants with higher aerobic fitness (n = 88 observations collected from 11 participants) than in participants with lower aerobic fitness (n = 70 observations collected from 9 participants) and (average minimum glucose 70.4 mg/dL vs 85.4 mg/dL, p = 3.3E-5)
(D) Interquartile range of sensor glucose obtained from participants during in-clinic study days 1 and 4. Participants with higher aerobic fitness exhibit significantly lower glucose during activities of daily living and aerobic exercise, and in the nighttime following exercise (p < 0.05). The lower aerobic fitness group is represented by grey area (n = 72 sensor traces collected from 9 participants). The higher aerobic fitness group is represented by magenta area (n = 88 sensor traces collected from 11 participants). During the in-clinic exercise study visits, activities of daily living were performed starting at 10 am, and exercise at 70% VO$_{2}$max was performed at 2 pm. The number of sensor traces from 9 pm – 12 am is lower for both groups (lower fitness, n = 36, higher fitness, n = 44), representing data only from study day 1, whereas participants exited the clinical study on day 4 and overnight sensor data is therefore not available.

(E) Interquartile range of sensor glucose across the entire 4-day study. The lower aerobic fitness group is represented by grey area (n = 36 sensor traces collected from 9 participants). The higher aerobic fitness group is represented by magenta area (n = 44 sensor traces collected from 11 participants).

Figure 3. Consensus Error Grid for models predicting minimum glucose at the end of exercise. The regions of the consensus error grid indicate the clinical impact of prediction errors. Observations that land in regions A and B indicate safe predictions. Observations that lay in regions C, D, and E may result in clinical errors such as missed hypoglycemia, or false positive hypoglycemia that results in excessive carbohydrate
intake. The percentage of observations falling within each region is listed below each figure.

(A) Population MARS model validation \( (n = 158 \text{ observations of exercise data collected from 20 participants}) \) without including prior exercise history.

(B) The MARS model predictions after personalization of population model coefficients \( (n = 158 \text{ observations of exercise data collected from 20 participants}) \).

