Introduction

Type 1 diabetes is a disease characterized by instability and unpredictability of blood glucose levels. The complex nature of this condition makes it challenging for patients to maintain optimal glucose control, often leading to frequent fluctuations between hypoglycemia and hyperglycemia.

Continuous Glucose Monitoring (CGM) systems represent a significant advancement in diabetes management, providing continuous rather than discrete point measurement of blood glucose levels. By continuously tracking blood glucose levels, CGM provides glucose patterns that offer information about the key behaviors that may be causing fluctuations. This detailed information is invaluable for making informed decisions regarding treatment plans and lifestyle modifications aimed at improving blood glucose control.

Despite the importance of CGM pattern recognition. There are some limitations. Most of the CGM datasets are too small to be effectively applied to machine learning or deep learning algorithms. This limitation necessitates the combination of multiple datasets. However, this approach is not without its challenges; there may be bias hidden within each dataset that could affect the accuracy and reliability of the results.

Another critical challenge is the shortage of healthy CGM datasets. To address this, there is a growing need for synthetic healthy CGM data based on real-world data. However, the reliability of these generated datasets remains a significant concern. Ensuring that the synthetic data accurately reflects real-world glucose patterns and behaviors is crucial for the efficacy of subsequent analyses and interventions.

Another aspect to consider is the relationship between HbA1c and Time in Range (TIR). HbA1c, a measure of average blood glucose levels over the past two to three months, has been found to have a significant correlation with TIR, which indicates the percentage of time a person's glucose levels remain within the target range. However, it has not yet been proven that TIR has a relationship with the diversity of CGM patterns.

Thus, in our study, we aim to verify the following four hypotheses:

- The diversity of daily CGM patterns has a linear relationship to HbA1c/TIR levels. We hypothesize that individuals with higher HbA1c/lower TIR levels will exhibit more diverse CGM patterns.
- When using multi-source datasets, there might be some bias patterns existed. This implies that different data sources could be separable through clustering techniques.
- Healthy subjects should have less diversity in CGM patterns. Even when compared to Type 1 diabetes subjects with similar HbA1c/TIR levels, healthy individuals are expected to show smaller diversity in their CGM patterns.
- We utilize dynamic time warping based techniques to synthesize healthy CGM data. This approach addresses the shortage of healthy CGM data. We aim to demonstrate that the patterns and distribution of generated CGM data are quite like those of original healthy data.

We propose using unsupervised learning to cluster daily CGM patterns and evaluate their distribution over a monitoring span. While traditional methods like k-means are susceptible to outliers, and DBSCAN struggles with border points, we aim to develop advanced techniques to mitigate these issues.

In “Identification of clinically relevant deglycation phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c.”, [9] the authors use

some advanced glucose variability indexes to evaluate and do CGM phenotype clustering. But these glucose variability indexes only represent the coarse grain variability. And both "Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study."[10] and “The Development and Potential Applications of an Automated Method for Detecting and Classifying Continuous Glucose Monitoring Patterns” [11] use the combination of short windows (< 4 hours) CGM pattern clustering and further hierarchical clustering to aggregate the short-term patterns. But this method may lose the general view about the daily CGM pattern. In “A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series.” [4], the authors use a new “motif based” clustering method which generates a set of dense, compact clusters of daily CGM pattern. The method avoids the pitfalls of k means method and DBSCAN method and provides more general view for daily CGM pattern than hierarchical clustering method. But the rigidity of RMSE distance measurements and hyperparameters setting makes the final set have over 400 motifs, which makes it too complex to be applied to the real world.

We propose our method to leverage the “motif based” method. But we replace RMSE distance measurements with dynamic time warping (DTW) distance measurements to solve the dynamic issue of daily CGM pattern. We also fine tune the hyperparameters to achieve the final motif set size under 100. Then based on the clusters result of daily CGM pattern, we evaluate the diversity and distribution of the daily CGM clusters of a proper monitoring span.

Methodology

Data Sources Introduction

We found an open-source data [1] which is a combination of multiple type1, type2 and healthy CGM datasets. Because the type 2 dataset it contains is quite small, we decided to ignore type 2 cases. We then selected several datasets based on subject numbers (ranging from dozens to hundreds), and it forms our multi-source CGM dataset. The multi-source dataset encompasses a total of 1236 subjects with Type 1 diabetes, with data collected over a period ranging from 3 to 6 months. Additionally, there are 418 healthy subjects with data spanning from one day to several days. (table 1)

Study summary of different datasets

- Aleppo : The purpose of this study was to determine whether the use of continuous glucose monitoring (CGM) without blood glucose monitoring (BGM) measurements is as safe and effective as using CGM with BGM in adults (25-40) with type 1 diabetes under insulin pump treatment.

