SAN Final Assignment - PLAN

Martin Bulant, Aneta Furmanová, Daniel Klamrt, Jonáš Kříž

December 14, 2023

1 Assignment

In healthcare, prevention and prediction play a key role and in this assignment, we are going to focus on cardiovascular (CV) diseases. An important cardiovascular diseases and morbidity predictor is arterial stiffness [1]. Arterial stiffness can be estimated by multiple markers [2], e.g. non-invasively via Pulse Wave Velocity (PWV) measurement. PWV is the speed, at which the pulse wave propagates through the arterial wall.

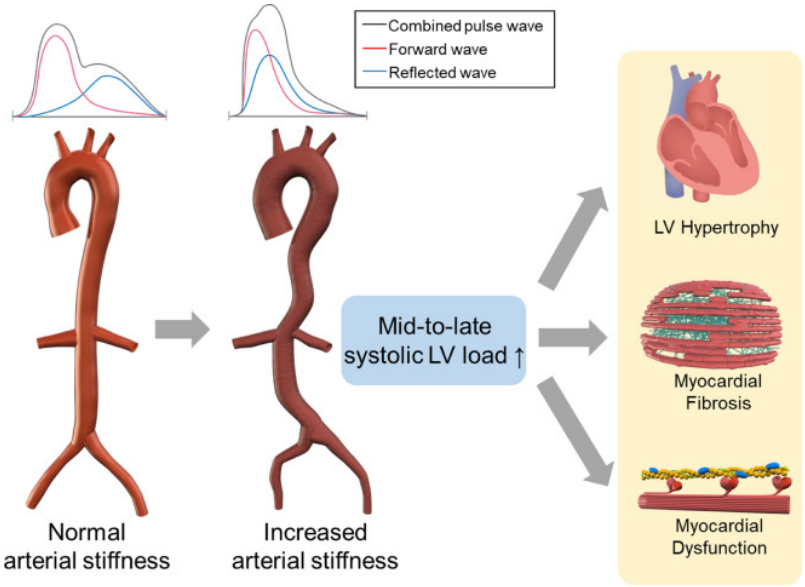

Figure 1: Arterial stiffness and pulse wave velocity [1]

For the aortic PWV estimation the distance jugulum - aortic bifurcation is needed. The real distance can be obtained from a chest MRI, but then it loses its advantage of being quite a cheap and fast method in daily practice. The currenctly used method for obtaining this distance is to use the arithmetic mean of the distances jugulum - umbilicus and jugulum - symphisis, where both are measured by the healthcare staff with a tailor's tape measure. Obviously, this is very impractical, as the measurement has a large inaccuracy, especially for obese patients. Hence, there is a need for more effective estimation of this distance in order to speed up the process of measurement and possibly reach better accuracy. The current state-of-the-art method will be used as the reference.

1.1 Question

How can be predicted the anatomical distance jugulum - aortic bifurcation without the need to measure the distances jugulum - umbilicus and jugulum - symphisis?

1.2 Dataset

The dataset was created during solving the project Apparatus for non-invasive automatic analysis of hemodynamic parameters (TH04010173) in the TAČR Starfos Programme. It contains data from 70 subjects as age, height, weight, sex, blood pressure, anamnesis, arm circumference above the elbow, measurements of anatomical distances (carotid - jugulum, jugulum - femoral artery, jugulum - umbilicus, jugulum - symphisis) and multiple PWV measurements by different devices. Possible limitations of used dataset will be discussed later in subsection 2.3.

2 Work Plan

2.1 Specific Instrumental Questions

1. What does the dataset look like?
2. What would be the optimal number of samples for this task? Can we some somehow enlarge the dataset?
3. Can be the anatomical distance jugulum - aortic bifurcation predicted only with predictors *height, weight, BMI*? (as those are the easiest to measure)
4. Are the other predictors from the dataset useful for obtaining a better model than only from predictors *height, weight, BMI*? Does it improve the model's performance (use the model from the previous question as a benchmark)?

2.2 Answering SIQs

1. Are the data homoscedastic? What is their variance? Should we employ some normalization?
2. Determine the optimal sample size for GLM from the power analysis. For enlarging dataset try following approaches:
   - Generative Adversarial Network (GAN)
   - data augmentation with Gaussian noise (where the σ of noise for each predictor depends on the σ of the predictor)
   - our naive approach:
     from $\mathcal{N}(X_0, \sigma_{X_0})$ generate x'_0
     from $\mathcal{N}(X_1(x'_0), \sigma_{X_1(x'_0)})$ generate x'_1
     ...
     from $\mathcal{N}(X_n(x'_0, x'_1, ...x'_{n-1}), \sigma_{X_n(x'_0, x'_1, ...x'_{n-1})})$ generate x'_n
     from $\mathcal{N}(Y(X), \sigma_X)$ generate y'
     where $X' = x'_0, x'_1, ...x_{n-1}$ are the predictors and y' is the output variable

3. Make a simple prediction model. Then, try multiple GLMs and estimate, if these predictors are sufficient.

4. Similarly as in previous question, try to evaluate the influence of other predictors. Employ backward stepwise selection (BSS) which starts with all available predictors and then gradually remove the predictors with the largest p-value. Then, instead of BSS try lasso and ridge regression. Since the dataset is very small, instead of test data the cross-validation can be used.

2.3 Risks and Limitations

- *Is there enough data? Are they properly sampled?*
  The dataset is relatively small, with measurements collected from only 70 subjects. Measuring one subject took quite a long time and the number of subjects was sufficient at that stage for the TAČR project. Hence, the size of the dataset is very limited, which we have to take into account.

  The data were not randomly sampled. The sample subjects were drawn from the Czech Technical University in Prague, Faculty of Electrical Engineering among students, researchers and employees, who were willing to participate. Therefore, the findings and conclusions derived from this dataset cannot be extrapolated to the entire population.

  Both the size of the obtained datased and the sampling method leads us to another question.

- *How to cope with the selection bias and causal inference?*

- *Our dataset is very small. Can we trust the result? How much?*

References

[1] Toru Miyoshi and Hiroshi Ito. "Arterial stiffness in health and disease: The role of cardio-ankle vascular index". In: *Journal of Cardiology* 78.6 (2021), pp. 493–501. ISSN: 0914-5087. DOI: https://doi.org/10.1016/j.jjcc.2021.07.011.

[2] Patrick Segers, Ernst R. Rietzschel, and Julio A. Chirinos. "How to Measure Arterial Stiffness in Humans". In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 40.5 (May 2020), pp. 1034–1043. ISSN: 1524-4636. DOI: 10.1161/atvbaha.119.313132.