Contemporary Clinical Trials Communications 11 (2018) 20–29

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Comparison of intent-to-treat analysis strategies for pre-post studies with loss to follow-up

Wenna Xi[a], Michael L. Pennell[a,*], Rebecca R. Andridge[a], Electra D. Paskett[b]

[a] Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH 43210, United States
[b] Department of Internal Medicine, College of Public Health, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, United States

ARTICLE INFO

Keywords:
ANCOVA
Missing data
Mixed model
Multiple imputation
Randomized trial

ABSTRACT

In pre-post studies when all outcomes are completely observed, previous studies have shown that analysis of covariance (ANCOVA) is more powerful than a change-score analysis in testing the treatment effect. However, there have been few studies comparing power under missing post-test values. This paper was motivated by the Behavior and Exercise for Physical Health Intervention (BePHIT) Study, a pre-post study designed to compare two interventions on postmenopausal women's walk time. The goal of this study was to compare the power of two methods which adhere to the intent-to-treat (ITT) principle when post-test data are missing: ANCOVA after multiple imputation (MI) and the mixed model applied to all-available data (AA). We also compared the two ITT analysis strategies to two methods which do not adhere to ITT principles: complete-case (CC) ANCOVA and the CC mixed model. Comparisons were made through analyses of the BePHIT data and simulation studies conducted under various sample sizes, missingness rates, and missingness scenarios. In the analysis of the BePHIT data, ANCOVA after MI had the smallest *p*-value for the test of the treatment effect of the four methods. Simulation results demonstrated that the AA mixed model was usually more powerful than ANCOVA after MI. The power of ANCOVA after MI dropped the fastest as the missingness rate increased; in most simulated scenarios, ANCOVA after MI had the smallest power when 50% of the post-test outcomes were missing.

1. Introduction

In a pre-post study a treatment is evaluated by measuring responses both before and after the study for each participant in a treatment group and a control group. Pre-post study designs have been widely used in clinical trials, psychology, education, and sociology. For example, our research was motivated by the Behavior and Exercise for Physical Health Intervention (BePHIT) Study, a pre-post study designed to compare two interventions intended to promote walking in postmenopausal women [1].

When there is complete follow-up, previous studies have shown that, in terms of testing the treatment effect, analysis of covariance (ANCOVA) is more powerful than a comparison of change scores [2–5]. However, in reality, missing data, in particular loss to follow-up, is very common in pre-post studies. For instance, in the BePHIT study, 17% of the participants did not finish the study. With unbalanced sample sizes for pre- and post-test levels in each treatment group, a regular ANCOVA or change score analysis cannot be conducted without dropping any subjects. Therefore the most straightforward method for handling missing values is to exclude all the subjects with missing data. This type of analysis is called the complete-case (CC) analysis. The CC analysis is usually not recommended, since it throws away information collected in the study and does not follow the intent-to-treat (ITT) principle for clinical trials [6,7]. Nowadays, one popular way to handle missing data is multiple imputation (MI) [6]. For instance, in pre-post studies, missing follow-ups can be simulated multiple times using the baseline outcome value and measured covariates and the results of the analysis of each complete data set are combined to account for the uncertainty introduced by the imputations [8]. Another approach often used for data with repeated measures is the mixed model, where all available pre- and post-test values are regressed over treatment and timepoint indicators, assuming some variance-covariance structure for the repeated measures.

The main goal of this study was to compare the power of two analysis methods which adhere to ITT principles: the mixed model and ANCOVA after MI for pre-post studies when missing post-test is present. We also wanted to compare these methods to two methods that do not adhere to an ITT principle: ANCOVA and the mixed model using only completely observed cases. These methods were first compared in the context of our motivating example (BePHIT) and then in simulation

[*] Corresponding author.
E-mail address: pennell.28@osu.edu (M.L. Pennell).

https://doi.org/10.1016/j.conctc.2018.05.008
Received 29 November 2017; Received in revised form 11 April 2018; Accepted 3 May 2018
Available online 09 May 2018

studies based on the BePHIT data. A parallel set of simulations comparing the type I error rate of the four methods were conducted as well.

2. Motivating example

The Behavior and Exercise for Physical Health Intervention (BePHIT) Study was a randomized controlled study of a 12-week walking intervention conducted on postmenopausal women between January 2008 and March 2009 [1]. The primary outcome was the change in time for women to finish a one-mile walk. In addition to one-mile walk time, anthropometric and psychometric measures were obtained at pre- and post-test.

After passing the selection criteria, 71 participants were stratified by BMI and randomized into either a coach group or a no-coach group. For women in the coach group, a trained coach was assigned. The role of the coach was to explain the intervention, provide the first week's steps goal, train subjects to use a pedometer and the Interactive Voice Response (IVR) system to collect data, and offer help during the intervention. Women in the no-coach group received similar instructions, training, and help, except that they were not informed that they had access to a coach. Although both groups received a treatment, to be consistent with the terminology in this paper, we will refer to the coach condition as the treatment and the no coach condition as the control.

Among the 71 randomized participants, 35 were assigned to the treatment group and 36 to the control group. For the control group, baseline walking time was only available for 35 patients. In total, 12 (17%) patients dropped out before the post walking test, 4 of whom were in the treatment group and 8 in the control group. The drop out rate did not differ significantly across the two groups ($p = 0.20$). The original study reported 2 withdrew and 12 did not complete in the treatment group, and 7 withdrew and 11 did not complete in the control group [1]. In that study, "completed" was defined as completing all post-test assessments within 30 days after the end of the walking intervention. These post-test assessments included the walk test and anthropometric and psychometric measures not considered in this paper. Those who had their post-test score recorded but did not finish all their anthropometric and psychometric measures were also included in this analysis, which led to fewer dropping out here.

3. Analysis methods for pre-post studies with complete data and missing data

3.1. Pre-post studies

A pre-post study is a randomized controlled study where outcome values are measured both before and after the study. As opposed to treatment-control studies where the outcome variable is only measured once, pre-post studies allow investigators to account for the level of the outcome variable before the treatment is applied. Different from a one-group pre-post design, a treatment-control pre-post study controls for secular trends [9,10]. In the BePHIT study, for instance, besides the intervention, the improvement of women's one-mile walk time may have been caused by some other factors, such as a national walking campaign, affecting the women during the same time period. A one-group pre-post study fails to consider these factors; however a treatment-control pre-post study accounts for secular trends by comparing the results from the treatment group to a control group observed over the same period of time.

