6-21-2017

Constructing a Confidence Interval for the Fraction Who Benefit from Treatment, Using Randomized Trial Data

Emily J. Huang
Johns Hopkins University School of Public Health, Department of Biostatistics, emhuang1@gmail.com

Ethan X. Fang
Penn State University, Department of Statistics

Daniel F. Hanley
Johns Hopkins Medical Institutions, Division of Brain Injury Outcomes

Michael Rosenblum
Johns Hopkins University School of Public Health

Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data

Emily J. Huang, Ethan X. Fang, Daniel F. Hanley, Michael Rosenblum

June 4, 2017

Abstract

The fraction who benefit from treatment is defined as the proportion of patients whose potential outcome under treatment is better than that under control. Statistical inference for this parameter is challenging since it is only partially identifiable, even in our context of a randomized trial. We propose and evaluate a new method for constructing a confidence interval for the fraction who benefit, when the outcome is ordinal-valued (with binary outcomes as a special case). This confidence interval procedure is proved to be pointwise consistent. Our method does not require any assumptions about the joint distribution of the potential outcomes, although it has the flexibility to incorporate a wide range of user-defined assumptions. A potential advantage of our approach is that, unlike existing confidence interval methods for partially identified parameters (such as m-out-of-n bootstrap and subsampling), we do not need to select m or the subsample size, which is generally a challenging problem. Our method is based on a stochastic optimization technique involving a second order, asymptotic approximation that, to the best of our knowledge, has not been applied to biomedical studies. This approximation leads to statistics that are solutions to quadratic programs, and so they can be computed efficiently using existing optimization tools. In all of our simulations, our method attains the nominal coverage probability or higher, and can have substantially narrower average width compared to the m-out-of-n bootstrap. We also apply our method to a completed trial data set of a new surgical intervention for severe stroke.

1 Introduction

The fraction who benefit from treatment is the proportion of patients whose potential outcome under treatment is better than that under control. In other words, it is the proportion who would be better off with treatment. This fraction may be of interest to patients and care providers deciding between treatment and control. It may also be informative to medical researchers; for example, a small fraction indicates that an exclusive subgroup benefits and resources should be devoted towards identifying it. We aim to draw inferences about the fraction who benefit, using a randomized trial.

In general, the fraction who benefit (sometimes abbreviated as "the fraction") is non-identifiable from observed data, even in the randomized trial context. This occurs because only one potential outcome can be observed per patient. Typically, identifiability of the fraction necessitates strong, untestable assumptions on the joint distribution of the potential outcomes, such as independence of the potential outcomes within a person. We do not require any assumptions on the joint distribution and only consider assumptions based on subject matter knowledge. Since the fraction is generally non-identifiable in this setting, constructing a confidence interval is a challenging problem.

An existing confidence interval procedure for our problem involves applying the m-out-of-n bootstrap to estimators of lower and upper bounds (which are identifiable) on the fraction who benefit. The m-out-of-n bootstrap is a generalization of the standard nonparametric bootstrap, where bootstrap replicate data sets are generated by resampling m patients with replacement for $m \leq n$. The m-out-of-n bootstrap is recommended because the bound estimators for our problem can be non-regular (Huang *and others*, 2017), and the standard bootstrap can be inconsistent in such cases. Another existing method for constructing confidence intervals is the subsampling approach of Romano and Shaikh (2008). Subsampling is similar to the m-out-of-n bootstrap, except resampling is done without replacement. A challenge in using m-out-of-n bootstrap or subsampling is how to select m to achieve good performance. We propose a new confidence interval method that avoids having to select m.

Through simulation, we compare our method to the m-out-of-n bootstrap with respect to coverage probability and average width. In all cases, our method has coverage probability at or above the nominal level, while the m-out-of-n bootstrap sometimes has coverage probability below the nominal level. In some cases, our method achieves substantially narrower average width than the m-out-of-n bootstrap, e.g., reduction in average width of 40%. Our method has good coverage probability even in cases where the lower and upper bound parameters are non-differentiable functions of the marginal distributions under treatment and control, as shown in Section 5.

We apply our method to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial of a new surgical treatment for stroke, which had a sample size of 500 patients (Hanley *and others*, 2017). Outcomes included disability measured by the modified Rankin Scale and death. As examples of the output of our confidence interval procedure, the 95% confidence interval for the fraction who benefit is [0.01,0.18] for the outcome 30-day mortality, [0.05,0.34] for 180-day mortality, [0,0.64] for 30-day disability, and [0.03,0.86] for 180-day disability.

Our confidence interval procedure is based on representing the problem as a stochastic optimization problem. Stochastic optimization involves maximizing or minimizing the expected value of a function of unknown parameters and random variables, based on repeated observations of the random variables (data). As a simple example, M-estimators can be represented in terms of solving stochastic optimization problems (van der Vaart, 1998, Chapter 5). Our problem is substantially harder, since its formulation as a stochastic optimization problem involves a set of additional constraints on the parameter space (specifically, that the parameter lies within a polyhedron). When the optimal solution converges to a point on the boundary of the parameter space, the resulting statistics are generally not asymptotically normal; this rules out standard confidence interval procedures, many of which require asymptotic normality.

Shapiro *and others* (2014) present general approaches for deriving the asymptotic distributions of such challenging stochastic optimization problems. To the best of our knowledge, these general approaches have not previously been used to solve problems arising in biomedical studies. We tailor one such approach to solve our problem, using a second order, asymptotic approximation of the objective function. We provide a self-contained proof of the validity of our method, which can be understood without requiring knowledge of stochastic optimization.

The statistic derived using the above approach can be computed using quadratic programming, i.e., minimizing a quadratic function of the data and parameters subject to linear equality and inequality constraints on the parameters. We used the "quadprog" solver in MATLAB 2013B. The computing time of our method (after preprocessing) is independent of the sample size, but dependent on the width of the confidence interval and the number of levels for the ordinal outcome (larger width and fewer levels take less time). Each confidence interval in the CLEAR III application was computed between 4 and 8 minutes. This running time can be further reduced through parallelization.

Section 2 provides an overview of the previous work on the fraction. In Section 3, we describe the data generating distribution and state assumptions that are used throughout the paper. Our new method is presented in Section 4, including proofs of its asymptotic properties. We evaluate the method through simulation in Section 5. It is applied to the CLEAR III randomized trial in Section 6. Future work is discussed in Section 7.

2 Previous Work

In general, the fraction who benefit is non-identifiable without making untestable assumptions about the joint distribution of the potential outcomes. There has been work on deriving and estimating bounds on this parameter in our context of ordinal outcomes (Borusyak, 2015; Lu *and others*, 2016; Huang *and others*, 2017).

To construct a confidence interval for the fraction who benefit, one could use the bound estimators proposed in Huang *and others* (2017) and apply the m-out-of-n bootstrap to them. Romano and Shaikh (2008) propose a general confidence interval method based on subsampling for partially identified parameters, such as the fraction who benefit. Under the subsampling condition (i) in Theorem 3.4 of their paper, Romano and Shaikh (2008) prove pointwise consistency of their confidence interval method. However, it is difficult to establish whether this condition holds in our problem.

Other parameters that contrast the distribution of an ordinal outcome under treatment versus control include the number needed to treat and the parameter in a responder analysis (Snapinn and Jiang, 2007). However, these parameters require that the ordinal outcome be dichotomized into "success" or "failure". For example, in one analysis of the CLEAR III trial, the modified Rankin Scale outcome was considered a "success" if it was in the range 0-3 (Hanley *and others*, 2017). The parameter of interest in a responder analysis is the difference between the population proportions who have a successful outcome under treatment versus control, where success can be a function of baseline variables. The number needed to treat is the reciprocal of this difference (Gordis, 2009). A downside to dichotomization of the outcome is that improvements not crossing the dichotomization threshold are ignored. The fraction who benefit considers the full ordinal scale.

