RECAP

"Research on European Children and Adults born Preterm"

Grant Agreement number: 733280

Deliverable 5.1

Report on a unifying missing data perspective to IPD problems

Workpackage:	WP 5
Task:	T 5.1
Due Date:	31st August 2017 (M8)
Actual Submission Date:	31st August 2017 (M8)
Last Amendment deliverable:	15th January 2018
Project Dates:	Project Start Date: January 01, 2017
	Project Duration: 51 months
Responsible partner:	TNO
Responsible author:	Prof. dr. Stef van Buuren
Email:	Stef.vanBuuren@tno.nl
Contributors:	Manon Grevinga, Sylvia van der Pal, Aurélie Piedvache, Jennifer Zeitlin

Project funded by the European Commission within H2020-SC1-2016-2017/H2020-SC1-2016-RTD		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	CO

Document History:

Version	Date	Changes	From	Review
V0.1		Deliverable template	Maaike Beltman	
V1.1	24 July 2017	First concept of report	Stef van Buuren	Manon Grevinga & Sylvia van der Pal
V1.2	31 July 2017	Second concept	Sylvia van der Pal	Manon Grevinga
V1.3	7 August 2017	Third concept	Manon Grevinga	Sylvia van der Pal
V1.4	10 August 2017	Fourth concept	Aurélie Piedvache	Manon Grevinga
V1.5	25 August 2017	Draft report	Stef van Buuren	Eero Kajantie & Manon Grevinga
Final	31 August 2017	Final report	Stef van Buuren	

Open Issues

No:	Date	Issue	Resolved
1			

SUMMARY

This Deliverable 5.1 (D5.1) of the RECAP project describes a unifying missing data perspective to problems related to combining Individual Patient Data (IPD) for cohorts that track children who were born very preterm (VPT) or with a very low birth weight (VLBW).

The report starts a summary of common problems when combining individual patient data (IPD). In order to deal with such issues, we use the missing-data perspective to propose a seven-step approach:

1. Ideal Data: envision what data we would like to have had to solve our problem given unlimited resources;
2. Ideal Analysis: define what analysis we would perform if we had the ideal data;
3. Available Data: evaluate which parts of the ideal data are actually available to us;
4. Missing data: determine why some parts of the ideal data are observed, and why others are not;
5. Replications: construct replications of the unseen ideal data;
6. Calculate: derive the answer from each replication by the method of point 2;
7. Summarize: aggregate the replication to obtain the answer.

The approach is illustrated on three hypothetical scientific questions from the RECAP project. We outline how multiple imputation is a generic solution. When properly executed, the seven-step approach results in appropriate statistical estimates.

We envisage that our approach may be especially beneficial to solve data combination problems for which no proper solution yet exists, or for augmenting existing procedure for which there is no good quantification of the uncertainty caused by data combination.

Table of contents

1 INTRODUCTION

1.1 Purpose and Scope

This document presents a unifying missing data perspective to problems related to combining Individual Patient Data (IPD). The perspective may contribute to finding principled solutions of statistical analysis problems arising from the combined analysis of data from a collection of cohorts that track children who were born very preterm (VPT) or with a very low birth weight (VLBW) as brought together by the RECAP project. This report is deliverable 5.1 (D5.1) of the RECAP project.

Work package 5 of the RECAP project consists of activities to develop adequate statistical methodology needed to solve analytic problems arising from the other work packages. Work package 5 focusses on three problems associated with IPD: *harmonisation*, *loss to follow up*, and *individual prediction*. On the surface these problems appear to differ, but they can all be framed as 'missing data problems'. In each problem, only part of the needed information is observed, whereas other needed information is missing, and the analytic objective to find the missing information based on what we have. Benefits of framing the three IPD problems as a missing data problem include:

1. It may stimulate the use of a common and precise vocabulary for seemingly different problems;
2. As opposed to models, everybody understands data, so it is easier to communicate what exactly the problem is, and how we can attack the problem;
3. There is a general solution of missing data problems – multiple imputation – that nearly always works.

This report outlines several problems that need to be solved when combining data from different sources. We then outline how a seven-step approach can be formulated from the missing-data perspective, illustrate how it can be applied to hypothetical questions of scientific interest in RECAP, and show how a generic quantitative solution can be obtained by multiple imputation.

1.2 References to other RECAP Documents

This document provides statistical concepts that will be implemented in the RECAP statistical analysis platform (WP4), in close collaboration with the other work packages.