(C) The predictions of the MARS model that incorporates exercise history features, with additional personalization of the model coefficients \( (n = 158 \text{ observations of exercise data collected from 20 participants}) \).
Table 1 Changes in glucose during exercise, and results of the best performing ML models to predict minimum glucose in the 4 hours following exercise.  a indicates participants with higher aerobic fitness.  b indicates the performance of the model designed to predict minimum glucose at the end of exercise, specifically, the MARS model designed with exercise history and adaptive personalization.  c indicates the performance of the model designed to predict minimum glucose at the end of exercise, specifically, the MARS model designed with exercise history and adaptive personalization.
minimum glucose within 4-hours following the start of exercise, specifically, the MARS model that underwent adaptive personalization. Participant 18 SMBG data was not available for the standard-care arm, and is not reported.
<table>
<thead>
<tr>
<th>Predicting Minimum Glucose at the end of Exercise</th>
<th>Population Model</th>
<th>Personalized Model, Coefficient Adaptation</th>
<th>Comparison between population model and personalized model</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (MAE) [mg/dL] [Sensitivity, Specificity] (Accuracy) [%]</td>
<td>RMSE (MAE) [mg/dL] [Sensitivity, Specificity] (Accuracy) [%]</td>
<td>Δ MAE [%]</td>
<td>Δ Accuracy [%]</td>
</tr>
<tr>
<td><strong>MARS Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>24.1 (19.2) [73, 67] (69)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>26.5 (23.4) [50, 86] (75)</td>
<td>23.1 (19.6) [70, 86] (81)</td>
<td>-16.2</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>24.6 (20.0) [63, 63] (67)</td>
<td>23.0 (18.1) [61, 78] (78)</td>
<td>-9.5</td>
</tr>
<tr>
<td><strong>MARS Model + Exercise History Features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>23.1 (18.2) [75, 65] (68)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>18.7 (14.3) [73, 86] (81)</td>
<td>19.7 (15.8) [73, 95] (88)</td>
<td>+10.1</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>22.6 (17.6) [66, 69] (70)</td>
<td>22.1 (17.5) [51, 83] (77)</td>
<td>-0.6</td>
</tr>
<tr>
<td><strong>AR Model: Population Model b</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>28.8 (22.7) [71, 94] (83)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>32.8 (28.6) [59, 87] (72)</td>
<td>27.6 (23.3) [59, 87] (72)</td>
<td>-18.7</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>29.6 (23.8) [71, 91] (81)</td>
<td>27.7 (22.0) [76, 90] (83)</td>
<td>-7.4</td>
</tr>
<tr>
<td><strong>Logistic Regression</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>--</td>
<td>[66, 67] (66)</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>--</td>
<td>[73, 76] (75)</td>
<td>--</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>--</td>
<td>[64, 56] (61)</td>
<td>--</td>
</tr>
<tr>
<td><strong>Predicting Minimum Glucose 4 hours after exercise</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MARS Model b</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>25.8 (19.7) [67, 68] (68)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>25.7 (21.6) [18, 76] (56)</td>
<td>21.5 (16.3) [33, 96] (78)</td>
<td>-24.8</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>25.1 (20.1) [62, 51] (56)</td>
<td>23.3 (18.3) [56, 70] (68)</td>
<td>-9.0 *</td>
</tr>
<tr>
<td><strong>MARS Model + Exercise History Features b</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>24.8 (18.6) [79, 61] (69)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>30.7 (26.1) [29, 61] (47)</td>
<td>23.0 (16.0) [56, 96] (84)</td>
<td>-38.8</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>26.3 (21.1) [74, 52] (57)</td>
<td>23.9 (18.2) [57, 70] (69)</td>
<td>-13.8</td>
</tr>
<tr>
<td><strong>Logistic Regression b</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training, 16-fold CV</td>
<td>--</td>
<td>[57, 72] (65)</td>
<td>--</td>
</tr>
<tr>
<td>Validation, Holdout Set</td>
<td>--</td>
<td>[32, 77] (60)</td>
<td>--</td>
</tr>
<tr>
<td>Validation, 20-fold CV</td>
<td>--</td>
<td>[63, 50] (58)</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 2 Comparing the effect of adaptation on the performance of models designed to predict exercise-related changes in glucose. Values represent the mean performance across participants. Training is performed with data from
n = 16 participants, while the holdout set includes data from n = 4 participants. The 20-fold validation includes data from all n = 20 participants. \(^a\) indicates that the significance p < 0.05 determined Wilcoxon signed-rank test for paired, non-parametric data comparing the change in error or accuracy on a per-participant basis. \(^b\) These models return predicted CGM, not SMBG values. The AR model is only designed to predict glucose approximately 43.2 minutes after the start of exercise, and the results for a 4 hour prediction horizon are not shown.
<table>
<thead>
<tr>
<th>Model</th>
<th>Personalization</th>
<th>Accuracy [%]</th>
<th>MARE [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower VO$_2$max</td>
<td>Higher VO$_2$max</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 70 obs</td>
<td>n = 88 obs</td>
</tr>
<tr>
<td>Predicting minimum glucose during exercise</td>
<td>Population</td>
<td>73</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>84</td>
<td>72</td>
</tr>
<tr>
<td>MARS Model</td>
<td>Population</td>
<td>74</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>86</td>
<td>70</td>
</tr>
<tr>
<td>MARS Model + Exercise History</td>
<td>Population</td>
<td>88</td>
<td>75 a</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>85</td>
<td>82</td>
</tr>
<tr>
<td>AR Model</td>
<td>Population</td>
<td>65</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>72</td>
<td>65</td>
</tr>
<tr>
<td>Predicting minimum glucose in the 4 hours following exercise</td>
<td>Population</td>
<td>57</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>70</td>
<td>69</td>
</tr>
<tr>
<td>MARS Model</td>
<td>Population</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>69</td>
<td>68</td>
</tr>
<tr>
<td>MARS Model+ Exercise History</td>
<td>Population</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>72</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 3 Comparing the effect of aerobic fitness on the performance of models designed to predict exercise-related changes in glucose. $^a$ indicates that the significance p < 0.05 as determined by Wilcoxon rank-sum test for non-parametric data, comparing algorithm performance on participants with higher aerobic and lower aerobic fitness rankings.
STAR METHODS

RESOURCE AVAILABILITY

Lead Contact
Further information and requests for data and code should be directed and will be fulfilled by the lead contact, Peter G. Jacobs (jacobsp@ohsu.edu).

