Regression without regrets

M. Baillie, G. Heinze & M. Huebner

5/9/23

Table of contents

# Preface

The focus of this report is to provide guidance on conducting initial data analysis in a reproducible manner in the context of intended regression analyses.

# 1. Bacteremia study

## 1.1 Bacteremia study overview

We will exemplify our proposed systematic approach to data screening by means of a diagnostic study with the primary aim of using age, sex and 49 laboratory variables to fit a diagnostic prediction model for the bacteremia status (= presence of bacteria in the blood stream) of a blood sample. A secondary aim of the study is to describe the functional form of each predictor in the model. Between January 2006 and December 2010, patients with the clinical suspicion to suffer from bacteremia were included if blood culture analysis was requested by the responsible physician and blood was sampled for assessment of hematology and biochemistry. An analysis of this study can be found in Ratzinger et al. (2014).

The data consists of 14,691 observations from different patients and 51 potential predictors. To protect data privacy our version of this data was slightly modified compared to the original, and this modified version was cleared by the Medical University of Vienna for public use (**DC 2019-0054**). Compared to the official results given in (Ratzinger et al. 2014), our results may differ to a negligible degree.

## 1.2 Where to access the data?

We refer to the **source** data as the raw data set available in this repository (**DC 2019-0054**). The data set is published on [Zenodo](https://doi.org/10.5281/zenodo.7554815) with the following doi: https://doi.org/10.5281/zenodo.7554815.

For simplicity, we have also stored the *source* data and accompanying materials such as the **data dictionary** the data-raw directory. The data dictionary provides an overview of the collected *source* data - see [Appendix A.1](#sec-data_dict) for further details. Within the appendix, we also display a short snapshot of source data set from the data-raw folder of the project directory. The snapshot provides a glimpse of the data and the data dictionary for more context. However, we refer you to the [Zenodo page](https://doi.org/10.5281/zenodo.7554815) for an interactive overview of the source data.

# 2. IDA plan

This document exemplifies the pre-specified plan for initial data analysis (IDA plan) for the bacteremia study.

## 2.1 Prerequisites for the IDA plan

### 2.1.1 Analysis strategy

We assume that the aims of the study are to fit a diagnostic prediction model and to describe the functional form of each predictor. These aims are addressed by fitting a logistic regression model with bacteremia status as the dependent variable.

Based on domain expertise, the predictors are grouped by their assumed importance to predict bacteremia. Variables with known strong associations with bacteremia are age (AGE), leukocytes (WBC), blood urea neutrogen (BUN), creatinine (CREA), thrombocytes (PLT), and neutrophiles (NEU) and these predictors will be included in the model as key predictors. Predictors of medium importance are potassium (POTASS), and some acute-phase related parameters such as fibrinogen (FIB), C-reactive protein (CRP), aspartate transaminase (ASAT), alanine transaminase (ALAT), and gamma-glutamyl transpeptidase (GGT). All other predictors are of minor importance.

Continuous predictors should be modelled by allowing for flexible functional forms, where for all key predictors four degrees of freedom will be spent, and for predictors of medium and minor importance, three or two degrees of freedom should be foreseen at maximum, respectively. The decision on whether to use only key predictors, or to consider predictors also from the predictor sets of medium or minor importance depends on results of data screening, but will be made before uncovering the association of predictors with the outcome variable.

An adequate strategy to cope with missing values will also be chosen after screening the data. Candidate strategies are omission of predictors with abundant missing values, complete case analysis, single value imputation or multiple imputation with chained equations.

### 2.1.2 Analysis Ready Data dictionary

The **source** data dictionary of the bacteremia data set consists of columns for variable names, variable labels, scale of measurement (continuous or categorical), units, plausibility limits, and remarks. See [Appendix A](#sec-source-data).

An additional [**analysis ready** data dictionary](https://docs.google.com/spreadsheets/d/1Ft5eyenvDnMBoLvJmcBaklfrYcwyW-rkt-ivIkaphdA/edit#gid=1128598743) will be created to capture additional derivation, transformations and metadata relevant to the research question.

### 2.1.3 Domain expertise

The demographic variables age and sex are are chosen as the structural variables in this analysis for illustration purposes, since they are commonly considered important for describing a cohort in health studies. Key predictors and predictors of medium importance are as defined above. Laboratory analyses always bear the risk of machine failures, and hence missing values are a frequent challenge. This may differ between laboratory variables, but no a priori estimate about the expected proportion of missing values can be assumed. As most predictors measure concentrations of chemical compounds or cell counts, skewed distributions are expected. Some predictors describe related types of cells or chemical compounds, and hence some correlation between them is to be expected. For example, leukocytes consist of five different types of blood cells (BASO, EOS, NEU, LYM and MONO), and the sum of the concentration of these types approximately (but not exactly) gives the leukocyte count, which is recorded in the variable WBC. Moreover, these variables are given as absolute counts and as percentages of the sum of the five variables, which creates some correlation. Some laboratory variables differ by sex and age, but the special selection of patients for this study (suspicion of bacteremia) may distort or alter the expected correlations with sex and age.

## 2.2 IDA data derivations

Based on the prerequisites the following data derivations will be performed on the **source data** to form an **analysis ready data set** stored in the **data** folder.

The outcome variable from the source data **BC** will be renamed to **BACTEREMIA** to be more informative. A numeric variable **BACTEREMIAN** will also be derived to enable modelling with the following coding, 0: no, 1: yes.

For the purpose of stratifying IDA results by age, **AGE** will be categorized into the following three groups: [16, 50], (50, 65], (65, 101], both with numeric (**AGEGR01**) and character coding (**AGEGR01C**).

Predictors will be grouped by importance to the research question. Indicator flags for the predictor grouping will be derived:

* **KEY\_PRED\_FL01** will indicate the key predictors: AGE, WBC, BUN, CREA, PLT, and NEU
* **MED\_PRED\_FL01** will indicate the medium importance predictors: POTASS, FIB, CRP, ASAT, ALAT, GGT
* **REM\_PRED\_FL01** will indicator all remaining predictors not defined as key or medium importance.

An additional metadata flag **PARCAT01** will de derived to indicate the blood cell variables which form the Leukocytes predictor (**WBC**): BASO, EOS, NEU, LYM and MONO.

## 2.3 IDA planned analyses

The following sections detail the IDA planned analyses.

### 2.3.1 M1: Prevalence of missing values

Numbers and proportions of missing values will be reported for each predictor separately (M1). Type of missingness has not been recorded.

### 2.3.2 M2: Complete cases

The number of available complete cases (outcome and predictors) will be reported when considering:

1. the outcome variable (BACTEREMIA)
2. outcome and structural variables (BACTEREMIA, AGE, SEX)
3. outcome and key predictors only (BACTEREMIA, AGE, WBC, BUN, CREA, PLT, NEU)
4. outcome, key predictors and predictors of medium importance (BACTEREMIA, AGE, WBC, BUN, CREA, PLT, NEU, POTASS, FIB, CRP, ASAT, ALAT, GGT)
5. outcome and all predictors.

### 2.3.3 M3: Patterns of missing values

Patterns of missing values will be investigated by:

1. computing a table of complete cases (for the three predictor sets described above) for strata defined by the structural variables age and sex,
2. constructing a dendrogram of missing values to explore which predictors tend to be missing together.

### 2.3.4 U1: Univariate descriptions: categorical variables

For sex and bacteremia status, the frequency and proportion of each category will be described numerically.

### 2.3.5 U2: Univariate descriptions: continuous variables

For all continuous predictors, combo plots consisting of high-resolution histograms, boxplots and dotplots will be created. Because of the expected skew distribution, combo plots will also be created for log-transformed predictors.

As numerical summaries, minimum and maximum values, main quantiles (5th, 10th, 25th, 50th, 75th, 90th, 95th), and the first four moments (mean, standard deviation, skewness, curtosis) will be reported. The number of distinct values and the five most frequent values will be given, as well as the concentration ratio (ratio of frequency of most frequent value and mean frequency of each unique value).

Graphical and parametric multivariate analyses of the predictor space such as cluster analyses or the computation of variance inflation factors are heavily influenced by the distribution of the predictors. In order to make this set of analyses more robust to highly influential points or areas of the predictor support, some predictors may need transformation (e.g. logarithmic). We will compute the correlation of the untransformed and log-transformed predictors with normal deviates. Since some predictors may have values at or close to 0, we will consider the pseudolog transformation (Johnson 1949) which provides a smooth transition from linear (close to 0) to logarithmic (further away from 0). The transformation has a parameter which we will optimize separately for each predictor in order to achieve an optimal approximation to a normal distribution monitored via the correlation of normal deviates with the transformed predictor. For those predictors for which the pseudolog-transformation increases correlation with normal deviates by at least 0.2 units of the correlation coefficient, the pseudolog-transformed predictor will be used in multivariate IDA instead of the respective original predictor. For those predictors, histograms and boxplots will be provided on both the original and the transformed scale.

### 2.3.6 V1: Multivariate descriptions: associations of predictors with structural variables

A scatterplot of each predictor with age, with different panels for males and females will be constructed. Associated Spearman correlation coefficients will be computed.

### 2.3.7 V2: Multivariate descriptions: correlation analyses

A matrix of Spearman correlation coefficients between all pairs of predictors will be computed and described numerically as well as by means of a heatmap.

### 2.3.8 VE1: Multivariate descriptions: comparing nonparametric and parametric predictor correlation

A matrix of Pearson correlation coefficients will be computed. Predictor pairs for which Spearman and Pearson correlation coefficients differ by more than 0.1 correlation units will be depicted in scatterplots.

### 2.3.9 VE2: Variable clustering

A variable clustering analysis will be performed to evaluate which predictors are closely associated. A dendrogram groups predictors by their correlation. Scatterplots of pairs of predictors with Spearman correlation coefficients greater than 0.8 will be created.

### 2.3.10 VE3: Redundancy

Variance inflation factors will be computed between the candidate predictors. This will be done for the three possible candidate models, and using all complete cases in the respective candidate predictor sets. Redundancy will further be explored by computing parametric additive models for each predictor in the three candidate models.

# 3. Analysis ready data for IDA

Based on the IDA plan, this section prepares the source data to be analysis ready: read, clean, tidy and transform. This section focuses on the steps prior to IDA (data screening) and the required additions to the **source data** in order to prepare the data. The aim is to produce an analysis ready data set for the research objective.

## 3.1 Analysis ready dataset

The aim of this section and the remaining chapters of the report are to document the steps taken towards transforming the *source* data set to an *analysis ready* data set. These are the steps prior to the IDA analysis plan being executed.

The steps taken in this section are guided by the [data set specification](https://docs.google.com/spreadsheets/d/1Ft5eyenvDnMBoLvJmcBaklfrYcwyW-rkt-ivIkaphdA/edit#gid=2082358047) for the analysis ready data set, which is based on review of the IDA and analysis strategy.

For example, additional meta-data, data derivations and indicator flags are added to the *source* data set.

To support IDA, it is important that we keep track of the changes to the source data including all new modifications, data derivations and transformations. Therefore, we store references to the source data in the data folder after adding additional meta-data for all variables.

The format of the analysis ready data set follows that of the analysis ready [CDISC data model](https://www.cdisc.org/standards).

### 3.1.1 Data set transformations

Important meta-data is added to the data set from the data dictionary. At the stage we could select the variables of interest to take in to the IDA phase by dropping variables we do not check in IDA.

First the source data set and corresponding data dictionary are loaded. The variable names are normalized.

A new variable is created called PARAMCD which stores the abbreviated variable name as a reference. This provides a link between the dictionary and source data.

The source data is then transform in to a *long* format to store the lab specific measurements called ADLB. A long format enables efficient data processing and also allows new transformed variables to be added during the course of IDA. Structural and demographic variables will be stored in a *wide* format in a separate data set named ADSL.

### 3.1.2 Add lab variable meta-data

The lab parameter variable information such as labels and units on to the transformed data are added directly from the data dictionary.

At this point, additional variables and metadata are derived per the analysis plan including:

* Units
* Variable type
* Categories for sex
* rename outcome to be more informative

### 3.1.3 Reorder variables

Select and re-order the variables as per the data set specification.

### 3.1.4 Add informative variable meta-data

Add variable metadata as label attributes.

### 3.1.5 Visual check outcome is correct

Visual check we have not introduced any errors with the outcome variable. First display marginal distribution from source data variable.

| BloodCulture | n |
| --- | --- |
| no | 675550 |
| yes | 59000 |

Second, display marginal distribution from transformed data variable.

| BACTEREMIA | n |
| --- | --- |
| no | 675550 |
| yes | 59000 |

### 3.1.6 Dervive indicator flags

The next step is to dervive indicator flags for predictors as per the IDA plan (see [Section 2.2](#sec-IDA_data_derivations)):

* age (AGE), leukocytes (WBC), blood urea neutrogen (BUN), creatinine (CREA), thrombocytes (PLT), and neutrophiles (NEU) and these predictors will be included in the model as key predictors
* Predictors of medium importance are potassium (POTASS), and some acute-phase related parameters such as fibrinogen (FIB), C-reactive protein (CRP), aspartate transaminase (ASAT), alanine transaminase (ALAT), and gamma-glutamyl transpeptidase (GGT).

Next step, add metadata flags to indicate relationship between blood cell parameters. See [Section 2.2](#sec-IDA_data_derivations).

### 3.1.7 Data dervivations

Now, derive age groups. For the purpose of stratifying IDA results by age, age will be categorized into the following three groups ([Section 2.2](#sec-IDA_data_derivations)):

* [16, 50],
* (50, 65],
* (65, 101].

### 3.1.8 Save analysis ready data for IDA

Save the analysis data sets in to two linked data sets following a structure similar to the CDISC ADaM data standard. Individual patient measurements are stored in a data set called ADSL. The lab specific data sets are stored in ADLB (a long format data set).

Saving ADSL into an intermediate location DATA/IDA/ADSL\_01.rds prior to IDA.

Saving ADLB into an intermediate location DATA/IDA/ADLB\_01.rds prior to IDA.

Note: At this stage of IDA, both ADSL and ADLB are *intermediate* files that will be used for further IDA. Findings in IDA may require updates to either data set.

# 4. Results of IDA: Missing values

THis seciton reports the IDA analyses for missing data. The Section headers (i.e. M1) correspond to the IDA analysis plan in [Chapter 2](#sec-IDA_plan).

## 4.1 M1: Prevalence of missing values

Number and percentage of missingness for each predictor, sorted by descending missingness proportion.

### 4.1.1 Outcome and Structural variables

Table 1: Summary of missing values for outcome and structural predictors

| Variable | Missing (count) | Missing (%) |
| --- | --- | --- |
| BACTEREMIA | 0 | 0.0 |
| AGE | 0 | 0.0 |
| SEX | 0 | 0.0 |

### 4.1.2 Lab parameters

Missingness patterns for key and medium importance predictors are reported here. The remaining predictors are reported in Appendix [Section B.1](#sec-M1-appendix) .