3 Notation, Parameter Definition, and Assumptions

3.1 Parameter Definition

Consider an ordinal outcome with a finite number of levels, L. Without loss of generality, we assume that the levels are numbered as integers from 1 to L, in order of least to most favorable. Denote Y_T as the potential outcome under treatment and Y_C as the potential outcome under control. Let P_0 denote the true, unknown joint distribution on (Y_C, Y_T). Let $\pi_{i,j}$ denote the probability that $Y_C = i$ and $Y_T = j$, i.e., $\pi_{i,j} = P_0(Y_C = i, Y_T = j)$. We say that a patient benefits from treatment compared to control if her/his potential outcome pair (y_C, y_T) satisfies $y_T > y_C$. She/he is harmed if $y_T < y_C$ and experiences no individual treatment effect if $y_T = y_C$. The fraction who benefit from treatment, our parameter of interest, is:

$$\psi_0 = P_0(Y_T > Y_C) = \sum_{j>i} \pi_{i,j}. \tag{1}$$

We propose a method to construct a confidence interval for the parameter ψ_0, which does not require assumptions about the joint distribution P_0. The method can incorporate restrictions on the support of P_0, supplied by the user based on subject matter knowledge. Support restrictions are assumptions that certain potential outcome pairs (i, j) are not possible, i.e., $\pi_{i,j} = 0$. The no harm assumption ($\pi_{i,j} = 0$ if $i > j$) is one example. For conciseness, we refer to support restrictions as restrictions. The user specifies restrictions through a function $g : \mathcal{L} \times \mathcal{L} \to \{0, 1\}$, where $\mathcal{L}$ is the set of integers from 1 to L. For any given input (i, j), the user sets $g(i, j)$ to 0 if she/he assumes that $\pi_{i,j} = 0$, and 1 otherwise. If no restrictions are made, the function g outputs 1 for all inputs. Let $\mathcal{R}$ be the set of all joint distributions P on (Y_C, Y_T) that satisfy the restrictions:

$$\mathcal{R} = \{P \text{ on } (Y_C, Y_T) : P(Y_C = i, Y_T = j) = 0 \text{ if } g(i, j) = 0\}. \tag{2}$$

Assumption 1. *The user-defined support restrictions are correct, i.e.,* $P_0 \in \mathcal{R}$.

Incorrect assumptions can lead to poor coverage probability of our method and the m-out-of-n bootstrap, as shown in Section 5.

3.2 Observed Data Distribution

We construct our confidence interval using data from a randomized trial. Let n be the number of participants in the trial. For each participant m, let A_m and Y_m denote the participant's treatment assignment (1 if treatment and 0 if control) and observed outcome, respectively. We assume that the vectors (A_m, Y_m), $i = m, .., n$, are fully observed. Other assumptions include the following:

Assumption 2. *For each participant* m*, her/his potential outcome pair* $(Y_{C,m}, Y_{T,m})$ *is an independent, identically distributed draw from the unknown joint distribution* P_0.

Assumption 3. *The treatment assignments,* A_m*,* $m = 1, .., n$*, are independent, identically distributed Bernoulli*(θ)*, where* $0 < \theta < 1$*. The treatment assignments* $\{A_m\}_{m=1}^n$ *are independent of the potential outcome pairs* $\{(Y_{C,m}, Y_{T,m})\}_{m=1}^n$.

Assumption 4. *For each participant* m*, we have* $Y_m = A_m Y_{T,m} + (1 - A_m) Y_{C,m}$.

Assumption 3 is satisfied by a simple randomized trial design (Friedman *and others*, 2010). The value θ is the probability of being assigned to treatment, which is known and should not be 0 or 1. Assumption 4 connects observed outcomes to potential outcomes and is called the consistency assumption.

3.3 Non-identifiability of the Fraction who Benefit

The assumptions above imply that the vectors (A_m, Y_m), $m = 1, .., n$, are independent and identically distributed. Let (A, Y) denote the random vector corresponding to a generic participant in the randomized trial. The vector (A, Y) for each participant is called the observed data, to distinguish it from the vector of potential outcomes (Y_C, Y_T) which is partially unobserved. Let P_{obs} denote the population distribution on the observed data vector (A, Y). By Assumption 3, we have $P_{obs}(A = a) = \theta^a(1-\theta)^{1-a}$. Let the vector $\boldsymbol{\gamma}^* = (\gamma^*_{01}, .., \gamma^*_{0L}, \gamma^*_{11}, .., \gamma^*_{1L})$ denote the marginal distributions of the potential outcomes under treatment and under control, where $\gamma^*_{0y} = P_0(Y_C = y)$ and $\gamma^*_{1y} = P_0(Y_T = y)$ for all y in $\mathcal{L}$. By Assumptions 3 and 4, we have that for all $y \in \mathcal{L}$:

$$\begin{aligned} \gamma^*_{0y} = P_0(Y_C = y) &= P_{obs}(Y = y|A = 0), \\ \gamma^*_{1y} = P_0(Y_T = y) &= P_{obs}(Y = y|A = 1). \end{aligned} \tag{3}$$

This implies that the marginal distributions of the potential outcomes are identifiable.

Because only one potential outcome is observed per participant, the fraction who benefit ψ_0 is typically non-identifiable from observed data. However, the marginal distributions $\boldsymbol{\gamma}^*$ and restrictions $\mathcal{R}$ may rule out certain possibilities. Let $\psi^{\mathcal{R}}_l(P_{obs})$ and $\psi^{\mathcal{R}}_u(P_{obs})$ denote the sharp lower and upper bounds on the fraction, given the marginal distributions and restrictions, i.e.,

$$\begin{aligned} &\psi^{\mathcal{R}}_l(P_{obs}) = \min\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\}, \\ &\psi^{\mathcal{R}}_u(P_{obs}) = \max\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\}. \end{aligned}$$

These bounds are functions of P_{obs} due to their dependency on $\boldsymbol{\gamma}^*$, and are identifiable because $\boldsymbol{\gamma}^*$ is identifiable. For conciseness, we suppress the dependency on P_{obs}. The bounds are discussed in Huang *and others* (2017). The fraction who benefit ψ_0 must be between the bounds, i.e., $\psi_0 \in [\psi^{\mathcal{R}}_l, \psi^{\mathcal{R}}_u]$. Moreover, for any $\psi \in [\psi^{\mathcal{R}}_l, \psi^{\mathcal{R}}_u]$, there exists some joint distribution $P \in \mathcal{R}$ that has marginals $\boldsymbol{\gamma}^*$ and with fraction who benefit $P(Y_T > Y_C)$ equal to ψ. This is proved in the Supplementary Materials. Intuitively, the marginal distributions and restrictions rule out candidates outside of the range $[\psi^{\mathcal{R}}_l, \psi^{\mathcal{R}}_u]$ but candidates inside the range are not ruled out.

We use the following definition for pointwise consistency from Romano and Shaikh (2008) tailored to our problem.

Definition 1. *A confidence set CS_n for ψ_0 is pointwise consistent at level $1-\alpha$ if, for any data generating distribution P_{obs} on (A, Y), we have for all $\psi \in \left[\psi^{\mathcal{R}}_l(P_{obs}), \psi^{\mathcal{R}}_u(P_{obs})\right]$:*

$$\liminf_{n\to\infty} P_{obs}(\psi \in CS_n) \geq 1 - \alpha. \tag{4}$$

Pointwise consistency is that, if one were to consider an arbitrary data generating distribution P_{obs} on (A, Y), then for all $\psi \in [\psi^{\mathcal{R}}_l(P_{obs}), \psi^{\mathcal{R}}_u(P_{obs})]$, the confidence set CS_n includes ψ with at least $1-\alpha$ probability when n is large. This is a desired property because the fraction who benefit ψ_0 must be somewhere in the range $[\psi^{\mathcal{R}}_l(P_{obs}), \psi^{\mathcal{R}}_u(P_{obs})]$ and the observed data distribution provides no information on where it lies within that range.

4 Proposed Method

We construct a 95% confidence set for the fraction who benefit ψ_0 through hypothesis test inversion. We consider candidate values of ψ on a grid on $[0, 1]$. In our simulations and data application (Sections 5 and 6), the grid that is used has a point at every hundredth, i.e., $\psi = 0, 0.01, 0.02, ..., 1$. A candidate value of ψ is excluded from the confidence set if and only if the hypothesis test for ψ rejects. If the confidence set is not an interval, we form a confidence interval using the smallest and largest points of the set. We present our hypothesis test in Section 4.1 and provide its implementation in Section 4.2. The MATLAB code is also provided with this paper. The asymptotic properties of the resulting confidence interval are presented in Section 4.3.

4.1 Hypothesis Test for Candidate Value of ψ

Let Π denote the set of all L by L matrices with nonnegative, real-valued entries that sum to 1. Define the set $\Gamma \in \mathbb{R}^{2L}$ as

$$\Gamma = \left\{ \boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t : \begin{array}{l} \text{For some } \boldsymbol{\pi} \in \Pi, \text{ we have} \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \\ \gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L} \\ \gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L} \end{array} \right\}. \tag{5}$$

This set is comprised of the pairs of marginal distributions (under treatment and under control) that are compatible with the restrictions. For example, if there are no restrictions, then Γ is the set of all vectors with nonnegative entries such that the sum of the first L entries equals 1 and the sum of the last L entries equals 1. If the no harm assumption is made and $L = 2$, the set Γ comprises all vectors with nonnegative entries that satisfy $\gamma_{12} \geq \gamma_{02}$ and the constraint in the previous sentence. Under Assumption 1, the pair of true marginal distributions $\boldsymbol{\gamma}^*$ is in the set Γ.

