Understanding and Predicting the Effect of Environmental Factors on People with Type 2 Diabetes

Kailas Vodrahalli* KAILASV@STANFORD.EDU
Stanford University, USA

Gregory D. Lyng GREGORY.LYNG@OPTUM.COM
Optum AI Labs

Brian L. Hill BRIAN.L.HILL@OPTUM.COM
Optum AI Labs

Kimmo Karkkainen KIMMO.KARKKAINEN@OPTUM.COM
Optum AI Labs

Jeffrey Hertzberg JEFFREY.HERTZBERG@OPTUM.COM
Optum AI Labs

James Zou*Work done at Optum AI Labs JAMESZ@STANFORD.EDU
Stanford University, USA

Eran Halperin ERAN.HALPERIN@UHG.COM
Optum AI Labs

Abstract

Type 2 diabetes mellitus (T2D) affects over 530 million people globally and is often difficult to manage leading to serious health complications. Continuous glucose monitoring (CGM) can help people with T2D to monitor and manage the disease. CGM devices sample an individual's glucose level at frequent intervals enabling sophisticated characterization of an individual's health. In this work, we leverage a large dataset of CGM data (5,447 individuals and 940,663 days of data) paired with health records and activity data to investigate how glucose levels in people with T2D are affected by external factors like weather conditions, extreme weather events, and temporal events including local holidays. We find temperature ($p=2.37 \times 10^{-8}$, n=3561), holidays ($p=2.23 \times 10^{-46}$, n=4079), and weekends ($p=7.64 \times 10^{-124}$, n=5429) each have a significant effect on standard glycemic metrics at a population level. Moreover, we show that we can predict whether an individual will be significantly affected by a (potentially unobserved) external event using only demographic information and a few days of CGM and activity data. Using random forest classifiers, we can predict whether an individual will be more negatively affected than a typical individual with T2D by a given external factor with respect to a given glycemic metric. We find performance (measured as ROC-AUC) is consistently above chance (across classifiers, median ROC-AUC=0.63). Performance is highest for classifiers predicting the effect of time-in-range (median ROC-AUC=0.70). These are important findings because they may enable better patient care management with day-to-day risk assessments based on external factors as well as improve algorithm development by reducing train- and test-time bias due to external factors.

Data and Code Availability This paper uses a proprietary dataset with patient-level information that cannot be released to the public. We summarize the dataset in the paper and the data may be available for research collaborations.

Institutional Review Board (IRB) IRB approval was obtained for the research in this paper through the Office of Human Research Affairs at UnitedHealth Group.

* Work done at Optum AI Labs

1. Introduction

Diabetes mellitus is characterized by issues with the body's ability to regulate blood glucose levels. It currently affects over 530 million people globally (Sun et al., 2022). Moreover, it can lead to numerous health complications including cardiovascular disease, chronic kidney disease, diabetic retinopathy, and diabetic neuropathy resulting in an annual cost of over $327 billion in the US alone in 2017 (Association, 2018). Diabetes can generally be classified into one of two categories – Type 1 diabetes (5-10% of patients (Colberg et al., 2016)), which is characterized by the body's inability to produce insulin (DiMeglio et al., 2018), and Type 2 diabetes (T2D) (90-95% of patients (Colberg et al., 2016)), which results from a heightened resistance to insulin (Nyenwe et al., 2011).

People with diabetes can rely on careful diet management, exercise, medication, and/or insulin doses depending on the type of diabetes and its severity (Colberg et al., 2016; DiMeglio et al., 2018; Nyenwe et al., 2011). People use daily measurements of blood glucose to help manage glucose levels. These measurements are typically infrequent and conducted through needle pricks where a small amount of blood is used to measure the blood glucose level. However, these pointwise measurements offer a limited window into characterizing the ability to regulate glucose levels.

Continuous glucose monitoring (CGM) offers an alternative that measures glucose levels throughout the day enabling more sophisticated analysis and characterization of an individual's glucose management. While CGM is typically recommended for people taking multiple daily doses of insulin and/or at high hypoglycemia risk (ElSayed et al., 2023), prior research has shown CGM data may be useful beyond the current recommended uses with a range of applications including predicting activity and sleep times (Karkkainen et al., 2022), identifying signs of prediabetes and cardiovascular disease (Hall et al., 2018), and providing insulin dose recommendations (Anderson et al., 2016). However, partly due to the difficulty in collecting large clinical datasets, research on CGM data has been limited and there may be many additional applications.

