An evaluation of the trimmed mean approach in clinical trials with dropout

Peer-reviewed author version

Wang, Ming-Dauh; Liu, Jiajun; MOLENBERGHS, Geert & Mallinckrodt, Craig (2018)
An evaluation of the trimmed mean approach in clinical trials with dropout. In:
PHARMACEUTICAL STATISTICS, 17(3), p. 278-289.

DOI: 10.1002/pst.1858
Handle: http://hdl.handle.net/1942/26190

An Evaluation of the Trimmed Mean Approach in Clinical Trials with Dropout

Ming-Dauh Wang[1], Jiajun Liu[2], Geert Molenberghs[3,4], and Craig Mallinckrodt[1]

1. *Lilly Research Labs, Eli Lilly and Co, Indianapolis, IN, USA*
2. *Biogen, Cambridge, MA, USA*
3. *I-BioStat, Hasselt University, Diepenbeek, Belgium*
4. *I-BioStat, Katholieke Universiteit, Leuven, Belgium*

**Correspondence to: Craig Mallinckrodt, Lilly Research Labs, Eli Lilly and Co, Indianapolis, IN, USA. E-mail: cmallinc@lilly.com*

Abstract

The trimmed mean is a method of dealing with patient dropout in clinical trials that considers early discontinuation to be a bad outcome rather than leading to missing data. The present investigation is the first comprehensive assessment of the approach across a broad set of simulated clinical trial scenarios. In the trimmed mean approach all patients who dropout are excluded from analysis by trimming an equal percentage of bad outcomes from each treatment arm. The untrimmed values are used to calculated means or mean changes. An explicit intent of trimming is to favor the group with lower dropout because having more completers can be a beneficial effect of the drug, or conversely, higher dropout can be a bad effect. In the simulation study treatment effects estimated from trimmed means were greater than the corresponding effects estimated from untrimmed means when dropout favored the experimental group, and vice versa. The trimmed mean estimates a unique estimand. Therefore, comparisons with other methods are difficult to interpret and the utility of the trimmed mean hinges on the reasonableness of its assumptions: dropout is an equally bad outcome in all patients and adherence decisions in the trial are sufficiently similar to clinical practice in order to generalize the results. Trimming might be applicable to other inter-current events such as switching to or adding rescue medicine. Given the well-known biases in some methods that estimate effectiveness, such as baseline observation carried forward and non-responder imputation, the trimmed mean may be a useful alternative when its assumptions can be justified.

Key words: Missing data, estimands, trimmed mean, clinical trials

1. Introduction

Missing data is an incessant problem in clinical trials. Fortunately, prevention and treatment of missing data has been an active area of investigation leading to new and updated guidance (1-4). Consensus exists that the framework for dealing with missing data begins with clear specification of trial objectives and causal estimands (1-5). Conceptually, an estimand is simply the true population quantity of interest to estimate that addresses the trial objective (5). Specification of estimands includes the population and time point (or time period) of interest, and a measure of the intervention effect (3). The intervention effect can often be described in terms of two components that yield four general categories of estimands: 1) whether interest is in the effects of an individual treatment or a treatment policy (often involving a specific treatment); and 2) whether interest is in the effects if the intervention taken as directed (efficacy) or as actually taken (effectiveness) (1). With the variety of clinical trial scenarios and missing data possibilities, consensus on a universally best estimand is neither realistic nor desirable and multiple estimands may be of interest in a single trial (1).

Conceptually, effectiveness estimands are some combination of efficacy and adherence; as such, dropout may be a relevant outcome rather than a cause of missing data (6). Baseline observation carried forward (BOCF) and non-responder imputation (NRI) use dropout as an outcome in that patients who discontinue are ascribed values indicating no improvement, thereby yielding data sets with no missing data. However, these methods ignore changes due to non- pharmacologic effects, such as study effects, placebo effects, and the natural progression or waxing and waning of the disease and can therefore be biased (2).

Recently, Permutt and Li proposed a trimmed mean approach as an alternative for estimands in which dropout is a bad outcome (6). In principle, post-randomization (inter-current) events such as adding or switching to rescue medication may also be bad outcomes. Therefore, extending the trimmed mean approach to include these events may also be useful. The trimmed mean approach integrates the observed data on completers (or more generally, patients not having relevant inter-current events) with the information that some patients dropped out (or more generally, patients with any relevant inter-current event).

The trimmed mean approach uses well-known statistics and distributional theory (6). The basic idea is to assign an arbitrarily bad score (or outcome) to all patients that had a relevant inter-current event such that the assigned outcomes are worse than any observed outcomes. An equal fraction of the worst outcomes is trimmed from each treatment group such that all patients with a relevant inter-current event are trimmed, thereby obviating concerns about sensitivity to the actual score assigned and yielding an analysis data set with no missing values. Means (medians) or mean (median) changes from the untrimmed values are calculated and permutation methods used to construct a reference distribution for testing and interval estimation (6).