If we let $\mathfrak{R}$ be the randomization process, $\mathfrak{T}$ be the treatment process, and $(Y_{\text{pre,t}}, Y_{\text{post,t}})$ and $(Y_{\text{pre,c}}, Y_{\text{post,c}})$ be the pre- and post-test measure of a treated and control participant, respectively, then a pre-post study design can be illustrated by the following:

$$\begin{tikzcd} \mathfrak{R} \arrow[r] & Y_{\text{pre,t}} \arrow[r, "\mathfrak{T}"] & Y_{\text{post,t}} \\ \mathfrak{R} \arrow[r] & Y_{\text{pre,c}} \arrow[r] & Y_{\text{post,c}} \end{tikzcd}$$

3.2. Analysis of pre-post studies with complete data

Many analysis approaches for pre-post studies have been discussed [2,3,5,9,11,12]. Arguably the two most common analysis methods are the change score analysis and Analysis of Covariance (ANCOVA) [2]. We discuss these two methods and their statistical power in the following.

3.2.1. Change score analysis

A change score analysis first obtains the difference in outcome values before and after the experiment, and then regresses the difference on the treatment assignment using the following model:

$$Y_i - X_i = \alpha_{C0} + \alpha_{C1} T_i + \varepsilon_i, \tag{1}$$

where Y_i is the post-test outcome level for subject i, X_i is the pre-test outcome level for subject i, T_i is the indicator variable for treatment assignment, and ε_i is the error term for subject i ($\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$). Note that $Y_i - X_i$ is the change score for subject i during the experiment. In the model above, α_{C1} quantifies the effect of treatment assignment on change in outcome level from pre to post. Since T_i is a binary variable, the change score test is equivalent to a two-sample t-test comparing the mean of $Y_i - X_i$ between treatment and control groups.

3.2.2. ANCOVA

Unlike the change score method, in ANCOVA, post-test value (Y_i) is treated as the outcome variable and pre-test value (X_i) is treated as a predictor. The ANCOVA model can be expressed as:

$$Y_i = \alpha_{A0} + \alpha_{A1} T_i + \alpha_{A2} X_i + \varepsilon_i, \tag{2}$$

where T_i and ε_i are as defined in the change-score model. ANCOVA assumes that pre-test values are measured without error [5]. This assumption holds for variables, such as weight and height, that can be measured precisely. However, it is often violated for self-reported measurements and educational or psychological tests. In the model above, α_{A1} is the effect of treatment assignment on the post-test scores adjusting for the pre-test scores.

The ANCOVA model (2) can also be written as

$$Y_i - X_i = \alpha_{A0} + \alpha_{A1} T_i + \alpha_{A2}^* X_i + \varepsilon_i, \tag{3}$$

where $\alpha_{A2}^* = \alpha_{A2} - 1$ from (2) [2,5]. Thus, ANCOVA can be viewed as an extension of the change score model (1) to include pre-test level X_i as a predictor.

3.2.3. Power comparison: change score analysis vs. ANCOVA

Oakes and Feldman compared the detectable treatment effects of the change score and ANCOVA models [5]. Assume

1. $Var(X_i) = Var(Y_i) = \sigma^2$ regardless of the experimental group,
2. Number of subjects in each group is the same, and
3. Pre-test is measured without error.

Under the assumption of normally distributed errors, the detectable treatment effect for the change score analysis at type-I and type-II error rates of α and β is

$$\Delta_C = \sqrt{\frac{4\sigma^2(1-\rho)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{m}},$$

and the detectable treatment effect for ANCOVA is

$$\Delta_A = \sqrt{\frac{2\sigma^2(1-\rho^2)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{m}},$$

where $\rho = Corr(X_i, Y_i)$, Z_x is the xth quantile of the standard normal distribution, and m is the number of subjects in each experimental group. Therefore we have

$$\frac{\Delta_C}{\Delta_A} = \sqrt{\frac{2}{1+\rho}}. \tag{4}$$

Assuming $0 \le \rho \le 1$,

$$1 \le \frac{\Delta_C}{\Delta_A} \le \sqrt{2}.$$

Thus, with complete data, the detectable treatment effect for ANCOVA is always less than or equal to the detectable treatment effect for the change score analysis. Also, from (4) we can see that the discrepancy in power increases as the correlation between pre- and post-test decreases.

3.3. Missing data

Both the change score analysis and ANCOVA require complete follow-up of subjects. Unfortunately, missing post-test data are very common in pre-post studies since participants may drop out because they moved, are unsatisfied with their performance in the study, etc. For instance, in the BePHIT study, 17% of the patients dropped out prior to study completion.

Post-test data could be missing completely at random, at random, or not at random [6]. Missing completely at random (MCAR) means the missing data does not depend on anything in the data collecting process (e.g., data missing due to a data entry error). We say a study has data missing at random (MAR) when the missingness only depends on observed data, but not those data that are missing (or would have been observed). A subject being removed from a study due to an unhealthy pre-test is an example of MAR. When MAR does not hold, we say that data are not missing at random (NMAR). For example, data would be NMAR if subjects drop out because they know their post-test outcome will be poor.

3.3.1. Statistical methods for missing data

Commonly used methods for handling missing data include: complete-case (CC) analysis, weighting adjustments, imputation, and model-based methods [6,13,14]. We focus in this paper on three natural approaches to handling loss to follow-up in a pre-post study. First, one can remove the data for subjects with post-test missing and carry out a complete-case (CC) analysis. Second, the change score model can be generalized to a mixed model, in which unequal sample sizes for pre- and post-test levels is allowed. Finally, ANCOVA could be applied following imputation of the missing data. These three approaches are summarized below.

3.3.2. Complete-case (CC) analysis

The most straightforward way to deal with missing data is to delete all subjects with any missing observations, no matter if data were partially collected or not. The dataset obtained after deleting all subjects with missing values is called a complete-case dataset and the analysis of these data is called a complete-case (CC) analysis. The CC analysis is simple to conduct and unbiased under MCAR. In certain situations, it is also unbiased under MAR. For instance, when estimating the regression model of Y on $X_1, X_2, \ldots, X_p$ where Y is incompletely observed, the estimation conditions on the values of the X's. Thus, the CC analysis is unbiased if the missingness only depends on the X's but not the Y [6,14]. As a result, the CC ANCOVA model is unbiased if the missingness of post-test values only depends on the pre-test values.

In addition to the possibility of bias, another disadvantage of the CC analysis is that information is thrown away by deleting subjects. Estimation based on a CC analysis may result in larger variances than methods for incomplete data. Also, since CC analysis does not include all subjects randomized in the final analyses, it does not adhere to the intent-to-treat (ITT) principle of clinical trials [7]. It is for these reasons that CC analyses are usually not employed in intervention studies.

3.3.3. Mixed models

The mixed model is a regression model that includes both population-level and subject-level effects. It assumes responses are MAR. For a pre-post study with one treatment group and one control group, a mixed model can be written as:

$$Y_{ij} = \alpha_{M0} + \alpha_{M1}T_i + \alpha_{M2}P_j + \alpha_{M3}T_i \times P_j + b_i + \varepsilon_{ij}. \tag{5}$$

Here, Y_{ij} is the response of subject i at time point j ($j = 1,2$), T_i and P_j are indicators of treatment group and post-test, respectively, and b_i is a random subject effect ($b_i \overset{iid}{\sim} N(0, \sigma_b^2)$), which is independent of the random error ε_{ij} ($\varepsilon_{ij} \overset{\text{indep}}{\sim} N\left(0, \sigma_{\varepsilon j}^2\right)$).