- Brown : In this study, the patients with type 1 diabetes were assigned in a 2:1 ratio to receive treatment with a closed-loop system (closed-loop group) or a sensor-augmented pump (control group).

- Lynch : To evaluate a transition from standard-of-care (SC) management of type 1 diabetes (any insulin delivery method including hybrid closed-loop systems plus real-time continuous glucose monitoring [CGM]) to use of the insulin-only configuration of the iLet® bionic pancreas (BP)

- Tamborlane : This study was designed to test CGM as a technology to assist in diabetes care. The randomized trial was intended to determine if CGM usage had a positive effect on diabetes management.

- Wadwa : The study is about young children (ages 2–6) with Type 1 diabetes using the t:slim X2 insulin pump with Control-IQ Technology and Dexcom G6 CGM

- Colas : This study includes 208 subjects all of whom were healthy at study start and 17 of whom developed type 2 diabetes by study end.

- Hall : This study analyzes how blood glucose fluctuates in healthy individuals by using a CGM to monitor glucose. Standardized meals (breakfast only) were given to a subset of patients in order to monitor the effect of meals on the glucose readings of healthy individuals.

- Shah : to evaluate glucose control in a mixed population (ages 6 and older) and to establish reference sensor glucose ranges in healthy, non-diabetic individuals

Trial name	Sample sizes	Diabetes Type	Population Group(age)	CGM device	duration
Aleppo	225	1	25-40	Dexcom G4	6 months
Brown	168	1	14+	Dexcom G6 & t:Slim X2 with Control-IQ Technology	6 months
Lynch	90	1	6-71	Dexcom G6	13 weeks
Tamborlane	451	1	8+	FreeStyle Navigator & Dexcom SEVEN & Medtronic Paradigm	6 months
Wadwa	102	1	2-6	Dexcom G6	13 weeks
Colas	208 (17)	Healthy(type 2)	18+	iPro	<=2 day
Hall	57	Healthy & Pre-diabetes	18+	Dexcom G4	2-4 weeks
Shah	168	healthy	6+	Dexcom G6	<=10 days

Table 1. Summary of Multi-Source Dataset

Additionally, we have another large dataset consisting of 736 patients with Type 1 Diabetes, monitored over a span of more than 4 years using the FreeStyle Libre device[2]. This dataset provides extensive insights into the long-term glucose control and variability in individuals with T1D.

We then filter out the subject with too much missing data with the following criteria [1]:

- >=10% missing data for 1-day records
- >=30% missing data for records spanning over 2 days

After filtering, we align and chunk each individual's data by:

- If there is a missing gap for over 3 hours, chop the data.
- Linearly interpolate data to every minute.
- For each continuous CGM time series, make it start at 00:00 and end at 23:59 to ensure each daily data can be aligned in tensor.
- For the healthy subjects, we

concatenate all data together given the minimal dynamic changes in daily patterns, allowing us to treat the data as continuous due to the lack of significant overnight variations.

We choose the 3 hours gap as our largest tolerance for linear interpolation although a review paper [5] suggests only interpolate gaps that are less than about 20 minutes. But since it may generate too many small chunks in pattern evaluation tasks and we are going to apply smoothing later, we make a trade-off and finally decide 3 hours.

Figure 1 illustrates the filtering process, alignment, and chunking of the data, as well as handling of missing data gaps and interpolation.

