A Comparison of Sample Size Calculations for Cluster Randomized Crossover Trials with a Binary Outcome

Erin Case

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2014

Committee:

Susanne May

Siobhan Brown

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

A Comparison of Sample Size Calculations for Cluster Randomized Crossover Trials with a Binary Outcome

Erin Case

Chair of the Supervisory Committee:
Dr. Susanne May
Biostatistics

A simplified sample size calculation for a cluster randomized crossover trial with a binary survival outcome (the “T-BOSS” method) was compared to the closed form equation and simulation methods currently available from Connolly, et. al. (2013, Canadian Journal of Cardiology, 29, p.652) and Reich, et. al. (2012, PLoS ONE, 7(4), p.E35564), respectively. When there was no period effect present, the T-BOSS method was consistently conservative with a sample size that is 11.4%, 14.8%, and 37.8% larger than the closed form sample size for intracluster correlations of 0.01, 0.04, and 0.20, respectively. When there was a period effect present, T-BOSS was nearly always anti-conservative unless the period effect was small or the average cluster size was very small (less than 20 patients at most for the settings we explored). The simulations were fairly close to the closed form equation, though showed more variability.

Acknowledgements

Thank you, Susanne May, for being such a positive, knowledgeable, and encouraging mentor. Thank you, Siobhan Brown, for serving on my thesis committee in a capacity above and beyond what was required of you. Thank you to the University of Washington Biostatistics Department for the academic support and the Resuscitation Outcomes Consortium for providing the motivation for this research. Finally, to my family and Kevin Kyyro: thank you for your continued emotional support, and for making me feel like no goal is too high.

TABLE OF CONTENTS

Manuscript

Abstract

Background:

In 2010, statisticians at the Resuscitation Outcomes Consortium (ROC) developed a simplified sample size calculation when planning a cluster randomized crossover trial with a binary survival outcome. We sought to compare this method, the “T-Test Approximation for Binary Outcome Sample Size Calculations (T-BOSS)”, to the closed form equation and simulation methods currently available from Connolly, et. al. (2013, Canadian Journal of Cardiology, 29, p.652) and Reich, et. al. (2012, PLoS ONE, 7(4), p.E35564), respectively.

Methods:

T-BOSS and the closed form equation were compared in two hypothetical study scenarios developed from the ROC setting. The study scenarios explored low, medium, and high settings of effect size (odds ratios of 1.1, 1.2, and 1.25 or 1.3) and intracluster correlation (0.01, 0.04, and 0.20). The scenarios assumed no period effect, and the number of clusters was fixed at 200. Two survival rates were explored: 5% and 25%. In addition to comparing the methods in these scenarios, it was also explored how cluster size, period effect, and intracluster correlation (ICC) relate to their sample size estimates. Additionally, the simulation approach was briefly explored in the study scenarios.

Results:

When there is no period effect present, the T-BOSS method was consistently conservative with a sample size that is 11.4%, 14.8%, and 37.8% larger than the closed form sample size for ICCs of 0.01, 0.04, and 0.20, respectively. When there was a period effect present, T-BOSS was nearly always anti-conservative unless the period effect was small or the average cluster size was very small (less than 20 patients at most for the settings we explored). For the ROC trial, the T-BOSS sample size calculated to detect a 9.4% survival rate in the treatment group (vs. 8.1% in the control) was 17.3% larger than the closed form

equation sample size (21,815 vs. 18,599), assuming an ICC of 0.06 and no period effect. The simulations were fairly close to the closed form equation, though showed more variability.

Conclusions:

With the closed form equation now available there is no need to use T-BOSS. Nevertheless, it was proven to have been a suitable alternative during the planning of the ROC study.

Background

When designing a clinical trial, individual randomization to treatment groups is not always possible. In such settings, cluster randomization, which randomly assigns groups (clusters) of people to treatment arms, is often considered as an alternative[1]. If each cluster only experiences one treatment assignment, it can be difficult to determine whether observed differences in treatment groups are due to the treatment or due to the inherent differences between clusters unless the number of clusters is large. In addition, the power of a cluster randomized trial is typically governed by the number clusters rather than the number of individuals. Incorporating a crossover design, where each cluster is assigned every treatment arm and is randomly assigned to a sequence of these treatment periods, can address these issues[1].

This research is motivated by a trial performed by the Resuscitation Outcomes Consortium (ROC)[2]. ROC is currently conducting a cluster randomized crossover trial called the CCC trial, which seeks to evaluate two different methods of CPR after out-of-hospital cardiac arrests. One method uses continuous chest compressions (CCC), while the other uses a ratio of thirty chest compressions for every two rescue breaths given (30:2). The clusters are defined as emergency medical service (EMS) agencies from participating regions across the United States and Canada, and each cluster is randomized to a sequence of CCC and 30:2 over a number of 6 month periods until the total sample size is accrued or the study is stopped after one of the planned interim analyses.

Due to the aforementioned advantages over the cluster randomized trial, the cluster randomized crossover trial is implemented fairly often in situations where individual randomization is infeasible. A closed form equation has been developed to calculate the sample size needed in a cluster-randomized crossover clinical trial using a continuous outcome measurement[3], and it has been proven to be useful for sample size calculations when the outcome is binary as well[4]. In addition, simulation programs have been developed to estimate the power of these study designs[5]. Before the publication of the Connolly paper in 2013[4], statisticians at ROC proposed a simplified method for calculating the sample size that was used when planning the CCC trial in 2010. We aim to compare the performance of this method, which we will call the "T-Test Approximation for Binary Outcome Sample Size Calculations (T-BOSS)", to the current gold-standard of the closed form equation when calculating the sample size for a variety of cluster randomized crossover designs. After a brief explanation of the T-BOSS and the closed form equation methods, their sample size estimates are compared in a variety of settings. We also comment on a comparison of these methods to the simulation approach by Reich[5].

Methods

T-BOSS Method

ROC statisticians proposed calculating the sample size for a trial that wishes to compare the difference between two proportions: p_1, the proportion of people experiencing the outcome in the treatment arm, and p_2, the proportion of people experiencing the outcome in the control arm using a normal approximation. Each proportion is a sample mean, so $\widehat{p_1}$ and $\widehat{p_2}$ are approximately normally distributed by the Central Limit Theorem. Therefore, the difference of proportions, $\widehat{p_1} - \widehat{p_2}$, is approximately normally distributed with mean equal to $p_1 - p_2$. With independent observations, the variance for the mean of the treatment arm is $\frac{1}{n}(p_1(1-p_1))$, and the variance for the mean of the control arm is $\frac{1}{n}(p_2(1-p_2))$, where n is the sample size in each arm (here we assume that the two sample sizes are equal). Thus, the variance for $\widehat{p_1} - \widehat{p_2}$ is:

$$\sigma^2 = \frac{1}{n}(p_1(1-p_1) + p_2(1-p_2)) \quad (1)$$

The two-sample t-test for this situation has the following test statistic[6]:

$$t = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\frac{1}{2}(\widehat{p_1}(1-\widehat{p_1}) + \widehat{p_2}(1-\widehat{p_2}))}\sqrt{\frac{2}{n}}} \quad (2)$$

The sample size equation for the two-sample t-test follows from the above test statistic:

$$N = 2 * \left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2 \frac{(p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2} \quad (3)$$

where $Z_{\alpha/2}$ is the critical value for a two-sided type-I error rate of α; Z_{β} is the critical value for a type-II error rate of β; and N is the total sample size for the treatment and the control group together, assuming equal sample sizes[7].

The T-BOSS method inflates the expected standard error of $\widehat{p_1} - \widehat{p_2}$ by 5% to account for the expected loss of efficiency due to the cluster randomized crossover study design. The expected variance for $\widehat{p_1} - \widehat{p_2}$ is:

$$\sigma^2_{TBOSS} = 1.05^2 * \frac{1}{n}(p_1(1-p_1) + p_2(1-p_2)) \quad (4)$$

Therefore, the corresponding t-test is:

$$t_{TBOSS} = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{1.05^2 * \frac{1}{2}(\widehat{p_1}(1-\widehat{p_1}) + \widehat{p_2}(1-\widehat{p_2}))}\sqrt{\frac{2}{n}}} \quad (5)$$

Using this new test statistic, the sample size needed to obtain a specific power follows:

$$N_{TBOSS} = 2 * \left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2 \frac{(p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2} * 1.05^2 \quad (6)$$

Note that this equation (6) is equation (3) multiplied by 1.05^2 to account for the expected loss of efficiency resulting from the cluster randomized crossover design. Also, note that the above variance in equation (4)

and equation (6) was used only in the calculation for sample size, and the corresponding test statistic in equation (5) will not be used to analyze the data upon study completion.

Gold Standard Closed Form Equation

The current gold standard for power and sample size calculations in cluster randomized crossover trials with a binary outcome is the following closed form equation[4]:

$$N_{CF} = 2km = 2\left(Z_{\frac{\alpha}{2}} + Z_{\beta}\right)^2 \frac{(p_1(1-p_1)+p_2(1-p_2))}{(p_1-p_2)^2}[1+(m-1)\rho - m\rho_{12}] \quad (7)$$

where k is the number of clusters, m is the average cluster size per treatment arm, and 2km is the total sample size. Here, ρ is the intraclass correlation (ICC): the correlation among patients within the same cluster and same period. Additionally, ρ_{12} is the inter-period correlation (IPC): the correlation among patients within the same cluster but within different periods. If we assume no period effect, the ICC will equal the IPC.

There are four important things to note about the above equation. First, this equation multiplies equation (3) by an inflation factor,$[1+(m-1)\rho - m\rho_{12}]$, instead of the 1.05^2 chosen in the T-BOSS equation[3]. Second, if there is no period effect, then the ICC will equal the IPC and the inflation factor reduces to$[1-\rho]$, implying that the cluster randomized crossover design is more efficient than an individually randomized trial in this case[3]. Third, if the ICC and IPC both equal 0, then the inflation factor equals 1 and the equation reduces to the original equation (3) used for studies with individual randomization[3]. This is the case if the subjects within a cluster are not correlated with each other (but does not necessitate that the subjects are independent). Finally, if the inflation factor is larger than 1, then the sample size needed for the cluster randomized crossover study will be larger than the sample size needed for an individually randomized study. For example, the T-BOSS inflation factor of 1.1025 consistently calls for a sample size that is 10.25% larger than the sample size needed in an individually randomized study. Inflation factors of less than 1 require a sample size smaller than that needed in an individually randomized study. This can happen if there is no period effect and the ICC equals the IPC, or

if there is a very small period effect and a very small average cluster size which causes $m * (\rho - \rho_{12})$ to be less than ρ.

Study Scenarios

The two sample size calculation methods were evaluated in two hypothetical study scenarios. The first is representative of a ROC study that has an outcome of functional survival to hospital discharge, defined as having a Modified Rankin Score of less than 4 or a Cerebral Performance Category of 1 or 2. In all treated cardiac arrest patients, the survival rate with satisfactory functional status is about 5%[8]. The second scenario is representative of a ROC study that has an outcome of survival to hospital discharge among patients with a shockable initial rhythm. The overall survival rate among this subset of cardiac arrest patients is about 25%[9]. There is no period effect expected in the ROC setting, so the ICC will be equal to the IPC. For both scenarios, intracluster correlation and inter-period correlation were estimated to be about 0.04 in previous ROC data[8]. Each scenario in the main analysis had several parameters fixed as outlined below. Both studies had two periods. All cluster-periods were equally sized. The settings focused on low, medium, and high values of effect size, corresponding to odds ratios of 1.1, 1.2, and 1.25 or 1.3. ICC and IPC (equal in this setting) were also explored at low, medium and high values of 0.01, 0.04, and 0.20. Neither T-BOSS nor the closed form equation is guaranteed to provide overall sample size calculations with equally sized cluster periods, so the following equation was used to provide a sample size estimate consistent with that requirement (with 2 periods and 200 clusters):

$$New\ Sample\ Size = Round(Original\ Sample\ Size/2/200) * 2 * 200 \quad (8)$$

This requirement of equally sized cluster periods was present in the simulation method, so it was implemented in these methods as well.

Results

In Figure 1, we examine the relationship of the inflation factor with cluster size when there is no period effect present, as is assumed for the ROC scenarios. This causes the intracluster correlation to equal the inter-period correlation. In this setting, the closed form inflation factor reduces to $[1-\rho]$, eliminating cluster size from the equation[3]. Thus, no matter what the cluster size, the closed form inflation factor remains the same. With positive ICC, the closed form sample size in this setting will always be less than the sample size in an individually randomized study, as the inflation factor is always less than 1. As the ICC and IPC increase and individuals in clusters become more correlated with each other, the inflation factor decreases and reduces the sample size needed. On the other hand, the T-BOSS inflation factor is fixed at 1.1025, consistently using a sample size 10.25% larger than the sample size for an individually randomized study – a conservative approach compared to the closed form method when there is no period effect.

Table 1 compares the sample size estimates for both scenarios using 2 periods and 200 clusters to calculate a sample size that ensures equally sized cluster-periods. Again, there is no period effect present, so regardless of cluster size, the T-BOSS sample size will always hold the same relative relationship to the closed form equation sample size. This relationship can be expressed in the following equation, using the individually randomized sample size (IRSS) as a constant:

$$\frac{TBOSS\ Sample\ Size - Closed\ Form\ Sample\ Size}{Closed\ Form\ Sample\ Size} = \frac{IRSS*1.1025 - IRSS*(1-\rho)}{IRSS*(1-\rho)} = \frac{0.1025+\rho}{1-\rho} \qquad (9)$$

This results in a relative increase in sample size for T-BOSS compared to the closed form equation of 11.4%, 14.8%, and 37.8% for ICCs of 0.01, 0.04, and 0.20, respectively.

When using equation (8) to guarantee equally sized cluster-periods, the resulting relative increase in sample size changes slightly but in most cases falls near the above values. One notable exception occurs for an odds ratio of 1.3, ICC of 0.01, and 25% baseline survival rate: instead of the 11.4% expected relative increase, the rounding results in a T-BOSS sample size that equals the closed form sample size. This is more likely to occur when there is a higher number of clusters and a smaller

overall sample size – to ensure equally sized cluster-periods, one must add one person to each cluster-period if the sample size needs to be increased. In the case of 200 clusters, this means the minimum step-wise increase in sample size is 400 patients. The unrounded sample size is 3,374 for T-BOSS and 3,029 for the closed form: with a difference of less than 400 patients, this causes the rounded sample sizes to be equal.

The simulation method generally provides sample size estimates that are quite close to the closed from equation method in the 5% baseline survival setting, while the estimates in the 25% baseline survival setting are less predictable compared to the closed form equation (results not shown). The differences between the simulations and the closed form equation sample size estimates may be due to mere sampling error, or they could be reflective of fundamental differences in the underlying models used for both methods. Specifically, the underlying model in the simulation method is a hierarchical model[5], while the underlying model in the closed form equation uses cluster-level analysis[3].

