An N=1 Type-1 Diabetes Dataset: Preliminary Insights from Personal Diet, Fitness, CGM, and Insulin Pump Data

Victor Goncalves
Harvard University

Abstract

This paper presents an N=1 study focused on long-term glycemic control in type-1 diabetes, offering a dataset and preliminary insights derived from personal diet, fitness, Continuous Glucose Monitoring (CGM), and insulin pump data. With the growing prevalence of diabetes worldwide, the need for advanced management strategies, particularly for Type-1 Diabetes, is critical. This study leverages diabetes care technology, exploring the potential of these tools in enhancing diabetes self-management and how it can apply with machine learning (ML). However, a significant challenge in this domain is the lack of comprehensive and accessible datasets, especially those that include detailed meal and fitness information. As a Type-1 diabetic, I created a personal 2.5-month dataset encompassing various health metrics. This dataset[1], integrating data from an insulin pump, CGM, a fitness tracker, and comprehensive meal data, aims to provide a foundational basis for preliminary analysis, with hopes to advance ML in diabetes management and treatment.

Keywords
Type-1 diabetes, machine learning, personal health data, glycemic control

1 Introduction

Diabetes, a group of conditions marked by elevated blood glucose levels, has been increasingly prevalent worldwide [1]. The disease manifests in two primary forms: type-1 and type-2. Insulin, a hormone crucial for blood glucose regulation produced by the pancreas, plays a central role in both types. In type-1 diabetes, an autoimmune response destroys the insulin-producing beta cells in the pancreas, leading to little or no insulin production and necessitating lifelong insulin therapy [2]. Type-2 diabetes, more common globally, involves insulin resistance and eventual pancreatic insufficiency [3].

Untreated/uncontrolled diabetes can lead to severe health complications. Hypoglycemia, a critical condition characterized by low blood sugar, can result from excess insulin or physical activity, sometimes leading to a diabetic coma, especially at night. Conversely, hyperglycemia, high blood sugar, can arise from insufficient insulin or excessive carbohydrate intake, leading to long-term complications, such as kidney failure, and limb amputation, if not managed properly.

Historically, diabetes treatment involved manual blood glucose testing and insulin injections. However, the stringent requirements for blood glucose regulation in type-1 diabetes have driven

[1] GitHub repository: https://github.com/Kiopsy/t1d_data

significant engineering advancements. Notable among these is the development of the Continuous Glucose Monitor (CGM) and artificial pancreas technologies, which automate blood glucose control and represent a major leap in diabetes management [7, 8].

With new technologies, we have seen advancements in combining machine learning (ML) with diabetes care research. However, a noticeable gap persists in the predictive capabilities of diabetes treatments, which I hypothesize primarily to a lack of comprehensive data. As of 2023, the availability of extensive, openly accessible type-1 diabetes datasets are limited, with many lacking data on fitness and meals due to its meticulous nature for collection [4, 5].

My own experience as a type-1 diabetic highlights these difficulties, effectively making me a living dataset. This paper contributes to this effort by presenting a comprehensive 2.5-month dataset that includes meal, fitness, insulin pump, and CGM data. Through this dataset, I aim to provide a foundational basis for preliminary analysis, thereby advancing the understanding and application of ML in diabetes management and treatment.

2 Literature Review

Artificial Pancreas

The artificial pancreas system is engineered to replicate the functionality of a healthy pancreas. Incorporating a CGM, an insulin pump, and a control program, often housed on a smartphone or integrated into the pump, these systems vary in design and function. Some systems are programmed to suspend insulin delivery when glucose levels fall below a certain threshold, while others dynamically adjust insulin delivery based on continuous glucose monitoring [7, 8]. The integration of these systems into standard diabetes care, particularly for type-1 diabetes, and their enhancement through machine learning algorithms, mark a significant advancement in the field [6, 7].

Machine Learning

The evolution of CGM and artificial pancreas technologies has catalyzed research in applying machine learning to diabetes management. This includes the development of models powered by recurrent neural networks (RNN) that utilize multi-modal data for glucose level prediction and management of hypo- and hyperglycemia events [12]. Additionally, neural networks have been employed to automate meal detection and insulin dosing, showcasing the potential of machine learning in enhancing diabetes self-management [6]. My research is driven by the aspiration to gather comprehensive data that can further refine these ML models.