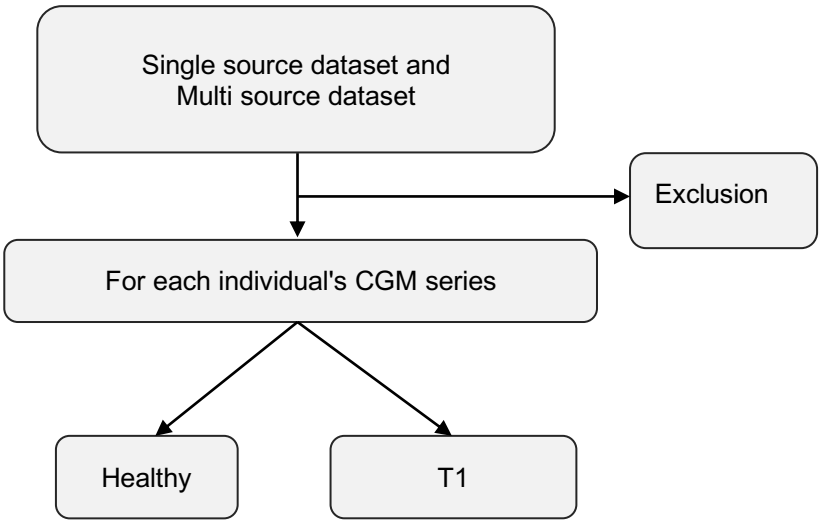

Figure 1. Flowchart of filtering and chunking CGM data

Exclusion

- >=10% missing data for 1-day records.
- >=30% missing data for records spanning over 2 days.

For each individual's CGM series

1. If there is a missing gap for over 3 hours, chop the data.
2. Linear interpolate each continuous data to every minute.
3. For each continuous CGM time series, make it start at 00:00 and end at 23:59 to ensure each daily data can be aligned in tensor.

Healthy

- Concatenate all data into one.

T1

- Don’t combine data together, as if there is a large missing gap between each continuous data, there might be some dramatic, unreasonable change when concatenate them together.

Generation of Healthy Data

We leverage two methods to generate healthy data [3]. If there is only one day data for a subject after above filtering and chunking, we first apply magnitude warping and time warping to generate over 5 days data. Then based on the data, use random guided warping to generate more data to expand total data span to 90 days. [figure 2] Random guided warping is an interpolation technique using dynamic time

warping to align two time series instead of definite time points.

