The Effect of Missing Continuous Glucose Monitoring Data on Time in Range

Kevin Nguyen

December 2, 2022

1 Executive Summary

Continuous glucose monitors (CGMs) are wearable devices that measure interstitial glucose levels continuously throughout the day, with some monitors taking measurements as often as every five minutes. The time series nature of CGM data provides a characterization of the temporal glucose profile. Thus, the use of CGMs has become more prevalent in clinical practice, specifically in diabetes. However, gaps in CGM data are fairly common and can occur for various reasons. One of the most popular CGM-derived metrics is time in range (TIR). TIR describes the percentage of glucose readings between 70 and 180 mg/dL and is the summary statistic of interest. According to the client, it is common practice to only require 110 out of 288 data points to calculate an acceptable TIR value. The client, on the other hand, argues that with more than half of the time series missing, some TIR estimates will be incorrect.

In response to the first question, while we cannot infer if missing data as a whole introduces a clinically significant difference in TIR, we can deduce from logistic regression that the amount of missing data does introduce a clinically significant difference in TIR. In response

to the second question, our analysis found the proportion of clinically significant differences increases as the amount of missing data increases. Therefore, greater amounts of missing data will result in a clinically significant difference. According to the logistic regression function, when the amount of missing data is increased to 178 ($288 - 110 = 178$), the maximum number of missing values permitted in common practice to calculate an acceptable TIR, the proportion of clinically significant differences climbs to 0.874. Thus, it appears our client is correct in arguing with more than half of the time series missing, some TIR estimates will be incorrect.

2 Introduction

2.1 General Background

Continuous glucose monitors (CGMs) are wearable devices that measure interstitial glucose levels continuously throughout the day, with some monitors taking measurements as often as every five minutes. The time series nature of CGM data provides a characterization of the temporal glucose profile. Thus, the use of CGMs has become more prevalent in clinical practice, specifically in diabetes, since they provide instantaneous responses to therapy decisions, lifestyle modifications, and identification of patterns of hypoglycemia. However, gaps in CGM data are fairly common and can occur for various reasons, such as the battery needing to be replaced or the sensor loosening after a few weeks. Additionally, a warm-up period is necessary for the CGM to calibrate, which also prevents readings during this time. One of the most popular CGM-derived metrics is time in range (TIR). TIR describes the percentage of glucose readings between 70 and 180 mg/dL and is the summary statistic of interest. According to the client, it is common practice to only require 110 out of 288 data points to calculate an acceptable TIR value. The client, on the other hand, argues that with more than half of the time series missing, some TIR estimates will be incorrect.

2.2 Objectives

This consultation's primary objective is to provide answers to the following two questions:

1. In general, does missing data introduce a clinically significant difference in TIR?
2. More specifically, can we characterize what amounts of missing data result in a clinically significant difference in TIR?

2.3 Data Description

The client's dataset consists of 9,614 rows and 304 columns. The dataset contains the following variables:

- **SID:** The subject ID.
- **Day:** The day of the trial the time series data corresponds to for the respective subject.
- **Motif_Idx:** An integer indicating a group that the full daily time series belongs to. In previous research, the client has identified 483 groups that represent almost any full daily time series.
- **mean_BG, TIR, TAR, TBR, LBGI, HBGI, SD, CV, AGP_VH, AGP_H, AGP_T, AGP_L, AGP_VL:** The value of the summary statistics for the full daily time series.
- **0, 1, . . . , 286, 287:** The index of the 5-minute interval that the blood glucose value belongs to in the full daily time series.

3 Approach to Project

3.1 Data Cleaning

The data cleaning process mainly consisted of removing irrelevant variables. We removed the **SID** and **Day** variables because they were primarily included for the client to be able to locate more information about the time series if necessary. The **mean_BG**, **TAR**, **TBR**, **LBGI**, **HBGI**, **SD**, **CV**, **AGP_VH**, **AGP_H**, **AGP_T**, **AGP_L**, and **AGP_VL** variables were removed because TIR is the main summary statistic we are interested in for this consultation. Finally, we converted the **Motif_Idx**, **TIR**, and **0, 1, . . . , 286, 287** variables to a numeric data type.

3.2 Methodology

3.2.1 Missing Data Algorithm

We began by developing an algorithm that simulates missing data to determine the proportion of clinically significant differences for each amount of missing data. For each combination of patient, amount of missing data, and starting point of the missing data, the algorithm calculates the TIR. Then, it calculates the percentage difference between the simulated TIR and the TIR in the dataset. Lastly, it calculates the proportion of the percentage differences that are greater than or equal to 5. According to the client, a 5% increase or decrease in TIR is considered a clinically significant difference. Ideally, we would like to test every possible combination. But, 9614 possible patients $\times$ 287 possible quantities of missing data $\times$ 288 possible starting points of missing data = 794,654,784 possible combinations. We needed to find a solution to substantially reduce the number of permutations the algorithm had to complete due to computing and time constraints. Therefore, we randomly sampled a patient from each of the 483 representative groups using **Motif_Idx**. Additionally, rather than using every number between 1 and 287 for the possible quantities of missing data, we used a sequence from 1 to 281, increasing by 10 at each step, i.e., 1, 11, 21, 31, . . . , 281.

