Johns Hopkins University, Dept. of Biostatistics Working Papers

10-25-2017

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*
Department of Biostatistics, Harvard University
Ethan X. Fang
Department of Statistics, Pennsylvania State University
Daniel F. Hanley
Division of Brain Injury Outcomes, Johns Hopkins Medical Institutions
Michael Rosenblum
Department of Biostatistics, Johns Hopkins University

October 18, 2017

Abstract

The fraction who benefit from treatment is the proportion of patients whose potential outcome under treatment is better than that under control. Inference on this parameter is challenging since it is only partially identifiable, even in our context of a randomized trial. We propose a new method for constructing a confidence interval for

*The authors gratefully acknowledge the U.S. Food and Drug Administration (U01 FD004977-01 and HHSF223201400113C), the National Institute on Aging (T32AG000247), the National Institute on Drug Abuse (P50 DA039838), and the Patient-Centered Outcomes Research Institute (ME-1306-03198). CLEAR III was supported by the grant 5U01 NS062851-05, awarded to Daniel Hanley from the U.S. National Institute of Neurological Disorders and Stroke. The MISTIE II trial was funded by R01NS046309 (PI: Daniel Hanley) U.S. National Institute of Neurological Disorders and Stroke. This paper's contents are solely the responsibility of the authors and do not represent the views of these organizations.

the fraction, when the outcome is ordinal or binary. Our confidence interval procedure is pointwise consistent. It does not require any assumptions about the joint distribution of the potential outcomes, although it has the flexibility to incorporate various user-defined assumptions. Unlike existing confidence interval methods for partially identifiable parameters (such as m-out-of-n bootstrap and subsampling), our method does not require selection of m or the subsample size. It 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, which can be computed efficiently using optimization tools. In simulation, our method attains the nominal coverage probability or higher, and can have substantially narrower average width than m-out-of-n bootstrap. We apply it to a trial of a new intervention for stroke.

Keywords: Causal inference; Potential outcome; Quadratic program; Treatment effect heterogeneity.

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 provides different information than the average treatment effect. For example, a positive average treatment effect could represent a small benefit to many or a large benefit to a small subpopulation. The fraction who benefit can help distinguish between these scenarios. It may be informative to medical researchers; for example, a small fraction indicates that an exclusive subgroup benefits and resources could be devoted toward identifying it. We aim to draw inferences about the fraction who benefit, using a randomized trial.

A parameter is partially identifiable if it cannot be determined from the data generating distribution and model assumptions, but we can deduce that it lies within a set (Manski, 2010, p.178). In general, the fraction who benefit (abbreviated as the fraction) is only partially identifiable from observed data, even in the randomized trial context. This occurs because only one potential outcome can be observed per patient. Generally, identification of the fraction necessitates strong, untestable assumptions about the joint distribution of the potential outcomes, such as independence of the potential outcomes within a person. We do not require any assumptions about the joint distribution, but do allow the user to incorporate certain types of assumptions if desired. Since the fraction is partially 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 (Fan and Park, 2010). 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 et al., 2017), and the standard bootstrap can be inconsistent in such cases. Fan and Park (2009, 2010) apply bootstrap-based methods to construct confidence intervals for bounds on the fraction. Another existing method for constructing confidence intervals for the fraction 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. Under the subsampling condition (i) in Theorem 3.4 of their paper, Romano and Shaikh (2008) prove pointwise consistency of their method. It is difficult to establish whether this condition holds in our problem. A challenge in using m-out-of-n bootstrap or subsampling is how to select m to achieve desired performance. We propose a new method that avoids having to select m.

Through simulation, we compare our method with the m-out-of-n bootstrap with respect to coverage probability and average width. In all cases, the coverage probability of our method is at or above the nominal level, while that of m-out-of-n bootstrap is sometimes below the nominal level. In some cases, our method achieves substantially narrower average width than the m-out-of-n bootstrap, e.g., reduction of 40%. Our method achieves the desired 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 4.

We apply our method to the CLEAR III (Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Haemorrhage III) randomized trial of a new treatment for severe stroke, which had a sample size of 500 patients (Hanley et al., 2017). Outcomes included disability measured by the modified Rankin Scale and death. As examples of the output of our procedure, the 95% confidence interval for the fraction 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. The widths of these confidence intervals depend on the data generating distribution and the sample size. We investigate this relationship in a variety of scenarios through the aforementioned simulation study.

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. As a simple example, M-estimators can be represented in terms of solving stochastic optimization problems (van der Vaart, 1998, chap. 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 et al. (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.

The statistic derived using the above approach can be computed with 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. Each confidence interval in the CLEAR III application was computed within 4 to 8 minutes.

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. The parameter 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, 2014, chap. 8). A limitation to dichotomization of the outcome is that improvements not crossing the dichotomization threshold are ignored. The fraction who benefit considers the full ordinal scale.

Our general approach can be used to construct confidence intervals for a variety of partially identifiable parameters, such as the fraction who are harmed by treatment, the fraction who benefit above a given threshold, and the average treatment effect among those who benefit by at least the clinically meaningful, minimum threshold.

In Section 2, we describe the data generating distribution and state assumptions that are used throughout the paper. Our proposed method is presented in Section 3, including theorems about its asymptotic properties. We evaluate the method through simulation in Section 4. It is applied to the CLEAR III randomized trial in Section 5. Future work is discussed in Section 6.

2 Notation, Parameter Definition, and Assumptions

2.1 Parameter Definition

Consider an ordinal outcome with a finite number of levels, L. We assume that the levels are numbered as integers from 1 to L, in order of least to most favorable. Let Y_T be the

potential outcome under treatment and Y_C be the potential outcome under control. Let P_0 denote the 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 with potential outcome pair (y_C, y_T) benefits from treatment compared to control, if $y_T > y_C$. She/he is harmed by treatment if $y_T < y_C$, and experiences no individual treatment effect if $y_T = y_C$. The fraction, our parameter of interest, is $\psi_0 = P_0(Y_T > Y_C) = \sum_{j>i} \pi_{i,j}$.

We propose a method to construct a confidence interval for the parameter ψ_0. The method does not require assumptions about the distribution P_0. However, it 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. Under no restrictions, 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:

$$\mathcal{R} = \{P \text{ on } (Y_C, Y_T) : P(Y_C = i, Y_T = j) = 0 \text{ for all } i, j \text{ with } g(i, j) = 0\}.$$

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

2.2 Observed Data Distribution

We consider the context of a randomized trial. Let n be the number of participants. For each participant m, denote her/his study arm assignment by $A_m \in \{0, 1\}$ (0 for control and 1 for treatment) and observed outcome by $Y_m \in \mathcal{L}$. We assume that the vectors (A_m, Y_m), $m = 1, \ldots, n$, are fully observed, i.e., no data is missing. Other assumptions include:

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 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 et al., 2015, chap. 6). The value θ is the probability of being assigned to treatment, which is known. Assumption 4 connects observed outcomes to potential outcomes and is called the consistency assumption.

2.3 Partial Identifiability of the Fraction who Benefit

The assumptions above imply that the vectors $\boldsymbol{V}_m = (A_m, Y_m)$, $m = 1, .., n$, are independent and identically distributed. Let $\boldsymbol{V} = (A, Y)$ denote the random vector corresponding to a generic trial participant. The vector $\boldsymbol{V}$ is called the observed data, to distinguish it from the vector of potential outcomes (Y_C, Y_T) which is partially unobserved. In the rest of the paper, P_0 applied to a function of $\mathbf{V}$ is understood as the induced distribution on the observed data vector $\mathbf{V}$ under P_0. Let the column vector $\boldsymbol{\gamma}^* = (\gamma_{01}^*, .., \gamma_{0L}^*, \gamma_{11}^*, .., \gamma_{1L}^*)^t$ denote the marginal distributions of the potential outcomes under control and treatment, 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 $y \in \mathcal{L}$:

$$\gamma_{0y}^* = P_0(Y_C = y) = P_0(Y = y \mid A = 0); \quad \gamma_{1y}^* = P_0(Y_T = y) = P_0(Y = y \mid A = 1). \quad (1)$$

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

Because only one potential outcome is observed per participant, the fraction ψ_0 typically is not point identified from observed data. However, the marginal distributions $\boldsymbol{\gamma}^*$ and restrictions $\mathcal{R}$ may rule out certain possibilities. Let $\psi_l^{\mathcal{R}}(P_0)$ and $\psi_u^{\mathcal{R}}(P_0)$ denote the sharp lower and upper bounds on the fraction, given the marginal distributions and restrictions:

$$\psi_l^{\mathcal{R}}(P_0) = \min\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\};$$
$$\psi_u^{\mathcal{R}}(P_0) = \max\{P(Y_T > Y_C) : P \text{ has marginal distributions equal to } \boldsymbol{\gamma}^* \text{ and } P \in \mathcal{R}\}.$$

These bounds, discussed in Huang et al. (2017), are functions of P_0 due to their dependence on $\boldsymbol{\gamma}^*$, and are identifiable since $\boldsymbol{\gamma}^*$ is identifiable. For conciseness, we suppress their dependence on P_0. The fraction ψ_0 must lie between the bounds, i.e., $\psi_0 \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$. Also, for any $\psi \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$, 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 Appendix A of the Supplementary Materials. Intuitively, the marginal distributions and restrictions rule out candidate values of ψ outside the range $[\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$, while candidates inside the range are not ruled out.

Previous work on the bounds for an ordinal outcome includes Gadbury et al. (2004); Borusyak (2015); Lu et al. (2016); Huang et al. (2017). Gadbury et al. (2004) derive closed-form expressions for the bounds in the case of $L = 2$ and no restrictions. Borusyak (2015) represents the bounds as solutions to linear programs. Lu et al. (2016) derive closed-form expressions in the case of $L \geq 2$ and no restrictions. Huang et al. (2017) propose consistent estimators of the bounds.

A confidence set for ψ_0 is defined as a measureable function that maps the observed data $\{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$ to a subset of the unit interval. 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 $P_0 \in \mathcal{R}$, we have for all $\psi \in \left[\psi_l^{\mathcal{R}}(P_0), \psi_u^{\mathcal{R}}(P_0)\right]$:*

$$\liminf_{n\to\infty} P_0(\psi \in CS_n) \geq 1-\alpha.$$

Roughly speaking, pointwise consistency is that, for any data generating distribution $P_0 \in \mathcal{R}$ and any $\psi \in [\psi_l^{\mathcal{R}}(P_0), \psi_u^{\mathcal{R}}(P_0)]$, 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 ψ_0 must be in the range $[\psi_l^{\mathcal{R}}(P_0), \psi_u^{\mathcal{R}}(P_0)]$ and the observed data distribution provides no information on where it lies within that range.

3 Proposed Confidence Interval Procedure

3.1 Overview

We construct a 95% confidence set for the fraction 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 4 and 5), the grid has a point at every hundredth. 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 as endpoints the smallest and largest points of the set. We present our hypothesis test in Sections 3.2-3.5 and provide its implementation in Section 3.6. The asymptotic properties of the resulting confidence interval are presented in Section 3.5. For simplicity, this section focuses on the case where the assignment probability $\theta = 0.5$.

3.2 Hypothesis Test for Candidate Value of ψ_0

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

$$\Gamma = \left\{ \boldsymbol{\gamma} = (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L})^t \in \mathbb{R}^{2L} : \begin{array}{l} \text{For some } \boldsymbol{\pi} \in \Pi, \text{ we have} \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0; \\ \gamma_{0i} = \sum_{j=1}^{L} \pi_{i,j} \text{ for all } i \in \mathcal{L}; \\ \gamma_{1j} = \sum_{i=1}^{L} \pi_{i,j} \text{ for all } j \in \mathcal{L}. \end{array} \right\}. \quad (2)$$

This set is comprised of the pairs of marginal distributions (under control and treatment) that are compatible with the restrictions encoded by g. If no restrictions are made, 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 $L = 2$ and the no harm assumption is made, the set Γ comprises all vectors satisfying $\gamma_{12} \geq \gamma_{02}$ and the property in the previous sentence. Under Assumption 1, the pair of marginal distributions $\boldsymbol{\gamma}^*$ is in the set Γ.

Consider any candidate value of $\psi \in [0, 1]$ for the parameter ψ_0. Define the set Γ^ψ as (2), but adding the constraint that the fraction who benefit equals ψ, i.e, $\sum_{j>i} \pi_{i,j} = \psi$. The set Γ^ψ is comprised of the pairs of marginal distributions (under control and treatment) that are compatible with the restrictions $\mathcal{R}$ and the fraction who benefit equal to ψ. The sets Γ and Γ^ψ are not random. Also, each is a compact, convex polyhedron.

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

$$\begin{aligned} H_0(\psi) &: \boldsymbol{\gamma}^* \in \Gamma^\psi \\ H_a(\psi) &: \boldsymbol{\gamma}^* \notin \Gamma^\psi. \end{aligned}$$

The null hypothesis means that the pair of marginals $\boldsymbol{\gamma}^*$ (which is a function of the population distribution on the observed data vector (A, Y)) is compatible with both the restric-

tions $\mathcal{R}$ 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 $\psi \notin [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$. The null hypothesis means that the candidate value of ψ is not ruled out by the marginals $\boldsymbol{\gamma}^*$ and the restrictions $\mathcal{R}$.

As an example, consider a binary outcome with failure = 1 and success = 2. Assume the marginals $\boldsymbol{\gamma}^* = (0.5, 0.5, 0.25, 0.75)$ and we make no restrictions on the joint distribution P_0 on (Y_C, Y_T). The purpose of this example is to illustrate which values of ψ can be ruled out just from knowledge of the marginals. The joint distribution P_0 on (Y_C, Y_T) is pictured in Figure 1 (left panel). The fraction ψ_0 equals $\pi_{1,2}$, in the upper right cell. For the candidate fraction who benefit $\psi = 0$, the alternative hypothesis $H_a(\psi)$ holds. There exists no joint distribution of the form in Figure 1 (middle panel), i.e., with its marginals equal to $\boldsymbol{\gamma}^*$ and 0 in its upper right cell. Satisfying this form would require the lower right cell to be 0.75, but this contradicts that the second row sum is 0.5. For the candidate fraction who benefit $\psi = 0.5$, the null hypothesis $H_0(\psi)$ is true. There is a joint distribution with its marginals equal to $\boldsymbol{\gamma}^*$ and its upper right cell equal to 0.5, as shown in Figure 1 (right panel). The challenge below in constructing a confidence interval for ψ_0 is to rule out candidate values in the setting where the marginal distributions are estimated from the data in a randomized trial; also, we consider ordinal-valued outcomes, which correspond to $L \times L$ matrices.

3.3 Statistic For Testing Null Hypothesis $H_0(\psi)$

If Γ^{ψ} is the empty set, this implies $\psi_0 = \psi$ is incompatible with the restrictions $\mathcal{R}$ and $H_0(\psi)$ is false, so we immediately reject $H_0(\psi)$. Else, if either arm has zero participants, then we fail to reject $H_0(\psi)$. Below, we consider the case where neither of these extreme situations occurs.

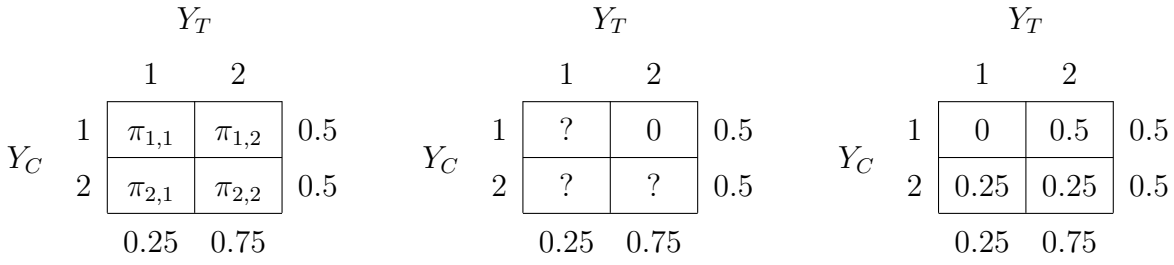

Figure 1: Example Showing Partial Identifiability of ψ_0 for Binary Outcome. The figure at the left depicts the unknown joint distribution on (Y_C, Y_T), under known marginals. The figure in the middle is used to illustrate why the alternative hypothesis $H_a(\psi)$ holds for candidate fraction who benefit $\psi = 0$. The figure at the right illustrates why the null hypothesis $H_0(\psi)$ holds for candidate fraction who benefit $\psi = 0.5$.

We use the notation P_0X to denote the expectation of X with respect to P_0. Define

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

where $1(S)$ denotes the indicator variable that has value 1 if S is true and 0 otherwise. The minimizer of $P_0F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to $\boldsymbol{\gamma}^*$ defined in (1), as proved in Appendix B of the Supplementary Materials.

We define the test statistic at sample size n corresponding to the null hypothesis $H_0(\psi)$ as

$$T_{n,\psi} = n \left\{ \min_{\boldsymbol{\gamma} \in \Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) - \min_{\boldsymbol{\gamma} \in \Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V}) \right\}, \tag{3}$$

where P_n denotes the empirical distribution, i.e., $P_nF(\boldsymbol{\gamma}, \boldsymbol{V}) = \frac{1}{n} \sum_{m=1}^{n} F(\boldsymbol{\gamma}, \boldsymbol{V}_m)$. Since $\Gamma^\psi \subseteq \Gamma$, we have $T_{n,\psi} \geq 0$. We use min instead of inf in the definition of $T_{n,\psi}$ since the minimum is always achieved, due to F being continuous and each of Γ and Γ^ψ being

compact. Each term in (3) has a unique minimizer, as proved in Appendix C of the Supplementary Materials.

Let $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_{01}, \ldots, \widehat{\gamma}_{0L}, \widehat{\gamma}_{11}, \ldots, \widehat{\gamma}_{1L})^t$ denote the empirical marginal distributions under control and treatment, i.e., for each arm $a \in \{0, 1\}$ and outcome value $i \in \mathcal{L}$,

$$\widehat{\gamma}_{ai} = \sum_{m=1}^{n} 1(A_m = a, Y_m = i) \Big/ \sum_{m=1}^{n} 1(A_m = a) \, ,$$

where we set $\widehat{\gamma}_{ai}$ to 0 if the rightmost sum equals 0. We next define the following function of $\widehat{\boldsymbol{\gamma}}$ and a generic vector $\boldsymbol{\gamma} \in \mathbb{R}^{2L}$:

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

The above function is a weighted sum of the squared differences between corresponding elements of the input $\boldsymbol{\gamma}$ and 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.

Lemma 1

$$T_{n,\psi} = n \left\{ \min_{\boldsymbol{\gamma} \in \Gamma^{\psi}} \textit{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \min_{\boldsymbol{\gamma} \in \Gamma} \textit{Discrep}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \right\} . \tag{4}$$

Lemma 1 is proved in Appendix C of the Supplementary Materials. It shows that the test statistic $T_{n,\psi}$ contrasts the minimum discrepancy between the empirical marginals $\widehat{\boldsymbol{\gamma}}$ and the set of vectors $\boldsymbol{\gamma} \in \Gamma^{\psi}$ versus the analogous discrepancy over $\boldsymbol{\gamma} \in \Gamma$.

3.4 Deriving the Null Distribution of Test Statistic $T_{n,\psi}$

Our method rejects the null hypothesis $H_0(\psi)$ for large values of $T_{n,\psi}$. Specifically, the hypothesis test involves approximating the asymptotic distribution of the statistic under the null hypothesis, and rejecting $H_0(\psi)$ if $T_{n,\psi}$ exceeds the 0.95 quantile of this distribution.

We give an overview of the second order approximation from Shapiro et al. (2014, pp. 166–169) that we applied to our problem to derive the asymptotic null distribution of $T_{n,\psi}$ given in Theorem 1 below. We also give a self-contained proof in Appendix D of the Supplementary Materials, which we constructed after deriving the asymptotic null distribution. The key approximation we use from Shapiro et al. (2014, pp. 166–169) is based on the second order, functional delta method, which is that under regularity conditions we have the following expansion:

$$\phi(f_n) - \phi(f) = \phi'_f(f_n - f) + (1/2)\phi''_f(f_n - f) + o_p(1/n), \tag{5}$$

for ϕ a second order, Hadamard differentiable function, and a sequence of random functions f_n and fixed function f such that $n^{1/2}(f_n - f)$ converges in distribution to a random element.