Materials Availability
This study did not generate unique reagents or materials.

Data and Code Availability
- De-identified human participant research data used in this analysis was granted for this analysis, and further data sharing is restricted and is not publically available. No standardized data types are reported in this manuscript. Data requests can be submitted to the lead contact. These requests are assessed on a case-by-case basis and require completion and signature of a sharing agreement, as defined by the Oregon Health & Science University Institutional Review Board (OHSU IRB). Summary statistics have been reported in the main manuscript.
- The algorithms designed in this manuscript are listed in the key resources table. The code used to perform the formal analysis of restricted participant data is available from the lead contact upon reasonable request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects and Study Setting
This analysis was performed upon approval of OHSU Institutional Review Board, study number 00019659. This analysis utilized data obtained during a previous clinical study (Castle et al., 2018). The data was collected from 20 adults with type 1 diabetes (N = 20, 14 F, Age 34.5 ± 4.7y, duration diabetes 19.7 ± 8.6 y, BMI 26 ± 5.7, HbA1C 7.5 ± 0.8, VO_{2\text{max}} 37.1 ± 9.6) who participated in a 4-arm study. Each study arm consisted of 4 days of either (1) single-hormone automated insulin therapy, (2) dual-hormone (insulin and glucagon) automated therapy, (3) predictive low glucose suspend CGM-augmented pump therapy, or (4) standard of care CGM-augmented pump therapy.

Participants visited the clinic on days 1 and 4 of each study arm. During in-clinic study visits, participants consumed a self-selected breakfast, lunch and dinner and performed aerobic exercise in the afternoon. Each participant consumed the same meals at the same time and performed the same physical activity at the same time for each of the 8 in-clinic visits (4 arms x 2 days). Participants underwent VO_{2\text{max}} testing using a modified Bruce protocol while wearing a gas-collecting mask. They performed aerobic exercise with graded work intensity every 3 minutes until volitional exhaustion or plateau of oxygen consumption. During the study, aerobic exercise was performed at 70% VO_{2\text{max}} and was designed to last for 40 minutes. Participants sometimes exercised for less than 40 minutes if for example, their glucose dropped below 70 mg/dL. Participants sometimes exercised for longer than 40 minutes if they needed to stop in the middle of exercise for some reason. The average length of exercise across all participants was 43.2 ± 14 minutes. Participant accelerometer and heartrate data were obtained using ZephyrLife BioPatch devices (Zephyr, Annapolis, MD). The automated insulin and glucagon delivery systems were controlled using a custom exercise-aware
algorithm (Jacobs et al., 2015) installed on a Google Nexus smart phone. This automated delivery system wirelessly communicated the t:slim pumps (Tandem, San Diego, CA) and G5 CGM sensors (Dexcom, San Diego, CA) via Bluetooth. During the control arm, participants used their own insulin pumps. The insulin pumps in this study were filled with aspart insulin (Novo Nordisk, Plainsboro, NJ). This secondary analysis utilized participant data obtained from G5 devices, t:slim devices, ZephyrLife BioPatch devices and self-monitored blood glucose (SMBG) Contour Next devices (Bayer, Whippany, NJ).

