UNIVERSITÄT
BERN

Development of a big data analysis pipeline for precise reference interval estimation using ICD-10 code labelled data

Master thesis

Faculty of Science, University of Bern

handed in by

David Schaer

2023

Supervisor

Prof. Alexander Leichtle

October 26, 2023

Abstract

Reference Intervals are an important tool in medicine for interpreting blood analyte test results and patient care. Traditional methods for reference interval estimation often assume the underlying distribution of many analytes to follow Gaussian form. In practice however, data often shows skewed and heavily tailed distributions, challenging the validity of the current concept of reference intervals. This project aims at introducing a novel technique to approximate the underlying distribution of non-pathological values from mixed data from which reference intervals are drawn.

This newly developed technique called “Minus Distribution Method” involves stratifying data based on patient diagnoses and identifying test value subpopulations distorting the global distribution. By excluding distorting subpopulations prior to inference, subsequent statistical processing may be improved leading to more precise reference interval estimates. Adding to this, we introduce an innovative approach to cluster diagnoses using natural language processing based on co-occurrence in order to potentially improve our methods performance.

The Minus Distribution Method is presented in form of an interactive R Shiny app, representing a user-friendly platform for patient data visualization and automated estimation of reference intervals. Our efforts were aimed at generating reference intervals based on a given patients age and biological sex. Furthermore, we implemented features allowing for in-depth analysis of the composition of test value distributions in terms of patient diagnoses. The app has been developed in an effort to create a user-friendly tool for everyday use in clinical settings as to facilitate test value interpretation and ultimately, patient care

This project aspires to contribute to the refinement of reference interval estimation by introducing a novel data mining approach and presenting a practical tool for automated inference.

Contents

1 Introduction

1.1 Background

In clinical practice, reference intervals are a common tool to interpret test values and help clinicians in diagnosing patients. Tests for commonly screened blood analytes are conducted before and after treatment and compared to reference intervals as to assess the patient's medical state. A reference interval is a range in which a test value should lie in in order for the patient to be considered "healthy". A reference interval is traditionally defined by the central 95% of a "healthy" reference group distribution.[1] The exact definition of the term "healthy" remains one of the core difficulties in reference interval estimation, as there is no clear distinction between healthy and diseased.[2] The term "Healthy" may be interpreted in this context as individuals not affected by a specific disease.

However, it is well established that these intervals can vary greatly by multiple factors like genetics, lifestyle choices and diet, but also non-biological variance sources like test device settings and differing hospital practices. While many factors can be corrected for by standardisation of the measurement techniques, the most significant source of variance comes from differences in age and biological sex.[3] As individuals progress through life stages, their metabolic and hormonal systems change as well as their physique. As a consequence, the variance coming from these factors must be taken into account when trying to establish personalized reference intervals.[3,4] Age and biological sex are recommended as the main stratification factors by the official guidelines to reference interval generation.[5]

When comparing reference interval estimates from independent laboratories, a great amount of variance in estimates is unaccounted for even when controlling for known variance sources. These disparate estimates may result in patients being diagnosed differently depending on hospital site. Differences in blood analyte test manufacturing have been proposed to account for this effect but have been found insufficient to fully explain inter-laboratory variance.[6] While efforts are aimed at standardizing testing practice it is recommended for each testing site to generate and rely on their own reference interval estimates.[3]

1.2 Methods for reference interval estimation

Conventional practice holds that the lower and upper reference interval boundaries can be acquired by determining the 2.5% and 9.75% quantile of a population consisting exclusively of individuals not affected by a specific disease.[1] Acquiring such a population poses a challenging endeavour for which current methods for inference can be categorized into two distinct approaches: the direct and the indirect approach.[7] Central philosophy of both methods consists of acquiring a population of non-

diseased individuals from which reference interval estimates may be drawn by inferring its Gaussian parameters. However, even stratified by age and biological sex clinical data rarely follows Gaussian form and is often right-skewed.[8] Common practice often involves transforming the data with the Box-Cox method,[9] a parametric power transformation technique designed to transform data into a more Gaussian form. An additional technique is the transformation of the data to Z-scores, [1,10] in which the data is transformed into a distribution with approximately N(0,1). The limitation of this method consists in the data being transformed into an independent unit which requires the data to be transformed back to its original unit after processing. Furthermore, Z-scores are distribution dependent and may not reflect the relative position of datapoints accurately when facing outliers or a skewed distribution.

The difficulties introduced by medical data often not meeting the criteria of a Gaussian distribution may be addressed by resorting to non-parametric approaches. In this approach, the 2.5th and the 97.5th empirical centile are to be regarded as the 95% reference interval. A common technique is the *Harrell-David bootstrap*[11] approach which represents a sophisticated, non-parametric approach to reference interval estimation. However, even though non-parametric bootstrap approaches are considered suitable methods for reference interval estimation and in some cases more robust when dealing with skewed distributions[12], other sources like Daly et al.[13] argue that parametric approaches provide the least biased and most precise estimates.

1.3 Direct approach

The direct approach consists of *a priori* sampling of individuals not affected by a specific disease and estimating the reference intervals directly from the acquired distribution.[14] In practice, this entails recruiting a large number of non-diseased study participants. A minimum number of 120 participants is recommended by both the IFCC and the CLSI guideline C28-A3c for reference interval generation.[5,15] Drawing reference intervals from sample sizes smaller than recommended may result in unreliable estimates and high standard errors for reference limits. The estimation of reference intervals is either done under the assumption that the distribution is of Gaussian form or using non-parametric methods all aimed at estimating the central 95%. While computationally simple and easily interpretable, this method requires substantial resources and planning to gather sufficient participants of all ages in order to reach the minimal sample size.[16] An additional limitation is the fact that when new biochemical methods are developed or improved, the study needs to be repeated.[17]

1.4 Indirect approach

In the indirect approach, diseased and non-diseased populations are separated post data acquisition. In practice, this means considering the entirety of hospital data entries regardless of health status

and only then inferring the non-diseased population from which reference intervals are estimated.[14] Methods for inference entail many different approaches at the heart of which typically lie the inference of the Gaussian element of the non-diseased population.

Mitigating time and resource constraints, this approach allows for acquisition of substantially more data as well as insight into the distribution pattern of values from diseased populations.[18] Large datasets are important when creating personalized reference intervals as population size must be sufficiently large even after stratification in order to render the statistical methods viable.

1.5 Reference intervals for creatinine values

Creatinine is a break down product of creatine phosphate and is usually released in the body at a constant rate. Creatine phosphate is a high-energy compound used to replenish the ATP supply during periods of high energy demand.[19] As creatine phosphate is broken down to form ATP from ADP, it eventually leads to the formation of creatinine as a waste product and is cleared from the blood by the kidneys. Creatinine is a fairly reliable indicator of kidney diseases, high creatinine blood levels thus being an biomarker for kidney insufficiency or malfunction.[20,21]

The work presented in this project has been developed primarily using data from tests for Creatinine [Moles/volume] in Serum or Plasma (LOINC 14682-9). With 247'081 data entries in the Swiss BioRef dataset and a balanced age distribution for adults, creatinine was an ideal candidate for developing the methodology and R Shiny app presented in this thesis.

1.6 Ichihara Method

In their 2005 paper, Kiyoshi Ichihara and Tadashi Kawai[22] developed an iterative method for reference interval estimation, which we will refer to as "Ichihara Method" in this work. It allows for approximation of an underlying Gaussian element in a distribution where Gaussian assumptions are not met and also serves as a sophisticated way to deal with extreme outliers.

The conventional method of outlier removal in reference interval estimation consists of two times removing values exceeding the mean by +/- 3 standard deviations $\boldsymbol{\sigma}$.[22] This method performs poorly when dealing with non-Gaussian distributions or populations where outliers mainly lie on one side. The Ichihara Method has been demonstrated to be less sensitive to these effects and identifies the underlying populations more accurately. It aims at estimating the parameters of an underlying Gaussian distribution from a non-Gaussian or skewed distribution. In the context of reference interval estimation, the Ichihara Method aims to separate the main non-diseased population from a population containing both diseased as well as non-diseased individuals.

It works by iteratively removing proportions of the population which exceed the mean by +/- *k* * standard deviations. The truncation factor *k*, usually set to 1.6 is a parameter which determines how big of a proportion is trimmed in each iteration. To compensate for the trimming, the standard deviation is multiplied by a correction factor *c(k)* depending on the truncation factor. This process is repeated until the change in standard deviation from step *n-1* to *n* stabilizes, being when $\frac{\sigma_n - \sigma_{n-1}}{\sigma_{n-1}}$ falls below 10^{-5}. The central 95% may be drawn from the resulting distribution.

The Ichihara Method is particularly valuable when dealing with extreme outliers or skewed distributions and is used to gain Gaussian parameter estimates for reference interval estimation from approximately non-diseased distributions in our method.

2 Study Approval

This study was conducted in accordance of the ethical guidelines set forth by the Cantonal Ethics Commission for Human Research Bern (KEK) Waiver number: Req -2020-0063. Due to the anonymization of the patient data provided by the Insel Data Coordination Lab (IDCL), the study doesn't require a full application for ethical approval by the KEK under the Human Rights Association.

3 Aim of the thesis

Our efforts in the scope of this project have been aimed at developing an interactive application ("app") able to generate personalized reference intervals in order to facilitate data exploration for researchers and patient diagnosis for clinicians. The codebase for the app was written in the R language for its intuitive nature and interactive features made possible by the extension R Shiny[23]. The user interface should feature options for the main stratification factors biological sex and age, as well as allow for choosing different statistical tests, significance levels and clustering options.

Two statistical tests were implemented, being the t-test for its robustness and wide applicability, and the Wilcox test for accounting for cases in which the assumptions for an underlying Gaussian distribution are not met. As these statistical tests were carried out in parallel and in large numbers, a multiple hypothesis testing correction was implemented by controlling the false discovery rate (FDR).[24] In the context of this project, statistical testing should identify diagnoses with distributions deviating from the "Global Distribution" which represents a data subset of a given blood analyte stratified by biological sex and an age range of 10 years. Through statistical testing diagnoses may be classified as distorting the Global Distribution, leading to their exclusion as part of our method. This process is aimed at the reduction of the negative effects skewed distributions and extreme outliers may have on the subsequent inference methods. Ultimately, this would allow us to more accurately approximate the underlying non-diseased population for which Gaussian parameters and reference interval estimates could be obtained with the Ichihara Method.