In the CCC trial, ROC statisticians hypothesized a survival to discharge rate of 8.07% in the 30:2 group and powered the study to detect a survival to discharge rate of 9.37% or higher in the CCC group. For this setting, T-BOSS provides a sample size of 21,815 for a study with 90% power and a 5% Type-I error rate. With an estimated ICC and IPC of 0.06, the closed form equation provides a sample size of 18,599. The T-BOSS sample size is 17.3% larger than the closed form sample size, just as expected for a study with an ICC of 0.06 and no period effect. The T-BOSS sample size of 21,815 matches the closed form sample size given with an ICC of 0.06, no period effect, and 94% power. With around 200 EMS agency clusters, the average cluster size is around 109 patients for T-BOSS and 93 for the closed form equation. If we instead choose to use an ICC and IPC of 0.09, the upper limit of the 95% confidence interval for the estimated ICC in this setting, the closed form equation provides a sample size of 18,006. This corresponds to T-BOSS sample size that is 21.2% larger than the closed form sample size. Using the lower limit of 0.04, the closed form equation sample size is 18,995, corresponding to a T-BOSS sample size that is 14.8% larger.

Although T-BOSS is conservative when there is no period effect, this is rarely the case when a period effect is present. As seen in Figures 2a, 2b, and 2c, the inflation factor escalates as the period

effect increases and the inter-period correlation correspondingly decreases compared to the ICC. The inflation factor rises at a faster rate when the cluster size is larger and the number of clusters is smaller. The largest inflation factor occurs when the cluster size is large, the ICC is high, and the IPC is low, with the inflation factor rising to over 200 when the ICC is 0.20, the IPC is below 0.04, and the average cluster size is 1,300 (Figure 2c). T-BOSS, comparatively, is nearly always anti-conservative with the exception of when the IPC is close to the ICC. For example, when the ICC is 0.20, T-BOSS is conservative only when the IPC is greater than 0.1998, 0.1985, and 0.1942 for cluster sizes of 1,300, 208, and 52, respectively (Table 2a).

We can see similar effects in Figures 3a, 3b, and 3c, where we hold ICC and IPC constant and vary cluster size. As cluster size increases and the number of clusters correspondingly decreases, the inflation factor rises in the closed form equation. The rate at which it rises increases as the IPC moves farther away from the ICC. Again, T-BOSS is nearly always anti-conservative except when cluster size is relatively small. For example, if the ICC is 0.04, T-BOSS is conservative only when the cluster size is less than 4, 5, 7, and 14 for IPCs of 0, 0.01, 0.02, and 0.03, respectively (Table 2b).

Discussion

For the CCC study in which it was used, the T-BOSS method was a conservative but acceptable approach for several reasons. First, since the study was implemented at such a large scale (roughly 20,000 patients across over 200 EMS agencies), the addition of 3,216 more patients in the sample size would not have made a significant impact on study time or cost. Second, the T-BOSS sample size corresponds to the sample size given by the closed form equation when the average cluster size is 93, the ICC is 0.06, and the IPC 0.056. Allowing for this small period effect in our setting seems reasonable. More importantly, the CCC sample size incorporated the T-BOSS method within a calculation that also allowed for several interim analyses, so the addition of more patients will only be seen if the differences observed between treatment arms are such that the study continues to maximum enrollment. The sample size comparisons here do not incorporate interim analyses, but are a good reference for the comparison

between these two methods in the interim analysis setting. This reduces the ethical issues that arise when potentially enrolling many more people than necessary, as it would only happen if the difference in efficacy between treatment arms was little to none. Additionally, we must remain aware that all sample size calculations are inherently imprecise due to their reliance on assumptions, so small differences in sample sizes do not necessarily reflect an unethical situation.

In other settings, the appropriateness of T-BOSS depends on a combination of study settings and value judgments. For instance, when the ICC and IPC are equal and both are low (0.01), and when the overall sample size is expected to be low due to low power or high effect size, a relative increase of 11.4% in the T-BOSS calculation may not be meaningfully different from the closed form sample size. However, if the treatment is expected to provide a substantial benefit and no interim analyses are planned, it may be unethical if too many additional patients are enrolled. When there is a period effect present, T-BOSS appears to almost never be an acceptable alternative unless the period effect is small and the IPC is close to the ICC, or average cluster size is very small. This is reassuring, because if there was a small period effect in the ROC setting, the T-BOSS method would likely still be acceptable.

Conclusions

In the present day, there is no longer a need for T-BOSS or the simulation approach in settings with two periods and approximately equal cluster sizes because the gold standard closed form equation is equally fast and capable of being incorporated into a study with planned interim analyses. The simulations were found to be somewhat close to the closed form solution depending on the study scenario, but they took a much longer time to compute and showed more variability in the estimates. However, they may still be useful in more complex settings where the cluster sizes vary widely, there are many periods, there are multiple treatment groups, or there are other factors that cannot be incorporated into the underlying model of cluster-level analysis that is used in the closed form equation. The assumption of equal cluster sizes that is present in this paper is a notable limitation, especially because

the ROC cluster sizes vary widely. The effect of unequal cluster sizes on study power should be explored in the future.

At the time of planning the ROC study, T-BOSS was an acceptable alternative to using simulations and the closed form equation to estimate the sample size needed for the CCC study. This is largely due to the fact that there is no period effect hypothesized for the study setting, and the ICC is expected to be low based on data from previous ROC studies. Additionally, the ICC is expected to be low because patients in the same cluster will never experience both treatments and may not even be treated by the same EMS providers, even though survival rates can vary substantially across sites[9]. While T-BOSS did provide a conservative sample size estimate compared to the closed form solution, any detrimental effects this could have caused are mitigated by the several interim analyses planned that can stop the study early if a clear difference is seen before all planned patients are enrolled.

Manuscript Table 1: T-BOSS vs. Closed Form Equation Comparison of Sample Sizes, 200 Clusters				
		Fixing Power = 90%		
5% Baseline Survival Rate		Sample Size		
	Odds Ratio	**T-BOSS**	**Closed Form**	**T-BOSS % Increase in Sample Size**
Low Correlation (ICC = IPC = 0.01)	1.1	102,800	92,400	11.3
	1.2	27,200	24,400	11.5
	1.3	12,800	11,200	14.3
Medium Correlation (ICC = IPC = 0.04)	1.1	102,800	89,600	14.7
	1.2	27,200	23,600	15.3
	1.3	12,800	10,800	18.5
High Correlation (ICC = IPC = 0.20)	1.1	102,800	74,800	37.4
	1.2	27,200	19,600	38.8
	1.3	12,800	9,200	39.1
25% Baseline Survival Rate		Sample Size		
	Odds Ratio	**T-BOSS**	**Closed Form**	**T-BOSS % Increase in Sample Size**
Low Correlation (ICC = IPC = 0.01)	1.1	26,400	24,000	10.0
	1.2	7,200	6,400	12.5
	1.3	3,200	3,200	0.0
Medium Correlation (ICC = IPC = 0.04)	1.1	26,400	23,200	13.8
	1.2	7,200	6,000	20.0
	1.3	3,200	2,800	14.3
High Correlation (ICC = IPC = 0.20)	1.1	26,400	19,200	37.5
	1.2	7,200	5,200	38.5
	1.3	3,200	2,400	33.3

Manuscript 2a: When T-BOSS Inflation Factor Equals Closed Form Inflation Factor			
	Inter-Period Correlation When		
Cluster Size	**ICC = 0.01**	**ICC = 0.04**	**ICC = 0.20**
1300	0.0099	0.0399	0.1998
208	0.0095	0.0393	0.1985
52	0.0078	0.0373	0.1942

Manuscript Table 2b: When T-BOSS Inflation Factor Equals Closed Form Inflation Factor			
	Cluster Size When		
Inter-Period Correlation		**ICC = 0.01**	
0		11.25	
0.0025		15	
0.005		22.5	
0.0075		45	
	Cluster Size When		
Inter-Period Correlation		**ICC = 0.04**	
0		3.56	
0.01		4.75	
0.02		7.13	
0.03		14.25	
	Cluster Size When		
Inter-Period Correlation		**ICC = 0.20**	
0		1.51	
0.05		2.02	
0.1		3.03	
0.15		6.05	
0.19		30.25	

Figure 1:

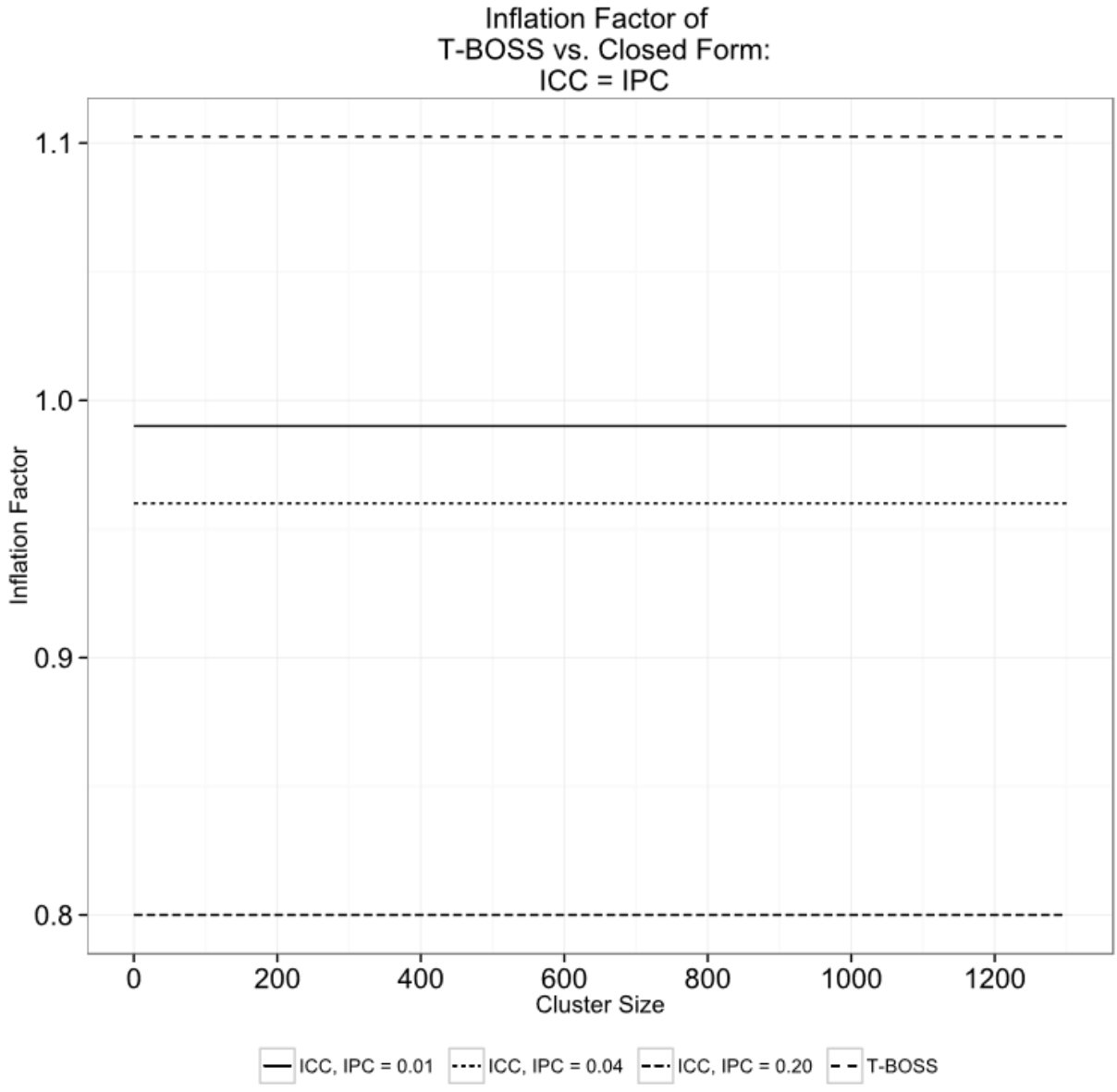

Inflation Factor of
T-BOSS vs. Closed Form:
ICC = IPC
1.1
1.0
0.9
0.8
Inflation Factor
0
200
400
600
800
1000
1200
Cluster Size
ICC, IPC = 0.01
ICC, IPC = 0.04
ICC, IPC = 0.20
T-BOSS

Figure 2a:

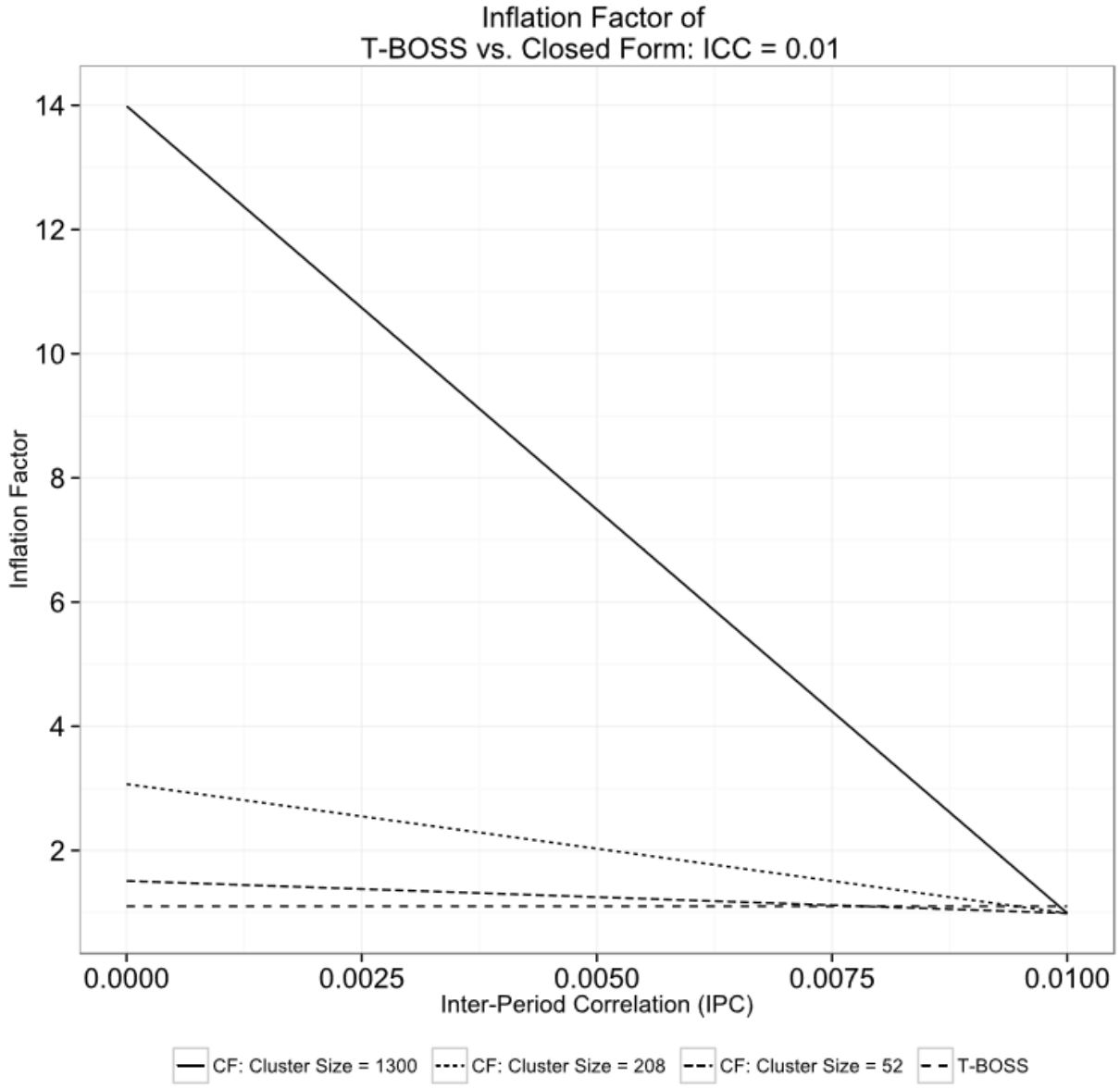

Inflation Factor of
T-BOSS vs. Closed Form: ICC = 0.01
Inflation Factor
14
12
10
8
6
4
2
0.0000
0.0025
0.0050
0.0075
0.0100
Inter-Period Correlation (IPC)
CF: Cluster Size = 1300
CF: Cluster Size = 208
CF: Cluster Size = 52
T-BOSS