Table 1: Summary of missing values for key predictors

| Predictor | Description | Missing (count) | Missing (%) |
| --- | --- | --- | --- |
| NEU | Neutrophiles (G/L) | 728 | 5.0 |
| WBC | White blood count (G/L) | 462 | 3.1 |
| BUN | Blood urea nitrogen (mg/dl) | 172 | 1.2 |
| CREA | Creatinine (mg/dl) | 159 | 1.1 |
| PLT | Blood platelets (G/L) | 42 | 0.3 |
| AGE | Patient Age (years) | 0 | 0.0 |

Table 1: Summary of missing values for medium importance predictors

| Predictor | Description | Missing (count) | Missing (%) |
| --- | --- | --- | --- |
| FIB | Fibrinogen (mg/dl) | 2567 | 17.5 |
| POTASS | Potassium (mmol/L) | 2008 | 13.7 |
| GGT | Gamma-glutamyl transpeptidase (G/L) | 1262 | 8.6 |
| ASAT | Aspartate transaminase (U/L) | 1154 | 7.9 |
| ALAT | Alanin transaminase (U/L) | 987 | 6.7 |
| CRP | C-reactive protein (mg/dl) | 155 | 1.1 |

## 4.2 M2: Complete cases

Number of available complete cases (outcome and predictors):

Table 1: Summary of complete cases by sets of predictors

| Set | Complete (count) | Complete (%) |
| --- | --- | --- |
| Outcome | 14691 | 100.0 |
| Outcome and structural variables | 14691 | 100.0 |
| Outcome and key predictors only | 13793 | 93.9 |
| Outcome key predictors and predictors of medium importance | 9389 | 63.9 |
| Outcome and all predictors | 3979 | 27.1 |

## 4.3 ME1: Patterns of missing values

### 4.3.1 Complete cases by strata defined by structural variables

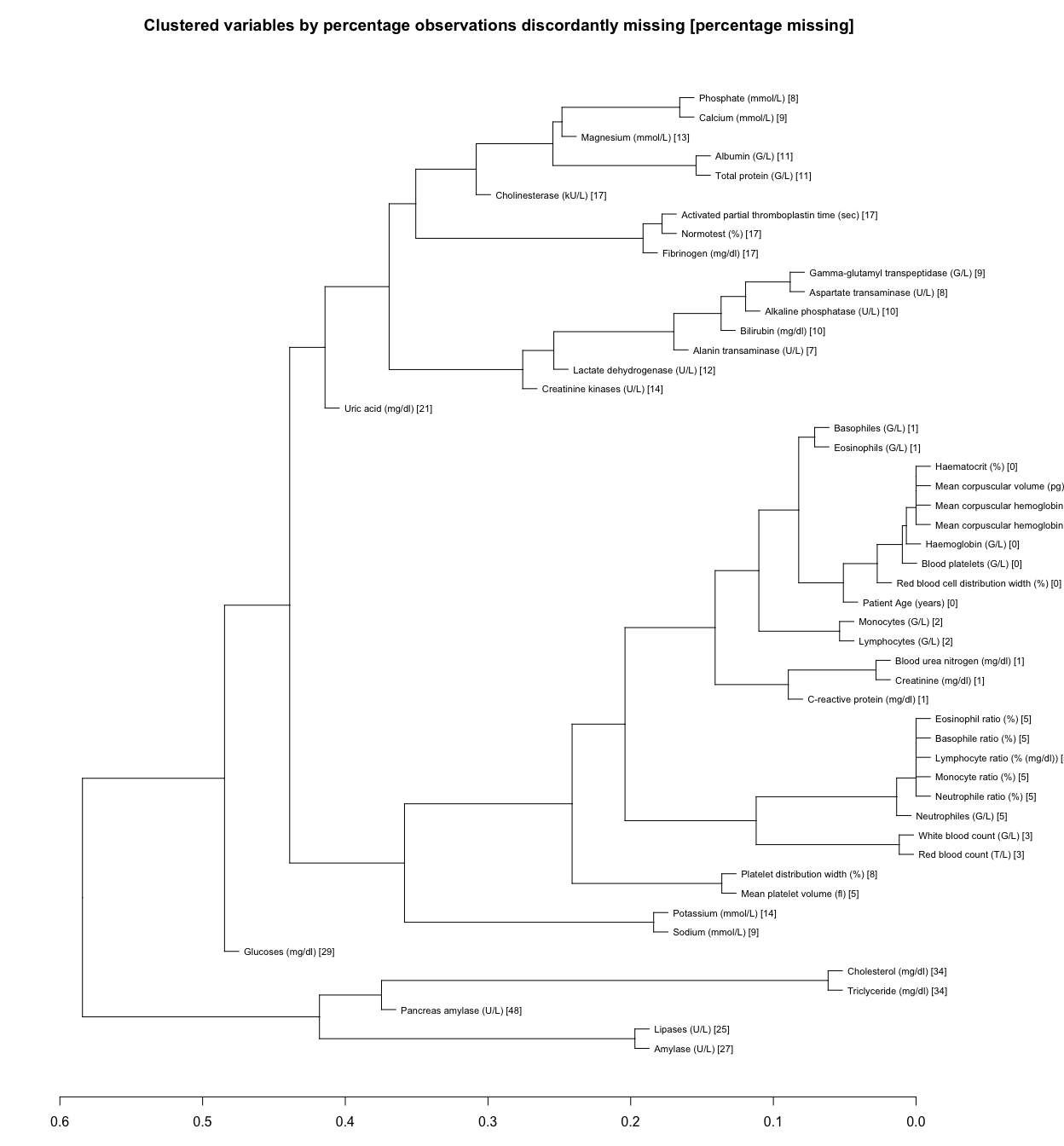
Table 1: Summary of complete cases by sets of predictors

| Set | Complete (count) | Complete (%) | AGEGR01 |
| --- | --- | --- | --- |
| female - [16, 50] |
| All predictors | 2309 | 93.8 | 1 |
| Key predictors | 2309 | 93.8 | 1 |
| Medium importance predictors | 1656 | 67.3 | 1 |
| female - (50, 65] |
| All predictors | 1468 | 93.0 | 2 |
| Key predictors | 1468 | 93.0 | 2 |
| Medium importance predictors | 1075 | 68.1 | 2 |
| female - (65, 101] |
| All predictors | 1975 | 93.4 | 3 |
| Key predictors | 1975 | 93.4 | 3 |
| Medium importance predictors | 1389 | 65.7 | 3 |
| male - [16, 50] |
| All predictors | 2744 | 94.5 | 1 |
| Key predictors | 2744 | 94.5 | 1 |
| Medium importance predictors | 1993 | 68.7 | 1 |
| male - (50, 65] |
| All predictors | 2504 | 93.7 | 2 |
| Key predictors | 2504 | 93.7 | 2 |
| Medium importance predictors | 1862 | 69.7 | 2 |
| male - (65, 101] |
| All predictors | 2793 | 94.3 | 3 |
| Key predictors | 2793 | 94.3 | 3 |
| Medium importance predictors | 2014 | 68.0 | 3 |

### 4.3.2 Dendrogram of missingness indicators

The dendrogram depicts the results of a cluster analysis using the complete linkage method based on the percentage of discordant missing indicators. (This percentage was computed via the squared Euclidian distance of missingness indicators between predictors.) The horizontal axis shows the distance between two clusters, which is given by the maximum distance between any element of the first and the second clusters. For example, if two clusters are merged at a height of 25 it means that in 25% of the observations the missingness indicators of the most discordant predictors contained in the two clusters are discordant.

The numbers in brackets are the percentages of missing observations for each predictor.



# 5. Univariate distribution checks

This section reports a series of univariate summary checks of the bacteremia dataset.

## 5.1 U1: Categorical variables

Age group, sex and bacteremia status are described by frequencies and proportions in each category.

| Category | Count | Proportion |
| --- | --- | --- |
| Age group |
| [16, 50] | 5365 | 0.37 |
| (65, 101] | 5076 | 0.35 |
| (50, 65] | 4250 | 0.29 |
| Sex |
| male | 8536 | 0.58 |
| female | 6155 | 0.42 |
| Presence of bacteremia |
| no | 13511 | 0.92 |
| yes | 1180 | 0.08 |

Also plot the categories.

|  |
| --- |
| Summary of categorical variables including outcome |

## 5.2 Continuous variables

### 5.2.1 U2: Univariate distributions of continuous variables

#### 5.2.1.1 U2: Structural variables

The only structural continuous variables is AGE. This variable is also a key predictor (see below).

#### 5.2.1.2 U2: Key predictors

Note: the structural variable Age is also considered a key predictor.

Distribution of key predictors. Lines indicate the 5-number summary including reported numerical values (where possible).

The remaining predictors are reported in the appendix [Section C.1](#sec-u2-remaining).

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

#### 5.2.1.3 U2: Predictors of medium importance

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

### 5.2.2 Numerical summaries

#### 5.2.2.1 Key predictors

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 718 | 1 | 220 | 130.1 | 50 | 81 | 140 | 204 | 277 | 369 | 445 |

lowest : 0 1 2 3 4 , highest: 1068 1211 1321 1639 2092

CREA: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14532 | 159 | 674 | 1 | 1.329 | 0.8518 | 0.620 | 0.690 | 0.810 | 1.000 | 1.350 | 2.160 | 3.144 |

lowest : 0.26 0.27 0.28 0.29 0.30 , highest: 15.24 15.40 15.67 16.64 20.75

BUN: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14519 | 172 | 947 | 1 | 22.66 | 16.92 | 7.1 | 8.6 | 11.6 | 16.6 | 26.9 | 44.8 | 60.8 |

lowest : 2.5 2.7 2.8 2.9 3.0 , highest: 160.6 171.3 171.9 176.0 184.8

NEU: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13963 | 728 | 374 | 1 | 8.367 | 5.776 | 1.60 | 2.70 | 4.60 | 7.30 | 10.80 | 15.08 | 18.40 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 54.0 56.4 63.7 71.6 83.8

WBC: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14229 | 462 | 2710 | 1 | 11.23 | 7.602 | 2.66 | 4.26 | 6.63 | 9.60 | 13.53 | 18.22 | 22.27 |

lowest : 0.00 0.01 0.02 0.03 0.04 , highest: 365.30 383.74 387.73 433.83 604.47

AGE: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14691 | 0 | 85 | 1 | 56.17 | 20.78 | 24 | 29 | 43 | 58 | 70 | 79 | 84 |

#### 5.2.2.2 Predictors of medium importance

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12124 | 2567 | 1084 | 1 | 547.4 | 231 | 247 | 301 | 397 | 529 | 674 | 816 | 892 |

lowest : 55 60 66 67 69 , highest: 1506 1508 1529 1537 1593

POTASS: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12683 | 2008 | 408 | 1 | 4.003 | 0.6004 | 3.20 | 3.39 | 3.66 | 3.95 | 4.29 | 4.67 | 4.92 |

lowest : 1.92 2.07 2.11 2.12 2.21 , highest: 8.57 11.34 13.55 14.60 36.62

Value 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5  
 Frequency 6 91 650 3385 5088 2455 692 201 73 26 7 2  
 Proportion 0.000 0.007 0.051 0.267 0.401 0.194 0.055 0.016 0.006 0.002 0.001 0.000  
   
 Value 8.0 8.5 11.5 13.5 14.5 36.5  
 Frequency 2 1 1 1 1 1  
 Proportion 0.000 0.000 0.000 0.000 0.000 0.000

For the frequency table, variable is rounded to the nearest 0.5

ASAT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13537 | 1154 | 650 | 1 | 86.9 | 115.6 | 15 | 17 | 22 | 31 | 56 | 121 | 218 |

lowest : 3 5 6 7 8 , highest: 10845 11928 12079 12380 13991

ALAT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13704 | 987 | 578 | 1 | 67.66 | 90.07 | 9 | 11 | 16 | 26 | 48 | 101 | 175 |

lowest : 0 1 2 3 4 , highest: 7109 9136 9314 12329 15059

GGT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13429 | 1262 | 858 | 1 | 115.1 | 141.3 | 13.0 | 16.0 | 25.0 | 49.0 | 117.0 | 262.2 | 429.0 |

lowest : 3 5 6 7 8 , highest: 2932 3303 3782 3919 5171

CRP: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14536 | 155 | 3328 | 1 | 10.92 | 10.39 | 0.29 | 0.77 | 2.87 | 8.57 | 16.45 | 24.49 | 29.61 |

### 5.2.3 Suggested transformations

Next we investigate whether a pseudolog transformation of continuous variables may substantially symmetrize the univariate distributions of the continuous variables, and may hence be useful for multivariate summaries. We employ a function ida\_trans for this purpose, which optimises the parameter sigma of the pseudo-logarithm for that purpose. The optimization targets the best possible linear correlation of the transformed values with normal deviates. If no better transformation can be found, or if the improvement in correlation is less than 0.2 correlation units, no transformation is suggested.

Display the proposed variable transformations and the new parameter codes.

| PARAMCD | n |
| --- | --- |
| ALAT\_T | 14691 |
| AMY\_T | 14691 |
| AP\_T | 14691 |
| ASAT\_T | 14691 |
| BASO\_T | 14691 |
| CK\_T | 14691 |
| CREA\_T | 14691 |
| EOS\_T | 14691 |
| GBIL\_T | 14691 |
| GGT\_T | 14691 |
| LDH\_T | 14691 |
| LIP\_T | 14691 |
| LYM\_T | 14691 |
| PAMY\_T | 14691 |
| WBC\_T | 14691 |

Register transformed variables in the data set. The updated data set with suggested log transformed data sets is saved at data/IDA/ADLB\_02.rds.

The IDA analysis plan and specifications are updated with the proposed variable transformations. A new flag is derived to indicate the categorization of predictors now including transformations.

### 5.2.4 Comparison of univariate distributions with and without pseudo-log transformation

The comparison is only shown for variables where a transformation is suggested. Note, all observed values, and the distribution min, max and interquartile range as reference lines, are displayed.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
|  |

# 6. Multivariate analyses

First load the required packages and data. Note:, from the univariate analyses the analysis data set for the lab parameters has been updated to include transformed variables. Therefore, we load the second iteration of the data data/ADLB\_02.rds.

## 6.1 V1: Association with structural variables

Attached the required structural variables to the lab data.

A scatterplot of each predictor with age, with different panels for males and females have been constructed. Associated Spearman correlation coefficients have been computed.

The remaining predicotrs are reported in appendix [Section E.1](#sec-V1-appendix).