In this paper, we utilize a large dataset of CGM data collected from a cohort of people with T2D over a 2.5-year period from October 2019-May 2022. All people in our cohort are adults residing in the US. In total, we have 12,909 people and 1,859,101 days of CGM data collected. This data is paired with health records including medical claims data, as well as activity data (including heart rate and step count). The data is also paired with datasets containing information on US weather conditions (Menne et al., 2012) and air quality (US Environmental Protection Agency) over the duration of the study. We use the paired data to understand the environmental conditions each person experienced on any given day. We then analyze the resulting dataset to understand how an individual's ability to manage glucose is affected by external factors including weather conditions and temporal events like holidays or weekends.

We find that people at a population level are affected by temperature, holidays, and weekends. This supports previous evidence that temperature can affect blood glucose levels (Kenny et al., 2016) and offers new evidence into how an individual's behavior is characterized by their environment and can affect their ability to manage glucose. We do not observe a significant effect at a population level due to other external factors, though in some cases (e.g., for rare events) we are still limited by amount of data due to the rarity of the events.

Furthermore, we find that using only an individual's medical history and a few days of CGM and activity data can help predict whether an individual's glycemic metrics are likely to be significantly affected by various external factors including temperature, rain, snow, holidays, weekends, and extreme weather events like heat waves. This result is of particular interest as it can enable doctors or health coaches to have early warnings for treating an individual. And importantly, it enables risk prediction without previously observing any of the events in question.

To summarize, our contributions are two-fold: (1) we demonstrate population-level significance for three external factors in inhibiting or supporting glucose management and (2) we develop a method to predict an individual's likelihood of being at risk from the external factors we investigate, further enabling the utility of CGM devices.

1.1. Related Works

Past research has shown CGM data has been shown to be useful in a range of applications. In (Karkkainen et al., 2022), the authors use a U-Net neural network architecture (Ronneberger et al., 2015) to predict various activities including sleep, walking, and exercising using only a CGM recording and medical

identifiers of a patient. CGM data can also be used for early disease prediction and risk calculations, as in (Hall et al., 2018) where the authors use spectral clustering applied to CGM data to identify signs of prediabetes and cardiovascular disease. CGM data has also been used to provide real-time insulin dose recommendations (Anderson et al., 2016) and early prediction of hypoglycemia using logistic regression and random forests applied to hand-selected features applied to CGM data (Dave et al., 2021; Duckworth et al., 2022).

In the medical literature, there has been much research dedicated to finding how external factors can influence various biological processes. Results consistently show that while many external factors adversely affect the general population, people with diabetes are more susceptible to severe consequences. Studies have shown that temperature affects blood glucose absorption, with higher temperature being associated with a decrease blood glucose levels (Kenny et al., 2016). Extreme temperatures, especially during heatwaves, can have significant, adverse consequences; people with diabetes are particularly susceptible to these weather conditions, with heightened risks of hospitalization or death (Xu et al., 2019; Vallianou et al., 2021). Other weather conditions like air pollution can also adversely effect people with diabetes (Vallianou et al., 2021). In addition to weather conditions, research has found seasonal variations due to a mix of diet, holiday seasons, and weather condition variation in biological metrics including fat, cholesterol, and HbA1C values (Ma et al., 2006; Ockene et al., 2004; Higgins et al., 2009).

2. Methods

We leverage an extensive dataset of CGM data that is paired with activity data (heart rate and step count) and medical records including demographic information, disease history, medications, and claims data for each person in our cohort. We also use a dataset sourced from the Global Historical Climatology Network (GCHN) containing historical daily weather data (Menne et al., 2012) and a dataset sourced from the US Environmental Protection Agency (EPA) for historical daily air quality data (US Environmental Protection Agency). Below, we provide details and preprocessing information about each of these datasets respectively, before describing how to join the datasets for our analysis. The analysis pipeline is visualized in Figure 1. A summary of the external events dataset indicating each external factor considered and a range of values for each external factor is provided in Appendix, Table 4.

2.1. CGM and Activity Tracker Data

The people in our cohort are part of an ongoing program to help them manage their diabetes. This program involves tracking glucose through a CGM device for at least 20 days every 3 months. People optionally also track their activity, recording their heart rate and step count. We provide general demographic information on the cohort (after filtering) in Table 1, and the distribution of days recorded for each individual is shown in the Appendix (Figure 4). The CGM devices used measured blood glucose level roughly once every 5 minutes over the duration the device is worn and transmits the data via Bluetooth to a nearby receiver or smartphone. As the device has finite memory, issues during data transmission can result in loss of data, though substantial loss of data is rare in our dataset. The CGM monitoring device lasts for 10 days, and once taken off cannot be put back on.