1.3 Definitions, Abbreviations and Acronyms

Table 1 List of Abbreviations and Acronyms

Abbreviation/ Acronym	DEFINITION
VPT	Very preterm
VLBW	Very Low birth weight
IPD	Individual Patient Data
QoL	Quality of Life
GA	Gestational age
BW	Birth weight
MAR	Missing at random

2 COMMON PROBLEMS IN COMBINING INDIVIDUAL PATIENT DATA

Meta-analysis attempts to combine the results from multiple studies with the objective to having a larger base of evidence. Meta-analysis increases statistical power, enables subgroup analysis, and can sort out possible inconsistencies between scientific studies. The analysis of individual participant data (IPD) is a form of meta-analysis that takes the raw data from each relevant study, and executes an analysis on the combined data. Debray et al. (2015) provide an impressive list of applications of IPD:

1. to disentangle subject-level and study-level sources of heterogeneity in treatment effect;
2. to study effect modification;
3. to adjust for confounding variables;
4. to improve data quality;
5. to standardize definitions and analyses;
6. to obtain complete follow-up data on all randomized participants;
7. to combine studies with different follow-up times;
8. to analyze multiple outcomes;
9. to investigate long-term outcomes;
10. to investigate rare exposures.

Most of these analytic objectives cannot be achieved by classic meta-analyses that are solely based on published estimates. IPD is therefore considered as a gold standard in evidence synthesis.

An important requirement for IPD is the combination of data sets. This sounds easy, but in practice, data combination is riddled with practical difficulties. The following list contains common problems in combining data from different studies:

1. Studies measure collect different sets of variables that measure the same construct;
2. Studies apply different measurement instruments;
3. The timing of the measurements varies widely between studies;
4. Study employ different designs to select units, or to allocate treatments;
5. Data are missing for different reasons, e.g. loss to follow up, not administered, skipped;
6. The key to link data from the same individual is imprecise, absent or contains duplicates;
7. The original data were collected for different analytic objectives;
8. Data may be sensitive, and at risk for de-identification after combining;
9. Definitions and classification may change over time;
10. Access to the original study sources is restricted.

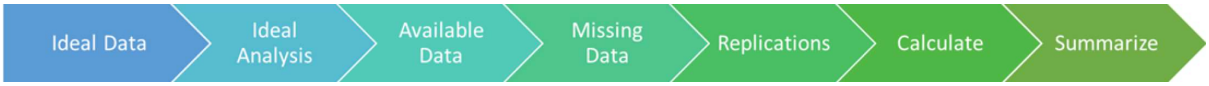
This list is by no means exhaustive and purely illustrative. Some issues are related, and multiple problems could co-existent for a given collection of study sources.

3 MISSING-DATA PERSPECTIVE IN SEVEN STEPS

3.1 Seven steps

This report explores the missing-data perspective to data combination. The data combination problems can be seen as manifestations of a deeper problem. In practice, we almost never see *all* data relevant to our question of interest, and we must usually simply deal with what we have. We often need some mastery to make the best use of available fragments of data.

The missing-data perspective is a conceptual framework for analyzing data as a missing data problem. Many statistical techniques address incomplete data problem. Suppose that we are interested in knowing the mean income in a population. If we take a sample from the population, then the units *not* in the sample will have missing values on income (because were not measured). It is therefore not possible to calculate the population mean directly since the mean is undefined if one or more values are missing. Rather, the population mean can only be estimated. One method to estimate the mean requires us to assume that the sample (for which we have income observed) is a random subset of all members of the population, so that -on average- the sample mean is equal to the population mean. The remaining uncertainty can be captured as a confidence interval. Thus, in order to solve this seemingly simple estimation problem, we need to think about the data that we have not observed, and think about what we would have done had the data been complete. This missing-data perspective covers sampling from a population, the counterfactual model of causal inference, prediction of outcomes, statistical modeling of missing data, and many other seemingly different statistical computation techniques. The book by Gelman and Meng (2004) and the paper by Little (2013) provide in-depth discussions of the generality and richness of the incomplete data perspective.



In this report, we implement the missing-data perspective as a sequence of seven steps:

1. Ideal Data: envision what data we would like to have had to solve our problem given unlimited resources;
2. Ideal Analysis: define what analysis we would perform if we had the ideal data;
3. Available Data: evaluate which parts of the ideal data are available to us;
4. Missing data: determine why some parts of the ideal data are missing;
5. Replications: construct replications of the unseen ideal data;
6. Calculate: our answer from each replication by the method of point 2;
7. Summarize: the answer over the replications.