Diabetes Datasets

The development of publicly accessible type-1 diabetes datasets is critical for advancing treatment methodologies, enhancing artificial pancreas systems, and informing machine learning applications. Prominent among these datasets are DiaTrend and the OhioT1DM dataset. DiaTrend is notable for its extensive longitudinal data, featuring 27,651 days of CGM and 8,220

days of insulin pump data from 54 patients [4]. In contrast, the OhioT1DM dataset provides an intensive eight-week period of data integration, including CGM, insulin pumps, physiological sensors, and self-reported events from 12 individuals with type 1 diabetes [5].

In my research, I aim to contribute to these datasets by compiling a long-term dataset that encompasses extensive diabetes management data (insulin, and blood sugar), and also integrates detailed meal and fitness information. This approach seeks to provide a more comprehensive understanding of type-1 diabetes management and its challenges, and provides better data for ML models and analysis.

3 Method

As a type-1 diabetic, I collected data on myself (N=1) for approximately 8 weeks to date. This dataset includes sources from an insulin pump, CGM, fitness tracker, and comprehensive meal data. I tracked data on myself. Specifically, I wore a Tandem t:slim X2 insulin pump, a Dexcom G6 CGM, a Fitbit Inspire 3 fitness tracker for the duration of study. In order to properly aggregate insulin pump and CGM data, I utilized the Tidepool website, a non-profit organization that created a website that connects to the Dexcom API and allows you to plug in your insulin pump and upload and display your data. Lastly, I collected meal data, using Bitesnap, following recommendations found in relevant literature [11]. The application allows you to take a photo of your food, and it predicts the label of each food item using a convolutional neural network. Additionally, the application has food items in its database that have macronutrient breakdown for each food item. The data encompassed both a detailed macronutrient breakdown and photographs of the food.