In our case, we let $f_n(\boldsymbol{\gamma}) = P_n F(\boldsymbol{\gamma}, \mathbf{V}), f(\boldsymbol{\gamma}) = P_0 F(\boldsymbol{\gamma}, \mathbf{V})$, and ϕ denote the min-function over $\boldsymbol{\gamma} \in \Gamma$, i.e., $\phi(\delta) = \min_{\boldsymbol{\gamma}\in\Gamma} \delta(\boldsymbol{\gamma})$, where the domain of ϕ is the set $\mathcal{F}$ of Lipschitz continuous functions δ from $\Gamma \to \mathbb{R}$. The min-function ϕ is second order Hadamard differentiable at f in direction $\delta \in \mathcal{F}$ with first order directional derivative $\phi'_f(\delta) = \delta(\boldsymbol{\gamma}^*)$ (recall $\boldsymbol{\gamma}^* = \arg\min_{\gamma\in\Gamma} f(\boldsymbol{\gamma})$) and second order directional derivative

$$\phi''_f(\delta) = \min_{\tilde{\boldsymbol{h}} \in C(\boldsymbol{\gamma}^*, \Gamma)} \left\{ 2\tilde{\boldsymbol{h}}^t \nabla\delta(\boldsymbol{\gamma}^*) + \tilde{\boldsymbol{h}}^t D^2 f(\boldsymbol{\gamma}^*) \tilde{\boldsymbol{h}} \right\}, \tag{6}$$

for set of column vectors $C(\boldsymbol{\gamma}^*, \Gamma) = \{r(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) : \boldsymbol{\gamma} \in \Gamma, r \in \mathbb{R}_+\}$, $\mathbb{R}_+$ the nonnegative reals, and ∇, D^2 the gradient and Hessian with respect to $\boldsymbol{\gamma}$ evaluated at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$; this follows from Theorem 7.23 of Shapiro et al. (2014, p. 352). Substituting the above functions into (5) and setting $\delta = f_n - f$ gives

$$\min_{\boldsymbol{\gamma}\in\Gamma} f_n(\boldsymbol{\gamma}) - f(\boldsymbol{\gamma}^*) = f_n(\boldsymbol{\gamma}^*) - f(\boldsymbol{\gamma}^*) + (1/2)\phi''_f(f_n - f) + o_p(1/n),$$

since $\phi'_f(\delta) = \delta(\boldsymbol{\gamma}^*) = f_n(\boldsymbol{\gamma}^*) - f(\boldsymbol{\gamma}^*)$. It follows from the above display and (6) that

$$\min_{\boldsymbol{\gamma}\in\Gamma} f_n(\boldsymbol{\gamma}) - f_n(\boldsymbol{\gamma}^*) = \min_{\tilde{\boldsymbol{h}} \in C(\boldsymbol{\gamma}^*, \Gamma)} \left\{ \tilde{\boldsymbol{h}}^t \nabla(f_n - f)(\boldsymbol{\gamma}^*) + \tilde{\boldsymbol{h}}^t D^2 f(\boldsymbol{\gamma}^*) \tilde{\boldsymbol{h}}/2 \right\} + o_p(1/n). \tag{7}$$

We next consider the limit distribution of $\nabla(f_n - f)(\boldsymbol{\gamma}^*)$ in the above display, which equals

$$\nabla(f_n - f)(\boldsymbol{\gamma}^*) = (P_n - P_0)\nabla F(\boldsymbol{\gamma}, \mathbf{V})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} = (P_n - P_0)\boldsymbol{W},$$

where $\boldsymbol{W} = (W_{01}, \ldots, W_{0L}, W_{11}, \ldots, W_{1L})^t$ with $W_{aj} = 2 \times 1(A = a)\left\{\gamma^*_{aj} - 1(Y = j)\right\}$ for each $a \in \{0, 1\}, j \in \mathcal{L}$. By the central limit theorem, the limit distribution of $\mathbf{Z}_n = n^{1/2}\nabla(f_n - f)(\boldsymbol{\gamma}^*)$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = P_0\boldsymbol{W}\boldsymbol{W}^t$. Let $\boldsymbol{Z} = (Z_{01}, \ldots, Z_{0L}, Z_{11}, \ldots, Z_{1L})^t$ denote a random vector with this limit distribution. The Hessian matrix $D^2 f$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ is the identity matrix for the case of $\theta = 1/2$. It follows from the above arguments, multiplying both sides of (7) by n, and substituting $\mathbf{h} = n^{1/2}\tilde{\mathbf{h}}$, that

$$n\left\{\min_{\boldsymbol{\gamma}\in\Gamma} f_n(\boldsymbol{\gamma}) - f_n(\boldsymbol{\gamma}^*)\right\} = \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma)}\left\{\boldsymbol{h}^t\mathbf{Z}_n + \boldsymbol{h}^t\boldsymbol{h}/2\right\} + o_p(1). \tag{8}$$

Under the null hypothesis $H_0(\psi)$, the analogous formula as above holds replacing Γ by Γ^ψ, i.e.,

$$n\left\{\min_{\boldsymbol{\gamma}\in\Gamma^\psi} f_n(\boldsymbol{\gamma}) - f_n(\boldsymbol{\gamma}^*)\right\} = \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma^\psi)}\left\{\boldsymbol{h}^t\mathbf{Z}_n + \boldsymbol{h}^t\boldsymbol{h}/2\right\} + o_p(1). \tag{9}$$

Taking the difference between (9) and (8) implies that the null distribution of the test statistic (3) has the second order approximation

$$\begin{aligned} T_{n,\psi} &= n\left\{\min_{\boldsymbol{\gamma}\in\Gamma^\psi} f_n(\boldsymbol{\gamma}) - \min_{\boldsymbol{\gamma}\in\Gamma} f_n(\boldsymbol{\gamma})\right\} \\ &= \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma^\psi)}\left\{\boldsymbol{h}^t\mathbf{Z}_n + \boldsymbol{h}^t\boldsymbol{h}/2\right\} - \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma)}\left\{\boldsymbol{h}^t\mathbf{Z}_n + \boldsymbol{h}^t\boldsymbol{h}/2\right\} + o_p(1). \end{aligned}$$

Taking the limit as n goes to infinity yields the following, for which the formal proof is given in Appendix D of the Supplementary Materials:

Theorem 1 *Under the null hypothesis $H_0(\psi)$, $T_{n,\psi}$ converges in distribution to T_ψ, where*

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

The distribution of T_ψ depends on the value ψ and on the data generating distribution P_0 through Σ defined above. We approximate the distribution of T_ψ as described in Section 3.6. As an example, Figure 2 shows this distribution in Setting A (described in Section 4) for $\psi = 0.5$. In this case, the distribution of T_ψ is a mixture of a point mass at 0 with probability 0.25 and a density with support on the positive reals.

Theorem 2 *Under the alternative hypothesis $H_a(\psi)$, for any $M \in \mathbb{R}$, $P(T_{n,\psi} > M) \to 1$ as $n \to \infty$.*

Under the alternative hypothesis $H_a(\psi)$, the test statistic grows arbitrarily large, by Theorem 2 which is proved in Appendix E of the Supplementary Materials. Since the statistic converges to a distribution (which can be approximated from the data) under the null hypothesis but grows arbitrarily large under the alternative hypothesis, our test can differentiate between the null $H_0(\psi)$ and alternative $H_a(\psi)$, as the sample size goes to infinity.

For any $\psi \in [0, 1]$, let $t_\psi^{0.95}$ denote the 0.95 quantile of T_ψ. Our test rejects the null hypothesis $H_0(\psi)$ if and only if $T_{n,\psi} > t_\psi^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The addition of ϵ accounts for the error tolerance in our computations below, which involve solving quadratic programs. Let CS_n be the 95% confidence set constructed by inverting our hypothesis test:

$$CS_n = \{\psi \in [0, 1] : T_{n,\psi} \leq t_\psi^{0.95} + \epsilon\}. \tag{11}$$

3.5 Properties of Confidence Set and Corresponding Confidence Interval

Using Theorems 1 and 2, we prove the following theorems in Appendices F and G of the Supplementary Materials:

Theorem 3 *The confidence set CS_n in (11) is pointwise consistent at level 0.95.*

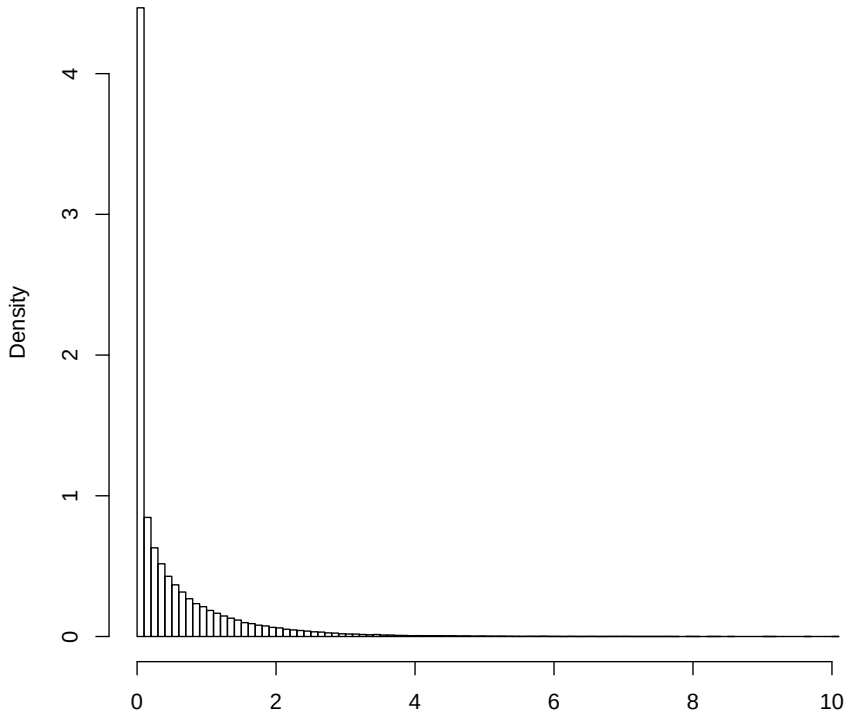

Figure 2: Histogram of T_ψ in Setting A (described in Section 4), with $\psi = 0.5$. This plot is obtained through simulation, using 100,000 draws from T_ψ. The empirical probability of $T_\psi = 0$ was 0.25.

Theorem 4 *For any ψ for which the null hypothesis $H_0(\psi)$ is false, i.e., for which $\boldsymbol{\gamma}^* \notin \Gamma^\psi$, the probability that ψ is excluded from CS_n in (11) converges to 1.*

The confidence set CS_n may not 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 pointwise consistent. We focus on the confidence interval because it is simpler to report than the corresponding set.

3.6 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 (3) or (4). We present how to use (3). This requires solving two optimization problems, $\min_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ and $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. We show that each is the minimization of a quadratic function subject to a finite number of linear equality and inequality constraints. This is called a quadratic program, which can be solved efficiently using existing software such as MATLAB.

Consider the optimization problem $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. 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 $\mathbf{x}$ denote the vector including all of these variables. The function to be minimized, $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, simplifies to

$$\sum_{a=0}^{1}\sum_{j=1}^{L}\left[\gamma_{aj}^2 P_n 1(A=a) - 2\gamma_{aj} P_n 1(A=a, Y=j) + P_n 1(A=a, Y=j)\right]. \qquad (12)$$

The expression (12) is a quadratic function of the variables $\mathbf{x}$. In the optimization problem $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, the function (12) is minimized under the constraint $\boldsymbol{\gamma} \in \Gamma$, which by (2) can be represented by linear equality and inequality constraints on the variables

$\mathbf{x}$. Thus, the constrained minimization problem $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$ can be solved by quadratic programming. The other optimization problem required to compute $T_{n,\psi}$ is $\min_{\boldsymbol{\gamma}\in\Gamma^\psi} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$. Its quadratic program is the same as for $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, except with the additional linear constraint $\sum_{j>i} \pi_{i,j} = \psi$.

We use Monte Carlo simulation to approximate the distribution of T_ψ. Each draw from T_ψ is computed as follows. Let $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ denote the minimizer over $\boldsymbol{\gamma} \in \Gamma$ of $P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$, that is, we have $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}} \in \Gamma$ and $P_n F(\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}, \boldsymbol{V}) = \min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma}, \boldsymbol{V})$; as noted above, this minimizer is unique. Next, we generate a random draw from the distribution of $\boldsymbol{Z}$, where we approximate the covariance matrix $\boldsymbol{\Sigma}$ by replacing $\boldsymbol{\gamma}^*$ by $\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ and P_0 by P_n in the definition of $\boldsymbol{\Sigma}$. We then solve the two quadratic programs in (10). To solve the second quadratic program $\min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma)}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$, 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 $\tilde{\boldsymbol{x}}$ denote the vector including all of these variables. Define the linear constraints: $\pi_{i,j} \geq 0$, $\pi_{i,j} = 0$ if $g(i,j) = 0$, $\gamma_{0i} = \sum_{j=1}^{L} \pi_{ij}$, $\gamma_{1j} = \sum_{i=1}^{L} \pi_{ij}$, $\boldsymbol{h} = \boldsymbol{\gamma} - (\sum_{i,j} \pi_{ij})\widehat{\boldsymbol{\gamma}}_{\mathcal{R}}$ (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 $\tilde{\boldsymbol{x}}$ and under the above linear constraints. To solve the first quadratic program $\min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma^\psi)}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$, do as above but add the constraint that $\sum_{i<j} \pi_{ij} = \psi \sum_{i,j} \pi_{ij}$.

We repeat the above procedure 1000 times, which generates 1000 independent draws of T_ψ. We then compute the empirical 0.95 quantile, denoted as $\widehat{t}_\psi^{0.95}$. The hat symbol indicates that we use estimates of $\boldsymbol{\gamma}^*$ and the covariance matrix $\boldsymbol{\Sigma}$ in generating draws from the distribution of T_ψ. We reject the null hypothesis $H_0(\psi)$ if $T_{n,\psi} > \widehat{t}_\psi^{0.95} + \epsilon$, where $\epsilon = 10^{-10}$. The confidence set computed from this procedure is $\widehat{CS}_n = \{\psi \in G[0,1] : T_{n,\psi} \leq \widehat{t}_\psi^{0.95} + \epsilon\}$, where $G[0,1]$ is a grid on $[0,1]$, e.g., $\{0, 0.01, \ldots, 0.99, 1\}$.

Let $\widehat{CI}_n$ denote the confidence interval computed from $\widehat{CS}_n$. That is, we have $\widehat{CI}_n = [\min \widehat{CS}_n, \max \widehat{CS}_n]$. To efficiently compute $\widehat{CI}_n$, we perform the hypothesis test for $\psi = 0$

and for successively larger ψ until failing to reject, in order to obtain the left endpoint. To obtain the right endpoint, we perform the hypothesis test for $\psi = 1$ and for successively smaller ψ until failing to reject. This reduces computation time because the hypothesis test need not be done for every candidate value of ψ in $G[0, 1]$. Run time can be further reduced by computing the tests for different ψ in parallel, with different computing nodes.

4 Simulation Studies

4.1 Confidence Interval Procedure Based on m-out-of-n Bootstrap

We use simulation to assess our method $\widehat{CI}_n$ at sample sizes n ranging from 200 to 2000. We also compare $\widehat{CI}_n$ to a competitor method that utilizes m-out-of-n bootstrap. The competitor method is to construct a one-sided 0.975 confidence interval denoted $[A, 1]$ for the lower bound $\psi_l^{\mathcal{R}}$ using m-out-of-n, percentile bootstrap, and analogously a one-sided 0.975 confidence interval denoted $[0, B]$ for the upper bound $\psi_u^{\mathcal{R}}$ using m-out-of-n, percentile bootstrap, and taking their intersection. If both of the confidence intervals are pointwise consistent, then asymptotically the coverage probability for any point in $[\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$ will be at least 0.95. Appendices H and I of the Supplementary Materials present how A and B are obtained. We consider subsample sizes $m \in \{0.25n, 0.5n, 0.75n, 0.9n, n\}$.

4.2 Data Generating Distributions

We consider four simulation settings, labeled A-D, shown in Table 1. In each setting, the number of levels L, the marginal distributions $\boldsymbol{\gamma}^*$, and the restrictions $\mathcal{R}$ are specified; these determine the corresponding bound parameters $[\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$.

Table 1: Simulation Settings

Setting	L	$\boldsymbol{\gamma}^*$	User-defined restrictions $\mathcal{R}$	$\psi_l^{\mathcal{R}}$	$\psi_u^{\mathcal{R}}$
A	2	$(0.5, 0.5, 0.5, 0.5)^t$	no restrictions	0	0.5
B	2	$(0.5, 0.5, 0.5, 0.5)^t$	no harm	0	0
C	2	$(0.5, 0.5, 0.25, 0.75)^t$	no restrictions	0.25	0.5
D	6	MISTIE II empirical marginals	no restrictions	0.82	0.96

Setting D is designed to mimic features from the MISTIE II (Minimally Invasive Surgery for Intracerebral Hemorrhage Evacuation Phase II) randomized trial, which compared a new surgical intervention for stroke to standard care (Hanley et al., 2016). The outcome is ordinal-valued with six levels, representing the reduction in clot volume (after discretizing into intervals of 5 mL). We define this outcome, abbreviated as RICV5, in Appendix J of the Supplementary Materials. In Setting D, we set the marginal distributions $\boldsymbol{\gamma}^*$ under treatment and control to be those observed in the MISTIE II trial, which are shown in Figure 1 of the Supplementary Materials.

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 consists of 5000 simulated trials. For Setting D, each study consists of 1000 trials, since the six-level ordinal outcome requires longer running times. The steps to generate and analyze data for a single simulated trial are as follows. First, we generate a data set consisting of the treatment assignments and observed outcomes of n participants, i.e., (A_m, Y_m) with $m = 1, \ldots, n$. Each participant is randomly assigned to treatment or control using the randomization probability $\theta = 0.5$. Given A_m, the outcome Y_m is a draw from the multinomial distribution on $(1, \ldots, L)$ with probabilities corresponding to the marginal distribution $(\gamma_{a1}^*, \ldots, \gamma_{aL}^*)$ for

$a = A_m$. Second, a 95% confidence interval for the fraction who benefit is computed using $\widehat{CI}_n$ defined in Section 3.6. Third, a 95% confidence interval for the fraction is computed using the m-out-of-n bootstrap, as described in Section 4.1.

For each simulation study, we compute the empirical coverage probability and average confidence interval width of each method. For any given $\psi \in [0, 1]$, the coverage probability of ψ equals the proportion of the confidence intervals that contain ψ.

4.3 Simulation Results

We present the coverage probabilities at $n = 500$ for Settings A and B in Figures 3 and 4, respectively. The plots for the other settings and sample sizes are given in the Supplementary Materials (Figures 2 - 15). 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 ψ_0 must be in the grey region and could be anywhere in this region. For any ψ in the grey region, the goal is to have the probability that the confidence interval contains ψ be ≥ 0.95.

In Figures 3 and 4, 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, for some values of m. This occurs in Figure 3 (Setting A, $n = 500$) for $\psi = 0.5$. The coverage probability is 0.95 for our method (as is desired since $H_0(\psi)$ is true), but for m-out-of-n bootstrap the coverage probability is 0.89 ($m = n$), 0.90 ($m = 0.9n$), 0.93 ($m = 0.75n$), 0.97 ($m = 0.5n$), and 1.00 ($m = 0.25n$). The choices $m = 0.5n$ and $m = 0.25n$ do not lead to undercoverage problems in our simulations. However, in all but one case, they have larger average widths than our method. For example, in Setting B at $n = 200$, the average width of our method is 0.09, while the

average widths are 0.20 and 0.28 for $m = 0.5n$ and $m = 0.25n$, respectively. In the exception case (Setting D at $n = 200$), m-out-of-n bootstrap with $m = 0.5n$ yields a slightly shorter average width (0.004 difference in absolute units) compared to our method.

In Setting B (Figure 4), the set of ψ for which the null hypothesis $H_0(\psi)$ is true is the single point $\{0\}$ (under the no harm assumption). 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. The m-out-of-n bootstrap excludes it only 6% of the time at best (with $m = n$). Our method's ability to exclude ψ outside of the grey region translates to large improvements in average width as described below.

In Figure 3, we show zoomed-in figures of the upper part of the grey region. They show that, at the left and right edges, our method (the solid line) has the nominal coverage probability 0.95, while the m-out-of-n bootstrap has coverage probability above 0.95. In Figure 4, the zoom-in on the upper left of the plot shows that, at $\psi = 0$ (the only point in the grey region), our method has the nominal coverage probability 0.95, while the m-out-of-n bootstrap again has probability above 0.95. Therefore, our method not only achieves the nominal coverage probability in all of our simulation cases, but also can be less conservative than the m-out-of-n bootstrap at the boundaries of the identified region $[\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$.

The average widths of our method and the m-out-of-n bootstrap are tabulated in Tables 1-4 of the Supplementary Materials. For each method, the average widths decrease with higher sample size. At the largest sample size $n = 2000$, the average width of our method is approximately the difference between the upper and lower bound parameters. Our method can have substantially shorter average width than the m-out-of-n bootstrap. In Setting B, the reduction in average width of our method (compared to the m-out-of-n bootstrap)

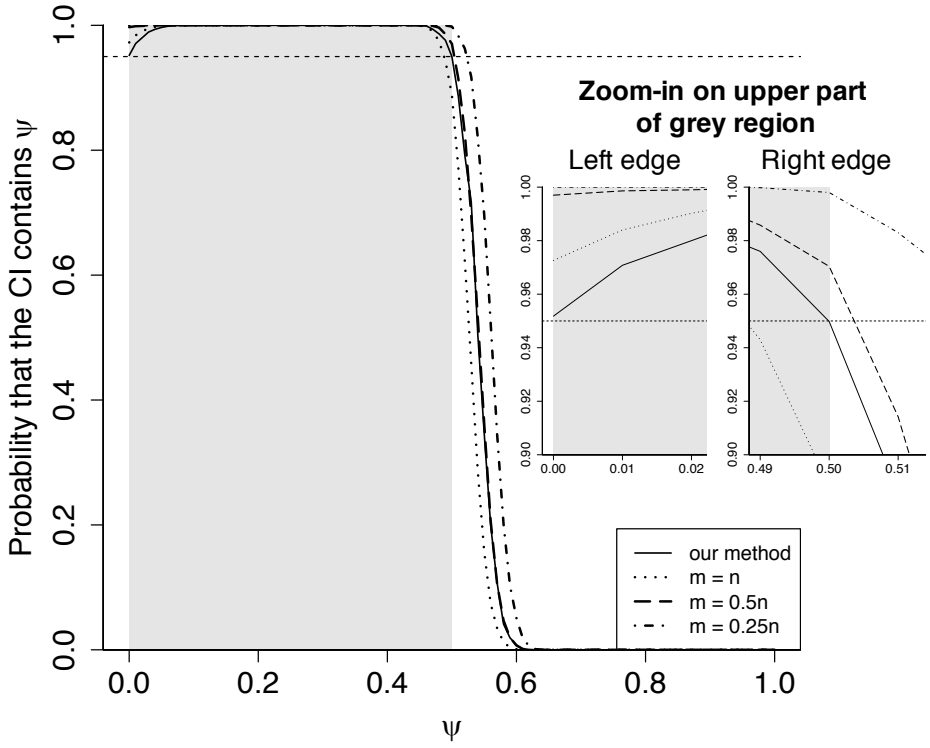

Figure 3: Coverage probabilities in Setting A at $n = 500$ for our method (solid) and the m-out-of-n bootstrap with $m = n$ (dots), $m = 0.5n$ (dashes), and $m = 0.25n$ (dot-dash). The grey region spans from $\psi = 0$ to $\psi = 0.5$, which are the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ 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 $m = n$. For $\psi > 0.5$, the curve for $m = 0.5n$ closely resembles the curve for our method.