Consider any candidate value of $\psi \in [0,1]$. Define the set Γ^ψ as (5), but adding the constraint that $\sum_{j>i} \pi_{i,j} = \psi$ on the right hand side. The set Γ^ψ is comprised of the pairs of marginal distributions (under treatment and under control) that are compatible with both the restrictions $\mathcal{R}$ and the fraction who benefit being equal to ψ. Note that Γ and Γ^ψ are sets of vectors and not random. Each of these sets is a bounded, closed, convex polyhedron.

The null and alternative hypotheses for the candidate value of ψ are

$$H_0(\psi) \quad : \quad \boldsymbol{\gamma}^* \in \Gamma^\psi \tag{6}$$

$$H_a(\psi) \quad : \quad \boldsymbol{\gamma}^* \notin \Gamma^\psi. \tag{7}$$

The null hypothesis means that the pair of marginals $\boldsymbol{\gamma}^*$ (which is a function of P_{obs}) is compatible with both the restrictions $\mathcal{R}$ and fraction who benefit being equal to ψ. That is, there exists a joint distribution P on (Y_C, Y_T) such that its marginals equal $\boldsymbol{\gamma}^*$, it satisfies the restrictions, and the fraction who benefit $P(Y_T > Y_C)$ equals ψ. The null hypothesis is equivalent to $\psi \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$, while the alternative hypothesis is equivalent to $\psi \notin [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$. Intuitively, the null hypothesis means that the candidate value of ψ is not ruled out by the marginals $\boldsymbol{\gamma}^*$ and the restrictions $\mathcal{R}$.

Let $\boldsymbol{V} = (A, Y)$. We use the notation $P_{obs}X$ to denote the expectation of X with respect to P_{obs}, the distribution on $\boldsymbol{V}$ induced by P_0 and Assumptions 3 and 4, as discussed in Section 3.3. Define

$$F(\boldsymbol{\gamma}, \boldsymbol{V}) = \sum_{a=0}^{1} \sum_{j=1}^{L} 1(A = a) \left\{ 1(Y = j) - \gamma_{aj} \right\}^2.$$

The minimizer of $P_{obs}F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to γ^* defined in (3), as proved in the Supplementary Materials.

Define the test statistic as

$$T_{n,\psi} = n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - \inf_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) \right\}, \tag{8}$$

where P_n denotes the empirical distribution (that is, $P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) = \frac{1}{n} \sum_{m=1}^{n} F(\boldsymbol{\gamma}, \boldsymbol{V}_m)$).

Let $\widehat{\boldsymbol{\gamma}}$ be the vector $(\widehat{\gamma}_{01}, \ldots, \widehat{\gamma}_{0L}, \widehat{\gamma}_{11}, \ldots, \widehat{\gamma}_{1L})$, where $\widehat{\gamma_{0i}} = P_n(A = 0, Y = i)/P_n(A = 0)$ and $\widehat{\gamma_{1j}} = P_n(A = 1, Y = j)/P_n(A = 1)$ for $i, j \in \mathcal{L}$. The vector $\widehat{\boldsymbol{\gamma}}$ represents the empirical marginal distributions of the potential outcomes under control and under treatment. Let Discrep be the following function of $\widehat{\boldsymbol{\gamma}}$ and a generic vector $\boldsymbol{\gamma}$ of length $2L$:

$$\text{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1} \sum_{j=1}^{L} \left[P_n 1(A = a) \right] (\gamma_{aj} - \widehat{\gamma}_{aj})^2.$$

The indicator function notation $1(S)$ equals 1 if S is true and 0 otherwise.

Lemma 1.

$$T_{n,\psi} = n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \textit{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma} \in \Gamma} \textit{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \right\}. \tag{9}$$

Lemma 1 is useful for interpreting the test statistic. The function $\text{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}})$ is a weighted sum of the squared differences between corresponding elements of the input $\boldsymbol{\gamma}$ and pair of empirical marginals $\widehat{\boldsymbol{\gamma}}$. Intuitively, $\text{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}})$ measures the discrepancy between $\boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\gamma}}$, with higher values indicating more discrepancy. The test statistic $T_{n,\psi}$ compares the minimum discrepancy from the empirical marginals $\widehat{\boldsymbol{\gamma}}$ attained by $\boldsymbol{\gamma} \in \Gamma^\psi$ versus that attained by $\boldsymbol{\gamma} \in \Gamma$.

We will reject the null hypothesis that $\boldsymbol{\gamma}^* \in \Gamma^\psi$ for large values of $T_{n,\psi}$, as described below. This involves computing the asymptotic distribution of the statistic under the null hypothesis, and rejecting if $T_{n,\psi}$ exceeds the 0.95 quantile of this distribution.

Let $\boldsymbol{W} = (W_{01}, \ldots, W_{0L}, W_{11}, \ldots, W_{1L})^t \in \mathbb{R}^{2L}$ be a random (column) vector with $W_{aj} = 2{\times}1(A = a)\left\{1(Y = j) - \gamma^*_{aj}\right\}$. Let $\boldsymbol{Z} = (Z_{01}, \ldots, Z_{0L}, Z_{11}, \ldots, Z_{1L})^t \in \mathbb{R}^{2L}$ be a random (column) vector having a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = P_0 \boldsymbol{W}\boldsymbol{W}^t$. Define $C(\boldsymbol{\gamma^*})$ and $C^\psi(\boldsymbol{\gamma^*})$ as

$$C(\boldsymbol{\gamma}^*) = \{r(\boldsymbol{\gamma} - \boldsymbol{\gamma^*}) : \boldsymbol{\gamma} \in \Gamma, r \in \mathbb{R}_+\}, \qquad C^\psi(\boldsymbol{\gamma}^*) = \{r(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma^\psi, r \in \mathbb{R}_+\},$$

where $\mathbb{R}_+$ is the set of nonnegative real numbers.

Theorem 1. *Under the null hypothesis $\boldsymbol{\gamma}^* \in \Gamma^\psi$, $T_{n,\psi}$ converges in distribution to T_ψ defined as*

$$T_\psi = \min_{\boldsymbol{h} \in C^\psi(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) - \min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2). \tag{10}$$

We prove Theorem 1 in the appendix of this paper. The proof also shows the connection between $T_{n,\psi}$ and T_ψ, under the null hypothesis. We derive that

$$n\left\{\inf_{\boldsymbol{\gamma} \in \Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V})\right\} = \min_{\boldsymbol{h} \in C_n^\psi(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \tag{11}$$

$$n\left\{\inf_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V})\right\} = \min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \tag{12}$$

where

$$C_n(\boldsymbol{\gamma}^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma\}, \qquad C_n^\psi(\gamma^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma^*}) : \boldsymbol{\gamma} \in \Gamma^\psi\}.$$

The test statistic $T_{n,\psi}$ is the difference between the left sides of (11) and (12). The limit distribution T_ψ, in (10), is the difference of the right sides, except with $\boldsymbol{D}_n$, $\boldsymbol{Z}_n$, C_n, and C_n^ψ replaced with their limits. Their limits are the identity matrix, $\boldsymbol{Z}$, C, and C^ψ, respectively. Under the alternative hypothesis, the test statistic $T_{n,\psi}$ does not converge to T_ψ, since $C_n^\psi(\boldsymbol{\gamma}^*)$ does not converge to $C^\psi(\boldsymbol{\gamma}^*)$ in this case.

Theorem 2. *Under the alternative hypothesis $\boldsymbol{\gamma}^* \notin \Gamma^\psi$, for any $M \in \mathbb{R}$, $P(T_{n,\psi} > M) \to 1$.*

Intuitively, Theorem 2 is that the test statistic goes to infinity under the alternative hypothesis. This theorem is proved in the Appendix. Since the test statistic converges to a distribution (which we can simulate) under the null hypothesis but to infinity under the alternative hypothesis, our test can differentiate between the null and alternative hypotheses, as the sample size goes to infinity.