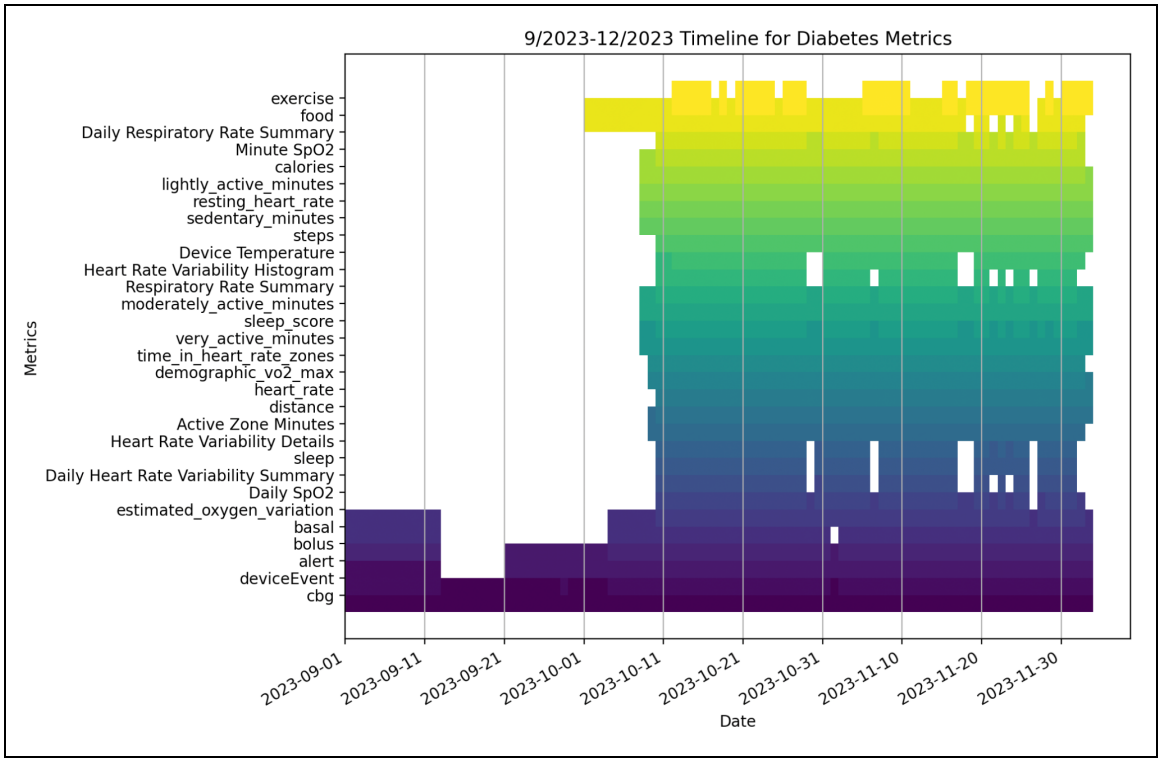


Figure 1. Available data from 9/2023-12/2023

Figure 1 includes a timeline breakdown of each metric collected between September 2023 to December 2023. The metrics collected during this time, along with the total amount of entries per metric, are listed in Tables 1-3. Note the gap in data during September for fitness and meal data compared to the insulin and CGM data. I began collecting meal data as of the start of October, and I started wearing the fitness tracker on October 10th.

Fitbit			
Metric	**Total Entries**	**Metric**	**Total Entries**
heart_rate	500036	sedentary_minutes	60
calories	80141	lightly_active_minutes	60
Device Temperature	55289	demographic_vo2_max	55
distance	42009	time_in_heart_rate_zones	55
steps	42009	Respiratory Rate Summary	53
Minute SpO2	24432	Heart Rate Variability Histogram	53
estimated_oxygen_variation	24115	sleep	52
Active Zone Minutes	5233	Computed Temperature	47
Heart Rate Variability Details	3412	Daily SpO2	45
swim_lengths_data	2195	Daily Heart Rate Variability Summary	45
resting_heart_rate	365	Daily Respiratory Rate Summary	45
exercise	81	sleep_score	44
very_active_minutes	60	height	8
moderately_active_minutes	60	weight	1

Table 1. Fitbit metrics and total entries

Bitesnap	
Metric	**Total Entries**
food	629

Table 2. Bitesnap metrics and total entries

Tidepool			
Metric	**Total Entries**	**Metric**	**Total Entries**
cbg	25560	pumpSettings.bgTargets	75
basal	10070	pumpSettings.basalSchedules	75
alert	1533	wizard	175
bolus	791	cgmSettings	21
deviceEvent	681	upload	3
pumpSettings.insulinSensitivities	75	smbg	2
pumpSettings.carbRatios	75		

Table 3. Tidepool metrics and total entries

Personal History

In terms of my personal history, I was diagnosed with type-1 diabetes in 2012 and have been a diabetic for 11 years to date. I have been on insulin pump therapy in conjunction with the CGM for 5 years now. My most recent A1c has been 6.4%, but it has ranged from 5.7%-6.4% from 2019-2023 as seen in Figure 2. For context, an A1c is a blood test that shows average blood sugar for about two or three months, and the American Diabetes Association recommends an A1c of less than 7% [10]. Moreover, based on a representative National Health and Nutrition Examination Survey (NHANES) from 2009-2020, only 25% of type-1 diabetics meet this range [9]. My blood sugars are on the more controlled range for the population of type-1 diabetics.

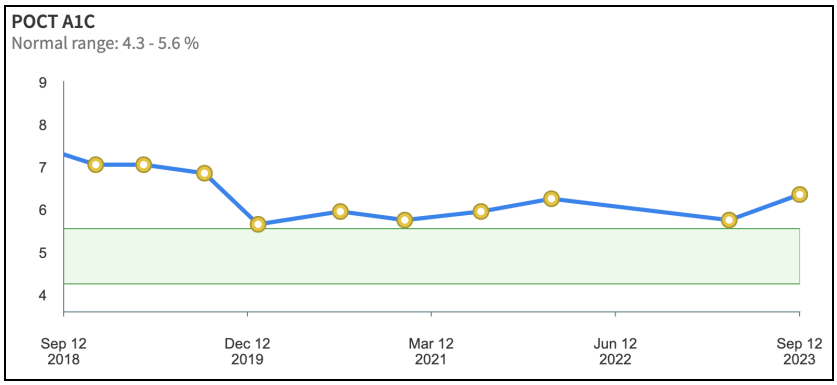


Figure 2. My personal A1c values from 2018-2023

4 Implementation

Python Script for Data Aggregation

A Python script was developed to integrate data from multiple sources into a consolidated timeline. Given the lack of direct API access for certain data sources, namely Bitesnap and Tidepool, a manual approach was necessitated. Data exports from these platforms were locally merged on my computer to construct a continuous timeline.