Note that if we subtract the pre- and post-test outcomes we get

$$Y_{i2} - Y_{i1} = \alpha_{M2} + \alpha_{M3}T_i + \tilde{\varepsilon}_i,$$

where $\tilde{\varepsilon}_i = \varepsilon_{i_2} - \varepsilon_{i_1}$. Thus, when there is no missing data, the maximum likelihood estimate of α_{M3} is equivalent to the treatment effect (α_{C1}) in the change score model (1). However, unlike the change score model, the mixed model can include data on subjects with just one of the two outcome values.

An alternative formulation of the mixed model may also be used to analyze pre-post data:

$$Y_{ij} = \alpha_{M0} + \alpha_{M1}T_i + \alpha_{M2}P_j + \alpha_{M3}T_i \times P_j + e_{ij}, \tag{6}$$

where $e_{i1} \overset{iid}{\sim} N(0, \sigma_{e1}^2)$, $e_{i2} \overset{iid}{\sim} N(0, \sigma_{e2}^2)$, $\mathrm{Cov}(e_{i1}, e_{i2}) = \tau_{e12}$, and $\mathrm{Cov}(e_{i1}, e_{i'2}) = 0$ for $i \neq i'$. The major practical difference between (5) and (6) is that (5) assumes a positive correlation between pre- and post-measures (which is expected in a pre-post study) while (6) permits positive and negative correlations. Expression (6) is often referred to as a marginal formulation of a mixed model since the regression function does not include subject-specific effects.

Sullivan et al. consider a different mixed model for studies with missing follow-up in which post-test and an auxiliary outcome (potentially the main outcome measured at an earlier time point following randomization) are modeled using a linear mixed model containing fixed effects of treatment and a baseline covariate (potentially the pre-test value of the outcome). Though a useful approach in some settings, the Sullivan et al. model cannot be applied in studies with a single post-test and loss-to-follow-up and hence will not be considered further.

3.3.4. Multiple imputation (MI)

The general idea of multiple imputation is to fill in the missing data several times and obtain several “complete” data sets, then conduct the statistical inferences based on those completed data sets [6,15]. Compared to single imputation, multiple imputation accounts for variability in the estimate for the missing value.

The general procedure for MI can be summarized as three steps:

Step 1: Fill in the missing data N times and get N complete data sets.
Step 2: Analyze each of the N data sets separately using analysis methods for complete data.
Step 3: Use Rubin's rules to combine the analysis results for the N data sets and obtain the final result [16].

For pre-post data with only post-test values Y_{i2} missing, a regression model of Y_{i2} on pre-test (Y_{i1}) and some selected covariates is first estimated using all observations with observed values of Y_{i2} [17]:

$$Y_{i2} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_{p-2}X_{p-2} + b_{p-1}Y_{i1}.$$

At imputation k ($k = 1, \ldots, N$), new coefficients $\boldsymbol{b}_k^{(*)} = (b_0, b_1, \ldots, b_{p-1})^T$ are sampled from the posterior predictive distribution $(\boldsymbol{b}_k | Y_{\text{obs}})$ and then the missing value Y_{i2} is sampled from $(Y_{i2_k} | \boldsymbol{b}_k^{(*)}, \boldsymbol{X}_i, Y_{i1})$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i,p-2})^T$ are the values of the covariates of subject i in the imputation model.

Consider a linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is a length n vector of outcomes, $\mathbf{X}$ is an $n \times p$ covariate matrix, $\boldsymbol{\beta}$ is a length p vector of regression parameters, and $\boldsymbol{\varepsilon}$ is a length n vector of errors ($\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$). Assume the outcomes $\boldsymbol{Y}$ are partially observed, the covariates $\mathbf{X}$ are completely observed, the missing data mechanism in $\boldsymbol{Y}$ is MAR, and that σ^2 is known. Suppose we are interested in estimating $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_{\text{CC}}$ and $\hat{\boldsymbol{\beta}}_{\text{MI}}$ be the complete-case (CC) ML estimator and MI estimator of $\boldsymbol{\beta}$, respectively. It has been shown that [18]:

$$\begin{aligned} &E(\hat{\boldsymbol{\beta}}_{\text{MI}}) = E(\hat{\boldsymbol{\beta}}_{\text{CC}}) = \boldsymbol{\beta}, \\ &Var(\hat{\boldsymbol{\beta}}_{\text{MI}}) = Var(\hat{\boldsymbol{\beta}}_{\text{CC}}) + \tfrac{1}{N}[Var(\hat{\boldsymbol{\beta}}_{\text{CC}}) - \sigma^2(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}]. \end{aligned} \tag{7}$$

Thus, both the CC ML and MI estimates are unbiased, but their variances differ. Since $Var(\hat{\boldsymbol{\beta}}_{\text{CC}}) - \sigma^2(\mathbf{X}^{\text{T}}\mathbf{X})^{-1}$ is positive-definite, we see that the variance of the MI estimator is always larger than the variance of the ML estimator using only complete cases. The above results can be generalized to unknown σ^2 [18]. Thus, when analyzing pre-post data using the ANCOVA model, estimates from both CC and MI analyses are unbiased, but the standard errors from the CC analysis should be smaller.

4. Analysis of the BePHIT data

4.1. Analyses

Logistic regression was used to determine whether or not dropping out was related to any pre-test measures. The estimated odds ratios (ORs) from univariate logistic models and their p-values are listed in Table 1. Only waist-hip ratio was associated with missing post-test ($p = 0.04$). People with larger waist-hip ratio were more likely to drop out of the study. Since waist-hip ratio was related to the dropout rate, the missing data mechanism was not MCAR.

David et al. used linear mixed models to analyze the pre-post changes [1]. Here we analyzed the data using a mixed model (with all available data), ANCOVA after MI, and the corresponding CC analyses (the mixed model with CC and ANCOVA with CC). Results from each method were compared. The ANCOVA model is valid here since the primary outcome, time for one-mile walk, can be assumed to be measured without error. For both the all-available-data mixed model and ANCOVA after MI, all 70 participants with a baseline walking time were included. One subject in the control group was excluded due to missing both pre- and post-test scores. For the two CC analyses, only participants with both baseline and post-intervention walking time were used ($n = 58$).

We used the marginal formulation of the mixed model (6) and the Kenward-Roger method was used for computing denominator degrees of freedom [19–21]. In MI, post-test values were imputed using a model that included the 12 baseline measures listed in Table 1, including treatment. To preserve the reproducibility of the imputation, missing values were imputed 20 times based on a rule of thumb proposed by White et al. [15]. Imputed data were analyzed using the ANCOVA model (2) and results were combined using the standard MI combining rules [6,15].

Table 1
Logistic regression analysis of drop out in BePHIT.