### 6.1.1 Key predictors

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |
| --- | --- |
|  |  |

### 6.1.2 Predictors of medium importance

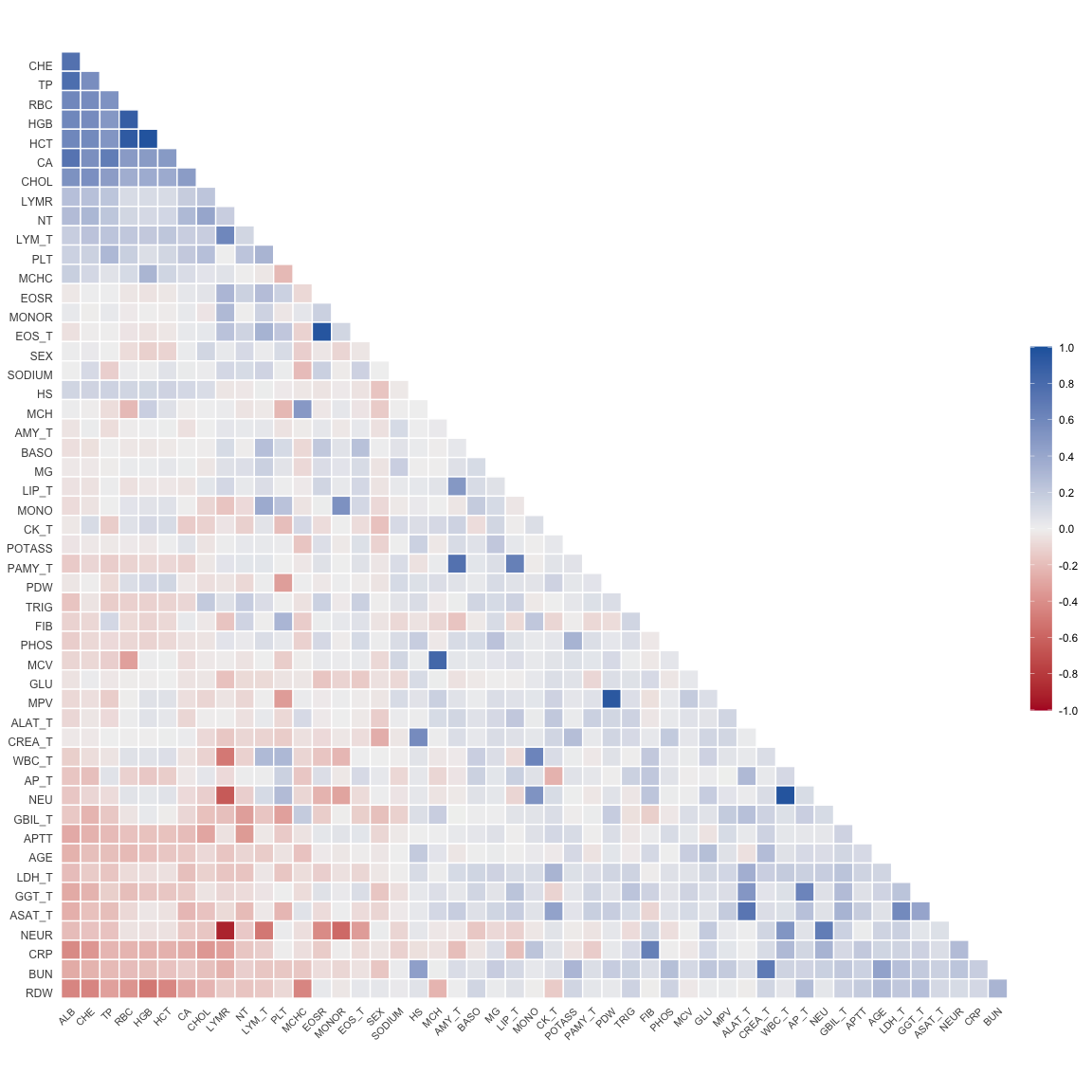
|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

## 6.2 V2: Correlation coefficients between all predictors

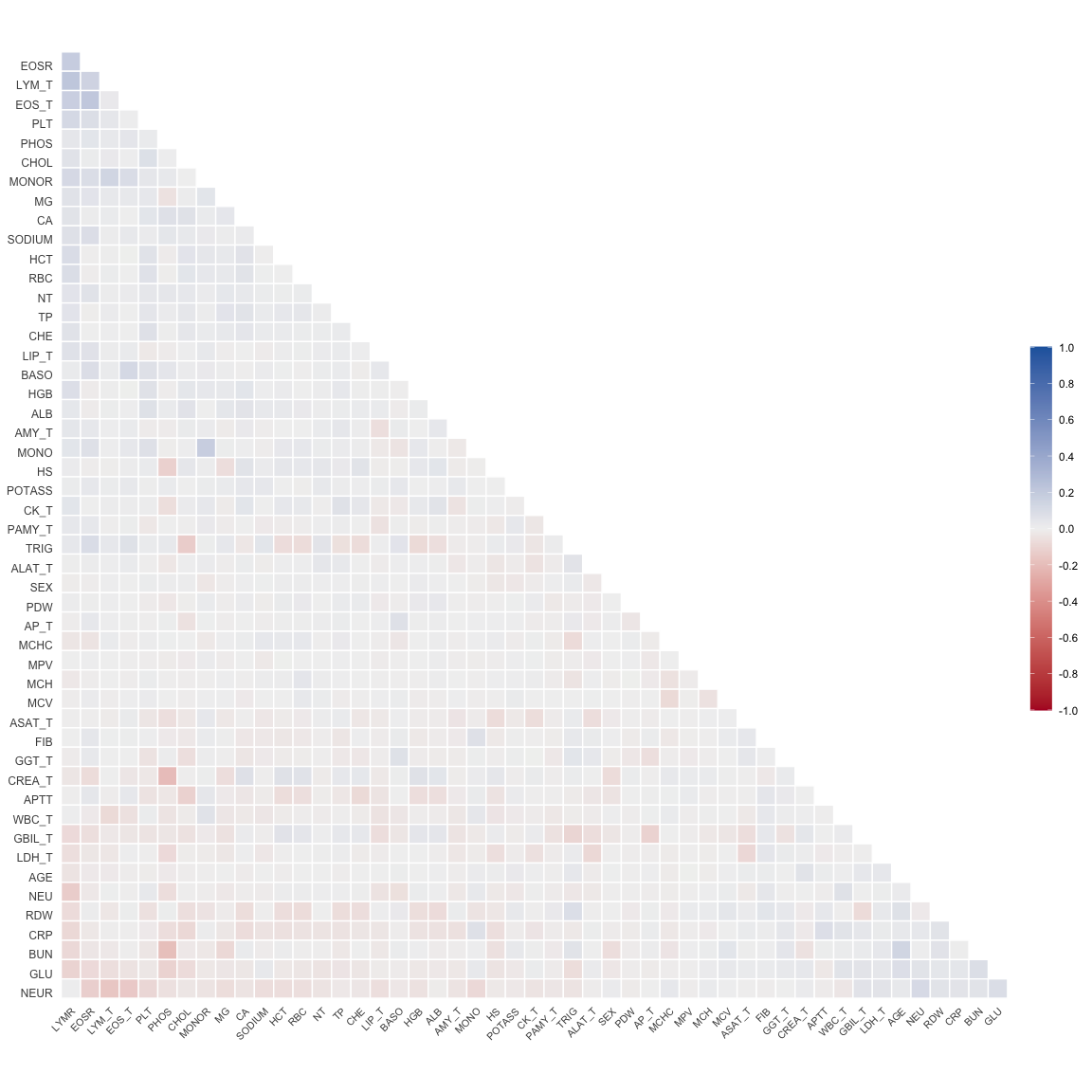
Calculate correlation matrix using Spearman correlation coefficient.

The Spearman correlation coefficients are depicted in a quadratic heat map:

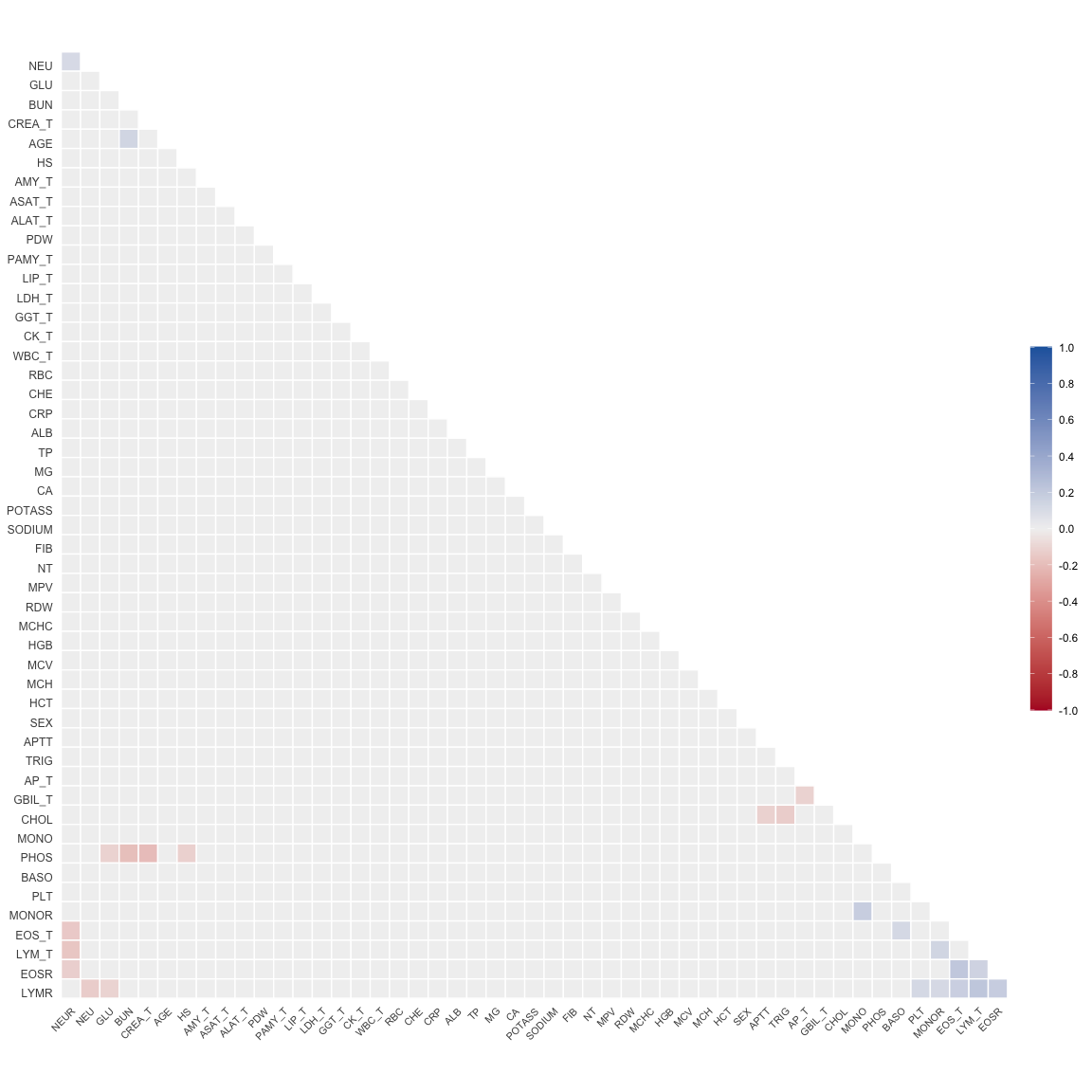


### 6.2.1 VE1: Comparing nonparametric and parametric predictor correlation

Plot the difference matrix



Plot the matrix but suppress differences less then 0.1.



Predictor pairs for which Spearman and Pearson correlation coefficients differ by more than 0.1 correlation units will be depicted in scatterplots. First report the table of predictor parirs where r is greater than 0.1.

| x | y | r |
| --- | --- | --- |
| PLT | LYMR | 0.12 |
| MONO | MONOR | 0.18 |
| BASO | EOS\_T | 0.12 |
| BUN | AGE | 0.13 |
| EOSR | LYMR | 0.18 |
| EOSR | EOS\_T | 0.21 |
| EOSR | LYM\_T | 0.14 |
| LYMR | MONOR | 0.11 |
| LYMR | EOS\_T | 0.17 |
| LYMR | LYM\_T | 0.22 |
| MONOR | LYM\_T | 0.13 |
| NEU | NEUR | 0.11 |

Then plot the scatter plots.

|  |  |  |
| --- | --- | --- |
|  |  |  |

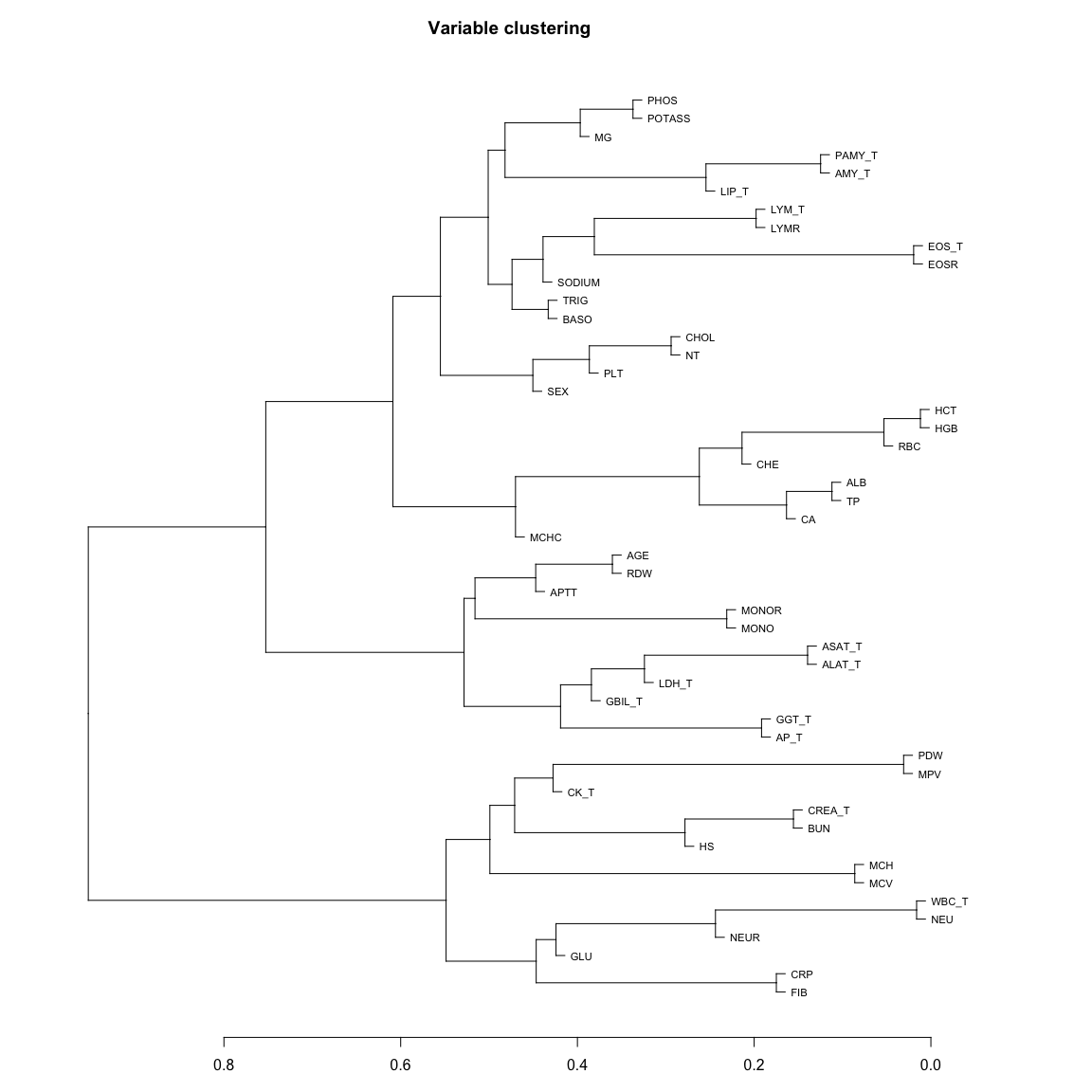
|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