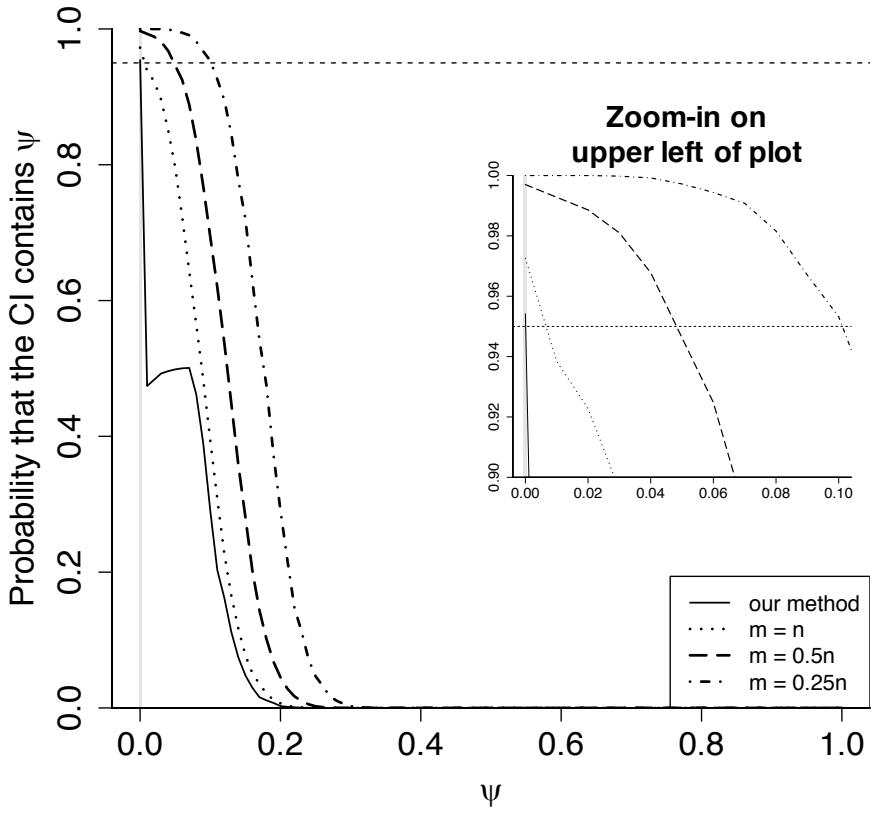

Figure 4: Coverage probabilities in Setting B at $n = 500$ for our method (solid) and the m-out-of-n bootstrap with $m = n$ (dots), $m = 0.5n$ (dashes), and $m = 0.25n$ (dot-dash). The grey region is the single point $\psi = 0$, since in Setting B the lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ 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 $m = n$.

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 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 $n = 2000$.

Unsurprisingly, our method and the m-out-of-n bootstrap can have poor coverage if Assumption 1 is violated, that is, if the assumed restrictions $\mathcal{R}$ fail to hold. Consider Setting B and suppose that the no harm assumption does not hold. Then the 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 4).

Our method achieves the nominal coverage probability 0.95 or higher, even in cases where the lower and upper bound parameters are non-differentiable functions of the marginal distributions under treatment and control. For binary outcomes and under no restrictions, the lower bound is non-differentiable when $\gamma^*_{02} = \gamma^*_{12}$; the upper bound is non-differentiable when $\gamma^*_{01} = \gamma^*_{12}$. Hence, the bounds $\psi^{\mathcal{R}}_l$ and $\psi^{\mathcal{R}}_u$ are non-differentiable in Setting A. Figure 3 shows that, despite the non-differentiability, our method has coverage probability 0.95 at $\psi = \psi^{\mathcal{R}}_l$ and $\psi = \psi^{\mathcal{R}}_u$.

5 Application to the CLEAR III Randomized Trial

5.1 Analysis Procedure

We apply our method to a Phase III randomized trial called CLEAR III (Hanley et al., 2017). This trial enrolled patients with intraventricular haemorrhage, or bleeding into the ventricles of the brain, due to a stroke. It tested whether using the drug alteplase (treatment) to remove the blood clot from the ventricles results in a better functional outcome than using saline (control). The trial had 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 ranging from 0 = no symptoms to 6 = death (Cheng et al., 2014). Based on CLEAR III, the probability of having 180-day mRS ≤ 3 was estimated as 0.48 under alteplase and 0.45 under saline (95% confidence interval 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 mRS, we utilize the full ordinal scale. A separate analysis is performed for each outcome. We apply our method $\widehat{CI}_n$ from Section 3 to compute a 95% confidence interval for the fraction who benefit from alteplase (relative to saline). Also, we compute 95% confidence intervals using m-out-of-n bootstrap, with choices of m as considered in Section 4. Participants with missing outcomes are excluded. Out of 500 participants, the number of participants excluded is 6 for 30-day mRS, 9 for 180-day mRS, 0 for 30-day mortality, and 5 for 180-day mortality.

The confidence intervals are constructed without any restrictions ($g = 1$), so Assumption 1 is met. In CLEAR III, a covariate adaptive randomization method was used to achieve balance between the alteplase and saline arms on two pre-selected baseline vari-

ables. For the purpose of demonstrating our method, we assume that simple randomization was performed throughout CLEAR III with $\theta = 0.5$. A potential area of future research is to extend our method to handle more randomization schemes.

5.2 Results

Table 2 shows the confidence intervals for each outcome. We discuss the results for 30-day mortality, shown in the third column. The 95% confidence interval computed using our method is [0.01, 0.18]. 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. Using m-out-of-n bootstrap, the 95% confidence interval computed when $m = n$ is [0, 0.19], which is very close to our result. The results for the other m are also comparable, except when $m = 0.25n$.

For 30-day mRS, the confidence interval for our method is 0.04 wider (absolute units) than those for m-out-of-n bootstrap when $m = n$ and $m = 0.9n$. However, the narrower width of m-out-of-n bootstrap could potentially be due to its coverage probability falling below the nominal rate 0.95, which we observed in Simulation Setting A for these choices of m at the same sample size as the CLEAR III trial. For m-out-of-n bootstrap, the choice of m can affect the result. For 180-day mortality, the confidence interval is [0.04, 0.35] for $m = n$, and [0, 0.38] for $m = 0.5n$. A feature of our method is that it does not require selecting m.

The confidence intervals for mortality are narrow, while those for mRS are wide. For 180-day mRS, the 95% confidence interval outputted by our method is [0.03, 0.86]. One possible explanation for the wide width is that the marginal distributions of the potential outcomes are not very informative about the fraction, so the lower and upper bound parameters span a wide range. Support restrictions can potentially reduce the width of the

Table 2: 95% Confidence Intervals for the Fraction who Benefit from Alteplase Compared to Saline, in the CLEAR III Data Analysis

	30-day mRS	180-day mRS	30-day mortality	180-day mortality
$\widehat{CI}_n$	[0 , 0.64]	[0.03, 0.86]	[0.01, 0.18]	[0.05, 0.34]
$m = 1.00n$	[0.03, 0.63]	[0.04, 0.84]	[0 , 0.19]	[0.04, 0.35]
$m = 0.90n$	[0.03, 0.63]	[0.03, 0.84]	[0 , 0.19]	[0.03, 0.36]
$m = 0.75n$	[0.02, 0.65]	[0.03, 0.84]	[0 , 0.20]	[0.02, 0.36]
$m = 0.50n$	[0.02, 0.67]	[0.02, 0.85]	[0 , 0.21]	[0 , 0.38]
$m = 0.25n$	[0.01, 0.71]	[0 , 0.87]	[0 , 0.24]	[0 , 0.41]

bounds (Huang et al., 2017). We do not make restrictions in this application due to lack of supporting subject matter knowledge. We discuss future directions to address wide bound parameters in Section 6.

6 Discussion

Our simulations and CLEAR III application show that the confidence interval constructed using the proposed method can be narrow. We also encountered some cases in which the confidence intervals are wide, possibly due to the lower and upper bound parameters being far apart. Our confidence interval is designed so that for any given value between the bounds, the coverage probability is at least 0.95. Consequently, if the bounds are far apart, the confidence intervals will tend to be wide. Huang et al. (2017) found that incorporating a baseline variable can in some cases substantially narrow the bounds. A potential area for future research is to incorporate a baseline variable into our method.

We can apply our general approach to any parameter ψ defined as a linear combination of $\pi_{i,j}$, which include the fraction who benefit (sum over $\pi_{i,j}$ with $j > i$), the fraction who are harmed (sum over $\pi_{i,j}$ with $j < i$), the fraction who benefit above a given threshold t (sum over $\pi_{i,j}$ with $j - i > t$), and the average treatment effect among those who benefit by at least the clinically meaningful, minimum threshold t (sum over $(j-i)\pi_{i,j}$ with $j - i \geq t$ and then divide by the sum $\sum_{j-i \geq t} \pi_{i,j}$). The same statistic can be used, except that Γ^{ψ} is redefined by replacing $\psi = \sum_{j>i} \pi_{i,j}$ by the new parameter definition. Alternative statistics could be constructed by adding positive weights to the definition of F in Section 3.3.

In the CLEAR III analysis, the proportion of participants with missing data was only 1-2% for each outcome. However, in other trials, there may be higher rates of missingness. A potential area of future research is to address the issue of missing outcome data, such as by using double robust estimators of the marginal distributions that account for dropout, instead of the empirical marginal distributions ignoring participants with missing outcomes.

References

Borusyak, K. (2015), ""Bounding the Population Shares Affected by Treatments," *Technical Report*, SSRN: 2473827.

Cheng, B., Forkert, N. D., Zavaglia, M., Hilgetag, C. C., Golsari, A., Siemonsen, S., Fiehler, J. et al. (2014), "Influence of Stroke Infarct Location on Functional Outcome Measured by the Modified Rankin Scale," *Stroke*, 45(6), 1695–1702.

Fan, Y., and Park, S. (2009), "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets," *Advances in Econometrics*, 25, 3–70.

Fan, Y., and Park, S. (2010), "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory*, 26(03), 931–951.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (2015), *Fundamentals of Clinical Trials*, Cham: Springer.

Gadbury, G., Iyer, H., and Albert, J. (2004), "Individual Treatment Effects in Randomized Trials with Binary Outcomes," *Journal of Statistical Planning and Inference*, 121(2), 163–174.

Gordis, L. (2014), *Epidemiology*, Philadelphia: Elsevier Saunders.

Hanley, D. F., Lane, K., McBee, N., Ziai, W., Tuhrim, S., Lees, K. R., Dawson, J., Gandhi, D., Ullman, N., Mould, W. A. et al. (2017), "Thrombolytic Removal of Intraventricular Haemorrhage in Treatment of Severe Stroke: Results of the Randomised, Multicentre, Multiregion, Placebo-controlled CLEAR III Trial," *The Lancet*, 389(10069), 603–611.

Hanley, D. F., Thompson, R. E., Muschelli, J., Rosenblum, M., McBee, N., Lane, K., Bistran-Hall, A. J. et al. (2016), "Safety and Efficacy of Minimally Invasive Surgery Plus Alteplase in Intracerebral Haemorrhage Evacuation (MISTIE): a Randomised, Controlled, Open-label, Phase 2 Trial," *The Lancet Neurology*, 15(12), 1228–1237.

Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017), "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?," *Biostatistics*, 18(2), 308–324.

Lu, J., Ding, P., and Dasgupta, T. (2016), "Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds," *Technical Report*, arXiv: 1507.01542.

Manski, C. F. (2010), "Partial Identification in Econometrics," in *Microeconometrics*, eds. S. N. Durlauf, and L. E. Blume, New York: Palgrave Macmillan, chapter 10, pp. 178–188.

Romano, J. P., and Shaikh, A. 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*, MOS-SIAM Series on Optimization, Philadelphia: Society for Industrial and Applied Mathematics.

Snapinn, S. M., and Jiang, Q. (2007), "Responder Analyses and the Assessment of a Clinically Relevant Treatment Effect," *Trials*, 8, 31–36.

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

Supplementary Material for "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

October 18, 2017

Appendix A proves that, for each value ψ between the bound parameters, there exists a joint distribution P that has marginals $\boldsymbol{\gamma}^*$, is in the set $\mathcal{R}$, and satisfies $P(Y_T > Y_C) = \psi$. Appendix B proves that the minimizer of $P_0 F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equals $\boldsymbol{\gamma}^*$. Appendix C proves Lemma 1 and uniqueness of the minimizer in each term in the test statistic $T_{n,\psi}$. Appendices D, E, F, and G prove Theorems 1, 2, 3, and 4, respectively. Appendices H and I present the competitor method using m-out-of-n bootstrap, which was implemented in the simulation studies and CLEAR III analysis of the main paper. Appendix J presents the RICV5 outcome from Setting D of the simulation studies. Tables 1-4 present the average confidence interval widths for our method $\widehat{CI}_n$ and the m-out-of-n bootstrap, in each simulation setting. Figure 1 shows the distribution of RICV5 observed in the MISTIE II trial. Figures 2-15 show the coverage probabilities of our method $\widehat{CI}_n$ and the m-out-of-n bootstrap, for each simulation setting and at each sample size.

A Proof of Claim in Section 2.3

Claim. *Let P_0 be any joint distribution on (Y_C, Y_T). Let $\boldsymbol{\gamma}^*$ denote the pair of marginal distributions of Y_C and Y_T, under P_0. Suppose the restrictions $\mathcal{R}$ satisfy $P_0 \in \mathcal{R}$. Then, for any given $\psi \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$, there exists some joint distribution P on (Y_C, Y_T) that has marginals $\boldsymbol{\gamma}^*$, is in the set $\mathcal{R}$, and satisfies $P(Y_T > Y_C) = \psi$.*

Proof. Consider any $\psi \in [\psi_l^{\mathcal{R}}, \psi_u^{\mathcal{R}}]$. The lower and upper bounds $\psi_l^{\mathcal{R}}$ and $\psi_u^{\mathcal{R}}$ on the fraction who benefit are sharp (Huang et al., 2017). Thus, there exists a joint distribution P_l on (Y_C, Y_T) that has marginals γ^*, is in the set $\mathcal{R}$, and satisfies $P_l(Y_T > Y_C) = \psi_l^{\mathcal{R}}$. Analogously, there exists a joint distribution P_u on (Y_C, Y_T) that has marginals γ^*, is in the set $\mathcal{R}$, and satisfies $P_u(Y_T > Y_C) = \psi_u^{\mathcal{R}}$. If $\psi_l^{\mathcal{R}} = \psi_u^{\mathcal{R}}$, the claim directly follows. The rest of the proof is for the case that $\psi_l^{\mathcal{R}} \neq \psi_u^{\mathcal{R}}$.

If $\psi_l^{\mathcal{R}} \neq \psi_u^{\mathcal{R}}$, we have $\psi = (1-\beta)\psi_l^{\mathcal{R}} + \beta\psi_u^{\mathcal{R}}$, where $\beta = (\psi - \psi_l^{\mathcal{R}})/(\psi_u^{\mathcal{R}} - \psi_l^{\mathcal{R}})$. β is well-defined since $\psi_u^{\mathcal{R}} - \psi_l^{\mathcal{R}} \neq 0$. Also, we have $\beta \geq 0$ because $\psi - \psi_l^{\mathcal{R}} \geq 0$ and $\psi_u^{\mathcal{R}} - \psi_l^{\mathcal{R}} > 0$. We have $\beta \leq 1$ because $\psi_l^{\mathcal{R}} \leq \psi \leq \psi_u^{\mathcal{R}}$, which implies that $\psi - \psi_l^{\mathcal{R}} \leq \psi_u^{\mathcal{R}} - \psi_l^{\mathcal{R}}$.

Define P as follows: For any given i, j, let $P(Y_C = i, Y_T = j) = (1-\beta)P_l(Y_C = i, Y_T = j) + \beta P_u(Y_C = i, Y_T = j)$.

For any given i, j, we have $P_l(Y_C = i, Y_T = j) \geq 0$, $P_u(Y_C = i, Y_T = j) \geq 0$ since P_l and P_u are probability distributions. Since $0 \leq \beta \leq 1$, we also have $1 - \beta \geq 0$ and $\beta \geq 0$. It follows that $P(Y_C = i, Y_T = j) \geq 0$. Moreover, the sum $P(Y_C = i, Y_T = j)$ over all i, j pairs satisfies

$$\begin{aligned}
\sum_{i,j} P(Y_C = i, Y_T = j) &= \sum_{i,j} [(1-\beta)P_l(Y_C = i, Y_T = j) + \beta P_u(Y_C = i, Y_T = j)] \\
&= (1-\beta)\sum_{i,j} P_l(Y_C = i, Y_T = j) + \beta \sum_{i,j} P_u(Y_C = i, Y_T = j) \\
&= 1 - \beta + \beta = 1,
\end{aligned}$$

since P_l and P_u are probability distributions. By the results in this paragraph, P is a valid probability distribution.

Now we will show that P has marginals γ^*, is in the set $\mathcal{R}$, and satisfies $P(Y_T > Y_C) = \psi$. For any $i = 1, ..., L$, we have

$$\begin{aligned} P(Y_C = i) &= \sum_j P(Y_C = i, Y_T = j) \\ &= (1-\beta)\sum_j P_l(Y_C = i, Y_T = j) + \beta \sum_j P_u(Y_C = i, Y_T = j) \\ &= (1-\beta)P_l(Y_C = i) + \beta P_u(Y_C = i) \\ &= (1-\beta)\gamma^*_{0i} + \beta\gamma^*_{0i} = \gamma^*_{0i}. \end{aligned}$$

Analogously, for any $j = 1, ..., L$, we have

$$P(Y_T = j) = (1-\beta)P_l(Y_T = j) + \beta P_u(Y_T = j) = (1-\beta)\gamma^*_{1j} + \beta\gamma^*_{1j} = \gamma^*_{1j}.$$

For any (i, j) such that $g(i, j) = 0$, we have

$$P(Y_C = i, Y_T = j) = (1-\beta)P_l(Y_C = i, Y_T = j) + \beta P_u(Y_C = i, Y_T = j) = (1-\beta) \times 0 + \beta \times 0 = 0,$$

since $P_l, P_u \in \mathcal{R}$. Thus, we have $P \in \mathcal{R}$.

Also, we have

$$\begin{aligned} P(Y_T > Y_C) &= \sum_{j>i} P(Y_C = i, Y_T = j) \\ &= (1-\beta)\sum_{j>i} P_l(Y_C = i, Y_T = j) + \beta \sum_{j>i} P_u(Y_C = i, Y_T = j) \\ &= (1-\beta)\psi^{\mathcal{R}}_l + \beta\psi^{\mathcal{R}}_u = \psi. \end{aligned}$$

In conclusion, the joint distribution P has marginals γ^*, is in the set $\mathcal{R}$, and satisfies $P(Y_T > Y_C) = \psi$.

□

B Proof that minimizer of $P_0F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to $\boldsymbol{\gamma}^*$

Claim 1. *The minimizer of $P_0F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to $\boldsymbol{\gamma}^*$.*

Proof. We have that

$$\begin{aligned} P_0F(\boldsymbol{\gamma}, \boldsymbol{V}) &= P_0\left[\sum_{a=0}^{1}\sum_{j=1}^{L} 1(A = a)\{1(Y = j) - \gamma_{aj}\}^2\right] \\ &= \sum_{a=0}^{1}\sum_{j=1}^{L} \theta^a(1-\theta)^{1-a}(\gamma_{aj} - \gamma^*_{aj})^2 + \sum_{a=0}^{1}\sum_{j=1}^{L} \theta^a(1-\theta)^{1-a}\gamma^*_{aj}(1-\gamma^*_{aj}). \end{aligned}$$

Since θ and $\boldsymbol{\gamma}^*$ are constants, the rightmost double sum is also a constant. Denote it as c, i.e.,

$$c = \sum_{a=0}^{1}\sum_{j=1}^{L} \theta^a(1-\theta)^{1-a}\gamma^*_{aj}(1-\gamma^*_{aj}).$$

Then we have

$$P_0F(\boldsymbol{\gamma}, \boldsymbol{V}) = c + \sum_{a=0}^{1}\sum_{j=1}^{L} \theta^a(1-\theta)^{1-a}(\gamma_{aj} - \gamma^*_{aj})^2.$$

Since $0 < \theta < 1$, we have that $\theta^a(1-\theta)^{1-a}(\gamma_{aj} - \gamma^*_{aj})^2 \geq 0$ for all a, j. It follows that $P_0F(\boldsymbol{\gamma}, \boldsymbol{V}) \geq c$ for all $\gamma \in \mathbf{R}^{2L}$. If $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$, we have that $P_0F(\boldsymbol{\gamma}, \boldsymbol{V}) = c$. If $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}^*$, we have that $P_0F(\boldsymbol{\gamma}, \boldsymbol{V}) > c$. This occurs since, if $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}^*$, there is a pair (a, j) with $\gamma_{aj} \neq \gamma^*_{aj}$, which implies that $\theta^a(1-\theta)^{1-a}(\gamma_{aj} - \gamma^*_{aj})^2 > 0$.