METHOD DETAILS

Model Input Features and Outcome Measures

The participant data collected by the study devices during each of the in-clinic exercise sessions was processed for predictive exercise features and glucose outcomes following exercise (N = 160 exercise sessions). No observations were excluded from analysis on the basis of artifacts in the time series data, such as noise in CGM data due to calibration or movement, or signal dropout. The input features derived from the clinical data are defined in Table S1. Additional features describing participant exercise history are defined in Table S2. The final input features for each model were determined from Greedy sequential variable selection (Whitney, 1971), or reproduced as described in previous publications (Breton et al., 2018; Romero-Ugalde et al., 2019). The algorithms were trained to predict (1) the minimum glucose from the start of exercise to the end of exercise as measured using self-monitored blood glucose...
(SMBG) or continuous glucose monitor (CGM), and (2) minimum glucose 4 hours following the start of exercise as measured by CGM. SMBG measurements were measured by all participants at the start and end of exercise per study protocol, however SMBG was not always measured in the 4 hour period following exercise therefore CGM is used for the 4-hour prediction model. Participant age, sex, and VO$_2$max were used to classify each participant into categories of higher (including good, excellent, and superior VO$_2$max) aerobic fitness or lower (including very poor, poor and fair VO$_2$max) aerobic fitness, as defined by the American Society of Sports Medicine VO$_2$max aerobic fitness norms (American College of Sports Medicine’s Complete Guide to Fitness & Health by Barbara Bushman, 2017).

**Development of the Population Models**

Three machine learning models were investigated to predict glucose outcomes during aerobic exercise. The first model is a MARS model was designed to predict minimum blood glucose during exercise, and minimum CGM-measured glucose in the 4 hours following exercise. The second model is a logistic regression model designed to predict hypoglycemia during exercise, and in the 4 hours following exercise. The third model developed was an AR model to predict CGM values approximately 43.2 minutes after the start of exercise. To investigate if exercise history is predictive of future glucose trends, a fourth model, a personalized MARS model was designed that incorporates participant exercise history features as inputs to the model (Table S2). Each population model was designed using a training set, which consisted of data from 16 participants. The population machine learning models were trained using leave-one-participant-out
cross-validation, meaning the input features and model parameters were selected using fifteen of the participants in the training set, and then performance was evaluated on the sixteenth held-out participant. The machine learning models were then evaluated on data from a holdout set, which consisted of data from the 4 participants who were not used in the training set. These 4 holdout participants were sampled to ensure that they were representative of the population and had the same frequency of hypoglycemia and minimum glucose as the training set. The general predictive accuracy of the models were also evaluated using a 20-fold leave-one-participant-out cross-validation, where the model parameters were retrained on 19 participants and the model performance evaluated on 1 held-out participant (Figure S1).

**MARS Model to Predict Low SMBG after Exercise**

A MARS model implements a linear regression framework that also considers the numerical range of the predictors. Each input feature (Table S1) was processed into paired hinge-functions, representing the feature values above and below a specific hinge point (i.e., SMBG values above and below a hinge point of 150 mg/dL are considered separate variables with separate model coefficients). Candidate hinge points for a given feature were determined by sorting observations within a feature and selecting every 5th value for efficiency. The optimal hinge points were determined from the set of candidate hinge points during supervised training of the algorithm. Next, *Greedy sequential variable selection (Whitney, 1971)* was used to iteratively identify optimal hinge-functions to predict minimum glucose during exercise. The MARS model coefficients were designed using a weighted regression; this approach places a penalty
on MARS model misestimation of observations with glucose < 70 mg/dL. This essentially minimizes predictive error as well as improves sensitivity and specificity of the algorithm to detect hypoglycemia. The final model structure used to predict the minimum glucose during aerobic exercise is shown in equation (1). The model coefficients (in this case, $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$) along with the hinge points are solved for each short-term and long-term prediction horizon model separately during model training.

$$
\text{Minimum glucose during exercise} = \beta_0 + \beta_1 \times \max(0, CGM_{\text{Start of Exercise}} - 254) \left[ \frac{mg}{dL} \right] \\
+ \beta_2 \times \max(0, CBG_{\text{Start of Exercise}} - 124) \left[ \frac{mg}{dL} \right] \\
+ \beta_3 \times \max(0, HR_{10 \text{ minutes prior to exercise}} - 97.15) [BPM] \\
+ \beta_4 \times CGM_{\text{Trend}_{25 \text{ minutes prior}}} \left[ \frac{mg}{dL/\text{min}} \right]
$$

