UNIVERSITÄT
BERN

MASTER THESIS

Awarding the academic title of Master of Medicine (M Med)

Medical Faculty, University of Bern

Multiple imputation to complete laboratory data as a foundation for computational evidence on diagnostic pathways

Scientific research in the medical field

Master Thesis submitted by

Zara Liniger

Immatriculation Nr. 10-126-332

Handed in 18th of December 2015

For the degree of

Master of Medicine (M med)

Supervisor: PD Dr. Alexander Leichtle,

Institute for Clinical Chemistry, Bern

Medical Faculty of the University of Bern

Table of Contents

Introduction

Laboratory diagnostics is an integral part of patient care. It is therefore astonishing that besides specific diagnostic guidelines for certain diseases {Hofmann:2012} there is no standardized procedure {AACC:2007} for the choice of laboratory diagnostic tests for all differential diagnoses. Different hospitals, books and experts recommend varying choices of laboratory tests for establishing diagnoses.

Putting such a standardized diagnostical system in place would not only improve patient care through more and more precise diagnostic pathways ("precision medicine") but would also contribute to reducing the continually rising costs of the health care system by saving unnecessary testing expenses. The theme of this study developed with these reasons in mind.

The overall goal therefore would be to create standardized diagnostic pathways, which are not only based on expert opinions but on solid evidence.

This master thesis and succeeding studies combined attempt to create computer based evidence for a few selected diagnosis as a pilot project. These selected diagnoses contain various ischemic heart diseases coded by the ICD-10-system from I20.0 to I25.19. Hence, laboratory data has to be analysed with the aim of calculating predictive parameters for these diagnoses.

An advantage for this study is that the data needed for the calculation was already available. Over years laboratory test results from the Emergency Department of the University Hospital of Bern have been routinely recorded. For this study neither the name, identifiers or date of birth of the patients were required, therefore the data set could be used in a completely anonymous form. In accordance with the Swiss Law on Human Research (Law № 810.30 Art. 2 [2]c) this anonymous form of study does not require approval of the Cantonal Ethics Committee {Leichtle:2014}.

To have access to such "big data" is a privilege as it has the potential of providing huge amounts of important and specific information. Therefore this potential should be put to use, as V. Mayer-Schönberger points out that "predictions based on correlations lie at the heart of big data" {Mayer-Schönberger:2013}.

However, one of the problems with almost all bigger (or sometimes even small) data sets is their incompleteness. An incomplete data set should not lead to disregarding all the valuable information in the rest of the data. These days with advanced computational methods on completing datasets this vast amount of data can be used.

The aim of this master's thesis is to complete the laboratory data set from the Emergency Department through applying the procedure of Multiple Imputation so that the data can be

used to create the computational evidence for laboratory diagnostic pathways. Also, different statistical diagnostics were applied to control the imputation for its appropriateness.
This complete data set will not only provide more information but is crucial for the creation of the models, from which the computational evidence can be derived. (Linear) model creation (e.g. by general linear models) cannot be done with missing values. In addition to that, the data should represent reality as good as possible. As each and every person has a “true” value for any parameter (even if not tested), a “missing value” for this parameter in the data is further form reality than any kind of thoughtfully recreated value.

Methods

Lab Data

The source of data is an extract of the 14 most tested laboratory parameters in the year 2009 until 2012 from the Emergency Department of the University Hospital Berne.
The selection criteria of patients was based on two parameters: All patients tested for blood Troponin (TnT) and Creatin Kinase (CK) where included in the study. The rationale for these criteria in our myocardial ischemia-focused study is to ensure that no patients with a suspected differential diagnosis for myocardial ischemia were missed.
Retrospectively the chosen patients could be divided into two groups: The first group had all the patients with one of the diagnosis ICD-10 I20.0-I25.19, and the second group consisted of all the rest (with TnT and CK measurements) and could therefore be used as the control group. All in all the study consists of 495 patients with one of the defined diagnosis and 2966 controls.

Missing data

The only two lab tests available in all patient cases were TnT and CK as they function as selection criteria (see above). All the other requested lab tests varied according to the patients' symptoms, medical history, the admitting physician and the differential diagnosis. Therefore *a priori* complete data for all the patients is not available.
Due to the missing data the focus was laid on those parameters, which were predominantly available for most patients. Therefore, in order to have valuable information and to limit the amount of missing data to impute, the analysis was reduced to the 14 most frequently tested parameters.
Although this selection reduced the amount of missing data, it did not provide data for all the 14 parameters for all the patients (see fig. 1).
To arrive at accurate conclusions for this study on the predictability value of the parameters (i.e. modelling), it would be necessary to have the complete set of data.
Missing Data is a common problem in most studies, which can lead to false conclusions if not interpreted with caution.
There are methods to partially overcome this problem. The choice of the method utilized depends on mechanism of "missingness".
Missing Data can be categorized into four different groups as described by Gelman et al: Data Analysis Using Regression and Multilevel-Hierachical Models (see discussion) {Gelman:2014wt}.
The problem in our study is that different mechanisms of "missingness" are likely to have played a role for the fact that certain data is missing (see discussion). As there is no

identifiable explanation or pattern, as to why some parameters where tested in few cases and others weren't, the process was simplified by presuming that all the parameters were selected at random and therefore the missing values are missing at random. Hence the missing data was put in the category MCAR ("missingness" completely at random).
Logically thinking, there must have been a reason why the physician chose these parameters, but due to the retrospective approach of this study this information cannot be included in the categorization of "missingness".
There are different methods of dealing with missing data with both advantages and disadvantages. If missing data is looked at completely at random, there would be the simple possibility of excluding all the cases with missing values.
This so-called "complete-case-analysis" would put a limit on the number of cases and the amount of information available. And as Gelman mentions: "[…] the sample of observations that have no missing data might not be representative of the full sample" therefore the evaluation will be biased. {Gelman:2014}
Looking at it from a practical point of view the complete-case analysis applied to this study would lead to a near to 100% loss of patient cases.
Plus, as discussed above, seeing that different kinds of mechanisms of missingness are involved and that MCAR is a simplified assumption, the complete-case analyse is even less suitable.
Instead of excluding cases with missing values there are possibilities to try to recreate/calculate these values with computational algorithms. For this study the method of multiple imputation of missing data was chosen as the calculation algorithm.

