Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The 9 months Type 1 Diabetes dataset was created specifically for personal reasons. It was made so that the creator could examine his fiancee's diabetes data and conduct research to derive personal and academic value.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The original creator of the dataset is Jeremy Chu, a Master of Information student at the University of Toronto at the time of the dataset's release in 2021

What support was needed to make this dataset?

Except for consent from the subject of the dataset, no other support was needed for the dataset's creation

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings;

A single person's blood sugar level readings, food consumption, and insulin intake for 9 months

Is there a label or target associated with each instance?

Data is grouped monthly. Cleaned dataset is aggregated by hour.

Is any information missing from individual instances?

The months January - March 2020 are missing. Dataset goes from December 2019 and jumps to April 2020.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

There are obvious date times for each month's data.

Are there recommended data splits (e.g., training, development/validation, testing)?

No recommended data split. If desired, choose specific months to use as testing months..

Are there any errors, sources of noise, or redundancies in the dataset?

No errors or redundancies. For the cleaned dataset, be aware that data is aggregated by hour. For the raw dataset, there are no errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

Data is an individual's diabetes data. Consent has been prior obtained, any further details must be requested.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Does the dataset relate to people?

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)?

Yes.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No direct identifiable information present in the dataset. The study however, speifies the relationship between the individual and the researcher.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

As sensitive as health data publicly disclosed after obtaining consent.

Any other comments?

Collection Process

How was the data associated with each instance acquired?

The data is collected from the Dexcom G6 CGM. One to one translated from the machine to csv.

Over what timeframe was the data collected? D

The data was continously collected by the CGM. Every momth the data is available to be downloaded directly from the machine to a computer. The data collection malfunctioned during the months January - March of 2020, therefore only 9 months of data remain instead of 1 year.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Dexcom G6 CGM sensor.

What was the resource cost of collecting the data?

No cost.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Sample of the individual's diabetes data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Just the individual whose data is collected. Obtaining consent was the only requirement.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Obtained permission to take directly from CGM.

Were the individuals in question notified about the data collection?

Yes. Individual must give me their CGM for data collection.

Did the individuals in question consent to the collection and use of their data?

Yes. Individual must give me their CGM for data collection.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Yes. Individual has access to the GitHub repository.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Yes. Individual has access to the GitHub repository.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Yes. The data has been aggregated by hour, with the CGM and insulin intake data combined. Any personal identifiable variables have been removed from both the raw and cleaned versions.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes. In the same GitHub repository.

Is the software used to preprocess/clean/label the instances available?

Any other comments?

Uses

Has the dataset been used for any tasks already?

Yes. The data has been used for a personal research already. The study can be found in the same GitHub repository.

Is there a repository that links to any or all papers or systems that use the dataset?

https://github.com/JeremyJChu/diabetes

What (other) tasks could the dataset be used for?

Combinded with larger repositories of datasets to compile a more comprehensive diabetes dataset. Applicable for meal frequency and/or overnight blood sugar control research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Know that the cleaned data aggregates by hour, if that is undesired, please turn to the raw dataset.

Are there tasks for which the dataset should not be used?

No.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

No specific distribution. Publicly available on GitHub.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

GitHub.

When will the dataset be distributed?

Sometime in April 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Creative Commons 4.0. Free to distribute, reuse given proper credits.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

Any other comments?

Maintenance

Who is supporting/hosting/maintaining the dataset?

Jeremy Chu, a Masters of Information student at the University of Toronto at the time of dataset distribution.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Please email jeremychuj@gmail.com for any and all inquiries.

Is there an erratum?

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes. Raw data will be uploaded annually. There will be a changelog on GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data.

No limits on data retenion. Have obtained consent for the data to remain on GitHub until told otherwise.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes in the sense that old datasets will not be removed. More new raw datasets will continuously be added.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

No mechanism for adding or building onto datasets.

Any other comments?

While there is no available mechanism for users to build onto the dataset, they are free to take the data and build onto their own datasets. If they would like to include a link to their project/datasets on this GitHub repository, please contact Jeremy Chu @ jeremychuj@gmail.com with the link and it will be added to this project's GitHub repositor