General attributes of the trimmed mean across a narrow set of conditions have been reported (6). The purpose of the present investigation is to assess the trimmed mean approach across a broader set of realistic clinical conditions than has been previously reported. Section 2 provides a summary of the trimmed mean methodology. Section 3 details a simulation study to assess power and type I error. Section 4 summarizes results of the simulation study. Section 5 includes a clinical trial example. Finally, Section 6 ties these ideas together in a discussion.

2. Description of the trimmed mean approach

The trimmed mean approach is applicable to clinical trials comparing treatments intended to improve the signs and symptoms of a disease (rather than alter the underlying disease mechanism) (6). The method assumes that all patients that dropped out have equally bad outcomes (6). Although the original authors did not speak to this, the approach could include other inter-current events such as switching to or adding rescue medicine. Another assumption, as with all methods that incorporate dropout as part of the outcome, is that adherence decisions in the clinical trial are sufficiently close to the decisions that made in clinical practice in order for the results to be generalized (1).

Using the nomenclature of the draft ICH E9 R1 addendum on estimands and sensitivity (7), the trimmed mean is a composite approach to dealing with inter-current events. The specific estimand tested by the trimmed mean is the difference between treatments in endpoint means in the X% subset of patients with the most favorable outcomes, where patients with a relevant inter-current event are considered to have had an outcome worse than any patient in the X% subset. The fraction of data to be trimmed can be determined using either an a priori chosen fixed percentage that based on previous trials ensures all bad outcomes resulting from assigning bad scores to patients with relevant inter-current are trimmed; or, trimming can be adaptive based on the actual results of the trial (6).

Permutt and Li (6) provided technical details for the trimmed mean approach, with additional description and software code to implement the method in Mehrotra et al (8). Permutt and Li (6) summarized the method as follows:

Stigler [9] gave distribution theory for the trimmed mean, including the asymmetric case. The trimmed mean is asymptotically normal under regularity conditions, and its expectation is the population trimmed mean (6). Tukey (10) gave a heuristic way of approximating the variance, at least in the symmetric case; alternatively, a permutation test can be sued to construct the exact distribution conditional on the observed values:

- Calculate the difference between the trimmed means.
- Keep the data (including the fact of discontinuation) the same and permute the treatment assignments in all possible ways, or in a sufficiently large random sample of possible ways.
- Calculate the difference between the trimmed means for each permutation. If the original difference falls in, say, the upper 2.5% of the permutation distribution, the difference is significant.

Exact CIs for location shifts are constructed by inverting the permutation test. That is, if Δ represents the shift, the hypothesis $\Delta = t$ can be tested by subtracting t from each value in the active treatment group and repeating the permutation test. The set of values of t for which the hypothesis $\Delta = t$ is not rejected is an exact CI for Δ. This procedure entails intensive computation because the permutation distribution has to be constructed for each value of t. Alternatively, CI can be constructed from the permutation distribution under $t = 0$. For example, if Y is the estimate and $Y_{\gamma/2}$ and $Y_{1-\gamma/2}$ are the lower and upper $\gamma/2$ quantiles of the permutation distribution, then $Y + \gamma_{\gamma/2}$, $Y + \gamma_{1-\gamma/2}$ is approximately a 1-γ CI for all percentiles above $_{\gamma/}$.

The quantiles of the permutation distribution may be estimated either by the empirical quantiles or by a normal approximation $\pm s\phi^{-1}(\gamma/2)$ where s is the sample standard deviation of the permutation distribution. Such intervals are conservative because the permutation distribution overestimates the variance of the test statistic when the null hypothesis is false (6).

Permutt and Li (6) further note that the difference between groups in the trimmed mean is proportional to the area between the empirical cumulative distribution curves (with changes of sign if they cross) and below $1 - \alpha$. The difference in medians is the horizontal difference between the curves at a height of 0.5. The difference between groups in any percentile is a horizontal distance, and the difference in trimmed means is the average of the differences for all percentiles above γ. Accordingly, it can be interpreted much the same as the median, as a summary of the distance between the CDFs in general, and as the constant difference in the special case of a location shift. Averaging many (differences in) order statistics will usually result in a less variable statistic than the (differences in) medians alone (6).

3. **Simulation study**

The goals of the simulation study were two-fold: 1) To characterize the impact of discontinuations for individual reasons, such as lack of efficacy or intolerability, on trimmed means; and, 2) to characterize the simultaneous impact of multiple reasons for discontinuation, such as would be anticipated in actual clinical trials, on trimmed means.

A base scenario with no missing data was constructed to contrast with results from otherwise similar scenarios that had varying rates of and reasons for missing data. The base scenario was

constructed as: Treatments $i = 1,2$ refer to experimental and control arms, respectively. Changes from baseline were simulated as $N(\mu_i, \sigma^2), i = 1,2$. With $\mu_1 = -40,\ \mu_2 = -20,\ d = \mu_1 - \mu_2 = -20$. In this parameterization, lower scores represented greater improvement (-40 is better than -20). Using a two-sided alpha = 0.05 yielded 90% power in a superiority test of experimental versus control with $\sigma = 30$ and N = 50 per arm, or if $\sigma = 43$ and N = 100 per arm.