For any given ψ, let $t_\psi^{0.95}$ denote the 0.95 quantile of T_ψ. Reject the null hypothesis $\gamma^* \in \Gamma^\psi$ if and only if $T_{n,\psi} > t_\psi^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The tiny perturbation ϵ is required for the proof of pointwise consistency in Section 4.3. Let CS_n be the 95% confidence set constructed by inverting our hypothesis test, i.e.,

$$CS_n = \{\psi : T_{n,\psi} \leq t_\psi^{0.95} + \epsilon\}. \tag{13}$$

4.2 Using Quadratic Programming to Implement the Hypothesis Test

We present how to compute $T_{n,\psi}$ and estimate $t_\psi^{0.95}$. The test statistic $T_{n,\psi}$ can be computed from its form in (8) or (9). We present how to use (8). This requires solving two problems: $\inf_{\boldsymbol{\gamma} \in \Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ and $\inf_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. We show that each is the minimization of a quadratic function subject to a finite number of linear equality and inequality constraints. This is known as a quadratic program. Quadratic programs can be solved efficiently using existing softwares, such as MATLAB or CPLEX.

Consider the problem $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. Define the following as the unknown variables: $\{\pi_{i,j} : i, j \in \mathcal{L}\}$, $\{\gamma_{0i} : i \in \mathcal{L}\}$, $\{\gamma_{1j} : j \in \mathcal{L}\}$. Let $\boldsymbol{H}_t$ denote the vector including all of these variables. The function to be minimized, $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, simplifies to

$$\sum_{a=0}^{1}\sum_{j=1}^{L}\left[P_n(A=a, Y=j) + \gamma_{aj}^2 P_n(A=a) - 2\gamma_{aj} P_n(A=a, Y=j)\right]. \tag{14}$$

Each term of the form $P_n(event)$ is a constant because it can be directly computed from the randomized trial. Note that (14) is a quadratic function of the variables $\boldsymbol{H}_t$. In the problem $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, the function (14) is minimized under the constraint $\boldsymbol{\gamma} \in \Gamma$. By (5), $\boldsymbol{\gamma} \in \Gamma$ means that: $\pi_{i,j} \geq 0$ for all $i, j \in \mathcal{L}$, $\sum_{i,j} \pi_{i,j} = 1$, $\pi_{i,j} = 0$ if $g(i,j) = 0$, $\gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j}$ for all $i \in \mathcal{L}$, $\gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j}$ for all $j \in \mathcal{L}$. These are linear equality and inequality constraints on the variables $\boldsymbol{H}_t$. Therefore, one can solve $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ by finding the minimum of (14) under the above constraints, using quadratic programming.

The other problem required to compute $T_{n,\psi}$ is $\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. Its corresponding quadratic program is the same as that for $\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, except with the additional linear constraint that $\sum_{j>i} \pi_{i,j} = \psi$.

We use simulation to estimate T_ψ, which is the limiting distribution of $T_{n,\psi}$ under the null hypothesis. Each draw from T_ψ is computed as follows. Let $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ denote the minimizer over $\boldsymbol{\gamma} \in \Gamma$ of $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, previously solved to get the test statistic $T_{n,\psi}$. The minimizer may not be unique. We simply let $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ be the minimizer that is returned by "quadprog" in MATLAB. Generate a random draw of $\boldsymbol{Z}$. This requires first estimating $\boldsymbol{\Sigma}$ by replacing $\boldsymbol{\gamma}^*$ by $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ and P_0 by P_n in the definition of $\boldsymbol{\Sigma}$. Next, solve the two quadratic programs in (10). To solve the second quadratic program $\min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*)}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$, define the following variables: $\{\pi_{ij} : i, j \in \mathcal{L}\}$, $\boldsymbol{h} = (h_{01}, \ldots, h_{0L}, h_{11}, \ldots, h_{1L})^t$, $\boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t$. Let $\boldsymbol{H}$ denote the vector including all of these variables. Define the linear constraints: $\pi_{ij} \geq 0$, $\pi_{i,j} = 0$ if $g(i,j) = 0$, $\gamma_{0j} = \sum_{i=1}^{L} \pi_{ij}$, $\gamma_{1j} = \sum_{j=1}^{L} \pi_{ij}$, $\boldsymbol{h} = \boldsymbol{\gamma} - (\sum_{i,j} \pi_{ij})\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ (note, this is a vector of equalities). Define the quadratic program to be $\min \boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2$, over the variables $\boldsymbol{H}$ and under the above linear constraints. To solve the first quadratic program $\min_{\boldsymbol{h}\in C^\psi(\boldsymbol{\gamma}^*)}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$, do as above but add the constraint: $\sum_{i<j} \pi_{ij} = \psi \sum_{i,j} \pi_{ij}$.

We take 1000 draws and compute their 0.95 quantile. Denote this quantile as $\widehat{t}_\psi^{0.95}$. The hat symbol is due to the finite number of draws and because we used an estimate of $\boldsymbol{\Sigma}$. Reject the null hypothesis $\gamma^* \in \Gamma^\psi$ if $T_{n,\psi} > \widehat{t}_\psi^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The confidence set computed from this procedure is denoted as $\widehat{CS}_n$, i.e.,

$$\widehat{CS}_n = \{\psi : T_{n,\psi} \leq \widehat{t}_\psi^{0.95} + \epsilon\}. \tag{15}$$

As the sample size n goes to infinity, the estimate $\widehat{t}_\psi^{0.95}$ converges to $t_\psi^{0.95}$, and therefore $\widehat{CS}_n$ converges to CS_n.

4.3 Properties of Confidence Set and Corresponding Confidence Interval

Theorem 1 implies the following:

Theorem 3. *The confidence set CS_n is pointwise consistent at level 0.95.*

Proof. Consider an arbitrary data generating distribution P_{obs} on (A, Y). Suppose the underlying distribution P_0 on (Y_C, Y_T) satisfies Assumption 1. Choose any ψ that is consistent with the marginal distributions and restrictions, i.e., $\boldsymbol{\gamma}^* \in \Gamma^\psi$. Then for all $\epsilon > 0$:

$$\begin{aligned}
\liminf_{n\to\infty} P_{obs}(\psi \in CS_n) &= \liminf_{n\to\infty} P_{obs}(T_{n,\psi} \leq t_\psi^{0.95} + \epsilon) \\
&\geq \liminf_{n\to\infty} P_{obs}(T_{n,\psi} < t_\psi^{0.95} + \epsilon) \\
&\geq P_{obs}(T_\psi < t_\psi^{0.95} + \epsilon) \\
&\geq P_{obs}(T_\psi \leq t_\psi^{0.95}) \\
&= 0.95,
\end{aligned}$$

where the second inequality follows from Theorem 1 and the Portmanteau Lemma (van der Vaart, 1998). □

Theorem 2 implies the following:

Theorem 4. *For any ψ satisfying $\boldsymbol{\gamma}^* \notin \Gamma^\psi$, the probability that ψ is excluded from CS_n converges to 1.*

Proof. Consider an arbitrary data generating distribution P_{obs} on (A, Y). Suppose the underlying distribution P_0 satisfies Assumption 1. Consider any ψ such that $\boldsymbol{\gamma}^* \notin \Gamma^\psi$. For any given $\epsilon > 0$,

$$\lim_{n\to\infty} P_{obs}(\psi \notin CS_n) \quad = \quad \lim_{n\to\infty} P_{obs}(T_{n,\psi} > t_\psi^{0.95} + \epsilon) = 1.$$

□

The confidence set CS_n generated through test inversion is not guaranteed to be an interval. A confidence interval, denoted as CI_n, is constructed by taking the minimum and maximum of CS_n, i.e., $CI_n = [\min CS_n, \max CS_n]$. From pointwise consistency of the confidence set CS_n, the confidence interval CI_n is also pointwise consistent. We focus on the confidence interval because it is simpler to report, compared to the corresponding set.

Let $\widehat{CI}_n$ denote the confidence interval computed from $\widehat{CS}_n$, where $\widehat{CI}_n = [\min \widehat{CS}_n, \max \widehat{CS}_n]$. To compute it efficiently, we implement the hypothesis test for $\psi = 0$ and for successively larger ψ only until failing to reject, in order to obtain the left endpoint of $\widehat{CI}_n$. To obtain the right endpoint, we implement the hypothesis test for $\psi = 1$ and for successively smaller ψ until failing to reject. This reduces computation time because the hypothesis test does not need to be done for every candidate value of ψ in the grid on $[0, 1]$. Wider intervals will take less time to run.