It follows that the minimizer of $P_0F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \mathbf{R}^{2L}$ is unique and equal to $\boldsymbol{\gamma}^*$. Since $\Gamma \subseteq \mathbf{R}^{2L}$ and $\boldsymbol{\gamma}^* \in \Gamma$ (by Assumption 1), we conclude that the minimizer of $P_0F(\boldsymbol{\gamma}, \boldsymbol{V})$ over $\boldsymbol{\gamma} \in \Gamma$ is unique and equal to $\boldsymbol{\gamma}^*$. □

C Proof of Lemma 1 and of Uniqueness of Minimizer of Each Term in $T_{n,\psi}$

C.1 Proof of Lemma 1

Claim. *The test statistic $T_{n,\psi}$ is equivalent to*

$$n\left\{\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}} Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) - \min_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})\right\},$$

where $\widehat{\theta}_n = P_n A$ and $Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2\,\widehat{\theta}_n^{a}(1-\widehat{\theta}_n)^{1-a}$.

Proof. We have

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

Thus,

$$\begin{aligned} P_nF(\boldsymbol{\gamma},V) &= \frac{1}{n}\sum_{m=1}^{n}\sum_{a=0}^{1}\sum_{j=1}^{L}\left[1(A_m=a,Y_m=j)+\gamma_{aj}^2 1(A_m=a)-2\gamma_{aj}1(A_m=a,Y_m=j)\right] \\ &= \sum_{a=0}^{1}\sum_{j=1}^{L} P_n\left[1(A=a,Y=j)+\gamma_{aj}^2 1(A=a)-2\gamma_{aj}1(A=a,Y=j)\right] \\ &= \sum_{a=0}^{1}\sum_{j=1}^{L}\left[P_n1(A=a,Y=j)+\gamma_{aj}^2 P_n1(A=a)-2\gamma_{aj}P_n1(A=a,Y=j)\right]. \end{aligned}$$

Let $\widehat{\boldsymbol{\gamma}}$ denote the empirical marginal distributions under control and treatment. In other words, $\widehat{\boldsymbol{\gamma}}$ is the vector $(\widehat{\gamma}_{01},..,\widehat{\gamma}_{0L},\widehat{\gamma}_{11},..,\widehat{\gamma}_{1L})$, where $\widehat{\gamma}_{aj} = \frac{P_n1(A=a,Y=j)}{P_n1(A=a)}$. We have

$$\begin{aligned} P_nF(\boldsymbol{\gamma},V) &= \sum_{a=0}^{1}\sum_{j=1}^{L}\left[(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n1(A=a)+\widehat{\gamma}_{aj}(1-\widehat{\gamma}_{aj})P_n1(A=a)\right] \\ &= \sum_{a=0}^{1}\sum_{j=1}^{L}\left[(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n1(A=a)\right]+\sum_{a=0}^{1}\sum_{j=1}^{L}\left[\widehat{\gamma}_{aj}(1-\widehat{\gamma}_{aj})P_n1(A=a)\right]. \end{aligned}$$

The second term $\sum_{a=0}^{1}\sum_{j=1}^{L}\left[\widehat{\gamma}_{aj}(1-\widehat{\gamma}_{aj})P_n1(A=a)\right]$ only depends on the data and not on $\boldsymbol{\gamma}$. Below, we denote it by $f(data)$.

The test statistic is defined as

$$T_{n,\psi} = n\left\{\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}} P_nF(\boldsymbol{\gamma},V) - \min_{\boldsymbol{\gamma}\in\Gamma} P_nF(\boldsymbol{\gamma},V)\right\}.$$

We have

$$\begin{aligned} \min_{\boldsymbol{\gamma}\in\Gamma^{\psi}} P_nF(\boldsymbol{\gamma},V) &= \min_{\boldsymbol{\gamma}\in\Gamma^{\psi}}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}[(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n1(A=a)]+f(data)\right\} \\ &= \min_{\boldsymbol{\gamma}\in\Gamma^{\psi}}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n1(A=a)\right\}+f(data). \end{aligned}$$

We can move $f(data)$ outside of the braces since it does not depend on $\boldsymbol{\gamma}$. Analogously,

$$\min_{\boldsymbol{\gamma}\in\Gamma} P_nF(\boldsymbol{\gamma},V) = \min_{\boldsymbol{\gamma}\in\Gamma}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n1(A=a)\right\}+f(data).$$

In conclusion, the test statistic $T_{n,\psi}$ simplifies to

$$\begin{aligned} & T_{n,\psi} \\ = \; & n\left[\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n 1(A=a)\right\} - \min_{\boldsymbol{\gamma}\in\Gamma}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 P_n 1(A=a)\right\}\right] \\ = \; & n\left[\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 (\widehat{\theta}_n)^a(1-\widehat{\theta}_n)^{1-a}\right\} - \min_{\boldsymbol{\gamma}\in\Gamma}\left\{\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2 (\widehat{\theta}_n)^a(1-\widehat{\theta}_n)^{1-a}\right\}\right]. \end{aligned}$$

□

C.2 Existence and Uniqueness of Minimizer of Each Term in $T_{n,\psi}$

Consider the terms $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma},\boldsymbol{V})$ and $\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}} P_n F(\boldsymbol{\gamma},\boldsymbol{V})$ in the definition of the statistic $T_{n,\psi}$ in (3) of the main paper; we refer to these as the first and second terms, respectively. It was claimed in Section 3.3 of the main paper that each term has a unique minimizer, under the assumptions in the first paragraph of that section (that Γ^{ψ} is non-empty and at least one participant is assigned to each arm). We now prove this claim.

By the arguments in Section C.1 of the Supplementary Material, it follows that the set of minimizers of the first term equals the set of minimizers of $\min_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})$, where $\widehat{\theta}_n = P_n A$, $\hat{c}_a = \widehat{\theta}_n^a(1-\widehat{\theta}_n)^{1-a} > 0$ for each $a\in\{0,1\}$, and $Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}}) = \sum_{a=0}^{1}\hat{c}_a\sum_{j=1}^{L}(\gamma_{aj}-\widehat{\gamma}_{aj})^2$.

Consider the change of variables $\underline{\gamma}_{aj} = \gamma_{aj}\hat{c}_a^{1/2}$ and $\underline{\gamma}^*_{aj} = \gamma^*_{aj}\hat{c}_a^{1/2}$. It follows that the vector $\underline{\boldsymbol{\gamma}} = \underline{\boldsymbol{B}}\boldsymbol{\gamma}$ and $\underline{\widehat{\boldsymbol{\gamma}}} = \underline{\boldsymbol{B}}\widehat{\boldsymbol{\gamma}}$ where $\underline{\boldsymbol{B}}$ denotes the $2L\times 2L$ diagonal matrix with first L diagonal elements equal to $\hat{c}_0^{1/2} > 0$ and last L diagonal elements equal to $\hat{c}_1^{1/2} > 0$. By the above arguments, there exists a unique minimizer of $\min_{\boldsymbol{\gamma}\in\Gamma} Discrep_{\widehat{\theta}_n}(\boldsymbol{\gamma},\widehat{\boldsymbol{\gamma}})$ if and only if there exists a unique minimizer of $\min_{\underline{\boldsymbol{\gamma}}\in(\underline{\boldsymbol{B}}\Gamma)}\sum_{a=0}^{1}\sum_{j=1}^{L}\left(\underline{\gamma}_{aj}-\underline{\widehat{\gamma}}_{aj}\right)^2$. The latter minimization problem has a unique minimum equal to the Euclidean projection of $\underline{\widehat{\boldsymbol{\gamma}}}$ on $\underline{\boldsymbol{B}}\Gamma$ (which is a closed, convex set since Γ has these properties). Here we used that the Euclidean projection of any point on a closed, convex set in Euclidean space both exists and is unique. This proves that there is a unique minimizer of the first term $\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma},\boldsymbol{V})$. The proof that there is a unique mimizer of the second term $\min_{\boldsymbol{\gamma}\in\Gamma^{\psi}} P_n F(\boldsymbol{\gamma},\boldsymbol{V})$ is analogous, replacing Γ by Γ^{ψ} throughout.

D Proof of Theorem 1

We use the general argument in Section 5.1.3 of Shapiro et al. (2014), except tailored to our specific problem. The proof here is self-contained.

The null hypothesis $\boldsymbol{\gamma}^*\in\Gamma^{\psi}$ implies that the minimizer $\boldsymbol{\gamma}^*$ of $\min_{\boldsymbol{\gamma}\in\Gamma} P_0 F(\boldsymbol{\gamma},\boldsymbol{V})$ is unique and satisfies $\gamma^*_{aj} = P_0(Y=j|A=a)$ for each $a\in\{0,1\}, j\in\{1,\ldots,L\}$. This implies $\nabla P_0 F(\boldsymbol{\gamma}^*,\boldsymbol{V}) = 0$, where the gradient is with respect to $\boldsymbol{\gamma}^*$.

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

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

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

$$n\left\{\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma},\boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*,\boldsymbol{V})\right\} = \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*,\Gamma)}(\boldsymbol{h}^t\boldsymbol{Z}_n + \boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2), \quad (1)$$

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

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

To show (1), we have

$$
\begin{aligned}
& n\left\{\min_{\boldsymbol{\gamma}\in\Gamma} P_n F(\boldsymbol{\gamma},\boldsymbol{V}) - P_n F(\boldsymbol{\gamma}^*,\boldsymbol{V})\right\} && (3)\\
= \;& n \min_{\boldsymbol{\gamma}\in\Gamma} P_n \left\{F(\boldsymbol{\gamma},\boldsymbol{V}) - F(\boldsymbol{\gamma}^*,\boldsymbol{V})\right\} \\
= \;& n \min_{\boldsymbol{\gamma}\in\Gamma} \sum_{a=0}^{1}\sum_{j=1}^{L} P_n 1(A=a)\left[\left\{1(Y=j)-\gamma_{aj}\right\}^2 - \left\{1(Y=j)-\gamma^*_{aj}\right\}^2\right] \\
= \;& n \min_{\boldsymbol{\gamma}\in\Gamma} \sum_{a=0}^{1}\sum_{j=1}^{L} P_n 1(A=a)\left[-2\left\{1(Y=j)-\gamma^*_{aj}\right\}(\gamma_{aj}-\gamma^*_{aj}) + (\gamma_{aj}-\gamma^*_{aj})^2\right] \\
= \;& \min_{\boldsymbol{\gamma}\in\Gamma}\left[n^{1/2}\sum_{a=0}^{1}\sum_{j=1}^{L} Z_{aj,n}(\gamma_{aj}-\gamma^*_{aj}) + \sum_{a=0}^{1} P_n 1(A=a)\sum_{j=1}^{L}\left\{n^{1/2}(\gamma_{aj}-\gamma^*_{aj})\right\}^2\right] \\
= \;& \min_{\boldsymbol{\gamma}\in\Gamma}\left[n^{1/2}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)^t \boldsymbol{Z}_n + \sum_{a=0}^{1} P_n 1(A=a)\sum_{j=1}^{L}\left\{n^{1/2}(\gamma_{aj}-\gamma^*_{aj})\right\}^2\right] && (4)\\
= \;& \min_{h\in C_n(\boldsymbol{\gamma}^*,\Gamma)} h^t \boldsymbol{Z}_n + h^t \boldsymbol{D}_n h/2, && (5)
\end{aligned}
$$

which proves (1). The proof of (2) is analogous, except replacing Γ by Γ^ψ

Taking the difference between the left sides of (2) and (1), we have

$$T_{n,\psi} = \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*,\Gamma^\psi)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) - \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*,\Gamma)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2).$$

Now we will prove that under the null hypothesis $H_0(\psi)$, we have

$$
\begin{aligned}
& \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*,\Gamma^\psi)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) - \min_{\boldsymbol{h}\in C_n(\boldsymbol{\gamma}^*,\Gamma)} (\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) && (6)\\
\to \;& \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma^\psi)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) - \min_{\boldsymbol{h}\in C(\boldsymbol{\gamma}^*,\Gamma)} (\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2), && (7)
\end{aligned}
$$

where the above convergence indicated by $\to$ is in distribution as $n \to \infty$. The expression (6) equals $T_{n,\psi}$, and (7) equals T_ψ. Therefore, showing the above result implies the convergence of $T_{n,\psi}$ to the null distribution T_ψ under $H_0(\psi)$.

Throughout, we assume the null hypothesis $H_0(\psi)$. For any column vectors $\boldsymbol{h}, \boldsymbol{A} \in \mathbb{R}^{2L}$ and positive definite matrix $\boldsymbol{B} \in \mathbb{R}^{2L\times 2L}$, define the function $f(\boldsymbol{h},\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{B}\boldsymbol{h}/2)$; for any set $C \subseteq \mathbb{R}^{2L}$, define $g(C,\boldsymbol{A},\boldsymbol{B}) = \min_{\boldsymbol{h}\in C}(\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{B}\boldsymbol{h}/2)$. Let $||\boldsymbol{A}||$ denote the Euclidean norm of $\boldsymbol{A}$. Define the ball of radius M by $B_M = \{\boldsymbol{h} \in \mathbb{R}^{2L} : ||\boldsymbol{h}|| \le M\}$.

By Skorokhod's representation theorem (Jakubowski, 1998), there exist $\boldsymbol{Z}_n, \boldsymbol{D}_n, \boldsymbol{Z}$ such that $(\boldsymbol{Z}_n, \boldsymbol{D}_n)$ converges almost surely to $(\boldsymbol{Z}, \boldsymbol{I})$ (under the sup-norm), where for each n these have the same distribution as the corresponding random vectors/matrices defined in the main text. We use this representation throughout this proof. We suppress the dependence of $C_n(\boldsymbol{\gamma}^*,\Gamma)$ and $C(\boldsymbol{\gamma}^*,\Gamma)$ on $\boldsymbol{\gamma}^*,\Gamma$; all results below hold as well replacing Γ by Γ^ψ under the null hypothesis $H_0(\psi)$. The key features of C_n and C used below are that they are each closed, convex, and contain $\boldsymbol{0}$.

We will prove that $|g(C_n,\boldsymbol{Z}_n,\boldsymbol{D}_n) - g(C,\boldsymbol{Z},I)|$ converges to 0 in probability. By the triangle inequality:

$$|g(C_n,\boldsymbol{Z}_n,\boldsymbol{D}_n) - g(C,\boldsymbol{Z},\boldsymbol{I})| \le |g(C_n,\boldsymbol{Z}_n,\boldsymbol{D}_n) - g(C_n,\boldsymbol{Z},\boldsymbol{I})| + |g(C_n,\boldsymbol{Z},\boldsymbol{I}) - g(C,\boldsymbol{Z},\boldsymbol{I})|. \quad (8)$$

In a series of steps involving the lemmas below, we prove that each term on the right side of the above inequality converges to 0 in probability.

Lemma 1. *For all $\epsilon > 0$, there exists an $M > 0$ such that the event $E = \{||\mathbf{Z}|| \le M\}$ has probability at least $1-\epsilon$.*

Proof. Consider any $\epsilon > 0$. We have

$$\begin{aligned} E\left[||\boldsymbol{Z}||^2\right] &= E\left[Z_{01}^2 + \ldots + Z_{0L}^2 + Z_{11}^2 + \ldots + Z_{1L}^2\right] \\ &= E[Z_{01}^2] + \ldots + E[Z_{0L}^2] + E[Z_{11}^2] + \ldots + E[Z_{1L}^2] \\ &= Var(Z_{01}) + \ldots + Var(Z_{0L}) + Var(Z_{11}) + \ldots + Var(Z_{1L}) \\ &= 4\sum_{a=0}^{1}\sum_{j=1}^{L}\left[(1-\gamma_{aj}^*)P_0(A=a, Y=j)\right]. \end{aligned}$$

Notice that $E\left[||\boldsymbol{Z}||^2\right]$ is finite because it is a sum of a finite number of terms, which are themselves finite (since $0 \leq \gamma_{aj}^* \leq 1$ and $0 \leq P_0(A=a, Y=j) \leq 1$). Choose M to be any positive number large enough that $E\left[||\boldsymbol{Z}||^2\right]/M^2 \leq \epsilon$. By Chebyshev's Inequality, we have

$$P(||\boldsymbol{Z}|| > M) \leq \frac{E\left[||\boldsymbol{Z}||^2\right]}{M^2} \leq \epsilon.$$

This implies that

$$P(||\boldsymbol{Z}|| \leq M) \geq 1 - \epsilon.$$

□

Lemma 2. *For any $M > 0$, there exists an $n_0 > 0$ such that $C_n \cap B_M = C \cap B_M$ for all $n > n_0$.*

Proof. Since Γ is a bounded polyhedron, it is the convex hull of a finite number $t > 0$ of extreme points $\{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_t\}$. Thus, we have that C_1 is the convex hull of the extreme points $\{\boldsymbol{e}_1 - \boldsymbol{\gamma}^*, \boldsymbol{e}_2 - \boldsymbol{\gamma}^*, ..., \boldsymbol{e}_t - \boldsymbol{\gamma}^*\}$ and so is also a bounded polyhedron. If C_1 is the set containing the single point $\mathbf{0}$, then so is C and each $C_n : n \geq 1$, so the lemma holds trivially. We assume $C_1 \neq \{\mathbf{0}\}$ for the remainder of the proof. Let $\tilde{E}$ denote the union of all facets of C_1 that do not contain $\mathbf{0}$. Then $\tilde{E}$ is closed, bounded, does not contain $\mathbf{0}$, and $\min_{\boldsymbol{h}\in\tilde{E}} \|\boldsymbol{h}\| > 0$. Denote $d = \min_{\boldsymbol{h}\in\tilde{E}} \|\boldsymbol{h}\| > 0$.

Let $n_0 = \lceil M/d \rceil + 1$. Consider any $\boldsymbol{h} \in C \cap B_M$. We will show that $\boldsymbol{h} \in C_{n_0} \cap B_M$. It suffices to show $\boldsymbol{h} \in C_{n_0}$. Since $\boldsymbol{h} \in C \cap B_M$, we have that $\boldsymbol{h} = r(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)$ for some $r > 0$ and $\boldsymbol{\gamma} \in \Gamma$, and $\|\boldsymbol{h}\| = r\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\| \leq M$. If $\boldsymbol{h} = \mathbf{0}$, then $\boldsymbol{h} \in C_{n_0} \cap B_M$ since $\boldsymbol{\gamma}^* \in \Gamma$. Next, consider the case of $\boldsymbol{h} \neq \mathbf{0}$. Define $\xi = \max\{\xi' \geq 0 : \xi'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_1\}$; the maximum is achieved and $\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_1$ since C_1 is closed, convex, and contains $\mathbf{0}$. We have $\xi \geq 1$ since $(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_1$. By construction, we have $\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in \tilde{E}$ and so has norm at least d; it follows that $\|n_0\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\| > (M/d)d = M$. It follows from $\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_1$ that $n_0\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_{n_0}$. Since C_{n_0} is convex and contains both $\mathbf{0}$ and $n_0\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)$, we have $y(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_{n_0}$ for every $y \geq 0 : \|y(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\| \leq \|n_0\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\|$. The value r satisfies this property since as shown above,

$$r\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\| \leq M < \|n_0\xi(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\|.$$

Therefore, $\boldsymbol{h} = r(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) \in C_{n_0}$, as desired. We have shown $C \cap B_M \subseteq C_{n_0} \cap B_M$.

We have that $C_{n_0} \subseteq C$ and so $C \cap B_M \supseteq C_{n_0} \cap B_M$. Combining this with the above result, we have $C \cap B_M = C_{n_0} \cap B_M$ as desired. □

Lemma 3. *i. For any $\boldsymbol{A} \in \mathbb{R}^{2L}$, we have* $\arg\min_{\boldsymbol{h}\in\mathbb{R}^{2L}}(\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{h}/2) = -\boldsymbol{A}$.

ii. For any $\boldsymbol{A} \in \mathbb{R}^{2L}$ and closed, convex $C' \subseteq \mathbb{R}^{2L}$, we have $\arg\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{h}/2)$ *is the Euclidean projection of $-\boldsymbol{A}$ on C'.*

iii. For any $\boldsymbol{A} \in \mathbb{R}^{2L}$, and closed, convex $C' \subseteq \mathbb{R}^{2L}$ with $\mathbf{0} \in C'$, we have

$$||\arg\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{h}/2)|| \leq ||\boldsymbol{A}||.$$

An important consequence is that if $||\boldsymbol{Z}|| \leq M$ then the minimizer $\boldsymbol{h}^*$ of $\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$ is in the ball B_M. In particular, this applies to $C' = C$ and $C' = C_n$ for all n.