3.2.2 Logistic Regression

We implemented logistic regression to answer our client's questions. The main feature of proportion data is that they are bounded by the interval [0, 1]. Consequently, the data often exhibit heterogeneity in variance. The generalized linear model naturally handles non-normality and heterogeneity issues, and the use of a link function guarantees that the fitted values will be exactly within the desired range [0, 1]. Logistic regression, a common generalized linear model, is suitable for a binomial outcome, where the proportion is computed as the ratio of the number of target events to the total number of trials. The expected proportion is then modeled as binomial, where the explanatory variables contribute to its prediction through the use of a logit link function.

Based on a given data set of explanatory variables, logistic regression calculates the likelihood that an event will occur. A logit transformation is applied on the odds, the probability of success divided by the probability of failure, in logistic regression. The transformed odds are often referred to as log odds. In this model, the beta parameters, or coefficients, are often estimated via maximum likelihood estimation. This algorithm evaluates various beta values over a number of iterations to get the best match for the log odds. To determine the most accurate parameter estimate, logistic regression aims to maximize the log-likelihood function, which is produced by all of these iterations. Following the identification of the best coefficients, the conditional probabilities for each observation can be computed, logged, and summed to provide a predicted probability.

4 Results

Prior to implementing logistic regression, I converted the simulated data from a wide format to a long format. The simulated dataset contains the following variables: **Motif_Idx**, **Num_Missing** (amount of missing data), and **Prop_Sig** (proportion of the percentage dif-

ferences that are greater than or equal to 5). We fitted a logistic regression model, with **Prop_Sig** as the response and **Num_Missing** as the predictor, using the `glm()` function with the argument `family = "binomial"`. Figure 1 displays the output of our logistic regression model. Figure 1 shows Num_Missing is very highly statistically significant. The estimated coefficient of 0.014 for Num_Missing indicates there is a positive relationship between Prop_Sig and Num_Missing. It also implies the odds of a clinically significant difference are multiplied by $e^{0.014}$, or 1.014, for each additional missing value. Furthermore, for each possible amount of missing data, we can estimate the proportion of clinically significant differences using the logistic regression function $f(x) = \frac{e^{-0.551+0.014x}}{1+e^{-0.551+0.014x}}$. For example, if we wish to estimate the proportion of clinically significant differences when the amount of missing data is 100, the estimated proportion is equal to $\frac{e^{-0.551+0.014\times100}}{1+e^{-0.551+0.014\times100}} \approx 0.70$.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.551      0.024 -23.338        0 ***
Num_Missing     0.014   0.000198  72.515        0 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Output of Logistic Regression Model

Figure 2 displays a plot of the logistic regression function $f(x) = \frac{e^{-0.551+0.014x}}{1+e^{-0.551+0.014x}}$ for all possible quantities of missing data. Figure 2 shows as the amount of missing data increases, the proportion of clinically significant differences increases as well. The black lines in Figure 2 indicate the trajectory for each group in Motif_Idx.