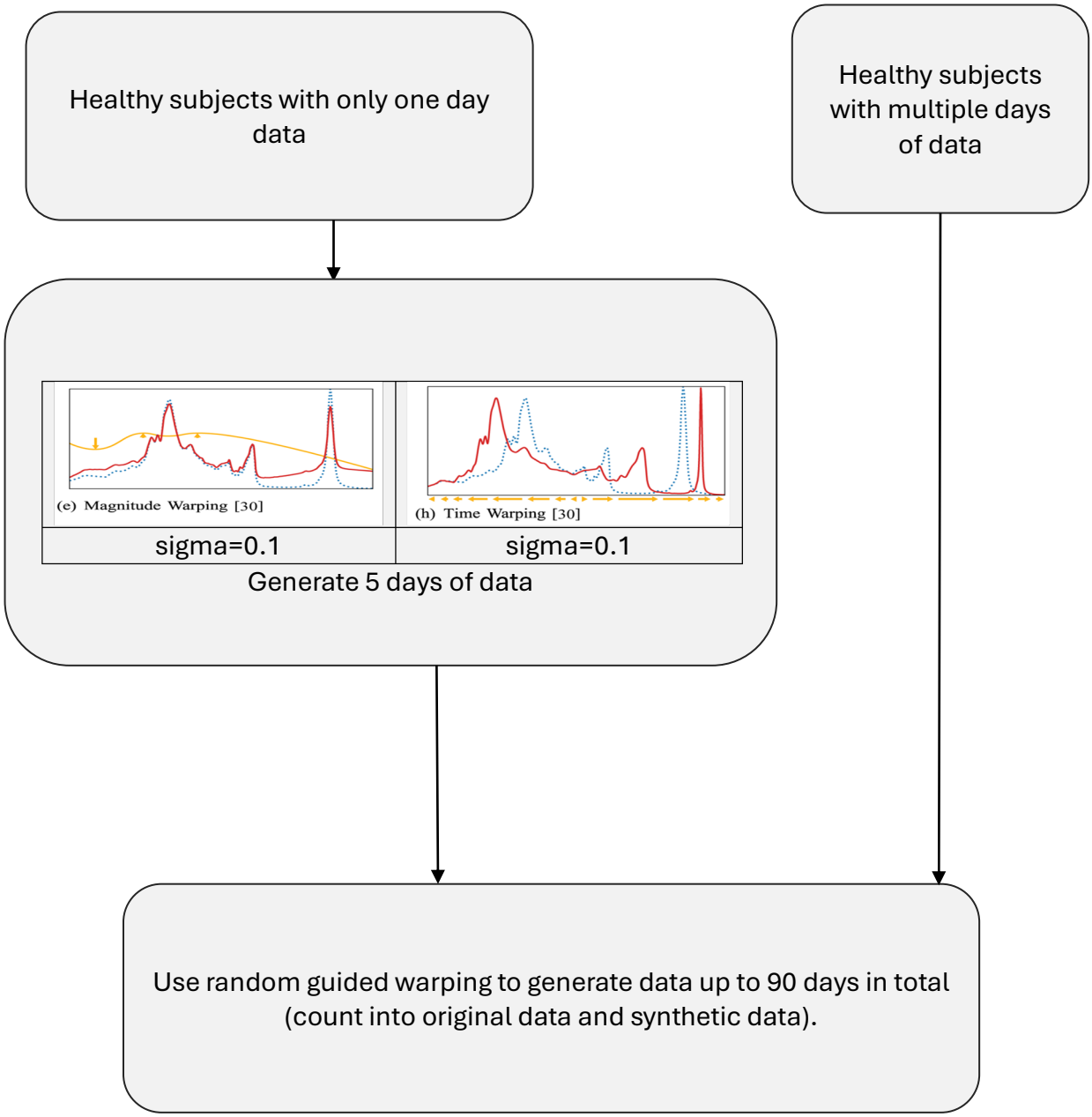

Figure 2. Process of generating synthetic healthy data from the existing records.

Clustering Methodology for CGM Data

Our approach to clustering Continuous Glucose Monitoring (CGM) data is inspired by the method proposed in the paper, "A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series." [4] While we largely adhere to the methodology outlined in the paper, we introduce several modifications.

Algorithm introduction

The same as original algorithm, we first transform the data using the following formula:

$$g(x_t) = \ln(x_t)^{1.084} - 5.381$$

Later, we calculate the distance between time series using dynamic time warping instead of original RMSE.

Then we follow the following steps to find out all the representative daily CGM motifs.

Given a set of daily profiles φ that have yet to be assigned to a cluster.

Find the pair of daily CGM profiles, say dp_x and dp_y, such that

$$f(dp_x, dp_y) \leq f(dp_j, dp_k)$$

and

$$f(dp_x, dp_y) < \gamma$$

for all $j,k \in \varphi$. The score $f(dp_x, dp_y)$ is the minimum score among all pairs of daily CGM profiles in φ that is also strictly less than γ. Where f is the dynamic time warping distance calculation function in our definition. The parameter γ serves as a threshold criterion – the *maximum* possible score between two daily CGM profiles that will define a motif m_i.

The dp_x and dp_y are then defined as a motif pair. Then for each daily CGM profile $dp_j \in \varphi \setminus \{dp_x, dp_y\}$, remove dp_j from φ if and only if

$$f(dp_x, dp_j) \leq f(dp_x, dp_y) + \tau$$

or

$$f(dp_y, dp_j) \leq f(dp_x, dp_y) + \tau$$

where τ is a tolerance value describing how close the match between dp_j and one of dp_x or dp_y must be for dp_j to be removed from φ.

Then do the above steps iteratively until no further motif found.

Dividing Training, Validation, and Testing Datasets

The original paper implements specific criteria for filtering out daily CGM data that exhibits abrupt increases or decreases, as well as data with excessive missing values. Our modifications to this process are as follows: we will use the filtering method described in Figure 1. And in place of the original criteria for filtering out abrupt increases or decreases, we apply a smoothing method to the data.

The original paper randomly assigns daily CGM data into training, validation, and testing sets in 15%, 20%, and 65% split respectively. However, there are some pitfalls hidden within this approach. Firstly, for a single subject, different daily data can be distributed into training, validation, and test sets, which may make the validation and test set results look remarkably satisfying. But because a subject's data could be in the training, validation, and test sets at the same time, the representativeness of the final motifs becomes questionable, as they could only represent the pattern distribution of the whole dataset.

To address this, if we ensure the data from a single subject could only be assigned to one of training, validation, and test sets, the result motif set will be more representative, making the conclusions more robust.[5] Secondly, the original splitting method does not employ stratification, which means some representative patterns related to certain demographics may be missing in the training set.[5] Therefore, instead of the original random assignment, we will implement a stratified splitting strategy to ensure that all demographic patterns are adequately represented.

Furthermore, because we have a multi-source dataset and a single-source dataset, if we do clustering on one of the datasets, we could use another dataset as an external validation set.[5] This approach will help verify the robustness and generalizability of our clustering results across different data contexts.