Variable	Coefficient	SE	Odds Ratio	p-value
Design and Outcome				
Treatment (Coach)	−0.83	0.67	0.44	0.21
Pre-test walk time	0.11	0.16	1.12	0.47
Baseline Anthropometrics				
Pulse rate	0.04	0.03	1.05	0.12
Waist/hip (+ 0.1 units)	1.23	0.59	3.42	0.04
BMI	0.11	0.08	1.12	0.16
Baseline Psychometrics†				
Negative exercise thoughts	0.32	0.52	1.38	0.54
Exercise stage of change	0.21	0.42	1.23	0.62
Social support from family	−0.16	0.48	0.86	0.75
Social support from friends	−0.94	0.71	0.39	0.19
Walking Self-efficacy	−0.17	0.17	0.84	0.32
Self-efficacy to walk 30 min	0.19	0.21	1.21	0.38
Exercise goals	0.12	0.36	1.12	0.74
Exercise planning	−0.36	0.54	0.70	0.50

† Measures under this category are all quantitive scores. More details about these measures can be found in Ref. [1].

Table 2
Regression Estimates for Imputation Model in BePHIT analysis.

Variable	Estimate	SE	p-value
Intercept	−3.07	4.23	0.47
Design and Outcome			
Treatment (Coach)	0.01	0.40	0.98
Pre-test walk time	0.78	0.11	<0.0001
Baseline Anthropometrics			
Pulse rate	−0.01	0.02	0.61
Waist/hip	2.56	4.01	0.52
BMI	0.16	0.05	<0.01
Baseline Psychometrics			
Negative exercise thoughts	−0.28	0.39	0.48
Exercise stage of change	0.15	0.28	0.60
Social support from family	0.29	0.30	0.35
Social support from friends	−0.26	0.36	0.48
Walking Self-efficacy	−0.13	0.14	0.37
Self-efficacy to walk 30 min	0.14	0.14	0.31
Exercise goals	0.29	0.34	0.41
Exercise planning	−0.54	0.55	0.33

The imputation model was fit using the completely observed cases to check whether the variables were predictive of post-test. The estimates from the imputation model are listed in Table 2. The R^2 of the imputation model was 0.6975. However, only two variables, pre-test walk time and BMI, were significant at the 0.05 level.

4.2. Results

Estimates and tests of a treatment effect for the four methods are summarized in Table 3. All four methods showed no coach effect on the change in one-mile walk time. This result is consistent with those reported by David et al. [1]. As we can see, the results from the two mixed models were very similar. However, there were obvious differences in both the estimates and hypothesis tests between the two ANCOVA analyses. When MI was used, the estimated effect for treatment changed from −0.08 to −0.24, and its corresponding standard error changed from 0.38 to 0.63.

ANCOVA after MI gave the most significant test of treatment among all four methods. The treatment effect tests from the two ANCOVA methods were more significant than from the two mixed models. Within each model, the test of treatment effect using all available cases was more significant than using complete cases only.

5. Simulation study

5.1. Data generation

Simulation studies were conducted to compare the power and the

Table 3
Comparison of the BePHIT treatment effects from the four analysis methods.

Method	Variable	Estimate	SE	t	p-value
Mixed Model, CC	Trt. × Post	−0.048	0.38	−0.13	0.8995
Mixed Model, AA	Trt. × Post	−0.049	0.38	−0.13	0.8959
ANCOVA, CC	Treatment	−0.084	0.38	−0.22	0.8251
ANCOVA, MI	Treatment	−0.242	0.63	−0.39	0.7005

Abbreviations: CC, complete-case analysis; AA, all-available-case analysis.

type I error rate of the ITT and complete-case analysis methods. The coefficient values used in the data generation process were, for the most part, based on the results from the BePHIT analysis. The data sets used to compare power were generated from the following model:

$$Y_{ij} = 17.946 - 0.811P_j - 2WH_i - T_i \times P_j + b_i + \varepsilon_{ij}, \tag{8}$$

$i = 1, \ldots, n;\ j = 1{,}2;\ b_i \overset{iid}{\sim} N(0{,}3.069);\ \varepsilon_{ij} \overset{iid}{\sim} N(0{,}1.005)$. Here Y_{ij} is the outcome of subject i at time j and P_j and T_i are indicator variables of post-test and treatment, respectively. The variable WH_i is the waist-hip ratio of subject i, which was generated from $N(0.938,\ 0.121)$ to mimic the distribution in the BePHIT data. Under our data generation model, $\mathrm{Corr}(Y_{i1},\ Y_{i2}) = 0.753$ which was the value in the BePHIT data.

Although the estimated main effect of treatment was not 0 in the BePHIT analysis, it was set to zero in (8) so that the expected outcome value of the two groups was the same at baseline. Also, a main effect of waist-hip ratio was added because waist-hip ratio was significantly related to missing follow-ups in the BePHIT study. Its coefficient was adjusted to -2 so that waist-hip ratio was, on average, significantly associated with the outcome. In the original analyses of the BePHIT study, the interaction effect of treatment and time was not significant. However, in order to conduct power comparisons in the simulation studies, the coefficient for this interaction effect was adjusted to -1 so that, on average, it was significant for all the analysis models.

The model used to generate the data for the type I error rate comparisons was similar to (8), except that the interaction between treatment and time was set to 0.

Two different group sizes were considered: $m = 35$ (i.e., the BePHIT sample size) and $m = 100$ per group ($2m = n$). For each group size, 500 data sets were simulated. Post data were then set to be missing at rates of 20%, 30%, 40%, and 50% missingness. At each percentage of missingness, missing data were generated as missing completely at random (MCAR) and missing at random (MAR). For MAR, missingness was generated under four different conditions: dependent on waist-hip ratio, dependent on both waist-hip ratio and pre-test level, dependent on waist-hip ratio, pre-test level, and treatment, and dependent on waist-hip ratio, pre-test level, treatment, and the interaction between pre-test level and treatment. To generate the missing data, a Bernoulli indicator was drawn for each subject with the probability defined by the following logistic regression model:

$$\begin{aligned}\log - \text{odds of missing follow} - \text{up} = \gamma_0 + \gamma_1 WH_i + \gamma_2 Y_{i1} + \gamma_3 T_i \\ + \gamma_4 (Y_{i1} \times T_i).\end{aligned} \tag{9}$$

The γ's were first estimated using the BePHIT data. However, since the analyses of the drop out rate in the original study showed very different effects of the three main effect terms (Table 1), the coefficients of WH_i and T_i were adjusted so that their effects would be similar to the pre-test level's. The coefficient of WH_i was adjusted by forcing the odds ratio of drop out to be two for a one standard deviation change in waist-hip ratio. The coefficient of T_i was adjusted by forcing the odds ratio of drop out for the treatment group versus the control group to be four. In the last simulated scenario, MAR dependent on waist-hip ratio, pre-test level, treatment, and the interaction between pre-test level and treatment, the coefficient of Y_{i1} was adjusted such that the average effect of Y_{i1} between the two groups was the same as in the previous scenarios, and the coefficient of T_i was adjusted such that the average treatment effect was the same as in the previous scenarios. Our choice of main effects and interaction coefficient resulted in control subjects with smaller Y_{i1} being more likely to drop out and intervention subjects with larger Y_{i1} being more likely to drop out. The γ's were set to 0 when the missingness did not depend on their corresponding variables. Finally, γ_0 was adjusted for each percentage of missingness. Table 4 provides the values of the coefficients for each scenario.