Multiple imputation of missing data

The advantage of using multiple imputation for missing data is that this algorithm takes into account the potential variability of a missing value and its relationship to other parameters. More precisely, multiple imputation constructs or "imputes" missing values through creating multiple regression models for the parameter in question as described in Jeffrey C. Wayman's paper {Wayman:2003}.
These different (iterative) regression models are based on observed parameters, which are influential on the prediction of the missing value. All of these models recalculate a slightly different value for the missing parameter and restore its natural variability. The more regression lines are created, the higher is the probability that the imputed value matches the actual missing value. {Wayman:2003}.
If values of different kinds of parameters are missing the mentioned approach can be modified to fit the circumstances. In this case the univariate regression model will be

replaced by a multivariate regression model, which requires advanced calculation methods {Gelman:2014}.
To avoid the difficulty of this process, another option would be to use an iterative regression approach {Liu:2013} to impute the missing data, which is used in this project.

Iterative regression imputation {Liu:2013}
The key to avoid the construction of a multivariate regression model is to create a situation, in which there is only one missing parameter and the rest of the data set has been completed. This will then allow a univariate model to impute this particular response parameter. To meet the prerequisite of a complete data set all the other missing parameters initially need to be given a value through a simple imputation method. Simple imputation methods would for example be to impute the missing value by using the mean of observed values of that particular parameter or simply by imputing a random value. After this simple imputation, these parameters can serve as predictors in the univariate regression model. This procedure is done repeatedly, however with each iteration the response parameter changes and the preceding imputed parameter serves as a predictor in the current univariate regression model. Note that the prior imputed value of the present response parameter needs to be overwritten by the resulting value of the current regression model. In addition to the iterations the algorithm runs in different chains parallel to each other. Each chain starts off with a different completed data set as a base, therefore each chain begins with a different regression model for a randomly picked initial response parameter.

The ultimate goal is that these chains eventually all reach the same posterior distribution {Berglund:2014, p. 24} for the response parameter. This state is called convergence, which is a mathematical expression for arriving at the stable state while having different starting points. In this study the stable endpoint is represented through a joint posterior distribution {Berglund:2014, p. 24} of the different chains for the imputed value of the missing parameter. To aim for a posterior distribution instead of a single value for the missing parameter not only reflects the natural variability of it's values but also avoids to "overstate precision" {Schafer:2005} of the imputed value.
For this study it means, that the value imputed for the missing parameter, is most likely to be as close as possible to the real value.

From a practical point of view the package "`mi`" {Gelman et al:2014} was used to calculate these iterative regression algorithms in "`R`" {R Development Core Team:2014}. The following basic codes were used:

1. Info_Matrix <- mi.info(data=MI)
2. Info_Matrix <- mi.info.update.include(object=Info_Matrix, list=list("HDIA" = FALSE, "Klasse" = FALSE))
3. Preprocessed_Data <- mi.preprocess(data=MI, info=Info_Matrix)
4. Imputed_Data <- mi(object=Preprocessed_Data, n.iter = 2000, max.minutes = 5000000)

The preceding codes contain the following information {Su:2011,p.3-6}:

- **MI** is the name of the base data undergoing multiple imputation.
- The function **mi.info ()** gives access to the information contained in the data (e.g. the number of missing values)
- The **first code line** merges "**mi.info**" with the base data **MI** into one term given the name **Info_Matrix**. This serves as a simplification of the coding process.
- The **second code line** defines the "Info_Matrix" more precisely. The term "**update**" enables the addition of information to our first definition of the "Info_Matrix". "**Include**" is a function, which defines in detail what should be included or excluded into/out of the multiple imputation procedure. Included is the **object** "Info_Matrix", and excluded is a list of items set on "**FALSE**". Generally speaking the inclusion of an object, which is set on "FALSE" turns the function around and leads to an exclusion of that particular object. The reason for excluding "HDIA" and "Klasse" is obviously necessary, as they classify the patient's diagnosis and therefore should not be used as predictors. HDIA stands for "Hauptdiagnose ICD-10" (main diagnosis).
- The **third code line** captures the pre-processing of the variables. As the package "`mi`" {Gelman et al:2014} can only work with a certain format of parameters, pre-processing is needed for all other formats. The formats recognized by the multiple imputation package are those of parameters with standard distribution. Non-negative, positive continuous and proportioned values (so called semi-continuous values) cannot be used by the 0.09-19 version of the package, as they are not of standard distribution and have bounds and truncations[1]. Through the function ("**typecast()**") included in the package the parameters are automatically classified. The function can sometimes misinterpret values thereby leading to a false classification. In this study the function should classify all parameters as positive-continuous variables, or else a manual correction needs to be done. This is important because not only the pre-processing method depends on this classification but also the regression model

[1] A number is truncated, if from its actual value, decimal places were cut off after a defined point, without rounding.

constructed by the multiple imputation. Through the function "**mi.info()**" the automatic classification of the parameters can be overviewed.
Pre-processing the positive-continuous parameters is done through a logarithmic transformation whereby standard distribution may be achieved. The imputed results then need to be post-processed (i.e. delogarithmized, **mi.postprocess**) to appear in there original form {Su:2011vm}.