Proof. i. The gradient of $\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{h}/2$ is $\boldsymbol{h} + \boldsymbol{A}$ and the Hessian is the identity matrix. Therefore, the unique mimimum occurs at $\boldsymbol{h} = -\boldsymbol{A}$.

ii. Since C' is closed and convex, the Euclidean projection of $-\boldsymbol{A}$ on C' exists and is unique; it is the solution to $\min_{\boldsymbol{h}\in C'} \|\boldsymbol{h} - (-\boldsymbol{A})\|^2$. Since the optimization problem $\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{A} + \boldsymbol{h}^t\boldsymbol{h}/2)$ is equivalent to the optimization problem $\min_{\boldsymbol{h}\in C'}\left\{\|\boldsymbol{h} - (-\boldsymbol{A})\|^2/2 - \boldsymbol{A}^t\boldsymbol{A}/2\right\}$, and the solution

to the latter is the Euclidean projection of $-\boldsymbol{A}$ on C', we have that the Euclidean projection of $-\boldsymbol{A}$ on C' is the unique minimizer of the former problem.

iii. Denote by $P_{C'}(\boldsymbol{h})$ the Euclidean projection of $\boldsymbol{h}$ onto the convex set C'. We have that by (ii),

$$\arg\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{A}+\boldsymbol{h}^t\boldsymbol{h}/2)=P_{C'}(-\boldsymbol{A}).$$

Then, we have that

$$\|\arg\min_{\boldsymbol{h}\in C'}(\boldsymbol{h}^t\boldsymbol{A}+\boldsymbol{h}^t\boldsymbol{h}/2)\|=\|P_{C'}(-\boldsymbol{A})-P_{C'}(\boldsymbol{0})\|\leq\|-\boldsymbol{A}-\boldsymbol{0}\|=\|\boldsymbol{A}\|,$$

where the inequality holds by the nonexpansiveness of convex projection (Proposition 1.1.9 by Bertsekas (2009)), and our claim holds as desired. □

In the lemma below, we define N_0, N_1, N_z, N_d which are random (i.e., functions of the underlying probability space).

Lemma 4. *i. On the event $||\boldsymbol{Z}||\leq M$, with probability 1 there exists an $N_0>0$ such that $n>N_0$ implies the following results: $||\boldsymbol{Z}_n||\leq 2M$; $\boldsymbol{D}_n$ is a diagonal matrix with each diagonal entry at least 0.9; and,*

$$||\arg\min_{\boldsymbol{h}\in C_n}(\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2)||\leq 5M,$$

i.e., the minimizer of $\min_{\boldsymbol{h}\in C_n}(\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2)$ is in the ball B_{5M}.

ii.

$$\sup_{\boldsymbol{h}\in B_{5M}}|f(\boldsymbol{h},\boldsymbol{Z}_n,\boldsymbol{D}_n)-f(\boldsymbol{h},\boldsymbol{Z},\boldsymbol{I})|\leq 5M||\boldsymbol{Z}_n-\boldsymbol{Z}||+(5M)^2|2P_n1(A=0)-1|.$$

iii. If $||\boldsymbol{Z}||\leq M$, then with probability 1 the first term on the right of (8) satisfies

$$|g(C_n,\boldsymbol{Z}_n,\boldsymbol{D}_n)-g(C_n,\boldsymbol{Z},\boldsymbol{I})|\to 0$$

almost surely as $n\to\infty$.

iv. If $||\boldsymbol{Z}||\leq M$, then with probability 1 the second term on the right of (8) satisfies

$$|g(C_n,\boldsymbol{Z},\boldsymbol{I})-g(C,\boldsymbol{Z},\boldsymbol{I})|\to 0$$

almost surely as $n\to\infty$.

v. $|g(C_n,\boldsymbol{Z}_n,\boldsymbol{D}_n)-g(C,\boldsymbol{Z},\boldsymbol{I})|$ converges to 0 in probability.

vi. (6) converges in distribution to (7).

Proof. **(i)** Assume that $||\boldsymbol{Z}||\leq M$. By our above use of the Skorokhod's representation theorem, we have $(\boldsymbol{Z}_n,\boldsymbol{D}_n)$ converges almost surely to $(\boldsymbol{Z},\boldsymbol{I})$. Let E' denote the zero probability event that this convergence does not occur; we work on the complement of this event throughout the proof of this lemma, so that all claims hold with probability 1. By the Continuous Mapping Theorem, $||\boldsymbol{Z}_n||$ converges almost surely to $||\boldsymbol{Z}||$. Since $||\boldsymbol{Z}||\leq M$, there exists an N_z such that for $n>N_z$ we have that $||\boldsymbol{Z}_n||\leq 2M$.

The matrix $\boldsymbol{D}_n$ is defined as a $2L\times 2L$ diagonal matrix with the first L diagonal entries equal to $2P_n1(A=0)$ and the last L diagonal entries equal to $2P_n1(A=1)$. Since $(\boldsymbol{Z}_n,\boldsymbol{D}_n)$ converges almost surely to $(\boldsymbol{Z},\boldsymbol{I})$, it follows that $2P_n1(A=0)$ converges almost surely to 1 and $2P_n1(A=1)$ converges almost surely to 1. Thus, there exists an $N_d>0$ such that $|2P_n1(A=0)-1|\leq 0.1$ and $|2P_n1(A=1)-1|\leq 0.1$ for $n>N_d$. It follows that for $n>N_d$, $\boldsymbol{D}_n$ is a diagonal matrix with each diagonal entry at least 0.9.

Let $N_0=\max\{N_z,N_d\}$. For any $n>N_0$, we have that

$$\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2\geq\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2-|\boldsymbol{h}^t\boldsymbol{Z}_n|\geq 0.9||\boldsymbol{h}||^2/2-||\boldsymbol{h}||\cdot||\boldsymbol{Z}_n||=||\boldsymbol{h}||\left\{0.9||\boldsymbol{h}||/2-||\boldsymbol{Z}_n||\right\}.$$

If $||\boldsymbol{h}||>5M$ and $n>N_0$, then we have $||\boldsymbol{Z}_n||\leq 2M$ and

$$0.9||\boldsymbol{h}||/2-||\boldsymbol{Z}_n||>0.9\cdot 5M/2-||\boldsymbol{Z}_n||\geq 0.9\cdot 5M/2-2M=M/4>0.$$

Thus, if $||\boldsymbol{h}||>5M$ and $n>N_0$, we have that $\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2>0$. At $\boldsymbol{h}=\boldsymbol{0}$, we have that $(\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2)=0$. This implies that, for $n>N_0$, the minimizer $\boldsymbol{h}^*$ of $\boldsymbol{h}^t\boldsymbol{Z}_n+\boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2$

over $\boldsymbol{h} \in C_n$ (where C_n contains $\mathbf{0}$) cannot be such that $||\boldsymbol{h}^*|| > 5M$. We conclude that $||\arg\min_{\boldsymbol{h}\in C_n}(\boldsymbol{h}^t\boldsymbol{Z}_n + \boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2)|| \leq 5M$.

(ii) Consider any $\boldsymbol{h} \in B_{5M}$. Denote the elements of $\boldsymbol{h}$ as $(h_{01}, \ldots, h_{0L}, h_{11}, \ldots, h_{1L})$. Then we have

$$\begin{aligned}
& |f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I})| \\
= \; & |\boldsymbol{h}^t\boldsymbol{Z}_n + \boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2 - (\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)| \\
= \; & |\boldsymbol{h}^t(\boldsymbol{Z}_n - \boldsymbol{Z}) + \boldsymbol{h}^t(\boldsymbol{D}_n - \boldsymbol{I})\boldsymbol{h}/2| \\
\leq \; & |\boldsymbol{h}^t(\boldsymbol{Z}_n - \boldsymbol{Z})| + |\boldsymbol{h}^t(\boldsymbol{D}_n - \boldsymbol{I})\boldsymbol{h}/2| \\
\leq \; & ||\boldsymbol{h}||||\boldsymbol{Z}_n - \boldsymbol{Z}|| + \left([2P_n 1(A=0) - 1] \sum_{j=1}^{L} h_{0j}^2 + [2P_n 1(A=1) - 1] \sum_{j=1}^{L} h_{1j}^2 \right) /2 \\
\leq \; & 5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + |P_n 1(A=0) - 1/2| \cdot ||\boldsymbol{h}||^2 \\
\leq \; & 5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + |P_n 1(A=0) - 1/2|(5M)^2 \\
\leq \; & 5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + (5M)^2|2P_n 1(A=0) - 1|.
\end{aligned}$$

Thus, for $\boldsymbol{h} \in B_{5M}$, the value $5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + (5M)^2|2P_n 1(A=0) - 1|$ is an upper bound on $|f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I})|$. We conclude that

$$\sup_{\boldsymbol{h}\in B_{5M}} |f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I})| \leq 5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + (5M)^2|2P_n 1(A=0) - 1|.$$

(iii) Consider any $\delta > 0$. We will show that if $||\boldsymbol{Z}|| \leq M$, then with probability 1 there exists an $N_1 > 0$ such that $n > N_1$ implies

$$|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C_n, \boldsymbol{Z}, \boldsymbol{I})| < \delta. \tag{9}$$

Assume that $||\boldsymbol{Z}|| \leq M$ holds. Then by Lemma 4(i), there exists an $N_0 > 0$ such that the minimizer of $\min_{\boldsymbol{h}\in C_n}(\boldsymbol{h}^t\boldsymbol{Z}_n + \boldsymbol{h}^t\boldsymbol{D}_n\boldsymbol{h}/2)$ is in the ball B_{5M}, if $n > N_0$. For any $n > N_0$, we have

$$\begin{aligned}
|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C_n, \boldsymbol{Z}, \boldsymbol{I})| &= \left| \min_{\boldsymbol{h}\in C_n} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - \min_{\boldsymbol{h}\in C_n} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) \right| \\
&= \left| \min_{\boldsymbol{h}\in C_n} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) \right| \\
&= \left| \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) \right| \\
&\leq \sup_{\boldsymbol{h}\in C_n\cap B_{5M}} |f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I})| \\
&\leq 5M||\boldsymbol{Z}_n - \boldsymbol{Z}|| + (5M)^2|2P_n 1(A=0) - 1|.
\end{aligned}$$

The first line follows from the definition of g. The second line follows from Lemma 3(iii), according to which the minimizer $\boldsymbol{h}^*$ of $\min_{\boldsymbol{h}\in C_n}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$ must be in the ball B_M and hence also in the ball B_{5M}. The third line follows from Lemma 4(i). The proof of the fourth line is as follows. First, consider the case where $\min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) \leq \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n)$. Let $\boldsymbol{h}^*$ denote the minimizer of $\min_{\boldsymbol{h}\in C_n\cap B_{5M}}(\boldsymbol{h}^t\boldsymbol{Z} + \boldsymbol{h}^t\boldsymbol{h}/2)$. Then

$$\begin{aligned}
\left| \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) \right| &= \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}^*, \boldsymbol{Z}, \boldsymbol{I}) \\
&\leq f(\boldsymbol{h}^*, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}^*, \boldsymbol{Z}, \boldsymbol{I}) \\
&\leq \sup_{\boldsymbol{h}\in C_n\cap B_{5M}} |f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n) - f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I})|.
\end{aligned}$$

The case of $\min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}, \boldsymbol{I}) > \min_{\boldsymbol{h}\in C_n\cap B_{5M}} f(\boldsymbol{h}, \boldsymbol{Z}_n, \boldsymbol{D}_n)$ is handled analogously. The fifth line follows from Lemma 4(ii).

Since $(\boldsymbol{Z}_n, \boldsymbol{D}_n)$ converges almost surely to $(\boldsymbol{Z}, \boldsymbol{I})$, there exists an $N_1 > 0$ such that for $n > N_1$ we have that $||\boldsymbol{Z}_n - \boldsymbol{Z}|| < \delta/(10M)$ and $|2P_n 1(A=0) - 1| < \delta/(2(5M)^2)$. Combining this with the results in the above displays, we have shown for any $n > \max\{N_0, N_1\}$, we have that (9) holds. It follows that on the event $||\boldsymbol{Z}|| \leq M$, we have $|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C_n, \boldsymbol{Z}, \boldsymbol{I})| \to 0$ almost surely as $n \to \infty$.

(iv) Assume that $||\boldsymbol{Z}|| \leq M$. By Lemma 2, there exists an $n_0 > 0$ such that $C_n \cap B_M = C \cap B_M$ for all $n > n_0$. By Lemma 3(iii), the minimizer $\boldsymbol{h}^*$ of $\min_{\boldsymbol{h} \in C}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)$ is in the ball B_M. Also, for all n, the minimizer $\boldsymbol{h}_n^*$ of $\min_{\boldsymbol{h} \in C_n}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)$ is in the ball B_M. Thus, we have that for $n > n_0$,

$$\min_{\boldsymbol{h} \in C}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) = \min_{\boldsymbol{h} \in C \cap B_M}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) = \min_{\boldsymbol{h} \in C_n \cap B_M}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) = \min_{\boldsymbol{h} \in C_n}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2).$$

It follows that $|g(C_n, \boldsymbol{Z}, \boldsymbol{I}) - g(C, \boldsymbol{Z}, \boldsymbol{I})| = 0$ for $n > n_0$. We conclude that $|g(C_n, \boldsymbol{Z}, \boldsymbol{I}) - g(C, \boldsymbol{Z}, \boldsymbol{I})| \to 0$ almost surely as $n \to \infty$.

(v) Consider any $\epsilon > 0$ and $\delta > 0$. By Lemma 1, there exists an $M > 0$ such that the event $E = \{||\boldsymbol{Z}|| \leq M\}$ has probability at least $1 - \delta/2$. By (iii), (iv), and (8), we have $1(E)|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| \to 0$ almost surely as $n \to \infty$, where $1(E)$ is the indicator of the event E. Thus, there exists a natural number n_1 such that for $n \geq n_1$, we have

$$P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E) < \delta/2.$$

It follows that for $n \geq n_1$, we have

$$\begin{aligned}
P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon) &= P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E) P(E) \\
&\quad + P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E^C) P(E^C) \\
&\leq P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E) \\
&\quad + P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E^C) \delta/2 \\
&\leq P(|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})| > \epsilon | E) + \delta/2 \\
&< \delta/2 + \delta/2 \\
&= \delta.
\end{aligned}$$

Since ϵ and δ were arbitrary, we conclude that $|g(C_n, \boldsymbol{Z}_n, \boldsymbol{D}_n) - g(C, \boldsymbol{Z}, \boldsymbol{I})|$ converges to 0 in probability.

(vi) By (v) and the definition of g, we have

$$\left| \min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) - \min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) \right| \to 0$$

in probability. It follows that $\min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2)$ converges to $\min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)$ in probability. Under the null hypothesis $H_0(\psi)$, Lemmas 1, 2, 3, and 4(i-vi) hold if Γ is replaced by Γ^ψ. It follows that $\min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*, \Gamma^\psi)}(\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2)$ converges in probability to $\min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*, \Gamma^\psi)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)$. By Theorem 2.7 in van der Vaart (1998) and the Continuous Mapping Theorem, we have that

$$\begin{aligned}
&\min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*, \Gamma^\psi)}(\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) - \min_{\boldsymbol{h} \in C_n(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z}_n + \boldsymbol{h}^t \boldsymbol{D}_n \boldsymbol{h}/2) \\
\to &\min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*, \Gamma^\psi)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2) - \min_{\boldsymbol{h} \in C(\boldsymbol{\gamma}^*, \Gamma)}(\boldsymbol{h}^t \boldsymbol{Z} + \boldsymbol{h}^t \boldsymbol{h}/2)
\end{aligned}$$

in probability. This implies (6) converges in distribution to (7), completing the proof. □

E Proof of Theorem 2

Claim 2. *Under the alternative hypothesis $H_a(\psi)$, for any $M \in \mathbb{R}$, we have $P(T_{n,\psi} > M) \to 1$ as $n \to \infty$.*

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

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

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

We have as $n \to \infty$

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

by the Weak Law of Large Numbers, Slutsky's lemma, and Theorem 2.7(vi) in van der Vaart (1998).

For any positive integer n, we have

$$\begin{aligned} P(T_{n,\psi} > M) &= P\left(n\left\{\min_{\boldsymbol{\gamma}\in\Gamma^\psi} \text{Discrep}_{\widehat{\theta}_n}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \min_{\boldsymbol{\gamma}\in\Gamma} \text{Discrep}_{\widehat{\theta}_n}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}})\right\} > M\right) \\ &= P\left(\min_{\boldsymbol{\gamma}\in\Gamma^\psi} \text{Discrep}_{\widehat{\theta}_n}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) - \min_{\boldsymbol{\gamma}\in\Gamma} \text{Discrep}_{\widehat{\theta}}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) > \frac{M}{n}\right). \end{aligned} \tag{10}$$

For conciseness, let

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

Thus, we have $P(T_{n,\psi} > M) = P(d_{n,\psi} > M/n)$. By Lemmas 5 and 6 (see end of this section) and the Continuous Mapping Theorem, we have $\min_{\boldsymbol{\gamma}\in\Gamma^\psi} \text{Discrep}_{\widehat{\theta}_n}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \min_{\boldsymbol{\gamma}\in\Gamma^\psi} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ and $\min_{\boldsymbol{\gamma}\in\Gamma} \text{Discrep}_{\widehat{\theta}_n}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) \xrightarrow{p} \min_{\boldsymbol{\gamma}\in\Gamma} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$. Let $c = \min_{\boldsymbol{\gamma}\in\Gamma^\psi} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ and $b = \min_{\boldsymbol{\gamma}\in\Gamma} \text{Discrep}_{\theta}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$. 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, as proved in Lemma 5 at the end of this section. By Slutsky's lemma, the random variable $d_{n,\psi}$ converges in probability to $c > 0$. Let $\epsilon = c/2$. Since the sequence M/n converges to 0 as $n \to \infty$, we have for sufficiently large n,

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

Because $d_{n,\psi}$ converges in probability to c, the probability on the left converges to 1. Since the right side of the above display is at most 1, we have $P(d_{n,\psi} > M/n) = P(T_{n,\psi} > M)$ converges to 1. □

Lemma 5. *The sets Γ and Γ^ψ, for any $\psi \in [0, 1]$, are compact.*

Proof. Let ψ be an arbitrary number in $[0, 1]$. We prove that the set Γ^ψ is compact. The proof for Γ is analogous. Define

$$\Pi^\psi = \left\{ \pi = (\pi_{1,1}, \ldots, \pi_{1,L}, \pi_{2,1}, \ldots, \pi_{2,L}, \ldots, \pi_{L,1}, \ldots, \pi_{L,L})^t : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i, j \in \mathcal{L} \\ \sum_{i=1}^L \sum_{j=1}^L \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i, j) = 0 \\ \sum_{j>i} \pi_{i,j} = \psi \end{array} \right\}. \tag{11}$$

Choose any vector $\pi \in \Pi^\psi$. By the definition of Π^ψ, the components of π satisfy $0 \leq \pi_{i,j} \leq 1$ for all i, j. Thus, we have

$$||\pi|| = \sqrt{\pi^2_{1,1} + \cdots + \pi^2_{1,L} + \pi^2_{2,1} + \cdots + \pi^2_{2,L} + \cdots + \pi^2_{L,1} + \cdots + \pi^2_{L,L}} \leq L.$$

It follows that the set Π^ψ is bounded. Also, the set Π^ψ is closed since it is a polyhedron.

Since Π^ψ is closed and bounded, it is compact. Define the mapping $F : \Pi^\psi \to \mathbb{R}^{2L}$, where

$$\begin{aligned} F(\pi) &= (\gamma_{01}, \ldots, \gamma_{0L}, \gamma_{11}, \ldots, \gamma_{1L}), \\ \gamma_{0i} &= \sum_{j=1}^L \pi_{i,j} \text{ for all } i \in \mathcal{L}, \\ \gamma_{1j} &= \sum_{i=1}^L \pi_{i,j} \text{ for all } j \in \mathcal{L}. \end{aligned}$$

The mapping F is continuous by Proposition 11.1, Theorem 11.2, and Theorem 11.4 in Fitzpatrick (1996). Let $F(\Pi^\psi)$ denote the image of $F : \Pi^\psi \to \mathbb{R}^{2L}$, i.e.,

$$F(\Pi^\psi) = \{\gamma | \gamma = F(\pi) \text{ for some point } \pi \in \Pi^\psi\}.$$

By Theorem 11.12 in Fitzpatrick (1996), $F(\Pi^\psi)$ is compact. Since $\Gamma^\psi = F(\Pi^\psi)$, the set Γ^ψ is compact. □

Lemma 6. *Let A be the set of vectors $\widetilde{\gamma} = (\widetilde{\gamma}_{01}, \ldots, \widetilde{\gamma}_{0L}, \widetilde{\gamma}_{11}, \ldots, \widetilde{\gamma}_{1L})$ that satisfy the following: (1) $\widetilde{\gamma}_{aj} \geq 0$ for any $a \in \{0,1\}$ and $j \in \mathcal{L}$; (2) $\sum_{j=1}^{L} \widetilde{\gamma}_{aj} = 1$ for any $a \in \{0,1\}$. The functions $g : (0,1) \times A \to \mathbb{R}$ and $g_\psi : (0,1) \times A \to \mathbb{R}$, defined as*

$$g(\widetilde{\theta}, \widetilde{\gamma}) = \min_{\gamma \in \Gamma} Discrep_{\widetilde{\theta}}(\gamma, \widetilde{\gamma}), \tag{12}$$

$$g_\psi(\widetilde{\theta}, \widetilde{\gamma}) = \min_{\gamma \in \Gamma^\psi} Discrep_{\widetilde{\theta}}(\gamma, \widetilde{\gamma}), \tag{13}$$

are continuous at (θ, γ^*).

Proof. We prove that g_ψ is continuous at (θ, γ^*). The proof for g is analogous. Consider any sequence $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ that converges to (θ, γ^*), where $(\widetilde{\theta}_n, \widetilde{\gamma}_n) \in (0,1) \times A$ for every n. We want to show that $g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ converges to $g_\psi(\theta, \gamma^*)$.