5.2. Analyses

For each scenario mentioned above, simulated data sets were analyzed by both mixed models and ANCOVA after multiple imputation (MI). Since outcome values were generated from a model containing waist-hip ratio, in addition to the mixed model and ANCOVA model mentioned in Section 3, each method was also conducted with waist-hip ratio in the analysis model. For comparison purposes, we also performed complete-case (CC) analyses using the same mixed models and ANCOVA models. The mixed model used for the analysis was:

$$Y_{ij} = \alpha_{M0} + \alpha_{M1} T_i + \alpha_{M2} P_{ij} + \alpha_{M3} T_i \times P_{ij} + \alpha_{M4} WH_i + e_{ij}, \tag{10}$$

where e_{ij} is as defined in (6). Similar to the BePHIT study analysis, the Kenward-Roger method was used for computing denominator degrees of freedom [19–21]. The ANCOVA model used for the analysis was:

$$Y_{i2} = \beta_{A0} + \beta_{A1} Y_{i1} + \beta_{A2} T_i + \beta_{A3} WH_i + \varepsilon_i, \tag{11}$$

where $\varepsilon_i \overset{iid}{\sim} N(0,\ \sigma_\varepsilon^2)$. When waist-hip ratio was not included in the analysis model, α_{M4} in the mixed model (10) and β_{A3} in the ANCOVA model (11) were set to 0. For MI, missing follow-ups were imputed 20 times using the following imputation model:

$$\begin{aligned}Y_{i2} = \beta_{I0} + \beta_{I1} Y_{i1} + \beta_{I2} T_i + \beta_{I3} WH_i + \beta_{I4} Y_{i1} \times T_i + \beta_{I5} Y_{i1} \times WH_i \\ + \beta_{I6} T_i \times WH_i.\end{aligned}$$

That is, the predictors in the imputation model were all the main effects that were used in generating the data and all their two-way interactions. The statistical software used for both data generation and analyses was SAS 9.4 (SAS Inc., Cary, NC). The DATA step was used to generate the data, the REG procedure was used to fit the ANCOVA model, the MI and MIANALYZE procedures were used to do the multiple imputation, and the MIXED procedure was used to fit the mixed model.

5.3. Results

Fig. 1 compares the power of the tests of a treatment effect for each method when waist-hip ratio was included in the analysis models and $m = 35$ subjects per group. Here α was set to 0.05. The power of each method decreased as the percentage of missingness increased. The complete-case (CC) ANCOVA was the best in terms of the power when percentage of missingness was smaller (20% and 30%). When missing proportion became larger (40% and 50%), the all-available-case mixed model yielded similar results to CC ANCOVA. Under MCAR, ANCOVA after multiple imputation (MI) was slightly more powerful than the mixed model using all available data when the percentage of missingness was smaller (20% and 30%). Under MAR, ANCOVA after MI and the mixed model using all available data were still comparable when post data were missing at 20%. However, the power of ANCOVA after MI decreased dramatically as the percentage of missingness increased and for most scenarios was the least powerful method when 50% of the post data were missing. When the true logit model for drop out included an interaction between treatment and pre-test (Y_{i1}), the power of the CC mixed model was considerably lower than all other methods when 20–40% were missing follow-up; when 50% were missing follow-up, the power of the CC mixed model was comparable to the power of ANCOVA after MI. When $m = 100$ per group, the power of each test was very high (mostly >0.9) and CC ANCOVA and the all-available-case mixed model had very similar power under each scenario (Fig. S1 in the Supplementary Materials). Similar trends in power were observed when waist-hip ratio was omitted from the analysis model (Figs. S2 and S3 in the Supplementary Materials).

Coefficient estimates and standard errors of the power simulations are provided in Table S1 – S8 in the Supplementary Materials. For the most part, all four methods produced unbiased estimates. The one exception to this trend was when the CC mixed model was applied to data

Table 4
Coefficients for log-odds of missing follow-up used in the simulation study.

% Missingness	Missing Mechanism	γ_0	γ_1	γ_2	γ_3	γ_4
20	MCAR	−1.386	0	0	0	0
	MAR (WH)	−3.252	1.989	0	0	0
	MAR (WH + Pre)	−4.910	1.989	0.094	0	0
	MAR (WH + Pre + Trt)	−5.603	1.989	0.094	1.386	0
	MAR (WH + Pre + Trt + Pre × Trt)	−3.251	1.989	−0.039	−3.319	0.265
30	MCAR	−0.847	0	0	0	0
	MAR (WH)	−2.713	1.989	0	0	0
	MAR (WH + Pre)	−4.371	1.989	0.094	0	0
	MAR (WH + Pre + Trt)	−5.064	1.989	0.094	1.386	0
	MAR (WH + Pre + Trt + Pre × Trt)	−2.712	1.989	−0.039	−3.319	0.265
40	MCAR	−0.405	0	0	0	0
	MAR (WH)	−2.271	1.989	0	0	0
	MAR (WH + Pre)	−3.929	1.989	0.094	0	0
	MAR (WH + Pre + Trt)	−4.622	1.989	0.094	1.386	0
	MAR (WH + Pre + Trt + Pre × Trt)	−2.270	1.989	−0.039	−3.319	0.265
50	MCAR	0	0	0	0	0
	MAR (WH)	−1.866	1.989	0	0	0
	MAR (WH + Pre)	−3.524	1.989	0.094	0	0
	MAR (WH + Pre + Trt)	−4.217	1.989	0.094	1.386	0
	MAR (WH + Pre + Trt + Pre × Trt)	−1.864	1.989	−0.039	−3.319	0.265

Abbreviations: MAR (WH), missingness dependent on waist-hip ratio; MAR (WH + Pre), missingness dependent on waist-hip ratio and pre-test level; MAR (WH + Pre + Trt), missingness dependent on waist-hip ratio, pre-test level, and treatment assignment; MAR (WH + Pre + Trt + Pre × Trt), missingness dependent on waist-hip ratio, pre-test level, treatment, and the interaction between pre-test level and treatment.

with missingness dependent on an interaction between treatment and pre-test; in that scenario the treatment-time interaction effect (i.e., the treatment effect of interest) was biased (between 6 and 13% across missingess rates). The all-available-case mixed model consistently had smaller standard errors of the estimates than the CC mixed model, while the reverse was true for the ANCOVA analyses due to the relationship defined in (7). Also, the estimates were more accurate and precise at the larger sample size ($m = 100$ per group).