In this project we decided to implement clustering techniques to investigate grouping effects based on data labelled with diagnoses from the International Classification of Diseases ("ICD-10").[25], which drew our interest for the following reasons: Working with groups of diagnoses rather than single diagnoses could mean working with more complete representations of diseases rather than just their individual symptoms. A group of diagnoses would represent a disease type and potentially model reality more closely. Furthermore, clustering diagnoses by co-occurrence could shine light on interesting connections between symptoms and expose trends as well as further our understanding of the behaviour of some diseases.

All these methods were implemented with the ambition to refine reference interval estimation by making use of ICD-10 code labelled data and summarizing our methodology in the form of a downloadable app.

4 Methods

4.1 The Minus Distribution Method

Most methods for indirect reference interval estimation base their approach on the assumption that with the help of statistical methods the proposed non-diseased population from the mixed dataset can be extrapolated. Our approach aims to identify subpopulations distorting the Global Distribution and to exclude them from the complete dataset prior to statistical inference. This is achieved by using data labelled with codes from the 10th revision of the International Classification of Diseases ("ICD-10").[25] As for the format used in this project, each test value is labelled with up to 5 codes describing the patient's symptoms and condition. These codes are appointed by clinicians when treating patients and are in the ICD-10 GM (German Modification) format.[26]

The ICD-10 code system is an internationally recognized classification system for clinical diagnoses first established in 1994 by the WHO and last updated in 2019.[27] It facilitates communication, identification and reporting of medical conditions across the globe by using a common language for pathological phenomena.

Information on ICD-10 diagnoses constitutes an additional factor data may be stratified by. Exploratory data analysis on creatinine data grouped by diagnosis showed that some of these subsets can be identified as significantly different from the remaining population using the t-test at a significance level of 0.05. These findings suggest that on the basis of significantly different diagnosis subpopulations, test values stemming from distorting subpopulations could be excluded from the dataset prior to statistical inference of reference intervals. Acting on these insights led us to develop the concept of the Minus Distribution.

Acquiring a Minus Distribution is initiated by stratifying the total dataset by one's parameters of interest, being biological sex and an age range e.g., male & aged 30-40 years. An age range of 10 years has been selected as the standard for this project as to consider both the fact that analyte means change significantly with age, as well as to increase population size and hence statistical power. A subset of the full dataset stratified by specific parameters for biological sex and age is what we call a "Global Distribution". Our methodology consists of generating a Minus Distribution from the Global Distribution which represents a distribution from which all diagnoses classified as distorting have been removed, as in Global Distribution *minus* significantly different diagnoses. For a given Global Distribution, all uniquely occurring diagnoses are registered. For each diagnosis two subsets from the Global Distribution are created, one containing only the values containing the given diagnose, and one containing everything but the given diagnosis. These two distributions are to be

compared with statistical tests with the intent of classifying them as either stemming from the same overarching distribution or being significantly different distributions at a given significance level. If a particular diagnosis is found to be significantly different from the rest of the distribution, said diagnosis is classified as distorting.

Finally, all values stemming from diagnoses classified as distorting are excluded from the Global Distribution which will leave only the Minus Distribution from which reference intervals can be drawn. These calculations are applied to every combination of age, biological sex, statistical test and so on, and present the main computational challenge in the final product further discussed in section 5.7.3.

4.2 Statistical Methods

4.2.1 Students t-test

The t-test has been selected as our main statistical test for our purposes for its simplicity and robustness. Its evaluation metric is based on calculating a t-value from the mean and standard deviation of two populations, in our case being the Global Distribution and the single diagnosis distribution. If the t-value exceeds a critical value which is estimated based on significance level and degrees of freedom the populations are deemed significantly different.[28]

For our purposes the Welch t-test has been chosen as it allows for comparison of groups with large differences in sample sizes, an important requirement as in our case group size differences can become very large.[29] The Welch t-test is an extension to the t-test in that it doesn't assume equal variance and is hence, deemed more robust in cases where sample size and variance are unequal.

One of the most important requirements a statistical test has to fulfil in order to be considered for our purposes is execution time. In order to prepare a library of statistical evaluations for all stratification factors the selected test is to be executed many thousand times. For this reason, together with its wide recognition and easy we deem the Welch t-test an ideal candidate as a core statistical tool for our Shiny application.

4.2.2 Wilcoxon test

The Mann-Whitney U test[30], also known as Wilcoxon signed-rank test is a non-parametric test statistic used to compare two groups in cases where the data distribution is not of Gaussian form. Instead of assuming normality and using parametric measures like mean and standard deviation, the Wilcoxon test relies on the U-statistic for hypothesis testing. Calculating the U-statistic consists of sorting the population values in order and pairing them by their respective rank. The U-statistic is equal to the sum of the paired differences, for which the distribution under the null hypothesis is

known.[30] In situations where assumptions of normality are not satisfied the Wilcoxon test represents a valuable alternative and has been added to our app as a secondary hypothesis testing method.

4.2.3 Multiple testing

In statistics, hypothesis testing practice involves a significance level as to describe the probability that the null hypothesis is rejected when it is in fact true. Hence at a significance level alpha (α) of 0.05 the probability of a statistical test indicating a significant difference between two populations when there is no difference is 5%. In the case of conducting a single statistical test a chance of 5% of obtaining a false result is an acceptable rate and widely applied standard practice. However, when conducting multiple statistical tests simultaneously a significance level of 5% must be interpreted as the entirety of test results containing 5% false positives. In our case this would mean that of e.g. 700 t-tests calculated in parallel 35 results would return positive by pure chance.[31] It is for this confounding effect that multiple testing correction techniques have been established like Holm-Bonferroni[32] correction and the False Discovery Rate.[24]

4.2.4 False Discovery Rate

The False Discovery Rate (FDR) is a concept or statistical method used in hypothesis testing first introduced by John D. Storey and Robert Tibshirani.[33] The FDR serves as a statistical measure to tune the significance cutoff when dealing with multiple hypothesis tests in parallel. The FDR is the expected proportion of falsely rejected null hypotheses divided by the total number of rejected hypotheses.[24]

Building on the concept of FDR, the Storey-Tibshirani procedure consists of obtaining q-values for each hypothesis test. A q-value is very similar to a p-value as they both serve as a significance threshold cut-off measure but a q-value incorporates the control of the FDR.[34] While the p-value assesses the certainty for rejecting the null hypothesis in a given test, the q-value extends on this concept by considering the expected proportion of falsely rejected null hypotheses among all rejected results. This approach is considered a sophisticated way to balance the trade-off between minimizing false positives and identifying significant findings.[34]

The most common way to obtain q-values is by the Storey-Tibshirani procedure, which operates as follows. Firstly, the p-values from the hypothesis tests run in parallel are ranked in ascending order. Using empirical Bayes method an estimate for $\boldsymbol{\pi_0}$ is generated which represents the proportion of true null hypotheses among all hypotheses. Using $\boldsymbol{\pi_0}$, a q-value is calculated for each tested hypothesis using the formula:

$$qvalue = \frac{(\boldsymbol{\pi}_0 \ * pvalue * total\ number\ of\ tests)}{rank}$$

Equation 1: Formula for q-value estimation

Having obtained the q-values, all q-values below a given significance threshold are considered significant. The Bioconductor package *qvalue*[35] has been used in the scope of this project which offers an efficient way to obtain q-values in R by providing a list of p-values.

4.3 Clustering

4.3.1 Introduction

Relying on stratifying the data by ICD-10 code diagnosis, the Minus Distribution Method was expanded on by the formation of diagnosis clusters and treating them as individual populations. Our efforts were aimed at grouping ICD-10 codes with similar behaviour based on their co-occurrence with clustering techniques.

The formation of cluster groups was achieved in two steps. First, a measure for co-occurrence was established in which each diagnosis is assigned a similarity value for every other diagnose describing how commonly they occur together. For this purpose, we chose the natural language processing technique word2vec.[36] In a second step, ICD-10 codes were clustered into groups of diagnoses by either k-means[37] or hierarchical clustering.[38] Having acquired cluster groups, downstream procedure remained the same, but with groups containing multiple diagnoses instead of single diagnoses.

4.3.2 Word2vec

Word2vec describes a natural language processing method developed in 2013 consisting of a 2-layer neural network.[36] The model can be trained with large amounts of text as input. For a text with N unique words the output will consist of a N-dimensional vector space representing each word (or in our case diagnosis) as a single vector. The direction of a N-dimensional vector may be interpreted in the context of comparing it to a second vector and allows for quantification in terms of similarity. In a nutshell, the more two vectors point in the same direction, the more similar their corresponding words.

By representing words as vectors the similarity between two words can be quantified by calculating the distance between two vectors. Common practice holds distance metrics like cosine, Euclidean and Manhattan distance. For our purposes the cosine similarity was chosen which is equal to the normalized dot product between two vectors.[39]

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Equation 2: Cosine Distance Formula. Describes the angular relationship between two vectors A and B

By determining the cosine similarity, a N x N similarity matrix could be obtained, expressing the similarity between every diagnose as a number between -1 and 1.[40] A matrix like this is equivalent to the distance metric information that can be used for clustering techniques discussed in the following sections.

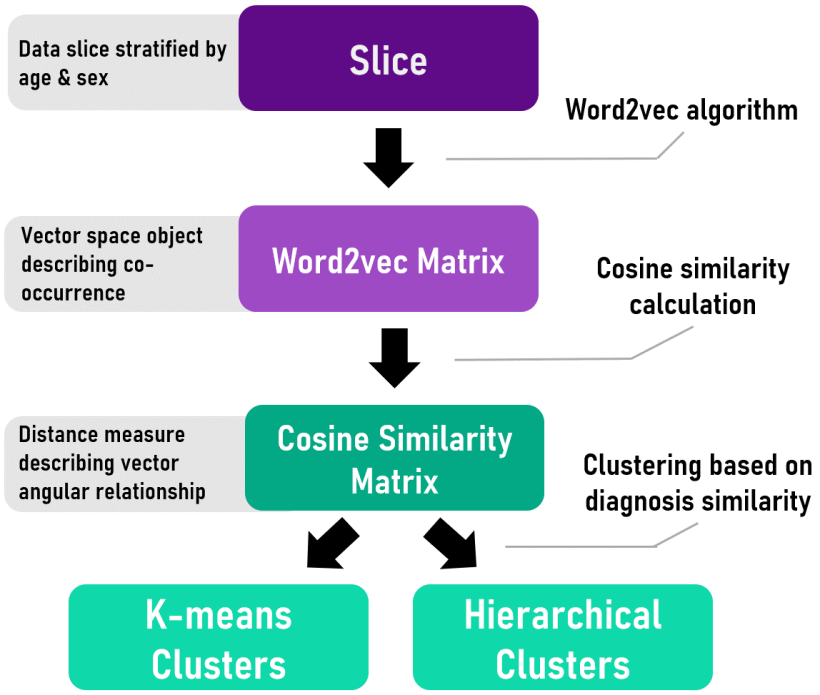

Figure 1: Diagnosis Clustering Workflow. Given N- diagnoses contained in a data slice stratified by age and biological sex a N-dimensional Word2vec vector space object is created with the word2vec algorithm. The angular relationship between each vector is quantified with the cosine similarity measure which once inverted, serves as a distance measure for subsequent clustering techniques.