- The **4th code line** contains the actual command for the imputation. The pre-processed data serves as base data. The imputation process stops when the value reaches convergence, this is a default setting ($\hat{R}$), which can be modified. As the state of convergence is not always attained, **n.iter** and **max.minutes** both control the termination of the process. Through **n.iter** the maximum number of iterations (repeats) of the imputation is defined. If either the maximum number of iterations or the maximum number of minutes of calculation time is reached, the process is brought to a halt.
  The number of separate imputation chains can be set with "**n.imput**". If it is not specifically defined multiple imputation runs on three different chains. This default setting applies to the imputation in this study. Version 0.09-19 only supports computation on a single core, however, using separate seedings, various processes can be started in parallel, e.g. in a linux environment (process scheduling).

Graphical diagnostics help to overview the multiple imputation process as they show the convergence and the accuracy of the created regression models.
The fitting of the model can be observed by methods given in the paper on "opening windows into the black box" {Su:2011}:

- The distribution of the missing values needs to be illustrated for initial analysis to avoid crucial mistakes in the model setup process. This overview shows for example if there is a parameter with very few or oddly distributed observed values. This information is necessary to then treat the imputed values of this parameter with more caution, as they may be less accurate.
- In addition to these basic analyses the regression models made by multiple imputation can be checked by comparing the imputed to the observed values.
  To illustrate this, four different plots are created for each parameter:
  - Histogram showing both observed and imputed values for a parameter, as well as the sum of those values
  - Scatterplot of the model predicted values against the residuals
  - Binned residual plot

- Bivariate scatterplot

The construction of these plots will be illustrated in the section “results”.
However the binned residual plot will be elaborated here due to its uniqueness.
Explanations will be based on the book “Data Analysis Using Regression and Multilevel-Hierarchical Models” {Gelman:2014}:
A residual is defined as “observed minus expected values” {Gelman:2014}.
The uniqueness of a binned residual plot is that the residuals are initially ordered into groups (so called “bins”) depending on their expected/predicted value. The average residual in these bins are then plotted against the expected value for that particular bin. Additionally two lines indicate ± 2 standard-error bounds, within which the binned residuals should be placed, if the model was accurate enough. {Gelman:2014}

For a visual control of the convergence “convergence-plots” were constructed. These illustrate the convergence of the mean and the standard deviation within and in-between the different imputation chains {Su:2011}. The parameters’ mean and the standard deviation are plotted against the number of iterations {StefvanBuuren:2009}

Results

This section of the thesis contains the results of:

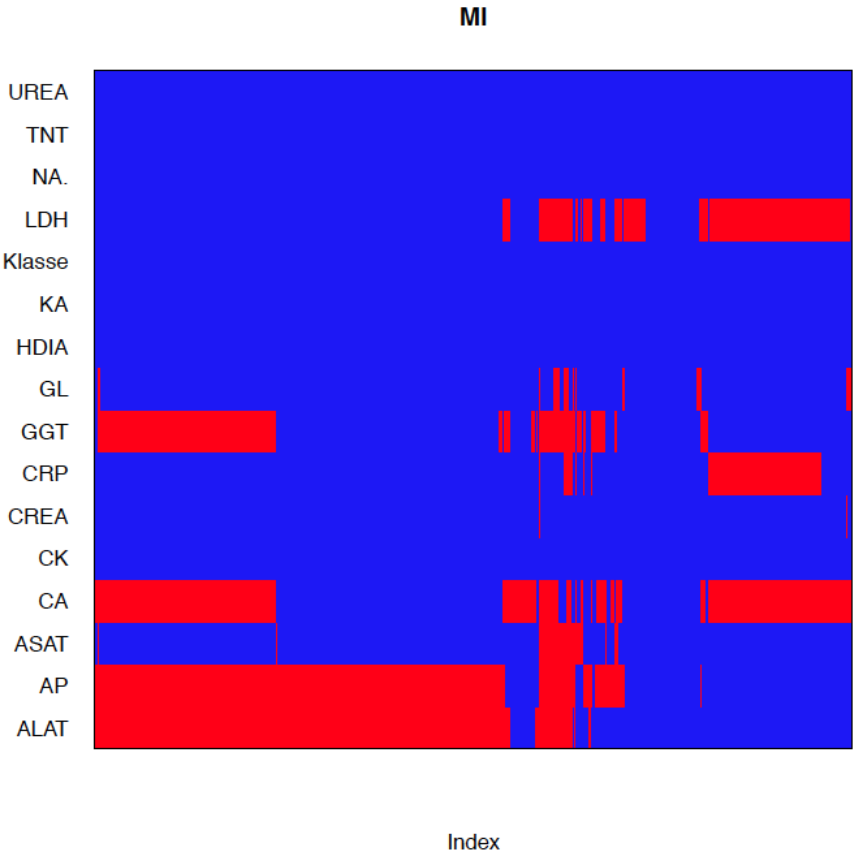
1. Overview plot of the missing data in the actual data set
2. Section of the completed laboratory data set
3. Convergence plots illustrating the last thirty-one iterations of the MI process
4. And the accompanying plots for the analysis of the model fitting

Overview plot

For basic analysis, as mentioned in the method section, an overview of the missing data is shown below.

.

Figure 1:



The 14 selected parameters are represented in rows. The patients are plotted along the column. Indicated in blue are the observed values and in red the missing ones. The patients are organized in such a way, that the same missing values are placed next to each other, for better visualisation {Su:2011}.
UREA (urea), TNT (troponin), NA. (sodium), LDH (lactate dehydrogenase), Klasse (patient's specific disease class), KA (potassium), HDIA ("Hauptdiagnose ICD-10", code of the patients main diagnosis based on the ICD-10 classification), GL (glucose), GGT (gamma-glutamyl transpeptidase), CRP (C-reactive

protein), CREA (creatinin), CK (creatin kinase), CA (calcium), ASAT (aspertate aminotransferase), AP (alkaline phosphatase), ALAT (alanine transaminase), Index (patient case number).