Define the mappings $f : \mathbb{R}^{2L} \to \mathbb{R}$ and $f_n : \mathbb{R}^{2L} \to \mathbb{R}$ as:

$$f(\gamma) = \text{Discrep}_\theta(\gamma, \gamma^*),$$
$$f_n(\gamma) = \text{Discrep}_{\widetilde{\theta}_n}(\gamma, \widetilde{\gamma}_n).$$

Thus, we have

$$g_\psi(\theta, \gamma^*) = \min_{\gamma \in \Gamma^\psi} f(\gamma),$$
$$g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) = \min_{\gamma \in \Gamma^\psi} f_n(\gamma).$$

Consider any $\gamma \in \Gamma^\psi$. Let $\delta_n(\gamma) = f_n(\gamma) - f(\gamma)$. Then we have

$$\begin{aligned}
\delta_n(\gamma) &= (1-\theta)\sum_{j=1}^{L}(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j})^2 + \theta\sum_{j=1}^{L}(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j})^2 \\
&\quad +(\theta - \widetilde{\theta}_n)\sum_{j=1}^{L}(\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 + (\widetilde{\theta}_n - \theta)\sum_{j=1}^{L}(\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \\
&\quad +2(1-\theta)\sum_{j=1}^{L}(\gamma_{0j} - \gamma_{0j}^*)(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j}) + 2\theta\sum_{j=1}^{L}(\gamma_{1j} - \gamma_{1j}^*)(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j}).
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
|\delta_n(\gamma)| &\leq (1-\theta)\sum_{j=1}^{L}(\gamma_{0j}^* - \widetilde{\gamma}_{n,0j})^2 + \theta\sum_{j=1}^{L}(\gamma_{1j}^* - \widetilde{\gamma}_{n,1j})^2 \\
&\quad + \left|\theta - \widetilde{\theta}_n\right|\sum_{j=1}^{L}(\gamma_{0j} - \widetilde{\gamma}_{n,0j})^2 + \left|\widetilde{\theta}_n - \theta\right|\sum_{j=1}^{L}(\gamma_{1j} - \widetilde{\gamma}_{n,1j})^2 \\
&\quad +2(1-\theta)\sum_{j=1}^{L}\left(\left|\gamma_{0j} - \gamma_{0j}^*\right|\left|\gamma_{0j}^* - \widetilde{\gamma}_{n,0j}\right|\right) + 2\theta\sum_{j=1}^{L}\left(\left|\gamma_{1j} - \gamma_{1j}^*\right|\left|\gamma_{1j}^* - \widetilde{\gamma}_{n,1j}\right|\right).
\end{aligned}$$

Since $0 < \theta < 1$, we have

$$\begin{aligned}
|\delta_n(\gamma)| &\leq \sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}^* - \widetilde{\gamma}_{n,aj})^2 + \left|\theta - \widetilde{\theta}_n\right|\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj} - \widetilde{\gamma}_{n,aj})^2 \\
&\quad +2\sum_{a=0}^{1}\sum_{j=1}^{L}\left(\left|\gamma_{aj} - \gamma_{aj}^*\right|\left|\gamma_{aj}^* - \widetilde{\gamma}_{n,aj}\right|\right).
\end{aligned}$$

Since $\gamma^* \in \Gamma$, $\gamma \in \Gamma^\psi$, and $\widetilde{\gamma}_n \in A$, we have that $|\gamma_{aj} - \gamma_{aj}^*|$ is bounded by 1 and $\sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj} - \widetilde{\gamma}_{n,aj})^2$ is bounded by $2L$. This implies that

$$|\delta_n(\gamma)| \leq \sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma_{aj}^* - \widetilde{\gamma}_{n,aj})^2 + 2L\left|\theta - \widetilde{\theta}_n\right| + 2\sum_{a=0}^{1}\sum_{j=1}^{L}\left|\gamma_{aj}^* - \widetilde{\gamma}_{n,aj}\right|.$$

Let α be any positive number. Since $(\widetilde{\theta}_n, \widetilde{\gamma}_n)$ converges to (θ, γ^*), there exists a positive integer N_α such that if $n \geq N_\alpha$,

$$\begin{aligned} |\theta - \widetilde{\theta}_n| &< \alpha, \\ \sum_{a=0}^{1}\sum_{j=1}^{L}(\gamma^*_{aj} - \widetilde{\gamma}_{n,aj})^2 &< \alpha, \\ |\gamma^*_{aj} - \widetilde{\gamma}_{n,aj}| &< \alpha \text{ for any } a \in \{0,1\}, j \in \mathcal{L}. \end{aligned}$$

It follows that if $n \geq N_\alpha$,

$$0 \leq |\delta_n(\gamma)| < (6L+1)\alpha.$$

Since this result holds for arbitrary $\gamma \in \Gamma^\psi$, it follows that if $n \geq N_\alpha$,

$$0 \leq \sup_{\gamma \in \Gamma^\psi} |\delta_n(\gamma)| \leq (6L+1)\alpha.$$

Since the choice of α was arbitrary, we have that $\sup_{\gamma\in\Gamma^\psi} |\delta_n(\gamma)| \to 0$.

We want to show that $g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) \to g_\psi(\theta, \gamma^*)$. For any n, we have

$$\begin{aligned} |g_\psi(\widetilde{\theta}_n, \widetilde{\gamma}_n) - g_\psi(\theta, \gamma^*)| &= \left| \min_{\gamma\in\Gamma^\psi} f_n(\gamma) - \min_{\gamma\in\Gamma^\psi} f(\gamma) \right| \\ &\leq \sup_{\gamma\in\Gamma^\psi} |f_n(\gamma) - f(\gamma)| \\ &= \sup_{\gamma\in\Gamma^\psi} |\delta_n(\gamma)|, \end{aligned}$$

which was shown above to converge to 0. This completes the proof of the lemma. □

F Proof of Theorem 3

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

Proof. Consider an arbitrary data generating distribution P_0 that satisfies Assumption 1. Choose any ψ that is consistent with the marginal distributions $\boldsymbol{\gamma}^*$ and restrictions $\mathcal{R}$, i.e., $\boldsymbol{\gamma}^* \in \Gamma^\psi$. Then for $\epsilon = 10^{-10}$,

$$\begin{aligned} \liminf_{n\to\infty} P_0(\psi \in CS_n) &= \liminf_{n\to\infty} P_0(T_{n,\psi} \leq t^{0.95}_\psi + \epsilon) \\ &\geq \liminf_{n\to\infty} P_0(T_{n,\psi} < t^{0.95}_\psi + \epsilon) \\ &\geq P_0(T_\psi < t^{0.95}_\psi + \epsilon) \\ &\geq P_0(T_\psi \leq t^{0.95}_\psi) \\ &= 0.95, \end{aligned}$$

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

G Proof of Theorem 4

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

Proof. Consider any data generating distribution P_0 that satisfies Assumption 1. Consider any ψ such that $\boldsymbol{\gamma}^* \notin \Gamma^\psi$. For $\epsilon = 10^{-10}$,

$$\lim_{n\to\infty} P_0(\psi \notin CS_n) = \lim_{n\to\infty} P_0(T_{n,\psi} > t^{0.95}_\psi + \epsilon) = 1.$$

The first equality follows from the definition of CS_n. The second equality follows from Theorem 2. □

H The m-out-of-n bootstrap

Let A and B denote the left and right endpoints of the confidence interval constructed using the m-out-of-n bootstrap, which we define next. For a given trial data set, to compute the value A, 10,000 replicate data sets are generated, each by sampling $m \leq n$ participants with replacement. Using each replicate 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 et al. (2017). These estimators are presented in the following section, entitled "Estimators from Huang et al. (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

$$\begin{aligned} P_0(A \leq \psi_0 \leq B) &\geq P_0(A \leq \psi_l^{\mathcal{R}} \leq \psi_0 \leq \psi_u^{\mathcal{R}} \leq B) \\ &= P_0(A \leq \psi_l^{\mathcal{R}} \leq \psi_u^{\mathcal{R}} \leq B) \\ &= 1 - P_0(A > \psi_l^{\mathcal{R}} \text{ or } B < \psi_u^{\mathcal{R}}) \\ &\geq 1 - P_0(A > \psi_l^{\mathcal{R}}) - P_0(B < \psi_u^{\mathcal{R}}) \\ &\approx 1 - 0.025 - 0.025 \\ &= 0.95. \end{aligned}$$

I Estimators from Huang et al. (2017)

For any $i \in \mathcal{L}$ and $j \in \mathcal{L}$, let

$$\begin{aligned} \widehat{F}_C(i) &= \frac{\sum_{m=1}^n 1(A=0, Y \leq i)}{\sum_{m=1}^n 1(A=0)}; \\ \widehat{F}_T(j) &= \frac{\sum_{m=1}^n 1(A=1, Y \leq j)}{\sum_{m=1}^n 1(A=1)}. \end{aligned}$$

Below is a direct excerpt from Appendix G in the Supplementary Materials of Huang et al. (2017): "The lower bound estimator $\overline{\psi}_l^{\mathcal{R}}$ is computed using a sequence of two linear programs:

$$\overline{\epsilon} = \min \left\{ \epsilon \geq 0 : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \epsilon \text{ for all } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \epsilon \text{ for all } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\}, \tag{14}$$

$$\overline{\psi}_l^{\mathcal{R}} = \min \left\{ \sum_{\substack{j>i \\ i,j \in \mathcal{L}}} \pi_{i,j} : \begin{array}{l} \pi_{i,j} \geq 0 \text{ for all } i,j \in \mathcal{L} \\ |\sum_{i'=1}^{i} \sum_{j=1}^{L} \pi_{i',j} - \widehat{F}_C(i)| \leq \overline{\epsilon} \text{ for all } i = 1,..,L-1 \\ |\sum_{j'=1}^{j} \sum_{i=1}^{L} \pi_{i,j'} - \widehat{F}_T(j)| \leq \overline{\epsilon} \text{ for all } j = 1,..,L-1 \\ \sum_{i=1}^{L} \sum_{j=1}^{L} \pi_{i,j} = 1 \\ \pi_{i,j} = 0 \text{ if } g(i,j) = 0 \end{array} \right\}. \tag{15}$$

The upper bound estimator $\overline{\psi}_u^{\mathcal{R}}$ is (15), with min replaced by max."

J Definition of Outcome in Setting D

We discretize reduction in clot volume using a bin length of 5 mL. The resulting ordinal outcome is as follows, where y represents the continuous reduction in clot volume:

Level 1: $y < 0$ mL
Level 2: 0 mL $\leq y < 5$ mL
Level 3: 5 mL $\leq y < 10$ mL
Level 4: 10 mL $\leq y < 15$ mL
Level 5: 15 mL $\leq y < 20$ mL
Level 6: $y \geq 20$ mL.

The list above is directly from Appendix L in the Supplementary Materials of Huang et al. (2017). In that paper, this outcome was referred to as RICV5.

References

Bertsekas, D. P. (2009), *Convex Optimization Theory*, Belmont: Athena Scientific.

Fitzpatrick, P. (1996), *Advanced Calculus: A Course in Mathematical Analysis*, Boston: PWS Publishing Company.

Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017), "Inequality in Treatment Benefits: Can We Determine if a New Treatment Benefits the Many or the Few?," *Biostatistics*, 18(2), 308–324.

Jakubowski, A. (1998), "The Almost Sure Skorokhod Representation for Subsequences in Non-metric Spaces," *Theory of Probability and Its Applications*, 42(1), 167–174.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014), *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Series on Optimization, Philadelphia: Society for Industrial and Applied Mathematics.

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

		m-out-of-n bootstrap				
n	$\widehat{CI}_n$	$m = n$	$m = 0.9n$	$m = 0.75n$	$m = 0.5n$	$m = 0.25n$
200	0.56	0.54	0.55	0.55	0.57	0.60
500	0.53	0.53	0.53	0.53	0.54	0.56
1000	0.52	0.52	0.52	0.52	0.53	0.54
2000	0.51	0.51	0.51	0.52	0.52	0.53

Table 1: Average widths of our method $\widehat{CI}_n$ and the m-out-of-n bootstrap in Setting A, at each sample size

		m-out-of-n bootstrap				
n	$\widehat{CI}_n$	$m = n$	$m = 0.9n$	$m = 0.75n$	$m = 0.5n$	$m = 0.25n$
200	0.09	0.14	0.15	0.16	0.20	0.28
500	0.05	0.09	0.09	0.10	0.12	0.18
1000	0.04	0.06	0.06	0.07	0.09	0.12
2000	0.03	0.04	0.05	0.05	0.06	0.09

Table 2: Average widths of our method $\widehat{CI}_n$ and the m-out-of-n bootstrap in Setting B, at each sample size

		m-out-of-n bootstrap				
n	$\widehat{CI}_n$	$m = n$	$m = 0.9n$	$m = 0.75n$	$m = 0.5n$	$m = 0.25n$
200	0.45	0.48	0.49	0.51	0.56	0.66
500	0.37	0.39	0.40	0.42	0.45	0.54
1000	0.33	0.35	0.36	0.37	0.39	0.45
2000	0.30	0.32	0.33	0.33	0.35	0.39

Table 3: Average widths of our method $\widehat{CI}_n$ and the m-out-of-n bootstrap in Setting C, at each sample size

		m-out-of-n bootstrap				
n	$\widehat{CI}_n$	$m = n$	$m = 0.9n$	$m = 0.75n$	$m = 0.5n$	$m = 0.25n$
200	0.29	0.25	0.25	0.26	0.29	0.33
500	0.22	0.21	0.22	0.22	0.24	0.28
1000	0.19	0.19	0.19	0.20	0.21	0.24
2000	0.17	0.18	0.18	0.18	0.19	0.21

Table 4: Average widths of our method $\widehat{CI}_n$ and the m-out-of-n bootstrap in Setting D, at each sample size

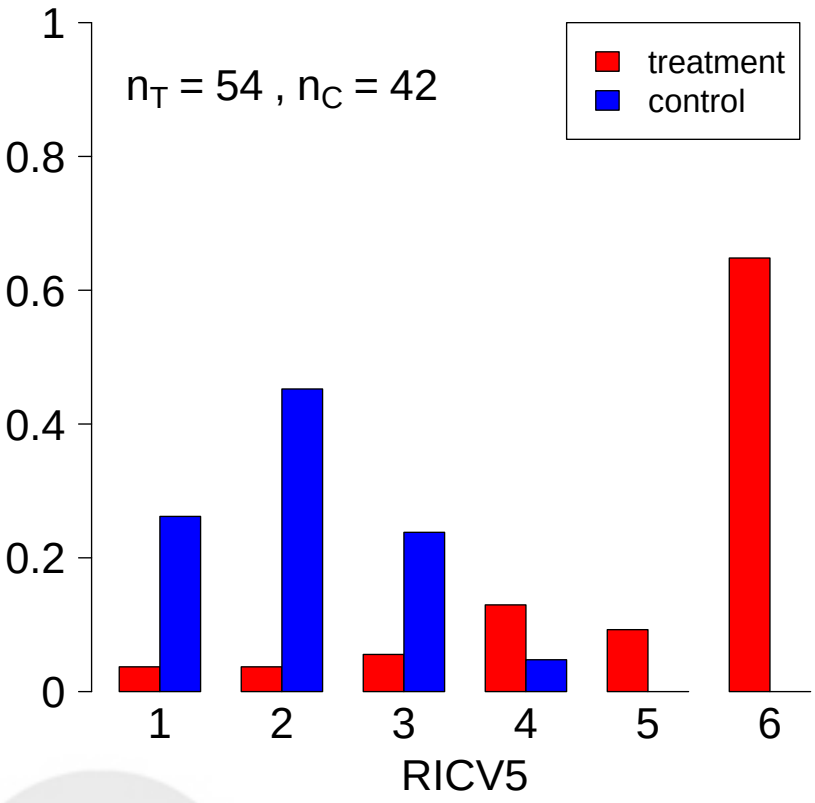

Figure 1: Empirical probability mass functions of RICV5 under treatment and control, computed using the MISTIE II randomized trial. The sample sizes of the treatment and control arms are shown in the upper left hand corner.

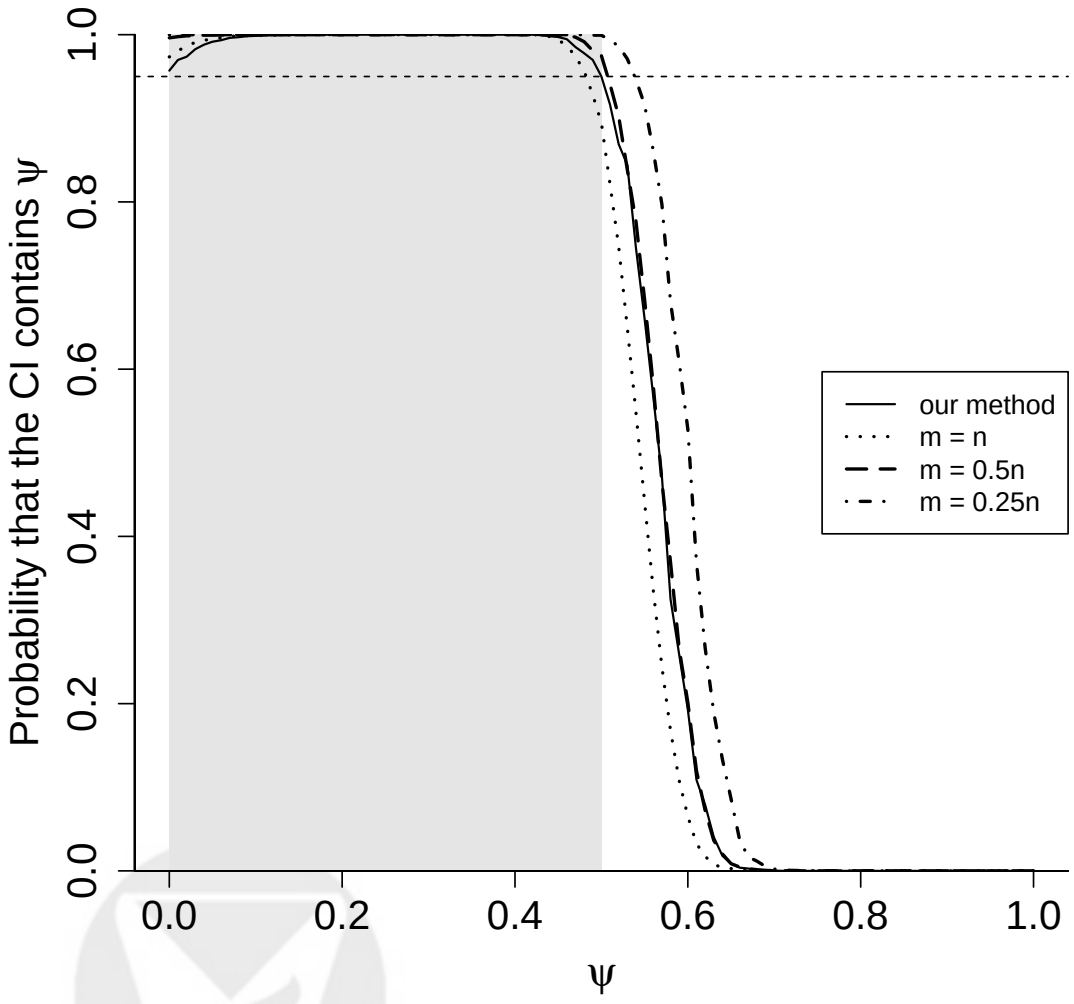

Figure 2: Coverage probabilities in Setting A at $n = 200$.

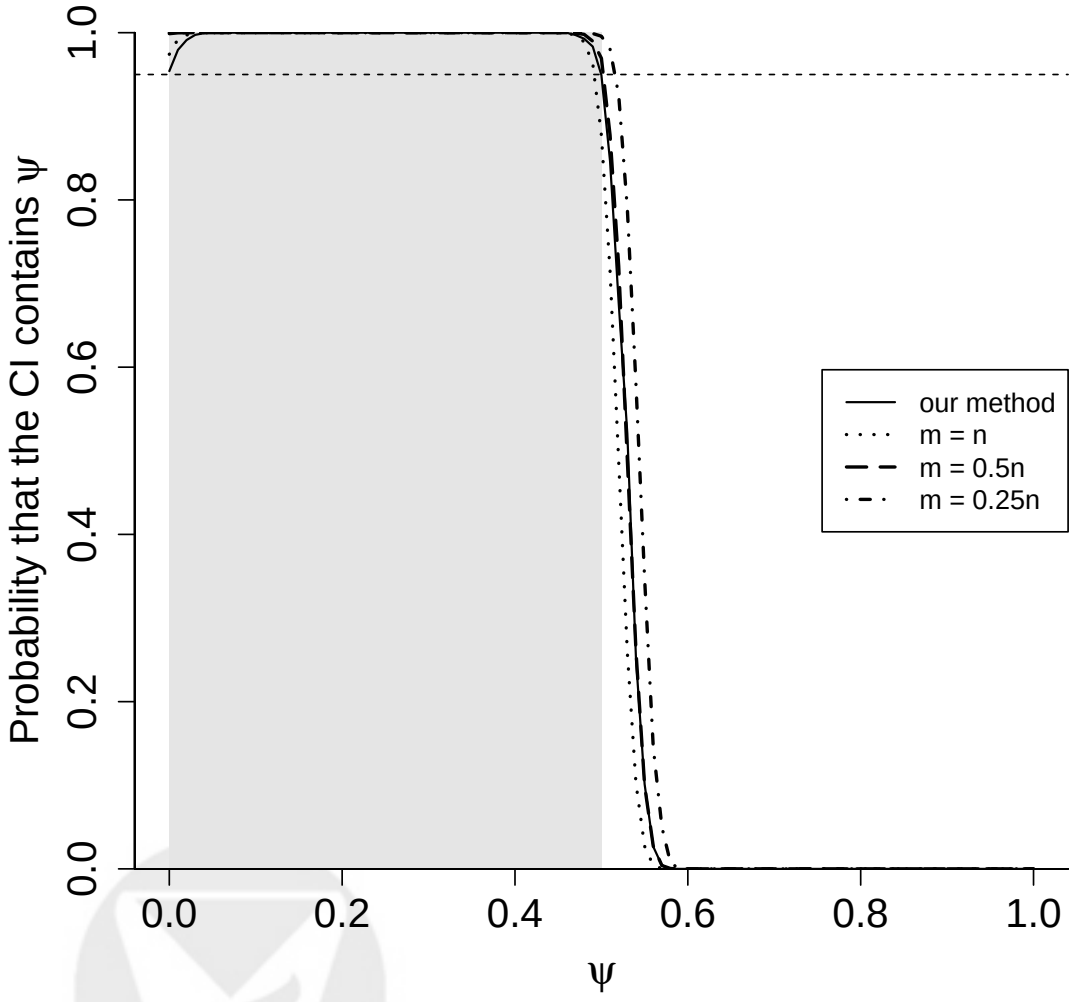

Figure 3: Coverage probabilities in Setting A at $n = 1000$.