4.3.3 K-means clustering

K-means is unsupervised machine learning technique to partition data into k number of groups by minimizing intra group variation and maximizing group to group differences. In our case the similarity matrix established with word2vec and cosine similarity in the chapters before serves as the distance metric for the clustering algorithm.

The algorithm starts by randomly generating k centroids and assigning each datapoint to the nearest centroid by minimizing the squared distance between them. For each cluster, the mean of all datapoints assigned to that cluster is set as the new centroid. The datapoints are again assigned to their nearest centroid. These two steps are repeated until the assignments no longer change. The final output are k clusters with the assignments of each datapoint to one of them. [37]

4.3.4 Hierarchical clustering

Hierarchical clustering is an unsupervised algorithm similar to k-means clustering to partition data into clusters with the difference that hierarchical clustering tries to organize all datapoints into a tree-like structure based on their similarity measure. Unlike k-means, hierarchical clustering contains no random number generation and the data may still be partitioned into a specified number of clusters.

The algorithm is initialized by computing the pairwise distance between each datapoint and treating each point like a separate cluster. In the first iteration each datapoint is merged with the closest datapoint, forming a centroid with a new centre. In the following iterations all centroids are again merged with their closest centroid and so on, until all datapoints are connected in a tree with the distances represented as the branch lengths.[38]

4.4 Dataset

The dataset used in this project contains 6 million blood analyte test results from patients of the University Hospital (Inselspital) Bern. The collected data contains entries from 2014 to 2022 and has been anonymized in accordance with the ethical waiver granted by the ethics committee of Bern (KEK). From this follows that each data entry is limited to 5 ICD-10 diagnoses per blood analyte measurement. The data contains a LOINC identifier, a universal standard for reporting medical laboratory observations.[41]

The dataset has been delivered by the Insel Data Science Centre (IDSC) and is part of the Swiss BioRef dataset. The data has been pre-processed and modified by removal of invalid ICD-10 codes (B59, A90, I84, A91) a data point labelled with a negative age and not available analyte test results

prior to analysis. Furthermore, the variables describing age of the patient have been rounded down to the closest whole number.

A single entry in this dataset contains information on blood analyte (Analyte), LOINC laboratory test number (LOINC), test result (Value), test unit (Unit), up to 5 diagnoses in ICD-10 code format (Diag01 - Diag05) and age (Age) and biological sex of the patient (Sex) as depicted in *Table 1*.

Analyte	LOINC	Value	Unit	Diag01	…	Diag05	Age	Sex
Creatinine	14682-9	70	µmol/L	N12		C02	55	m
Glucose	14749-6	5.4	mmol/L	N12		Z95	55	f
Creatinine	14682-9	94	µmol/L	A14		N18	55	f
Potassium	2823-3	3.9	mmol/L	N12		F20	55	f

Table 1: Schematic depiction of the Swiss BioRef dataset with fictional values. Not depicted in this graphic are additional factors: Device udi, Testkit udi, Device type, Testkit type, and RFIKey device.

4.5 Software

4.5.1 R Shiny

R Shiny is an extension to the open-source programming language R.[42] Developed for statistical computation and visualization, R is one of the most widely used languages next to Python among statisticians across many fields.[43] R Shiny is a free package available for both R and Python, adding tools for dynamic web application development. It allows for turning a simple R Script into an application featuring a user interface (UI) with input choices and reactive graphs. Reactivity in this context is achieved through the feature that graphs and calculations may be manipulated based on user input.

When a standard R script is run, computational processes are terminated once finished. In the reactive environment of R Shiny the environment is kept active as long as the application is running. During that time, each variable and object is subject to change based on user input. If the user changes a variable, all calculations involving that variable, along with all subsequent objects depending on it, are executed again. It is important to note that R Shiny doesn't rerun every single piece of code after any change. Only code directly associated with variables or downstream thereof are executed anew, in doing so limiting computational cost and increasing speed.

4.5.2 R Shiny app architecture

An R Shiny application is generally structured in two components, the UI and the server. The UI script contains code specifying the user interface layout, input values and graphs positions. Here, reactive values are defined and coupled to input options to be manipulated by the user. The R Shiny

community maintains an extensive library of input options like sliders, buttons and selectors, all customizable with options for shape, colour and themes.

Each input is labelled with a specified identifier from which the input value can be accessed in the server part of the application. The server is where instructions for input processing as well as the graph plotting or object creation are encoded. Code for reactive values or objects is enclosed in a designated bracket (*object <- ({ Code })*) as to instruct R Shiny to rerun said code in case an input value or upstream dependency changes. Objects created in the server can be created as an output object containing an identifiable label. This enables communication with the UI part of the system for the object to be rendered and displayed in the final form of the application. Hence, the server receives input information from the UI which is sent back after processing to be displayed by the UI.

4.6 App development

The R Shiny app was developed in R version 4.1.2 with RStudio version 2021.09.2. In addition to the base R installation, several packages were used to expand the functionality of R. A list of all R packages with their respective version is provided in *Table 2.*

Package name	Version	Description
dplyr [44]	1.1.0	Library for data manipulation
plotly [45]	4.10.1	Library for interactive, publication-quality graph generation
data.table [46]	1.14.8	Fast aggregation and manipulation of large data
reshape2 [47]	1.4.4	Flexible restructuring and aggregation of data
cowplot [48]	1.1.1	Plot generation, annotation and restructuring
lsa [49]	0.73.3	Natural language processing
word2vec [50]	0.3.4	Vector representations of words
text2vec [51]	0.6.3	Text vectorisation and topic modelling
matrix [52]	1.5-3	Matrix handling and manipulation tools
mclust [53]	6.0.0	Gaussian finite mixture models for clustering and classification,
infotheo [54]	1.2.0.1	Information theory measures based on entropy estimators
pheatmap [55]	1.0.12	Heatmap generation
qvalue[35]	2.26.0	Q-value estimation for false discovery rate control
scrutiny [56]	0.2.4	Summary statistics and error detection techniques
shiny [42]	1.7.4	Interactive web application building tools
shinyWidgets [23]	0.7.6	Additional Rshiny widget library

Table 2: List of all R packages used in the R Shiny app with their respective version and description

4.6.1 The Minus Distribution Method in practice

This section explores how the theoretical concepts were put into practice as an app in the computational sense. Based on our methodology, reference intervals may be obtained by deciding on a given biological sex and age range *a priori* (e.g., female, 40-50). Subsets from the full dataset stratified by age and sex are called a *slice* and can be acquired with the function *get_slice()*. Such a *slice* contains 1 to 5 ICD-10 Codes per lab test entry. The function *get_diags()* returns a list of all unique occurrences of diagnoses in lexicographic order. This list serves as an input to the function *tt_pval()*, using it as a dependency to run an apply function designed to run statistical analysis for every diagnosis once. At this point two more factors have to be decided on, being the statistical test used (t-test/Wilcoxon and the significance level alpha). The final output of *tt_pval()* is a table containing information on statistical testing results called a *p-table*.

A *p-table* contains all the information from statistical testing for every diagnose (Diag) in a given *slice*, being population size (n), population mean (mu), standard deviation (sd) and statistical test output raw p-value (pval) (See *Figure 2*). A *p-table* contains an entry for all diagnoses with a population size larger than 5. We considered the removal of the values stemming from these diagnoses, but as diagnoses are coupled to up to 4 other diagnoses, this could potentially remove valid data points from other diagnoses. Taking this into account together with how rarely a diagnosis population size falls below 5 the decision has been made for them to remain in the dataset.

Diag	n	mu	sd	pval
F20	163	73.84049	14.529632	1.018914e-32
F10	296	72.60473	23.198828	2.096063e-32
N18	181	393.08840	342.699614	6.198955e-25
X59	310	79.80323	15.429451	3.281000e-23
K50	74	73.62162	13.884753	8.035977e-19
S02	236	79.60169	17.282208	1.250499e-18
S01	137	78.33577	14.608256	1.703321e-18
E11	100	71.92000	20.212347	5.437485e-17
Z98	163	77.43558	19.072324	7.885794e-17

Figure 2: Depiction of a p-table object with columns 4diagnosis subset (Diag), population size (n), population mean (mu), standard deviation(sd), statistical test output p-value (pval)

Statistical test output information is stored in the *p-table* format for the sake of information density as well as facilitating multiple testing. In FDR / q-value multiple testing correction depends on the total number of tests conducted, significant or not.[24] This means that storing the entire statistical testing information is necessary. Adding to this is the advantage that storing a raw *p-table* allows for all necessary downstream manipulations and options, making use of the reactivity of the program. In other words, by storing the raw p-values calculation output in its complete form we may use the same *p-table* for different significance levels selected by the user.

After FDR multiple testing correction every diagnosis is labelled either significant or not significant. The function *globalminus()* takes as input a *slice* as well as a labelled *p-table* containing information on which diagnoses were deemed significantly different from the Global Distribution. The function

now excludes every datapoint from the Global Distribution which contains at least one of said diagnoses. The remaining data constitutes an approximation of the underlying non-diseased population, called Minus Distribution. Using the Ichihara Method, the Gaussian parameters and reference interval estimates from the Minus Distribution are calculated and visualized in a histogram (*Figure 3:* Minus RefInt). As to assess how the estimated parameters differ from the global to the Minus Distribution, Ichihara estimates are generated from the Global Distribution as well and added to the histogram for comparison (Global RefInt). The final output from *globalminus()* is depicted in *Figure 3* which shows both distributions, as well as both Ichihara Method reference interval estimates.