We preprocess this data by first partitioning the dataset into individual days (each day starts at midnight). All our analyses will be conducted relative to a day of data. While we could consider longer or shorter intervals, we opt for a single day as there is a clear glucose cycle within a single day period. After this partition, we have 12,909 individuals and 1,859,101 days of data.

On each day, we interpolate the recorded data at exactly 5-minute intervals. This gives us exactly 288 datapoints per day. In cases where there is too much missing data ($\geq$ 15 missing measurements over any given day), we exclude this day from our dataset. We exclude people who have $<$ 5 days of data. We similarly preprocess the activity data. We do require individuals have at least one day of activity data as we will use features derived from the activity data in our analysis. All individuals also have medical history data associated with them. We do not explicitly preprocess this data, but we do use it later in our analysis to help predict events that significantly affect an individual.

We make one final exclusion based on matching the weather and air quality data to individuals. Individuals have at least one zip code associated with them based on the location of medical care received. We exclude people who have zip codes that differ in loca-

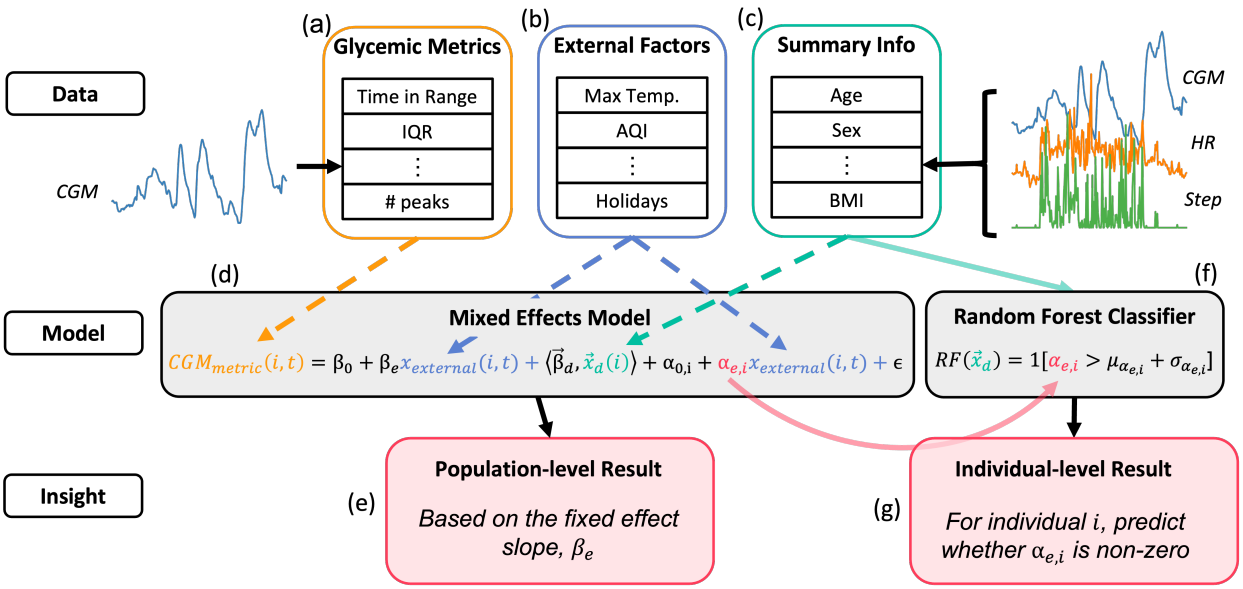

Figure 1: Overview of analysis procedure. We collect a dataset containing (a) standard glycemic metrics computed using a CGM trace, (b) external factors related to temporal and environmental conditions, and (c) "summary" demographic and medical information; we also compute summary statistics on the CGM and activity (heart rate (HR) and step count (Step)) data. (d) We fit a set of mixed effects models; here, we do not use the summary statistics computed on CGM or activity data. (e) This lets us assess the effect of each external factor. (f) A random forest (RF) classifier is trained to predict whether the external factor effect on each individual using the full set of summary information, giving us (g) the individual-level results.

tion by more than 10 miles during the 2.5-year period used for our analysis. We assume all other individuals generally spend their time close to the location where they receive care.

After filtering, the size of our dataset is reduced to 5,447 individuals and 940,663 days of data. Note that for each external factor we analyze, we may have to exclude additional datapoints (e.g., some individuals may not have any air quality data). For this filtering, we require each person has at least 2 days of data labeled with the external factor of interest.