5 Simulation Studies

We use simulation to assess $\widehat{CI}_n$ at sample sizes n ranging from 200 to 2000. We compare it to the m-out-of-n bootstrap, with respect to coverage probability and average width. Let A and B denote the left and right endpoints of the confidence interval constructed using the m-out-of-n bootstrap. To compute the value A, 10,000 bootstrap data sets are generated, each by sampling $m \leq n$ participants with replacement from the trial data set. Using each bootstrap data set, the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are estimated using the consistent estimators $\overline{\psi}_l^{\mathcal{R}}$ and $\overline{\psi}_u^{\mathcal{R}}$ proposed in Huang *and others* (2017). These estimators are defined in the Supplementary Materials. For their intuition and the proof of consistency, refer to Huang *and others* (2017). The value A is taken to be the 0.025 quantile of the 10,000 lower bound estimates. The value B is the 0.975 quantile of the 10,000 upper bound estimates. The rationale behind the choice of A and B is

$$\begin{aligned} P_{obs}(A \leq \psi_0 \leq B) &\geq P_{obs}(A \leq \psi_l^{\mathcal{R}} \leq \psi_0 \leq \psi_u^{\mathcal{R}} \leq B) \\ &= P_{obs}(A \leq \psi_l^{\mathcal{R}} \leq \psi_u^{\mathcal{R}} \leq B) \\ &= 1 - P_{obs}(A > \psi_l^{\mathcal{R}} \text{ or } B < \psi_u^{\mathcal{R}}) \\ &\geq 1 - P_{obs}(A > \psi_l^{\mathcal{R}}) - P_{obs}(B < \psi_u^{\mathcal{R}}) \\ &\geq 1 - 0.025 - 0.025 \\ &= 0.95. \end{aligned}$$

As a sensitivity analysis, we vary m between $m = n$, $0.9n$, $0.75n$, $0.5n$, and $0.25n$.

5.1 Setup

We compare our method to the m-out-of-n bootstrap in four settings. The settings, labeled A-D, are outlined in Table 1. Each setting is a unique choice of the number of levels L, the marginal distributions $\boldsymbol{\gamma}^*$, the restrictions $\mathcal{R}$, and the true bound parameters $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}})$. For each setting, we conduct a simulation study at each of the following sample sizes: $n = 200$, 500, 1000, and 2000. For Settings A-C, each simulation study includes 5000 simulations. For Setting D, each study includes 1000 simulations since the six-level ordinal outcome results in longer running times. The randomization probability θ is set to 0.5 in all simulations. The steps to run a single simulation are as follows:

1. We generate a data set consisting of the treatment assignments and observed outcomes of n participants, i.e., (A_m, Y_m) with $m = 1, ..., n$. Each participant is randomly assigned to treatment or control using the randomization probability $\theta = 0.5$. Her/his observed outcome is a random draw from either $\boldsymbol{\gamma}^*_{0\cdot}$ or $\boldsymbol{\gamma}^*_{1\cdot}$, depending on the assigned treatment.

2. A 95% CI for the fraction who benefit is computed using the method we proposed in Section 4.

3. A 95% CI for the fraction who benefit is computed using the m-out-of-n bootstrap, as described at the beginning of Section 5.

In each simulation study, we plot the coverage probability of each method as a mapping from $[0, 1]$ to $\mathbb{R}$. For any given $\psi \in [0, 1]$, the coverage probability of ψ equals the proportion of the confidence intervals that contain ψ. In addition, we compute the average width for each method.

5.2 Results

We present the coverage probabilities at $n = 500$ for Settings A and B in Figures 1 and 2, respectively. In each figure, we shade the region from $\psi = \psi_l^{\mathcal{R}}$ to $\psi = \psi_u^{\mathcal{R}}$ in grey. In Setting B, we have $(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}) = (0, 0)$, so the grey region is the thin line at $\psi = 0$. In general, under Assumption 1, the fraction who benefit ψ_0 must be in the grey region and could be anywhere in this region. For ψ in the grey region, the probability that the confidence interval contains ψ should be ≥ 0.95.

In Figures 1 and 2, our method has coverage probabilities ≥ 0.95 for all ψ in the grey region. Moreover, our method achieves this in all four settings and at all sample sizes $n = 200, 500, 1000, 2000$. In contrast, the m-out-of-n bootstrap can have coverage probability < 0.95 in the grey region. This occurs in Setting A for the choices $m = n$, $m = 0.9n$ and $m = 0.75n$. For instance, the coverage probability of $\psi = 0.5$, which lies in the grey region, is 0.95 for our method but 0.89 for m-out-of-n bootstrap ($m = n$). In contrast, for Settings B-D, the choice $m = n$ yields at least the nominal coverage and the smallest average width among the choices of m. This reflects the challenge of choosing m.

In Setting B (Figure 2), the set of ψ for which the null hypothesis $\gamma^* \in \Gamma^\psi$ is true is the single point 0 (under the no harm assumption). An impressive result is that, using our method, the confidence interval is $[0, 0]$ in 50% of the simulations. In other words, our method gives the best possible confidence interval 50% of the time, up to the precision of 0.01. The first point in the grid that should be excluded is $\psi = 0.01$. Our method excludes $\psi = 0.01$ 53% percent of the time. On the other hand, the m-out-of-n bootstrap excludes $\psi = 0.01$ only 6% of the time at best (with $m = n$). Our methods's ability to exclude ψ outside of the grey region translates to large improvements in average width.

Our method can have substantially shorter average width than the competitor. We observe this in Settings B and C. In Setting B, the reduction in average width of our method (compared to the m-out-of-n bootstrap) ranges from 37-69% at $n = 200$, 40-70% at $n = 500$, 41-71% at $n = 1000$, and 43-71% at $n = 2000$. (The ranges are due to trying different options for the choice of m.) In Setting C, the reduction in average width of our method ranges from 6-33% at $n = 200$, 7-32% at $n = 500$, 6-28% at $n = 1000$, and 6-23% at $n = 2000$.

In Settings A and D, the m-out-of-n bootstrap sometimes has narrower average width than our method. In Setting A, this occurs only when the m-out-of-n bootstrap undercovers, i.e., has coverage probability < 0.95 in the grey region. In Setting D, the m-out-of-n bootstrap achieves narrower average width at $n = 200$, with an improvement ranging from 2-14%. At $n = 500$, the m-out-of-n bootstrap offers an improvement of 2% when $m = n$. However, our method has narrower average width at the higher sample sizes, with reductions in average width ranging from 3-23% at $n = 1000$ and 6-22% at $n = 2000$.

Our method and the m-out-of-n bootstrap can have poor coverage if Assumption 1 is violated. Consider Setting B and suppose that the no harm assumption does not hold. Then the true fraction who benefit could be larger than zero, but both our method and the m-out-of-n bootstrap have coverage probabilities below 0.95 for candidate values of $\psi > 0$ (Figure 2. To avoid violating Assumption 1, restrictions should either be based on subject matter knowledge or no restrictions should be made.

6 Application to the CLEAR III trial

6.1 Analysis Procedure

We apply our method to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial (Hanley *and others*, 2017). This was a Phase III trial from 2009-2016 about intraventricular haemorrhage (IVH), which is bleeding into the ventricles of the brain, due to a stroke. CLEAR III tested whether using the drug alteplase (treatment) to remove the blood clot from the ventricle results in a better functional outcome than using saline (control). The trial included 500 participants, with 249 assigned to alteplase and 251 to saline. The primary outcome was the modified Rankin scale (mRS) score at 180 days post-stroke. The mRS score is an ordinal rating of functional outcome with seven levels. The levels are defined as follows (this list is quoted directly from Cheng *and others* (2014)):

0. no symptoms at all
1. no significant disability: despite symptoms, able to perform all usual duties and activities
2. slight disability: unable to perform all previous activities but able to look after own affairs without assistance
3. moderate disability: requiring some help but able to walk without assistance
4. moderately severe disability: unable to walk without assistance and unable to attend to own bodily needs without assistance
5. severe disability: bedridden, incontinent, and requiring constant nursing care and attention
6. death.

Based on CLEAR III, the proportion of patients with 180-day mRS ≤ 3 was estimated as 0.48 under alteplase and 0.45 under saline (95% CI for difference in proportions: [-0.04, 0.12]).

We consider the primary outcome 180-day mRS, as well as the outcomes 30-day mRS, 30-day mortality, and 180-day mortality. For the mRS outcomes, we utilize the full ordinal scale. A separate analysis is performed for each outcome. First, a 95% CI for the fraction of patients who benefit from alteplase (relative to saline) is computed using the method proposed in Section 4. Participants whose outcome is missing are excluded. Second, we attempt to answer the question of who benefits from alteplase compared to saline. An existing hypothesis was that patients with large baseline IVH volumes are more likely to benefit than patients with small baseline IVH volumes. We define 17.5 mL as the threshold between "small" and "large", per the suggestion of neurologist Daniel Hanley (co-author). By this definition, 188 of the 500 participants have small baseline volumes and 312 participants have large baseline volumes. We compute a 95% CI for the fraction of patients with a small baseline IVH volume who benefit from alteplase compared to saline. This is done by applying our method using solely the participants with baseline IVH clot volume ≤ 17.5 mL. The participants whose outcome is missing are excluded. Analogously, we compute a separate 95% CI for patients with a large baseline IVH volume.