Simulation Input parameters were chosen to test results across a broad set of realistic scenarios. However, inputs were not from actual clinical trials. The first set of simulations, referred to as scenario A, had treatment arms with equal rates of missing data that arose from a completely random mechanism. Scenarios A1-A4 assumed the means and variances as described in the previous paragraph. Scenarios A5-A8 were otherwise identical to A1-A4, except the mean changes were -20 for both groups, yielding $d = \mu_1 - \mu_2 = 0$.

Therefore, scenarios A1-A4 assessed power and scenarios A5-A8 assessed "something like" Type I error. The qualifier "something like" is important because this assessment was based on the rate of significant differences between *trimmed* means in scenarios where the *untrimmed* means did not differ, whereas typical Type I error assessment would be based on scenarios where the *trimmed* means did not differ. In other words, it is assessing the Type I error of a different estimand. The estimand associated with the untrimmed mean is the difference between treatment means based on all patients if all patients completed the trial. As such, a significant difference in trimmed means when the untrimmed means are equal is not necessarily a false positive result. However, the chosen comparisons are necessary to understand the impact of assigning arbitrarily

and equally bad scores for all non-adherent patients in situations where no difference between treatments exist if all patients adhere. Table 1 includes additional details on these simulations.

Table 1: Input parameters for simulations scenarios A1-A8, with equal rates of missing data arising from a completely random mechanism.

Scenario	Treatment arm[1]	Missing Rate (%)	Endpoint means
A1	Exp	5	-40
	Con	5	-20
A2	Exp	10	-40
	Con	10	-20
A3	Exp	15	-40
	Con	15	-20
A4	Exp	20	-40
	Con	20	-20
A5	Exp	5	-20
	Con	5	-20
A6	Exp	10	-20
	Con	10	-20
A7	Exp	15	-20
	Con	15	-20
A8	Exp	20	-20
	Con	20	-20

1. Exp = experimental arm, Con = control arm

Details on the second set of simulations are provided in Table 2. The only difference from the first scenario was that data were deleted from the experimental arm only. These simulations can be viewed in two contexts: 1) as extension of the first set except with deletions applied to only the experimental arm; or, 2) as an assessment of the impact of discontinuations for intolerability, assuming the tolerability is independent of efficacy.

Table 2: Input parameters for simulations scenarios B1-B8, with unequal rates of missing data arising from a completely random mechanism.

Scenario	Treatment arm[1]	Missing Rate (%)	Endpoint means
B1	Exp	5	-40
	Con	0	-20
B2	Exp	10	-40
	Con	0	-20
B3	Exp	15	-40
	Con	0	-20
B4	Exp	20	-40
	Con	0	-20
B5	Exp	5	-20
	Con	0	-20
B6	Exp	10	-20
	Con	0	-20
B7	Exp	15	-20
	Con	0	-20
B8	Exp	20	-20
	Con	0	-20

1. Exp = experimental arm, Con = control arm

A third set of simulations was conducted in order to mimic data missing due to discontinuation for lack of efficacy. For these simulations the probability an observation was missing was a function of the outcome variable; that is, missing probability for a response x was, $F_{N(m,s^2)}(x)$, where F is the distribution function of the normal distribution with mean m and variance s^2. For example, if $m = 50$, and $s = 10$, if a patient had a change from baseline of 35, the probability for this patient to drop out = $F_{N(50,10^2)}(35) = 0.067$. This process mimicked situations where patients with better responses (changes of greater negative magnitude) were less likely to dropout, and vice versa.

The same function was applied to both treatment arms. Values of m and s were manipulated to generate rates of missing data similar to those in scenarios A and B. With only one observation per subject, the missing data due to lack of efficacy was a missing not at random mechanism. However, given the same parameters were used for deleting data in each treatment arm, it was not possible to create equal rates of missing data when efficacy differed between treatment arms, nor was it possible to generate different rates of missing data when efficacy was the same in both treatments.

Table 3: Input parameters for simulations scenarios C1-C8, with varying rates of missing data arising from lack of efficacy.

Scenario	Treatment[1]	Missing Rate (%)	Endpoint means
C1 m=50,s=10	Exp	2	-40
	Con	6	-20
C2 m=50,s=35	Exp	5	-40
	Con	10	-20
C3 m=35,s=35	Exp	9	-40
	Con	16	-20
C4 m=15,s=8	Exp	10	-40
	Con	21	-20
C5 m=50,s=10	Exp	6	-20
	Con	6	-20
C6 m=50,s=35	Exp	10	-20
	Con	10	-20
C7 m=35,s=35	Exp	16	-20
	Con	16	-20
C8 m=15,s=8	Exp	21	-20
	Con	21	-20

1. Exp = experimental arm, Con = control arm

A fourth set of simulations was conducted in order to mimic data missing from actual clinical trials, with a combination of discontinuations for lack of efficacy, intolerability, and completely

random reasons. These simulations involved deleting observations using various combinations of the deletion strategies used in scenarios A, B, and C. Specific details are provided in Table 4.