The script's codebase comprises three primary directories: 'used', 'export', and 'cleaned'. Within these directories, there are sub-folders for each data source, including Tidepool, Bitesnap, and Fitbit. The process begins with adding newly pulled data into the appropriate 'export' sub-folder. Subsequently, a data parsing script is executed to integrate these datasets.

Upon execution, the data parser scans each file within the 'export' directory. It systematically deconstructs the data, organizing it by individual metrics and month. In instances where a single file contains multiple metrics over an extended period (e.g., a Tidepool file encompassing blood sugar and insulin data for three months), the script segregates this data into distinct dictionaries, further categorizing it by month.

The parser then assesses whether a file corresponding to a specific metric and month already exists. If it does, the script merges the new export data with the existing data, carefully avoiding duplication and ensuring chronological sorting. In cases where no corresponding file exists, the script generates a new file, depositing a month's worth of data into it.

Upon completion, the 'cleaned' directory houses folders for each metric, containing monthly data files. For instance, the heart rate data from my Fitbit is stored as follows:

For example here is the heart rate data stored on my computer:

data/cleaned_data/fitbit/heart_rate
- heart_rate-2023-10.json
- heart_rate-2023-11.json
- heart_rate-2023-12.json

Each source presents different data formats, so it was important to clean each metric to provide a standardized way to record entries across metrics. A notable challenge was the inconsistency in timestamp formats. Some files used UTC time, while others recorded local time. Additionally, all of the files had different timestamp headers. To analyze across different metrics, it is essential to unify these time formats. Additionally, I segment the data by month which offers a practical balance between time and space complexity, particularly when handling these large files. For instance, heart rate data alone, with approximately 14,000 entries per day. In total, the entire dataset amounted to around 115 MB over a span of 2.5 months. By breaking the data down monthly, I found it was a reasonable bound in size for these files. Especially if I wanted to create

a website to display the data in the future, it would have reasonable loading times, unless maybe comprehensive yearly or monthly trend calculations were needed.

Data Format

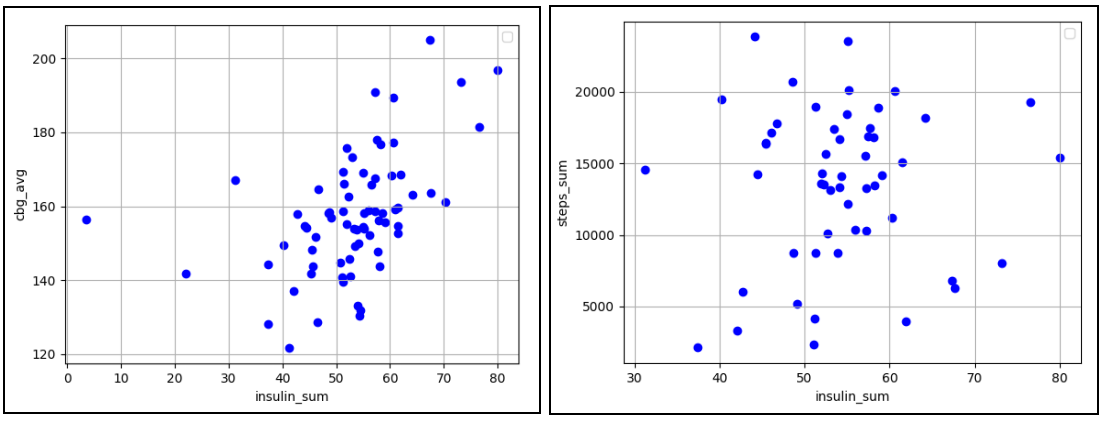
The structured data is available on my GitHub repository, although it will be removed eventually to protect patient confidentiality. The data organization follows a consistent structure: For each source (Fitbit, Tidepool, Bitesnap), there is a 'cleaned' folder containing metric-specific folders, with files organized by month. Each file is a JSON file, maintaining structure across different metrics. Here is an example of a heart rate entry:

- "value": {"bpm": 74, "confidence": 1}
- "utc_Time": "2023-12-03T01:28:34Z"
- "local_Time": "2023-12-02T20:28:34"
- "timezoneOffset": -300.0

While the keys may vary depending on the metric, each entry consistently includes both UTC and local time, with the appropriate timezone offset. Currently, the script is hard-coded to EST time, as I am the only user of this script for now. Future modifications will aim to accommodate various time zones to enhance the script's adaptability and broader applications.