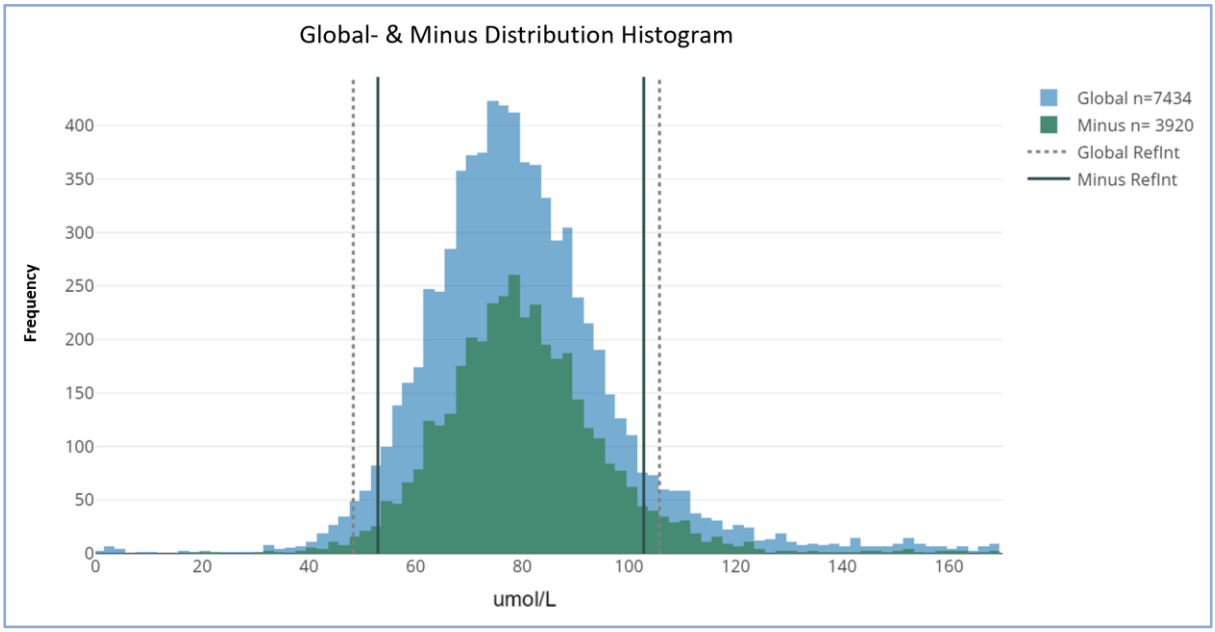

Figure 3: Main shiny app output figure depicting Global and Minus Distribution histograms with their respective Ichihara reference interval estimates (Global RefInt and Minus RefInt respectively) for creatinine values. The data stems from a population slice containing men, ages 30 to 40 years

4.6.2 Core Computational Challenge Outline

The architecture of R Shiny allows for interactive exploration of datasets by recalculating and re-rendering objects every time a new input is passed by the user. When dealing with simple computations reactive changes appear seamless for the user and all background computations may be done on the server part of the app. However, following our current model for reference interval estimation computation time lies between 30-40 seconds per *p-table* calculation. A delay of this magnitude forces us to stick to alternative methods to make the app viable and user friendly. Finally, the app is designed as to concentrate all extensive calculations into an initiation step when first setting up the app. In the initiation step, calculations for all possible input option combinations are executed and stored in a library accessible to the app. In doing so again reaching seamless reactivity once set up.

Running statistical tests (t-test/Wilcox-test) for every diagnose remains the most computationally expensive process in our method averaging at 800 diagnoses per *slice*. By pre-calculating the results of these tests and storing them locally we circumvent the issue of running them in a reactive environment every time the input changes. The information is stored as a *p-table* in a *.txt* file, containing population size, mean, standard deviation and p-value for every diagnose of a given *slice*. One such *p-table* contains the necessary information to calculate reference intervals of one specific factor combination of age, biological sex, statistical test and clustering. Hence, one *p-table* file needs to be generated for every possible combination of these factors for the app to function as intended. Ultimately, a complete p-value library is generated by the function *createDB()* and stored in a separate folder for each blood analyte relevant for a given session. Downstream analysis like multiple testing correction as well as sub-setting and graph rendering require little computation time and are handled by the app server interactively.

4.6.3 App file structure

The final product can be downloaded as a single folder of 45 kB with structure and files as depicted in *Figure* 4. The *R Scripts* folder contains all functions necessary to preprocess the data and run the app.

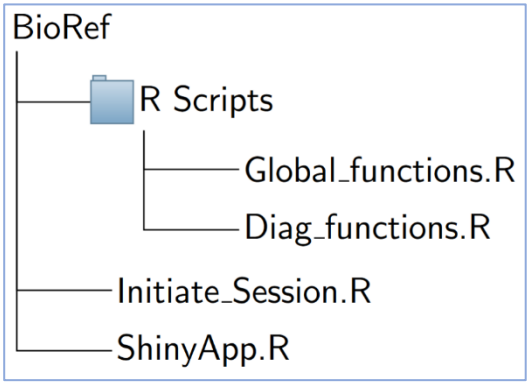

Figure 4: Folder architecture of the downloadable R Shiny app

Initiate_Session.R is responsible for sub-setting the full dataset by lab tests and converting them to *.rds* files. Segmenting the data allows for faster read-in and saves memory space. Separate folders for each test analyte are created, each of which contains a separate *p-table* library. *ShinyApp.R* contains all code relevant to the app itself and executing it will open an interactive session with a UI.

4.6.4 Initiating a session

Prior to using the app, a *p-table* library needs to be created using the data provided by the user. This is initiated by copying the dataset into the BioRef folder in .csv format and sourcing the file *Initiate_Session.R*. Said script will run the function *createDB()* for every lab test specified which fulfils three major tasks.

The function *createDB()* will create the folder *Labtests_subsets* and a subdirectory for each analyte (See *Figure 5)*. It will filter the full dataset by analyte and save said subset as an .rds in its respective folder. Two more folders will be created next to the data subset. *BigSliceClusters* will store the information on cluster groups which are calculated based on subset of the dataset containing only

ages from 20 to 80. The clustering methods chosen are hierarchical and k-means clustering, creating cluster groups with size 400 and 800 for both methods.

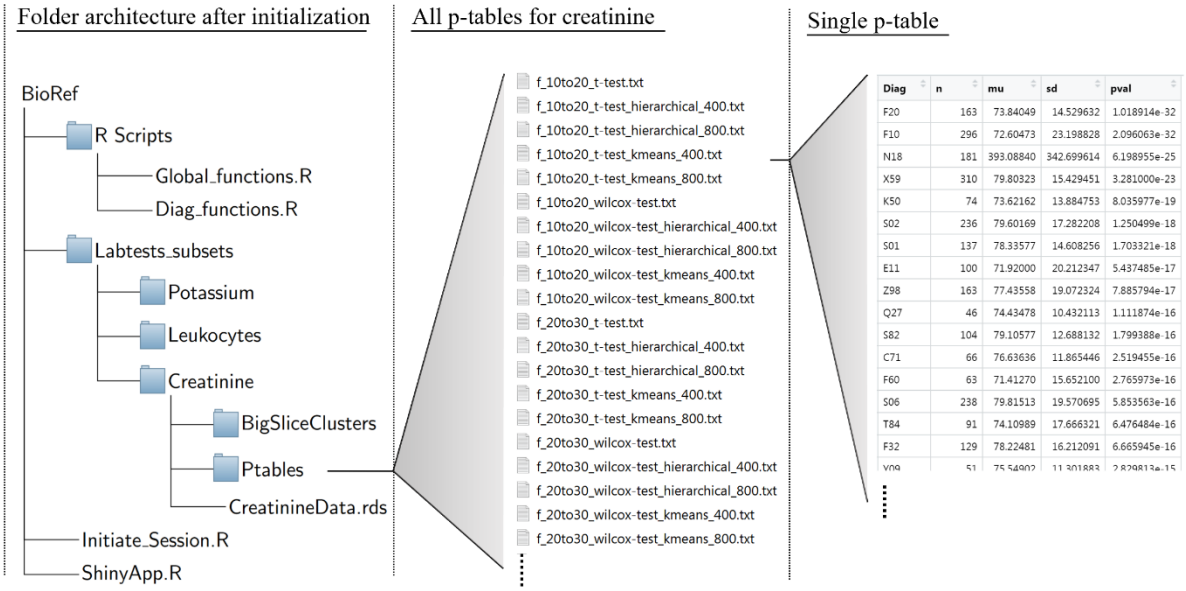

Figure 5: Schematic depiction of the BioRef folder architecture after running the initiation step. Initiate_Session.R creates a Folder in Labtests_subsets for each test analyte of interest. Inside each Folder a subset with the relevant data in .rds format is stored as well as a complete p-table library. Folder Architecture (Left). Single blood analyte p-table library (Middle), single p-table (Right)

The *Ptables* folder will contain all *p-table*s for a given lab test constituting the essential information for reference interval estimation in the app. The main UI allows for several factors to be specified, some of which contain multiple options for selection. Each individual combination of these factors requires a separate *p-table* file, amounting to 96 files per lab test. The factors that are subject to be manipulated by the user are listed in *Table 3*.

Factor:	Biological Sex	Age	Statistical test	Clustering method	Number of cluster groups
	Female	20-30	T-test	Hierarchical	400
	Male	30-40	Wilcoxon test	K-means	800
		40-50			
		50-60			
		60-70			
		70-80			

Table 3: List of all factors and their respective settings.

As to create 96 *p-table* files, *createDB()* depends on a parameter matrix containing all possible combinations of the input factors as rows. Each row is then fed into an apply function running the

statistical analysis and saving it as a file for each one. The name of each file is generated by pasting all input parameters together enabling the correct file to be read by the app.

One may note that the subset *.rds* files created for each analyte serve as the data source for the app in order to save memory and allow for faster reading of the data. The full dataset provided by the user may be the deleted after the initiation step is completed.

5 Results

5.1 Swiss BioRef App Main UI

When deciding on the main UI layout efforts were aimed at having a visual representation of the reference interval estimation in a central position at all times (See *Figure 6)*. Parameter selection is facilitated by keeping all input option in a sidebar window on the left. Advanced options for inspection and statistical analysis summaries may be found in the tabs below.