Completed laboratory data set

As a result of the multiple imputation, the data set was completed. Due to the large number of patients, only a small section of the data set is shown for illustration.

Table 1:

MDIA	KA	NA.	TNT	UREA	LDH_MI	ALAT_MI
A084	3.1	139	0.009	6.6	363	13
A084	4.1	132	0.01	15.1	1023	41
A084	3.2	136	0.009	20.7	***281.63***	25
A09	3.7	135	0.009	6.8	***332.72***	17
A09	2.9	138	0.012	7.2	***411.44***	14
A09	4.5	147	0.009	17	505	37
A09	3.5	139	0.009	18.1	***528.06***	***50.85***

MDIA stands for ICD-10 Main Diagnosis. The columns show the different parameters (only a selection shown in this table). The rows display different patients with their diagnosis (indicated by the ICD-10 diagnosis codes). According to the Swiss law on human research (Law № 810.30 Art. 2 [2]c), this study is therefore not subject to approval by the Cantonal Ethics Committee (Dispensation № Z023/2014).
The numbers in black are observed values of the parameters and those in red are the imputed ones.

However the following graphics have a greater interpretational value, as they serve as visual inspection of the model fitting and the convergence.

Convergence plots

The following illustrations (fig. 2) are a selection of convergence plots.

Figure 2:

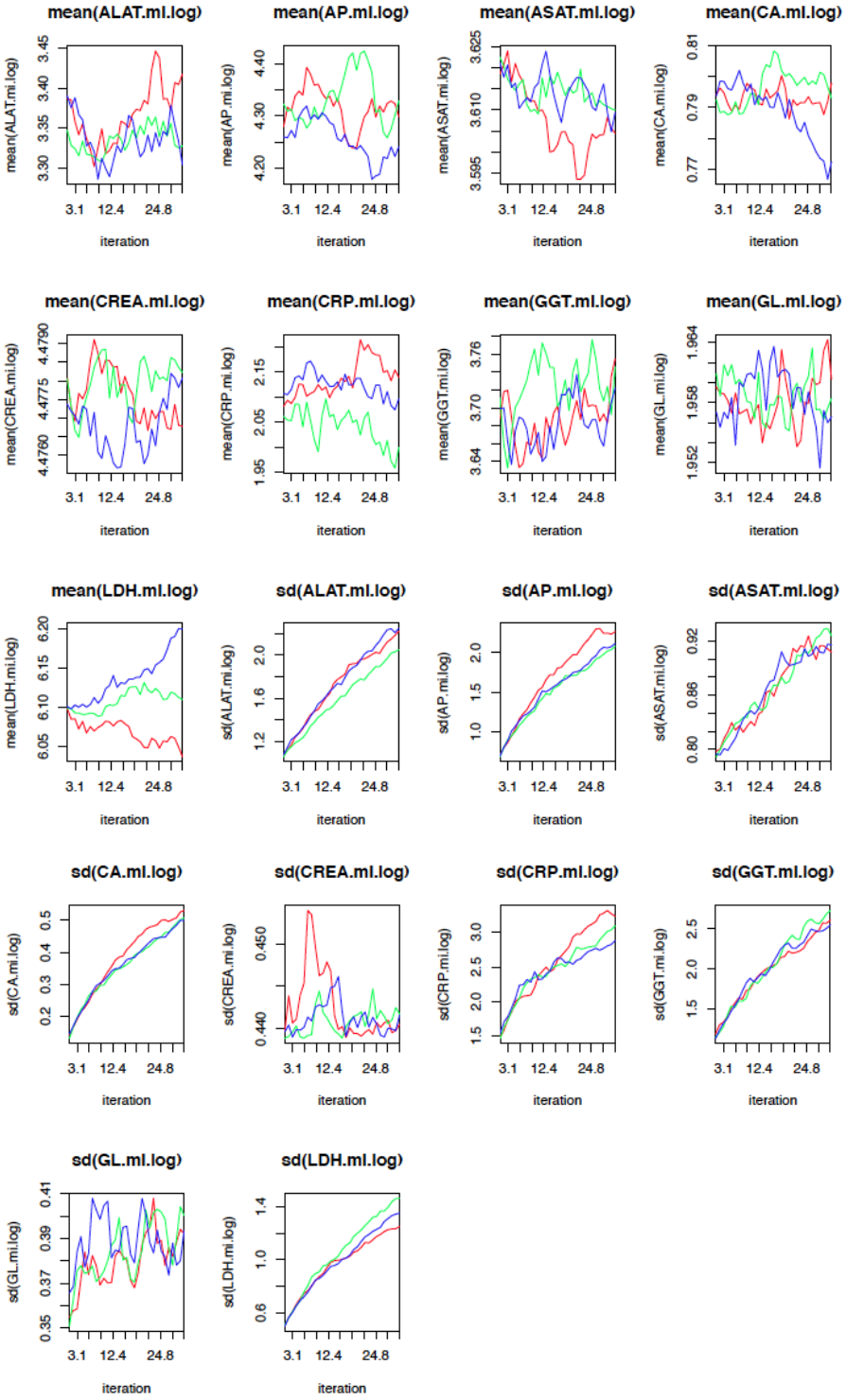


Figure 2: Due to the large amount of different parameters, only a few have been selected to illustrate the convergence plots. These convergence plots show the last 31 iterations. These plots are created directly after the multiple imputation process by "`mi`" {Gelman et al:2014} itself. The number of iterations plotted is defined prior to the beginning of this process. Therefore the plots cannot be changed retrospectively