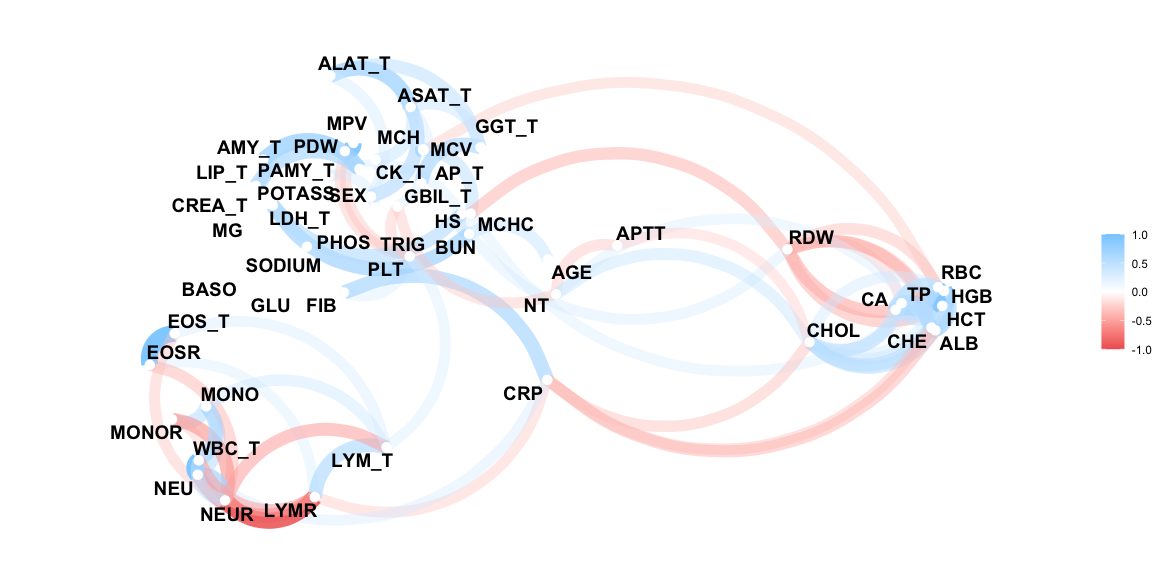
|  |  |  |
| --- | --- | --- |
|  |  |  |

### 6.2.2 VE2: Variable clustering

A variable clustering analysis has been performed to evaluate which predictors are closely associated. The dendrogram groups predictors by their correlation.



This could also be displayed as a network plot



In the following scatterplots we show predictor pairs with Spearman correlation coefficients greater than 0.8:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

### 6.2.3 VE3: Redundancy

Variance inflation factors (VIF) will be computed between the candidate predictors. This will be done for the three possible candidate models, and using all complete cases in the respective candidate predictor sets. Since , we also report the multiple R-squared values.

Redundancy was further explored by computing parametric additive models for each predictor in the key predictor model and the extended predictor model. VIFs and multiple are reported from those models, again for the three predictor sets.

Note, the all predictor model is reported in appendix [Section E.2](#sec-VE3-appendix).

#### 6.2.3.1 VIF for key predictor model

The available sample size is 13793 (1.44 %).

| Parameter code | Variance inflation factor | Multiple R-squared |
| --- | --- | --- |
| SEX | 1.1 | 0.05 |
| PLT | 1.2 | 0.15 |
| BUN | 2.4 | 0.58 |
| NEU | 5.4 | 0.82 |
| AGE | 1.1 | 0.09 |
| CREA\_T | 2.3 | 0.56 |
| WBC\_T | 5.6 | 0.82 |

#### 6.2.3.2 VIF for model with key predictors and predictors of medium importance

The available sample size is 9389 (0.98 %).

| Parameter code | Variance inflation factor | Multiple R-squared |
| --- | --- | --- |
| SEX | 1.1 | 0.09 |
| PLT | 1.4 | 0.30 |
| FIB | 2.3 | 0.56 |
| POTASS | 1.1 | 0.10 |
| BUN | 2.6 | 0.61 |
| CRP | 2.2 | 0.54 |
| NEU | 5.9 | 0.83 |
| AGE | 1.2 | 0.15 |
| ALAT\_T | 3.3 | 0.69 |
| ASAT\_T | 3.2 | 0.69 |
| CREA\_T | 2.4 | 0.58 |
| GGT\_T | 1.4 | 0.30 |
| WBC\_T | 6.0 | 0.83 |

#### 6.2.3.3 VIF for all predictor model

See appendix [Section E.2](#sec-VE3-appendix).

#### 6.2.3.4 Redundancy by parametric additive model: key predictor model

The available sample size is 13793 (1.44 %).

| Parameter code | Variance inflation factor | Multiple R-squared |
| --- | --- | --- |
| SEX | 1.1 | 0.11 |
| PLT | 1.2 | 0.16 |
| BUN | 2.6 | 0.61 |
| NEU | 11.5 | 0.91 |
| AGE | 1.3 | 0.21 |
| CREA\_T | 2.4 | 0.58 |
| WBC\_T | 13.7 | 0.93 |

#### 6.2.3.5 Redundancy by parametric additive model: key predictors and predictors of medium importance

The available sample size is 9389 (0.98 %).

| Parameter code | Variance inflation factor | Multiple R-squared |
| --- | --- | --- |
| SEX | 1.2 | 0.15 |
| PLT | 1.5 | 0.32 |
| FIB | 2.4 | 0.58 |
| POTASS | 1.1 | 0.11 |
| BUN | 2.8 | 0.65 |
| CRP | 2.3 | 0.56 |
| NEU | 14.1 | 0.93 |
| AGE | 1.4 | 0.29 |
| ALAT\_T | 3.4 | 0.70 |
| ASAT\_T | 3.5 | 0.71 |
| CREA\_T | 2.5 | 0.60 |
| GGT\_T | 1.5 | 0.35 |
| WBC\_T | 16.4 | 0.94 |

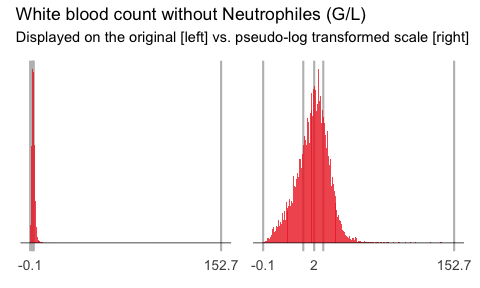
# 7. Final Analysis ready data

## 7.1 Final dervivations and transformations

From the IDA exercise a decision was made to derive a new variable based on White Blood Count in the absence of Neutrophiles:

* WBC\_NEU: WBC - NEU

The new derivation will also be checked as a candidate for pseudo-log transformation.



The new derivation and transformation is added to the analysis ready data, setting the correct indicator flags to enable final analyses and modelling.

Visual check of indicator flag and new derviation added to the analysis data set.

[1] "c(\"PLT\", \"CREA\", \"BUN\", \"NEU\", \"AGE\", \"WBC\_noNEU\")"

## 7.2 Save final analysis ready data sets

The final analysis ready data sets are then stored in data/.

# References

Johnson, N. L. 1949. “Systems of Frequency Curves Generated by Methods of Translation.” *Biometrika* 36 (1/2): 149–76. <http://www.jstor.org/stable/2332539>.

Ratzinger, Franz, Michel Dedeyan, Matthias Rammerstorfer, Thomas Perkmann, Heinz Burgmann, Athanasios Makristathis, Georg Dorffner, Felix Lötsch, Alexander Blacky, and Michael Ramharter. 2014. “A Risk Prediction Model for Screening Bacteremic Patients: A Cross Sectional Study.” *PLOS ONE* 9 (9): 1–10. <https://doi.org/10.1371/journal.pone.0106765>.

# Appendix A — Source data

The **source** data available in this repository (**DC 2019-0054**) can be located in the data-raw folder. For persistance, the data set is also published on [Zenodo](https://doi.org/10.5281/zenodo.7554815) with the following doi: https://doi.org/10.5281/zenodo.7554815.

The following sections provide a short overview of the data dictionary which accompanies the source data, and a short description of the data.

## A.1 Data dictionary

The data dictionary provides an overview of the collected data. First, we read and display the data dictionary below providing an overview of the collected measurements.

The variable name and label are displayed alongside the measurement scale and units as well as remarks and relevant study information from\_paper.

| variable | label | scale\_of\_measurement | units |
| --- | --- | --- | --- |
| ID | Patient Identification | nominal | 1-14691 |
| SEX | Patient sex | nominal | 1=male, 2=female |
| AGE | Patient Age | continuous | years |
| MCV | Mean corpuscular volume | continuous | pg |
| HGB | Haemoglobin | continuous | G/L |
| HCT | Haematocrit | continuous | % |
| PLT | Blood platelets | continuous | G/L |
| MCH | Mean corpuscular hemoglobin | continuous | fl |
| MCHC | Mean corpuscular hemoglobin concentration | continuous | g/dl |
| RDW | Red blood cell distribution width | continuous | % |
| MPV | Mean platelet volume | continuous | fl |
| LYM | Lymphocytes | continuous | G/L |
| MONO | Monocytes | continuous | G/L |
| EOS | Eosinophils | continuous | G/L |
| BASO | Basophiles | continuous | G/L |
| NT | Normotest | continuous | % |
| APTT | Activated partial thromboplastin time | continuous | sec |
| FIB | Fibrinogen | continuous | mg/dl |
| SODIUM | Sodium | continuous | mmol/L |
| POTASS | Potassium | continuous | mmol/L |
| CA | Calcium | continuous | mmol/L |
| PHOS | Phosphate | continuous | mmol/L |
| MG | Magnesium | continuous | mmol/L |
| CREA | Creatinine | continuous | mg/dl |
| BUN | Blood urea nitrogen | continuous | mg/dl |
| HS | Uric acid | continuous | mg/dl |
| GBIL | Bilirubin | continuous | mg/dl |
| TP | Total protein | continuous | G/L |
| ALB | Albumin | continuous | G/L |
| AMY | Amylase | continuous | U/L |
| PAMY | Pancreas amylase | continuous | U/L |
| LIP | Lipases | continuous | U/L |
| CHE | Cholinesterase | continuous | kU/L |
| AP | Alkaline phosphatase | continuous | U/L |
| ASAT | Aspartate transaminase | continuous | U/L |
| ALAT | Alanin transaminase | continuous | U/L |
| GGT | Gamma-glutamyl transpeptidase | continuous | G/L |
| LDH | Lactate dehydrogenase | continuous | U/L |
| CK | Creatinine kinases | continuous | U/L |
| GLU | Glucoses | continuous | mg/dl |
| TRIG | Triclyceride | continuous | mg/dl |
| CHOL | Cholesterol | continuous | mg/dl |
| CRP | C-reactive protein | continuous | mg/dl |
| BASOR | Basophile ratio | continuous | % |
| EOSR | Eosinophil ratio | continuous | % |
| LYMR | Lymphocyte ratio | continuous | % (mg/dl) |
| MONOR | Monocyte ratio | continuous | % |
| NEU | Neutrophiles | continuous | G/L |
| NEUR | Neutrophile ratio | continuous | % |
| PDW | Platelet distribution width | continuous | % |
| RBC | Red blood count | continuous | T/L |
| WBC | White blood count | continuous | G/L |
| BloodCulture | Blood culture result for bacteremia | nominal | no, yes |

## A.2 Source data

We also display a short snapshot of source data set from the data-raw folder of the project directory. The snapshot provides a glimpse of the data, giving the data dictionary more context.

We do not display all observations measured as it is too wide and long to fit reasonably in to the report. However, we refer you to the [Zenodo page](https://doi.org/10.5281/zenodo.7554815) for an interactive overview of the source data.