Figure 2b:

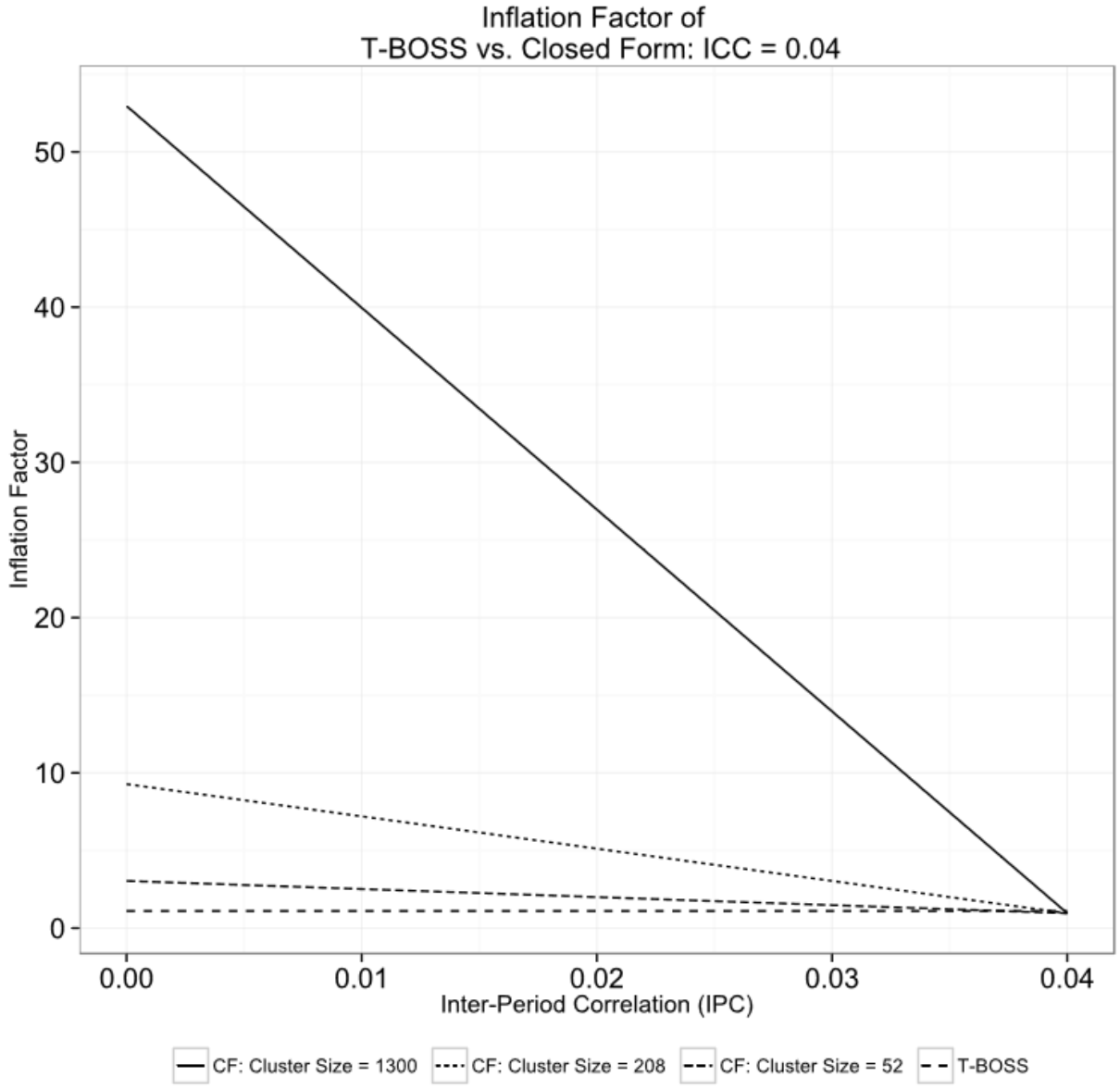

Inflation Factor of
T-BOSS vs. Closed Form: ICC = 0.04
Inflation Factor
50
40
30
20
10
0
0.00
0.01
0.02
0.03
0.04
Inter-Period Correlation (IPC)
CF: Cluster Size = 1300
CF: Cluster Size = 208
CF: Cluster Size = 52
T-BOSS

Figure 2c:

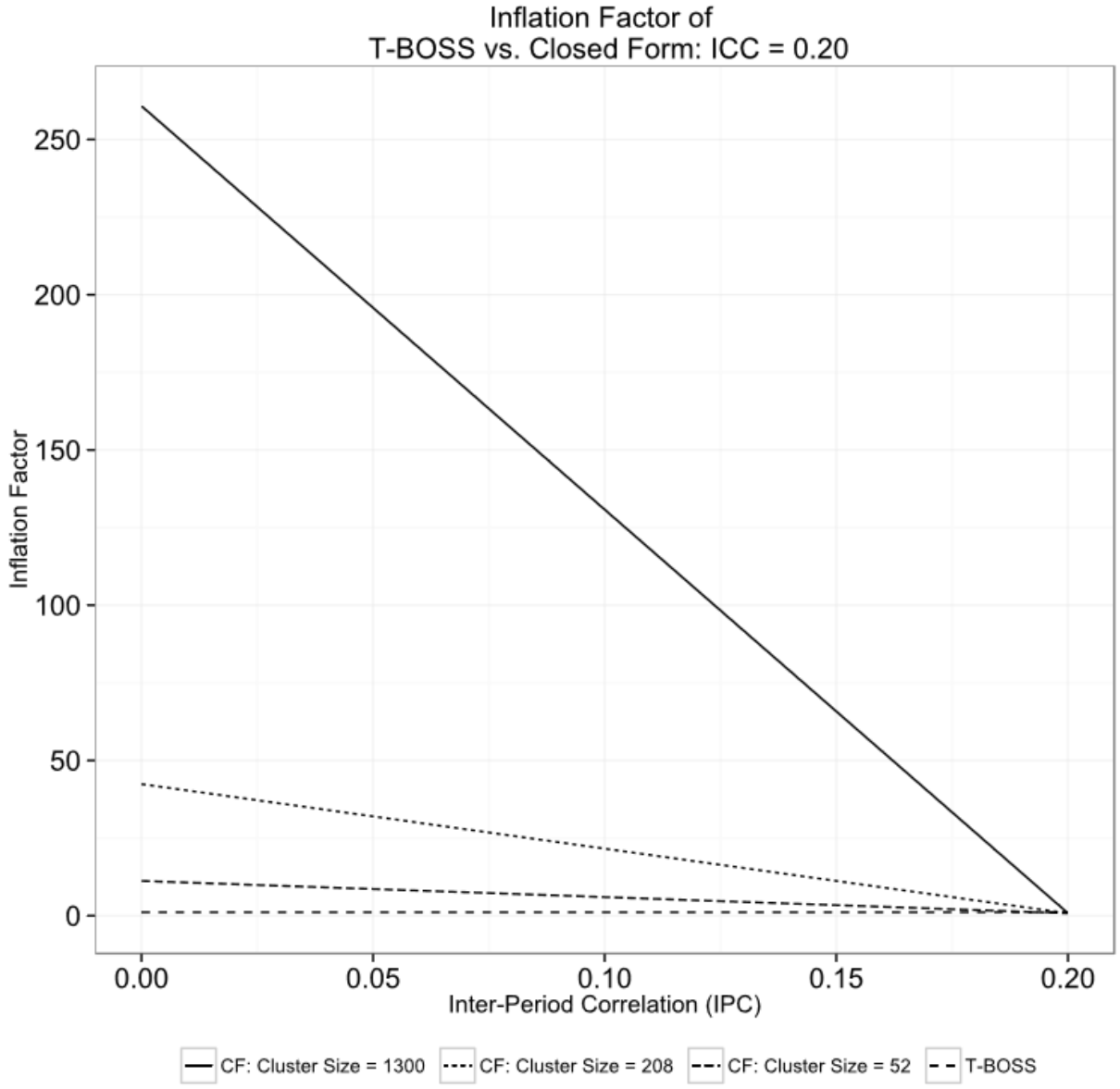

Inflation Factor of
T-BOSS vs. Closed Form: ICC = 0.20
Inflation Factor
250
200
150
100
50
0
0.00
0.05
0.10
0.15
0.20
Inter-Period Correlation (IPC)
CF: Cluster Size = 1300
CF: Cluster Size = 208
CF: Cluster Size = 52
T-BOSS

Figure 3a:

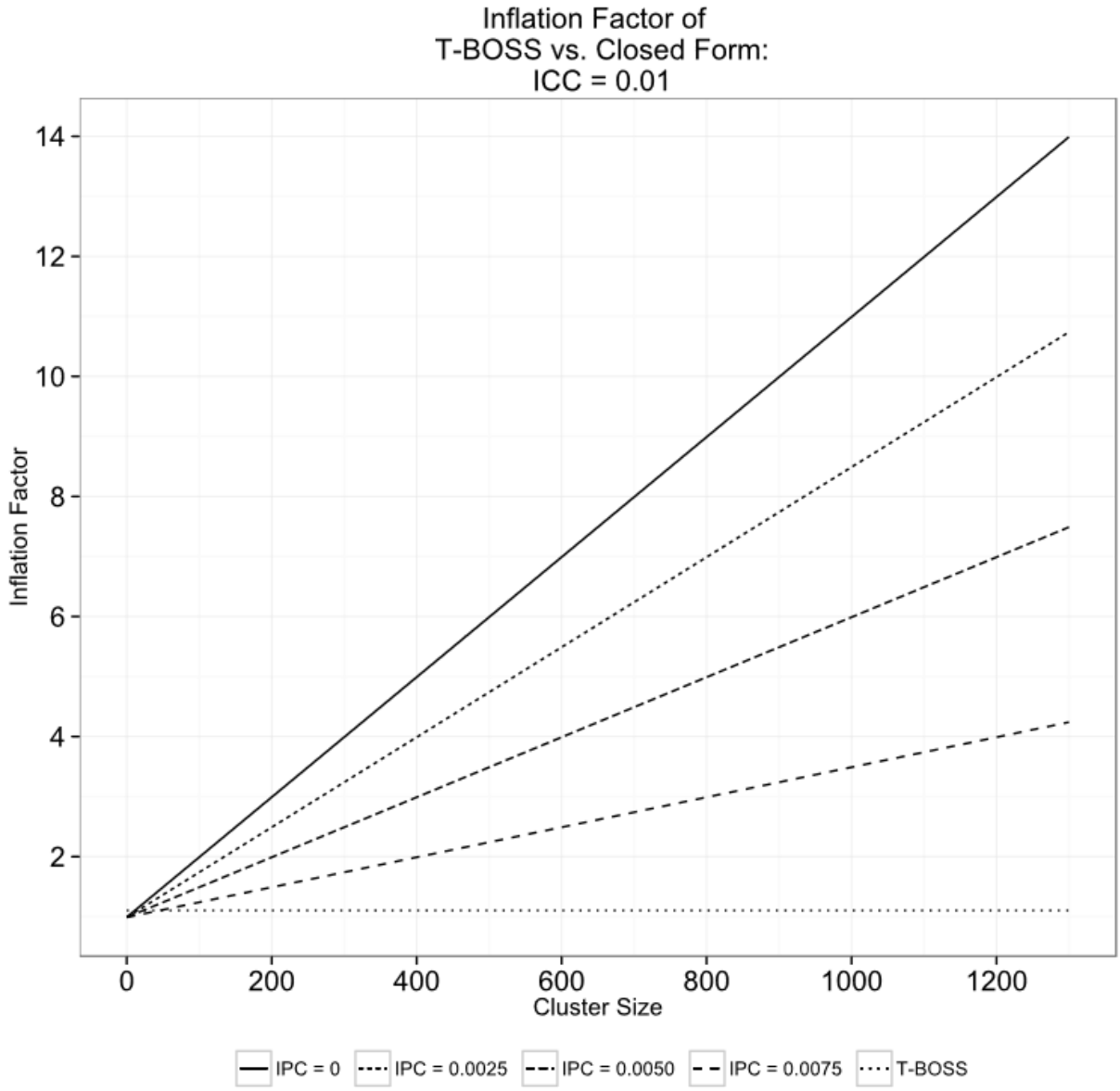

Inflation Factor of
T-BOSS vs. Closed Form:
ICC = 0.01
Inflation Factor
14
12
10
8
6
4
2
0
200
400
600
800
1000
1200
Cluster Size
IPC = 0
IPC = 0.0025
IPC = 0.0050
IPC = 0.0075
T-BOSS

Figure 3b:

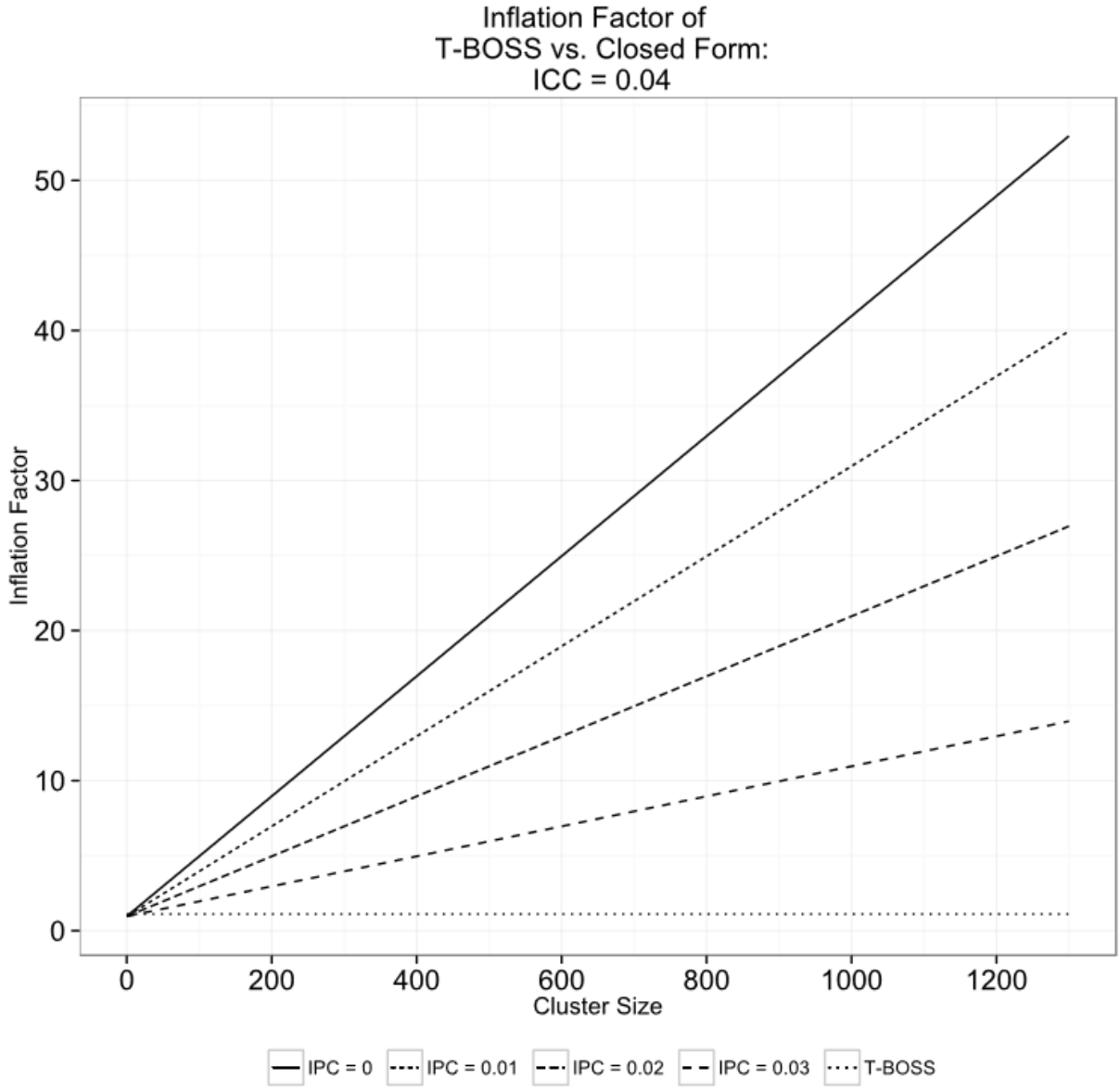
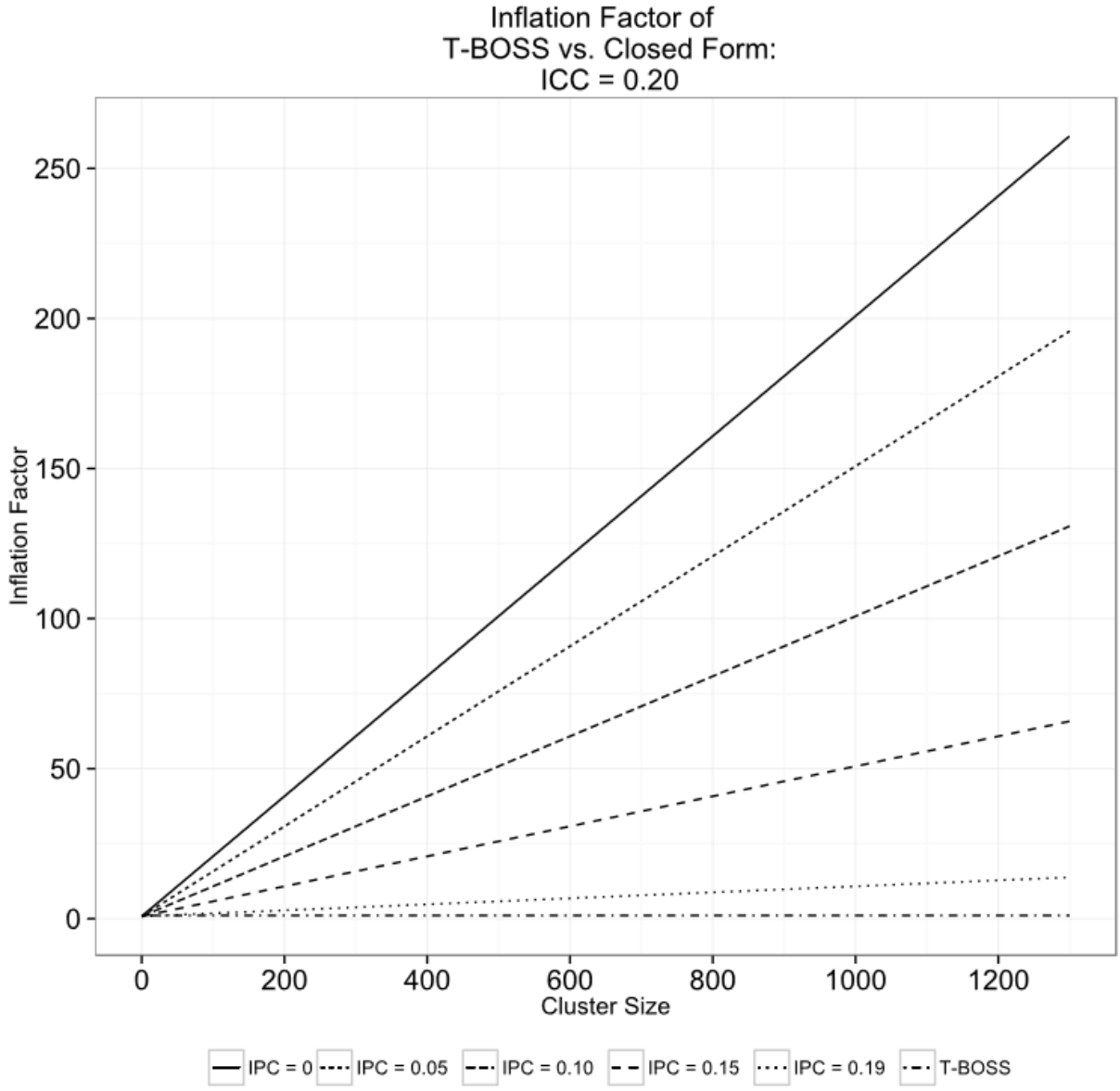

Inflation Factor of
T-BOSS vs. Closed Form:
ICC = 0.04
Inflation Factor
0
10
20
30
40
50
0
200
400
600
800
1000
1200
Cluster Size
IPC = 0
IPC = 0.01
IPC = 0.02
IPC = 0.03
T-BOSS

Figure 3c:


Inflation Factor of
T-BOSS vs. Closed Form:
ICC = 0.20
Inflation Factor
250
200
150
100
50
0
0
200
400
600
800
1000
1200
Cluster Size
IPC = 0
IPC = 0.05
IPC = 0.10
IPC = 0.15
IPC = 0.19
T-BOSS

Appendix

Description of Cluster Randomized Crossover Design[10]

In a cluster randomized crossover trial, each cluster is assigned to every treatment arm at different points in time ("periods"). The clusters are randomized to one of the possible sequences of treatments given. In the simplest example, a study comparing treatment A to treatment B using two periods would randomly assign each cluster to either receive treatment A first (and "crossover" to treatment B later) or to receive treatment B first (and crossover to treatment A later). In crossover studies without cluster randomization, each individual experiences both treatments. Some cluster randomized crossover trials will also have each individual experience both treatments. For example, a study of patients with a chronic disease at different treatment centers may have each cluster (treatment center) provide a sequence of treatments to these patients during the study periods, with a washout period in between each treatment. Alternatively, a cluster can experience both treatments but an individual will only experience one treatment, depending on when they enter the cluster. For example, a cluster can be defined as a hospital. For 6 months, the hospital staff will administer treatment A to all incoming patients. Then, they will administer treatment B to all incoming patients for the next 6 months. Assuming a patient only visits the hospital once in this 12-month period, each patient will only be assigned to one treatment arm.

Description of the Resuscitation Outcomes Consortium Setting

The Resuscitation Outcomes Consortium (ROC) is a network of Regional Clinical Centers and associated Emergency Medical Service (EMS) agencies across ten regions in the United States and Canada[11]. Each region is composed of numerous EMS agencies, which typically serve as the units of randomization in the cluster randomized crossover trials that are implemented by ROC. A current cluster randomized crossover trial seeks to determine whether there is a difference between two different methods of cardiopulmonary resuscitation (CPR)[2]. In the first method, continuous chest compressions

("CCC"), EMS responders continually provide chest compressions without stopping to provide rescue breaths. In the second, abbreviated "30:2", EMS responders provide thirty chest compressions followed by two rescue breaths and repeat the sequence until the study period ends or return of spontaneous circulation.

Previously, ROC implemented a cluster randomized crossover trial to evaluate two strategies of implementing CPR after out-of-hospital cardiac arrest[8]. The first strategy, "early analysis", analyzed a patient's cardiac rhythm after 30 to 60 seconds of CPR administered by EMS agents. The second, "late analysis", analyzed a patient's cardiac rhythm after 180 seconds of CPR. This trial, part of a larger partial factorial ROC trial called the PRIMED trial, used the same outcomes as the current CCC trial: survival to hospital discharge and functional survival. It also used the same ROC network of EMS agencies that served as clusters in the cluster randomization. These data were used to determine the between-cluster variance/ICC for both outcomes that were used as the best estimate of these parameters in the main analysis of this research.

Barriers to Implementation of Individual Randomization

In the ROC setting, both methods of CPR require EMS responders to be trained in the method before they can perform it successfully in the field and it is thought to be difficult to switch between approaches as would be required with individual randomization. Therefore, individual randomization at the time of cardiac arrest is not feasible. Instead, each EMS agency is randomized to one of the two methods during a given study period and all EMS responders in an agency are trained in the assigned method. A typical cluster randomized crossover design incorporates a "washout period", which in the ROC setting potentially could be a two week period in which no treatment is given, to prevent carryover of the effect between the treatment groups[10]. Ideally, the EMS agencies would have these two weeks to retrain the EMS responders to use the newly assigned CPR method before data collection resumed. However, it is not possible in the ROC setting to not use one of the treatment methods in a washout period. Instead, for a pre-specified length of time, the data following the end of a treatment assignment

could be excluded from analysis as a conceptual "washout" period. However, no significant carryover effect was found in the ROC PRIMED study described above[17], so cardiac arrests occurring after a crossover were not excluded. As a result, there may be some contamination in compliance to treatment assignment at the beginning of a new study period.

In addition to providing an alternative to individual randomization when the intervention requires training of treatment providers, cluster randomized crossover trials are also advantageous when the intervention must be applied to a group of people. One example is the use of this design to assign treatments to classrooms. As an example, a cluster randomized crossover trial was conducted in Australia in 22 schools to compare the respiratory health effects of two different types of heaters[12]. Clusters were defined as each classroom, and each classroom was randomized to have one of two heaters on during a given 6-week period. Here, individual randomization of students to a given heater would not be possible due to the heater serving the entire classroom.

Analyzing a Cluster Randomized Crossover Trial

Below, we describe three of the most common models that can be used for the analysis of correlated data from a cluster randomized crossover trial. The first, cluster-level analysis, examines the association of the treatment with the outcome using cluster-level summary measures[1]. The second, a Generalized Estimating Equation (GEE), is a marginal model that examines the association of the treatment with the outcome over the entire study population[13]. The third, a Mixed Effects Model, is a conditional model that examines the association of the treatment with the outcome within each cluster[14]. A description of the three methods of analysis is provided below.

Cluster-Level Analysis[1]

Instead of fitting a model to data points each representing an individual in the study, cluster-level analysis fits a model using independent data points for each cluster. In the simple setting of a two-period

cluster-randomized crossover trial, each cluster is assigned to a period of treatment followed by a period of control, or a period of control followed by a period of treatment. The difference in outcome rates for cluster j is calculated as:

$$d_j = w_j * (p_{j1} - p_{j2})$$

where d_j is the difference in outcome rates for cluster j; w_j is 1 if period 1 is assigned to treatment, -1 if period 1 is assigned to control; and p_{jk} is the outcome rate in cluster j during period k.

Once all $d_j's$ are calculated, a linear regression is fit to the data points with the following form:

$$d_j = \beta_0 + \beta_1 * w_j + \epsilon_j$$

The estimate of the treatment effect is β_0 and is calculated by ($\overline{d_1}$ + $\overline{d_2}$)/2, where $\overline{d_k}$ is the average difference in outcome rates among all clusters with the treatment assigned during period k. The test statistic for this estimate is then compared to the t-distribution with J-2 degrees of freedom, with J as the total number of clusters in the study. Alternative forms of the regression include weights for each cluster based on cluster size or the combination of cluster size and intracluster correlation.

Description of Mixed Effects Models[14]

In order to describe the form of a generalized mixed effects regression, we must define the following parameters that describe the observed data:

- Y_{ij}: outcome for subject i in cluster j

  In our application, the outcome is binary. Examples can include survival/death, cancer recurrence/remission, or disease/no disease.

- X_{ij}: predictor of interest for subject i in cluster j

  In many cases, this parameter is also binary and indicates treatment/control group assignment (typically coded 1/0, respectively).