2.2. External Factor Data

2.2.1. Weather and Air Quality Data

There are five weather measurements we use from the GCHN dataset: maximum temperature (TMAX), minimum temperature (TMIN), amount of rainfall (PRCP), amount of snowfall (SNOW), and depth of snow (SNWD). Note SNOW refers to snowfall on a given day, while SNWD refers to how much snow is on the ground on the given day (so it may include a previous day's snowfall). The EPA dataset consists of one measurement for the air quality index (AQI), which is calculated based on the concentration of small particulates (Lemeš, 2018). We show general statistics for these datasets in Table 2.

Measurements are recorded at base stations daily. The base stations have fixed, known locations (i.e., latitude and longitude are known). However, each base station may not provide measurements every day. So, to join these datasets with our CGM and activity measurements, we find the closest measurement to an individual's zip code on each day. We limit our search range to 10 miles and exclude the measurement if we cannot find a base station.

Table 1: Demographic information of Type II Diabetes cohort after filtering. In total, we have 5,447 individuals and 940,663 days of CGM data.

Category	External Factor	# Individuals	# Days	Age ($\mu \pm \sigma$)	Sex (%M / %F)
Weather	Max Temperature	3,561	506,260	55.72 ($\pm$8.79)	51.92% / 48.08%
	Temperature Range	3,559	503,374	55.72 ($\pm$8.79)	51.91% / 48.09%
	Precipitation	3,079	350,064	55.87 ($\pm$8.62)	52.09% / 47.91%
	Snow Depth	1,192	164,762	55.61 ($\pm$8.64)	51.74% / 48.26%
	AQI	3,262	448,042	55.87 ($\pm$8.82)	52.40% / 47.60%
Extreme Weather	Max Temperature	365	7,930	56.98 ($\pm$8.49)	48.54% / 51.46%
	Temperature Range	182	1,962	54.56 ($\pm$8.67)	56.52% / 43.48%
	Precipitation	50	263	56.80 ($\pm$8.75)	44.11% / 55.89%
	Snow Depth	35	187	57.97 ($\pm$7.80)	60.43% / 39.57%
	AQI	172	2,222	55.94 ($\pm$9.36)	55.58% / 44.42%
Temporal Events	Weekends	5,429	940,549	55.75 ($\pm$8.76)	48.89% / 51.11%
	Holidays	4,079	99,106	55.58 ($\pm$8.78)	48.84% / 51.16%

2.2.2. Extreme Weather and Air Quality

In addition to the continuous weather measurements, we would also like to understand glucose behavior relative to extreme weather events e.g., heatwaves. While there is no standard definition for what constitutes an extreme weather event, it is relatively common to consider heatwaves as anomalous high temperatures relative to location and time of year (Heo et al., 2019). We use this as a standard to calculate extreme weather events for each of the five weather conditions as well as for air quality.

For each condition, we first define regional clusters. This was done as the thresholds for determining extreme events are typically calculated over a geographic region (e.g., a country) (Heo et al., 2019). These clusters are based on clustering our population into 20 groups using k-means and defining these clusters as center points for regions. Within each cluster, we then defined an extreme weather event as outside two standard deviations of the mean over the 2.5-year period of the collected data. Furthermore, to label a time period as an extreme event, we required at least three consecutive days of the extreme event within a 10-mile radius of an individual.

As these events are sparse, we selected a subset of non-extreme days to use as controls in our analysis. These control days were defined as the three days immediately surrounding the extreme event (both three days before and after). This led to a roughly even split in the number of control and extreme days in our analysis.

2.2.3. Temporal Events

We also include two events associated with certain dates during a year: holidays and weekends. We include the following holidays: July 4th, Halloween, Thanksgiving, New Year, and the individual's birthday. In addition, for each holiday we include the preceding and following weeks, allowing for "holiday-like" behavior over multiple days (e.g., eating Thanksgiving leftovers later in the week). As was done for the extreme weather events, we selected a subset of non-holiday days to use as controls in our analysis – both a week prior and after a holiday were kept as controls (due to the week-long padding marking the holiday, this is two weeks before and after the holiday).

2.3. Glycemic Metrics

To evaluate the effect of external factors, we considered several widely used metrics for quantifying blood glucose management. These include time in range (TIR) (Maiorino et al., 2020), interquartile range (IQR) (Bergenstal, 2015), mean amplitude of glycemic excursions (MAGE) (Bergenstal, 2015), percent coefficient of variation (%CV) (Bergenstal, 2015). Definitions of each are provided below. Here $g(t)$ refers to an individual's glucose measurement (in mg/dL) at time t on a given day, μ_g, σ_g refer to the

mean and standard deviation of $g(t)$ on a given day, and Q_1, Q_3 refer to the 25th and 75th percentiles respectively. $E[\cdot]$ refers to the expected value of a measurement.

$$\begin{aligned} TIR &= 100\% \cdot E[1[70mg/dL \leq g(t) \leq 180mg/dL]] \\ IQR &= Q_3(g(t)) - Q_1(g(t)) \\ MAGE &= E[|(g(t) - \mu_g) \cdot 1|g(t) - \mu_g)| > \sigma_g|] \\ \%CV &= 100\%\sigma_g/\mu_g \end{aligned}$$

All measurements are calculated with respect to a single day for a single person. We provide recommended ranges for each of these metrics in Table 2. We also include the range of values observed in our dataset.