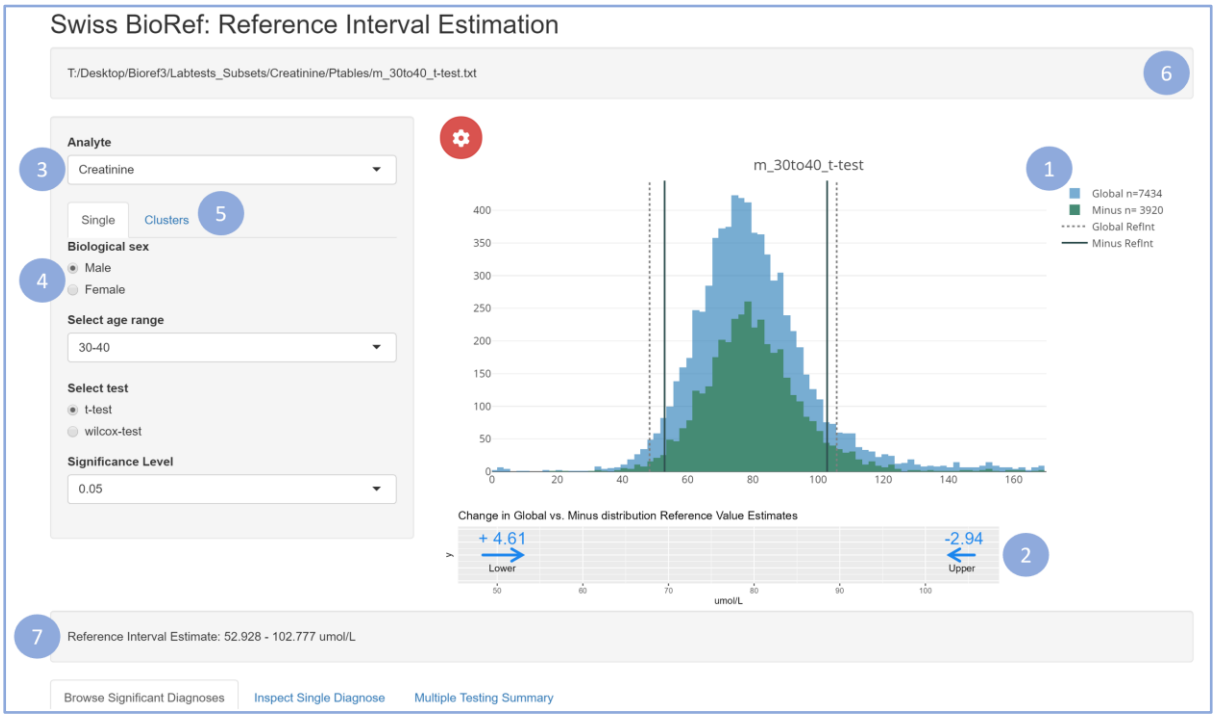

Figure 6: Main UI of the Swiss BioRef R Shiny App. Input options for reference interval estimation may be selected by the user in the window on the left. The Minus Distribution Method results are summarized visually in the main plot. Features: Histogram and Reference intervals (1-2), analyte and parameter selection (3-5), path to p-table file (6) and final reference interval values (7)

Main component of the UI may be found in window 1 and summarizes the Minus Distribution acquisition in one plot. Here, the Global Distribution (Global) stratified by selected inputs on laboratory test, biological sex and age is displayed as a histogram. In the same plot the Minus Distribution (Minus) is plotted in direct comparison to the Global Distribution. For both distributions a 95% confidence interval Ichihara estimate is shown as vertical lines (straight line for Global and dotted for Minus Distributions). This representation allows for direct comparison in terms of location and group sizes of both distributions and may facilitate inspection of adequate approximation of the non-diseased population inside the global population.

Right below the main plot window 2 may be found which is dedicated to summarizing the change in reference interval estimates from the Global to the Minus Distribution using the Ichihara Method. The colour of the arrows will be blue if the arrow points in the direction of the mean and red otherwise. This is to highlight if the estimate of the Minus Distribution is narrower than the Global one, being a measure for a better estimate. In a nutshell, the arrows represent the distance from the vertical dotted line to the straight line in plot 1 for the lower and upper reference limits.

Window 3 allows to select the test analyte of interest and window 4 features all possible parameter selections for said analyte. Possible selections are summarized in *Table 3*. Switching the selection of 5 enables the analysis to be run with or without clustering. Analysis without clustering is selected by default, selecting the clustering tab will expand the parameter selection window 4 with options for cluster technique and cluster size. As to monitor whether the correct *p-table* file has been selected the UI features a window 6 showing the exact path to the file. The file name itself consists of the parameters selected in the UI. Finally, the calculated reference interval values are displayed in window 7 as to summarize the estimation.

While the main UI features input options and the main plot, the app features three additional tabs containing supplementary information and allows for browsing of particular diagnoses.

5.2 Tab I: Browse Significant Diagnoses

The tab *Browse Significant Diagnoses* depicted in *Figure 7* serves as a complete summary of the statistical analysis conducted for a given *slice*. Window 1 informs about the number of diagnoses occurring in the *slice* set in the sidebar window above, as well as how many of which have been classified as significantly different from the Global Distribution. Window 2 essentially lists all entries contained in a given *p-table* file. For every diagnose the number of entries per *slice* are counted (n), the mean (mu), the standard deviation (sd), as well as the p-value from the statistical test are found here.

This tab allows the user to gain a quick and insightful overview of the diagnoses excluded from the Global Distribution to form the Minus Distribution. The table may be filtered by each variable independently to search for diagnoses with highest deviation from the global mean or highest significance.

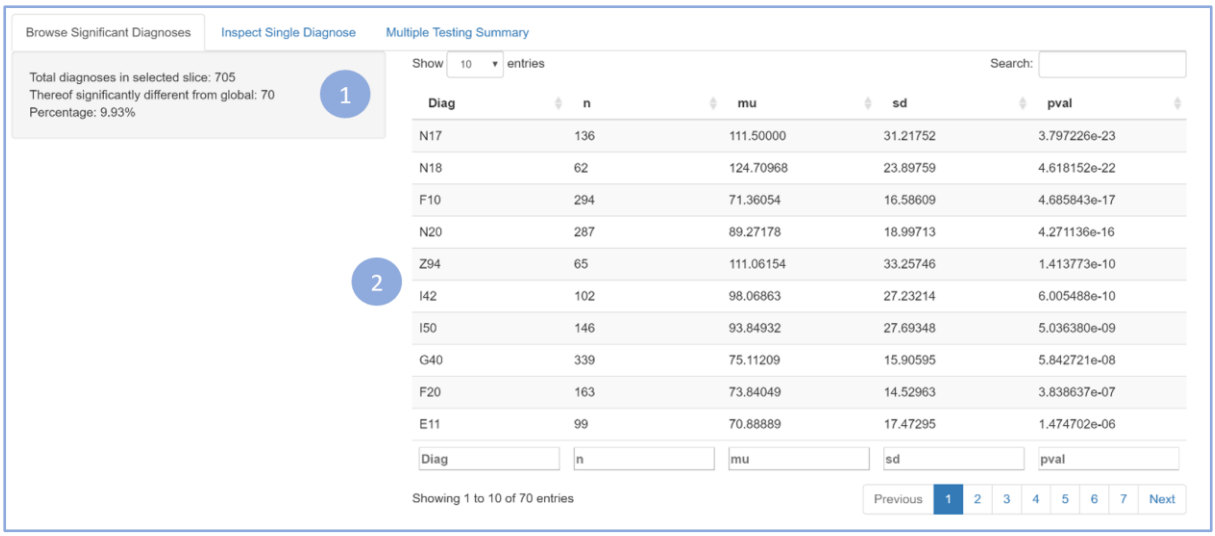

Figure 7: "Browse Significant Diagnoses" Supplementary Tab. Here each diagnose from a given slice can be inspected, containing information on all significantly different diagnoses. The table features the option to sort by all columns, offering the user to browse for information like the highest deviation from the Global Distribution and test significance.

5.3 Tab II: Inspect Single Diagnose

The second tab shown in *Figure* 8 serves the purpose of inspecting a single diagnose and its distribution. In window 1 a diagnose occurring in the selected *slice* may be chosen to be put into relation with the Global Distribution. The density distribution of said diagnose is plotted next to a histogram of the Global Distribution in window 2. The vast population size difference of our testing groups makes it challenging to visualize both groups in a single plot. Therefore, a density plot has been chosen to illustrate the position of the single diagnose distribution relative to the Global Distribution. As to keep density smoothing bias at a minimum, the density bandwidth may be adjusted in window 1.

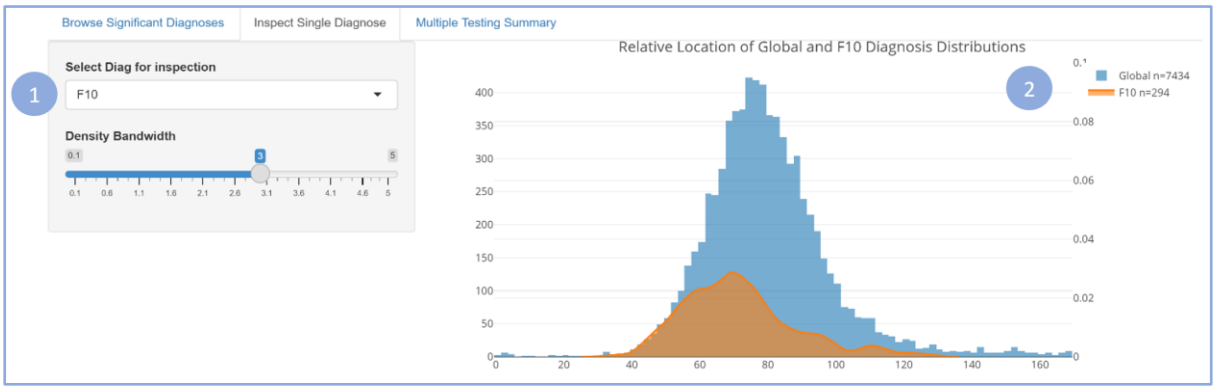

Figure 8: "Inspect Single Diagnosis" Second Supplementary Tab. Allowing for comparison of a particular diagnosis in relation to the Global Distribution. Depicted is the diagnosis F10 (Alcohol dependency), showing a notably lower distribution mean than the Global Distribution.

5.4 Tab III: Multiple Testing Summary

This tab depicted in *Figure 9* serves to inspect if multiple testing correction using FDR has been carried out up to standards. Using the Bioconductor package *qvalue*, the plots 2 - 5 are generated when plotting a *qvalue* object and represent an overview of the main steps of the procedure. Plot 2 serves to assess π_0 estimate reliability in which π_0 is plotted against a tuning parameter λ. With an increasing λ the bias of the π_0 estimate decreases at the cost of higher variance.[24] Plots 3 - 5 summarize the behaviour of significant tests and expected false positives in the context of using different q-value cut-offs. These features may be inspected even more explicitly when switching to the rejection overview tab in window 1. Here, a table is shown containing the expected number of significant findings depending on different significance cut-offs.