The mean model for a mixed effects logistic regression, used often for cluster-randomized crossover trials with a binary outcome, has the following form:

$$Log\left(odds\left(Y_{ij}\middle|X_{ij}\right)\right) = \beta_0 + \beta_1 X_{ij} + w_j$$

The following parameters are estimated by the model:

- β_0: log(odds of experiencing the outcome in the control group $in\ an\ average\ cluster$)
- β_1: $\log\left(\begin{matrix}\text{odds ratio of experiencing the outcome in the treatment group compared to the control group,}\\ in\ a\ particular\ cluster\end{matrix}\right)$
- w_j: random effect for cluster j

  This parameter allows for a different intercept on the log odds scale for each cluster, with the assumption that some clusters may experience higher or lower odds of experiencing the outcome than average. $w_j \sim N(0, \sigma_w^2)$

Description of Generalized Estimating Equations[13,14]

In order to describe the form of a generalized GEE regression, the following parameters that describe the observed data are the same as described in the mixed effects model:

- Y_{ij}: outcome for subject i in cluster j
- X_{ij}: predictor of interest for subject i in cluster j

The mean model for a GEE Logistic Regression, also used often for cluster-randomized crossover trials with a binary outcome, has the following form:

$$Log\left(odds\left(Y_{ij}\middle|X_{ij}\right)\right) = \beta_0 + \beta_1 X_{ij}$$

The following parameters are estimated by the model:

- β_0: $\log(\text{odds of experiencing the outcome in the control group}, \textit{among the whole sample})$

- β_1: $\log\left(\begin{matrix}\text{odds ratio of experiencing the outcome in the treatment group compared to the control group,} \\ \textit{among the whole sample}\end{matrix}\right)$

Other options for modeling a binary outcome include using a log link or an identity link instead of the logit link, which model the relative risk and risk difference, respectively. These links can be used in the Mixed Effects Model as well.

To use a GEE regression, a working correlation matrix must be specified to describe how the observations are assumed to be correlated with each other in a given cluster. This working correlation matrix can take one of the following forms: independent, exchangeable, auto-regressive, stationary, and unstructured. Each of these correlation structures will affect the weighting of the regression estimates, as well as the standard errors that are estimated.

Independent correlation structures assume that all observations within a cluster are independent. Because the GEE weights the treatment effect estimate differently depending on the correlation structure used, this correlation structure uses no weights and allows us to estimate the raw risk difference between the treatment and control group. The following is an example of the correlation matrix used for a cluster with three observations:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Exchangeable correlation structures assume that all observations within a cluster are equally correlated with each other. This is the correlation structure that is most believable when we are modeling cluster randomized studies.

$$\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$$

Auto-regressive correlation structures assume that observations farther apart are exponentially less correlated with each other. This is useful for when each cluster represents multiple observations on the same person, with observations further apart representing observations further apart in time.

$$\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}$$

Stationary correlation structures assume that observations next to each other are equally correlated with each other, and observations not next to each other are independent of each other.

$$\begin{pmatrix} 1 & \alpha & 0 \\ \alpha & 1 & \alpha \\ 0 & \alpha & 1 \end{pmatrix}$$

Unstructured correlation structures estimate a correlation for each pair of observations separately, with no pattern assumed.

$$\begin{pmatrix} 1 & \alpha & \epsilon \\ \beta & 1 & \gamma \\ \rho & \phi & 1 \end{pmatrix}$$

Comparison of Mixed Effects to GEE[13,14]

Note that GEE regression does not estimate a random effect, w_j, for each cluster. This random effect requires the mixed effects model to produce "conditional estimates". The β estimates provided by mixed effects models are interpreted as the estimated parameter within a particular cluster. In contrast, GEE regression provides "marginal estimates". The β estimates are interpreted as the estimated parameter within the entire study population.

Description of the Simulation Program Used to Calculate Power

In addition to the closed form equation, simulation programs have been developed for estimating the power of a study given several parameters such as number of clusters, number of periods, between

cluster variation, and period effect[5]. Our simulation, written in R, was based on the underlying model used by the existing simulation package.

The simulations calculate power using a mixed effects model to generate and analyze the data. To conduct power simulations, one must first simulate data that conforms to the assumptions that are made in a cluster-randomized crossover trial. First, data in clusters must be correlated, and clusters must be independent of each other. Further, each treatment arm must be applied in each cluster at different "times", with the potential for a period effect. A period effect describes the phenomenon that observations within the same treatment period may be correlated due to time-dependent effects, such as seasonal differences that are sometimes observed in event rates. The simulation uses the following model to generate the probability of an event for a given individual:

$$Log\left(odds\left(Y_{jk}\middle|X_{jk}\right)\right) = \beta_0 + \beta_1 X_{jk} + w_j + v_k$$

This model is similar to the one described in the mixed effects section, with a few notable differences:

- $w_j \sim N(0, \sigma_w^2)$, where σ_w^2 must be specified as the expected between cluster variation on the logistic scale.
- v_k, a period effect, is added as an additional random effect. In many cases, there will be no period effect expected. In all other cases, this parameter will have to be specified, with $v_k \sim N(\mu_v, \sigma_v^2)$.

Once this model is used to generate the probability of an event for all individuals in a given period and cluster, outcomes for each individual are generated randomly from a Bernoulli distribution with mean equal to the probability of Y_{jk}.

The intra-class correlation coefficient (ICC) can be approximated from the between cluster variation, σ_w^2, in this model by the following equation[15]:

$$ICC \approx \frac{\sigma_w^2}{\sigma_w^2 + \frac{\pi^2}{3}} \qquad (10)$$

Each power simulation conducted here consists of 1,000 trials. After a dataset is randomly generated using the above model in each trial, a logistic mixed effects model is fit using the same form. The resulting power estimate for a given sample size is the proportion of trials with a statistically significant estimate for β_1. To estimate the sample size needed for a given power, one must repeat the power simulation multiple times, each time adjusting the sample size to get closer to the desired power. The algorithm used to determine the sample size was based on the binary search algorithm[16], and consisted of the following steps:

1. Set first guess to the T-BOSS sample size estimate, rounded down to ensure equally sized cluster-periods by replacing "round" with "floor" in equation (8) in the manuscript.
2. Set upper limit to $1.5 * [TBOSS\ Sample\ Size]$.
3. Set lower limit to $0.1 * [TBOSS\ Sample\ Size]$.
4. Calculate power for first guess.
5. If power is between 89% and 91%, then stop. The current guess is the sample size estimate.
   If the power is not between 89% and 91%, then continue with steps 6-9.
6. If power is below 90%, then set lower limit to current guess.
   If power is above 90%, then set upper limit to current guess.
7. Set new guess to:
$$\frac{high.guess + low.guess}{2}$$
8. Round the new guess to ensure equally sized cluster-periods using equation (8) in the manuscript.
9. Calculate the power for the new guess, then repeat steps 5-8 until power is between 89% and 91% for the current guess, or 24 iterations have passed.

It is important to note that for 50 and 200 clusters in the main table for the second scenario (Appendix Table 5b), the sample size calculations had different parameters in an attempt to increase the precision of

the estimates. Namely, step 5 was changed to allow a stop only if the power was between 89.5% and 90.5%, and the maximum number of iterations was set to 50.

Specification of Parameters

The parameters necessary for each method are outlined in Appendix Table 3. All three methods require the type-I and type-II error rates to be specified in a sample size calculation. In addition, the T-BOSS method only requires the effect size and baseline outcome rate (the treatment outcome rate can be calculated from these two parameters). Optionally, to ensure equally sized cluster-periods, the number of clusters *k* can be used in the sample size calculation as shown in equation (8) in the manuscript.

In contrast, the simulation method requires the following parameters to be specified for each sample size calculation: effect size, baseline outcome rate, number of clusters, between cluster variance, and period effect. Because each trial in the simulation generates a dataset with equally sized cluster-periods, no post-hoc modification to the sample size estimate is necessary. The closed form equation essentially requires the same information as the simulation method, except it requires the intracluster correlation instead of the between cluster variance and the inter-period correlation instead of the period effect and period effect variation. This method, too, requires the use of equation (8) to ensure equally sized cluster-periods.

Appendix Tables 4a and 4b outline the specific values used for each method in both scenarios. For each sample size calculation, the type-I error rate was fixed at 5% and the power was fixed at 90%. Each combination of low, medium, and high values of effect size, number of clusters, and between cluster variation (corresponding to the ICC and IPC) were explored, and the values were based on what was most likely to be seen in real life study settings. Because the ROC setting does not have a period effect and our main goal was to explore the implications for ROC specifically, the period effect was set to zero for both scenarios in the main analysis.

Extended Main Results and Discussion

As we described in the manuscript, T-BOSS is always conservative compared to the closed form equation when there is no period effect. Since the cluster size variable is eliminated from the inflation factor when there is no period effect, the comparison of T-BOSS to the closed form equation is the same regardless of the number of clusters used. While the power calculations were similar between the simulations and the closed form solution, the simulations showed more variability and made it harder to detect patterns among the power and sample size estimates.

The power calculations in Tables 5a and 5b show the power estimate resulting from solving the given sample size equation for power, or, as is the case with the simulations, show the proportion of trials in the simulation that resulted in a rejection of the null hypothesis when the null hypothesis was false. For the power calculations in the first scenario, T-BOSS shows a range of values for the absolute decrease in power compared to the closed form equation, from 1.5% when the odds ratio was 1.1 and the ICC was 0.01, to 14% when the odds ratio was 1.2 and the ICC was 0.21 (Appendix Table 5a). For a given between cluster variance (and ICC, and IPC), the largest difference in power estimates occurs when the power was near 50% as seen in Appendix Figures 4a and 4b. In the tables and in Appendix Figure 5, we can see that for a given odds ratio, the difference between power estimates increases as the between cluster variance increases. This make sense, as the inflation factor for the closed form, $[1 - \rho]$, moves farther away from 1.1025 as the ICC, ρ, increases. Finally, Appendix Figure 6 reinforces that, when holding the between cluster variation and the odds ratio constant, the number of clusters has no relationship to the difference in power between T-BOSS and the closed form solution.

Issues Related to the Simulation Program

As we can also see from Appendix Tables 5a and 5b and Appendix Figures 4-6, the simulations usually do a good job of getting close to the power and sample size estimated by the closed form equation. However, the natural variability from the simulations makes it harder to distinguish patterns when comparing the simulations to T-BOSS, especially in the second scenario. While the relationship

between the simulation and T-BOSS sample size estimates is unclear in Appendix Tables 5a and 5b, the closed form solution clearly shows an unchanging relative relationship between the its sample size estimates and the T-BOSS estimates for a given ICC.

In extreme cases, the simulations can produce figures that may be misleading, such as Appendix Figure 7a and to a lesser extent, Appendix Figure 7b. In Appendix Figure 7a, the figure shows the power estimates for all three methods as the between cluster variation increases and other parameters are held constant. First, it is important to note that the scale of this is small, so the differences between the methods look larger than usual. With this in mind, we can see that the simulations have a slightly downward trend, while the closed form solution is increasing. Without the closed form solution, one may interpret this plot as showing a decrease in power as the between cluster variation rises – the opposite of what is happening as told by the closed form equation. These differences could be due to sampling error. Alternatively, it could be caused by differences in the underlying models used for the closed form equation and the simulation method, as the closed form equation is based on a cluster-level analysis model[3] and the simulations are based on a hierarchical model[5].

In Appendix Figure 7b, the trend is the same, but the magnitude of the difference in power between the simulations and T-BOSS is much smaller than the difference between the closed form equation and T-BOSS. This departure of the simulations from the closed form equation could be happening for a number of reasons, a few of which are: the variability of the simulations is causing us to see patterns that are not truly there; the simulations have a numerical precision error hidden deep in the code that are causing the estimates to be less reliable in the second scenario; or, least likely, the closed form equation for continuous outcomes cannot be extended to all binary outcome scenarios.

Additionally, the variability of the simulation sample size estimates occasionally results in a sample size that is different enough from the closed form equation that it could be detrimental to study planning. For example, the simulations in the first scenario for an odds ratio of 1.1, a between cluster variance of 0.15, and 8 clusters resulted in a sample size estimate of 97,120 – 7,504 more patients than the closed form solution estimate of 89,616 (Appendix Table 5a). While this is rare, it could result in a longer running, more expensive study that could also potentially be unethical if there are no interim

analyses planned. These differences may not be detrimental, however, if they are rooted in the differences between the underlying models – in this case, the appropriate sample size should be determined based on the planned analysis method.

An unexpected downfall of the simulations was seen when they didn't converge to the given power in a sample size calculation, as seen in the second scenario simulations with 200 clusters, an odds ratio of 1.3, and between cluster variances of 0.05 and 0.15. This is likely due to the issue explored in the manuscript, where increasing the sample size for 200 clusters and 2 periods can only be done in increments of 400 to ensure equally size cluster-periods. While T-BOSS and the closed form equation have the benefit of estimating the sample size and then rounding to ensure equally sized cluster-periods, the simulations can only estimate the power after ensuring equally sized cluster-periods. This also may have been influenced by the changing of the precision required in the sample size algorithm noted above, where the power estimates for a given sample size were changed from being required to be between 89% and 91% to being between 89.5% and 90.5%. Of course, the power of the T-BOSS and closed form equation sample sizes changes when they are rounded to ensure equally sized cluster-periods and may not lie between 89.5% and 90.5% either. Usually, the rounding causes a minimal difference in power. A more extreme example can be seen in the closed form equation sample sizes for a 25% baseline survival rate, odds ratio of 1.25, and ICC of 0.21: rounding to ensure equally sized cluster periods for 8 clusters results in a sample size of 3,376 with 90% power, but rounding for 200 clusters gives a sample size of 3,200 with 88.4% power.

