Consulting Report

Date: 10/21/2022, Friday
Client name: Stephen Gan
Position: MD, Pediatrics Specialist
Affiliation: UC Hospitals
Consulting team: (Friday Team) Aabesh Bhattacharyya, Alan Zhong, Amber Lee, Chenfeng Li, Rohan Hore

Project Aim:

Currently the CPR is performed usually in the middle of the chest, while in theory it should be done where the left ventricle is largest. If we call the distance between these two points as 'hemi-thorax' distance, the goal of the project is to show that this distance is indeed significant across different age groups of children while taking into account the effect of other physical factors like height, weight etc.

Project Description:

The client currently has no dataset at this point but has a framework for collecting the data. The data will be collected on a population of children coming from 5-6 different age groups. For each child various covariates like height, weight, BMI, Body surface area (BSA) etc. will be collected. Along with that, they will also compute the hemithorax distance as have been defined above, which is the variable of primary interest in this study.

What are the issues?

- The client currently has no dataset and wanted to have a rough idea about how many samples should be collected in order to have enough statistical power and make statistically sound conclusions.
- Given this study is still in the earliest stage, the client is broadly interested in learning statistically best approaches to tackle the problem.

Specificities and particularities of the project :

We, as a group, noticed the covariates are, by intuition, highly correlated, with some even as a direct function of others (Such as BMI is a function of height and weight).

Suggestions:

1. In order to study the statistical significance of hemithorax distance after controlling the other covariates, the most simple approach could be doing a linear regression and checking significance of the intercept (i.e. the mean effect after

controlling for other variables). Considering height and weight as our other primary variables we can fit any of these two models. In the first model, we just put a linear effect of age group in the model. But the client and we suspect the effect of height and weight on our response might also vary across age groups. So we can also fit the model (b) in each age group separately and check significance.

   a. hemothorax distance ~ age group + height + weight

   b. hemithorax distance(i) ~ height(i) + weight(i), for each i in {age group}

2. We have dropped the BMI in BSA in our models, because they are supposedly highly correlated with the other variables. If we want to make this variable exclusion process more statistically sound, we can instead start with all covariates in our model and then exclude the ones with high VIF (variance inflation score) score.

3. For inferring statistical significance of the intercept, we would not just look at the raw estimate and decide manually. The linear model summary in **R** will print out the corresponding p-value for it (a probabilistic measure indicating how far the estimate is away from 0). A general rule of thumb is to infer significance when this p value is less than 0.05.

4. Now for making proper statistical inference, we should have enough data. Since there are 5-6 age groups, we suggest having at-least 300 data points with 50-60 observations in each group.
   Note: To see whether our inference through the notion of p-value is correct or not, we can also look at the qq-plot of residuals to check the normality assumption for linear models.