UNIVERSITÄT
BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

AI in Health and Nutrition

Data Driven Diabetes Management

Course Project 6: Imputation Model for Glucose Values Out of Measuring Range for Continuous Glucose Monitors

Title	**Imputation Model for Glucose Values Out of Measuring Range for Continuous Glucose Monitors**
Author	Maria Panagiotou, MSc, PhD Fellow Knut Joar Strømmen, MSc, PhD Fellow
Editor	Stavroula Mougiakakou, PhD

UNIVERSITÄT
BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

Bern, September 2024
Course: Data Driven Diabetes Management
Master of Science in Biomedical Engineering

1. Scope of the Project

Continuous glucose monitoring (CGM) has revolutionised out-hospital diabetes management and increased patient satisfaction and is an established technology in diabetes research. The advantages of CGM compared to point-of-care glucose testing are numerous and include the continuous measure of glucose levels every 5 minutes, alarms on hypo- and hyperglycemia, and trend arrows indicating rapidly changing glucose levels.

Despite the advantages of CGM, challenges regarding the accuracy, practical implementation, and technical challenges of CGMs have been reported. The CGM time series data can be challenging to handle and interpret. Therefore, great effort has been made to derive useful summary statistics of CGM data like the Ambulatory Glucose Profile [1]. However, most CGMs have a limited measuring range which has not been addressed statistically.

Most CGMs have an upper detection limit of 22.2 mmol/L (400 mg/dL) and a lower detection limit of 2.2 mmol/L (40 mg/dL). The upper and lower detection limits result in censoring of the CGM data which biases standard CGM metrics recommended to be reported by international consensus in all clinical studies in diabetes utilizing CGM. The detection limits of CGMs might lead to a downward bias for metrics of glycemic variability, i.e., standard deviation (SD) of all CGM-glucose levels and coefficient of variation (CV). Mean glucose level and estimated A1c (i.e., glucose management index) can also be downward biased during severe hyperglycemia and to a lesser extent upward biased during severe hypoglycemia both exceeding the detection limits of CGMs.

You will in this project develop and validate machine learning models to impute (i.e., substitute) censored CGM data above the upper detection limit of all CGMs currently available. This enables more accurate quantification of CGM metrics set by international consensus.

2. Data

You will be working with recorded data from 12 different individuals with T1D. The data was released in the OhiaT1DM [2] dataset. You will have access to information such as continuous glucose monitoring (CGM), Blood glucose values obtained through self-monitoring by the patient (finger stick), basal insulin rate, bolus injection, the self-reported time and type of a meal, plus the patient's carbohydrate estimate for the meal and more. The measurements are provided at intervals of minutes.

3. Experiment

As we do not have access to data that is above the upper detection limit, we need to make an artificial upper limit to validate whether the model generalizes to ranges not observed in the training data. To do this, you will first censor the CGM samples above the 80th percentile. These censored samples will be used to test how well the algorithms perform at imputing data in ranges they have not seen during training. The remaining data will be used to train the models. Note, the actual values above the 80th percentiles should only be used for testing and should **not** be included in the training process. The performance of the imputation model will be evaluated using bias and mean squared error (MSE) on key CGM metrics, such as mean glucose level, standard deviation (SD), and coefficient of variation (CV).

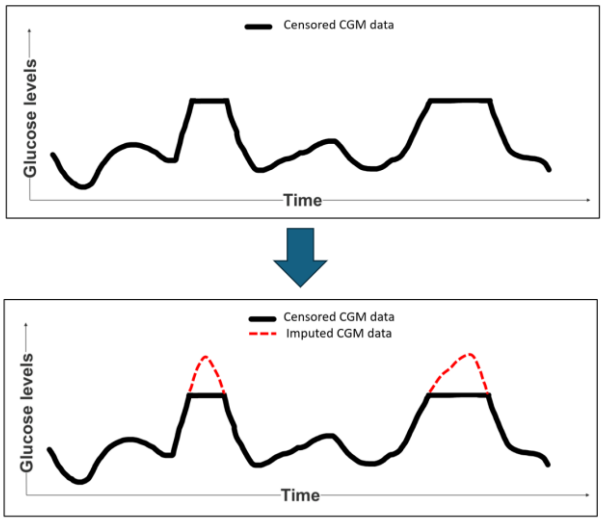
Hint 1: After removing samples above the 80th percentile, you will have segments of data with missing patches. To train a model for imputation, a straightforward approach is to "randomly" mask additional portions of the remaining data. The model's inputs are the masked sequences from the time series, while the outputs are the original unmasked sequences.