2.4. Population-level Analysis

To understand whether any of the external factors have a statistically significant effect at a population-level effect, we employ a linear mixed effects model. Our model assumes the following generative formula:

$$\begin{aligned} CGM_{metric}(i,t) = \beta_0 + \beta_e + x_{external}(i,t) + \\ \langle \boldsymbol{\beta}_d, \boldsymbol{x}_d(i) \rangle + \alpha_{0,i} + \\ \alpha_{e,i} x_{external}(i,t) + \epsilon, \end{aligned}$$

where $CGM_{metric}(i,t)$ is the metric of interest for individual i on day t, α, β are the fitted random and fixed effect parameters respectively, and $x_{external}(i,t), \boldsymbol{x}_d(i)$ are the external factor and demographic information for an individual on day t (only the external factor is time -dependent; we assume demographic information is static over the duration of our study). In particular, β are the fixed effects with β_0, β_e, $\boldsymbol{\beta}_d$ being the population intercept, effect due to the external factor, and effect due to demographic information. Similarly, $\alpha_{0,i}$and $\alpha_{e,i}$ are the random intercept and slope specific for individual i. Together, these terms can be thought of as a person-specific adjustment to better model individual i. Finally, ϵ is a Gaussian noise term.

We use a mixed effects model to correct for the fact that our data is not independent. Recall that each datapoint is a single day of data from a single person. As we have multiple days of data coming from each person, we will have dependence between data coming from the same person. Therefore, the standard statistical tests based on a linear model are not valid; a mixed effects model corrects for these repeated measures.

We fit a separate model for each external factor and each glycemic metric under consideration. Models are fit using the `Statsmodels` Python package (Seabold and Perktold, 2010). This results in 48 separate mixed effects models (one for each pair of glycemic metric and external factor). We determine which external factors have a significant population-level effect by identifying when the null hypothesis that the effect is zero can be rejected (p-value<0.05). Here, we use the p-value provided by `Statsmodels` which is based on a z-test where mean and β_e are estimated as part of fitting the mixed effects model. We correct for multiple hypothesis testing using the Bonferroni correction.

2.5. Individual-level Analysis

Given the population-level models, we construct a set of classifiers to identify which people are most significantly affected by each external factor. Note that we can identify such people regardless of whether the external factor has a significant population-level effect (people are highly heterogeneous, and we can identify some of that heterogeneity here).

For each external factor and glycemic metric, people are separated into three groups – those who are more negatively affected, more positively, or similarly affected compared to the population effect for the given external factor. This separation is based on the random slopes for the corresponding mixed effects model (recall, the random slopes are individual-level corrections to the population-level effect due to the external factor). We take the distribution of random slopes and partition the population into three groups – those at least one standard deviation below the mean, above the mean, or within one standard deviation of the mean. One standard deviation was chosen as a method for classifying outliers, as it ensures we have some data in each group (we need sufficient data to adequately evaluate the classifier performance). We found that choosing another threshold (e.g., 2 standard deviations) resulted in insufficient data to evaluate the classifier for most external factors.

We then assign a label to people based on their partition and train a random forest classifier (split into train, validation, and test sets using 60%, 20%, and 20% of the data respectively) to predict this label. To train the random forests, we used the `scikit-learn`

Table 2: Summary of glycemic metrics and how to interpret them. Glycemic metrics are computed on 24 hours of data. Percentiles computed across our filtered cohort.

		Percentiles in our dataset		
Metric	Target Range	25^{th}	50^{th} (median)	75^{th}
TIR	> 0.70 (Maiorino et al., 2020)	0.62	0.85	0.97
IQR	13-29 mg/dL (Bergenstal, 2015)	26.0	37.0	52.5
MAGE	41-48 mg/dL (Bergenstal, 2015)	39.7	53.4	71.8
%CV	19-25% (Bergenstal, 2015)	14.9	18.8	23.6

Python library (Pedregosa et al., 2011). Classifiers were trained using 100 trees, each with a max depth of 5. In total, we trained 48 random forest classifiers, one for each external factor–glycemic metric pair. Each classifier is input basic demographic and medical information as well as a summary of the individual's CGM and activity data (see Table 5 in the Appendix for more details). We also evaluate performance as a function of how much CGM and activity data is used for generating the summary – 0 days (i.e., exclude these features), 3 days (i.e., the first three days of recorded data we have), 10 days, and using all available data.