5 Results

After gathering my data, I conducted preliminary analyses to explore potential patterns. As depicted in Figure 3, I compared various metrics to identify correlations. Notably, a correlation emerged between average blood sugar levels and daily total insulin intake. Interestingly, this correlation is positive, which initially seems counterintuitive given that insulin typically reduces blood sugar levels, but it can make sense that I have higher blood sugar on days where I am insulin resistant and need more insulin. Also, the data revealed no correlation between step count and these variables (which could be influenced by exercise). These initial findings suggest the need for more analysis, potentially beyond a daily resolution, to uncover trends.



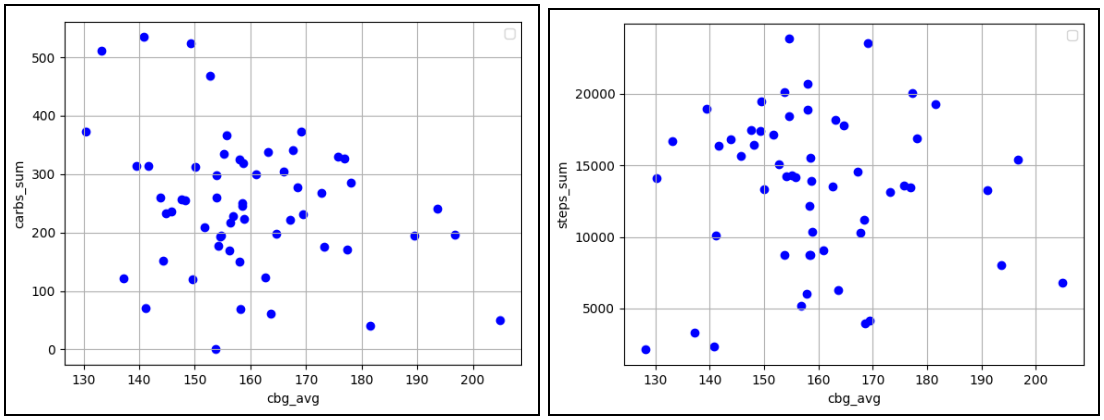


Figure 3. Correlation between metrics

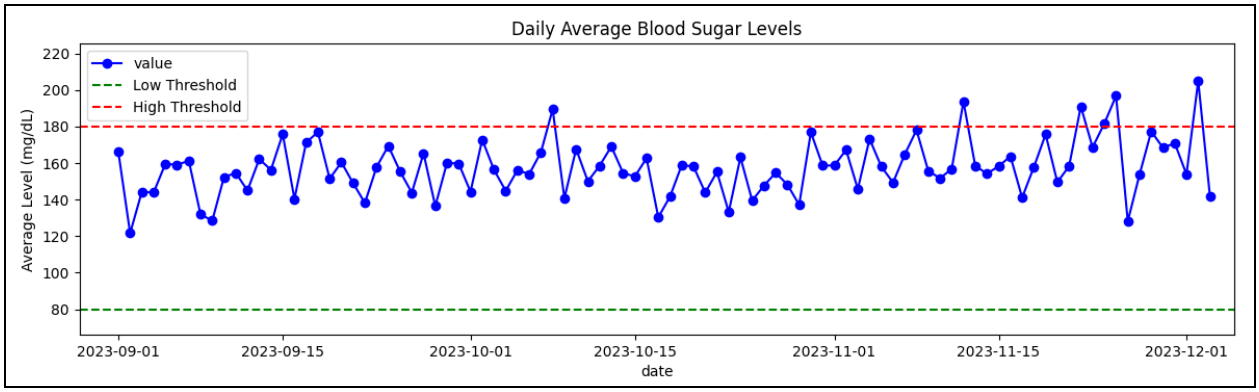


Figure 4. Daily average blood sugar levels for the month of October.