Rows: 14,691  
Columns: 53  
$ ID <dbl> 1, 3, 5, 7, 9, 10, 11, 12, 13, 19, 21, 22, 23, 25, 26, 27…  
$ SEX <dbl> 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, …  
$ AGE <dbl> 62, 72, 46, 84, 38, 68, 55, 55, 67, 52, 47, 29, 59, 51, 4…  
$ MCV <dbl> 99.3, 85.1, 96.3, 91.3, 85.1, 104.5, 99.3, 77.0, 95.3, 83…  
$ HGB <dbl> 11.5, 10.3, 7.4, 10.3, 13.7, 15.7, 14.6, 10.8, 10.9, 10.3…  
$ HCT <dbl> 35.9, 34.7, 22.8, 31.1, 38.7, 46.9, 43.5, 34.8, 30.4, 30.…  
$ PLT <dbl> 307, 182, 64, 309, 183, 144, 242, 38, 88, 105, 216, 188, …  
$ MCH <dbl> 31.5, 26.0, 31.2, 30.4, 30.2, 34.8, 33.1, 23.8, 33.6, 28.…  
$ MCHC <dbl> 31.8, 30.6, 32.4, 33.3, 35.3, 33.5, 33.4, 30.5, 35.3, 34.…  
$ RDW <dbl> 19.5, 15.0, 19.7, 13.8, 12.6, 13.9, 13.1, 16.8, 13.3, 13.…  
$ MPV <dbl> 10.8, 9.7, 11.1, 8.5, 10.0, 10.9, 10.3, NA, 10.7, 11.3, 1…  
$ LYM <dbl> 0.4, 0.4, 1.5, 1.3, 0.8, 2.2, 2.1, 0.4, 0.4, 0.9, 0.7, 1.…  
$ MONO <dbl> 1.7, 0.2, 1.2, 0.8, 0.4, 0.9, 1.6, 0.1, 0.2, 0.9, 0.6, 0.…  
$ EOS <dbl> 0.0, 0.1, 0.1, 0.0, 0.0, 0.1, 0.3, 0.1, 0.0, 0.3, 0.0, 0.…  
$ BASO <dbl> 0.1, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.1, 0.…  
$ NT <dbl> 86, 90, 58, 67, 95, 61, NA, 93, 57, 69, 108, 86, 93, 83, …  
$ APTT <dbl> 28.8, 29.8, 36.3, 38.2, 33.1, 41.8, NA, 36.3, 33.8, 28.1,…  
$ FIB <dbl> 578, NA, 313, 487, 490, 400, NA, 413, 431, 407, 604, 476,…  
$ SODIUM <dbl> 137, 141, 147, 141, 137, 141, 139, 142, 143, 136, 131, 13…  
$ POTASS <dbl> 3.88, NA, 4.61, 4.71, NA, 4.41, 3.69, 4.67, 2.35, 3.80, 5…  
$ CA <dbl> 2.29, 2.21, 1.92, 2.05, 2.34, 2.08, NA, 2.31, 2.10, 1.92,…  
$ PHOS <dbl> 1.20, 0.58, 1.51, 2.17, 0.97, 0.99, NA, 1.16, 0.51, 0.72,…  
$ MG <dbl> 0.66, NA, 1.03, 0.83, 0.74, 0.56, NA, 0.87, 0.36, 0.53, 0…  
$ CREA <dbl> 0.65, 0.76, 1.25, 2.78, 0.65, 0.82, 1.21, 1.77, 1.00, 0.5…  
$ BUN <dbl> 5.7, 19.9, 50.6, 47.5, 8.5, 15.3, 13.0, 29.8, 15.0, 14.0,…  
$ HS <dbl> 5.3, NA, NA, 9.7, 3.0, 5.5, NA, 6.2, 4.7, 4.0, 4.0, 4.1, …  
$ GBIL <dbl> 0.59, 0.48, 8.42, 0.35, 0.42, 2.40, 1.13, 0.45, 1.21, 2.4…  
$ TP <dbl> 67.0, 65.3, 40.5, 61.2, 78.4, 57.5, NA, 70.8, 67.4, 53.8,…  
$ ALB <dbl> 36.7, 37.4, 22.1, 33.2, 43.8, 30.1, NA, 43.6, 35.4, 24.8,…  
$ AMY <dbl> 30, NA, 146, 92, 84, 95, 117, 177, NA, 35, 79, 16, 25, 32…  
$ PAMY <dbl> 16, NA, NA, 28, 50, 57, NA, 43, NA, 35, 63, 14, 15, 20, 3…  
$ LIP <dbl> 10, NA, 89, 18, 50, 25, 73, 30, NA, 38, 52, 19, 14, 26, 5…  
$ CHE <dbl> 5.12, 5.61, 2.52, 4.10, 6.91, 6.79, NA, 7.40, NA, 2.64, 2…  
$ AP <dbl> 85, 80, 119, 94, 108, 68, 51, 153, 239, 146, 180, 64, 74,…  
$ ASAT <dbl> 22, 28, 124, 774, 35, 32, 29, 26, 91, 97, 24, 13, 25, 31,…  
$ ALAT <dbl> 14, 25, 135, 72, 22, 11, 20, 32, 57, 156, 63, 23, 27, 53,…  
$ GGT <dbl> 48, 61, 134, 23, 72, 68, 138, 96, 446, 192, 266, 19, 66, …  
$ LDH <dbl> 284, NA, 696, 1787, NA, 263, 303, 181, 183, 277, 221, 299…  
$ CK <dbl> 23, 36, 40, 2422, 79, 75, 230, 87, 53, 87, 30, 118, 17, 1…  
$ GLU <dbl> 107, 84, 107, 105, 93, 89, 91, 96, 86, 104, 104, 102, 161…  
$ TRIG <dbl> 105, NA, NA, 134, 152, 85, NA, 129, 62, 207, 292, 221, 12…  
$ CHOL <dbl> 175, NA, NA, 141, 167, 144, NA, 156, 118, 123, 194, 151, …  
$ CRP <dbl> 3.94, 1.42, 12.09, 3.78, 11.17, 5.89, 17.84, 1.29, 1.36, …  
$ BASOR <dbl> 0.4132231, 0.0000000, 0.5681818, 0.0000000, 0.0000000, 0.…  
$ EOSR <dbl> 0.0000000, 0.8264463, 0.5681818, 0.0000000, 0.0000000, 1.…  
$ LYMR <dbl> 1.652893, 3.305785, 8.522727, 11.016949, 8.333333, 22.000…  
$ MONOR <dbl> 7.024793, 1.652893, 6.818182, 6.779661, 4.166667, 9.00000…  
$ NEU <dbl> 22.0, 11.4, 14.7, 9.7, 8.4, 6.8, 8.9, 1.2, NA, 3.8, 8.2, …  
$ NEUR <dbl> 90.90909, 94.21488, 83.52273, 82.20339, 87.50000, 68.0000…  
$ PDW <dbl> 10.6, 11.4, 14.1, 8.7, 12.2, 12.9, 12.5, NA, NA, 13.2, 12…  
$ RBC <dbl> 3.7, 3.9, 2.5, 3.5, 4.4, 4.3, 4.5, 4.7, NA, 3.5, 3.3, 2.5…  
$ WBC <dbl> 24.10, 12.17, 17.45, 11.58, 9.86, 9.94, 13.06, 1.78, NA, …  
$ BloodCulture <chr> "no", "no", "no", "no", "no", "no", "no", "no", "yes", "n…

# Appendix B — Results of IDA: Missing values Appendix

## B.1 M1: Prevalence of missing values for remaining predictors

Table 1: Summary of missing values for remaining predictors

| Predictor | Description | Missing (count) | Missing (%) |
| --- | --- | --- | --- |
| PAMY | Pancreas amylase (U/L) | 7114 | 48.4 |
| TRIG | Triclyceride (mg/dl) | 5061 | 34.4 |
| CHOL | Cholesterol (mg/dl) | 5045 | 34.3 |
| GLU | Glucoses (mg/dl) | 4192 | 28.5 |
| AMY | Amylase (U/L) | 3913 | 26.6 |
| LIP | Lipases (U/L) | 3699 | 25.2 |
| HS | Uric acid (mg/dl) | 3061 | 20.8 |
| APTT | Activated partial thromboplastin time (sec) | 2549 | 17.4 |
| NT | Normotest (%) | 2467 | 16.8 |
| CHE | Cholinesterase (kU/L) | 2447 | 16.7 |
| CK | Creatinine kinases (U/L) | 2080 | 14.2 |
| MG | Magnesium (mmol/L) | 1869 | 12.7 |
| LDH | Lactate dehydrogenase (U/L) | 1714 | 11.7 |
| ALB | Albumin (G/L) | 1676 | 11.4 |
| TP | Total protein (G/L) | 1583 | 10.8 |
| GBIL | Bilirubin (mg/dl) | 1441 | 9.8 |
| AP | Alkaline phosphatase (U/L) | 1400 | 9.5 |
| SODIUM | Sodium (mmol/L) | 1282 | 8.7 |
| CA | Calcium (mmol/L) | 1276 | 8.7 |
| PHOS | Phosphate (mmol/L) | 1242 | 8.5 |
| PDW | Platelet distribution width (%) | 1102 | 7.5 |
| BASOR | Basophile ratio (%) | 732 | 5.0 |
| EOSR | Eosinophil ratio (%) | 732 | 5.0 |
| LYMR | Lymphocyte ratio (% (mg/dl)) | 732 | 5.0 |
| MONOR | Monocyte ratio (%) | 732 | 5.0 |
| NEUR | Neutrophile ratio (%) | 732 | 5.0 |
| MPV | Mean platelet volume (fl) | 702 | 4.8 |
| RBC | Red blood count (T/L) | 461 | 3.1 |
| LYM | Lymphocytes (G/L) | 262 | 1.8 |
| MONO | Monocytes (G/L) | 246 | 1.7 |
| BASO | Basophiles (G/L) | 146 | 1.0 |
| EOS | Eosinophils (G/L) | 135 | 0.9 |
| RDW | Red blood cell distribution width (%) | 56 | 0.4 |
| MCV | Mean corpuscular volume (pg) | 42 | 0.3 |
| HCT | Haematocrit (%) | 42 | 0.3 |
| MCH | Mean corpuscular hemoglobin (fl) | 42 | 0.3 |
| MCHC | Mean corpuscular hemoglobin concentration (g/dl) | 42 | 0.3 |
| HGB | Haemoglobin (G/L) | 41 | 0.3 |

# Appendix C — Univariate distribution checks

This section reports a series of univariate summary checks of the bacteremia dataset.

## C.1 U2: Remaining predictors

### C.1.1 Visual summary

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |
| --- | --- |
|  |  |

### C.1.2 Descriprive summary

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 506 | 1 | 88.35 | 6.992 | 78.2 | 81.1 | 84.7 | 88.3 | 92.0 | 95.9 | 99.0 |

lowest : 51.0 52.6 54.9 56.3 57.5 , highest: 121.0 121.8 124.6 127.9 128.7

HGB: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14650 | 41 | 157 | 1 | 11.57 | 2.558 | 8.2 | 8.8 | 9.9 | 11.4 | 13.2 | 14.6 | 15.4 |

lowest : 3.0 3.1 3.5 3.9 4.1 , highest: 19.5 20.5 20.7 20.8 21.0

HCT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 404 | 1 | 34.48 | 7.316 | 24.6 | 26.4 | 29.8 | 34.3 | 39.1 | 42.9 | 44.8 |

lowest : 0.0 0.1 0.2 9.7 9.8 , highest: 61.4 61.9 63.2 65.3 66.6

MCH: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 232 | 1 | 29.58 | 2.693 | 25.3 | 26.7 | 28.4 | 29.7 | 31.0 | 32.4 | 33.4 |

lowest : 14.9 15.6 15.9 16.0 16.5 , highest: 42.0 42.3 42.4 42.5 47.4

MCHC: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 124 | 0.999 | 33.47 | 1.546 | 31.1 | 31.7 | 32.6 | 33.5 | 34.4 | 35.2 | 35.6 |

lowest : 23.7 24.4 24.8 25.1 26.1 , highest: 38.3 38.4 38.9 39.3 43.5

RDW: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14635 | 56 | 173 | 1 | 15 | 2.385 | 12.4 | 12.7 | 13.4 | 14.5 | 16.0 | 18.0 | 19.5 |

lowest : 10.6 11.1 11.2 11.3 11.4 , highest: 28.6 28.9 29.1 29.7 31.8

MPV: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13989 | 702 | 71 | 0.999 | 10.38 | 1.132 | 8.9 | 9.2 | 9.7 | 10.3 | 11.0 | 11.7 | 12.2 |

lowest : 7.3 7.7 7.8 7.9 8.0 , highest: 14.2 14.3 14.6 14.8 15.0

LYM: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14429 | 262 | 114 | 0.998 | 1.366 | 1.162 | 0.2 | 0.4 | 0.7 | 1.0 | 1.6 | 2.1 | 2.6 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 149.9 357.5 366.8 375.1 578.1

Value 0 5 10 15 20 30 35 40 45 75 100 115  
 Frequency 13675 703 27 7 4 2 1 1 1 1 1 1  
 Proportion 0.948 0.049 0.002 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000  
   
 Value 150 360 365 375 580  
 Frequency 1 1 1 1 1  
 Proportion 0.000 0.000 0.000 0.000 0.000

For the frequency table, variable is rounded to the nearest 5

MONO: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14445 | 246 | 67 | 0.996 | 0.8527 | 0.5965 | 0.1 | 0.3 | 0.5 | 0.8 | 1.1 | 1.5 | 1.8 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 13.9 14.6 16.2 17.3 20.4

EOS: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14556 | 135 | 36 | 0.867 | 0.1148 | 0.1585 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.3 | 0.4 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 3.8 5.3 9.6 11.5 15.8

BASO: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14545 | 146 | 18 | 0.337 | 0.01725 | 0.03111 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 1.3 1.4 1.5 2.2 6.5

Value 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1  
 Frequency 12671 1636 109 59 31 14 6 7 1 2 1 2  
 Proportion 0.871 0.112 0.007 0.004 0.002 0.001 0.000 0.000 0.000 0.000 0.000 0.000  
   
 Value 1.2 1.3 1.4 1.5 2.2 6.5  
 Frequency 1 1 1 1 1 1  
 Proportion 0.000 0.000 0.000 0.000 0.000 0.000

NT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12224 | 2467 | 149 | 1 | 83.22 | 30.56 | 35 | 48 | 67 | 83 | 101 | 118 | 128 |

lowest : 4 5 6 7 8 , highest: 148 149 150 151 152

APTT: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12142 | 2549 | 631 | 1 | 40.06 | 9.533 | 30.1 | 31.4 | 34.1 | 37.7 | 42.7 | 49.9 | 56.6 |

lowest : 21.4 21.6 23.4 23.5 23.6 , highest: 160.7 163.0 168.7 171.6 176.1

SODIUM: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13409 | 1282 | 58 | 0.994 | 137.2 | 5.034 | 129 | 132 | 135 | 137 | 140 | 142 | 144 |

lowest : 106 108 109 110 112 , highest: 161 165 166 168 170

CA: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13415 | 1276 | 185 | 1 | 2.214 | 0.2213 | 1.89 | 1.96 | 2.09 | 2.22 | 2.35 | 2.45 | 2.51 |

lowest : 1.03 1.15 1.18 1.20 1.23 , highest: 3.84 3.88 3.96 4.18 4.40

PHOS: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13449 | 1242 | 306 | 1 | 1.048 | 0.3993 | 0.55 | 0.64 | 0.81 | 0.99 | 1.20 | 1.47 | 1.74 |

lowest : 0.30 0.31 0.32 0.33 0.34 , highest: 4.36 4.43 4.53 5.48 6.22

MG: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12822 | 1869 | 146 | 0.999 | 0.8136 | 0.1609 | 0.59 | 0.64 | 0.72 | 0.81 | 0.89 | 0.98 | 1.06 |

lowest : 0.20 0.21 0.22 0.26 0.28 , highest: 1.83 1.88 1.96 2.07 2.22

HS: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11630 | 3061 | 169 | 1 | 5.413 | 2.625 | 2.2 | 2.7 | 3.7 | 5.0 | 6.6 | 8.5 | 10.0 |

lowest : 1.3 1.4 1.5 1.6 1.7 , highest: 19.8 20.2 22.2 22.3 22.7

GBIL: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13250 | 1441 | 885 | 1 | 1.406 | 1.477 | 0.33 | 0.39 | 0.53 | 0.77 | 1.23 | 2.34 | 3.96 |

lowest : 0.11 0.12 0.13 0.14 0.15 , highest: 42.82 43.83 45.10 51.72 51.77

TP: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13108 | 1583 | 649 | 1 | 64.9 | 12.97 | 45.20 | 49.47 | 56.90 | 65.70 | 73.30 | 78.80 | 82.00 |

lowest : 29.9 30.0 30.3 30.5 30.6 , highest: 107.8 108.1 108.7 112.8 120.9

ALB: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13015 | 1676 | 401 | 1 | 33.42 | 8.513 | 21.3 | 23.6 | 27.9 | 33.6 | 39.1 | 43.2 | 45.2 |

lowest : 10.0 10.2 10.5 10.6 10.7 , highest: 52.9 53.2 53.7 54.0 55.7

AMY: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10778 | 3913 | 488 | 1 | 90.83 | 100.5 | 18 | 23 | 33 | 49 | 76 | 125 | 187 |

lowest : 8 9 10 11 12 , highest: 4984 5248 40372 43970 56146

Value 0 500 1000 1500 2000 2500 4000 4500 5000 40500 44000 56000  
 Frequency 10432 268 39 14 12 4 2 2 2 1 1 1  
 Proportion 0.968 0.025 0.004 0.001 0.001 0.000 0.000 0.000 0.000 0.000 0.000 0.000