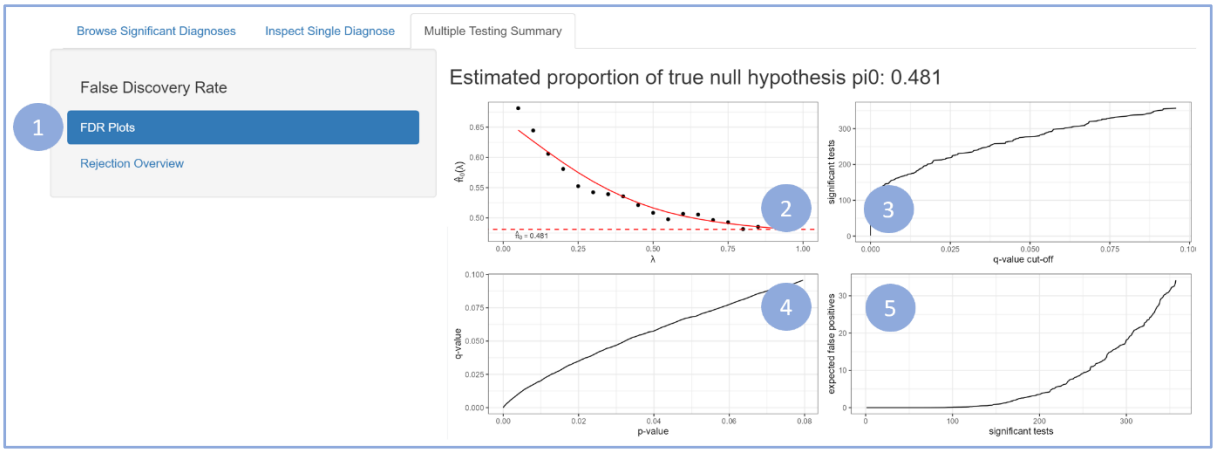

Figure 9: "Multiple Testing Summary" Third Supplementary Tab. This tab provides information on the FDR correction like π_0 estimation and q-value to p-value relationship, offering an overview on the multiple testing correction.

5.5 Reference Interval Inference

In the example of creatinine data, the Global Distribution consists of a heavily right-skewed normal distribution with a long tail of extreme values. Testing for significantly different diagnoses, we identified many diagnoses corresponding to liver deficiencies in the ICD-10 coding system. All these subpopulations were highly elevated in comparison to the global mean and greatly contributed to the formation of the distribution tail in the Global Distribution. Removing these values resulted in the Global Distribution taking on a more Gaussian form, suggesting that the information contained in the ICD-10 code labelling could be used to curate the Global Distribution before reference interval estimation.

Reference interval estimates generated with the Minus Distribution Method are summarized in *Table 4* on the example of *Creatinine (in µmol/L) extracted from blood serum or plasma* (LOINC: 14682-9). Parameters have been set to use the t-test with a significance level α = 0.05. Shown are 3

sets of reference intervals for age groups of 10 years from 10 to 90 with their respective standard error, as well as the population size they are drawn from. All reference intervals are drawn from their respective distributions with the Ichihara Method, being the unmanipulated Global Distribution (Left), the Minus Distribution without clustering (Middle) and the Minus Distribution generated using hierarchical clustering with 800 clusters (Right).

One may note how Minus Distribution reference intervals are generally narrower compared to ones generated from the Global Distribution. However, at the same time they show larger standard errors which may be attributed to the large reduction in sample size, particularly in older patients. Significant changes in reference intervals are often more prevalent in the 9.75th limit, which can be attributed to the reduction of the large distribution tails and extreme outliers on the right-hand side of the distributions.

	Global Distribution			Minus Distribution			Minus Distribution with Clustering		
Age*	n§	2.5th†	9.75th‡	n§	2.5th†	9.75th‡	n§	2.5th†	9.75th‡
Female									
20-30	8858	29 (28.2-29.1)	82 (81.6-82.5)	984	38 (37.2-76.9)	77 (75.9-77.9)	837	38 (38.1-39.2)	77 (75.8-77.9)
30-40	13505	29 (29-29.7)	82 (81.2-81.9)	1087	38 (36.6-80.3)	80 (79.2-81.3)	888	37 (36.9-38.1)	83 (81.6-84)
40-50	10379	36 (35.3-36.1)	86 (86.1-86.9)	4821	40 (39.7-85.6)	86 (85.1-86.1)	4164	40 (40.3-40.9)	86 (85.1-86.2)
50-60	14862	37 (37.1-37.8)	88 (87.2-87.8)	4465	40 (39.3-87.2)	87 (86.6-87.7)	3349	40 (40.1-40.7)	87 (86.3-87.6)
60-70	18037	36 (35.4-36.1)	93 (92.9-93.7)	2311	41 (40-90.6)	91 (89.7-91.4)	1481	40 (40.4-41.5)	90 (89.5-91.6)
70-80	22113	36 (35.2-35.9)	100 (100-100.7)	860	37 (35.5-97.9)	98 (96.2-99.5)	657	40 (40.5-42.3)	96 (94.7-98.3)
80-90	17240	35 (34.4-35.4)	112 (111.8-112.8)	564	41 (38.9-102.9)	103 (100.7-105.1)	313	39 (39.3-42.3)	103 (100.3-106.3)
Male									
20-30	5827	50 (49.9-51)	103 (102.3-103.4)	3087	55 (54.3-55.7)	101 (100.3-101.6)	2638	55 (54.5-55.9)	101 (100.3-101.7)
30-40	7434	49 (48.4-49.4)	105 (104.6-105.7)	3920	54 (52.9-54.2)	103 (102.8-104.1)	3354	52 (51.8-53.2)	104 (103.8-105.2)
40-50	11854	49 (48.3-49.2)	105 (105-105.9)	4260	51 (50.5-51.8)	106 (105.2-106.6)	3489	52 (51.5-52.9)	105 (104.2-105.6)
50-60	23049	46 (45.4-46.1)	110 (110-110.7)	3216	50 (48.7-50.3)	109 (107.7-109.4)	2402	51 (49.8-51.6)	107 (106.4-108.2)
60-70	31055	45 (44.6-45.3)	115 (114.8-115.4)	1636	50 (48.8-51.2)	112 (110.8-113.2)	711	53 (51.4-54.7)	108 (106.5-109.8)
70-80	31754	47 (46.4-47.1)	123 (122.8-123.5)	719	54 (51.9-55.5)	111 (109.6-113.2)	708	51 (49-53.1)	119 (116.8-120.9)
80-90	15743	48 (46.9-48.1)	137 (136.2-137.4)	673	49 (46.7-51.4)	123 (120.3-125)	641	48 (46.1-51)	124 (121.4-126.3)

General: All Reference Interval units in umol/L

* Age in years † 2.5th quantile with standard error ‡ 9.75th quantile with standard error § Population Size

Table 4: Creatinine reference interval estimates. The table compares estimates drawn from the Global Distribution (Left), the Minus Distribution (Middle) and Minus Distribution generated using hierarchical clustering with 800 clusters (Right).

6 Discussion

6.1 Implementation and execution

The core axiom of the Minus Distribution Method consists of using ICD-10 code labelled data to approximate an underlying non-diseased population from mixed data for reference interval estimation. In the context of implementing our methodology in an app to be used for diagnosis of patients, we deem our method to be a success. Once set up, our app provides quickly accessible personalized reference intervals.

Regarding the app, we are confident to have developed an accessible application able to communicate reference interval estimation in a concise manner. The app is set up in a matter of 2 clicks by providing the dataset and sourcing an R script. Once set up, the interactivity of R Shiny allows for quick browsing for relevant reference interval estimates as well as exploring the behaviour and differences of different populations. The way the app is designed allows for easy changes to the dataset whenever more recent data is available.

As for the method itself, the Minus Distribution Method is able to generate reference intervals estimates for all ages and biological sexes. By excluding significantly different subpopulations from the dataset, the remaining population shows clear reduction of long distribution tails in the example of creatinine extracted from blood serum or plasma (LOINC: 14682-9). As the Minus Distribution takes on average a more Gaussian form, we are confident to obtain improved estimates for reference intervals moving forward.

Evaluation metrics for improvement in this context may consist of narrower estimates for reference intervals as well as smaller standard errors.[13] Judging on behalf of these two metrics we deem our estimates superior over generating them directly from the Global Distribution. The removal of outliers and reduction of distribution tails through the Minus Distribution Method results in narrower intervals in almost all cases. However, the Minus Distribution Method greatly reduces population size, sometimes by a factor of 10. Particularly populations aged over 60 years are substantially reduced by the Minus Distribution Method due to *a priori* larger standard deviations on both Global Distribution and single diagnosis subset level, longer distribution tails and more extreme outliers.[4,57] Narrower reference intervals could in this case be attributed to reducing the variance by working with smaller population sizes. Further research is needed to assess how much the narrower estimates may be regarded as an accomplishment of the Minus Distribution Method.

Judging by the evaluation metric of small standard errors for the reference limits, the values generated by the Minus Distribution Method show much larger standard errors compared to the

ones generated directly from the Global Distribution. This decline may be attributed to the large reduction in population size as well. The standard error has an inversely proportional relationship to the sample size which may serve as an explanation to this effect.[58] Smaller population sizes leading to larger standard errors of reference interval estimates has been reported before as in George G. Klee et al.[59]

Both evaluation metrics being narrow reference intervals and small reference limit standard errors are negatively influenced by small sample size, which can be significantly reduced by the application of the Minus Distribution Method. Adjusting the significance level alpha to a stricter level, such as $\alpha = 0.01$, could help improve both metrics by increasing sample size in the Minus Distribution. However, it is important to note that sample size reduction is disproportionate with age, leading to much greater reductions in population sizes in older age groups. This follows that Minus Distribution sample size may need to be controlled for age groups individually. Further research needs to be conducted in order to optimize reference interval estimation with the Minus Distribution Method.