Hint 3: The masked patches should be of a similar length to the missing patches in the missing data.

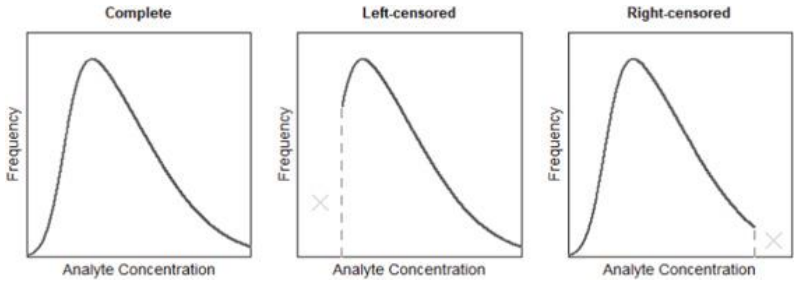
Hint 3: Choose input sequence lengths that matches or exceed the longest missing patches in the censored data.

Hint 4: Explore time series imputation models, such as those in the PyPOTS library. Start with simple models like Linear Regression to establish baseline performance. For an overview of the performance of some recent approaches for time series imputation take a look at TSI-Bench [3]

Hint 6: Test models on the censored data (above the 80th percentile) to assess their ability to impute values in unseen ranges.



Bonus question: CGM values above the upper detection limit are right-censored, meaning that higher values in the distribution are not observed due to the artificial cut-off. Randomly removing patches of data from the training set does not effectively simulate this right-censoring, as it doesn't specifically target the higher values that are missing due to the detection limit. **How can we mask the input data to more closely mimic the effects of the actual cut-off limits?**



4. Report

We encourage you to include the following sections in your report:

- Introduction: This section should include a brief presentation of the project's aims, objectives, and its clinical importance. You should briefly explain your basic approach and your main conclusions. If needed add a figure.
- Related work: This section should highlight previous work related to your problem and should put your work in a broader context.

- Methods: Here you describe the method/s you implemented in detail.
- Data and Experiment setup: Data description, preprocessing. Add a table with characteristics of the data, or an example of the data available for a specific individual, before and after any pre-processing. Describe your benchmarks.
- Results: Present the results of your analyses (use graphs and/or tables). Comment on these results: are they statistically significant? Are there interesting trends?
- Discussion: Highlight how your results relate to your original question formulation. Do they support your hypothesis? Discuss limitations with your analyses and how they might motivate future research directions.

References

[1] Johnson, Mary L., Thomas W. Martens, Amy B. Criego, Anders L. Carlson, Gregg D. Simonson, and Richard M. Bergenstal. 'Utilizing the Ambulatory Glucose Profile to Standardize and Implement Continuous Glucose Monitoring in Clinical Practice'. Diabetes Technology & Therapeutics 21, no. S2 (June 2019): S217–25. https://doi.org/10.1089/dia.2019.0034.

[2] Marling, Cindy, and Razvan Bunescu. 'The OhioT1DM Dataset for Blood Glucose Level Prediction: Update 2020', n.d.

[3] Du, Wenjie, Jun Wang, Linglong Qian, Yiyuan Yang, Fanxing Liu, Zepu Wang, Zina Ibrahim, et al. 'TSI-Bench: BenchmarkingTime Series Imputation'. arXiv, 18 June 2024. http://arxiv.org/abs/2406.12747.