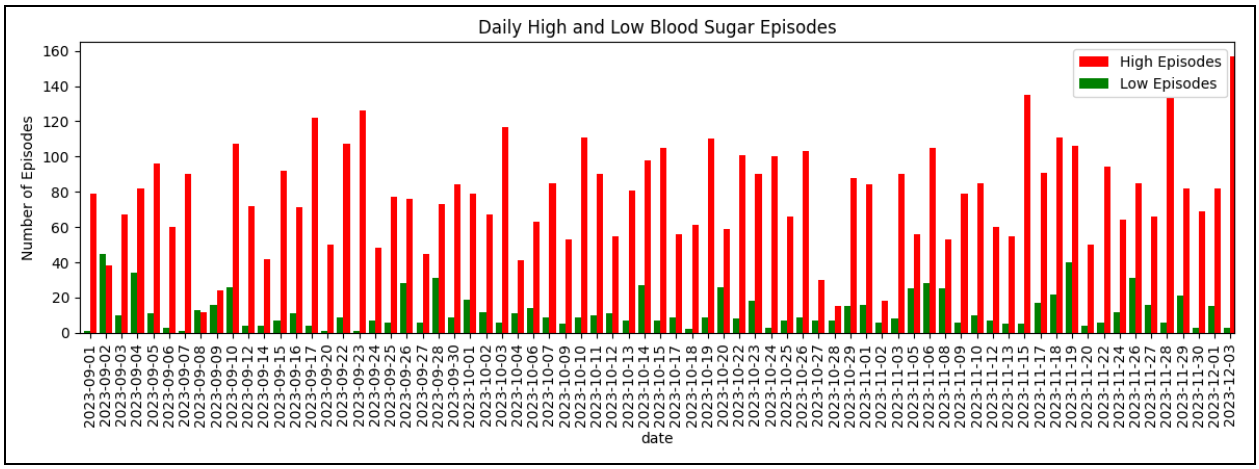


Figure 5. Daily high and low blood sugar episodes

As seen in Figure 4, my daily average blood sugar consistently remains below the high threshold, typically ranging from 120-170. Interestingly, my blood sugar levels in November were significantly higher compared to September and October. Figure 5 plots the frequency of blood sugar readings in the high range (>180) and low range (<70). Days with more high and low episodes may indicate periods of heightened insulin sensitivity.

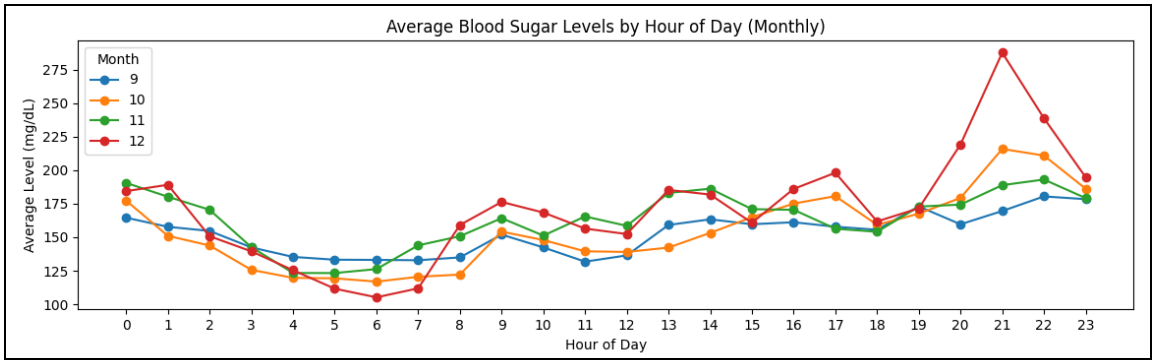


Figure 6. Average blood sugar by hour of day (monthly)

Figure 6 demonstrates that my blood sugar tends to be higher post-dinner, a trend consistent across months. This suggests a need to either increase my pre-dinner insulin dosage or reduce carbohydrate intake at that meal. Additionally, the data shows approximately 3-4 peaks during the day, likely corresponding to my meal times.