Fig. 2 compares the type I error rate of the tests of a treatment effect for each method when $m = 35$ per group and waist-hip ratio was included in the analysis models. Type-I error rates were similar under MCAR and when missingness only depended on WH, though in the latter scenario the type-I error rate for CC ANCOVA was inflated under 50% missingness. When missingness depended on treatment alone or treatment and pre-test, the mixed models generally exhibited deflated type-I error rates and with the exception of one case (ANCOVA after MI when missingess depended on WH, pre-test, treatment, and the pre-test by treatment interaction) the type-I error rates for the ANCOVA models were close to the nominal level. When there was interaction between treatment and pre-test in the missingness model, type-I error rates were deflated under ANCOVA after MI at the higher missingness rates and were inflated under the mixed models when 50% were missing post-test. Differences in type-I error rate were more variable when $m = 100$ per group with no single method consistently providing the lowest or highest type-I error rates (Fig. S4 in the Supplementary Materials). Similar trends in type-I error rate were observed when WH was excluded from the regression models (Figs. S5 and S6 in the Supplementary Materials).

6. Discussion

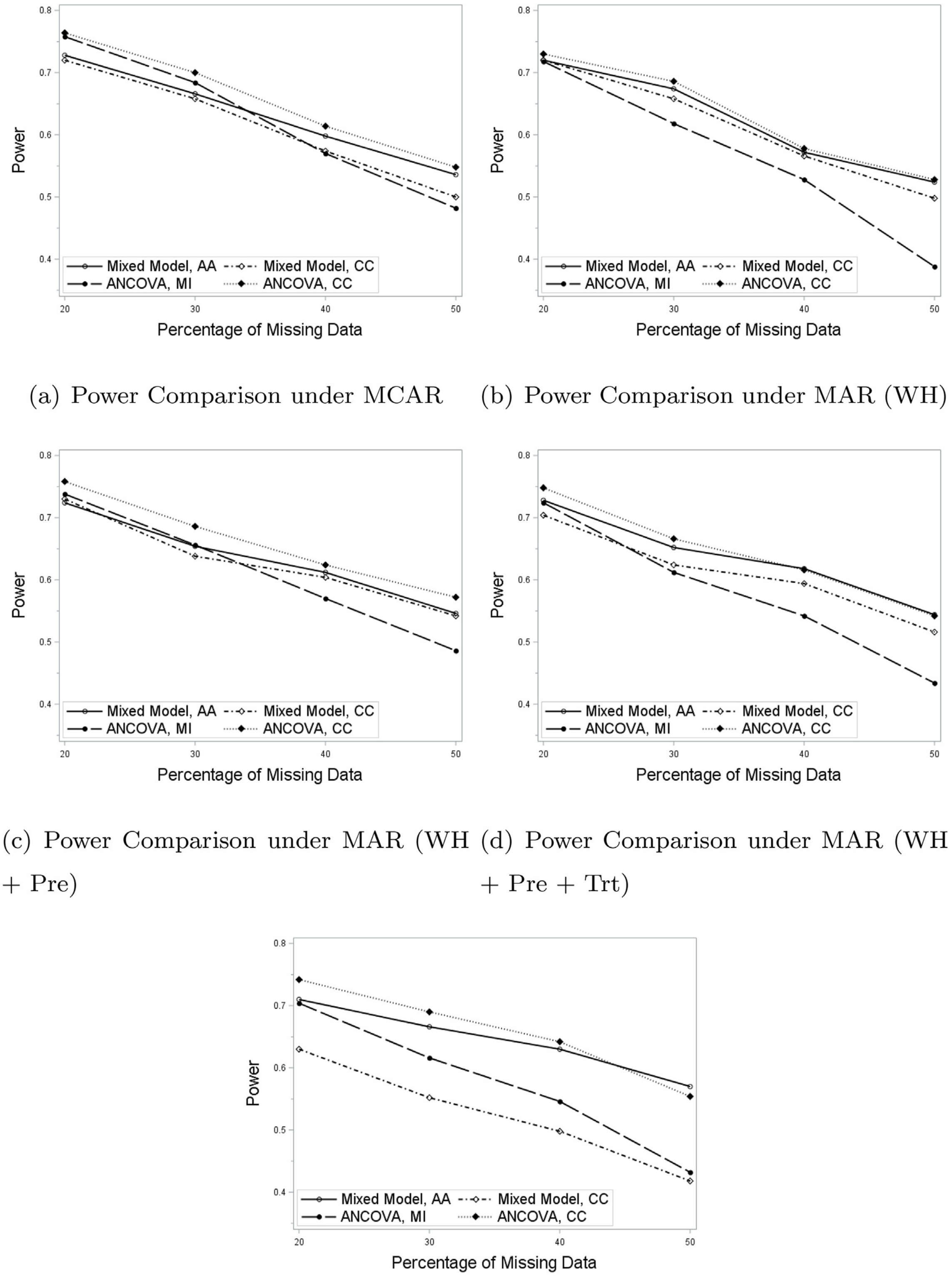
6.1. Summary

In the BePHIT data analyses, p-values were smaller for ANCOVA than for mixed models when testing the treatment effects. Also, for both ANCOVA and mixed models, p-values were smaller for ITT analyses (ANCOVA after MI and the AA mixed model) than CC analyses. However, the treatment effects estimated by ANCOVA after MI and CC ANCOVA were remarkably different (−0.08 for CC vs. −0.24 for MI), as were the standard errors (0.38 for CC vs. 0.63 for MI). It is also interesting that the p-value from CC ANCOVA was smaller than the p-value from the all-available-data mixed model. Thus, for the BePHIT data, the benefit of switching the CC mixed model to the CC ANCOVA model outweighed the benefit of switching the CC mixed model to the all-available-data mixed model.

The simulation studies showed that all four methods, ANCOVA after multiple imputation (MI), ANCOVA for complete cases (CC), the all-available-case mixed model, and the CC mixed model, generally produced unbiased estimates of the treatment effect when waist-hip ratio was included in the analysis model. The only exception was the CC mixed model under MAR dependent on waist-hip ratio, pre-test value, treatment, and the interaction between pre-test value and treatment. In this scenario, the treatment effect was confounded by pre-test in complete case analyses since the presence of an interaction in the missingness model resulted in an inverse association between assignment to treatment and pre-test. When there is no missing data, the mixed model (5) is equivalent to the change score model (1), which does not account for pre-test. Thus, it is not surprising that the complete case change score analysis produced a biased treatment effect while the CC ANCOVA analysis provided an unbiased treatment effect since it adjusts for pre-test. This result implies that among CC analyses, treatment effect estimates from an ANCOVA model are more robust to missingness pattern than the change score analysis.