Equation 1

AR Model to Predict CGM Following Exercise

Romero-Ugalde et al. developed predictive models to forecast CGM measurements during aerobic exercise (Romero-Ugalde et al., 2019). We used the methods and features described by Romero-Ugalde et al. to reproduce the population AR model that utilizes CGM data. In this approach, the CGM data is smoothed using a 1st-order simple moving average, whereby data-at time $t$ is averaged with the preceding data at time $t-5$. 
We found that the AR model using only CGM data and no exogenous inputs performed better than when including exogenous inputs. The AR with exogenous inputs (ARX) described in Romero-Ugalde et al. utilized raw activity data metrics from a different activity sensor than the one used in our study, and this may explain why they got better performance using an ARX model than using an AR model. We present the design and results of the AR model that achieved the highest accuracy during model validation.

The exercise sessions in our dataset lasted on average for 43.2 ±14 minutes, therefore the AR model was designed to predict CGM at approximately 43.2 minutes following the start of exercise. The final model structure is shown below in equation (2) where the coefficients $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are solved for during model training.

$$\text{CGM 40 minutes after the start of exercise}$$

$$= \beta_0 + \beta_1 \text{SmoothCGM}_{\text{Start of Exercise}} \left[ \frac{mg}{dl} \right]$$

$$+ \beta_2 \text{SmoothCGM}_{10 \text{ minutes preceding exercise}} \left[ \frac{mg}{dl} \right]$$

$$+ \beta_3 \text{SmoothCGM}_{20 \text{ minutes preceding exercise}} \left[ \frac{mg}{dl} \right]$$

Equation 2

Logistic Regression to Predict Hypoglycemia

Breton et al. published a logistic regression model to predict hypoglycemia during exercise. We used the identical variables described by Breton et al. (Breton et al., 2018) to train a population logistic regression model to predict the occurrence of hypoglycemia during aerobic exercise and in the 4 hours following exercise. The inputs
to this model were the CGM at the start of exercise, the average CGM trend in the hour preceding exercise, and the ratio of the active insulin (IOB) at the start of exercise to the participant’s total daily insulin requirement (TDIR). The participant TDIR is defined as the total insulin dosed per day on average. The model is shown in equation (3) where the coefficients $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are solved for during model training.

$$\text{Probability of Hypoglycemia}$$

$$= \logit (\beta_0 + \beta_1 CGM_{\text{Start of Exercise}} \frac{mg}{dL})$$

$$+ \beta_2 Average CGM Trend_{\text{Prior Hour}} \frac{mg}{dL \text{ min}}$$

$$+ \beta_3 \frac{IOB_{\text{Start of Exercise}}}{TDIR} \frac{\text{Units}}{\text{Units}}$$

Equation 3

**MARS Model Personalized with Exercise History**

The methods described above were used to create a second personalized MARS model that incorporates exercise history from a given participant. The model was designed by identifying the optimal features included in Table S1, and also exercise history features included in Table S2 that describe participants’ glucose dynamics during prior exercise sessions. The population model to detect minimum CGM-measured glucose during exercise is shown below in equation (4) whereby the coefficients $\beta_0$-$\beta_6$ were solved for each short-term and long-term prediction horizon model separately during training of the model.
Minimum glucose during exercise

\[
\begin{align*}
    &= \beta_0 + \beta_1 \max(0, \text{CGM}_{\text{Start of Exercise}} - 254) \text{[mg/dL]} \\
    &+ \beta_2 \max(0, 254 - \text{CGM}_{\text{Start of Exercise}}) \text{[mg/dL]} \\
    &+ \beta_3 \max(0, \text{HR}_{10 \text{ minutes prior to exercise}} - 97.15) \text{[BPM]} \\
    &+ \beta_4 \max(0, \text{Average } \Delta \text{CGM}_{\text{other exercise sessions}} + 84.92) \text{[mg/dL]} \\
    &+ \beta_5 \max(0, -84.92 - \text{Average } \Delta \text{CGM}_{\text{other exercise sessions}}) \text{[mg/dL]} \\
    &+ \beta_6 \max(0, 5.97 - \text{Average } \text{MET}_{\text{other exercise sessions}}) \text{[MET]} \\
\end{align*}
\]