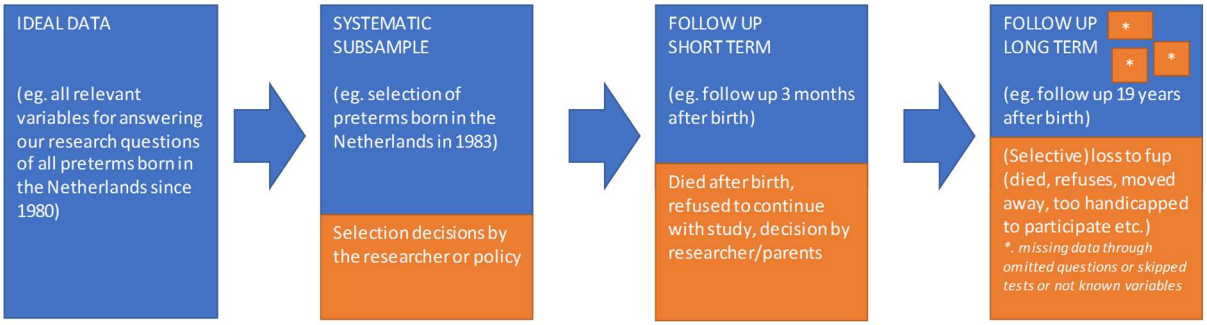
The scheme is designed to yield unbiased results of the parameters of interest in a broad set of practical cases. In addition, it can provide an estimate of the uncertainty of each parameter estimate. Many classic statistical techniques can be implemented using this generic scheme. Our hope is that the scheme may inspire solutions and techniques for problems that do not yet have a solution, especially for data combination problems.

We now describe three examples in which we demonstrate how this perspective works and is applicable to different missing data problems.

3.2 Example 1

Suppose we want to answer to following research question:

What is the distribution of Health-Related Quality of Life (HRQoL) at 19 years by gestational age at birth, per year in The Netherlands since 1980?



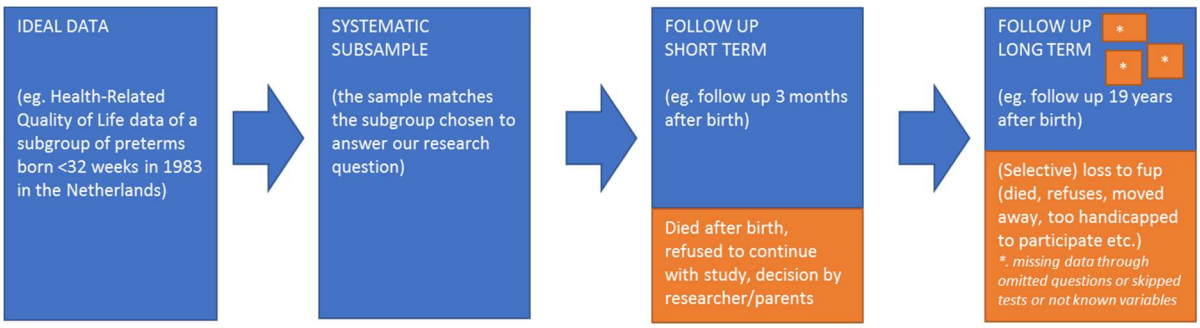
In that we case, we have attempt to apply the seven steps as follows:

1. Ideal data: We would like to have gestational age, birth year, one-week survival, and birth weight and Health-Related Quality of Life outcomes at age 19 of all children born the Netherlands since 1980;
2. Ideal analysis: For each year, we select all children surviving up to a least one week, calculate the median birth weight, visualize the median against gestational age, and summarize this relation be a linear regression model with one slope;
3. Available data: The POPS data. Full cohort of VLWB/VPT from in The Netherlands in the year 1983. A lot is missing, > 98% of children born in 1983.
4. Missing data: Investigator decided to restrict to GA < 32w or BW < 1500 g, and we are missing all other years except 1983. There are some missing outcome data, potentially related to the health of the infant;
5. Replications: We should find plausible values (presumably from other sources) of gestational age, birth year, one-week survival, and birth weight for the 98% of missing children in 1983, and the 100% of missing children in the other years since 1980. Although theoretically bona fide, this is not a manageable problem, so the analysis stops here. These data are not suitable to do the analysis.