Because we aim to evaluate the diversity of CGM patterns over a monitoring span, we will first filter out subjects who lack continuous data for over 7 days, which is our primitive monitoring span. The data

from these subjects will be evenly divided into training and validation sets. We will then stratify the T1 subjects included according to four demographics:

- Sex
- age group (<10, <20, 20-65, ≥65)
- mean HbA1c group (<7%, <8.5%, ≥8.5%)
- treatment group (multiple daily injection, basic pump, sensor-augmented pump, closed loop, bionic pancreas)

Healthy subjects will be stratified based on:

- sex
- age group (<10, <20, 20-65, ≥65)

With age groups roughly representing prepuberty, adolescence, adulthood, and elderly stages. The mean HbA1c group stratification is based on guidelines, with <7% indicating good control, 7-8.5% moderate control, and ≥8.5% poor control. [6][7]. And we can only stratify the t1 subjects from single source dataset based on sex, age and hba1c because the single source dataset lacks the information about treatment group.

Based on the above ideas, we form our flow chart of dataset choosing and splitting based on our four main objectives. [figure 3]

Apply algorithm and CGM pattern distribution evaluation

We then apply different data sources to the motif clustering algorithm according to different objectives. For the training, we apply both the final training set and the CGM data that has been filtered out at the initial in figure 3. For objective 1.: to verify the relationship between diversity of daily CGM patterns and TIR/HbA1C. We are going to apply single source dataset to avoid possible bias hidden inside multi source dataset. And for the remaining objectives: 2. To find out if there is any possible bias hidden inside multi source dataset 3. To verify for the same TIR/HbA1C background, is t1 subjects' CGM pattern more dynamic than healthy subjects? 4. To verify if there are any pattern differences between real healthy data and synthetic healthy data. We are going to apply multi source dataset, with synthetic healthy data included.

TIR for each chunked data. For further evaluation, we will randomly select out some data according to the following multi-labels.

- TIR group (<=50%, >50%, >70%, >90%) [7]
- Sex
- Age group
- Treatment group

We then go for two directions. The first is to evaluate the diversity of CGM pattern distribution, which we can consider as a similar idea to the amount of uncertainty in an entire probability distribution of CGM patterns.[8] In this case, we consider using Shannon entropy to represent diversity. The second is to evaluate the similarity between two distributions of CGM patterns. For this task, we can leverage the idea of relative entropy, like Kullback-Leibler (KL) Divergence, with $D_{KL}(P\|Q)$, which represents how many extra "bits" are needed on average to transmit a value drawn from distribution P when we use a code that was designed for another distribution Q.[8] And because of the asymmetric nature of KL divergence, we define final dissimilarity between the two CGM pattern distribution as

$$\min\{D_{KL}(P|Q), D_{KL}(Q|P)\}$$

Where P and Q represent CGM pattern distribution comes from different test cases.

For the similarity evaluation, we aim to calculate both intra-trial dissimilarity and inter-trial dissimilarity over several pairs of test cases coming from the same multi-labeling class.

And based on the logic of the motif-based algorithm, there might be some daily pattern being classified as "other". We

treat each “other” case as a different, additional pattern. [figure 4]