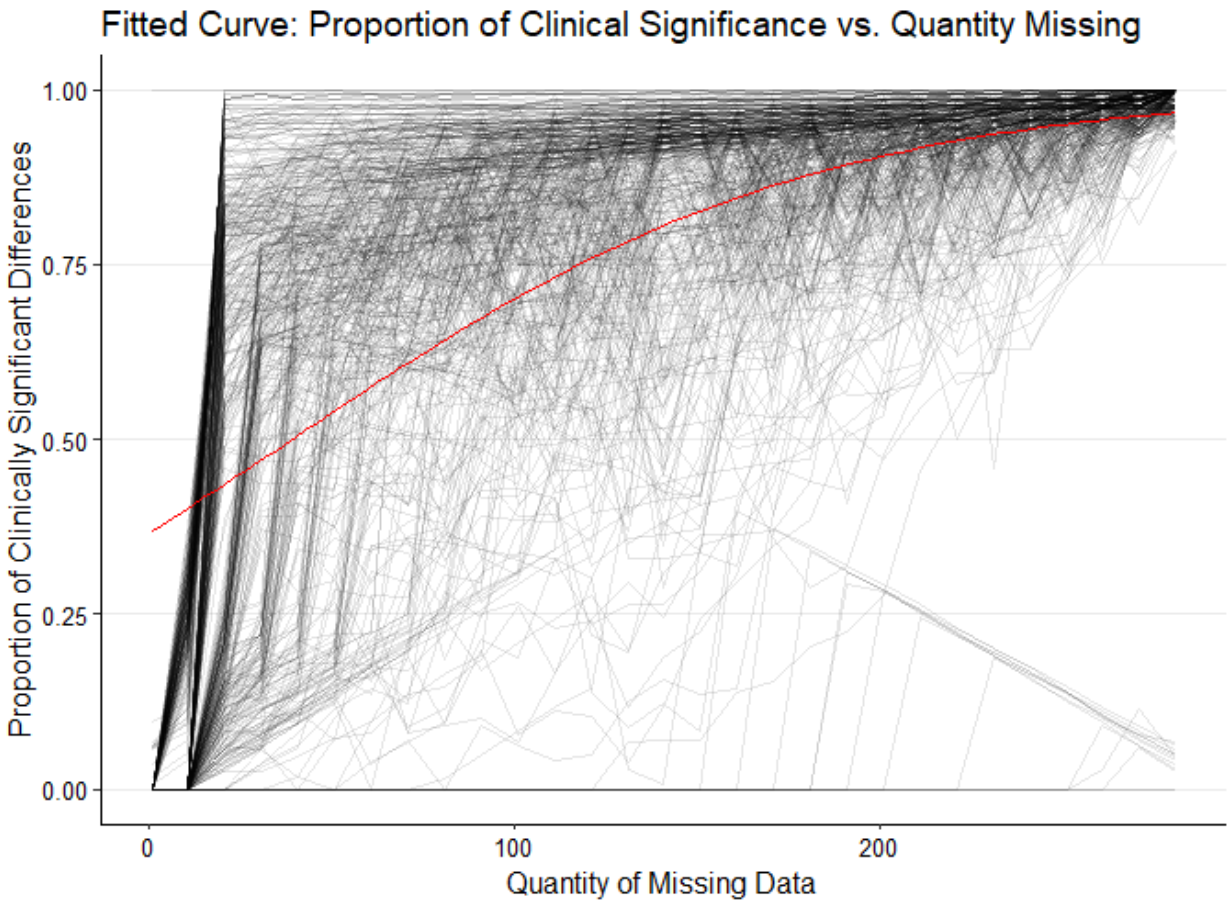

Figure 2: Logistic Regression Function Fitted Curve

Then, we implemented a likelihood ratio test to see if the model with Num_Missing as a predictor is a better fit for the data than the intercept-only model using the `anova()` function with the argument `test = "LRT"`. Figure 3 displays the output of the likelihood ratio test. Figure 3 shows Num_Missing is very highly statistically significant, which implies we should keep Num_Missing as a predictor because it is a better fit for the data than the intercept-only model.

```
Analysis of Deviance Table

Model: quasibinomial, link: logit

Response: Prop_Sig

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        14006     7184.4
Num_Missing  1   2761.1     14005     4423.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Output of Likelihood Ratio Test

5 Conclusions and Recommendations

5.1 Conclusions

In response to the first question, while we cannot infer if missing data as a whole introduces a clinically significant difference in TIR, we can deduce from logistic regression that the amount of missing data does introduce a clinically significant difference in TIR. In response to the second question, our analysis found the proportion of clinically significant differences increases as the amount of missing data increases. Therefore, greater amounts of missing data will result in a clinically significant difference. It should be noted that the logistic regression function estimates the proportion of clinically significant differences will be 0.502 for an amount of missing data as low as 40. When the amount of missing data is increased to 178 ($288 - 110 = 178$), the maximum number of missing values permitted in common practice to calculate an acceptable TIR, the proportion of clinically significant differences climbs to 0.874. Thus, it appears our client is correct in arguing with more than half of the time series missing, some TIR estimates will be incorrect.

5.2 Recommendations

If given more time and computational power, there is another approach I believe would yield more insightful results. Instead of programming the algorithm to calculate the proportion of the percentage differences that are greater than or equal to 5, the algorithm should store every permuted percentage difference and record the starting point of the missing data for each permutation. Then, we should convert the stored percentage differences into a binary variable: 1 if the percentage difference is greater than or equal to 5, and 0 otherwise. This approach enables us to implement binary logistic regression with two predictors: Num_Missing, the amount of missing data, and a new variable Start_Pt, the starting point of the missing data. Furthermore, we would be able to validate the accuracy of our model through the use of a confusion matrix. We recommend running the algorithm through all 794,654,784 possible combinations, whether it is for the approach described in this section or the approach used for this consultation, for improved results.