As expected, the simulation studies showed that ANCOVA was more powerful than the mixed model for CC analyses, and the mixed model was generally more powerful when all available data were used. Also, since CC ANCOVA provides more precise estimates than ANCOVA after MI [18] and ANCOVA is more powerful than the mixed model [5], it was not surprising that CC ANCOVA had the largest power under almost all scenarios. In some situations, the difference in power between the ANCOVA and mixed model analyses could be explained, in part, by the mixed models exhibiting type-I error rates below nominal levels, but this was not true in general. Also, the benefit in power from CC ANCOVA generally did not come at the cost of an inflated type-I error rate.

The results from the simulation studies were not consistent with the analyses of the BePHIT data. The simulation studies suggested that ANCOVA after MI was almost always the least powerful method for testing the treatment effect, while this approach resulted in the smallest p-value in the BePHIT analysis. This discrepancy may be the result of



(a) Power Comparison under MCAR

(b) Power Comparison under MAR (WH)

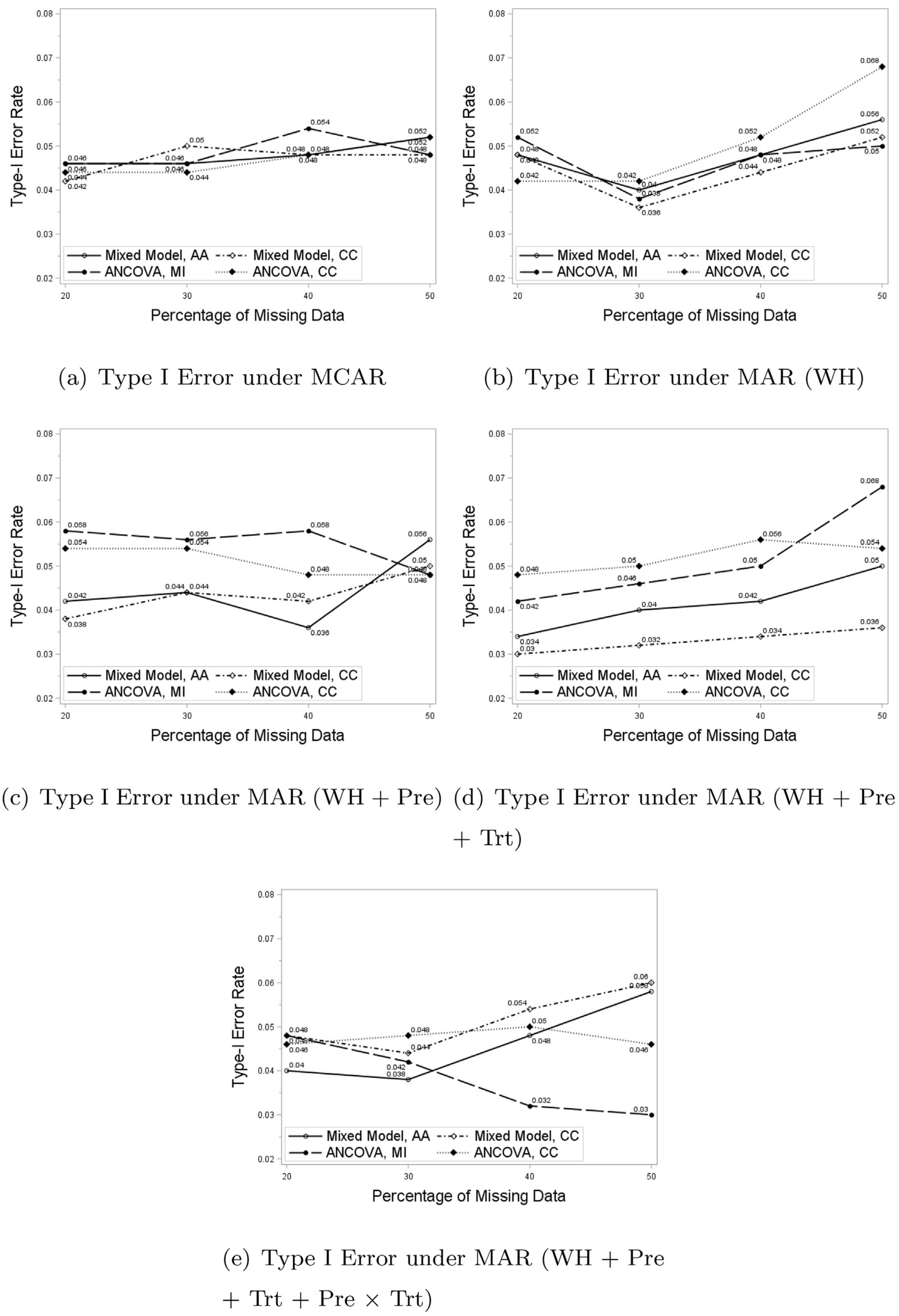
(c) Power Comparison under MAR (WH + Pre)

(d) Power Comparison under MAR (WH + Pre + Trt)

(e) Power Comparison under MAR (WH + Pre + Trt + Pre × Trt)

Fig. 1. Power comparisons for analysis models including WH and $m = 35$ per group.

using different imputation models between the simulation studies and the BePHIT analyses. The imputation model in the simulations included only the variables used in generating the data and their two-way interactions, while the imputation model for the BePHIT studies also included some baseline anthropometric and psychometric measures. This resulted in a difference in the R^2 of the imputation model for the



(a) Type I Error under MCAR

(b) Type I Error under MAR (WH)

(c) Type I Error under MAR (WH + Pre)

(d) Type I Error under MAR (WH + Pre + Trt)

(e) Type I Error under MAR (WH + Pre + Trt + Pre × Trt)

Fig. 2. Type I Error Rate Comparisons for Analysis Models with WH and $n = 35$.

two studies: $R^2 = 0.6975$ for the BePHIT study, while the average R^2 was smaller for each simulated scenario when 20% of the data were missing (between 0.61 and 0.65).

6.2. Comparison with similar studies

Although many papers have compared analysis methods for pre-post studies when data are completely observed [2–5,22] considerably fewer have compared methods under missing post-test values. Liu et al. [23] and Hyer and Waller [24] compared mixed models to ANCOVA models applied to complete cases. Mehrotra et al. imputed missing post-test values, applied a robust regression model to the pre-post changes, and compared the approach to more standard analyses for missing longitudinal data (weighted generalized estimating equations and mixed models) [25]. Mehrotra et al.'s work was similar to ours in that they compared ITT analysis methods, one of which involved imputation. However, the authors did not consider ANCOVA models which many consider the preferred analysis approach for complete data [5]. Furthermore, Mehotra et al. considered studies with several follow-ups and thus imputations were based on more observable outcome data than in two group pre-post studies.