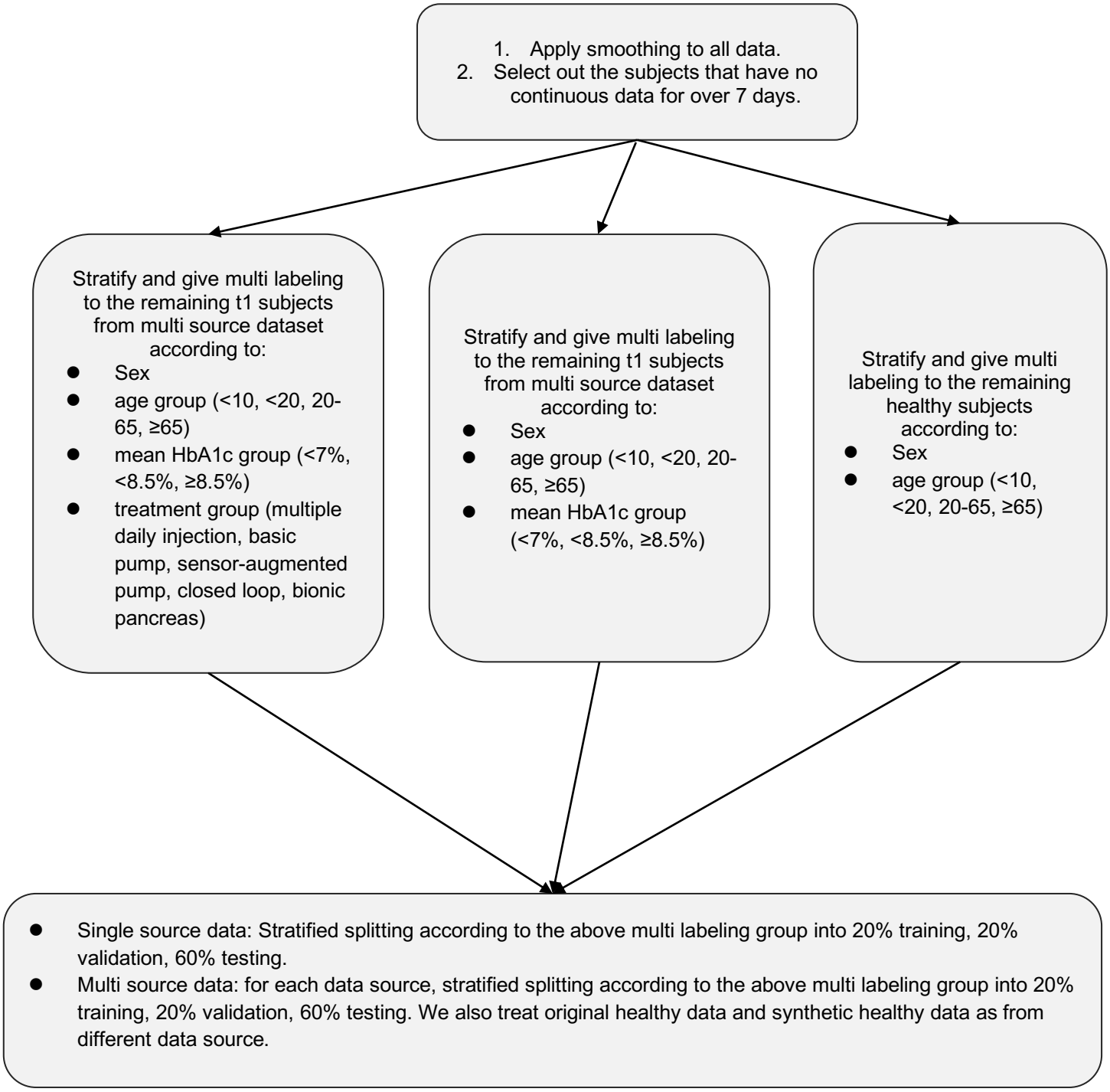

Figure 3. process of splitting training, validation and test set

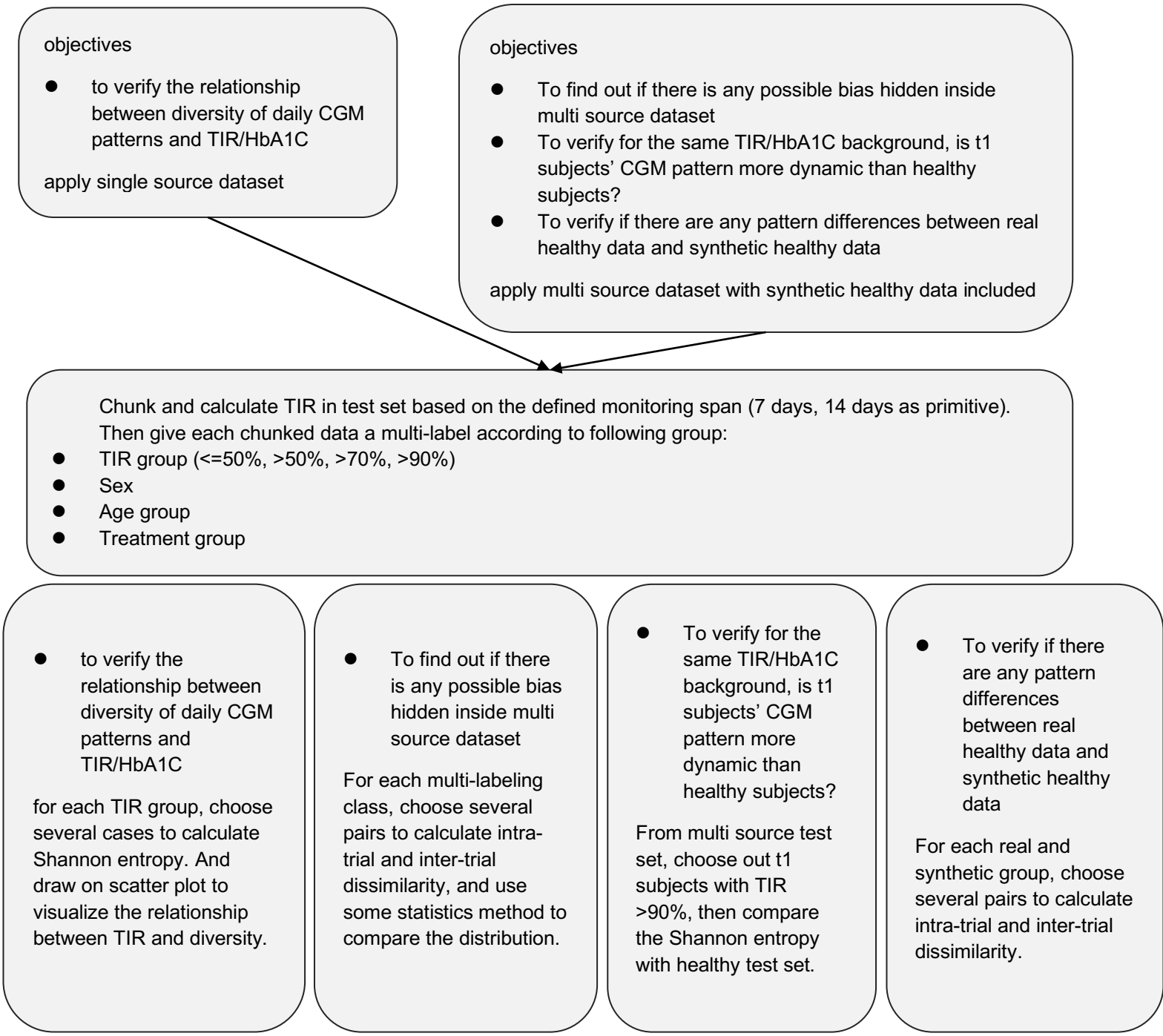

Figure 4. process of applying algorithm and pattern evaluation

References

[1] Xinran Xu, Neo Kok, Junyan Tan, Mary Martin, David Buchanan, Elizabeth Chun, Rucha Bhat, Shaun Cass, Eric Wang, Sangaman Senthil, & Irina Gaynanova. (2024). IrinaStatsLab/Awesome-CGM: Updated release with additional public CGM dataset and enhanced processing (v2.0.0). Zenodo. https://doi.org/10.5281/zenodo.14541646

[2] Rodriguez-Leon, C., Aviles Perez, M. D., Banos, O., Quesada-Charneco, M., Lopez-Ibarra, P. J., Villalonga, C., & Munoz-Torres, M. (2023). T1DiabetesGranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus

[Data set]. Zenodo. https://doi.org/10.5281/zenodo.10050944

[3] Iwana, Brian Kenji, and Seiichi Uchida. "An empirical survey of data augmentation for time series classification with neural networks." *Plos one* 16.7 (2021): e0254841.

[4] Lobo, Benjamin, et al. "A data-driven approach to classifying daily continuous glucose monitoring (CGM) time series." *IEEE Transactions on Biomedical Engineering* 69.2 (2021): 654-665.

[5] Jacobs, Peter G., et al. "Artificial intelligence and machine learning for improving glycemic control in diabetes: Best practices, pitfalls, and opportunities." *IEEE reviews in biomedical engineering* 17 (2023): 19-41.

[6] 6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes—2025 | Diabetes Care | American Diabetes Association

[7] Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range | Diabetes Care | American Diabetes Association

[8] https://nthu-datalab.github.io/ml/slides/03_Probability_Info-Theory.pdf

[9] Kahkoska, Anna R., et al. "Identification of clinically relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c." *Pediatric diabetes* 20.5 (2019): 556-566.

[10] Chan, Nicholas Berin, et al. "Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study." *JMIR AI* 2.1 (2023): e45450.

[11] Shomali, Mansur, et al. "The development and potential applications of an automated method for detecting and classifying continuous glucose monitoring patterns." *Journal of Diabetes Science and Technology* (2024): 19322968241232378.