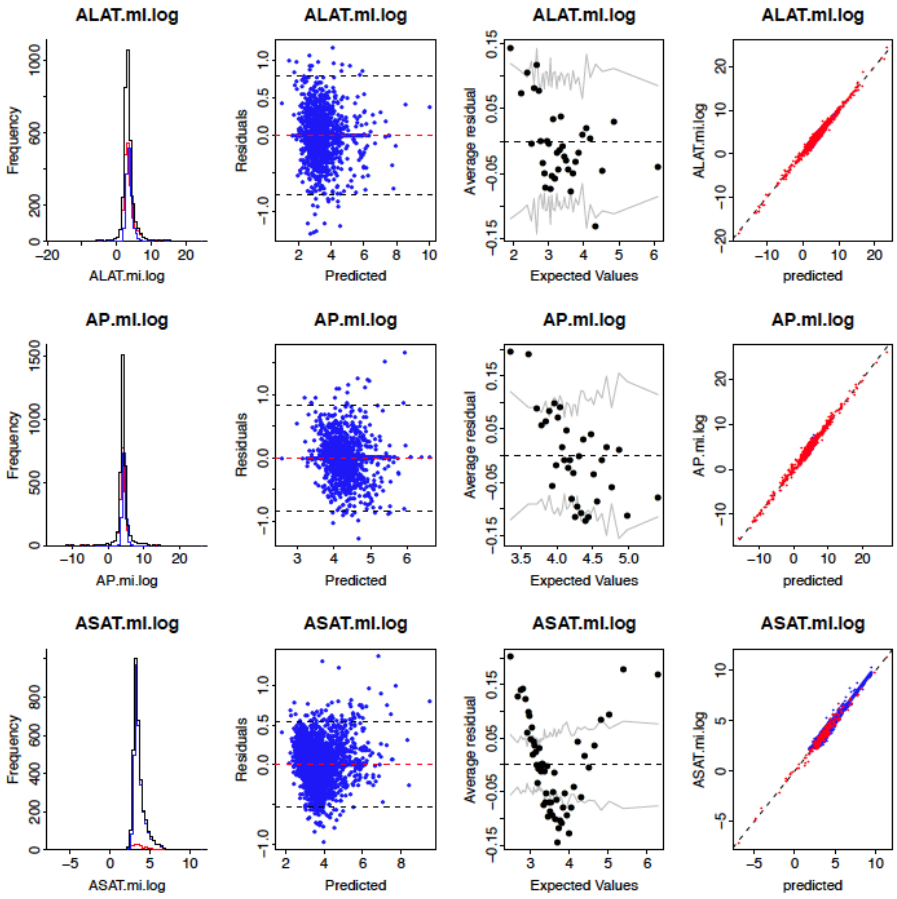
without repeating the whole multiple imputation process again. Why it would be of benefit to change the number of iterations plotted see “interpretations of the results” below.

The x-axis indicated the iteration number, while the y-axis shows the mean or standard deviation of the imputed parameter for each iteration {SSCC, 2009-2015}. The mean results from the different imputed values in each patient case during that specific iteration.

The colours red, blue and green represent one of the imputation chains.

Graphs for the visual control of the model fitting

Figure 3:



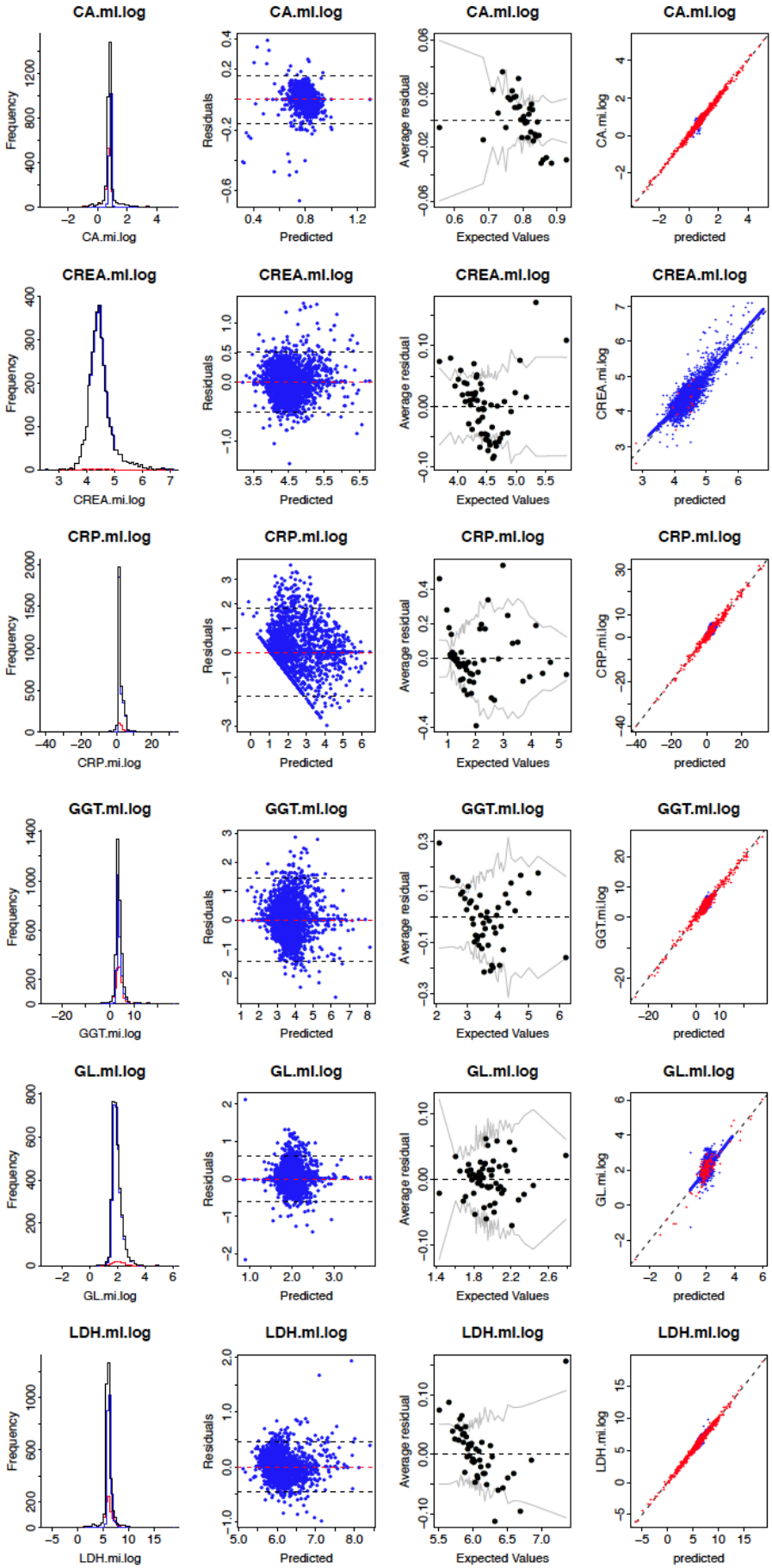

CA.ml.log
Frequency
CA.ml.log
Residuals
Predicted
Average residual
Expected Values
CA.ml.log
predicted
CREA.ml.log
Frequency
CREA.mi.log
Residuals
Predicted
Average residual
Expected Values
CREA.mi.log
predicted
CRP.ml.log
Frequency
CRP.mi.log
Residuals
Predicted
Average residual
Expected Values
CRP.mi.log
predicted
GGT.ml.log
Frequency
GGT.mi.log
Residuals
Predicted
Average residual
Expected Values
GGT.mi.log
predicted
GL.ml.log
Frequency
GL.mi.log
Residuals
Predicted
Average residual
Expected Values
GL.mi.log
predicted
LDH.ml.log
Frequency
LDH.mi.log
Residuals
Predicted
Average residual
Expected Values
LDH.mi.log
predicted

Figure 3 shows four different diagnostic plots of 9 parameters of one imputation chain as an illustrative example.

1. Histogram of the observed (blue), imputed (red) and completed (black outlining representing the sum of both predicted and observed values) values of the indicated parameter {Su:2011vm}. On the x-axis the logarithm of value of the parameter is indicated and on the y-axis the frequency of a certain value is shown. Note that through the use of the logarithm the scale on the x-axis consists of positive as well as negative numbers.
2. Scatterplot of the model predicted values against the residuals.
3. Binned residual plot
4. Bivariate scatterplot: the blue colour represents the observed values where as the red colour stands for the imputed ones.

If the model fits well the distribution of the observed and the imputed values match well {StefvanBuuren:2009}.

Discussion

This chapter of the master thesis explains the mechanisms of “missingness” in further detail and contains the interpretation of the results (see above), possible improvements and limitations as well as the continuation of this study.

Categorise of “missingness”

In the following passage the definition of the four categories of “missingness” will be briefly illustrated with examples from this study.
TnT and CK being a selection criterion (therefore without missing values) are excluded in these examples. The term “parameters” will refer to the remaining parameters.

MCAR (“Missingness” completely at random):

The values in the data would be missing completely at random, if the choice of the parameters to be tested was made randomly. This for example would be the case if the physician filled out the laboratory form without looking at it. In other words there would be no way to reason the choice of the parameters and therefore no way to find a reason for the missing data.

MAR (“Missingness” at random):

MAR is described in a lecture by Dr. Matechou on “Missing Data” as “Missingness” that depends on the observed data itself {Matechou:2013wa}.
In this study the differential diagnosis (DD) of the admitting physician would be observed data.
If data is missing at random, “missingness” can be explained by the differential diagnosis (observed data) e.g. if hepatitis was not included in the DD, the “missingness” of liver function tests could be explained by this fact.
The retrospective nature of the study doesn’t allow for access to these differential diagnoses. Therefore it can’t be known for certain whether the data is missing at random.
In general as mentioned in Gelman’s work, to prove that Data is missing at random is near to impossible, as there may be many additional unobserved parameters playing a role in the mechanism of missingness {Gelman:2014}.

MNAR (“Missingness” not at random):

MNAR can be divided into two groups:

- “Missingness” due to an unknown predictor:

In this case the parameter was not tested (therefore missing) because there was no known relation to the suspected diagnosis. For example if all the patients had the same a suspected diagnosis, if they had all been admitted by the same physician, who had had a precise panel of parameters tested for this diagnosis, then a certain parameter might have been systematically left out because the physician did not see the relevance of the parameter for the mentioned diagnosis. Obviously this is not the case in this study.
The patients had different suspect diagnoses, and they were admitted to different physicians with varying knowledge.
All these factors do not allow categorizing the data as "missingness" due to an unknown predictor.

- "Missingness" that depends on the missing value itself:
  If the testing of a parameter is not possible then the "missingness" of this parameter depends on the specific parameter it's self. E.g. the value of the parameter doesn't exceed the detection limit.

Interpretation of the results

The presentation of the interpretations will be done in the same format as the results:

1. Concerning figure 1:
   At first glance there are apparently few parameters with no missing values. However two of these are TnT and CK, which as selection criterion for the differential diagnoses were available for all patients. Furthermore "R" {R Development Core Team:2014} did not distinguish "Klasse" and "HDIA" from the parameters therefore they are displayed in the plot. As all patients had "Klasse" and "HDIA" there would be no missing values in those rows. This could be corrected, by setting these two terms on "FALSE" while creating the plot in "R" {R Development Core Team:2014}.
   The only two other parameters with no missing values are potassium and urea. Potassium can be explained through the general recommendation to test electrolytes if myocardial ischemia is a part of the differential diagnosis {Hofmann:2012}. For the fact however that urea was tested in all patients no plausible explanation can be found.
   The parameters with the most missing values are calcium, alkaline phosphatase and ALAT. It would therefore be interesting if the multiple imputation could find appropriate values for these parameters.