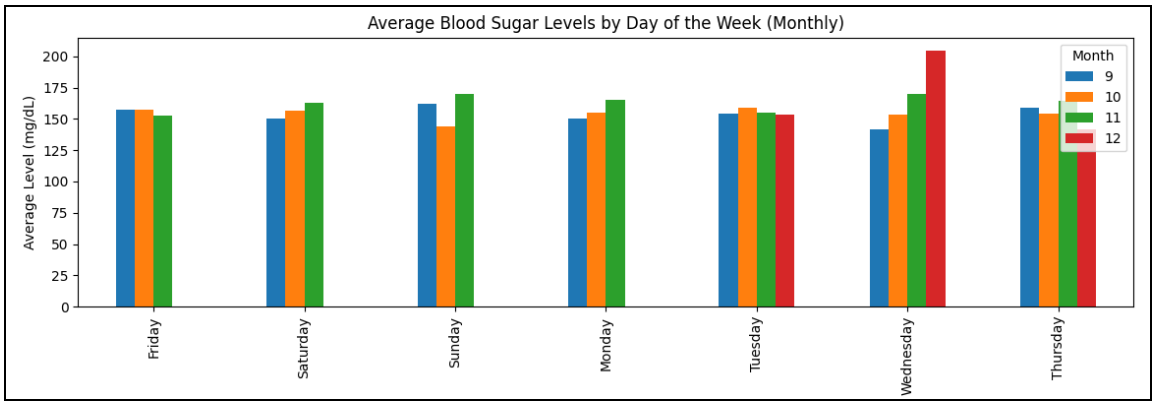


Figure 7. Average blood sugar by day of the week (monthly)

Lastly, I analyzed trends based on the day of the week, considering that Harvard offers similar meal options on specific days. However, it's important to note that the data for December only covers three days, limiting its utility for average comparisons with other months.

In summary, a more comprehensive data analysis is essential for a detailed understanding. A granular examination within days will help discern if factors like exercise or specific meals significantly impact my blood sugar levels or insulin requirements.

6 Challenges & Limitations

Throughout the semester, I spend a good amount of time implementing the parsing script for the data collection. Additionally, planning and setting up for the project took a while as well. It also took a while to decide on the food tracking application. While Bitesnap is user-friendly, I am questionable about its macronutrient counting accuracy. In contrast, MyFitnessPal allows

everyone to contribute to its food database. By nature of the extensiveness, it offers more data and I believe it was a little more precise in terms of macronutrient data. However, I strayed from this as the data export feature from MyFitnessPal required a premium subscription. I tried to create a web scraper for MyFitnessPal in September to no avail. During this development phase, monitoring progress, particularly in food tracking, was very challenging. It felt like I was blindly tracking data, without having context of what it would look like in the end. The completed parser and analysis script now facilitate better data tracking and the creation of a more robust database for future research. I can collect data knowing what to expect. Accurately tracking every meal and snack to provide precise macronutrient counts was also very meticulous, but I argue is essential for controlling blood sugar levels.

The dataset has inherent limitations, including gaps in exercise, sleep, or meal data. For example, sometimes the fitness tracker would die overnight, or I would forget to log some meals. Additionally, I went to Dubai from November 20th to 27th, which presented challenges including inconsistent food tracking and difficulties adjusting data for significant time zone changes. The current software version does not accommodate time zone differences, complicating the data aggregation process across geographic locations.

The data volume for 2.5 months stood at approximately 115mb, which could easily become hard to manage over longer periods or with additional participants. I am considering transitioning from JSON to CSV files to save space, but this would require significant reengineering due to the complex nested structures in the source JSON data.

7 Discussion

This project lays the infrastructure for a comprehensive blood sugar analysis. I am now equipped to consistently collect and analyze my data, contributing to an ongoing dataset. This endeavor initially began with the intention of applying machine learning to my diabetes data, exploring possibilities such as:

- RNNs for insulin dosing strategies based on a time series of blood sugar data, incorporating variables like insulin, fitness, heart rate, etc.
- Supervised learning models to predict meals from blood sugar changes observed over 30 minutes.

An intriguing prospect is the application of LLMs in analyzing diabetes data. For instance, with a standardized data structure, patients could upload their data to an LLM platform, enabling them to ask questions about their diabetes management. Nowadays, GPT-4-like models could even generate immediate graphical representations and analyses.

Example queries could include:

- Q: "Why was my blood sugar high yesterday?"
- A: "Your data indicates recurring post-dinner blood sugar spikes. Consider adjusting your insulin dosage or reducing carbohydrate intake during dinner."