Table 4: Input parameters for simulations scenarios D1-D8, with varying rates of missing data arising from a combination reasons.

Scenario	Treatment[1]	Missing Rates (%)[2]				
		R1	R2	R3	Overall	Endpoint Means
D1 m=50,s=10	Exp	5	24	2	30	-40
	Con	5	0	6	10	-20
D2 m=50,s=35	Exp	5	16	5	25	-40
	Con	5	0	10	15	-20
D3 m=35,s=35	Exp	5	8	9	20	-40
	Con	5	0	16	20	-20
D4 m=15,s=8	Exp	5	0	10	15	-40
	Con	5	0	21	25	-20
D5 m=50,s=35	Exp	5	0	5	10	-40
	Con	5	17	10	30	-20
D6 m=50,s=10	Exp	5	21	6	30	-20
	Con	5	0	6	10	-20
D7 m=50,s=35	Exp	5	11	10	25	-20
	Con	5	0	10	15	-20
D8 m=35,s=35	Exp	5	0	16	20	-20
	Con	5	0	16	20	-20
D9 m=50,s=35	Exp	5	0	10	15	-20
	Con	5	11	10	25	-20
D10 m=50,s=10	Exp	5	0	6	10	-20
	Con	5	21	6	30	-20

1. Exp = experimental arm, Con = control arm
2. R1 and R2 specify rates of dropout from a completely random dropout mechanism. The sum of R1 + R2 = dropout for completely random reasons + dropout due to intolerability that is unrelated to outcome; R3 = dropout depending on outcome, thereby mimicking lack of efficacy.

For each scenario 5,000 trials were simulated and for each trial 1,000 permutations were constructed. In order to confirm that this level of replication was sufficient to yield stable results, Monte Carlo errors were calculated for scenarios B2 and B6 as Monte Carlo errors were calculated as (INSERT) . The impact of this level of Monte Carlo error was assessed by

replicating these scenarios were each conducted 10 times. Results across these 10 replications yielded rates of significant difference that all fell within a range of approximately 1%, thereby confirming that results were stable and replication sufficient to eliminate findings due to chance alone.

Adaptive trimming was used in analyses of each permuted data; that is, only dropouts were trimmed for the group with the higher dropout, and the same percentage of observations were trimmed from the group with lower dropout, resulting in all dropouts plus the worst observed values being trimmed from that group.

4. **Simulation study results**

Results from simulation scenarios A1-A4 are summarized in Table 5. In these scenarios with equal rates of dropout arising from a completely random mechanism, the average estimates of the trimmed means were equal to the corresponding untrimmed means that were input for simulation. As the rate of missing data increased, more data were trimmed and the sample sizes decreased, which reduced power. Differences in the variance of untrimmed values did not influence trimmed mean estimates.

Table 5: Trimmed means and power from simulation scenarios A1-A4, with untrimmed treatment means of -40 and -20, and equal rates of missing data arising from a completely random mechanism.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation							
			n=50 ($\sigma = 30$)				n=100 ($\sigma = 43$)			
			Exp Mean	Con Mean	Mean Diff	Power	Exp Mean	Con Mean	Mean Diff	Power
A1	Exp	5	-40.00	-20.00	-20.00	0.87	-40.12	-19.98	-20.14	0.84
	Con	5								
A2	Exp	10	-39.89	-20.13	-19.75	0.78	-39.99	-20.07	-19.92	0.77
	Con	10								
A3	Exp	15	-39.88	-20.12	-19.76	0.73	-39.90	-19.99	-19.92	0.70
	Con	15								
A4	Exp	20	-40.05	-19.87	-20.18	0.67	-40.04	-19.95	-20.09	0.65
	Con	20								

Results from simulation scenarios A5-A8 are summarized in Table 6. In these scenarios with equal untrimmed means and equal rates of complete random dropout, the percentages of significant differences (something like Type I error, as described in Section 3) was less than the nominal rate of 5%. Differences in variance of untrimmed values did not influence estimates of trimmed means.

Table 6: Trimmed means and rate of significant differences from simulation scenarios A5-A8, with untrimmed treatment means of -20 and -20, and equal rates of missing data arising from a completely random mechanism.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation							
			n=50 ($\sigma = 30$)				n=100 ($\sigma = 43$)			
			Exp Mean	Con Mean	Mean Diff	α	Exp Mean	Con Mean	Mean Diff	α
A5	Exp	5	-20.06	-19.97	-0.09	0.04	-20.15	-19.93	-0.02	0.03
	Con	5								
A6	Exp	10	-20.10	-20.03	-0.07	0.02	-20.06	-20.05	-0.01	0.02
	Con	10								
A7	Exp	15	-19.91	-19.93	0.02	0.02	-20.13	-20.01	-0.12	0.02
	Con	15								
A8	Exp	20	-20.02	-20.02	0.002	0.01	-20.07	-19.97	-0.10	0.01
	Con	20								

Results from simulation scenarios B1-B4 are summarized in Table 7. In these scenarios with untrimmed treatment means of -40 and -20 and unequal rates of completely random dropout, estimates of the difference between treatments for trimmed means decreased and therefore power decreased as the rate of missing data increased. Trimmed means within the experimental group were not affected by trimming because the trimmed values were a random sample from the distribution. Trimmed means within the control group increased as trimming increased because the trimmed values were the worst values from that group. The difference between trimmed and untrimmed means for the control group were greater in the scenarios with greater variance because the values trimmed from the tail of the distribution were more different from the mean than when the variance was smaller. Therefore, in the scenarios with greater variance the difference between the experimental and control groups were smaller than in the scenarios with smaller variance.