To report results, we use the receiver operating characteristic area under curve (ROC-AUC) and precision recall area under curve (PR-AUC) metrics. For each metric, we identified whether an increase or decrease in value is generally less desirable for someone with T2D. We report the ROC-AUC and PR-AUC values for identifying these individuals who are more negatively affected.

3. Results

3.1. Population-level Results

Here we select all external factors that have a significant effect (p-value < 0.05) on TIR. Once selecting those factors, we present the results from our models with respect to all glycemic metrics for those factors. These results are shown in Table 3. There are three external factors that have a significant effect at a population level: maximum temperature, holidays, and weekends.

There are several main takeaways. (1) Higher temperature leads to better glucose management. This result is not surprising – as has been previously noted in the literature, higher temperature has a physiological effect that can increase glucose absorption by the body (Kenny et al., 2016). (2) Holidays harm an individual's ability to manage glucose. While previous studies have shown seasonal variations in fat, cholesterol, and HbA1c values (Ma et al., 2006; Ockene et al., 2004; Higgins et al., 2009), this is the first direct evidence in a large-scale study to show how holidays specifically affect people using CGM data. (3) People are also worse at managing glucose on weekends. This is the first direct evidence for this result in a large-scale study using CGM data. (4) These conclusions largely hold across all glycemic metrics we measure. While each metric offers only a one-dimensional view into the complexity of an individual's glucose signal, the fact that our analysis shows significant change related to the external events further suggest the effect of the external event is significant and has a measurable impact on health

3.2. Individual-level Results

Classifier performance, measured as ROC-AUC is shown in Figure 2. Here we see that classifier performance is consistently above chance (across classifiers, median ROC-AUC=0.63). Performance is highest for classifiers predicting the effect of TIR (median ROC-AUC=0.70). Similar results are found when measuring PR-AUC (Appendix, Figure 5). These models are based on using all available CGM and activity data to compute summary statistics for an individual's behavior.

We also analyze performance as a function of the availability of CGM and activity data. Results are plotted in Figure 3. Here, models are trained using one of four feature sets: (1) just demographic features, demographic features and summary statistics based on (2) three, (3) ten, or (4) all days of CGM and activity data. Performance consistently improves when given more CGM and activity data. Moreover, performance is already high given only ten days of

Table 3: Population-level results. Reporting the fixed effect slope from the corresponding mixed effects model. Bolded when statistically significant ($\alpha < 0.05$, with the Bonferroni correction). A positive value for TIR and a negative value for IQR, MAGE, %CV generally indicates an improvement in health for the individuals in this dataset.

	Temp. Max [°C]	Weekends	Holidays
TIR	**6.34×10^{-4}**	**-1.47×10^{-2}**	**-1.90×10^{-2}**
IQR [mg/dL]	**-5.45×10^{-2}**	**9.11×10^{-1}**	**7.14×10^{-1}**
MAGE [mg/dL]	-5.95×10^{-2}	**1.08×10^{0}**	4.60×10^{-1}
%CV	**-9.27×10^{-3}**	**2.06×10^{-1}**	-4.92×10^{-2}

data indicating a small amount of data may be sufficient for high performance. As the features related to the CGM and activity data are intended to represent an individual baseline rather than relate information about any specific external event, increasing the number of days used provides a better estimate of the true individual baseline. For similar results with the PR-AUC metric, see Appendix, Figure 6.

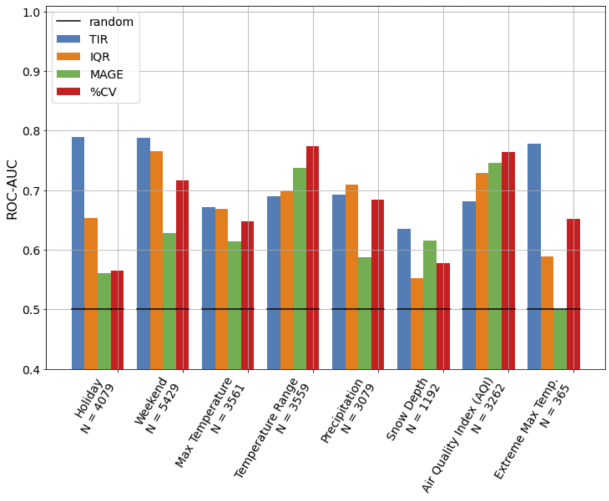

Figure 2: Performance (ROC-AUC) of models at predicting the effect of an external factor individuals. Each bar represents a classifier. External factors that have < 1000 datapoints (days of data) for training and evaluation combined are excluded.