The Minus Distribution Method is based on stratifying the data by ICD-10 diagnosis where each diagnosis distribution is compared to the full dataset. Handling ICD-10 code labelled data, it soon became evident that rather than occurring independently of each other some diagnoses seemed to frequently occur in pairs or groups. This was to be expected as clinicians may diagnose patients based on observable symptoms rather than the underlying cause. As to give an example, all patients with a cold would have been diagnosed by clinicians with a similar set of ICD-10 codes describing their symptoms like a runny nose, sneezing and headaches. In theory, these groups should be quantifiable by introducing a measure for co-occurrence between diagnoses. Investigating these effects was deemed relevant for our work for the reason that clusters may serve as a more accurate model of disease than individual symptoms. Treating them as such may improve our method, potentially leading to more accurate reference intervals.

Put into practice, no significant improvement of reference intervals could be observed using clustering as notable in *Table 5*. In fact, working with groups of diagnosis seemed to produce very insignificant changes in estimated reference interval values altogether. Variance in reference intervals introduced by changing age and sex exceeded clustering variance by far, even when comparing hierarchical to k-means clustering. We take from this that clustering diagnoses doesn't appear to improve reference intervals estimated by the Minus Distribution Method. However, working with diagnose groups revealed interesting inter-diagnosis connections and may hold great potential for future research.

When comparing our estimates with recent literature, the female creatinine reference intervals estimated with the Minus Distribution method without clustering were generally shifted a few units lower compared to the reference intervals estimated by Martinez-Sanchez et al., 2022.[60] Similarly for males, our reference intervals were slightly shifted downwards while being narrower across all ages. When comparing the estimates from Arzideh et al., 2010 [61] their estimates for males aged 18-49 were slightly shifted upwards. The reference intervals for 50-65 agree well with each other though our intervals are slightly narrower. As for values for females, ages 18-49 agree with each other as well, though our estimates for ages 50-65 are slightly higher. Overall, our method produced reference values similar to results from recent literature and were found to be narrower in some cases. Additionally, our intervals were often found to be slightly lower compared to other results which may be attributed to the reduction in right-hand distribution tails introduced by the Minus distribution Method, often found in creatinine data distributions.

6.2 Limitations

A core difficulty of reference interval estimation is the extrapolation of a Gaussian element from non-Gaussian distributions. This difficulty is met again in the Minus Distribution Method when conducting parametric statistical tests on the Global Distribution and its subsets in form of diagnosis distributions. Introducing the Wilcox-test as a non-parametric test bypasses this problem but may still be affected by large distribution tails and extreme outliers. Adding to this is the problem of finding the optimal parameters for the statistical analysis like significance level and cluster options. The perfect settings as well as how to assess them from the results is at this point uncertain. At the heart of the indirect approach in reference interval estimation lies the unresolved distinction between healthy and diseased. The imprecise definition of these terms results in a lack of metrics to test novel approaches like ours in terms of performance and precision. Our method requires the data to be labelled with information on ICD-10 codes, which only few labs have access to. In light of these limitations and the scarcity of similar studies, finetuning parameters and interpreting the results has proven to be a somewhat ambiguous task.

Judging by reference interval estimate consistency and continuity we consider our method to work best with creatinine extracted from blood serum or plasma data. Having been developed mostly using data from said analyte this comes to no surprise, but highlights the question if the method is applicable to any analyte the same way. Exploratory data analysis yielded the observation that different analytes may follow entirely different distributions. This implies that the method may have to be adjusted and optimized for each analyte of interest individually in order to reach its full potential.

7 Conclusion

In conclusion, this project introduces a novel method for reference interval inference using ICD-10 code labelled data to facilitate test result interpretation and improve accuracy in patient diagnosing. We are confident to have demonstrated the potential of using an app to generate personalized reference intervals from clinical data with the Minus Distribution Method. With an intuitive user interface and its ability to function reactively the app represents a widely applicable tool, practicable for everyday use. Adding to this is the implementation of the Minus Distribution Method which highlights the benefits of using ICD-10 code labelled data. Stratifying data based on diagnoses has proven to yield novel insights and much potential for future research.

Creating clusters from ICD-10 codes highlights the benefits of clinical diagnoses as an additional stratification factor. Ascertaining similar behaviour of diagnoses on the basis of co-occurrence may enhance reference interval inference by identifying distorting sub-populations. While showing promise, further work is necessary to fully understand the implications of clustering ICD-10 codes.

The implementation of an interactive app for reference interval estimation demonstrates much potential, however future research is necessary to determine the degree to which the results may be called superior to results from traditional techniques. While confident that the methods used in this work yield enhanced, narrower estimates, the extent to which the Minus Distribution Method contributes to this effect remains to be ascertained.

In conclusion, this master thesis presents a novel approach to approximate a “healthy” population from mixed data. Implemented in form of an interactive app, this project demonstrates the advantage of using big data and sophisticated statistical methods for reference interval estimation and ultimately patient care.

8 Outlook and future directions

This project was developed as part of the “Swiss BioRef” project. Swiss BioRef is a research project aimed at establishing a data infrastructure where data from multiple Swiss hospitals are combined to form a standardized data pool in order to generate better reference intervals. In this context, the work presented in this thesis represents the last element of the BioRef workflow, being reference interval estimation and communication of the results.

The ultimate vision for this project is an intuitive app for clinicians for everyday use which may provide personalized reference intervals for a given patient. Following this vision, the Minus Distribution Method may be refined and simplified in ways that allow users unfamiliar with the concepts to grasp its inner workings and interpret the results. Furthermore, the method may be expanded on and tweaked to guarantee its applicability to all blood analytes of the Swiss BioRef dataset. Future work could consist of developing a method for validating the improvements of our reference interval estimates in relation to traditional methods.

Once implemented Swiss BioRef could be a powerful medical tool allowing for more accurate diagnosis of patients and perhaps earlier detection of disease. Quantifying medical data is a difficult endeavour, but ultimately worth the effort if through them patient care can be improved in the future.

9 Acknowledgements

I would like to thank Alexander Leichtle for giving me the opportunity to work on this exciting project. As a PI he created a very pleasant work environment and never missed an opportunity to lighten the mood with joking remarks about researcher's struggles and life itself.

My special thanks to Tobias Blatter as my direct supervisor and mentor. In moments of unclarity or impostor syndrome a meeting with Tobias got me on the right track again. His elaborate visions of our projects but also his ability to eloquently communicate them always did the trick. Right after I found myself again full of motivation and energy to master the challenges at hand.

I want to thank the rest of the Leichtle lab being Harald Witte and Priyanka Nagabhushana for giving me feedback on my work in all those lab meetings on Wednesday morning. And Christos Nakas I guess.

And how could I forget James Ackermann which reminded me every Thursday that there is a life outside of the home office.

For purposes of readability and spell-checking ChatGPT has been used to assist in the generation of some formulations. ChatGPT has not been used to generate any content.

10 References

1. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. *Stat Methods Med Res*. 1999;8(2):93-112. doi:10.1177/096228029900800202

2. Ozarda Y. Reference intervals: current status, recent developments and future considerations. *Biochem Medica*. 2016;26(1):5-11. doi:10.11613/BM.2016.001

3. McPherson K, Healy MJ, Flynn FV, Piper KA, Garcia-Webb P. The effect of age, sex and other factors on blood chemistry in health. *Clin Chim Acta Int J Clin Chem*. 1978;84(3):373-397. doi:10.1016/0009-8981(78)90254-1

4. Weaving G, Batstone GF, Jones RG. Age and sex variation in serum albumin concentration: an observational study. *Ann Clin Biochem*. 2016;53(1):106-111. doi:10.1177/0004563215593561

5. EP28A3C: Define and Verify Reference Intervals in Lab. Clinical & Laboratory Standards Institute. Accessed September 26, 2023. https://clsi.org/standards/products/method-evaluation/documents/ep28/

6. Zardo L, Secchiero S, Sciacovelli L, Bonvicini P, Plebani M. Reference Intervals: Are Interlaboratory Differences Appropriate? 1999;37(11-12):1131-1133. doi:10.1515/CCLM.1999.165

7. Haeckel R, Wosniok W, Streichert T, Dgkl M of the SGL of the. Review of potentials and limitations of indirect approaches for estimating reference limits/intervals of quantitative procedures in laboratory medicine. *J Lab Med*. 2021;45(2):35-53. doi:10.1515/labmed-2020-0131

8. Elveback LR, Guillier CL, Keating FR. Health, normality, and the ghost of Gauss. *JAMA*. 1970;211(1):69-75.

9. Sakia RM. The Box-Cox Transformation Technique: A Review. *J R Stat Soc Ser Stat*. 1992;41(2):169-178. doi:10.2307/2348250

10. Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. *Stat Med*. 2000;19(21):2943-2962. doi:10.1002/1097-0258(20001115)19:21<2943::AID-SIM559>3.0.CO;2-5

11. Harrell FE, Davis CE. A New Distribution-Free Quantile Estimator. *Biometrika*. 1982;69(3):635-640. doi:10.2307/2335999

12. Coskun A, Ceyhan E, Inal TC, Serteser M, Unsal I. The comparison of parametric and nonparametric bootstrap methods for reference interval computation in small sample size groups. *Accreditation Qual Assur*. 2013;18(1):51-60. doi:10.1007/s00769-012-0948-5

13. Daly CH, Higgins V, Adeli K, Grey VL, Hamid JS. Reference interval estimation: Methodological comparison using extensive simulations and empirical data. *Clin Biochem*. 2017;50(18):1145-1158. doi:10.1016/j.clinbiochem.2017.07.005

14. Jones GRD, Haeckel R, Loh TP, et al. Indirect methods for reference interval determination – review and recommendations. *Clin Chem Lab Med CCLM*. 2019;57(1):20-29. doi:10.1515/cclm-2018-0073

15. Solberg HE, Stamm D. IFCC recommendation: The theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values. *J Autom Chem*. 1991;13(5):231-234. doi:10.1155/S146392469100038X

16. Zierk J, Arzideh F, Kapsner LA, Prokosch HU, Metzler M, Rauh M. Reference Interval Estimation from Mixed Distributions using Truncation Points and the Kolmogorov-Smirnov Distance (kosmic). *Sci Rep*. 2020;10(1):1704. doi:10.1038/s41598-020-58749-2

17. Pottel H, Vrydags N, Mahieu B, Vandewynckele E, Croes K, Martens F. Establishing age/sex related serum creatinine reference intervals from hospital laboratory data based on different statistical methods. *Clin Chim Acta*. 2008;396(1):49-55. doi:10.1016/j.cca.2008.06.017

18. Cook MG, Levell MJ, Payne RB. A method for deriving normal ranges from laboratory specimens applied to uric acid in males. *J Clin Pathol*. 1970;23(9):778-780. doi:10.1136/jcp.23.9.778