Sullivan et al. considered missing data in a univariate outcome, a multivariate outcome, and a baseline covariate and compared MI, applied both overall and by randomized group, with a complete case analysis and a mixed model. While there are some similarities between our study and Sullivan et al.'s, there are two key differences. First, Sullivan et al. only considered a large sample size in their simulations ($n = 600$ total). While large sample sizes such as these may be common in clinical trials, many studies with pre-post designs involve much fewer subjects. For instance, our motivating study BePHIT consisted of 35 women per treatment group. Another example is a recent crossover study on depression and healthier food choices, where a total of 58 women were recruited [26]. Thus, motivated by the BePHIT study, we considered scenarios with small sample sizes. The second major difference between our study and Sullivan et al.'s is the type of mixed model considered in the ITT analysis; we considered a model for pre-test and post-test with a time-dependent treatment effect while Sullivan et al. considered a mixed model for post-test and an auxiliary outcome adjusting for pre-test. Our mixed model provides a treatment effect on the pre-post change in the outcome and thus is an extension of the commonly used change-score analysis to accommodate missing data. From our experience, the Sullivan et al. model is not as common since, in a study with a single post-randomization follow-up, if a subject is lost-to-follow-up, they will not have any information on the main outcome or the auxiliary outcome.

Despite the differences in the settings of the simulations, Sullivan et al. also observed that complete case analysis (CCA) was more efficient than multiple imputation (MI) for pre-post studies [27]. When there were multiple follow-ups, Sullivan et al. found that the mixed model was more efficient than MI, though MI outperformed CCA since the intermediate outcome was ignored in CCA but used in the imputation model.

6.3. Limitations

The simulation settings were based on the BePHIT data and thus may not be reflective of studies whose outcome variables and covariates come from different distributions or with different designs (e.g., unequal numbers in each treatment group). Only four missing data mechanisms were considered in the simulations. These scenarios were chosen based on the analyses of the BePHIT study. However, many other missing mechanisms are possible in biomedical studies. For example, if the subject dropped out the study because she knew that her post-test one-mile walk would be very bad, the data would be not missing at random (NMAR). However, this scenario, along with many other scenarios of missing data were not considered.

6.4. Recommendations

Even though CC ANCOVA appeared to be the most powerful method, it is usually not recommended since it does not adhere to the ITT principle of clinical trials [7]. In most simulated scenarios, the all-available-data (AA) mixed model provided a power benefit over ANCOVA after MI. Even in those scenarios where the power of ANCOVA after MI was larger than the power of the AA mixed model, the two powers were very close (e.g., Fig. 1(a) at 20 and 30% missingness). Besides the power loss in ANCOVA after MI, ANCOVA requires no measurement error in the outcome [5] and MI is more time consuming than fitting a mixed model. Therefore, for ITT analyses of pre-post studies with loss to follow-up, the AA mixed model is recommended.

Acknowledgements

BePHIT was supported by a grant from the Breast Cancer Research Foundation (BCRF-17-122). This paper does not represent the official views of the Breast Cancer Research Foundation.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2018.05.008.

References

[1] P. David, J. Buckworth, M.L. Pennell, M.L. Katz, C.R. DeGraffinreid, E.D. Paskett, A walking intervention for postmenopausal women using mobile phones and interactive voice response, J. Telemed. Telecare 18 (1) (2012) 20–25.

[2] P.D. Allison, Change scores as dependent variables in regression analysis, Socio. Meth. 20 (1990) 93–114.

[3] L.J. Cronbach, L. Furby, How we should measure "change": or should we? Psychol. Bull. 74 (1) (1970) 68.

[4] D.V. Glidden, S.C. Shiboski, C.E. McCulloch, Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, Springer, 2011.

[5] J.M. Oakes, H.A. Feldman, Statistical power for nonequivalent pretest-posttest designs the impact of change-score versus ancova models, Eval. Rev. 25 (1) (2001) 3–28.

[6] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, 2002.

[7] J.M. Lachin, Statistical considerations in the intent-to-treat principle, Contr. Clin. Trials 21 (3) (2000) 167–189.

[8] I.R. White, J.B. Carlin, Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, Stat. Med. 29 (28) (2010) 2920–2931.

[9] D.T. Campbell, J.C. Stanley, N.L. Gage, Experimental and Quasi-experimental Designs for Research, Houghton Mifflin Boston, 1963.

[10] U.J. Pape, C. Millett, J.T. Lee, J. Car, A. Majeed, Disentangling secular trends and policy impacts in health studies: use of interrupted time series analysis, J. R. Soc. Med. 106 (4) (2013) 124–129.

[11] C.W. Harris, et al., Problems in Measuring Change, University of Wisconsin Press Madison, 1963.

[12] C.E. Werts, R.L. Linn, A general linear model for studying growth, Psychol. Bull. 73 (1) (1970) 17–22.

[13] Panel on Handling Missing Data in Clinical Trials; National Research Council, The Prevention and Treatment of Missing Data in Clinical Trials, The National Academies Press, 2010 ISBN 9780309158145 http://www.nap.edu/openbook.php?record_id=12955.

[14] J.G. Ibrahim, H. Chu, M.H. Chen, Missing data in clinical studies: issues and methods, J. Clin. Oncol. 30 (26) (2012) 3297–3303.

[15] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Stat. Med. 30 (4) (2011) 377–399.

[16] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley, New York, 1987.

[17] Y. Yuan, Multiple imputation using SAS software, J. Stat. Software 45 (6) (2011) 1–25.

[18] J. Carpenter, M. Kenward, Multiple Imputation and its Application, John Wiley & Sons, 2012.

[19] E.V. Gomez, G.B. Schaalje, G.W. Fellingham, Performance of the Kenward–Roger method when the covariance structure is selected using AIC and BIC, Commun. Stat. Simulat. Comput. 34 (2) (2005) 377–392.

[20] M.G. Kenward, J.H. Roger, Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53 (3) (1997) 983–997.

[21] R.C. Littell, SAS for Mixed Models, SAS institute, 2006.

[22] A. Salim, A. Mackinnon, H. Christensen, K. Griffiths, Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates, Psychiatr. Res. 160 (3) (2008) 335–345.

[23] G.F. Liu, K. Lu, R. Mogg, M. Mallick, D.V. Mehrotra, Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Stat. Med. 28 (2009) 2509–2530.

[24] J.M. Hyer, J.L. Waller, Comparison of five analytic techniques for two-group pre-post repeated measures designs using sas, SAS Global Forum Proceedings, 2014 Paper 1798–2014.

[25] D.V. Mehrotra, X. Li, J. Liu, K. Lu, Analysis of longitudinal clinical trials with missing data using multiple imputation in conjunction with robust regression, Biometrics 68 (2012) 1250–1259.

[26] J.K. Kiecolt-Glaser, C.P. Fagundes, R. Andridge, J. Peng, W.B. Malarkey, D. Habash, et al., Depression, daily stressors and inflammatory responses to high-fat meals: when stress overrides healthier food choices, Mol. Psychiatr. 22 (3) (2017) 476–482.

[27] T.R. Sullivan, I.R. White, A.B. Salter, P. Ryan, K.J. Lee, Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat. Meth. Med. Res. (2016) 0962280216683570.