Table 7: Trimmed means and power from simulation scenarios B1-B4, with untrimmed treatment means of -40 and -20, and unequal rates of missing data arising from a completely random mechanism.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation							
			n=50 ($\sigma = 30$)				n=100 ($\sigma = 43$)			
			Exp Mean	Con Mean	Mean Diff	Power	Exp Mean	Con Mean	Mean Diff	Power
B1	Exp	5	-40.04	-22.55	-17.49	0.78	-39.99	-24.64	-15.35	0.64
	Con	0								
B2	Exp	10	-40.01	-25.66	-14.37	0.53	-40.08	-28.29	-11.79	0.36
	Con	0								
B3	Exp	15	-40.00	-27.61	-12.39	0.38	-40.07	-31.66	-8.41	0.17
	Con	0								
B4	Exp	20	-40.00	-30.31	-9.69	0.20	-40.07	-34.87	-5.19	0.06
	Con	0								

Results from simulation scenarios B5-B8 are summarized in Table 8. In these scenarios with

untrimmed treatment means of -20 and -20 and unequal rates of completely random dropout, estimates of the difference between treatments for trimmed means increased and therefore the rate of significant differences (something like Type I error, as described in Section 3) increased as the rate of missing data increased. Trimmed means within the experimental group were not affected by trimming because the trimmed values were a random sample from the distribution. Trimmed means within the control group increased as trimming increased because the trimmed values were the worst values from that group. The difference between trimmed and untrimmed means for the control group were greater in the scenarios with greater variance because the values trimmed from the tail of the distribution were more different from the mean than when the variance was smaller. Therefore, in the scenarios with greater variance the difference between the experimental and control groups were greater than in the scenarios with smaller variance.

Table 8: Trimmed means and type I error from simulation scenarios B5-B8, with untrimmed treatment means of -20 and -20, and unequal rates of missing data arising from a completely random mechanism.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation							
			n=50 ($\sigma = 30$)				n=100 ($\sigma = 43$)			
			Exp Mean	Con Mean	Mean Diff	α	Exp Mean	Con Mean	Mean Diff	α
B5	Exp	5	-19.92	-22.70	2.77	0.06	-20.06	-24.61	4.55	0.09
	Con	0								
B6	Exp	10	-20.06	-25.68	5.62	0.11	-19.96	-28.30	8.34	0.21
	Con	0								
B7	Exp	15	-19.94	-27.56	7.62	0.17	-19.99	-31.57	11.58	0.35
	Con	0								
B8	Exp	20	-19.95	-30.17	10.22	0.27	-20.09	-34.92	14.83	0.50
	Con	0								

Results from simulation scenarios C1-C4 are summarized in Table 9. In these scenarios with untrimmed treatment means of -40 and -20 and unequal rates of missing data arising from lack of efficacy, trimmed means within each group increased because the trimmed values tended to be

bad outcomes. The increase in trimmed means was relatively equal in both groups resulting in relatively small changes to the average estimates of the difference between trimmed means and power.

Table 9: Trimmed means and power from simulation scenarios C1-C4, with untrimmed treatment means of -40 and -20, and unequal rates of missing data arising from lack of efficacy.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation			
			Exp Mean	Con Mean	Mean Diff	Power
C1 m=50,s=10	Exp	2	-45.11	-25.18	-19.96	0.899
	Con	6				
C2 m=50,s=35	Exp	5	-48.07	-26.86	-21.20	0.890
	Con	10				
C3 m=35,s=35	Exp	9	-51.70	-29.72	-21.98	0.884
	Con	16				
C4 m=15,s=8	Exp	10	-55.61	-35.59	-20.02	0.878
	Con	21				

Results from simulation scenarios C5-C8 are summarized in Table 10. In these scenarios with untrimmed treatment means of -20 and -20 and equal rates of missing data arising from lack of efficacy, trimmed means within each group increased because the trimmed values tended to be bad outcomes. The increase in trimmed means was relatively equal in both groups resulting in negligible changes to the average estimates of the difference between trimmed means and correspondingly negligible changes in the rate of significant differences (something like Type I error, as described in Section 3).

Table 10: Trimmed means and type I error from simulation scenarios C5-C8, with untrimmed treatment means of -40 and -40, and equal rates of missing data arising from lack of efficacy.