Draft Appendix Table 3: Parameters Required for Each Sample Size Calculation Method			
	T-BOSS	Simulation	Closed Form Equation
Type-I Error Rate	✓	✓	✓
Type-II Error Rate	✓	✓	✓
Effect Size	✓	✓	✓
Baseline Outcome Rate	✓	✓	✓
Treatment Outcome Rate	✓	✓	✓
(Calculated from Effect Size)			
Number of Clusters	Optional	✓	✓
Between Cluster Variance		✓	
Period Effect		✓	
Variance of Period Effect		✓	
Intracluster Correlation (ICC)			✓
Inter-Period Correlation (ICC)			✓

Appendix Table 4a: Scenario 1 Parameters (Stars indicate scenario-specific parameters)			
	T-BOSS	Simulation	Closed Form Equation
Type-I Error Rate	0.05	0.05	0.05
Type-II Error Rate	0.9	0.9	0.9
Effect Size*	0.004, 0.009, 0.014	1.1, 1.2, 1.3	0.004, 0.009, 0.014
	(risk difference)	(odds ratio)	(risk difference)
Baseline Outcome Rate*	0.05	0.05	0.05
Treatment Outcome Rate*	0.055, 0.059, 0.064	0.055, 0.059, 0.064	0.055, 0.059, 0.064
Number of Clusters	8, 50, 200	8, 50, 200	8, 50, 200
Between Cluster Variance		0.05, 0.15, 0.90	
Period Effect		0	
Variance of Period Effect		0	
Intracluster Correlation (ICC)			0.01, 0.04, 0.21
Inter-Period Correlation (ICC)			0.01, 0.04, 0.21

Appendix Table 4b: Scenario 2 Parameters (Stars indicate scenario-specific parameters)			
	T-BOSS	Simulation	Closed Form Equation
Type-I Error Rate	0.05	0.05	0.05
Type-II Error Rate	0.9	0.9	0.9
Effect Size*	0.018, 0.036, 0.044	1.1, 1.2, 1.25	0.018, 0.036, 0.044
	(risk difference)	(odds ratio)	(risk difference)
Baseline Outcome Rate*	0.25	0.25	0.25
Treatment Outcome Rate*	0.268, 0.286, 0.294	0.268, 0.286, 0.294	0.268, 0.286, 0.294
Number of Clusters	8, 50, 200	8, 50, 200	8, 50, 200
Between Cluster Variance		0.05, 0.15, 0.90	
Period Effect		0	
Variance of Period Effect		0	
Intracluster Correlation (ICC)			0.01, 0.04, 0.21
Inter-Period Correlation (ICC)			0.01, 0.04, 0.21

Appendix Table 5a: Extended Main Table with Simulation Results. Baseline Survival Rate = 5%						
		Fixing Sample Size = 10,400				
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Small Number of Clusters (n=8)						
Low Between-Cluster Variance	1.1	17.6	18.2	19.1	0.6	1.5
(0.05, ICC ≈ 0.01)	1.2	52.0	54.4	56.4	2.4	4.4
	1.3	83.7	86.7	87.4	3.0	3.7
Medium Between-Cluster Variance	1.1	17.6	19.8	19.6	2.2	2.0
(0.15, ICC ≈ 0.04)	1.2	52.0	58.0	57.6	6.0	5.6
	1.3	83.7	89.4	88.3	5.7	4.6
High Between-Cluster Variance	1.1	17.6	24.3	22.9	6.7	5.3
(0.90, ICC ≈ 0.21)	1.2	52.0	64.6	66.0	12.6	14.0
	1.3	83.7	90.9	93.5	7.2	9.8
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Medium Number of Clusters (n=50)						
Low Between-Cluster Variance	1.1	17.6	22.6	19.1	5.0	1.5
(0.05, ICC ≈ 0.01)	1.2	52.0	55.2	56.4	3.2	4.4
	1.3	83.7	89.4	87.4	5.7	3.7
Medium Between-Cluster Variance	1.1	17.6	19.3	19.6	1.7	2.0
(0.15, ICC ≈ 0.04)	1.2	52.0	60.3	57.6	8.3	5.6
	1.3	83.7	87.4	88.3	3.7	4.6
High Between-Cluster Variance	1.1	17.6	25.1	22.9	7.5	5.3
(0.90, ICC ≈ 0.21)	1.2	52.0	68.2	66.0	16.2	14.0
	1.3	83.7	92.3	93.5	8.6	9.8
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
High Number of Clusters (n=200)						
Low Between-Cluster Variance	1.1	17.6	19.5	19.1	1.9	1.5
(0.05, ICC ≈ 0.01)	1.2	52.0	55.2	56.4	3.2	4.4
	1.3	83.7	87.4	87.4	3.7	3.7
Medium Between-Cluster Variance	1.1	17.6	18.3	19.6	0.7	2.0
(0.15, ICC ≈ 0.04)	1.2	52.0	58.6	57.6	6.6	5.6
	1.3	83.7	87.4	88.3	3.7	4.6
High Between-Cluster Variance	1.1	17.6	24.3	22.9	6.7	5.3
(0.90, ICC ≈ 0.21)	1.2	52.0	69.6	66.0	17.6	14.0
	1.3	83.7	93.7	93.5	10.0	9.8

(continued on next page)

Appendix Table 5a: Extended Main Table with Simulation Results. Baseline Survival Rate = 5%						
		Fixing Power = 90%				
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Small Number of Clusters (n=8)						
Low Between-Cluster Variance	1.1	102,928	85,536	92,416	20.3	11.4
(0.05, ICC ≈ 0.01)	1.2	27,072	25,552	24,304	5.9	11.4
	1.3	12,624	11,552	11,344	9.3	11.3
Medium Between-Cluster Variance	1.1	102,928	97,120	89,616	6.0	14.9
(0.15, ICC ≈ 0.04)	1.2	27,072	25,552	23,568	5.9	14.9
	1.3	12,624	10,496	10,992	20.3	14.8
High Between-Cluster Variance	1.1	102,928	91,328	74,688	12.7	37.8
(0.90, ICC ≈ 0.21)	1.2	27,072	21,760	19,648	24.4	37.8
	1.3	12,624	10,144	9,168	24.4	37.7
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Medium Number of Clusters (n=50)						
Low Between-Cluster Variance	1.1	102,900	94,300	92,400	9.1	11.4
(0.05, ICC ≈ 0.01)	1.2	27,100	25,500	24,300	6.3	11.5
	1.3	12,600	10,200	11,300	23.5	11.5
Medium Between-Cluster Variance	1.1	102,900	85,600	89,600	20.2	14.8
(0.15, ICC ≈ 0.04)	1.2	27,100	23,200	23,600	16.8	14.8
	1.3	12,600	10,600	11,000	18.9	14.5
High Between-Cluster Variance	1.1	102,900	75,000	74,700	37.2	37.8
(0.90, ICC ≈ 0.21)	1.2	27,100	19,400	19,600	39.7	38.3
	1.3	12,600	7,800	9,200	61.5	37.0
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
High Number of Clusters (n=200)						
Low Between-Cluster Variance	1.1	102,800	96,800	92,400	6.2	11.3
(0.05, ICC ≈ 0.01)	1.2	27,200	24,000	24,400	13.3	11.5
	1.3	12,800	12,000	11,200	6.7	14.3
Medium Between-Cluster Variance	1.1	102,800	88,400	89,600	16.3	14.7
(0.15, ICC ≈ 0.04)	1.2	27,200	23,200	23,600	17.2	15.3
	1.3	12,800	10,400	10,800	23.1	18.5
High Between-Cluster Variance	1.1	102,800	70,800	74,800	45.2	37.4
(0.90, ICC ≈ 0.21)	1.2	27,200	24,000	19,600	13.3	38.8
	1.3	12,800	8,800	9,200	45.5	39.1

Appendix Table 5b: Extended Main Table with Simulation Results. Baseline Survival Rate = 25%						
		Fixing Sample Size = 5,200				
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Small Number of Clusters (n=8)						
Low Between-Cluster Variance	1.1	29.9	32.2	32.7	2.3	2.8
(0.05, ICC ≈ 0.01)	1.2	79.1	83.0	83.3	3.9	4.2
	1.25	92.6	95.1	94.9	2.5	2.3
Medium Between-Cluster Variance	1.1	29.9	31.2	33.6	1.3	3.7
(0.15, ICC ≈ 0.04)	1.2	79.1	84.6	84.4	5.5	5.3
	1.25	92.6	94.8	95.5	2.2	2.9
High Between-Cluster Variance	1.1	29.9	29.9	39.5	0.0	9.6
(0.90, ICC ≈ 0.21)	1.2	79.1	80.9	90.6	1.8	11.5
	1.25	92.6	93.0	98.1	0.4	5.5
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Medium Number of Clusters (n=50)						
Low Between-Cluster Variance	1.1	29.9	33.3	32.7	3.4	2.8
(0.05, ICC ≈ 0.01)	1.2	79.1	83.6	83.3	4.5	4.2
	1.25	92.6	95.5	94.9	2.9	2.3
Medium Between-Cluster Variance	1.1	29.9	32.9	33.6	3.0	3.7
(0.15, ICC ≈ 0.04)	1.2	79.1	80.7	84.4	1.6	5.3
	1.25	92.6	94.5	95.5	1.9	2.9
High Between-Cluster Variance	1.1	29.9	30.0	39.5	0.1	9.6
(0.90, ICC ≈ 0.21)	1.2	79.1	79.1	90.6	0.0	11.5
	1.25	92.6	93.4	98.1	0.8	5.5
		Power			T-BOSS Absolute Decrease of Power	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
High Number of Clusters (n=200)						
Low Between-Cluster Variance	1.1	29.9	32.6	32.7	2.7	2.8
(0.05, ICC ≈ 0.01)	1.2	79.1	82.1	83.3	3.0	4.2
	1.25	92.6	94.5	94.9	1.9	2.3
Medium Between-Cluster Variance	1.1	29.9	32.2	33.6	2.3	3.7
(0.15, ICC ≈ 0.04)	1.2	79.1	80.8	84.4	1.7	5.3
	1.25	92.6	95.5	95.5	2.9	2.9
High Between-Cluster Variance	1.1	29.9	32.2	39.5	2.3	9.6
(0.90, ICC ≈ 0.21)	1.2	79.1	80.3	90.6	1.2	11.5
	1.25	92.6	92.2	98.1	-0.4	5.5

(continued on next page)

Appendix Table 5b: Extended Main Table with Simulation Results. Baseline Survival Rate = 25%						
		Fixing Power = 90%				
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Small Number of Clusters (n=8)						
Low Between-Cluster Variance	1.1	26,576	25,088	23,872	5.9	11.3
(0.05, ICC ≈ 0.01)	1.2	7,120	5,904	6,384	20.6	11.5
	1.25	4,704	4,384	4,224	7.3	11.4
Medium Between-Cluster Variance	1.1	26,576	25,088	23,136	5.9	14.9
(0.15, ICC ≈ 0.04)	1.2	7,120	3,904	6,192	82.4	15.0
	1.25	4,704	4,320	4,096	8.9	14.8
High Between-Cluster Variance	1.1	26,576	14,624	19,040	81.7	39.6
(0.90, ICC ≈ 0.21)	1.2	7,120	7,552	5,104	-5.7	39.5
	1.25	4,704	5,152	3,376	-8.7	39.3
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
Medium Number of Clusters (n=50)						
Low Between-Cluster Variance	1.1	26,600	22,100	23,900	20.4	11.3
(0.05, ICC ≈ 0.01)	1.2	7,100	6,600	6,400	7.6	10.9
	1.25	4,700	4,400	4,200	6.8	11.9
Medium Between-Cluster Variance	1.1	26,500	25,000	23,100	6.0	14.7
(0.15, ICC ≈ 0.04)	1.2	7,100	6,600	6,200	7.6	14.5
	1.25	4,700	4,200	4,100	11.9	14.6
High Between-Cluster Variance	1.1	26,500	26,800	19,000	-1.1	39.5
(0.90, ICC ≈ 0.21)	1.2	7,100	7,000	5,100	1.4	39.2
	1.25	4,700	4,800	3,400	-2.1	38.2
		Sample Size			T-BOSS % Increase in Sample Size	
	Odds Ratio	**TBOSS**	**Simulation**	**Closed Form**	**Simulation**	**Closed Form**
High Number of Clusters (n=200)						
Low Between-Cluster Variance	1.1	26,400	24,800	24,000	6.5	10.0
(0.05, ICC ≈ 0.01)	1.2	7,200	6,400	6,400	12.5	12.5
	1.25	4,800	4,000*	4,400	20.0*	9.1
Medium Between-Cluster Variance	1.1	26,400	25,200	23,200	4.8	13.8
(0.15, ICC ≈ 0.04)	1.2	7,200	6,400	6,000	12.5	20.0
	1.25	4,800	4,000*	4,000	20.0*	20.0
High Between-Cluster Variance	1.1	26,400	14,400	19,200	83.3	37.5
(0.90, ICC ≈ 0.21)	1.2	7,200	7,200	5,200	0.0	38.5
	1.25	4,800	5,200	3,200	-7.7	50.0
*Simulations could not converge to 90% power for these settings. These sample sizes have an estimated 88% power.						

Figure 4a:

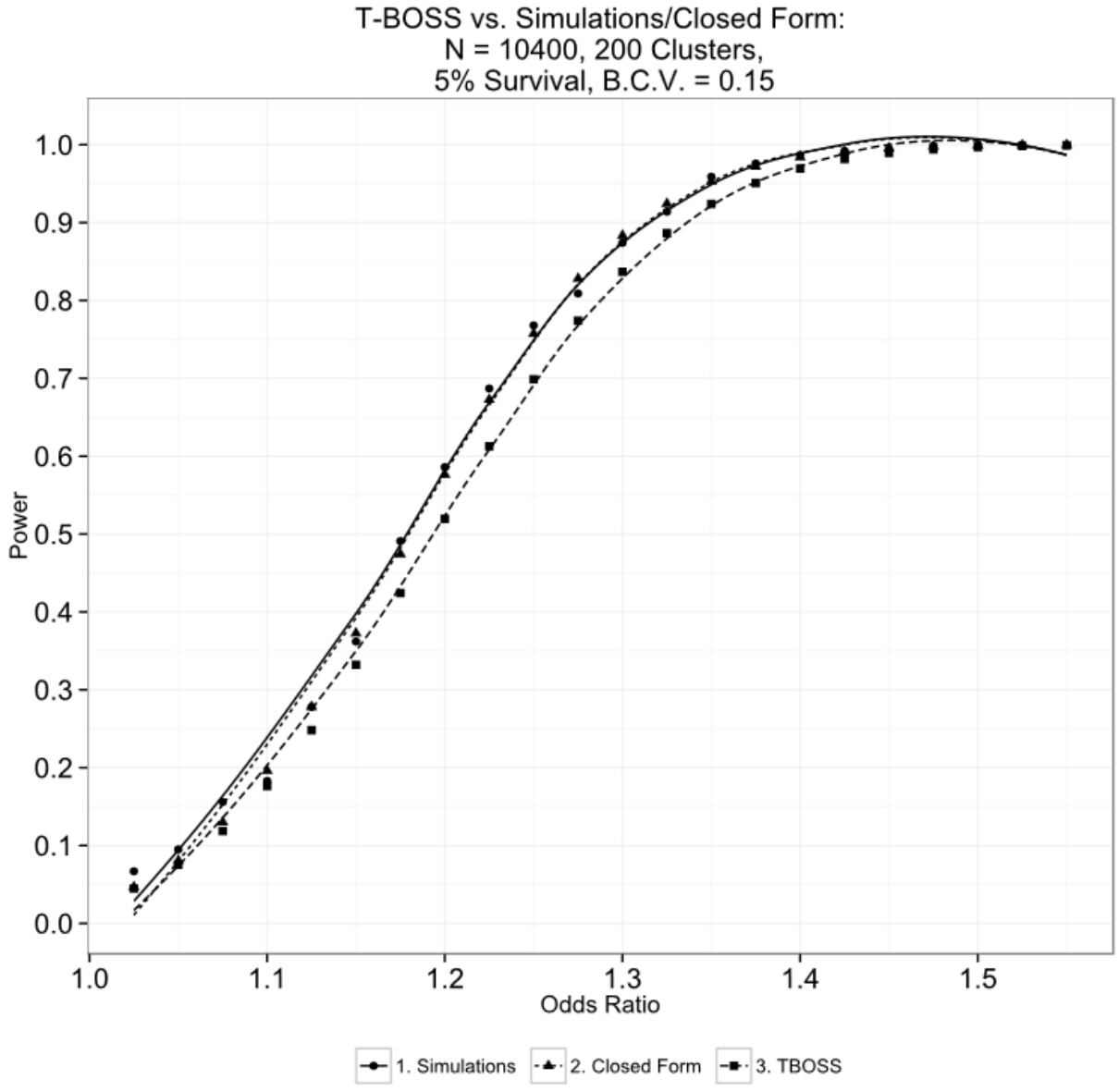

T-BOSS vs. Simulations/Closed Form:
N = 10400, 200 Clusters,
5% Survival, B.C.V. = 0.15
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
1.1
1.2
1.3
1.4
1.5
Odds Ratio
1. Simulations
2. Closed Form
3. TBOSS