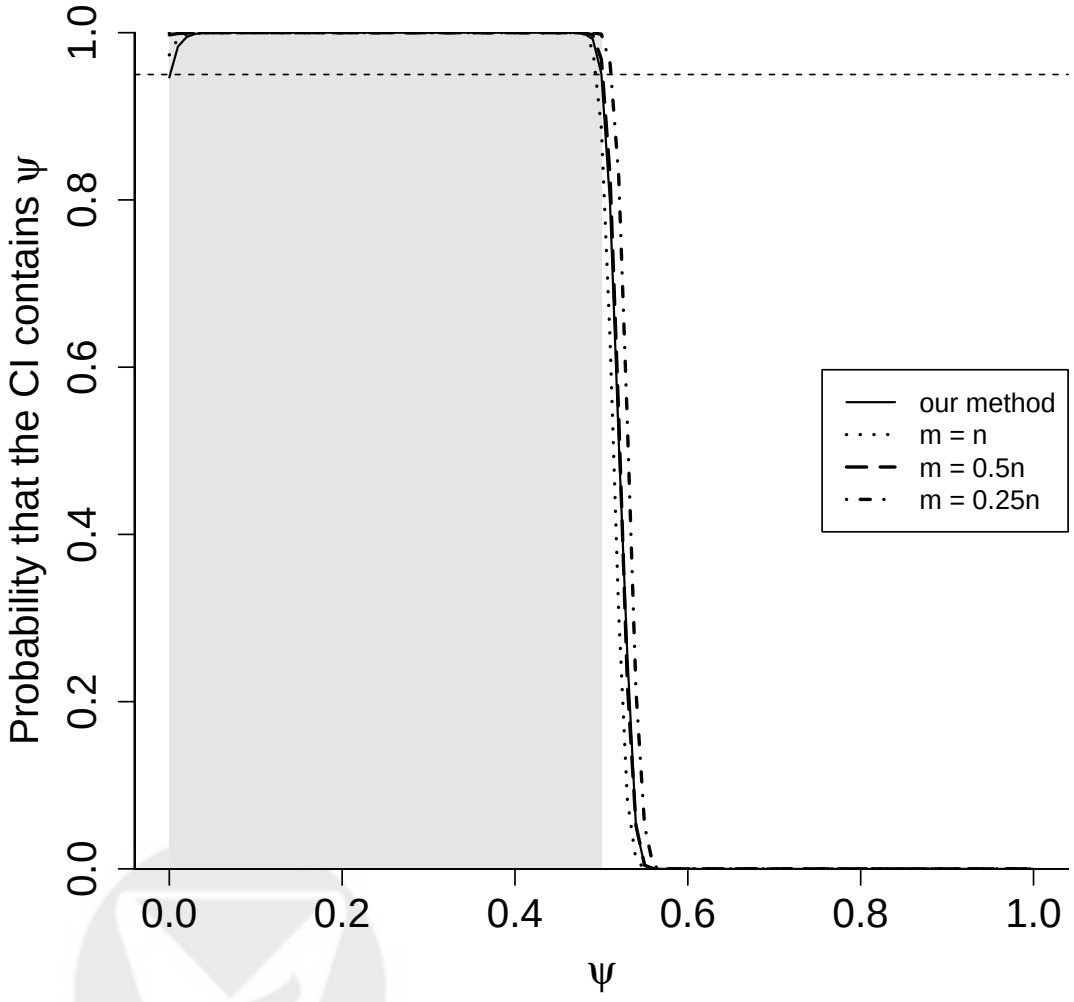

Figure 4: Coverage probabilities in Setting A at $n = 2000$.

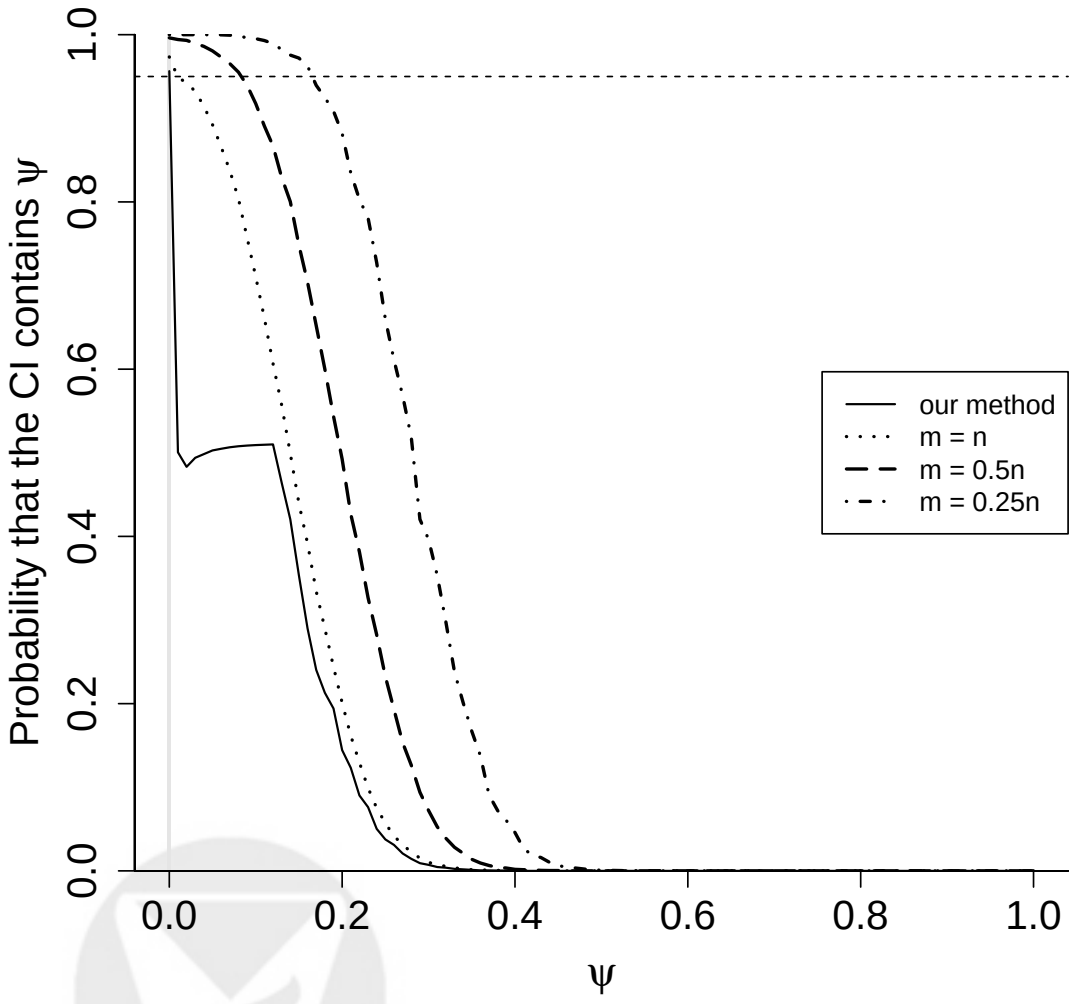

Figure 5: Coverage probabilities in Setting B at $n = 200$.

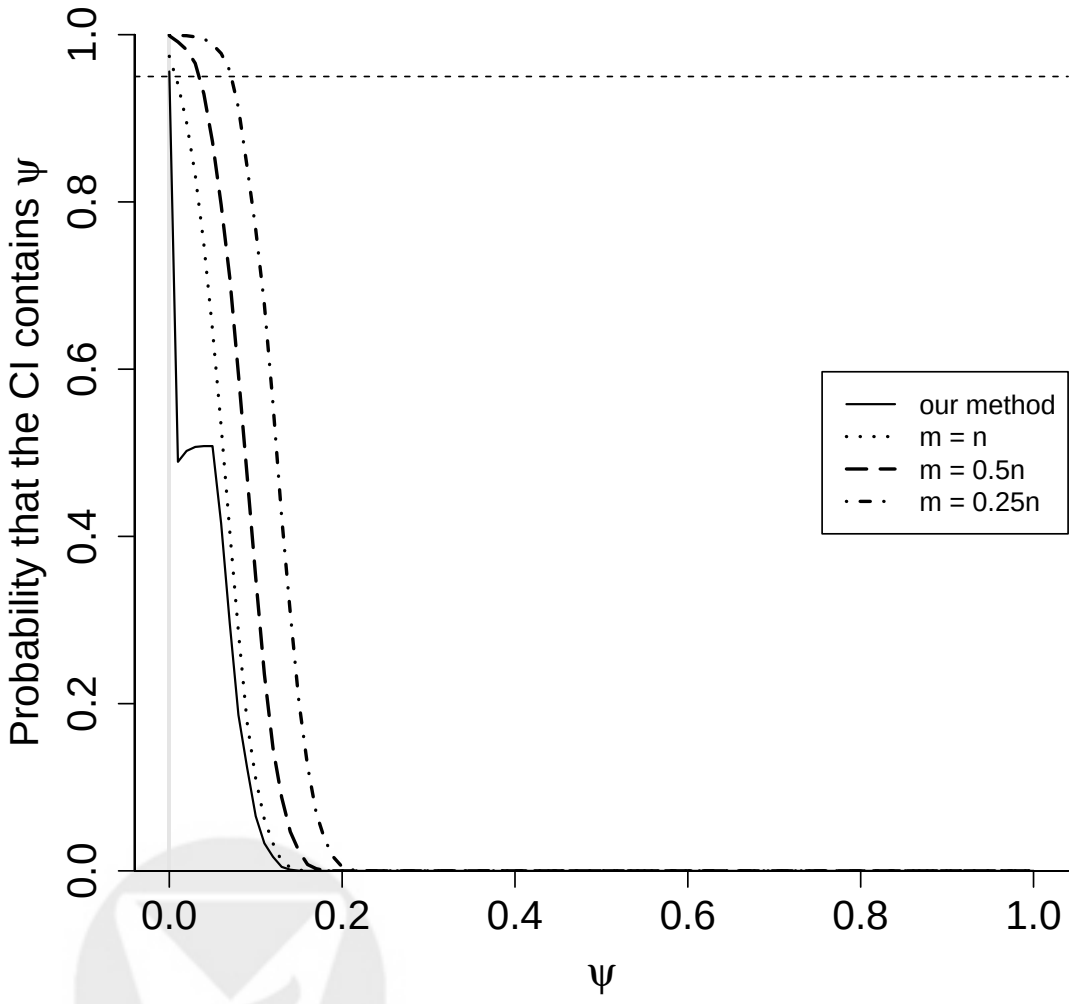

Figure 6: Coverage probabilities in Setting B at $n = 1000$.

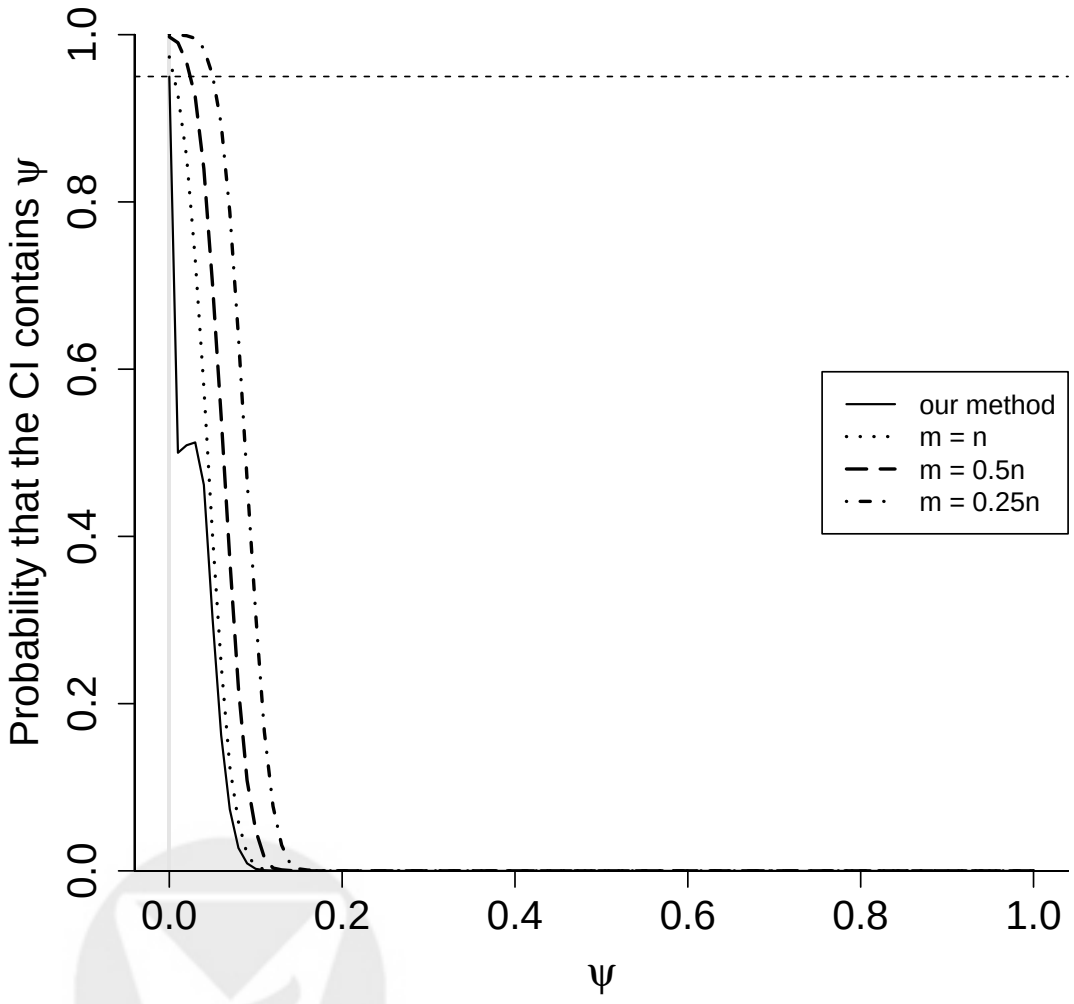

Figure 7: Coverage probabilities in Setting B at $n = 2000$.

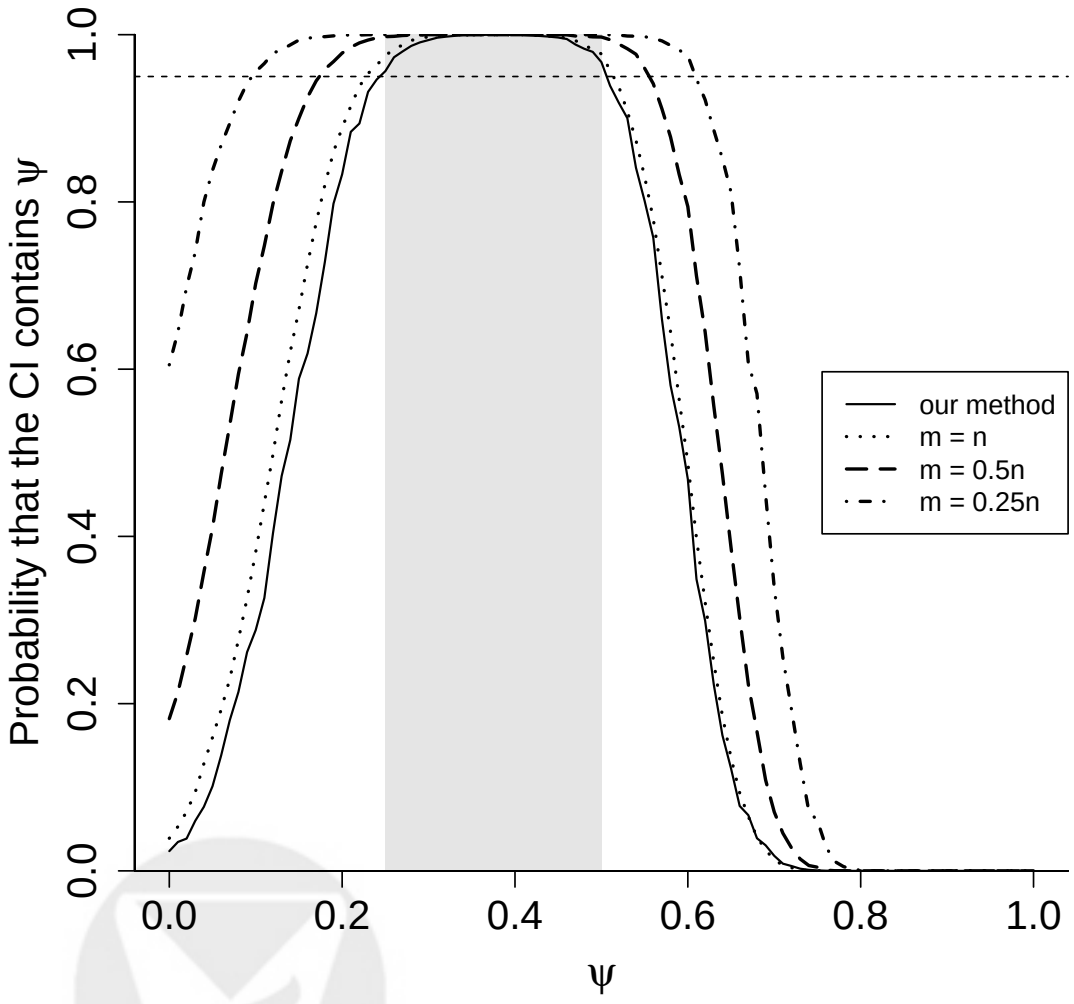

Figure 8: Coverage probabilities in Setting C at $n = 200$.

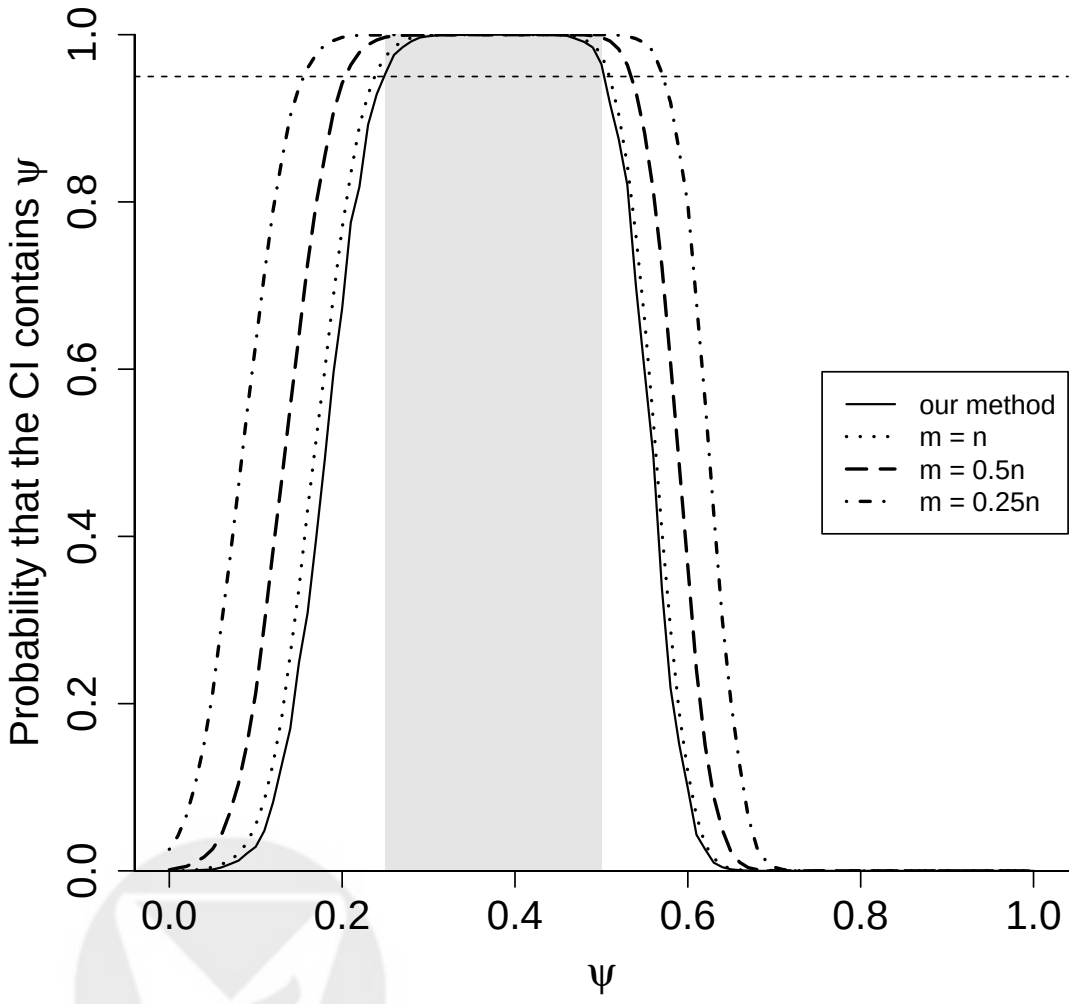

Figure 9: Coverage probabilities in Setting C at $n = 500$.