For the frequency table, variable is rounded to the nearest 500

PAMY: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 7577 | 7114 | 280 | 0.999 | 41.66 | 47.28 | 7 | 9 | 14 | 22 | 36 | 64 | 97 |

lowest : 1 2 3 4 5 , highest: 1673 2083 2116 3066 38369

Value 0 500 1000 1500 2000 3000 38500  
 Frequency 7495 65 7 6 2 1 1  
 Proportion 0.989 0.009 0.001 0.001 0.000 0.000 0.000

For the frequency table, variable is rounded to the nearest 500

LIP: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10992 | 3699 | 444 | 1 | 63.82 | 89.88 | 6 | 8 | 14 | 23 | 40 | 79 | 135 |

lowest : 0 1 2 3 4 , highest: 11469 15843 18560 22339 45991

CHE: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12244 | 2447 | 997 | 1 | 4.79 | 2.378 | 1.70 | 2.17 | 3.15 | 4.60 | 6.22 | 7.65 | 8.49 |

lowest : 0.98 0.99 1.00 1.01 1.02 , highest: 12.39 12.55 12.97 13.32 13.89

AP: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13291 | 1400 | 672 | 1 | 118.8 | 91.51 | 42 | 49 | 63 | 84 | 123 | 206 | 302 |

lowest : 11 14 15 16 17 , highest: 1980 2132 2549 2596 2995

LDH: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12977 | 1714 | 1137 | 1 | 331.2 | 240.9 | 136 | 152 | 187 | 239 | 332 | 508 | 724 |

lowest : 39 46 54 55 56 , highest: 10473 10784 10822 11246 13906

CK: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12611 | 2080 | 1506 | 1 | 385 | 615.4 | 18 | 25 | 42 | 80 | 184 | 577 | 1155 |

lowest : 8 9 10 11 12 , highest: 60799 63011 82180 83880 98801

GLU: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10499 | 4192 | 389 | 1 | 126.4 | 48.3 | 78 | 85 | 97 | 113 | 138 | 177 | 216 |

lowest : 19 22 23 26 28 , highest: 843 848 890 1349 1403

TRIG: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9630 | 5061 | 538 | 1 | 141.7 | 90.33 | 54 | 64 | 83 | 115 | 165 | 241 | 307 |

lowest : 14 15 16 20 22 , highest: 1796 2247 2662 2918 5440

CHOL: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9646 | 5045 | 339 | 1 | 150.8 | 59.23 | 74 | 89 | 113 | 145 | 182 | 219 | 243 |

lowest : 25 26 27 28 29 , highest: 646 662 676 710 1104

BASOR: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

n missing distinct Info Mean Gmd .05 .10 .25   
 13959 732 419 0.322 0.145 0.2679 0.0000 0.0000 0.0000   
 .50 .75 .90 .95   
 0.0000 0.0000 0.5501 1.0526

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lowest : | 0.0000000 | 0.1358696 | 0.1545595 | 0.1652893 | 0.1818182 |
| highest: | 11.1111111 | 15.2173913 | 16.6666667 | 18.4210526 | 23.6559140 |

EOSR: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

n missing distinct Info Mean Gmd .05 .10 .25   
 13959 732 927 0.891 1.297 1.825 0.0000 0.0000 0.0000   
 .50 .75 .90 .95   
 0.5882 1.7857 3.4900 5.0000

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lowest : | 0.0000000 | 0.1834862 | 0.2028398 | 0.2178649 | 0.2188184 |
| highest: | 39.1752577 | 46.6019417 | 46.9026549 | 50.0000000 | 73.4883721 |

LYMR: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13959 | 732 | 3121 | 1 | 14.61 | 11.87 | 2.752 | 4.000 | 6.757 | 11.340 | 18.182 | 27.869 | 36.620 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lowest : | 0.0000000 | 0.3215434 | 0.4484305 | 0.4608295 | 0.4636785 |
| highest: | 97.2413793 | 97.4193548 | 98.0000000 | 99.1847826 | 100.0000000 |

MONOR: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13959 | 732 | 2334 | 1 | 8.793 | 5.4 | 2.000 | 3.390 | 5.634 | 8.000 | 10.870 | 14.141 | 17.021 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lowest : | 0.0000000 | 0.2747253 | 0.3412969 | 0.3448276 | 0.4566210 |
| highest: | 68.5446009 | 69.2307692 | 70.3703704 | 72.7272727 | 100.0000000 |

NEUR: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13959 | 732 | 3850 | 1 | 75.15 | 15.6 | 47.42 | 57.88 | 69.23 | 78.33 | 85.32 | 90.13 | 92.63 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| lowest : | 0.000000 | 1.484829 | 1.935484 | 1.960784 | 2.413793 |
| highest: | 99.122807 | 99.166667 | 99.476440 | 99.484536 | 100.000000 |

PDW: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13589 | 1102 | 167 | 1 | 12.29 | 2.375 | 9.3 | 9.8 | 10.8 | 12.0 | 13.4 | 15.1 | 16.4 |

lowest : 6.6 6.8 6.9 7.0 7.1 , highest: 24.1 24.7 24.9 25.2 25.3

RBC: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14230 | 461 | 65 | 0.999 | 3.936 | 0.8772 | 2.7 | 2.9 | 3.4 | 3.9 | 4.5 | 4.9 | 5.2 |

# Appendix D — Descriptive summaries

here() starts at /Users/markbaillie/Documents/ida-tmp-set-up

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

Loading required package: lattice

Loading required package: survival

Loading required package: Formula

Attaching package: 'Hmisc'

The following object is masked from 'package:gt':  
  
 html

The following objects are masked from 'package:dplyr':  
  
 src, summarize

The following objects are masked from 'package:base':  
  
 format.pval, units

Attaching package: 'sjlabelled'

The following object is masked from 'package:ggplot2':  
  
 as\_label

The following object is masked from 'package:dplyr':  
  
 as\_label

kurtosis max max\_1 max\_2 max\_3 max\_4 max\_5   
 50 50 50 50 50 50 50   
 mean median min min\_1 min\_2 min\_3 min\_4   
 50 50 50 50 50 50 50   
 min\_5 mode n\_distinct qt\_05 qt\_10 qt\_25 qt\_50   
 50 50 50 50 50 50 50   
 qt\_75 qt\_90 qt\_95 range sd skewness   
 50 50 50 50 50 50

| PARAM | PARAMCD | mean | min | median |
| --- | --- | --- | --- | --- |
| Activated partial thromboplastin time (sec) | APTT | 40.05639928 | 21.40 | 37.7000000 |
| Alanin transaminase (U/L) | ALAT | 67.66221541 | 0.00 | 26.0000000 |
| Albumin (G/L) | ALB | 33.42457933 | 10.00 | 33.6000000 |
| Alkaline phosphatase (U/L) | AP | 118.77955007 | 11.00 | 84.0000000 |
| Amylase (U/L) | AMY | 90.83225088 | 8.00 | 49.0000000 |
| Aspartate transaminase (U/L) | ASAT | 86.89525006 | 3.00 | 31.0000000 |
| Basophile ratio (%) | BASOR | 0.14504559 | 0.00 | 0.0000000 |
| Basophiles (G/L) | BASO | 0.01724991 | 0.00 | 0.0000000 |
| Bilirubin (mg/dl) | GBIL | 1.40627774 | 0.11 | 0.7700000 |
| Blood platelets (G/L) | PLT | 220.02546249 | 0.00 | 204.0000000 |
| Blood urea nitrogen (mg/dl) | BUN | 22.65653971 | 2.50 | 16.6000000 |
| C-reactive protein (mg/dl) | CRP | 10.91585237 | 0.00 | 8.5700000 |
| Calcium (mmol/L) | CA | 2.21379650 | 1.03 | 2.2200000 |
| Cholesterol (mg/dl) | CHOL | 150.79991706 | 25.00 | 145.0000000 |
| Cholinesterase (kU/L) | CHE | 4.79039611 | 0.98 | 4.6000000 |
| Creatinine (mg/dl) | CREA | 1.32930498 | 0.26 | 1.0000000 |
| Creatinine kinases (U/L) | CK | 385.01435255 | 8.00 | 80.0000000 |
| Eosinophil ratio (%) | EOSR | 1.29742747 | 0.00 | 0.5882353 |
| Eosinophils (G/L) | EOS | 0.11481863 | 0.00 | 0.1000000 |
| Fibrinogen (mg/dl) | FIB | 547.36481359 | 55.00 | 529.0000000 |
| Gamma-glutamyl transpeptidase (G/L) | GGT | 115.05838112 | 3.00 | 49.0000000 |
| Glucoses (mg/dl) | GLU | 126.40565768 | 19.00 | 113.0000000 |
| Haematocrit (%) | HCT | 34.48307734 | 0.00 | 34.3000000 |
| Haemoglobin (G/L) | HGB | 11.56801365 | 3.00 | 11.4000000 |
| Lactate dehydrogenase (U/L) | LDH | 331.15165292 | 39.00 | 239.0000000 |
| Lipases (U/L) | LIP | 63.82141557 | 0.00 | 23.0000000 |
| Lymphocyte ratio (% (mg/dl)) | LYMR | 14.61410741 | 0.00 | 11.3402062 |
| Lymphocytes (G/L) | LYM | 1.36577725 | 0.00 | 1.0000000 |
| Magnesium (mmol/L) | MG | 0.81360084 | 0.20 | 0.8100000 |
| Mean corpuscular hemoglobin (fl) | MCH | 29.57993037 | 14.90 | 29.7000000 |
| Mean corpuscular hemoglobin concentration (g/dl) | MCHC | 33.47141784 | 23.70 | 33.5000000 |
| Mean corpuscular volume (pg) | MCV | 88.35160762 | 51.00 | 88.3000000 |
| Mean platelet volume (fl) | MPV | 10.38182858 | 7.30 | 10.3000000 |
| Monocyte ratio (%) | MONOR | 8.79255066 | 0.00 | 8.0000000 |
| Monocytes (G/L) | MONO | 0.85273797 | 0.00 | 0.8000000 |
| Neutrophile ratio (%) | NEUR | 75.15086887 | 0.00 | 78.3333333 |
| Neutrophiles (G/L) | NEU | 8.36677648 | 0.00 | 7.3000000 |
| Normotest (%) | NT | 83.22087696 | 4.00 | 83.0000000 |
| Pancreas amylase (U/L) | PAMY | 41.65830804 | 1.00 | 22.0000000 |
| Patient Age (years) | AGE | 56.16806208 | 16.00 | 58.0000000 |
| Phosphate (mmol/L) | PHOS | 1.04768756 | 0.30 | 0.9900000 |
| Platelet distribution width (%) | PDW | 12.29323718 | 6.60 | 12.0000000 |
| Potassium (mmol/L) | POTASS | 4.00301191 | 1.92 | 3.9500000 |
| Red blood cell distribution width (%) | RDW | 14.99738299 | 10.60 | 14.5000000 |
| Red blood count (T/L) | RBC | 3.93573436 | 1.00 | 3.9000000 |
| Sodium (mmol/L) | SODIUM | 137.21135059 | 106.00 | 137.0000000 |
| Total protein (G/L) | TP | 64.90141135 | 29.90 | 65.7000000 |
| Triclyceride (mg/dl) | TRIG | 141.70799585 | 14.00 | 115.0000000 |
| Uric acid (mg/dl) | HS | 5.41254514 | 1.30 | 5.0000000 |
| White blood count (G/L) | WBC | 11.22715651 | 0.00 | 9.6000000 |

# Appendix E — Multivariate analyses

First load the required packages and data. Note:, from the univariate analyses the analysis data set for the lab parameters has been updated to include transformed variables. Therefore, we load the second iteration of the data data/IDA/ADLB\_02.rds.

## E.1 V1: Association with structural variables

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |
| --- |
|  |

## E.2 VE3: Redundancy

### E.2.1 Linear model

VIF for the all predictor model.

The available sample size is 3979 (0.42 %).

| Parameter code | Variance inflation factor | Multiple R-squared |
| --- | --- | --- |
| SEX | 1.3 | 0.25 |
| MCV | 129.5 | 0.99 |
| HGB | 254.3 | 1.00 |
| HCT | 249.4 | 1.00 |
| PLT | 1.9 | 0.48 |
| MCH | 179.4 | 0.99 |
| MCHC | 47.5 | 0.98 |
| RDW | 1.9 | 0.47 |
| MPV | 9.2 | 0.89 |
| MONO | 4.3 | 0.77 |
| BASO | 2.6 | 0.62 |
| NT | 1.5 | 0.32 |
| APTT | 1.2 | 0.18 |
| FIB | 2.9 | 0.66 |
| SODIUM | 1.4 | 0.30 |
| POTASS | 1.4 | 0.29 |
| CA | 2.1 | 0.52 |
| PHOS | 1.6 | 0.36 |
| MG | 1.3 | 0.23 |
| BUN | 3.2 | 0.69 |
| HS | 1.8 | 0.43 |
| TP | 4.5 | 0.78 |
| ALB | 6.3 | 0.84 |
| CHE | 2.8 | 0.65 |
| GLU | 1.3 | 0.20 |
| TRIG | 1.4 | 0.29 |
| CHOL | 1.9 | 0.47 |
| CRP | 2.8 | 0.65 |
| EOSR | 47.3 | 0.98 |
| LYMR | 1,321.2 | 1.00 |
| MONOR | 321.4 | 1.00 |
| NEU | 8.9 | 0.89 |
| NEUR | 2,179.1 | 1.00 |
| PDW | 9.1 | 0.89 |
| RBC | 40.1 | 0.98 |
| AGE | 1.5 | 0.31 |
| ALAT\_T | 3.9 | 0.74 |
| AMY\_T | 2.7 | 0.63 |
| AP\_T | 2.3 | 0.57 |
| ASAT\_T | 6.6 | 0.85 |
| CK\_T | 2.2 | 0.55 |
| CREA\_T | 2.9 | 0.65 |
| EOS\_T | 3.5 | 0.71 |
| GBIL\_T | 1.6 | 0.38 |
| GGT\_T | 2.6 | 0.62 |
| LDH\_T | 2.2 | 0.55 |
| LIP\_T | 2.2 | 0.54 |
| LYM\_T | 4.5 | 0.78 |
| PAMY\_T | 3.4 | 0.71 |
| WBC\_T | 15.7 | 0.94 |

# Appendix F — Supplementary Example

## F.1 Overview

Define key predictors without pseudo-log transform (‘orig’) and transformed (‘trans’).