19. Clinical Chemistry (Chemical Analysis: A Series of Mono…. Accessed September 25, 2023. https://www.goodreads.com/book/show/9687140-clinical-chemistry

20. Creatinine Blood Test: Normal, Low, High Levels, Causes & Symptoms. MedicineNet. Accessed September 25, 2023. https://www.medicinenet.com/creatinine_blood_test/article.htm

21. Pierre CC, Marzinke MA, Ahmed SB, et al. AACC/NKF Guidance Document on Improving Equity in Chronic Kidney Disease Care. *J Appl Lab Med*. 2023;8(4):789-816. doi:10.1093/jalm/jfad022

22. An iterative method for improved estimation of the mean of peer-group distributions in proficiency testing. Accessed August 16, 2023. https://www.degruyter.com/document/doi/10.1515/CCLM.2005.074/html

23. shinyWidgets package - RDocumentation. Accessed August 17, 2023. https://www.rdocumentation.org/packages/shinyWidgets/versions/0.7.6

24. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat*. 2000;25(1):60-83. doi:10.2307/1165312

25. ICD- 10 - CM International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Published June 29, 2023. Accessed August 16, 2023. https://www.cdc.gov/nchs/icd/icd-10-cm.htm

26. BfArM - ICD-10-GM. Accessed September 19, 2023. https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html

27. WHO | International Classification of Diseases (ICD) Information Sheet. Published November 4, 2012. Accessed August 16, 2023. https://web.archive.org/web/20121104064222/http://www.who.int/classifications/icd/factsheet/en/

28. Student. The Probable Error of a Mean. *Biometrika*. 1908;6(1):1-25. doi:10.2307/2331554

29. Ahad NA, Yahaya SSS. Sensitivity analysis of Welch's t-test. *AIP Conf Proc*. 2014;1605(1):888-893. doi:10.1063/1.4887707

30. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biom Bull*. 1945;1(6):80-83. doi:10.2307/3001968

31. Romano JP, Shaikh AM, Wolf M. Multiple Testing. In: *The New Palgrave Dictionary of Economics*. Palgrave Macmillan UK; 2016:1-5. doi:10.1057/978-1-349-95121-5_2914-1

32. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat*. 1979;6(2):65-70.

33. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci*. 2003;100(16):9440-9445. doi:10.1073/pnas.1530509100

34. Lai Y. A statistical method for the conservative adjustment of false discovery rate (q-value). *BMC Bioinformatics*. 2017;18(3):69. doi:10.1186/s12859-017-1474-6

35. Storey JD, Bass AJ, Dabney A, Robinson D, Warnes G. qvalue: Q-value estimation for false discovery rate control. Published online 2023. doi:10.18129/B9.bioc.qvalue

36. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. Published online September 6, 2013. doi:10.48550/arXiv.1301.3781

37. MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Vol 5.1. University of California Press; 1967:281-298. Accessed August 16, 2023. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

38. Nielsen F. Hierarchical Clustering. In: ; 2016:195-211. doi:10.1007/978-3-319-21903-5_8

39. Singhal A. Modern Information Retrieval: A Brief Overview. http://singhal.info/ieee2001.pdf

40. IEEE Technical Committee on Data Engineering. Accessed August 23, 2023. http://sites.computer.org/debull/bull_issues.html

41. Home. LOINC. Accessed September 19, 2023. https://loinc.org/

42. Chang W, Cheng J, Allaire JJ, et al. shiny: Web Application Framework for R. Published online August 12, 2023. Accessed August 17, 2023. https://cran.r-project.org/web/packages/shiny/index.html

43. R: The R Project for Statistical Computing. Accessed August 18, 2023. https://www.r-project.org/

44. dplyr package - RDocumentation. Accessed August 17, 2023. https://www.rdocumentation.org/packages/dplyr/versions/1.0.10

45. Plotly. Accessed August 17, 2023. https://plotly.com/r/

46. Dowle M, Srinivasan A, Gorecki J, et al. data.table: Extension of “data.frame.” Published online February 17, 2023. Accessed August 17, 2023. https://cran.r-project.org/web/packages/data.table/index.html

47. Wickham H. reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. Published online April 9, 2020. Accessed August 17, 2023. https://cran.r-project.org/web/packages/reshape2/index.html

48. Wilke C, Fox SJ, Bates T, et al. wilkelab/cowplot: 1.1.1. Published online January 2, 2021. doi:10.5281/ZENODO.2533860

49. Wild F. lsa: Latent Semantic Analysis. Published online May 9, 2022. Accessed August 17, 2023. https://cran.r-project.org/web/packages/lsa/index.html

50. wrapper) JW (R, wrapper) B (R, src/word2vec) MF (Code in. word2vec: Distributed Representations of Words. Published online July 2, 2021. Accessed August 17, 2023. https://cran.r-project.org/web/packages/word2vec/index.html

51. Selivanov D, models) MB (Coherence measures for topic, code) QW (Author of the WC. text2vec: Modern Text Mining Framework for R. Published online November 30, 2022. Accessed August 17, 2023. https://cran.r-project.org/web/packages/text2vec/

52. Bates D, Maechler M, Jagan M, et al. Matrix: Sparse and Dense Matrix Classes and Methods. Published online August 14, 2023. Accessed August 17, 2023. https://cran.r-project.org/web/packages/Matrix/index.html

53. Fraley C, Raftery AE, Scrucca L, Murphy TB, Fop M. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. Published online October 31, 2022. Accessed August 17, 2023. https://cran.r-project.org/web/packages/mclust/index.html

54. infotheo package - RDocumentation. Accessed August 17, 2023. https://www.rdocumentation.org/packages/infotheo/versions/1.2.0.1

55. Kolde R. pheatmap: Pretty Heatmaps. Published online January 4, 2019. Accessed August 17, 2023. https://cran.r-project.org/web/packages/pheatmap/index.html

56. Jung L, Allard A. scrutiny: Error Detection in Science. Published online August 8, 2023. Accessed August 17, 2023. https://cran.r-project.org/web/packages/scrutiny/index.html

57. MAHLKNECHT U, KAISER S. Age-related changes in peripheral blood counts in humans. *Exp Ther Med*. 2010;1(6):1019-1025. doi:10.3892/etm.2010.150

58. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ*. 2005;331(7521):903. doi:10.1136/bmj.331.7521.903

59. Klee GG, Ichihara K, Ozarda Y, et al. Reference Intervals: Comparison of Calculation Methods and Evaluation of Procedures for Merging Reference Measurements From Two US Medical Centers. *Am J Clin Pathol*. 2018;150(6):545-554. doi:10.1093/ajcp/aqy082

60. Martinez-Sanchez L, Cobbaert CM, Noordam R, et al. Indirect determination of biochemistry reference intervals using outpatient data. *PLOS ONE*. 2022;17(5):e0268522. doi:10.1371/journal.pone.0268522

61. Arzideh F, Wosniok W, Haeckel R. Reference limits of plasma and serum creatinine concentrations from intra-laboratory data bases of several German and Italian medical centres: Comparison between direct and indirect procedures. *Clin Chim Acta Int J Clin Chem*. 2010;411(3-4):215-221. doi:10.1016/j.cca.2009.11.006

11 Appendix

Swiss BioRef dataset specifications

Analyte	Entries
Creatinine [Moles/volume] in Serum or Plasma	247081
Glucose [Moles/volume] in Serum or Plasma	256423
C reactive protein [Mass/volume] in Serum or Plasma	257481
Potassium [Moles/volume] in Serum or Plasma	289459
Sodium [Moles/volume] in Serum or Plasma	277022
Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P	142244
Prothrombin time (PT) actual/Normal	260125
Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area]	234722
INR in Platelet poor plasma by Coagulation assay	246904
Platelet mean volume [Entitic volume] in Blood by Automated count	245798
Hematocrit [Volume Fraction] of Blood by Automated count	247002
INR in Capillary blood by Coagulation assay	12150
Leukocytes [#/volume] in Blood by Automated count	247009
Hemoglobin [Mass/volume] in Blood	247004
Platelets [#/volume] in Blood by Automated count	247005
MCH [Entitic mass] by Automated count	246971
MCHC [Mass/volume] by Automated count	246971
MCV [Entitic volume] by Automated count	246973
Erythrocyte distribution width [Ratio] by Automated count	246860
Erythrocytes [#/volume] in Blood by Automated count	247004
Prothrombin time (PT)	167029
Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P	147427
Oxygen [Partial pressure] in Blood	90710
Carbon dioxide [Partial pressure] in Blood	91063
pH of Blood	92126
aPTT in Platelet poor plasma by Coagulation assay	76137
Oxygen saturation in Blood	74883
Chloride [Moles/volume] in Serum or Plasma	78583
Urea [Moles/volume] in Serum or Plasma	140991
Creatinine [Moles/volume] in Blood	63437

Cholesterol in HDL [Moles/volume] in Serum or Plasma	50303
Cholesterol [Moles/volume] in Serum or Plasma	50546
Triglyceride [Moles/volume] in Serum or Plasma	52758
Cholesterol in LDL [Moles/volume] in Serum or Plasma by calculation	47606
Hemoglobin A1c/Hemoglobin.total in Blood by IFCC protocol	40562
25-Hydroxyvitamin D2+25-Hydroxyvitamin D3 [Moles/volume] in Serum or Plasma by Immunoassay	9441
25-Hydroxyvitamin D2+25-Hydroxyvitamin D3 [Moles/volume] in Serum or Plasma	1136
Cholesterol in LDL [Moles/volume] in Serum or Plasma by Direct assay	2508
Insulin [Units/volume] in Serum or Plasma	332

Figure 4: Blood test analytes and their respective population size contained in the Swiss BioRef dataset used in this project

Swiss BioRef R Shiny App Github Repository

Link to repository: https://github.com/eldavidsc/Swiss-BioRef-R-Shiny-App

Declaration of consent

on the basis of Article 30 of the RSL Phil.-nat. 18

Name/First Name: Schaer David

Registration Number: 17-053-521

Study program: Computational Biology and Bioinformatics

Bachelor ☐ Master ☑ Dissertation ☐

Title of the thesis: Development of a big data analysis pipeline for precise reference interval estimation using ICD-10 code labelled data

Supervisor: Alexander Leichtle

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of 5 September, 1996 is authorized to revoke the title awarded on the basis of this thesis.

For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with future theses submitted by others.

Bern, 26 October 2023

Place/Date

Signature David Schaer