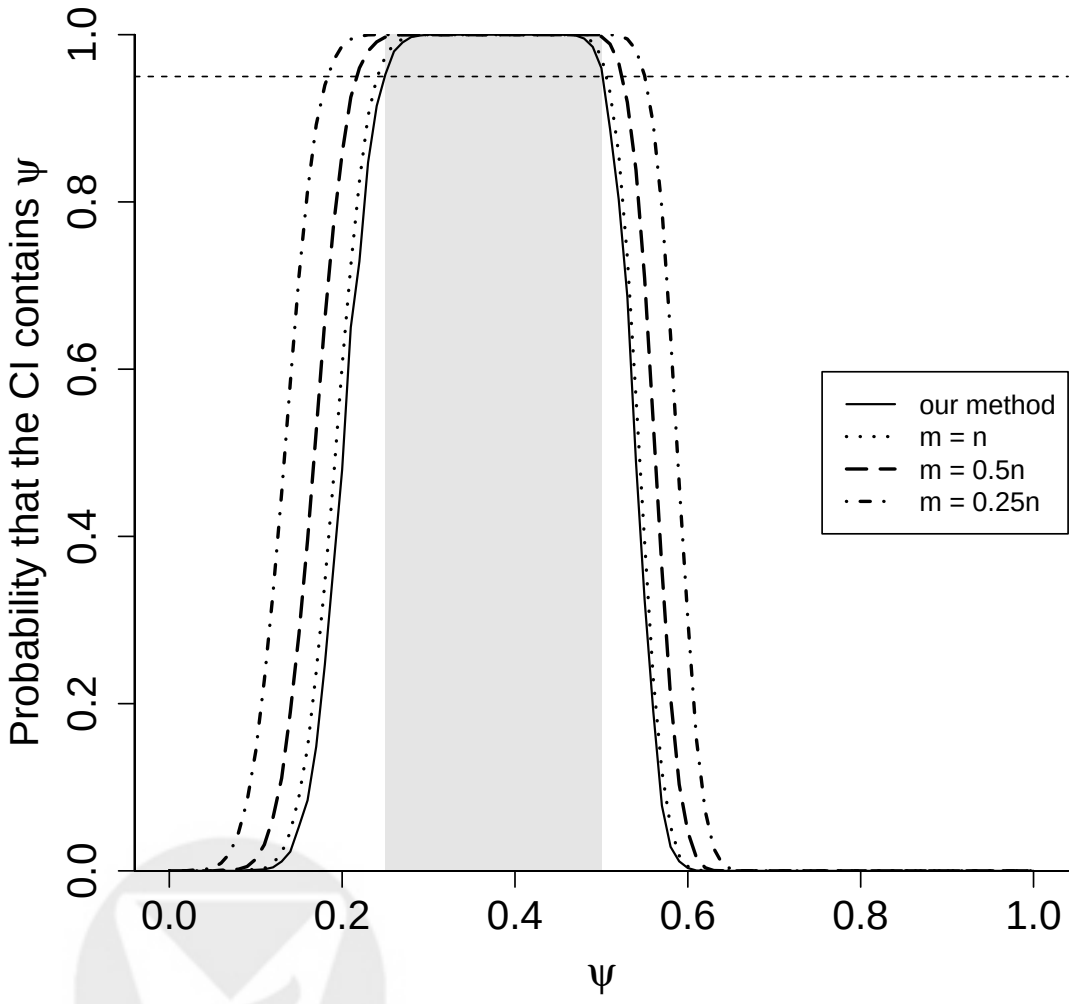

Figure 10: Coverage probabilities in Setting C at $n = 1000$.

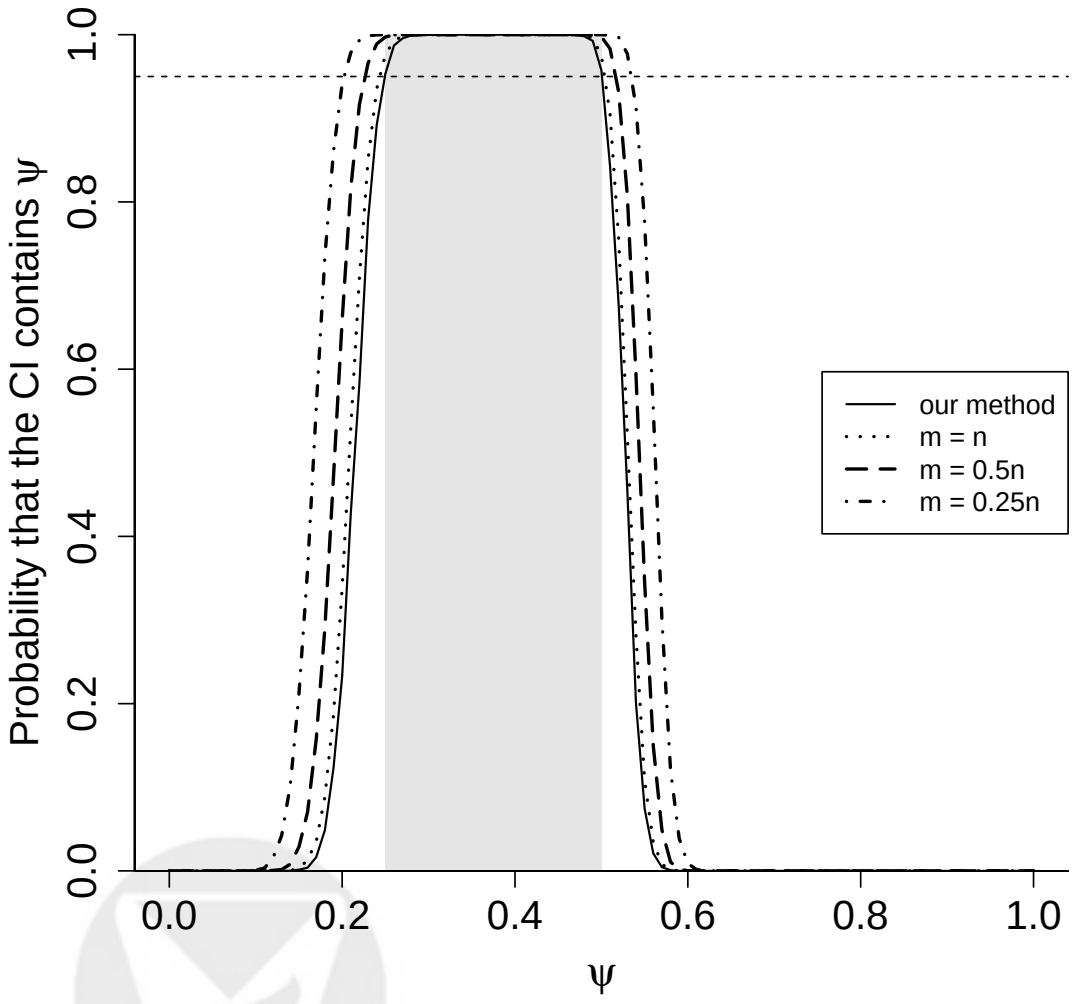

Figure 11: Coverage probabilities in Setting C at $n = 2000$.

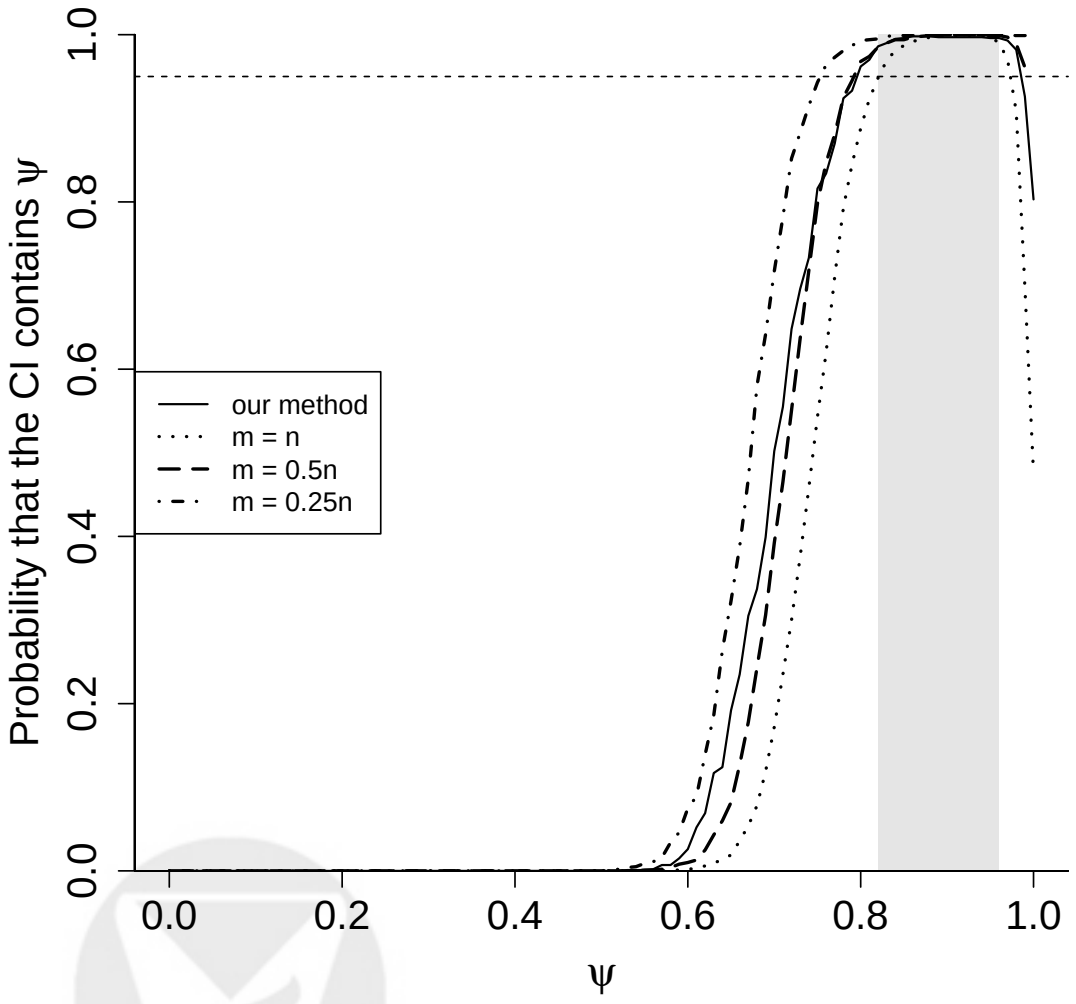

Figure 12: Coverage probabilities in Setting D at $n = 200$.

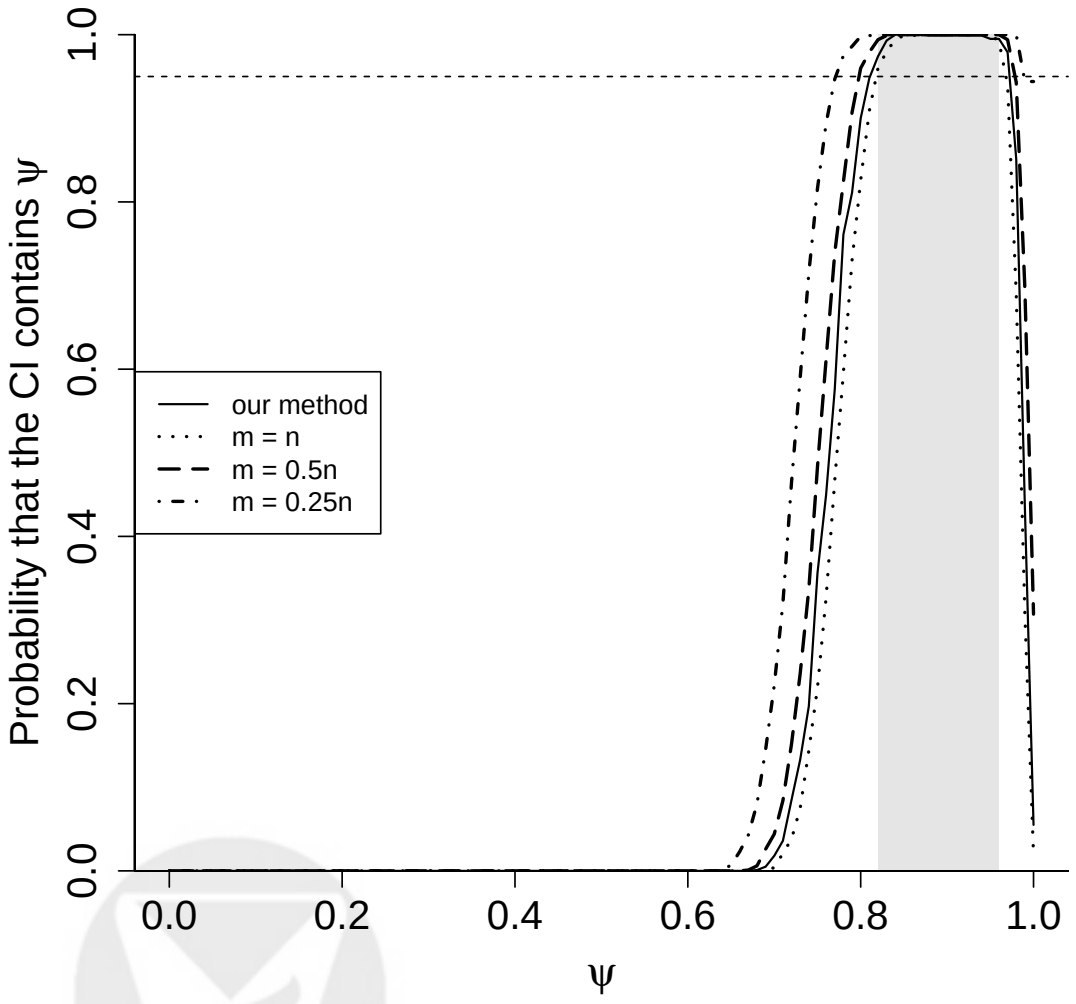

Figure 13: Coverage probabilities in Setting D at $n = 500$.

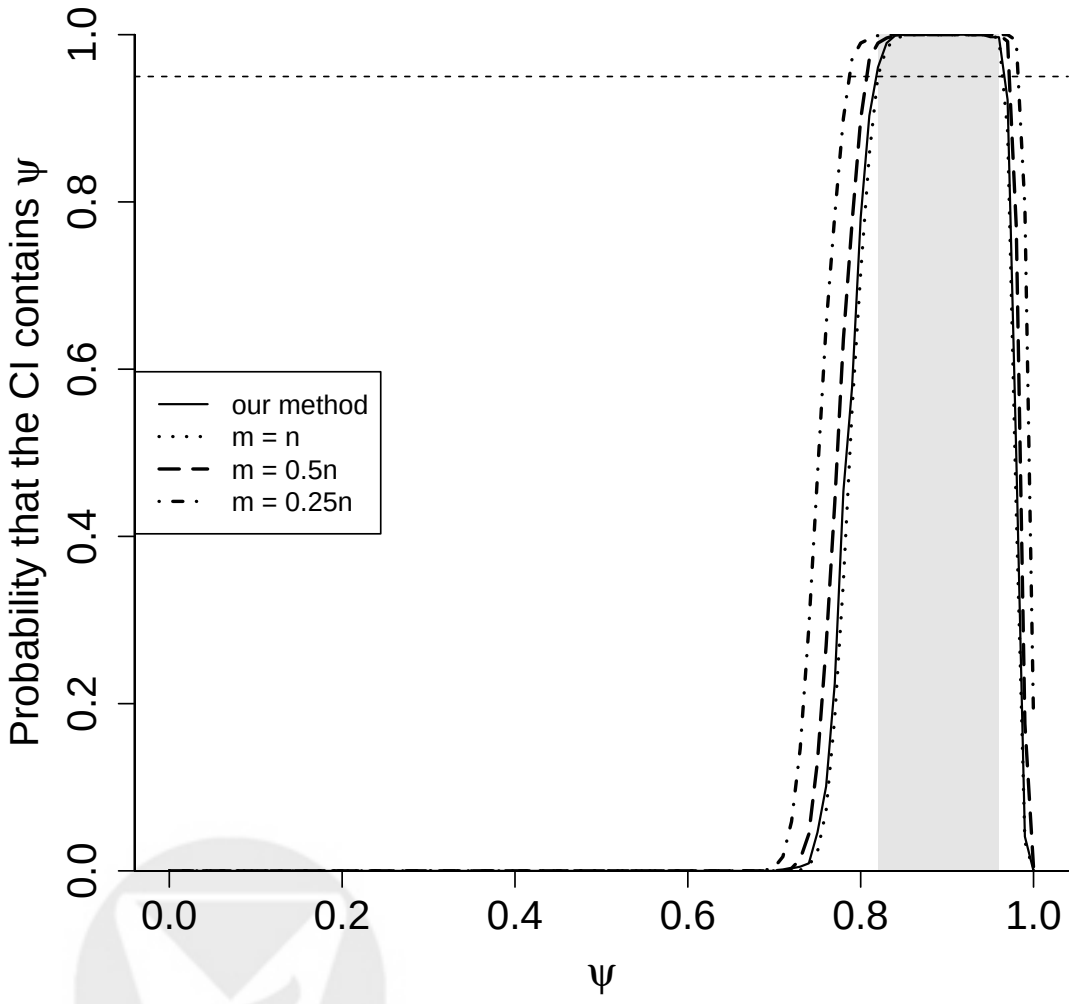

Figure 14: Coverage probabilities in Setting D at $n = 1000$.

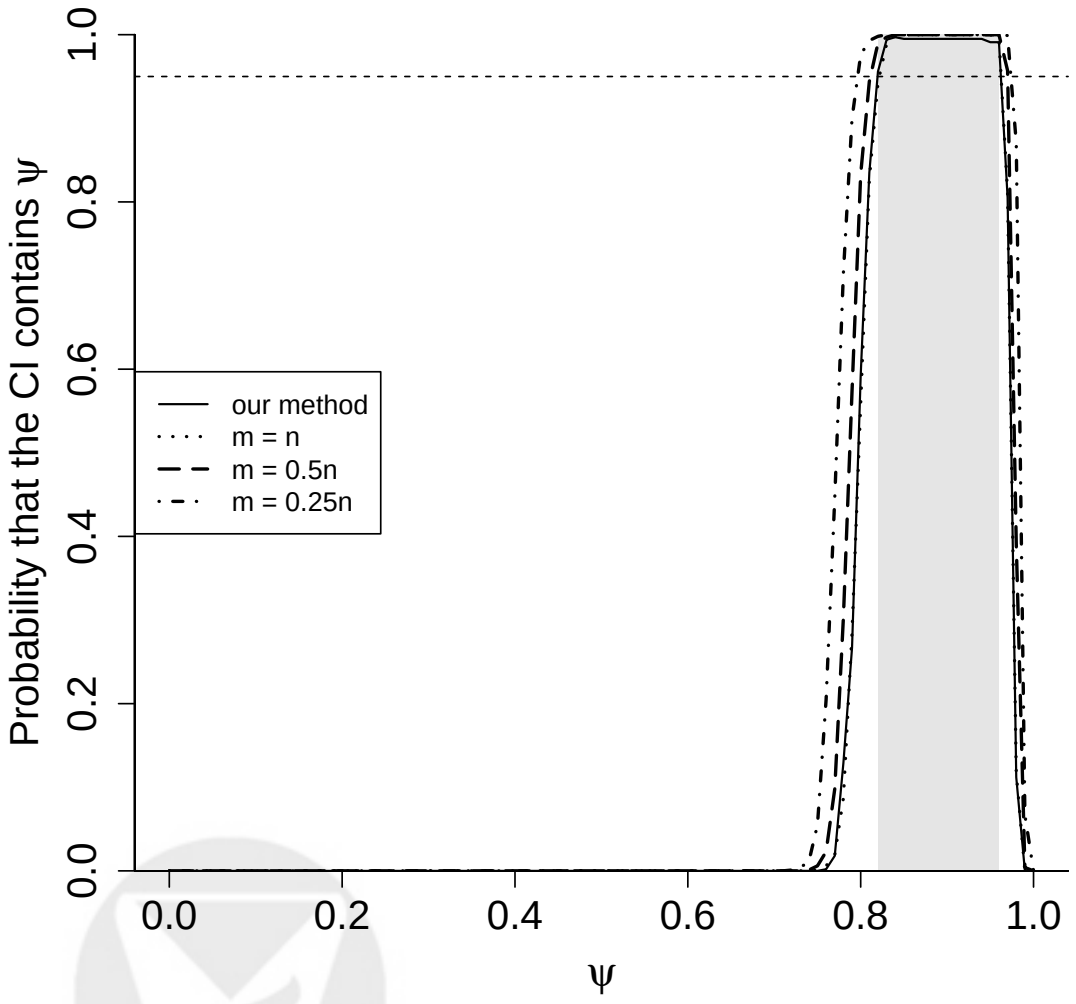

Figure 15: Coverage probabilities in Setting D at $n = 2000$.