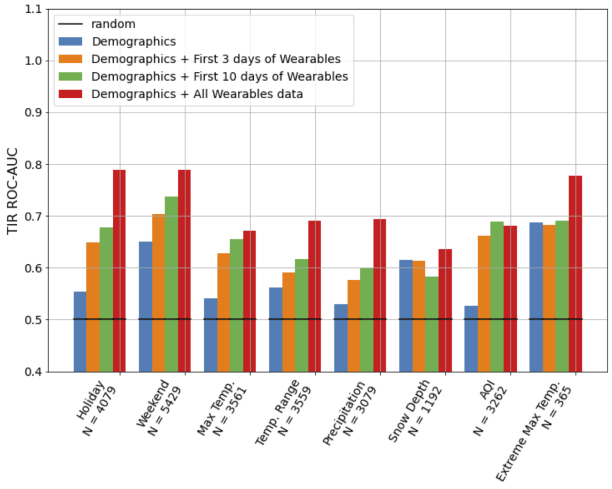

Figure 3: Performance increases with additional CGM and activity data. Showing ROC-AUC of models predicting TIR effect on individuals. External factors that have < 1000 datapoints (days of data) for training and evaluation combined are excluded.

4. Discussion, Limitations, and Future Work

In this paper, we demonstrated that external factors, including weather conditions and temporal events like holidays, can affect glucose management. We showed these effects on a day-to-day basis, where an individual's daily environment affects their glucose levels. We also showed that these effects can be predicted with limited to no observed CGM or activity data; when CGM and activity data are available, we see improved performance.

This second result is important. While with sufficient data, it is possible to directly analyze the effects of external factors on an individual, this data is not available prior to long-term monitoring. Moreover, for certain events like extreme weather situations, we may rarely or even never observe these events for any given individual. However, our method is still able to infer the effect of these external factors.

There are two key limitations in our findings: (1) our analysis cannot distinguish between the behavioral and biological causes for variation in and glycemic outcomes, and (2) we do not investigate interactions between environmental effects. Despite these limitations, our findings are important. Our ability to predict individual negative responses to external factors is an important finding and has several implications. It may enable better care management with day-to-day risk assessments based on external factors, as well as improve algorithm development by reducing train-time and test-time bias due to external factors.

There are several future directions for our work. At the population level, we used a linear model, while the relation between our features and glucose management metrics are likely non-linear and may benefit from non-linear modeling. Using non-linear modeling, we may also be able to detect population-level significance for certain subgroups of people we were not able to find in this paper. At the individual level, we limited our modeling tuning for our classifier models. While we demonstrate we can predict the effects due to external factors at an individual level, it is likely we can significantly boost performance by further tuning our model choice; given our large dataset, performance may benefit from deep learning.

In conclusion, we found that day-to-day variation in glucose level can be partially explained by external factors like the weather or temporal events like holidays. These effects are significant at a population level. We also found that we could predict which specific individuals will see the largest negative effects based on summary data available prior to monitoring and a limited amount of CGM and activity data. This second result is important and could potentially lead to more personalized care for diabetes.

References

Stacey M Anderson, Dan Raghinaru, Jordan E Pinsker, Federico Boscari, Eric Renard, Bruce A Buckingham, Revital Nimri, Francis J Doyle III, Sue A Brown, Patrick Keith-Hynes, et al. Multinational home use of closed-loop control is safe and effective. *Diabetes care*, 39(7):1143–1150, 2016.

American Diabetes Association. Economic costs of diabetes in the us in 2017. *Diabetes care*, 41(5): 917–928, 2018.

Richard M Bergenstal. Glycemic variability and diabetes complications: does it matter? simply put, there are better glycemic markers! *Diabetes care*, 38(8):1615–1621, 2015.

Sheri R Colberg, Ronald J Sigal, Jane E Yardley, Michael C Riddell, David W Dunstan, Paddy C Dempsey, Edward S Horton, Kristin Castorino, and Deborah F Tate. Physical activity/exercise and diabetes: a position statement of the american diabetes association. *Diabetes care*, 39(11): 2065–2079, 2016.

Darpit Dave, Daniel J DeSalvo, Balakrishna Haridas, Siripoom McKay, Akhil Shenoy, Chester J Koh, Mark Lawley, and Madhav Erraguntla. Feature-based machine learning model for real-time hypoglycemia prediction. *Journal of Diabetes Science and Technology*, 15(4):842–855, 2021.