Equation 4

Real-time Model Adaptation

To determine the impact of adaptation on prediction accuracy, the population model parameters were adapted to each participant left-out of model training using data from the participant’s exercise observations. Stochastic gradient descent (An overview of gradient descent optimization algorithms, Ruder, 2016) was used to update the population model parameters using the participant’s most recent observed exercise session, and the adapted model was then used to predict the same participant’s outcomes of the next exercise session. This adaptation procedure was repeated successively for each held-out exercise observation, updating the population model parameters over time to better reflect a held-out participant’s glucose dynamics as each exercise session was observed. In order to determine if the order of the exercise sessions impacted prediction accuracy, the order of the 8 identical exercise sessions were shuffled four times and the adaptation procedure was repeated.
Statistical parameter details are included here; additional parameters, including the number of observations and the precise statistical tests, are included in the figure and table legends. Significance testing was performed on a per-participant level, df = 19, to compare the change in error and accuracy to detect hypoglycemia before and after personalization. Normality of data was assessed using the Kolmogorov-Smirnov test to determine the appropriate statistical tests. The differences in model error before and after personalization were evaluated using a two-tailed paired t-test for parametric data, and a two-tailed Wilcoxon signed-rank test for non-parametric data, significance level of alpha = 0.05. The differences in glucose outcomes for participants in different physical fitness categories were evaluated using a two-tailed students t-test for parametric data, and a two-tailed Wilcoxon rank-sum test for non-parametric data, significance level alpha = 0.05. Glucose outcomes measured during exercise for each participant was explored with a boxplot, whereby the centerline of the boxplot indicates the median measurement and box edges represent the 25th and 75th percentiles, cross symbols represent outlier values and each whisker extends to the most extreme data point that is not an outlier. Model performance was assessed using root mean squared error (RMSE), mean absolute error (MAE), as well as the sensitivity, specificity and accuracy to detect observations with level 1 hypoglycemia (< 70 mg/dL). Leave-4-participant-out cross-validation was used to create a receiver operating curve for each algorithm to determine the optimal predictive threshold to detect hypoglycemia. The optimal threshold for each algorithm was then used to evaluate algorithm sensitivity, specificity, and accuracy to detect hypoglycemia for left-out participant data (Figure S1).
Parkes consensus error grid analysis (Parkes et al., 2000) was used to determine the clinical impact of the algorithm predictions. Model design and assessment, and statistical analysis were performed in Matlab 2019b (MathWorks, Natick, MA). A power analysis was performed previously for the published clinical study; a study size of 20 participants was sufficient to detect a -3.3% change in % time-in-hypoglycemia and a 16.3% change in % time-in-target glucose (70-180 mg/dL), for >80% power and an alpha = 0.0125 (Castle et al., 2018).
REFERENCES


A  Population model

B  With personalization

C  Exercise history and personalization

Reference glucose (mg/dL)

Predicted glucose (mg/dL)

<table>
<thead>
<tr>
<th>Reference glucose (mg/dL)</th>
<th>Predicted glucose (mg/dL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>250</td>
<td>250</td>
</tr>
</tbody>
</table>

A = 69.6%
B = 29.1%
C = 1.3%
D = 0%
E = 0%
A+B = 98.7%

A = 72.8%
B = 25.9%
C = 1.3%
D = 0%
E = 0%
A+B = 98.7%

A = 74.7%
B = 24.7%
C = 0.6%
D = 0%
E = 0%
A+B = 99.4%
Highlights (3-4 with no more than 85 characters, including spaces)

- People with type 1 diabetes exercised in 8 identically-designed treadmill sessions
- Intra-person glycemic response varies even under controlled and repeated conditions
- Glucose trends downward more quickly in people with higher aerobic fitness
- Adaptive ML algorithms predict exercise-related nadir glucose with high accuracy
<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Participant Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software and Algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matlab 2019b</td>
<td>Mathworks, Natick, MA. <a href="http://www.mathworks.com">www.mathworks.com</a></td>
<td>N/A</td>
</tr>
</tbody>
</table>