In the following examples we use the Bacteremia data with complete observations regarding the key predictors PLT, BUN, NEU, AGE, CREA\_T, WBC\_noNEU\_T, which represent 93.9% of the dataset **w.r.t key predictors**. We will fit a global logistic regression model with the outcome **BACTEREMIAN** (i.e. presence of bacteremia) and the key predictors as covariates. We will use pseudo-log transformations as suggested in the IDA. Within the model, all key predictors will be transformed by fractional polynomials of order 1 (df = 2).

The aim of the examples is to showcase how decisions derived from IDA influence the results of the fitted model.

## F.2 Global Model

The global model will be fitted by the *mfp* function. If not indicated otherwise, we will use the fp-transformations of the key predictors determined in global model in all consecutive models. For all models we report McFaddens’s R² and the AUC, i.e. the area under the ROC curve, and boxplots comparing “**BACTEREMIAN**” predictions with outcomes.

### F.2.1 Model Summary

The global model is specified as print(global\_formula, quote = TRUE). The model is fit to the complete cases data on the **transformed** scale stored in the data set data\_model\_trans.

Report the global model fit.

Table 1: global model - including transformed variables

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.69 | 0.59 | -4.58 | 4.59 × 10^-6 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.21 | 0.25 | -20.52 | 1.56 × 10^-93 |
| I(((NEU + 0.1)/10)^0.5) | 1.53 | 0.12 | 13.09 | 3.76 × 10^-39 |
| I((AGE/100)^1) | 1.69 | 0.20 | 8.49 | 2.12 × 10^-17 |
| I(CREA\_T^1) | 0.71 | 0.20 | 3.60 | 3.22 × 10^-4 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.57 | 1.03 × 10^-2 |
| I((BUN/10)^1) | 0.01 | 0.02 | 0.43 | 6.66 × 10^-1 |

Plot predictions by outcome and report the model AUC.

|  |
| --- |
| Global model - including covariates on the transformed psuedo log scale |

### F.2.2 Functional forms of global model

We now take a look at the functional forms of the covariates in the global model, which are determined by the fp algorithm. Besides scaling factors, only for WBC\_noNEU\_T the fp algorithm chose a non-linear transformation (note the ‘^0.5’ in the term column). This means all other covariates enter the model in a linear fashion. In the following effect plots, each variable is adjusted to the median of the other variables in the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | (a) | | |  | | --- | | (b) | | |  | | --- | | (c) | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | (d) | | |  | | --- | | (e) | | |  | | --- | | (f) | |

Figure F.1: Functional forms of global model - repeats over plots

## F.3 Example 1: to transform or not to transform

Only for one out of the six key predictors did the fp algorithm chose a non-linear transformation. But out of those six variables, four were pseudo-log transformed before entering the model. In the first example we want to compare the global model to a model using the key predictors on their original scale.

Table 1: Global model without pseudo-log tranformations

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -3.18 | 0.28 | -11.17 | 5.48 × 10^-29 |
| log((WBC\_noNEU + 0.2)) | -1.25 | 0.06 | -20.22 | 5.92 × 10^-91 |
| I(((NEU + 0.1)/10)^0.5) | 1.52 | 0.12 | 12.99 | 1.34 × 10^-38 |
| I((AGE/100)^1) | 1.64 | 0.20 | 8.26 | 1.51 × 10^-16 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.53 | 1.13 × 10^-2 |
| I((BUN/10)^1) | 0.02 | 0.02 | 0.68 | 5.00 × 10^-1 |
| I(CREA^-0.5) | -0.77 | 0.21 | -3.69 | 2.20 × 10^-4 |

Note the different fp-transformation arising when the key predictors are not pseudo-log transformed. On the original scale, three covariates instead of one now enter the model via a non-linear fp-transformation. This suggests that a transformation prior to the regression model ‘outsources’ the need for transformations within the model. Now let us compare the model performances.

|  |
| --- |
| Global model(s) with covariates on original and transformed scale. |

With regards to McFadden’s R² and the AUC, the differences between the two approaches is marginal.

Next we will compare the differences of the functional forms in the two models for those covariates where a pseudo-log transformation was suggested in IDA. We will look at the log odds for bacteremia by each covariate on the original and the transformed scale, and compare the global model using the original and the pseudo-log transformed covariates. Each variable is adjusted for the median of all other variables used.

|  |  |
| --- | --- |
|  |  |

|  |
| --- |
| Comparing functional form of both global models. |

## F.4 Example 2: the support of a model determines what it can explain

Next we compare the global model to a model were for an important variable, in our case we chose age, the variable support is reduced to the central 50% of the data (i.e. data within the 25% and 75% quantiles). Again, in the reduced models we use the same fp-transformations as in the global model.

| data | AUC | scaled Brier score |
| --- | --- | --- |
| complete | 0.731 | 0.06483 |
| central 50% | 0.722 | 0.05284 |
| 50% sample | 0.717 | 0.05619 |

|  |
| --- |
| Model predictions based on different missing data scenarios. |

## F.5 Example 3: the limits of mulitiple imputation

To show the effect of multiple imputation if the number of missing values is high, we construct a dataset with 50% artificially generated missing values in one variable. First, recall the output of the complete model, relying on the Bacteremia data with complete cases regarding the key predictors.

Table 1: global model

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.69 | 0.59 | -4.58 | 4.59 × 10^-6 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.21 | 0.25 | -20.52 | 1.56 × 10^-93 |
| I(((NEU + 0.1)/10)^0.5) | 1.53 | 0.12 | 13.09 | 3.76 × 10^-39 |
| I((AGE/100)^1) | 1.69 | 0.20 | 8.49 | 2.12 × 10^-17 |
| I(CREA\_T^1) | 0.71 | 0.20 | 3.60 | 3.22 × 10^-4 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.57 | 1.03 × 10^-2 |
| I((BUN/10)^1) | 0.01 | 0.02 | 0.43 | 6.66 × 10^-1 |

Creatinine (‘CREA’) is significant at a level that might not survive substantial missingness. We thus create a dataset were we artificially introduce 50% missing creatinine values, missing completely at random.

Next we fit a ‘complete case’ model in the case of missing creatinine data, using the fp-transformations from the global model.

Now we impute the missing creatinine data using MICE with 50 imputations, fit the model using the fp-transformations from the global model and pool the results.

We now can compare the outputs of the complete model, the complete model with missing data (i.e. only half of the original complete data is used), and the imputed model.

Table 1: Comparison of the complete model, the complete model with missing data

| term | model | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | global model | -2.694 | 0.588 | -4.583 | 4.59 × 10^-6 |
|  | missing, complete cases | -2.014 | 0.841 | -2.396 | 1.66 × 10^-2 |
|  | missing, imputed | -1.799 | 0.792 | -2.272 | 2.41 × 10^-2 |
| I(((NEU + 0.1)/10)^0.5) | global model | 1.525 | 0.117 | 13.090 | 3.76 × 10^-39 |
|  | missing, complete cases | 1.580 | 0.168 | 9.391 | 5.95 × 10^-21 |
|  | missing, imputed | 1.524 | 0.117 | 13.076 | 7.70 × 10^-39 |
| I(((PLT + 1)/100)^1) | global model | -0.082 | 0.032 | -2.566 | 1.03 × 10^-2 |
|  | missing, complete cases | -0.042 | 0.045 | -0.946 | 3.44 × 10^-1 |
|  | missing, imputed | -0.081 | 0.032 | -2.545 | 1.09 × 10^-2 |
| I((AGE/100)^1) | global model | 1.689 | 0.199 | 8.487 | 2.12 × 10^-17 |
|  | missing, complete cases | 2.161 | 0.284 | 7.601 | 2.94 × 10^-14 |
|  | missing, imputed | 1.674 | 0.199 | 8.412 | 4.43 × 10^-17 |
| I((BUN/10)^1) | global model | 0.010 | 0.024 | 0.432 | 6.66 × 10^-1 |
|  | missing, complete cases | 0.033 | 0.034 | 0.984 | 3.25 × 10^-1 |
|  | missing, imputed | 0.038 | 0.029 | 1.321 | 1.87 × 10^-1 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | global model | -5.210 | 0.254 | -20.516 | 1.56 × 10^-93 |
|  | missing, complete cases | -5.339 | 0.364 | -14.657 | 1.22 × 10^-48 |
|  | missing, imputed | -5.192 | 0.254 | -20.459 | 1.15 × 10^-91 |
| I(CREA\_T^1) | global model | 0.712 | 0.198 | 3.597 | 3.22 × 10^-4 |
|  | missing, complete cases | 0.362 | 0.280 | 1.291 | 1.97 × 10^-1 |
|  | missing, imputed | 0.388 | 0.274 | 1.418 | 1.58 × 10^-1 |

The z-statistic for creatinine drops from 2.98 to 1.49 when half the data is missing. Also in other variables the z-statistic is less extreme in the ‘missing, complete case analysis’ compared to the global model. The interesting observations is that MI recreates estimates and standard errors very close to the global model in most variables, but not in the one that was being imputed, namely creatinine. In variable selection, creatinine, which is highly significant in the ‘true’ model, is likely to be dropped, based on the imputed data.

## F.6 Example 4: Interpretation of regression coefficient ‘size’

The variables WBC\_noNEU and WBC\_noNEU\_T are on two very different scales:

Plot

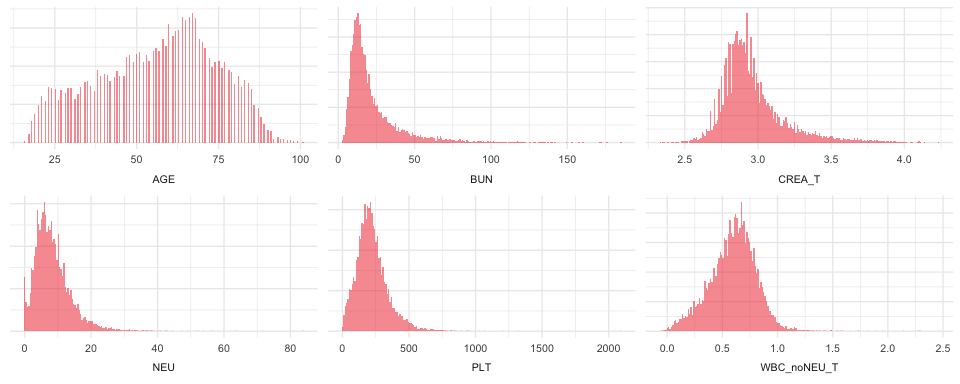


Table the standardized regression coefficients.

Table 1: Global model - standardized regression estimates

| term | estimate |
| --- | --- |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.210 |
| I(((NEU + 0.1)/10)^0.5) | 1.525 |
| I((AGE/100)^1) | 1.689 |
| I(CREA\_T^1) | 0.712 |
| I(((PLT + 1)/100)^1) | -0.082 |
| I((BUN/10)^1) | 0.010 |

Table standardized coefficients.

Table 1: Global model - standardized regression coefficients

| variable | standardized beta |
| --- | --- |
| WBC\_noNEU | 0.7019 |
| NEU | 8.5632 |
| AGE | 0.3072 |
| CREA | 0.1562 |
| PLT | 0.0993 |
| BUN | 0.1797 |

Let us recall the two estimates to the covariates WBC\_noNEU and WBC\_noNEU\_T.

Table 1: Comparing original scale and transformed

| term | estimate |
| --- | --- |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.21 |
| log((WBC\_noNEU + 0.2)) | -1.25 |

(Suggestion: show this with models without fp transform?)

Because the fp-transformations further complicate the interpretation of the regression coefficients, let us consider two logisitc regression models with WBC\_noNEU and t\_WBC\_noNEU as single covariate, respectively.

Table 1: Comparing original scale and transformed

| term | estimate |
| --- | --- |
| WBC\_noNEU | -0.56 |
| WBC\_noNEU\_T | -2.87 |

The estimates -0.56 and -2.87 the change in log odds for the outcome when the ‘term’ variable changes by 1 unit, but cannot be compared directly. A unit change is only a small step on the original scale, where WBC\_noNEU covers values from -0.15 up to 152.74. In comparison, WBC\_noNEU\_T lies between up to -, so change of unit cover almost half the range of the variable.

## F.7 Example 5: Plot of functional form should be resticted to areas with high density

The functional forms have wide confidence intervals when the data is sparse. In presentations of the effects, plots of the functional forms can be limited to areas with high density. In this analysis, PLT was very sparse above ~800 [UNITS], which is reflected in a large confidence interval for high PLT values. In the effect plot PLT values could be limited to values <800 [UNITS].

|  |
| --- |
| Model predictions |

|  |
| --- |
| WBC\_noNEU comparison |

|  |
| --- |
| PLT comparison |

# Appendix G — Pseudo-log transformations

## G.1 Introduction

This supplemental section illustrates how pseudo-log transformations can be used to transform skewed distributions towards normality.

The transformation is a pseudo-logarithmic transformation mentioned by (Johnson 1949). It has the following advantages over ordinary logarithmic transformations:

* it is defined also for as
* it is a signed logarithmic transformation, and is defined also for negative values as

Of course, this comes at the cost of deviation from the logarithmic transformation in terms of interpretability.

The parameter may be used to adapt the transformation to a specific range of an empirical distribution.