Linda A DiMeglio, Carmella Evans-Molina, and Richard A Oram. Type 1 diabetes. *The Lancet*, 391(10138):2449–2462, 2018.

Christopher Duckworth, Matthew J Guy, Anitha Kumaran, Aisling Ann O'Kane, Amid Ayobi, Adriane Chapman, Paul Marshall, and Michael Boniface. Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations. *Journal of Diabetes Science and Technology*, page 19322968221103561, 2022.

Nuha A ElSayed, Grazia Aleppo, Vanita R Aroda, Raveendhara R Bannuru, Florence M Brown, Dennis Bruemmer, Billy S Collins, Marisa E Hilliard, Diana Isaacs, Eric L Johnson, et al. 7. diabetes technology: Standards of care in diabetes—2023. *Diabetes Care*, 46(Supplement_1):S111–S127, 2023.

Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS biology*, 16(7):e2005143, 2018.

Seulkee Heo, Michelle L Bell, and Jong-Tae Lee. Comparison of health risks by heat wave definition: Applicability of wet-bulb globe temperature for heat wave criteria. *Environmental research*, 168: 158–170, 2019.

Trefor Higgins, Sharon Saw, Ken Sikaris, Carmen L Wiley, George C Cembrowski, Andrew W Lyon, Annu Khajuria, and David Tran. Seasonal variation in hemoglobin a1c: is it the same in both hemispheres?, 2009.

Kimmo Karkkainen, Gregory D Lyng, Brian L Hill, Kailas Vodrahalli, Jeffrey Hertzberg, and Eran Halperin. Sleep and activity prediction for type 2 diabetes management using continuous glucose monitoring. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Glen P Kenny, Ronald J Sigal, and Ryan McGinn. Body temperature regulation in diabetes. *Temperature*, 3(1):119–145, 2016.

Samir Lemeš. Air quality index (aqi)—comparative study and assesment of an appropriate model for b&h. In *2th Scientific/Research Symposium with International Participation 'Metallic And Nonmetallic Materials*, 2018.

Yunsheng Ma, Barbara C Olendzki, Wenjun Li, Andrea R Hafner, D Chiriboga, James R Hebert, M Campbell, M Sarnie, and Ira S Ockene. Seasonal variation in food intake, physical activity, and body weight in a predominantly overweight population. *European journal of clinical nutrition*, 60 (4):519–528, 2006.

Maria Ida Maiorino, Simona Signoriello, Antonietta Maio, Paolo Chiodini, Giuseppe Bellastella, Lorenzo Scappaticcio, Miriam Longo, Dario Giugliano, and Katherine Esposito. Effects of continuous glucose monitoring on metrics of glycemic control in diabetes: a systematic review with meta-analysis of randomized controlled trials. *Diabetes Care*, 43(5):1146–1156, 2020.

Matthew J Menne, Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S Vose, Byron E Gleason, et al. Global historical climatology network-daily (ghcn-daily), version 3. *NOAA National Climatic Data Center*, 10:V5D21VHZ, 2012.

Ebenezer A Nyenwe, Terri W Jerkins, Guillermo E Umpierrez, and Abbas E Kitabchi. Management of type 2 diabetes: evolving strategies for the treatment of patients with type 2 diabetes. *Metabolism*, 60(1):1–23, 2011.

Ira S Ockene, David E Chiriboga, Edward J Stanek III, Morton G Harmatz, Robert Nicolosi, Gordon Saperia, Arnold D Well, Patty Freedson, Philip A Merriam, George Reed, et al. Seasonal variation in serum cholesterol levels: treatment implications and possible mechanisms. *Archives of internal medicine*, 164(8):863–870, 2004.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.

Hong Sun, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B Duncan, Caroline Stein, Abdul Basit, Juliana CN Chan, Jean Claude Mbanya, et al. Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes research and clinical practice*, 183: 109119, 2022.

US Environmental Protection Agency. Air quality system data mart. https://www.epa.gov/outdoor-air-quality-data, 2022. Accessed: 2022-07-01.

Natalia G Vallianou, Eleni V Geladari, Dimitris Kounatidis, Chara V Geladari, Theodora Stratigou, Spyridon P Dourakis, Emmanuel A Andreadis, and Maria Dalamaga. Diabetes mellitus in

the era of climate change. *Diabetes & Metabolism*, 47(4):101205, 2021.

Zhiwei Xu, Shilu Tong, Jian Cheng, James Lewis Crooks, Hao Xiang, Xiangyu Li, Cunrui Huang, and Wenbiao Hu. Heatwaves and diabetes in brisbane, australia: a population-based retrospective cohort study. *International Journal of epidemiology*, 48(4):1091–1100, 2019.