All of the CI's are constructed without using any restrictions ($g = 1$), so Assumption 1 is met. In CLEAR III, simple randomization with $\theta = 0.5$ was implemented for the first 100 participants. After this point, a covariate adaptive method was used to achieve balance between the alteplase and saline arms on two pre-selected baseline variables. Our method is currently designed for trials using simple randomization. For the purpose of demonstrating our method, we assume that simple randomization was performed throughout CLEAR III with $\theta = 0.5$. As future work, we will extend our method to handle more randomization schemes.

6.2 Results

The 95% CI's are presented in Table 2. For every CI, we specify the outcome of interest and whether the CI is for all patients, only those with baseline clot volume ≤ 17.5 mL, or only those with baseline clot volume > 17.5 mL. The corresponding sample sizes are presented in Table 3. We discuss the results for 30-day mortality. As shown in Table 2, the 95% CI for this outcome is [0.01, 0.18] for all patients. In words, we are 95% confident that the fraction of patients who benefit with respect to 30-day mortality (i.e., the proportion who would be alive under alteplase but dead

under saline, at 30 days) is between 0.01 and 0.18. Our result contributes new knowledge that was not provided by the difference in proportions ATE. The proportion dead at 30 days was estimated to be 0.06 higher under saline compared to alteplase (95% CI: [-0.00,0.11]). The difference in proportions is a lower bound on the fraction who benefit. Using the 95% CI for the difference in proportions, we could only infer that the fraction who benefit is 0 or above. As shown in Table 2, the 95% CI for the fraction who benefit is $[0.03, 0.13]$ for those with baseline clot volume ≤ 17.5 mL, and $[0, 0.23]$ for those with baseline clot volume > 17.5 mL. The first result suggests that there is a small proportion who benefit among those with ≤ 17.5 mL. However, it is inconclusive whether those with > 17.5 mL are more likely to benefit than those with ≤ 17.5 mL since the 95% CI for > 17.5 mL is wider in both directions. Further work is required to identify the subgroup that benefits with respect to 30-day mortality. One could develop a scalar score using multiple baseline covariates, and stratify patients using the score rather than a single baseline covariate.

As shown in Table 2, the 95% CI's for the mortality outcomes are narrow, while those for the mRS outcomes are very wide. The 95% CI for 180-day mRS and all patients is [0.03, 0.86]. The finite sample size of $n = 491$ (Table 3) contributes to the CI width. However, we believe the driving factor behind the wide width is that the true lower and upper bound parameters span a wide range. In other words, the marginal distributions of the potential outcomes alone are not very informative about the fraction who benefit. Support restrictions can potentially reduce the width of the bounds (Huang *and others*, 2017). In this setting, we are not willing to make restrictions due to lack of supporting subject matter knowledge. We discuss future directions to address wide bound parameters in Section 7.

7 Discussion

We have developed a new method for constructing a 95% confidence interval for the fraction who benefit. It offers the user flexibility to define support restrictions based on subject matter knowledge or to make no assumptions at all on the joint distribution of the potential outcomes. Our confidence interval is proved to be pointwise consistent. In simulation tests, it had empirical coverage of at least 95% when the sample size was varied from 200 to 2000. The method is computationally efficient because it uses quadratic programming to compute the test statistic and the distribution of the test statistic under the null hypothesis. It also avoids having to choose m.

Our simulations and CLEAR III application show that the confidence interval constructed using the method can be narrow and informative. However, we also encountered cases in which the confidence intervals are wide, likely due to the true lower and upper bound parameters being far apart. Our confidence interval is designed so that for any given value between the lower and upper bounds, the coverage of the value is at least 0.95. Consequently, if the bounds are far apart, the confidence intervals will tend to be wide even for extremely large sample sizes. Huang *and others* (2017) found that incorporating a baseline variable can substantially narrow the bounds, without requiring assumptions. As future work, we will incorporate a baseline variable into our method. Also, we will provide guidance on how to select a baseline variable that will be effective in tightening the bounds.

8 Funding

CLEAR III was supported by the grant 5U01 NS062851-05, awarded to D.F.H. from the National Institutes of Health, National Institute of Neurological Disorders and Stroke. E.J.H. was supported by the U.S. Food and Drug Administration (U01 FD004977-01) and the National Institute on Aging, USA (T32AG000247). This paper's contents are solely the responsibility of the authors and do not represent the views of these organizations.

9 Supplementary Materials and Code

The supplementary materials and MATLAB code are available upon request. If interested, please email Emily Huang at ehuang19@jhu.edu.

A Asymptotic distribution of test statistic under the null hypothesis

Claim 1. *Under the null hypothesis $\boldsymbol{\gamma}^* \in \Gamma^\psi$, the statistic T_n converges in distribution to T.*

Proof. We use the general argument in Section 5.1.3 of Shapiro *and others* (2014), except tailored to our specific problem. The proof here is self-contained. The null hypothesis $\psi_0 = \psi$ implies that the minimizer $\boldsymbol{\gamma}^*$ of $\min_{\boldsymbol{\gamma}\in\Gamma} P_0 F(\boldsymbol{\gamma}, \boldsymbol{V})$ is unique, satisfies $\gamma^*_{aj} = P_0(Y = j|A = a)$ for each $a \in \{0, 1\}, j \in \{1, \ldots, L\}$, and also $\boldsymbol{\gamma}^* \in \Gamma^\psi$. This implies $\nabla P_0 F(\boldsymbol{\gamma}^*, \boldsymbol{V}) = 0$.

Define $\boldsymbol{Z}_n = n^{1/2}\{\nabla P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V}) - \nabla P_0 F(\boldsymbol{\gamma}^*, \boldsymbol{V})\}$, where the gradient is with respect to $\boldsymbol{\gamma}^*$. It follows that $\boldsymbol{Z}_n = (Z_{01,n}, \ldots, Z_{0L,n}, Z_{11,n}, \ldots, Z_{1L,n})^t$, where

$$Z_{aj,n} = -2n^{1/2} P_n \left[1(A = a)\left\{1(Y = j) - \gamma^*_{aj}\right\}\right],$$

for each $a \in \{0, 1\}, j \in \{1, \ldots, L\}$. By the multivariate central limit theorem, $\boldsymbol{Z}_n$ converges in distribution to $\boldsymbol{Z}$ defined above. Let $\boldsymbol{D}_n$ denote the $2L \times 2L$ diagonal matrix with first L diagonal elements equal to $2P_n 1(A = 0)$ and last L diagonal elements equal to $2P_n 1(A = 1)$. Recall we assume that $P_0(A = a) = 1/2$ for each $a \in \{0, 1\}$. It follows that $(\boldsymbol{Z}_n, \boldsymbol{D}_n)$ converges in distribution to $(\boldsymbol{Z}, \boldsymbol{D})$, for $\boldsymbol{D}$ the $2L \times 2L$ identity matrix.