2. Concerning table 1:
   There's a limited amount of information that can be retrieved by looking at the imputed values. However there are some key points to be examined (according to the paper on "MICE"{StefvanBuuren:2009}):
   The first question needed to be asked is whether the imputed values are plausible. Do they for example fit into the expected range of the parameters or do they take values which could not possibly be tested in patients. {StefvanBuuren:2009}

3. Convergence plots (fig. 2):
   According to Stef van Buuren {StefvanBuuren:2009} in the state of convergence the imputation chains intermix freely with each other and no specific trend of the chain can be recognized. Convergence is reached then, when the variance within a chain is no larger than the variance between the chains. {StefvanBuuren:2009}.
   As mentioned in the paper MICE, "Plotting somewhat longer iteration sequences will generally convey a good idea whether the between-imputation variability has stabilized and whether the estimates are free of trend." {StefvanBuuren:2009}
   Therefore creating plots with e.g. the last 100-200 iterations would provide a better overview to assess the convergence state.
   In about two thirds of the mean convergence plots in figure 2 the above mentioned criteria for the state of convergence can be expected to be met.
   There is however a particular tendency of the standard deviation convergence plots to show strong trends. At the moment there is no obvious reason for the recognizable phenomena. Hence, further investigations, maybe with a bigger number of iterations, have to be done. Or, as discussed later on, the regression models have to be adjusted.
   Only the plots of "sd(CREA.mi.log)" and "sd(GL.mi.log)" don't show strong trends in the chains as seen in all the other standard deviation convergence plots. This may be explained by the nearly no missing values of CREA and GL in the data set.
   What can be taken into considerations for the problem investigation are following examples of causes for failing convergence:
   a. A study in Stef van Buurens book on "Flexible imputation on missing data" shows that the state of convergence is reached slowly if there is a high percentage of missing values {StefvanBuuren, 2012, p. 115}.
   b. A common cause of failed convergence seems to lie in insufficient data for the estimation of a certain parameter {Enders:2010, p. 255-256}. As further explained in "Applied missing data analysis" by Craig K. Enders, this problem then leads to including too many variables into the prediction-model for a

specific parameter. As a result a possible way to overcome this problem is to reduce the number of predictors in a model or eliminate the questionable predictors {Enders:2010, p. 255-256}. A more elaborate description of the problem can be seen under “Possibilities to improve the method” (see below).

c. This is by far not a complete list of possible obstacles in reaching convergence, but it should give a sufficient starting point to dealing with failing convergence. If convergence is still not achieved further possibilities need to be looked at.

4. Examining the binned residual plots (being the most useful plot to evaluate the model fitting) all the parameters have average residuals beyond the 95% standard error bound. Residuals beyond the 95%-standard error bound “show that there is room for improvement” {Su:2011} in the model creating process.
   It is interesting to see that ASAT, without many initial missing values, shows an inaccurate model fitting. This might suggest, that in the specific models for ASAT there are predictive parameters with a lot of missing values, but for the model high predictive quality.

Possibilities to improve the method

As the state of convergence is the procedural goal of multiple imputation, deeper analysis should be done when some of the parameters never converge. As already mentioned above, one of the factors that play a role in this process, is a sufficient number of iterations. The more missing values the bigger the number of iterations needed to reach convergence/for the multiple imputation process to succeed.

Another possibility to improve reaching the state of convergence faster, is mentioned in the paper “ MICE: Multivariate Imputation by Chained Equation in R”. {StefvanBuuren:2009} Although this publication uses MICE to complete a data set as opposed to this study using multiple imputation, there should be a possibility to use some of the valuable functions shown in the paper. The authors describe a function “**visitSequence**” which has the ability to “visit” the parameters for the imputation according to their number of missing values (in increasing order). By bringing in such a structure into the imputation process it has been shown, that convergence can be attained faster in certain situations. {StefvanBuuren:2009}. This function might not apply to this study, as “`mi`” {Gelman et al:2014} does not support sequence visiting, but it could be an inspiring thought experiment with the order of imputation to achieve better convergence.

A very crucial and probably the most challenging part of multiple imputation is the creation of fitting models. As seen in the results (e.g. ALAT), the models of this study need to be adjusted to improve the imputation.

As indicated in Stef van Buuren's publication {StefvanBuuren:2009}, a model should fit to the following description:
It "should account for the process that created the missing data, preserve the relations in the data and preserve the uncertainty about these relations."
A possible way to improve models used in the MICE method, is to define precise predictors for each parameter amongst the other parameters. The created model for the parameter should include as many predictors as possible and their interaction, however too many predictors can provide an inaccurate, overfitted model. As one can see, this process is a balancing act. Stef van Buuren gives an elaborate description on the choice of predictors on pages 22-25 of "MICE: Multivariate Imputation by Chained Equations in R". The selection of the predictors for each model would open a whole new chapter and is therefore not discussed further in this master thesis.
The social science computing cooperation from UW-Madison provides interesting options on improving model fitting in the article "Multiple imputation in Stata: Imputing" {SSCC:2009-2015}. It recommends a time-consuming approach for checking each model to see if convergence is achieved. Another interesting approach, concerning the selection of predictors for the models, is to create a working model with the least number of predictors and then gradually include more predictors until the model stops to work. {SSCC:2009-2015}
Generally, it would also be interesting to define the maximum number and patterns of missing values, for which the multiple imputation process could still find fitting models and therefore converge. A possible advantage of including parameters not generally associated with a diagnosis, and therefore more likely with a larger number of missing values, is that possible new association (maybe even negative association) could be found.