3.3 Example 2

Suppose we want to answer to following research question:

What is the distribution of Health-Related Quality of Life at 19 years, by gestational age at birth in a subgroup of preterms who were born in 1983 in The Netherlands with a gestational age <32 weeks at birth?



For answering question 2, we apply the seven steps as follows:

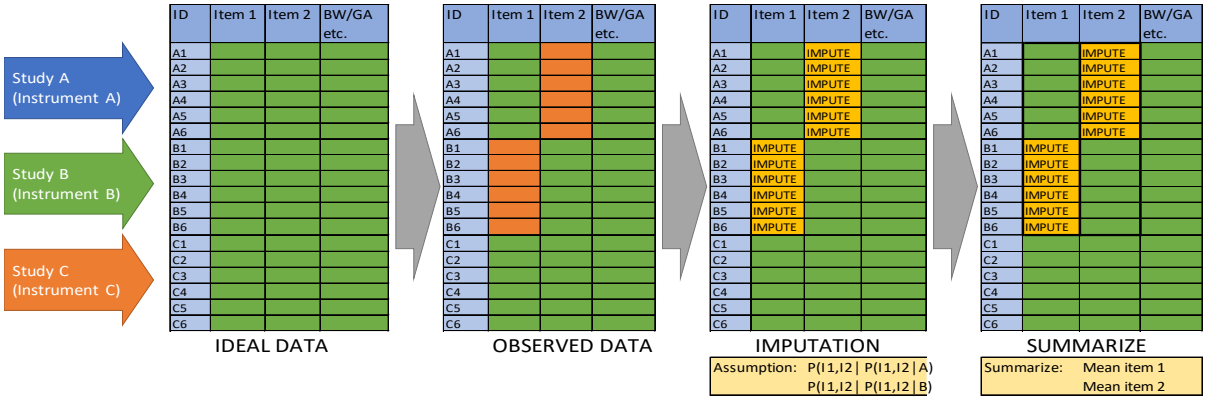
1. Ideal data: We would like to have gestational age, one-week survival, and birth weight and Health-Related Quality of Life at age 19 of all children born the Netherlands in 1983;
2. Ideal analysis: We select all children surviving up to at least one week, and estimate the outcome distribution at 19y per GA week, with stabilization across GA using the LMS method;
3. Available data: The POPS data. Full cohort of VLWB/VPT from in The Netherlands in the year 1983. In principle, the coverage of the population of interest in complete;
4. Missing data: 67 children between 1y-19y died. Missing outcomes at 19y are related to health and well-being;
5. Replications: In the ideal data, plausible imputations should depend on (previous) health and other factors that inform participation at 19; we need to decide on whether to impute outcomes for the 67 deceased children (both are defensible);
6. Calculate: for each replication, perform the analysis specified in step 2;
7. Summarize: aggregate the results from step 6; estimate the uncertainty introduced by the missing data.

3.4 Example 3

Suppose we want to answer to following research question:

Is the distribution of Health-Related Quality of Life at 19 years by gestational age, different for two population of preterms, e.g., in NL and UK?

The complication in answering the above question stems from the fact that the NL (study A) and UK (study B) may have used *different instruments* (instrument A or B) for measuring QoL at 19y, so comparison of the distribution of QoL in their original scales is not sensible. We may however attempt to transform the responses onto a common scale, and do the comparison on that scale. The common scale could be the UK-scale (so the NL data needs to be transformed), the NL-scale (so the UK data needs to be transformed), or a third scale (so both NL and UK data need to be transformed). In the scenario below, we assume that we have access to a third study (study C) that collected measurements in both NL and UK instruments, thus both instrument C = {A, B}.

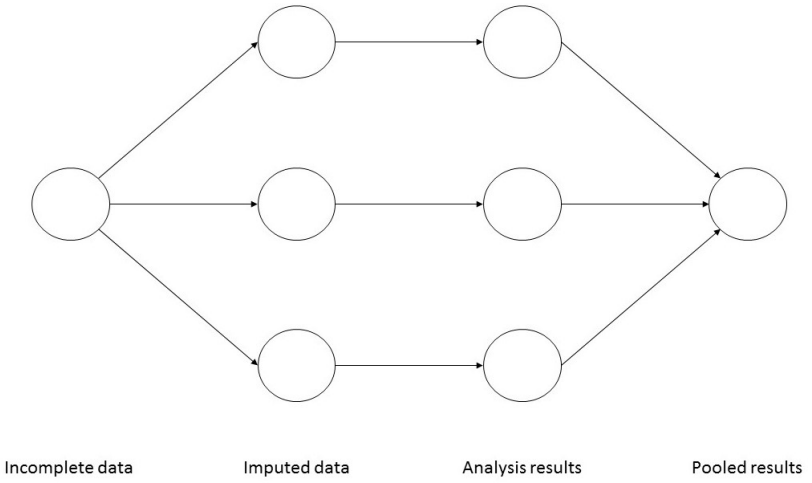


For answering question 3, we apply the seven steps as follows:

