Data Preparation Report

Utrecht University

Mike Vink

April 28, 2021

Contents

1 Data selection

The data that we use in this work is based on the data used in the simon manuscript (Tomic et al., 2019). This subset of the fluprint database comprises data from 5 clinical studies, most importantly the longitudal study SLVP015. It only uses the first visits of donors, as the classification is the most complete in this dataset (Figure 1). We will use this dataset to model the high vaccine response with as predictors the in total 3285 features measured in assays done by the clinical studies. The data will be prepared in the same way as in the original work, using the mulset algorithm (Tomic et al., 2019).

In addition to repeating a similar procedure as in Tomic et al., 2019, we will compare the values of features selected by the models trained on the first visit data to those of second visit data. Initially the plan was to train new models on the second visit data, however the classes are extremely unbalanced in the data of repeat visits (Figure 1). For example in the first visit there are 65 high responders and 130 low responders, in the second visit there is only data available for 6 high responders and 44 low responders. Therefore we will only train model on first visit data, and use the knowledge gained to explore second visit data.

2 Clean data

The goal is to obtain one or more tables from the simon data suitable to train models, thus we are looking to change the data from the long format as in the database in a wide format where each column is a features measured in an assay. When attempting to do this it was discovered that some assay data contained duplicate readouts (Table 1). Since the values were all similar it was decided to aggregate the values to unique features using the mean value.

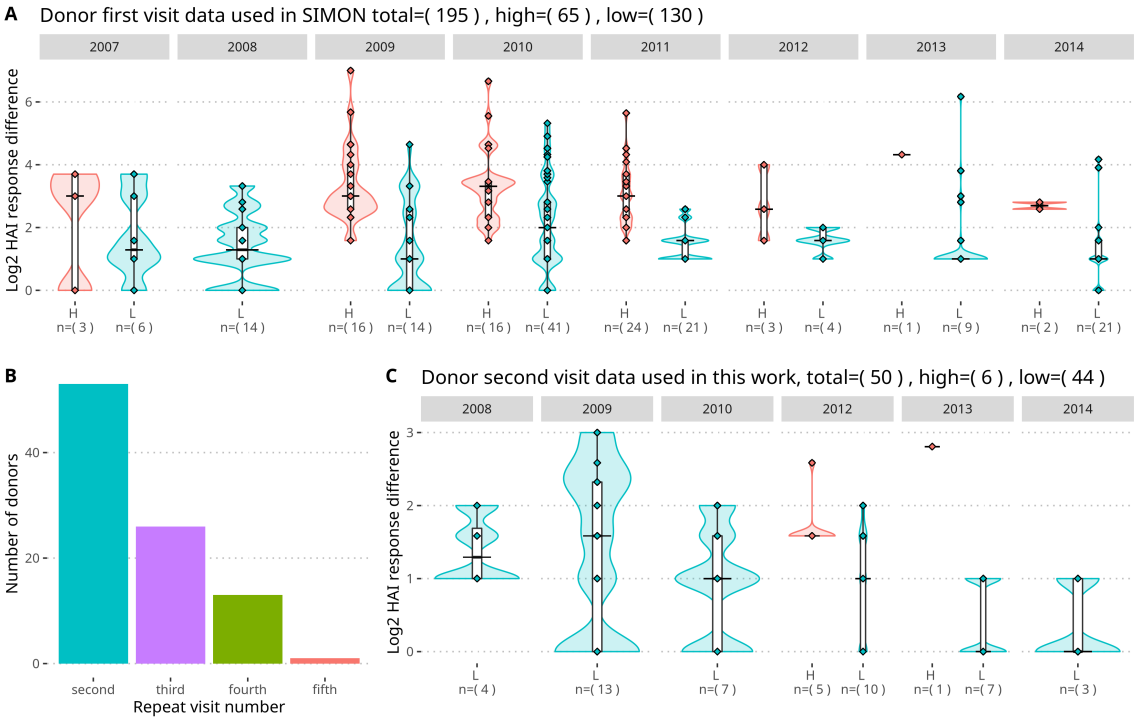

Figure 1: caption

donor_id	study	age	outcome	year	type	hai_response	name	data_name	assay	data	dup
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	33.8	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	34.1	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	34.3	TRUE
285	18	9.47	0	2009	pre	1	CD4+ T cells	CD4_pos_T_cells	13	33.0	TRUE

Table 1