Figure 4b:

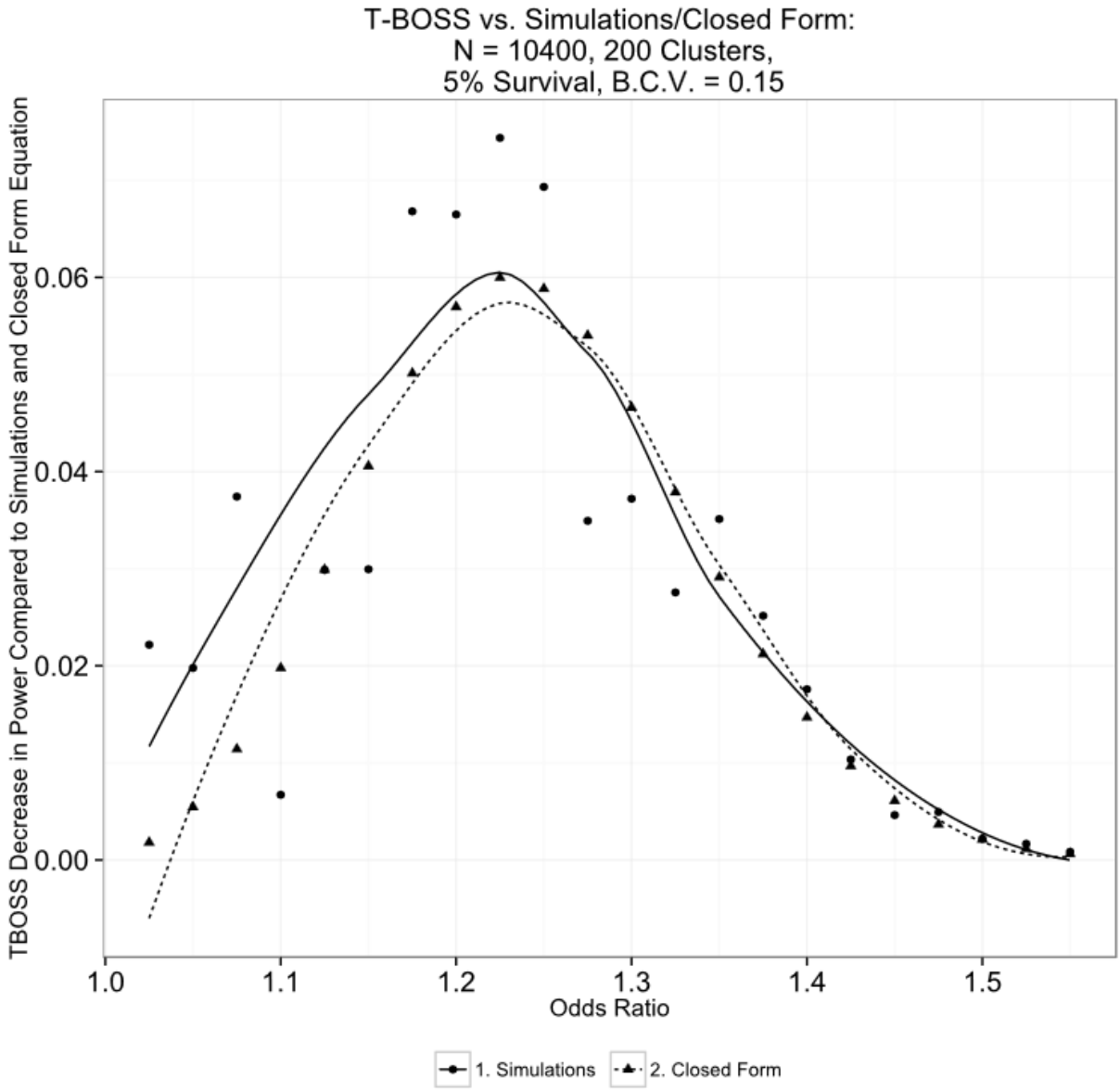

T-BOSS vs. Simulations/Closed Form:
N = 10400, 200 Clusters,
5% Survival, B.C.V. = 0.15
TBOSS Decrease in Power Compared to Simulations and Closed Form Equation
0.00
0.02
0.04
0.06
1.0
1.1
1.2
1.3
1.4
1.5
Odds Ratio
1. Simulations
2. Closed Form

Figure 5:

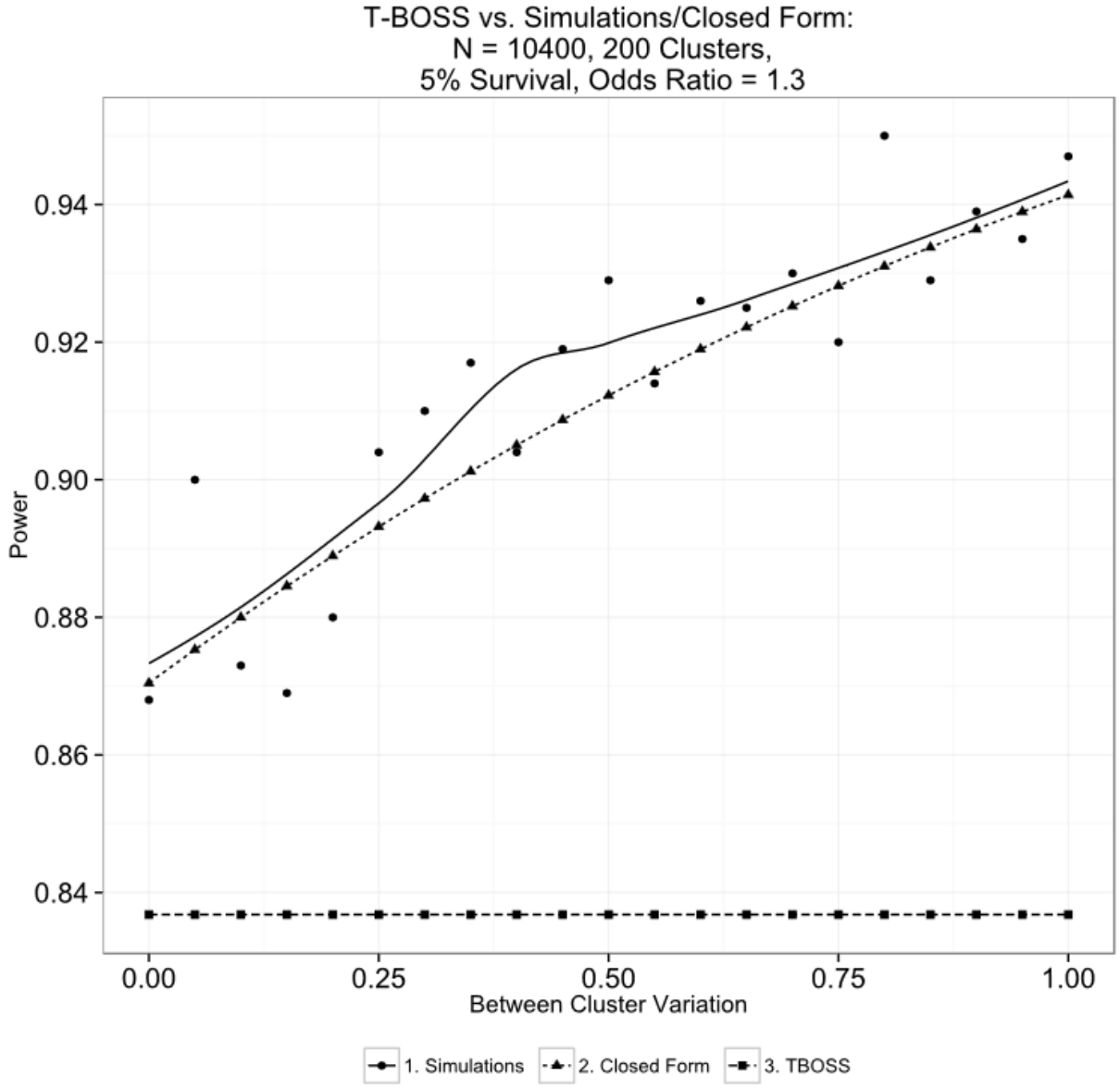

T-BOSS vs. Simulations/Closed Form:
N = 10400, 200 Clusters,
5% Survival, Odds Ratio = 1.3
Power
0.94
0.92
0.90
0.88
0.86
0.84
0.00
0.25
0.50
0.75
1.00
Between Cluster Variation
1. Simulations
2. Closed Form
3. TBOSS

Figure 6:

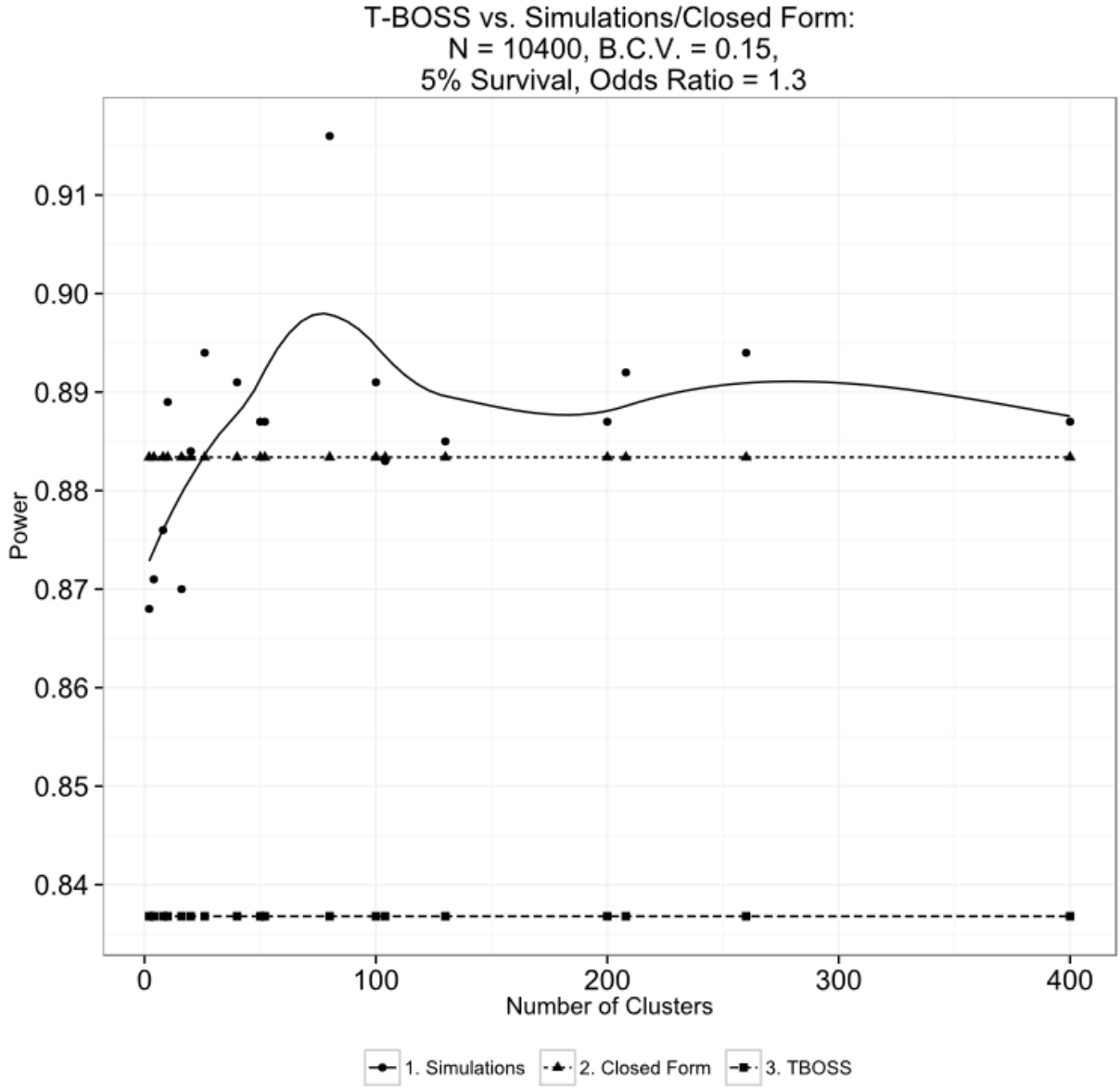

T-BOSS vs. Simulations/Closed Form:
N = 10400, B.C.V. = 0.15,
5% Survival, Odds Ratio = 1.3
Power
0.91
0.90
0.89
0.88
0.87
0.86
0.85
0.84
0
100
200
300
400
Number of Clusters
1. Simulations
2. Closed Form
3. TBOSS

Figure 7a:

