SAN: Example Final Assignment and Work Plan

Jan Blaha

November 9, 2023

Contents

1 About

This document serves as an example of the topic for the final assignment of the SAN course. It also contains an example of how to approach the work plan. It is not a full example; you can be more specific about what you want to do and how, but it should help you get an idea of how to think about your problem. You can take it as an inspiration for how you can prepare your work and submission.

In your work plan, you are not expected to know all the answers, but you should document the questions.

2 Assignment

2.1 Question

Developments in the management of Type 1 diabetes in terms of time-in-range metrics (TIR) for patients using continuous glucose monitoring (CGM) in the last 20 years. Did the quality of compensation metrics improve for patients with newer versions of CGM sensors?

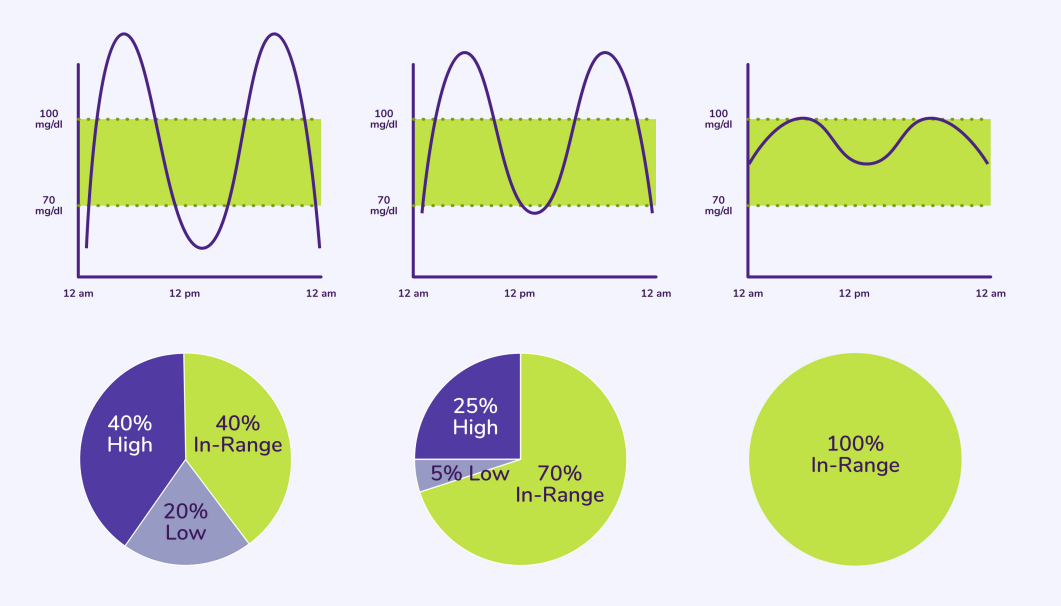

Figure 1: Different profiles and time-in-range metrics. Taken from `https://www.snaq.ai/blog/time-range-after-meal`.

2.2 Data

Datasets of glucose monitoring in patients with T1 diabetes and healty people.
`https://github.com/irinagain/Awesome-CGM`

3 Work Plan

3.1 Specific Instrumental Questions

smaller, technical questions to tackle in order to answer the big one

1. What do the datasets look like? (some exploration, ...)
2. Could we identify some cohorts in the data? It may be useful to know in further work.
3. Is the target metric lower for people from later datasets than from the earlier ones? Is the variability within individual datasets/years high?
4. Is there some specific group for which there is a significant improvement in diabetes control? How does it look?
5. Some studies focus on night/sleep time to disregard the effect of meals. When looking at time-in-target, this seems a bit contraproductive (we want to know the general quality of compensation, and that includes meals and activity). Still, is there some time of day, looking at which the results of 3. are different?
6. Do the results correspond to observations from glucose profiles in healthy people?

3.2 Datasets

`https://github.com/irinagain/Awesome-CGM`

Most of the datasets for T1 diabetes are accessible and span 2004 - 2020. They include time-series data for individual patients (usually in American mg/l, not European mmol/l), with patient information, and some even more additional data like treatments with controlled meals, etc. In size, they range from 9 to 451 patients, which seems at least somewhat sufficient. In length, they range from 4 days to 6 months.

As a reference/control group, there is also a dataset with healthy people.

3.3 Answering IQs

1. Self-explanatory. Get data together, compute TIR, plot some basic information about patients and try to see how the TIR behave. It will likely be necessary to compute it at shorter intervals than full-length datasets. Identify all extra information in the datasets that might be interesting.
2. Try to apply some dimensionality reduction techniques to your patient data. For time-series, one could precompute some features on the glucose profiles and use table-data methods from SAN lectures or see if there are any methods for time-series (maybe based on FFT?). Visualize your results; what do you see?
3. Make a linear model. Starting with $TIR \sim year\_of\_study$ see what kind of model would be appropriate—linear, non-linear, maybe even GLM? But what about different

sizes of data in each year? Due to the different sizes of study, the model would favour making fewer errors in larger studies. It would make sense to employ some reweighting. Are the data homoscedastic, and is the variance large?

4. Make a classification model that would predict based on patient data whether or not such a patient is likely to have a good TIR. Maybe calculate TIR per day, set a threshold for what is good (e.g. 80%) and try to learn a classifier that would be able to predict good/bad compensation based on patient data and other extra information (e.g. controlled meals in the day). Could you identify some patient groups specifically that would respond well?
5. Redo 3., but focus on specific parts of the day individually.
6. Compare results with healthy people for control. Are there similar relationships between patient data and TIR? Healthy people will have pretty much perfect TIR on common thresholds for diabetics, but what about some tighter targets?

3.4 Risks and Limitations

Examples of questions you can/should ask about your work.

- *Is there enough data? Are they properly sampled?*

  Data should be enough; the sampling might not be great. Most of the datasets are from the USA, and the healthcare system works differently than in Europe. This would not affect the results themselves but might limit their generalizability.

  Maybe because of the private nature of American healthcare, only wealthier people could get a CGM. It is imperative to look into the dataset documentation to know how the data were sampled from the general population.

- *Confounding? How would it affect the results?*

  Confounding could happen in the models, and it is therefore dependent on the models which will be created. This question has to be kept in mind during the work itself.

- *Sensitivity of the analysis to the TIR thresholds defining the target range.*

  The target in the time-in-range is based on some threshold. To be consistent, it makes sense to start from those used in clinical practice, but maybe one should perform some sensitivity analysis to see if the results are stable over some range of values for these thresholds.