Dealing with missing data in randomized and cluster randomized trials

Chapter 7 - Summary

Dealing with missing data is a prominent problem in randomized controlled trials (RCTs) and cluster randomized trials (CRTs). Indeed, statistical analysis of RCTs and CRTs should account for missing data, since not doing so may lead to biased and/or inefficient results (White and Thompson 2005; Kayembe et al. 2020). Further, statistical analysis of CRTs should also account for clustering of the data, since not doing so in data analysis and in sample size calculation leads respectively to underestimation of the sampling variance of the treatment effect and to underpowered trials even after accounting for missing data (Fiero et al. 2016; Turner et al., 2020). In this thesis, we explore statistical methods for dealing with missing data in RCTs and CRTs by comparing the performance of various missing data methods under various missingness scenarios, in terms of estimation of the treatment effect and its standard error (SE), which are converted into various performance criteria.

In chapter 1, we introduce the notion of RCTs and CRTs, describe the problem of missing data in both studies and why this should appropriately be dealt with in statistical analysis, and provide a non-technical review on the statistical analysis models for RCTs and CRTs, the various missing-data mechanisms, and different methods that are available for dealing with missing data under various missingness mechanisms assumed in this thesis. Chapter 1 is concluded with an outline of the present thesis.

In chapter 2, we first review the literature on dealing with missing values on a covariate in randomized studies and summarize what had been done and what was lacking in the literature on this topic before the beginning of this thesis. We then investigate the situation with a continuous outcome and a missing binary covariate in more details through simulations, comparing the performance of multiple imputation (MI) with various simple

alternative methods. This is finally extended to the case of time-to-event outcome. The simulations consider five different missingness scenarios: missing completely at random (MCAR), at random (MAR) with missingness depending only on the treatment, and missing not at random (MNAR) with missingness depending on the covariate itself (MNAR1), missingness depending on both the treatment and covariate (MNAR2), and missingness depending on the treatment, covariate and their interaction (MNAR3). Here, we distinguish two different cases: (1) when the covariate is measured before randomization (best practice), where only MCAR and MNAR1 are plausible, and (2) when it is measured after randomization but before treatment (which sometimes occurs in nonpharmaceutical research), where the other three missingness mechanisms can also occur. The proposed methods are compared based on the treatment effect estimate and its standard error. The simulation results suggest that the patterns of results are very similar for all missingness scenarios in case (1) and also in case (2) except for MNAR3. Furthermore, in each scenario for continuous outcome, there is at least one simple method that performs at least as well as MI, while for time-to-event outcome MI is best.

In chapter 3, we extend the problem of handling missingness on a single covariate in RCTs considered in chapter 2 to the situation of missingness on multiple baseline covariates in RCTs. Specifically, we evaluate the performance of sophisticated methods like MI and maximum likelihood (ML)-based methods compared with simple alternative methods under various missingness scenarios in RCTs with a quantitative outcome, in terms of bias and efficiency of treatment effect estimation. We first derive asymptotic relative efficiencies of the simple methods under the missing completely at random (MCAR) scenario and then perform a simulation study for non-MCAR scenarios. Finally, a trial on chronic low back pain is used to illustrate the implementation of the methods. The results show that all simple methods give unbiased treatment effect estimation, but with increased mean squared residual.

It also turns out that mean imputation and the missing-indicator method are most efficient under all covariate missingness scenarios and perform at least as well as MI and LM in each scenario.

Chapter 4 compares different methods for handling missing data in randomized controlled trials (RCTs) with a continuous outcome, focusing on the case of joint missingness in the outcome and one or more covariates. For this case, it was not yet obvious in the literature how advanced missing data methods like the linear mixed model (LMM) and multiple imputation (MI) perform in comparison with simple methods with respect to the estimation of the treatment effect and its standard error (SE) for various realistic missingness mechanisms. This chapter therefore compares the performance of LMM and MI with simple alternative methods through a wide range of simulation scenarios for various plausible missingness mechanisms. This comparison is made with respect to five performance criteria, namely the bias of the treatment effect estimate, coverage of the 95% confidence interval (CI), mean squared error (MSE), false positive rate (FPR) and power. The results show that no missing data method is universally superior, but LMM followed by MI has a better performance under most missingness scenarios. Complete case analysis (CCA) performs similarly to these advanced methods only when the missingness rate is low (10%). More interestingly, a simple method that uses CCA for the outcome missingness combined with mean imputation (ME) for covariate missingness (CCAME) performs similarly to LMM and MI in most simulation scenarios. All methods are furthermore compared in an RCT on chronic obstructive pulmonary disease (COPD).

In chapter 5, we compare the performance of various methods for handling missing data in cluster randomized trials (CRTs) with a baseline and post-test measurement of a quantitative outcome. We specifically consider the setting with missingness in both the pre- and post-test outcome for various plausible missingness scenarios, as this appeared to have

not been given much attention, if any at all, in the literature. The methods compared are based on the 3-level linear mixed effects model (LMM) with repeated measures nested in persons nested in clusters, and on its counterpart 2-level analysis of covariance (ANCOVA) with the baseline measurement as a covariate and allowing for a contextual effect. The LMM methods handle the missing values during the analysis by the direct likelihood method, whereas the ANCOVA methods do so prior to the analysis by leaving out cases or imputing missing values. The comparison is made through extensive simulation with respect to bias, coverage, false positive rate and power for the treatment effect. The results show that the best performing ANCOVA methods, which use complete-case analysis (CCA) for outcome missingness combined with mean imputation overall (ME) or per cluster (MEC) for covariate missingness, are generally superior to the LMM based methods as implemented in R, because they give smaller bias for SE's of the treatment effect estimate, resulting in better coverage and false positive rate (FPR). All methods are furthermore compared in a CRT on mental health among primary school pupils.

In chapter 6, we discuss the findings and conclusions of the investigations conducted in this thesis. The conclusions are given in terms of practical recommendations regarding the choice of methods to be used for appropriately handling missing data in RCTs and CRTs depending on the scenarios considered in this thesis. Furthermore, we highlight the limitations of this thesis and provide possible topics for future work.