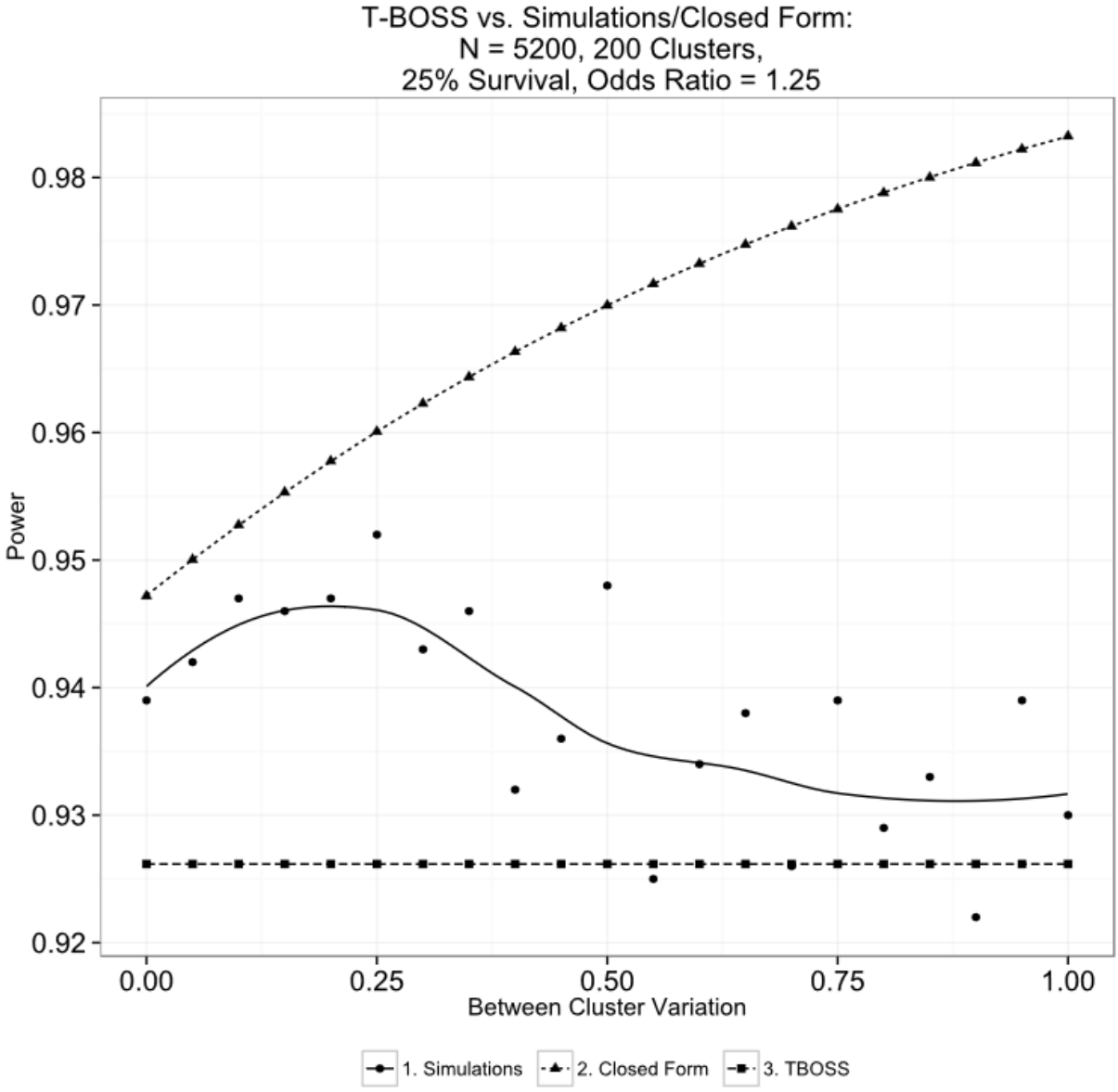
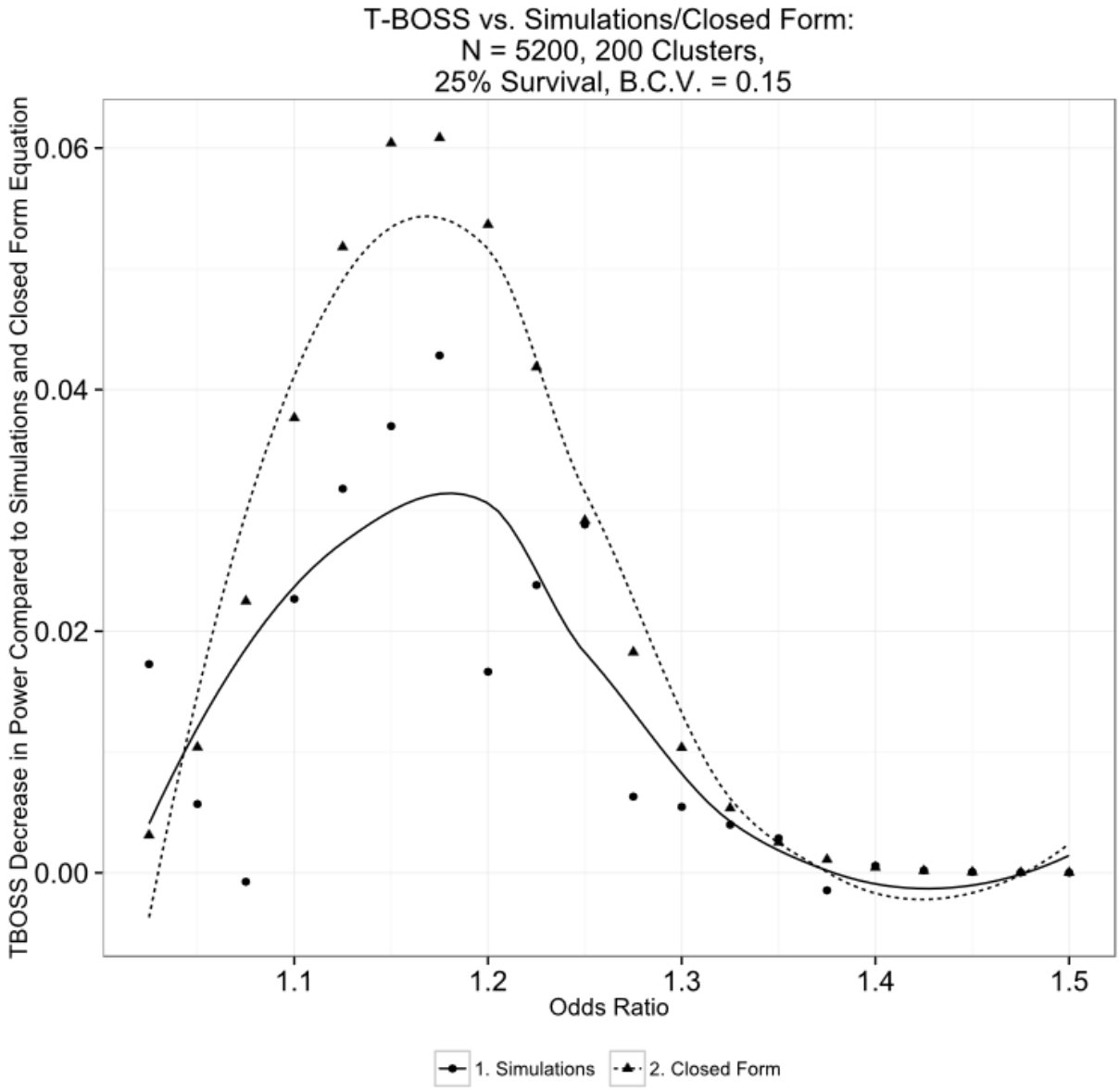

T-BOSS vs. Simulations/Closed Form:
N = 5200, 200 Clusters,
25% Survival, Odds Ratio = 1.25
Power
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.00
0.25
0.50
0.75
1.00
Between Cluster Variation
1. Simulations
2. Closed Form
3. TBOSS

Figure 7b:


T-BOSS vs. Simulations/Closed Form:
N = 5200, 200 Clusters,
25% Survival, B.C.V. = 0.15
TBOSS Decrease in Power Compared to Simulations and Closed Form Equation
0.06
0.04
0.02
0.00
1.1
1.2
1.3
1.4
1.5
Odds Ratio
1. Simulations
2. Closed Form

Simulation R Code

```
sample.size.function = function (n.clusters,
                                 n.sims = 1000,
                                 n.periods = 2,
                                 alpha = 0.05,
                                 odds.ratio,
                                 null.mean,
                                 btw.clust.var)
{

  mylogit = function(x){log(x/(1-x))}
  myexpit = function(x){exp(x)/(1+exp(x))}



  set.seed(42)



  ######################################## Test parameters
  effect.size = log(odds.ratio)
  alt.mean  = myexpit(effect.size + mylogit(null.mean))

  if(alt.mean > 1) stop("alt.mean > 1")



  power.sim <- function (sample.size) {

    ######################################## Specify Parameters, Error Check
```

```
    n = sample.size
    n.incluster = n/n.clusters

    if(round(n.clusters/2) != n.clusters/2) stop("number of clusters must be even")
    if(round(n.incluster) != n.incluster) stop("n.incluster not an integer")
    if(round(n/n.clusters/n.periods) != n/n.clusters/n.periods) stop("data cannot be broken into equal cluster
sizes. change n, n.clusters, or n.periods to fix.")

    ##################################### Start Simulation

    significant = rep(NA, n.sims)
    error = rep(NA, n.sims)

    for (i in 1:n.sims) {

      x <- matrix(NA, nrow=n, ncol=5)
      colnames(x) = c("baseline", "treatment", "clustereffect", "clusterid", "period")


      x[,1] = mylogit(null.mean) #baseline

      x[,2] = rep(c(rep(rep(c(1,0),n.periods/2), each=n.incluster/n.periods),
                    rep(rep(c(0,1),n.periods/2), each=n.incluster/n.periods)), n.clusters/2) #trt

      x[,3] = rep(rnorm(n.clusters, 0, sd=sqrt(btw.clust.var)), each=n.incluster) #cluster effect


      x[,4] = rep(c(1:n.clusters), each=n.incluster) #clusterid
```

```
    x[,5] = rep(c(1:n.periods), each=n.incluster/n.periods, times=n.clusters) #period

    design.mat <- x[,1:3]
    full.beta <- c(1, effect.size, 1)
    mean.y <- design.mat %*% full.beta
    y = rbinom(n, size = 1, prob = myexpit(mean.y))
    final.data = as.data.frame(cbind(y, x))

    test <- glmer(y ~ treatment + (1|clusterid) + (1|period), data=final.data, family=binomial,
control=glmerControl(optimizer="bobyqa"))

    Est <- fixef(test)[2]
    Ste <- sqrt(vcov(test)[2,2])
    significant[i] = prod(Est + c(-1,1) * qnorm(1-0.05/2) * Ste) > 0
    error[i] = sum(is.na(final.data$y)) > 0

  }

  power = sum(significant)/n.sims

  power

}
```

```
 ######################################## T-BOSS Method
tb.effect.size <- alt.mean - null.mean
tb.variance   <- 1.05^2*(null.mean*(1-null.mean)+ alt.mean*(1-alt.mean))/2

TBOSS <- power.t.test(delta=tb.effect.size, sd=sqrt(tb.variance), power=0.9)

tboss.sample.size <- round(TBOSS$n*2/n.clusters/n.periods)*n.clusters*n.periods




######################################## Binary Search for Gold Standard Sample Size

# starting values for the simulation
power=0.79
sample.size <- floor(TBOSS$n*2/n.clusters/n.periods)*n.clusters*n.periods
high.guess <- sample.size*1.5
low.guess <- sample.size*0.1
iteration = 0

while (abs(power-0.9) > 0.005 & iteration < 50) {

  # simulation function - fixed cluster size
  power <- power.sim(sample.size = sample.size)

  if(power < 0.9) {low.guess = sample.size}
  else if(power > 0.9) {high.guess = sample.size}

  new.sample.size = (high.guess + low.guess)/2
```

```
    sample.size = round(new.sample.size/n.clusters/n.periods)*n.clusters*n.periods


    iteration = iteration + 1
  }


  results = c(tboss.sample.size, sample.size, power, odds.ratio, n.clusters, n.periods, null.mean,
btw.clust.var, iteration)
  results = as.data.frame(results)
  rownames(results) = c("tboss.sample.size", "sample.size", "power", "odds.ratio",
                        "n.clusters", "n.periods", "null.mean", "btw.clust.var", "n.iterations")
  results
}
```

References

1. Turner, R., White, I., & Croudace, T. (2007). Analysis of cluster randomized cross-over trial data: A comparison of methods. *Statistics in Medicine, 26*, 274-289.

2. University of Washington. Continuous Chest Compressions vs AHA Standard CPR of 30:2 (CCC). In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000. [cited 2014 Aug 08]. Available from http://clinicaltrials.gov/show/NCT01372748 NLM Identifier: NCT01372748.

3. Giraudeau, B., Ravaud, P., & Donner, A. (2008). Sample size calculation for cluster randomized cross-over trials. *Statistics in Medicine, 27*, 5578-5585.

4. Connolly, S., Philippon, F., Longtin, Y., Casanova, A., Birnie, D., Exner, D., ... Krahn, A. (2013). Randomized Cluster Crossover Trials for Reliable, Efficient, Comparative Effectiveness Testing: Design of the Prevention of Arrhythmia Device Infection Trial (PADIT). *Canadian Journal of Cardiology, 29*, 652-658.

5. Reich, N., Myers, J., Obeng, D., Milstone, A., & Perl, T. (2012). Empirical Power and Sample Size Calculations for Cluster-Randomized and Cluster-Randomized Crossover Studies. *PLoS ONE, 7*(4), E35564.

6. Illowsky, B. & Dean, S. Introductory Statistics. OpenStax College, 2013. OpenStax College CNX. Web. 3 Aug. 2014. http://openstaxcollege.org/textbooks/introductory-statistics

7. Hayes, R., & Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology, 28*, 319-326.

8. Stiell IG, Nichol G, Leroux BG, et al. (2011) Early versus later rhythm analysis in patients with out of hospital cardiac arrest. *New England Journal of Medicine*, *365*(9), 787–797.

9. Nichol, G., Thomas, E., Callaway, C., Hedges, J., Powell, J., Aufderheide, T., ... Stiell, I. (2008). Regional Variation in Out-of-Hospital Cardiac Arrest Incidence and Outcome. *JAMA: The Journal of the American Medical Association, 300*, 1423-1431.

10. Rietbergen, C., & Moerbeek, M. (2011). The Design of Cluster Randomized Crossover Trials. *Journal of Educational and Behavioral Statistics, 36*(4), 472-490.

11. Davis, D., Garberson, L., Andrusiek, D., et. al. (2007). A Descriptive Analysis of Emergency Medical Service Systems Participating in the Resuscitation Outcomes Consortium (ROC) Network. *Prehospital Emergency Care, 11*, 369-382

12. Marks, G., Ezz, W., Aust, N., Toelle, B., Xuan, W., Belousova, E., ... Smith, W. (2010). Respiratory Health Effects of Exposure to Low-NOx Unflued Gas Heaters in the Classroom: A Double-Blind, Cluster-Randomized, Crossover Study. *Environmental Health Perspectives, 118*(10), 1476-1482.

13. Zeger, S., & Liang, K. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics, 42*, 121-130.

14. McCulloch, C. (2003), Generalized Linear Mixed Models. *NSF-CBMS Regional Conference Series in Probability and Statistics, 7,* i-v+vii-viii+1-84.

15. Eldridge, S., Ukoumunne, O., & Carlin, J. (2009). The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review, 77*, 378-394.

16. Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2009). *Introduction to Algorithms* (Revised/Expanded ed.). Cambridge, Mass.: MIT Press.

17. Schmicker, R., Leroux, B., Sears, G., Stiell, I., Morrison, L., Aufderheide, T., ... Cheskes, S. (2012). Temporal compliance trends in a cluster randomization with crossover trial of out-of-hospital cardiac arrest. *Clinical Trials, 9*, 314-321.