Limitations of this study

There are various steps in multiple imputation which are still a "black box" {Su:2011}. The less "black boxes" there are, the more possibilities exist to improve the method. The package, being a computer program with no prior knowledge of the study, has to make certain assumptions to be able to perform the algorithms. All these assumptions generally speaking can be possible sources of error. "Responsible use of MI requires basic understanding of these assumptions and the possible implications for subsequent analyses if they are violated" (Olsen & Schafer:1998). Without opening these "black boxes" the full potential of multiple imputation cannot be achieved.
Although Stef van Buuren's paper seems to provide keys to these "black boxes", the question still remains, if these particular functions can be modified adapting the "`mi`" {Gelman et al:2014} package.

The continuation and potential of the study

As the aim of the whole project is to improve diagnostic pathways, here is a brief overview of the methods used to achieve this aim:

The data set will be used to create logit models through Bayesian model averaging. The creation of these models must be done with a complete data set, therefore this master thesis is an essential part for the continuation of the study.

The generated models will be constructed out of the parameters in such a way that they predict a certain diagnosis. Models with little predictive value will be left out of the process by a penalization term.

To be able to look at the predictive value of each parameter separately, the inclusion probability of the parameter in the predictive models is calculated. The parameters with the highest inclusion probabilities are those with the largest predictive value.

To estimate the precision of the calculated inclusion probability, the method of bootstrapping is used for the several individual imputation chains to create a confidence interval for the inclusion probabilities.

If all in all this method proves to be sustainable it should be used to predict other diagnoses than those of myocardial ischemia (ICD-10 I20.0-I25.19). However for diagnoses other than myocardial ischemia, the number of inputs in the laboratory data of the emergency room decreases rapidly. Therefore the methods need to be improved in such a way, that they work for smaller amounts of data and cases too.

Executive Summery

With the aim of providing computational evidence for diagnostic pathways a large amount of laboratory data was requested as study database from the emergency ward at the University Hospital of Bern. The project aims at mathematically finding laboratory parameters with a high diagnostic value.
For this project the diagnoses of myocardial ischemia ICD-10 I20.0-I25.19 were chosen as a pilot study subject for developing a suitable method to achieve this goal.
With these diagnoses in mind, the patient selection from the laboratory data set was based on the availability of the parameters TnT and CK. By this selection a group of 495 patients with one of the defined diagnoses and 2966 controls could be included in this study.
The laboratory data, although available in large quantities, was incomplete as far as the parameters were concerned. Therefore the method of iterative regression imputation included in the package "`mi`" was used to complete the data set.
This method constructs or "imputes", missing values through creating multiple univariate regression models for the parameter in question.
These different regression models, which are based on influential parameters, all calculate a slightly different value for the missing parameter and therefore restore its natural variability.
Through specific coding in "`R`" the method can be defined more precisely.
With diagnostic graphics the process of multiple imputation could be overviewed to see if the imputed values are plausible.
The results showed that although the results are largely plausible there is still room for improvement of this method, especially for some inaccurately fitting regression models. But it also shows that the values imputed generally matched the expectations. Therefore this method has an unique potential that can be further optimized.
The possibility to create computational evidence with this completed data set can revolutionize diagnostic pathways by contributing to a standardized procedure for lab testing e.g. in a emergency room {Clinical Chemistry:2014}.
Putting such a system in place would not only improve patient care through more precise diagnostic pathways but would also contribute to reducing the continually rising costs of the health care system.

Acknowledgement

I would like to thank my supervisors Prof. Dr. Martin Fiedler and PD Dr. Alexander Leichtle for their guidance and mentorship as well as for providing me with the opportunity to attend conferences both in Germany and the US.
These have broadened my horizons tremendously and exposed me to what can be done with “big data”. I enjoyed learning about this new field.
I also wish to thank my mother, Dr. Naim Liniger-Janmohamed, for helping me to present the tremendous amount of information in a reader friendly way!

References

AACC (2007). Evidence-Based Laboratory Medicine: Principles, Practice, and Outcomes, 2nd Edition.

Berglund, P. & Heeringa, S. (2014). Multiple Imputation of Missing Data Using SAS.

CLINICAL CHEMISTRY, Vol. 60, No. 10, Supplement, 2014 S133, B-008

Gelman, A. & Hill, J. (2014). Gelman A., Hill J. Data Analysis Using Regression and Multilevel-Hierarchical Models (CUP, 2006)(ISBN 9780521867061)(O)(651s)_MVsa_. *Data Analysis Using Regression and Multilevel-Hierarchical Models*, 1–651.

Gelman, A., Hill, J., Su, Y., Yajima, M. & Grazia Pittau, M. (2014), Version `mi0.09-19`

Hofmann, W., Aufenanger, J. & Hoffmann, G. (2012). Klinikhandbuch Labordiagnostische Pfade: Einführung-Screening-Stufendiagnostik, 1–214

Leichtle, A. B. (2014). Gottfried und Julia Bangerter-Rhyner-Stiftung, 1–16.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2013). On the Stationary Distribution of Iterative Imputations, 1–18.

Mayer-Schönberger, V. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think

Matechou, E. (2013). Missing Data, 1–19.

Olsen, M. K., & Schafer, J. L. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective,1–42.

R, R Develoment Core team (2014), Version `R`-3.1.2, www.r-project.org

Schafer, J. L. (2005). *Multiple Imputation: a primer* (pp. 1–13).

SSCC (2009-2015), http://www.ssc.wisc.edu/sscc/pubs/stata_mi_impute.htm#table

Stef van Buuren, K. G.-O. (2009). MICE: Multivariate Imputation by Chained Equations in R, 1–68.

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011a). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box, 1–31.

Su, Y.-S., Yajima, M., Gelman, A., & Hill, J. (2011b). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*, *45*(2), 1–31.

Wayman, J. C. (2003). Multiple Imputation For Missing Data: What Is It And How Can I Use It?, 1–16.

Erklärung

"Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, habe ich als solche kenntlich gemacht. Mir ist bekannt, dass andernfalls der Senat gemäß dem Gesetz über die Universität Bern zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist."

Datum und Unterschrift des Studierenden