We next show

$$n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V})\right\} = \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \tag{16}$$

$$n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V})\right\} = \min_{\boldsymbol{h}\in C^\psi_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2), \tag{17}$$

for

$$C_n(\boldsymbol{\gamma}^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma\}, \qquad C^\psi_n(\boldsymbol{\gamma}^*) = \{n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma^\psi\}.$$

To show (16), we have

$$\begin{aligned}
& n\left\{\inf_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*, \boldsymbol{V})\right\} \\
= \; & n \inf_{\boldsymbol{\gamma}\in\Gamma} P_n \{F(\boldsymbol{\gamma}, \boldsymbol{V}) - F(\boldsymbol{\gamma}^*, \boldsymbol{V})\} \\
= \; & n \inf_{\boldsymbol{\gamma}\in\Gamma} \sum_{a=0}^{1}\sum_{j=1}^{L} P_n 1(A = a)\left[\{1(Y = j) - \gamma_{aj}\}^2 - \left\{1(Y = j) - \gamma^*_{aj}\right\}^2\right] \\
= \; & n \inf_{\boldsymbol{\gamma}\in\Gamma} \sum_{a=0}^{1}\sum_{j=1}^{L} P_n 1(A = a)\left[-2\left\{1(Y = j) - \gamma^*_{aj}\right\}(\gamma_{aj} - \gamma^*_{aj}) + (\gamma_{aj} - \gamma^*_{aj})^2\right] \\
= \; & \inf_{\boldsymbol{\gamma}\in\Gamma}\left[n^{1/2}\sum_{a=0}^{1}\sum_{j=1}^{L} Z_{aj,n}(\gamma_{aj} - \gamma^*_{aj}) + \sum_{a=0}^{1} P_n 1(A = a)\sum_{j=1}^{L}\left\{n^{1/2}(\gamma_{aj} - \gamma^*_{aj})\right\}^2\right] \\
= \; & \inf_{\boldsymbol{\gamma}\in\Gamma}\left[n^{1/2}(\boldsymbol{\gamma} - \boldsymbol{\gamma})^t \boldsymbol{Z}_n + \sum_{a=0}^{1} P_n 1(A = a)\sum_{j=1}^{L}\left\{n^{1/2}(\gamma_{aj} - \gamma^*_{aj})\right\}^2\right] \\
= \; & \min_{h\in C_n(\boldsymbol{\gamma}^*)} h^t \boldsymbol{Z}_n + h^t \boldsymbol{D}_n h/2,
\end{aligned} \tag{18}$$

which proves (16). The proof of (17) is analogous, except replacing Γ by Γ^ψ and $C_n(\boldsymbol{\gamma}^*)$ by $C^\psi_n(\boldsymbol{\gamma}^*)$.

Taking the difference between the left sides of (17) and (16), we have

$$T_n = \min_{\boldsymbol{h}\in C^\psi_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) - \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2).$$

Since $C_n(\boldsymbol{\gamma}^*) \uparrow C(\boldsymbol{\gamma}^*)$, $C^\psi_n(\boldsymbol{\gamma}^*) \uparrow C^\psi(\boldsymbol{\gamma}^*)$, and $(\boldsymbol{Z}_n, \boldsymbol{D}_n)$ converges in distribution to $(\boldsymbol{Z}, \boldsymbol{D})$, it follows from the continuous mapping theorem that T_n converges in distribution to T. □

B Asymptotic distribution of test statistic under the alternative hypothesis

Claim 2. *Under the alternative hypothesis* $H_a(\psi) : \boldsymbol{\gamma}^* \notin \Gamma^\psi$*, for any* $M \in \mathbb{R}$*,* $P(T_{n,\psi} > M) \to 1$*.*

Proof. Assume the alternative hypothesis $H_a(\psi) : \boldsymbol{\gamma}^* \notin \Gamma^\psi$ holds. Choose any $M \in \mathbb{R}$. By Lemma 1,

$$T_{n,\psi} = n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \right\},$$

where $\widehat{\theta} = P_n A$ and $\text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1} \sum_{j=1}^{L} \left[(\gamma_{aj} - \widehat{\gamma}_{aj})^2 \left(\widehat{\theta}\right)^a \left(1 - \widehat{\theta}\right)^{1-a} \right]$.

Let $\text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \sum_{a=0}^{1} \sum_{j=1}^{L} \left[\left(\gamma_{aj} - \gamma^*_{aj}\right)^2 \theta^a (1-\theta)^{1-a} \right]$. We have as $n \to \infty$

$$\left(\widehat{\boldsymbol{\gamma}}, \widehat{\theta}\right) \xrightarrow{p} (\boldsymbol{\gamma}^*, \theta),$$

by the Weak Law of Large Numbers, Slutsky's lemma, and Theorem 2.7(vi) in van der Vaart (2000).

For any given $n \in \mathbb{N}$, we have

$$\begin{aligned} P(T_{n,\psi} > M) &= P\left(n \left\{ \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \right\} > M \right) \\ &= P\left(\inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) > \frac{M}{n} \right). \end{aligned} \tag{19}$$

For conciseness, let

$$d_{n,\psi} = \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}).$$

Thus, we have $P(T_{n,\psi} > M) = P(d_{n,\psi} > M/n)$. By the two lemmas for this proof (see Supplementary Materials) and the Continuous Mapping Theorem, we have $\inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \textit{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ and $\inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \inf_{\boldsymbol{\gamma} \in \Gamma} \textit{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$. Let $c = \inf_{\boldsymbol{\gamma} \in \Gamma^\psi} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ and $b = \inf_{\boldsymbol{\gamma} \in \Gamma} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$. If $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, $\text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = 0$ because $\gamma_{aj} = \gamma^*_{aj}$ for all (a, j) pairs. If $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}^*$, we have $\text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) > 0$ since $\gamma_{aj} \neq \gamma^*_{aj}$ for some (a, j) pair and $0 < \theta < 1$. Since $\boldsymbol{\gamma}^* \in \Gamma$, we have that $b = 0$. We have $c > 0$ since $\boldsymbol{\gamma}^* \notin \Gamma^\psi$ and Γ^ψ is compact, which is proved in the Supplementary Materials. By Slutsky's lemma, the random variable $d_{n,\psi}$ converges in probability to positive number c. Let $\epsilon = c/100$. Since the sequence M/n converges to 0 as $n \to \infty$, we have for sufficiently large n,

$$P(d_{n,\psi} \geq c - \epsilon) \leq P(d_{n,\psi} > M/n) \leq 1.$$

Because $d_{n,\psi}$ converges in probability to a, the probability on the left converges to 1. By the Squeeze Theorem, $P(d_{n,\psi} > M/n) = P(T_{n,\psi} > M)$ converges to 1. □

References

Borusyak, K. (2015). Bounding the population shares affected by treatments. *Technical Report: SSRN: http://ssrn.com/abstract=2473827*.

Cheng, B., Forkert, N.D., Zavaglia, M., Hilgetag, C.C., Golsari, A., Siemonsen, S., Fiehler, J., Pedraza, S., Puig, J., Cho, T.H. *and others*. (2014). Influence of stroke infarct location on functional outcome measured by the modified Rankin Scale. *Stroke* **45**(6), 1695–1702.

Friedman, Lawrence M, Furberg, Curt D and DeMets, David L. (2010). *Fundamentals of Clinical Trials Fourth Edition*. Springer.

Gordis, Leon. (2009). Epidemiology (fourth edition).

Hanley, Daniel F, Lane, Karen, McBee, Nichol, Ziai, Wendy, Tuhrim, Stanley, Lees, Kennedy R, Dawson, Jesse, Gandhi, Dheeraj, Ullman, Natalie, Mould, Andrew *and others*. (2017). Thrombolytic removal of intraventricular haemorrhage in treatment of severe stroke: results of the randomised, multicentre, multiregion, placebo-controlled clear iii trial. *Lancet*.

Huang, Emily J, Fang, Ethan X, Hanley, Daniel F and Rosenblum, Michael. (2017). Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics* **18**(2), 308–324.

Lu, Jiannan, Ding, Peng and Dasgupta, Tirthankar. (2016). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *arXiv preprint arXiv:1507.01542*.

Romano, Joseph P and Shaikh, Azeem M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference* **138**(9), 2786–2807.

Shapiro, A., Dentcheva, D. and Ruszczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*, MOS-SIAM Series on Optimization.

Snapinn, Steven M and Jiang, Qi. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials* **8**, 31–36.

van der Vaart, AW. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A.W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Table 1: Simulation Settings.

	L	γ^*_{01}	γ^*_{02}	γ^*_{11}	γ^*_{12}	User-defined restrictions $\mathcal{R}$	$(\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}})$
A	2	0.5	0.5	0.5	0.5	no restrictions	(0,0.5)
B	2	0.5	0.5	0.5	0.5	no harm	(0,0)
C	2	0.5	0.5	0.25	0.75	no restrictions	(0.25, 0.5)
D	6	empirical marginals in MISTIE II				no restrictions	(0.82,0.96)

Table 2: 95% CI's for the fraction who benefit from alteplase compared to saline. For each outcome, we present the 95% CI's for all patients, only patients with baseline IVH volume $\leq$ 17.5 mL, and only patients with baseline IVH volume $>$ 17.5 mL.

	All patients	$\leq$ 17.5 mL only	$>$ 17.5 mL only
30-day mRS	[0.00,0.64]	[0.00,0.79]	[0.00,0.61]
180-day mRS	[0.03,0.86]	[0.00,0.92]	[0.01,0.83]
30-day mortality	[0.01,0.18]	[0.03,0.13]	[0.00,0.23]
180-day mortality	[0.05,0.34]	[0.02,0.23]	[0.04,0.44]

Table 3: Sample Sizes for 95% CI's in Table 2. For every pairing of outcome with patient group, we present the sample sizes in the format, total (saline, alteplase).