Scenario	Treatment	Missing Rate (%)	Monte Carlo Estimation			
			Exp Mean	Con Mean	Mean Diff	α
C5 m=50,s=10	Exp	6	-25.93	-25.95	0.03	0.049
	Con	6				
C6 m=50,s=35	Exp	10	-28.25	-28.19	-0.06	0.056
	Con	10				
C7 m=35,s=35	Exp	16	-31.30	-31.57	0.27	0.051
	Con	16				
C8 m=15,s=8	Exp	21	-36.94	-36.97	0.03	0.048
	Con	21				

Results from simulation scenarios D1-D5 are summarized in Table 11. These scenarios were intended to match realistic clinical trial scenarios with untrimmed treatment means of -40 and -20 and varying rates of missing data arising from multiple reasons. When dropout was higher in the experimental group than in the control group the difference between trimmed means was smaller than in the corresponding untrimmed means. The magnitude of this decrease, and the corresponding decrease in power, was greater as the difference in dropout rates became greater as a result of higher rates in the experimental group. When dropout was higher in the control group than in the experimental group the difference between trimmed means was greater than in the corresponding untrimmed means. The magnitude of this increase, and the corresponding increase in power, was greater as the difference in dropout rates became greater as a result of higher rates in the control group.

Table 11: Trimmed means and power from simulation scenarios D1-D4, with untrimmed treatment means of -40 and -20, and varying rates of missing data arising from multiple reasons.

Scenario	Treatment	Missing Rates (%)				Monte Carlo Estimation			
		R1	R2	R3	Overall	Exp Mean	Con Mean	Mean Diff	Power
D1 m=50,s=10	Exp	5	24	2	30	-42.15	-38.59	-3.56	0.04
	Con	5	0	6	10				
D2 m=50,s=35	Exp	5	16	5	25	-43.77	-35.10	-8.67	0.14
	Con	5	0	10	15				
D3 m=35,s=35	Exp	5	8	9	20	-47.39	-31.37	-16.02	0.51
	Con	5	0	16	20				
D4 m=15,s=8	Exp	5	0	10	15	-55.61	-35.49	-20.12	0.85
	Con	5	0	21	25				
D5 m=50,s=35	Exp	5	0	5	10	-58.85	-26.82	-32.03	0.99
	Con	5	17	10	30				

Results from simulation scenarios D6-D10 are summarized in Table 12. These scenarios were intended to match realistic clinical trial scenarios with untrimmed treatment means of -20 and -20 and varying rates of missing data arising from multiple reasons. When dropout was higher in the experimental group than in the control group the difference between trimmed means favored the control group. The magnitude of the difference in trimmed means, and the corresponding increase in the rate of significant differences (something like Type I error, as described in Section 3) was greater as the difference in dropout rates became greater as a result of higher rates in the control group. When dropout was lower in the control group than in the experimental group the difference between trimmed means favored the control group. The magnitude of the difference in trimmed means, and the corresponding increase in the rate of significant differences (something like Type I error, as described in Section 3) was greater as the difference in dropout rates became greater as a result of higher rates in the experimental group.

Table 12: Trimmed means and type I error from simulation scenarios D5-D8, with untrimmed treatment means of -20 and -20, and varying rates of missing data arising from multiple reasons.

Scenario	Treatment	Missing Rates (%)				Monte Carlo Estimation			
		R1	R2	R3	Overall	Exp Mean	Con Mean	Mean Diff	α
D6 m=50,s=10	Exp	5	21	6	30	-25.07	-39.00	13.94	0.43
	Con	5	0	6	10				
D7 m=50,s=35	Exp	5	11	10	25	-26.73	-34.74	8.01	0.15
	Con	5	0	10	15				
D8 m=35,s=35	Exp	5	0	16	20	-31.46	-31.52	0.05	0.047
	Con	5	0	16	20				
D9 m=50,s=35	Exp	5	0	10	15	-34.75	-26.88	-7.87	0.14
	Con	5	11	10	25				
D10 m=50,s=10	Exp	5	0	6	10	-38.95	-25.00	-13.96	0.43
	Con	5	21	6	30				

5. **Clinical trial example**

A re-examination of a real clinical trial is used to reinforce findings from the simulation study. This was a phase 3, double-blind, placebo- and active-controlled trial in which 1307 patients with active rheumatoid arthritis (RA) were randomly assigned in a 3:3:2 ratio to placebo, an experimental drug, or an active comparator, a well-characterized standard of care in RA (1305 patients had post-baseline data for analysis).

The primary endpoint was the binary outcome of 20% improvement according to the criteria of the American College of Rheumatology (ACR20 response). For purposes of this re-examination only the placebo and active comparator arms are included and focus is on one of the key secondary endpoints, the HAQ-DI, which is a measure of physical function. Although this is an ordinal outcome with a range of 0 to 3 (higher scores indicating greater disability), it is commonly analyzed as a continuous outcome and that convention is adopted here. The time point of interest is week 24, with rescue treatment becoming available at week 16.