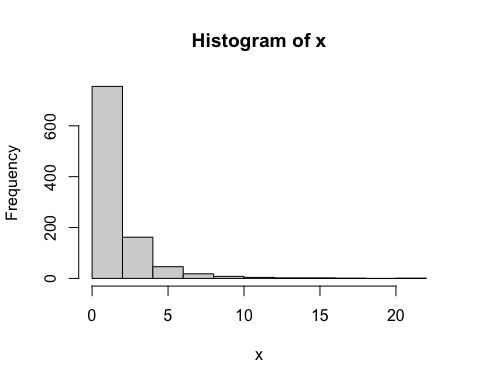
We define the pseudo-logarithmic transformation in R as:

pseudo\_log <- function(x, sigma=1, base=10) asinh(x/(2 \* sigma))/log(base)  
inv\_pseudo\_log <- function(x, sigma=1, base=10) 2 \* sigma \* sinh(x \* log(base))

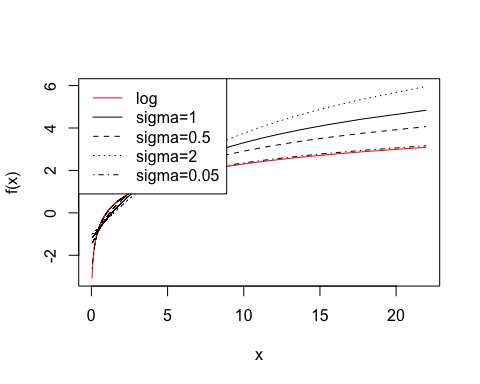
## G.2 Impact of choice of

Next, we investigate how the parameter impacts the result of the transformation. We assume that follows a log-normal distribution, and we will show results of with different choices of . We center and scale before plotting.

p <- seq(0.001, 0.999, 0.001)  
x <- exp(qnorm(p, mean=0, sd=1))  
  
hist(x)

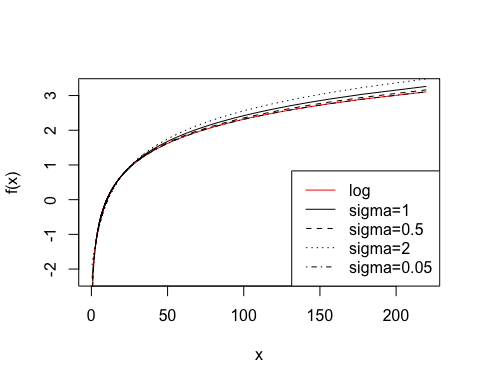


y <- cbind(log(x), scale(pseudo\_log(x, 1)), scale(pseudo\_log(x,0.5)), scale(pseudo\_log(x, 2)), scale(pseudo\_log(x, 0.05)))  
  
  
plot(x, y[,2], type="l", ylab="f(x)", ylim=range(y))  
lines(x, y[,1], col="red")  
lines(x, y[,3], type="l", lty=2)  
lines(x, y[,4], type="l", lty=3)  
lines(x, y[,5], type="l", lty=4)  
legend("topleft", lty=c(1,1,2,3,4), col=c("red","black","black","black","black"),   
 legend=c("log", "sigma=1", "sigma=0.5", "sigma=2", "sigma=0.05"))



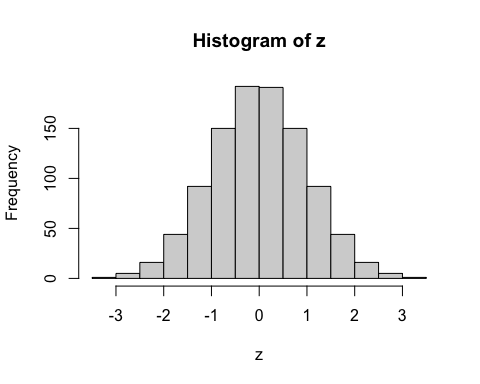
In the next code chunk, we multiply by 10 and repeat the exercise. We learn that the transformations become more similar and the choice of less relevant.

x <- x\*10  
  
y <- cbind(scale(log(x)), scale(pseudo\_log(x, 1)), scale(pseudo\_log(x,0.5)), scale(pseudo\_log(x, 2)), scale(pseudo\_log(x, 0.05)))  
  
  
plot(x, y[,2], type="l", ylab="f(x)")  
lines(x, y[,1], col="red")  
lines(x, y[,3], type="l", lty=2)  
lines(x, y[,4], type="l", lty=3)  
lines(x, y[,5], type="l", lty=4)  
legend("bottomright", lty=c(1,1,2,3,4), col=c("red","black","black","black","black"),   
 legend=c("log", "sigma=1", "sigma=0.5", "sigma=2", "sigma=0.05"))

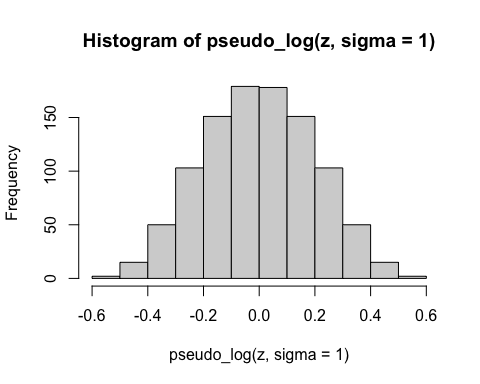


Finally, we apply the pseudo-logarithmic transformation to the original normal deviates. We learn that a higher value for the parameter makes the distribution ‘slimmer’ while a lower value makes it ‘fatter’, and it is even possible to induce bimodality with low values of sigma:

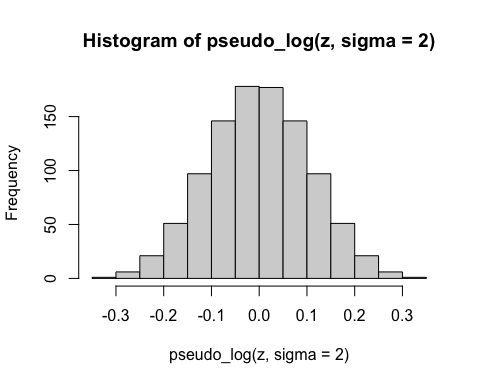
z <- qnorm(p, mean=0, sd=1)  
hist(z)



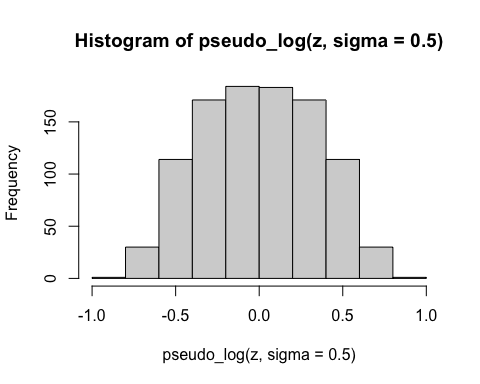
hist(pseudo\_log(z, sigma=1))



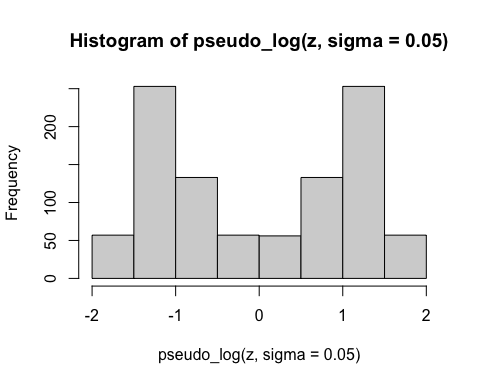
hist(pseudo\_log(z, sigma=2))



hist(pseudo\_log(z, sigma=0.5))



hist(pseudo\_log(z, sigma=0.05))



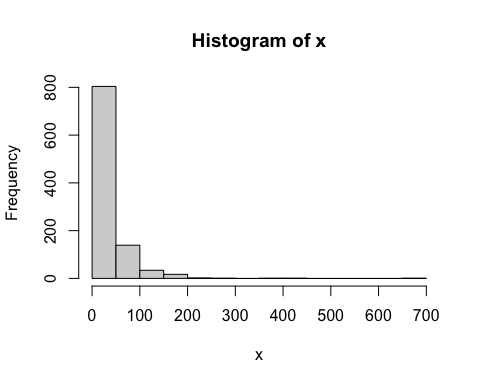
## G.3 Finding a parameter that best achieves normality

Any test statistic for testing normality could be chosen to find a suitable value of that induces normality into the transformed values. Here we use the (Pearson) correlation coefficient to compare the empirical distribution with normal deviates.

### G.3.1 Deviates from log normal distribution

We simulate from a shifted log normal distribution and evaluate the value of that optimizes agreement with a normal distribution:

x<-sort(exp(rnorm(1000)+3))  
  
hist(x)



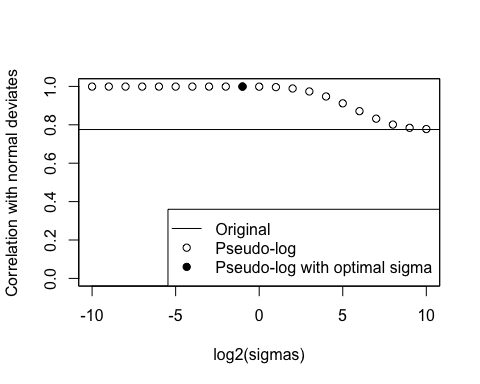
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

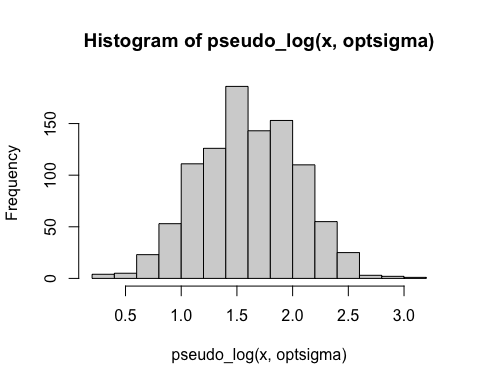
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 0.5

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



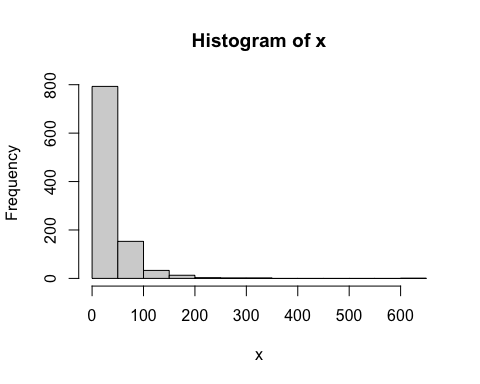
hist(pseudo\_log(x, optsigma))



### G.3.2 Deviates from exponential distribution

Also with an exponential distribution, the pseudo-logarithm may achieve a better agreement with a normal:

x<-sort(exp(rnorm(1000)+3))  
  
hist(x)



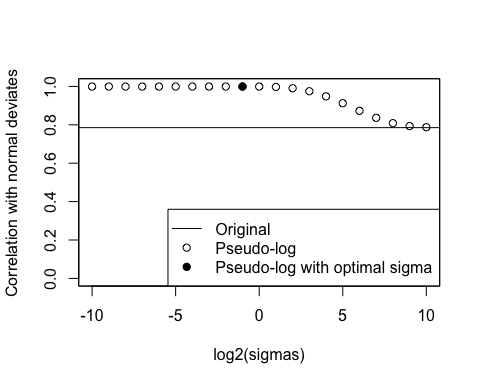
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

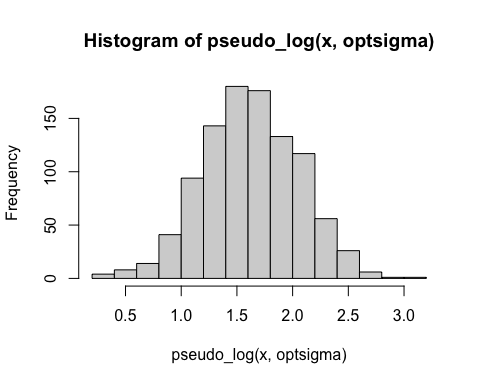
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 0.5

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



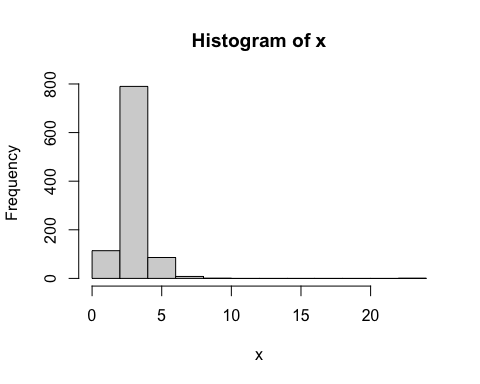
hist(pseudo\_log(x, optsigma))



### G.3.3 Deviates from mixture distribution

Now we mix a normal, lognormal and exponential distribution:

x1 <- scale(rnorm(1000))  
x2 <- scale(rexp(1000))  
x3 <- scale(exp(rnorm(1000)))  
  
p1 <- rbinom(1000, 1, 0.33)  
p2 <- rbinom(1000, 1, 0.5)  
  
x<-p1\*x1 + (1-p1)\*(p2\*x2+(1-p2)\*x3)  
x <- x-min(x)  
x<-sort(x)  
  
hist(x)



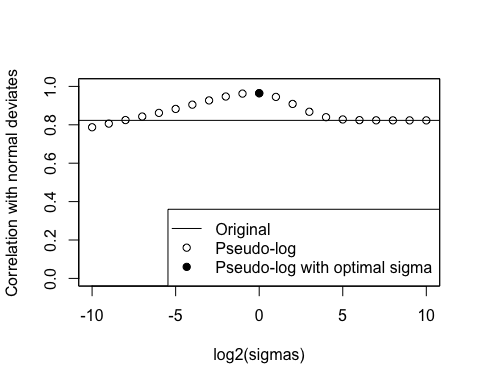
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

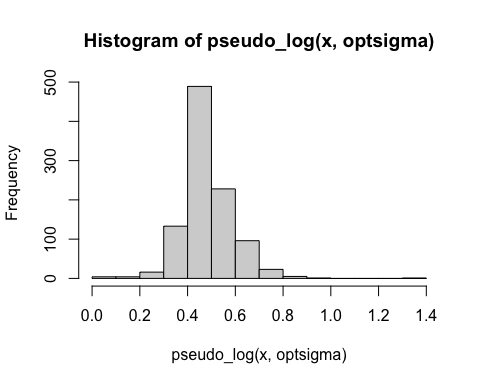
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 1

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



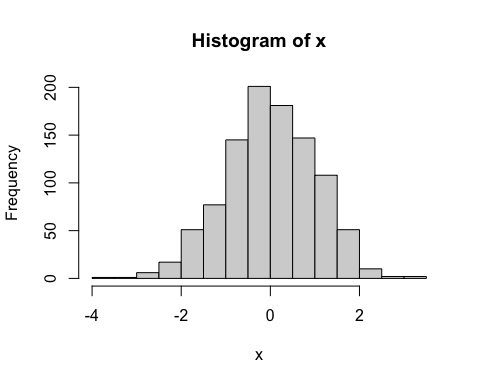
hist(pseudo\_log(x, optsigma))



### G.3.4 Normal deviates

With simulated normal deviates, the pseudo-logarithm cannot improve the already perfect normality.

x<-sort(rnorm(1000))  
  
hist(x)



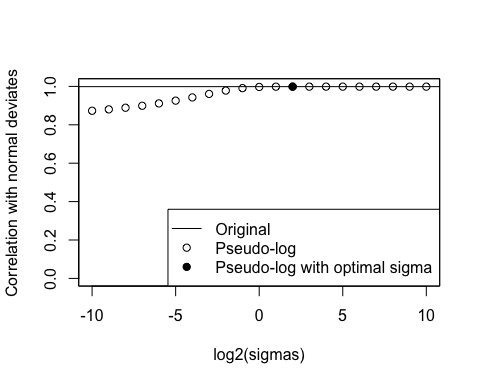
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 4

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
abline(h=origcor)  
box()



hist(pseudo\_log(x, optsigma))