- Q: "What is the trend in my insulin usage this month?"
- A: [Displays a graph of insulin usage trends]

Currently, no type-1 diabetes management software offers such flexibility and customizability in data representation. The development of such a tool could significantly enhance blood sugar regulation for diabetics. A lot more analysis is needed on my data in order to see how machine learning can apply, and if I can find any more specific trends, say with exercise, heart rate, or type of meal.

Conclusion

This study presents an N=1 dataset that captures comprehensive personal health data for type-1 diabetes management, integrating Continuous Glucose Monitoring (CGM), insulin pump, fitness, and dietary information over 2.5 months. The findings show the potential of using integrated, personalized data to enhance diabetes self-management and inform machine learning models. Despite its limited scope, the study demonstrates the feasibility and value of such data in understanding and optimizing type-1 diabetes management. Future research should focus on expanding this dataset with more participants to develop sophisticated, more analysis and ways to display the data, and personalized machine learning applications for diabetes care. Overall, this project represents a significant step towards data-driven and individualized diabetes management.

Acknowledgements

I would very much like to thank Dr. Finale Doshi-Velez for advising this CS91r project. Having lost some family members to diabetes, this project holds a special place in my heart. I hope to continue the project to find easier ways to better my blood sugar and find interesting insights on my own personal data. Eventually, I hope to extend these benefits to other individuals living with diabetes

References

[1] G. A. Gregory et al., “Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: A modelling study,” The Lancet Diabetes & Endocrinology, vol. 10, no. 10, pp. 741–760, 2022. doi:10.1016/s2213-8587(22)00218-2

[2] “What is type 1 diabetes?,” Centers for Disease Control and Prevention, https://www.cdc.gov/diabetes/basics/what-is-type-1-diabetes.html#:~:text=Type%201%20diabetes%20is%20thought,years%20before%20any%20symptoms%20appear. (accessed Dec. 12, 2023).

[3] “Insulin resistance and diabetes,” Centers for Disease Control and Prevention, https://www.cdc.gov/diabetes/basics/insulin-resistance.html (accessed Dec. 12, 2023).

[4] T. Prioleau, A. Bartolome, R. Comi, and C. Stanger, “Diatrend: A dataset from advanced diabetes technology to enable development of novel Analytic Solutions,” Scientific Data, vol. 10, no. 1, 2023. doi:10.1038/s41597-023-02469-5

[5] C. Marling and R. Bunescu, “The ohiot1dm dataset for blood glucose level prediction: Update 2020,” CEUR workshop proceedings, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/ (accessed Dec. 12, 2023).

[6] C. Mosquera-Lopez et al., “Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence,” npj Digital Medicine, vol. 6, no. 1, 2023. doi:10.1038/s41746-023-00783-1

[7] F. J. Doyle, L. M. Huyett, J. B. Lee, H. C. Zisser, and E. Dassau, “Closed-loop artificial pancreas systems: Engineering the algorithms,” Diabetes Care, vol. 37, no. 5, pp. 1191–1197, 2014. doi:10.2337/dc13-2108

[8] “Artificial pancreas - NIDDK,” National Institute of Diabetes and Digestive and Kidney Diseases, https://www.niddk.nih.gov/health-information/diabetes/overview/managing-diabetes/artificial-pancreas (accessed Dec. 12, 2023).

[9] E. R. Hankosky et al., “Gaps remain for achieving hba1c targets for people with type 1 or type 2 diabetes using insulin: Results from NHANES 2009–2020,” Diabetes Therapy, vol. 14, no. 6, pp. 967–975, 2023. doi:10.1007/s13300-023-01399-0

[10] “A1C looks back,” A1C: What It Is, Test & Chart | ADA, https://diabetes.org/about-diabetes/a1c#:~:text=The%20goal%20for%20most%20adults,that%20is%20less%20than%207%25.&text=If%20your%20A1C%20level%20is,were%20in%20the%20diabetes%20range. (accessed Dec. 11, 2023).

[11] S. Gioia et al., “Mobile apps for dietary and Food Timing Assessment: Evaluation for use in clinical research,” JMIR Formative Research, vol. 7, 2023. doi:10.2196/35858

[12] T. Zhu et al., “Enhancing self-management in type 1 diabetes with wearables and deep learning,” npj Digital Medicine, vol. 5, no. 1, 2022. doi:10.1038/s41746-022-00626-5