Although the active comparator is well-known and well-characterized, the true difference between it and placebo is not known. Therefore, results from the trimmed mean cannot be compared to true values. Therefore, results from the trimmed mean are compared to results from commonly used methods, which estimate differing estimands. Trimmed means were calculated as previously described for the simulation study. Other analyses included modified baseline observation carried forward (mBOCF), modified last observation carried forward (mLOCF), and a likelihood-based repeated measures analysis commonly referred to as MMRM (mixed-effects model for repeated measures).

In mBOCF analyses, for patients who discontinue the study or permanently discontinue the study treatment because of an AE, including death, the baseline observation is used as the week 24 observation, indicating no improvement. For patients who receive rescue, the last nonmissing observation at or before rescue is used as the week 24 observation

In mLOCF analyses, for patients who discontinue the study or permanently discontinue the study treatment for any reason, the last nonmissing postbaseline observation before discontinuation is used as the week 24 observation. For patients who receive rescue, the last nonmissing observation at or before rescue is used as the week 24 observation.

In both mBOCF and mLOCF, data were analyzed using ANCOVA with a model that included treatment and baseline values. The MMRM analysis included all postbaseline observations (weeks 1, 2, 4, 8, 12, 14, 16, 20, 24), Data were analyzed using likelihood-based estimation with

a fixed effects model that included treatment, baseline values, visit and treatment by visit interaction, and the within-patient errors were modeled using an unstructured covariance matrix.

In the placebo group, 333 of 488 Week 24 observations were available, leaving 155 (31.8%) of the observations that were either missing, imputed, or trimmed, depending on the analytic method. For the active comparator, 272 of 330 observations were available, leaving 58 (17.6%) of the observations that were either missing, imputed or trimmed, depending on the analytic method. Disposition is further summarized in Table 13.

Table 13. Patient disposition for the example clinical trial

	Placebo	Comparator
Number randomized	488	330
Number rescued	105	35
Discontinuations		
AE	15	7
Lack of efficacy	15	3
All other reasons	20	13
Total	50	23

Results from the various analyses are summarized in Table 14. It is not surprising that all analytic methods yielded highly significant results given the robust, proven efficacy of the active comparator and the large sample sizes. Therefore, focus is on the magnitude of the point estimates and the width of the confidence intervals.

The point estimates from MMRM, mBOCF, and mLOCF were -0.256, -0.275, and -0.277, respectively. In contrast, the trimmed mean was considerably larger, 0.435. Therefore, trimmed mean results from the example clinical trial were similar to results from the simulation study in

scenarios where dropout was higher on the control arm. That is, the trimmed mean yielded a larger difference between treatments because all dropout outcomes and the worst actually observed outcomes were trimmed in the active arm, whereas only dropout outcomes were trimmed in the placebo arm. Also similar to simulation findings, the estimated difference between trimmed means was less precise than estimates of means based on all randomized patients. Specifically, the width of the 95% confidence intervals from mBOCF, mLOCF, and MMRM were approximately 0.15, whereas the confidence interval width from the trimmed mean approach was around 0.26.

Table 14: Results from example clinical trial

		COMP N=330	PBO N=488
MMRM Censored at Discon/Rescue	n (missing %)	272 (17.6%)	333 (31.8%)
	Mean	-0.710	-0.443
	LS mean	-0.709	-0.452
	LSMean difference (95% CI) Pvalue	-0.256 [-0.328, -0.184] P<0.0001	
Adaptive Trimmed Mean	Trimmed mean	-0.878	-0.443
	Mean difference (95% CI) Pvalue	-0.435 [-0.565, -0.304] P<0.0001	
ANCOVA + mBOCF	N	330	488
	Mean	-0.625	-0.340
	LS mean	-0.620	-0.345
	LSMean difference (95% CI) Pvalue	-0.275 [-0.351, -0.199] P<0.0001	
ANCOVA + mLOCF	n	330	488
	Mean	-0.640	-0.350
	LS mean	-0.632	-0.355
	LSMean difference (95% CI) Pvalue	-0.277 [-0.354, -0.201] P<0.0001	

6. **Discussion**

The trimmed mean is an alternative to existing methods for continuous endpoints. The estimand it assesses is the difference between treatments in endpoint means in the best X% of patients. This is a unique estimand not addressed by other methods currently in common use. An explicit intent of the trimmed mean is to favor the group with lower dropout because having more completers can be a beneficial effect of the drug, or conversely, higher dropout can be a bad effect. In the simulation study, the difference between treatments in trimmed means did indeed influence completion rates. In the simulation study the difference between groups in trimmed means was greater than the corresponding difference in untrimmed mean changes when completion rates were higher in the experimental group; the difference between groups in trimmed mean changes was less than the corresponding differences between untrimmed mean changes when completion rates were higher in the control group. In scenarios where the untrimmed means were equal in the experimental and control groups, higher completion rates in the experimental group yielded trimmed mean changes that favored the experimental group, and the difference between groups in trimmed mean changes favored the control group when completion rates were higher in the control group.