1. Ideal Data: In additional to common variates, we would like to have data from all subjects from study A and B measured on either instrument A or instrument B (or both);
2. Ideal Analysis: Linear regression analysis using participants in study A and B, with measurement A (or B) as outcomes, and gestational age and study as predictors. Test for study effect;
3. Available Data: Study A measures QoL using instrument A, study B measures QoL using instrument B. There is also a third study (C) that measured both A and B;
4. Missing data: The reason for the missingness is that different instruments were chosen in studies A and B;
5. Replications: Assume that the relation between A and B in study A and study B is the same as the relation between A and B in study C (MAR assumption). Find plausible values of A (given B) in study B, and of B (given A) in study A;
6. Calculate: Perform regression analysis as specified in step 2 to each replication;
7. Summarize: Aggregate the estimates of the study effect from the replicates. Derive appropriate measure of uncertainty caused by the transformation to the common scale.

4 A GENERIC SOLUTION: MULTIPLE IMPUTATION

The seven steps as described in section 3 have strong theoretical underpinning. Each of the seven steps has an analogue within the theory of multiple imputation, a statistical method developed by Rubin (1987) for solving missing data problems.



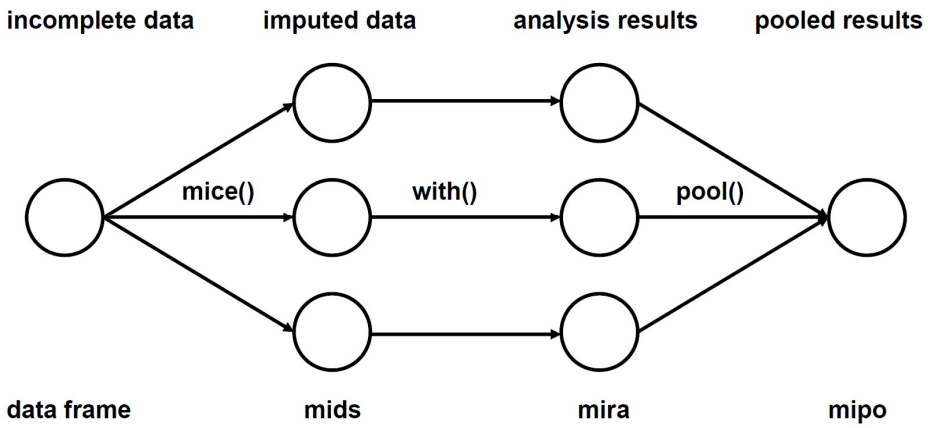
Starting from one incomplete data with missing data, multiple imputation creates several ($m > 1$) complete datasets, where the missing values are replaced by plausible values. These plausible values are drawn from a distribution specifically modeled for each missing entry based on the relations in the observed data. Each of these datasets is analyzed using standard software, and the m results are then pooled in to a final point estimate. A schematic visualization of this process with $m = 3$ is given in the figure. Normally a larger value for m is chosen, however $m = 3$ allows for a nice visualization of the process.

The variation of the imputed data in a given cell tells us something about the uncertainty about the imputed value: the bigger the variation the more uncertain we are. Then when each of these 'new' datasets is analyzed, the result are likely to be different, perhaps slightly, which is the result of our uncertainty about the true (but missing) value. The final step aggregates the m estimates into one combined estimate, and calculates its associated variance. This variance not only takes the conventional uncertainty caused by sampling in account, but also uncertainty caused by the missing data. Rubin has shown that under general conditions the pooled estimates are unbiased and have correct statistical properties.

The number of imputations (m) can be taken quite small. Usual values are m = 5 or m = 10. If many cells are missing, we may need to set m larger, e.g., m = 50 or m = 100.

The MICE package (Van Buuren & Oudshoorn, 2011; Van Buuren 2012) implements the theory of Rubin for multivariate missing data, i.e., where the missing data can appear anywhere in the data. The most important functions in the package, mice(), with() and pool(), corresponds to the major conceptual steps in Rubin's theory.