	All patients	$\leq$ 17.5 mL only	$>$ 17.5 mL only
30-day mRS	494 (249,245)	186 (95,91)	308 (154,154)
180-day mRS	491 (245,246)	185 (93,92)	306 (152,154)
30-day mortality	500 (251,249)	188 (95, 93)	312 (156,156)
180-day mortality	495 (247,248)	187 (94,93)	308 (153,155)

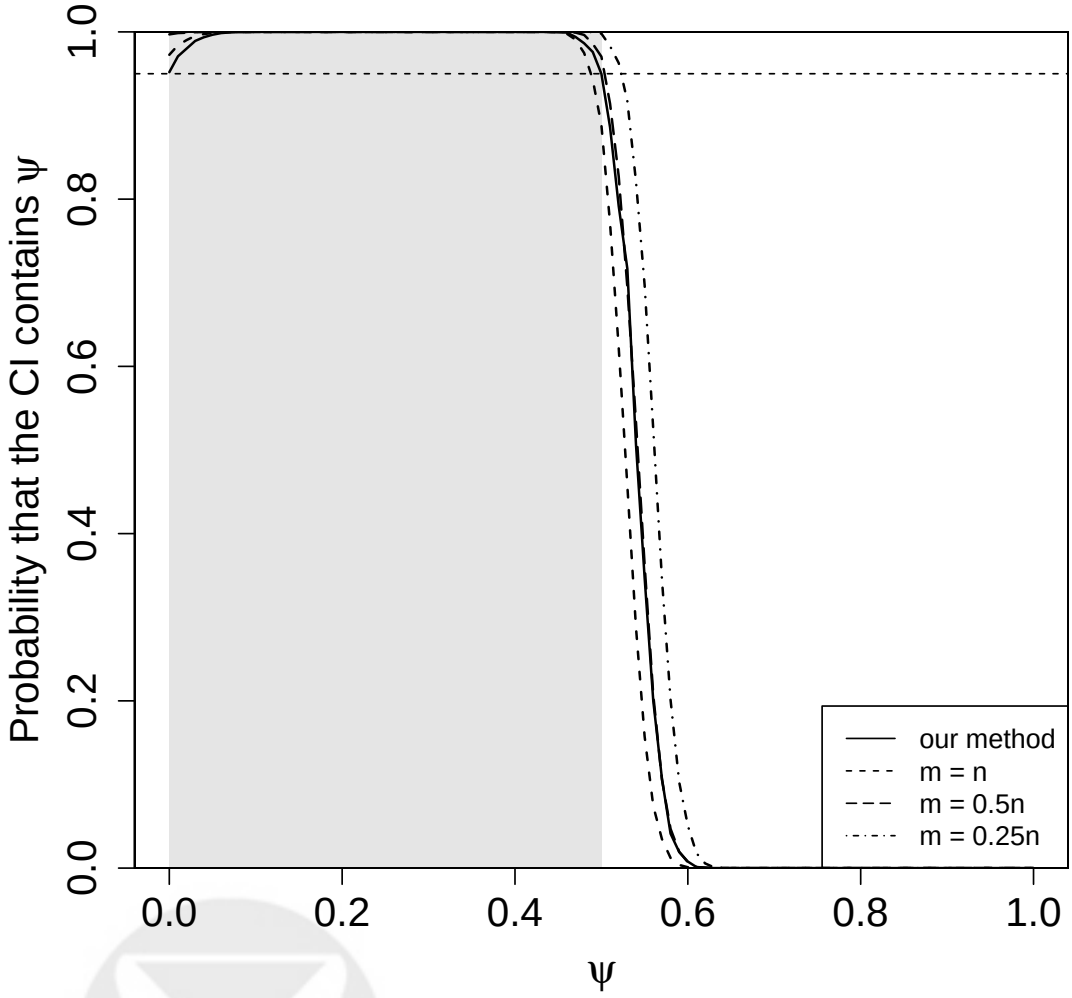

Figure 1: Coverage probabilities in Setting A at $n = 500$. *The grey region spans from* $\psi = 0$ *to* $\psi = 0.5$, *which are the lower and upper bounds* $\psi_l^{\mathcal{R}}$ *and* $\psi_u^{\mathcal{R}}$ *in Setting A. To achieve good coverage under Assumption 1, coverage probabilities should be* ≥ 0.95 *for all* ψ *in the grey region. For legibility of the plot, the curves for* $m = 0.9n$ *and* $m = 0.75n$ *are not shown. They lie between the curves for* $m = n$ *and* $m = 0.5n$, *but closely resemble the curve for* $m = n$.

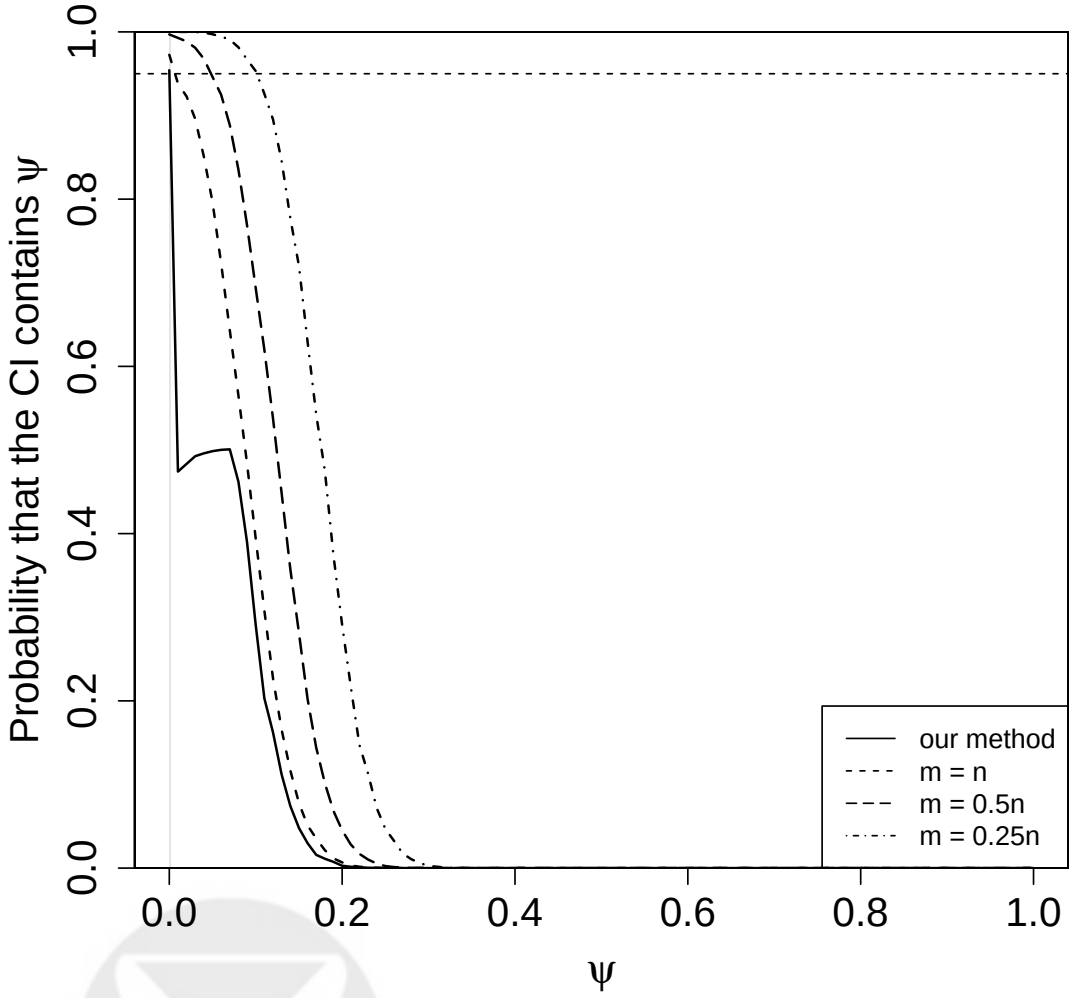

Figure 2: Coverage probabilities in Setting B at $n = 500$. *The grey region is the single point $\psi = 0$, since in Setting B the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ are both zero. To achieve good coverage under Assumption 1, coverage probabilities should be ≥ 0.95 at $\psi = 0$. For legibility of the plot, the curves for $m = 0.9n$ and $m = 0.75n$ are not shown. They lie between the curves for $m = n$ and $m = 0.5n$, but closely resemble the curve for $m = n$.*