The magnitude of the difference between untrimmed and trimmed means depended not just on the relative rates of dropout but also on the dropout mechanism. Trimming dropouts for lack of efficacy tended to have less impact than trimming for reasons unrelated to efficacy. Dropouts for lack of efficacy would have had a bad outcome if they had been observed; therefore, assigning bad outcomes to dropouts did not cause much re-ranking of outcomes. In contrast, dropout for reasons unrelated to efficacy resulted in a greater re-ranking of the data because bad outcomes

could be assigned to patients that would have had good outcomes if they had been observed and not trimmed.

Several limitations influence interpretations of these results. For example, dropout mechanism influenced the effects of trimming and the simulations used a mechanism in which adverse events were unrelated to the outcomes. If a relationship existed, the results of trimming would be different from those observed here. Therefore, results are not a specific guide of what to expect in actual practice, although the general trends should be a useful. Further, it would be interesting to explore the trimmed mean with longitudinal data as that would allow exploration of a larger variety of missing-data mechanisms.

The trimmed mean is estimating something different than existing methods. Therefore, direct comparisons of the trimmed mean with other methods are difficult to interpret. For example, we cannot say the trimmed mean works better than or worse than method X because the methods target different estimands. Similarly, we cannot say that a significant difference in trimmed means in scenarios where no difference existed in untrimmed means is a Type I error because the trimmed mean does not target inferences from the untrimmed values. However, it is important to note that the trimmed mean can yield significant differences between treatments in situations where there would be no difference between treatments if all patients completed.

Therefore, the utility of the trimmed mean hinges on the reasonableness of its assumptions. These assumptions include that dropout is an equally bad outcome in all patients and that adherence decisions in the trial are sufficiently similar to what is expected in clinical practice in

order to generalize the results. Permutt and Li (6) noted scenarios where these assumptions do not hold. For example, it is crucial to distinguish incomplete data from drop out due to intolerability versus incomplete data from death. These authors also noted that increased adherence is not always beneficial if that adherence arises from a pleasant side effect of the drug that is unrelated to the outcome. However, in situations where assumptions are reasonable, the trimmed mean is an intuitive approach that does not require modeling assumptions or assumptions about the missing data mechanism, nor does it rely on explicit imputation of missing values.

It must be remembered, though, that the trimmed mean assesses a unique estimand. This estimand may not be relevant in all situations. For example, trimmed means estimate benefit in a subset of patients. However, medications have cost for everyone who takes the drug. Therefore, trimmed means may not be relevant to Health Technology Assessors. Similarly, for those wishing to make across study comparisons the lack of historical use of the trimmed mean could be problematic.

For those situations where the trimmed mean is an appropriate option, power compared with untrimmed means is an important consideration. Results of the current study show that, all else equal, the reduction in power from using a subset of the data can be considerable and therefore the impact of trimming must be carefully considered in study planning. These considerations should include the anticipated rates of and reasons for early discontinuation. Although previous studies on the same compound in similar settings can be a useful guide, uncertainty in rates of

and reasons for dropout may add additional uncertainty into sample size estimation that would need to be taken into account.

Given the well-known biases and strong recommendations against using some methods that estimate effectiveness estimands, such as BOCF and NRI, the trimmed mean may be a useful alternative when the assumptions are justifiable.

References

1. Mallinckrodt CH, Molenberghs G, Rathmann S, Molenberghs G. Choosing estimands in clinical trials with missing data. *Pharmaceut. Statist.* ***2017***, 16 29–36. DOI: 10.1002/pst.1765
2. Mallinckrodt CH and Ilya Lipkovich. A practical guide to analyzing longitudinal clinical trial data. CRC press. New York. 2016
3. Phillips A, Abellan-Andres J, Andersen S, *et al.* Estimands: discussion points from the PSI estimands and sensitivity expert group. *Pharm.Stat.* 2017; **16**:6-11. **DOI:** 10.1002/pst.1745
4. Mallinckrodt CH, Roger J, Chuang-Stein C, *et al.* Recent developments in the prevention and treatment of missing data. *Therapeutic Innovation and Regulatory Science* 2014; **48**(1):68–80.
5. National Research Council. *The prevention and treatment of missing data in clinical trials. Panel on handling missing data in clinical trials. Committee on National Statistics, Division of Behavioural and Social Sciences and Education*. The National Academies Press: Washington, DC, 2010.
6. Permutt T, Li F. Trimmed means for symptom trials with dropouts. *Pharmaceut. Statist.* ***2017***, 16 20–28.
7. *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. EMA/CHMP/ICH/436221/2017. 30 August 2017*
8. Mehrotra, D, Liu F, Permutt T. Missing data in clinical trials: Control-based mean imputation and sensitivity analyses. Pharmaceutical Statistics. 2017;16:378‑392.
9. Stigler SM. The asymptotic distribution of the trimmed mean. *Annals of Statistics* 1973; **1**:472–477.
10. Tukey J. *Memorandum reports 31-34*. Statistical Research Group, Princeton University, 1946.