Steps in the R-package 'mice'



5 DISCUSSION

Bringing together data from different sources with IPD is very useful. The availability of combined data opens a myriad of analytic possibilities, but also pose challenges, both on the analytics as well as on harmonization. Data combination sounds easier than it is. There are many practical hurdles that we need to overcome. Moreover, our decisions may have consequences on the conclusions drawn from the combined data.

The report proposes a seven-step approach derived from the missing-data perspective. In a nutshell: visualize what the ideal data look like, and imagine how the analysis is to be done. Once this is done, assess how far you are from the ideal, try to build realizations of how the ideal data could have looked like, perform your ideal analysis, and aggregate. The seven-step approach conforms closely to the concept of multiple imputation. Where properly executed, the resulting estimates from the seven step are expected to have good statistical properties.

When is our approach likely to be useful? Many existing statistical techniques (e.g. estimating mean of a population) can be derived from a missing-data perspective, but the algorithms and recipes produced in this way are likely to be less efficient than the prevalent technique. The reason is that it is often

possible to cut corners, make simplifying assumptions, use special properties of the data collection design, and so on. In contrast, our approach is more generic, and meant for cases where it is not yet obvious how the solution can be formulated, or useful in situations where uncertainty is not properly quantified.

A disadvantage of the seven-step approach is that it is more work, especially for creating the replications and repeating the analysis on each replication. On the other hand, as computers become more capably, such issues are likely to be less of an issue over time. The missing-data perspective also has relatively little to offer for purely exploratory analysis and data discovery.

On the other hand, the missing-data perspective is a principled approach, capable of producing trustworthy estimates of uncertainty caused by sampling and missing data on quantities of scientific interest. In order to work well, the investigator should have a reasonable idea of the ideal data would be, and how these should be analyzed.

Data protection is an important requirement because of privacy concern. However, data protection is also misused as an argument to restrict bona fida access to data for scientific purposes. Rubin (1993) proposed to replace all observed data by a set of synthetic data sets. These data sets cannot be traced back to any individual (since all data are synthetic), but still contain the relations that are of scientific interest. In this way, the missing-data perspective may also provide useful for improving access to source data.

6 LITERATURE

1. Harron K et al (2016). Methodological developments in data linkage.
2. Vd Vijver and Kok Leung (1997), Method and data analysis for cross-cultural studies.
3. Van Deth, JW (1998). Comparative politics. The problem of equivalence.
4. Van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2005). Improving Comparability of Existing Data by Response Conversion. Journal of Official Statistics, 21, 53-72.
5. Kunkel, D., & Kaizar, E. E. (2017). A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Statistics in Medicine. doi:10.1002/sim.7388
6. Debray, T., Moons, K. G. M., Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., & Reitsma, J. B. (2015). Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Research synthesis methods, 6(4), 293-309. doi:10.1002/jrsm.1160.
7. Gelman, A., & Meng, X.-L. (2004). Applied Bayesian modeling and causal inference from incomplete-data perspectives. John Wiley & Sons.
8. Little, R. J. (2013). In Praise of Simplicity not Mathematistry! Ten Simple Powerful Ideas for the Statistical Scientist. Journal of the American Statistical Association, 108(502), 359-369.
9. Hille E T M, Weisglas-Kuperus N, Goudoever van J B, Jacobusse G W, Ens-Dokkum M H, Groot de L, Wit J M, Geven W B, Kok J H, Kleine de M J K, Kollée L A A, Mulder A L M, Straaten van H L M, Vries de L S, Weissenbruch van M M, Verloove-Vanhorick S P,Dutch POPS-19 Collaborative Study Group. Functional Outcomes and Participation in Young Adulthood for Very Premature and Very Low Birth Weight Infants: the Dutch POPS-study at 19 years of age. Pediatrics 2007; 120 (3); e587-e595
10. van Lunenburg A, van der Pal SM, van Dommelen P, van der Pal-de Bruin KM, Bennebroek Gravenhorst J, Verrips GH. Changes in quality of life into adulthood after very preterm birth and/or very low birth weight in the Netherlands. Health Qual Life Outcomes. 2013:26;11:51. DOI: 10.1186/1477-7525-11-51.
11. Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468.
12. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.
13. van Buuren, S. (2012). Flexible Imputation of Missing Data. Boca Raton, FL: Chapman & Hall/CRC Press.