Efficient Nonparametric Estimation of Causal Effects in Randomized Trials with Noncompliance

Jing Cheng  
Division of Biostatistics, University of Florida,  
Gainesville, FL 32608, U.S.A.  
jcheng@biostat.ufl.edu

Dylan S. Small  
Department of Statistics, University of Pennsylvania,  
Philadelphia, PA 19104, U.S.A.  
dsmall@wharton.upenn.edu

Zhiqiang Tan  
Department of Biostatistics, Johns Hopkins University,  
Baltimore, MD 21205, U.S.A.  
ztan@jhsph.edu

Thomas R. Ten Have  
Division of Biostatistics, University of Pennsylvania,  
Philadelphia, PA 19104, U.S.A.  
ttenhave@cceb.med.upenn.edu
Abstract

Causal approaches based on the potential outcome framework provide a useful tool for addressing noncompliance problems in randomized trials. Various estimators, e.g. the instrumental variable (IV) estimator, have been proposed for causal effects of treatment. In this paper, we propose a new empirical likelihood-based estimator of causal treatment effects in randomized clinical trials with noncompliance. By using the empirical likelihood approach to construct a profile random sieve likelihood and taking into account the mixture structure in outcome distributions, our estimator is robust to parametric distribution assumptions and more efficient than the standard IV estimator. Our estimator can be applied to outcome variables with a continuous, ordinal or binary scale. We apply our method to data from a randomized trial of an intervention to improve the treatment of depression among depressed elderly patients in primary care practices.

Keywords: Efficient nonparametric estimation; causal effects; randomized trials; non-compliance; empirical likelihood
1. Introduction

Randomized trials provide a powerful tool for estimating the effect of a treatment. However, noncompliance is a common problem in randomized trials. When there is noncompliance, there is often interest in estimating the causal effect of actually receiving the treatment compared to receiving the control. Knowledge of this effect is useful for predicting the impact of the treatment in a setting for which compliance patterns might differ from the randomized trial and for scientific understanding of the treatment (Sommer and Zeger, 1991; Sheiner and Rubin, 1995; Small et al., 2005; Cheng and Small, 2006).

Note that intention-to-treat (ITT) analysis is not suitable for estimating the causal effect of actually receiving the treatment when there is noncompliance because it estimates the effect of assignment to the treatment group. An as-treated analysis seeks to estimate the causal effect of receiving the treatment but is biased if compliers are not comparable to noncompliers. Imbens and Angrist (1994) and Angrist et al. (1996) show that the causal effect of actually receiving the treatment for the subgroup of subjects who would receive the treatment if assigned to the treatment group and would receive the control if assigned to the control group (called the Complier Average Causal Effect (CACE) or the Local Average Treatment Effect in the econometrics literature (Imbens and Angrist, 1994)) is nonparametrically identified under certain (often plausible) assumptions that do not require compliers and noncompliers to be comparable. These assumptions, henceforth referred to as the IV assumptions, are discussed in Section 2. The CACE can be consistently estimated under the IV assumptions by the standard two stage least squares instrumental variables (IV) estimator. Imbens and Rubin (1997a, b) demonstrate that under the IV assumptions, the standard IV estimator is an inefficient estimator of the CACE because it does not make full use of the mixture structure of the outcome distributions of the four observed groups defined by the cross-classification of the randomization and treatment received; see Section 2.4 for further discussion. Imbens and Rubin (1997b) present three new alternatives to the standard IV estimator. One is based on a normal approximation to the outcome distributions in the four groups, and two are based on multinomial approximations to the outcome distributions. In
a simulation study with normally distributed outcomes, Imbens and Rubin (1997b) show that all three alternative estimators are more efficient than the standard IV estimator. The estimator that is based on a normal approximation to the outcome distributions is efficient when the outcomes are in fact normal but can have substantial bias when the outcomes are not normal; this is demonstrated in Section 4. The estimators based on multinomial approximations to the outcome distributions are in principle nonparametric. However, an approach for choosing the multinomial approximations is needed. Imbens and Rubin (1997b) provide reasonable approximations for their example but do not present any systematic approach for choosing the multinomial approximations.

Multinomial approximations to the outcome distributions are a type of sieve. A sieve is a sequence of approximations \( \{F_n\} \) to a space \( F \) of distributions such that \( F_n \to F \) as \( n \to \infty \) (Grenander, 1981). Maximizing the likelihood over a sieve rather than the whole parameter space often leads to desirable statistical properties, especially when the underlying parameter space is large (Shen and Wong, 1994). However, the construction of sieves is not an easy task. One approach to constructing sieves is to use a random approximation \( \hat{F}_n \) that depends on the data (a random sieve). The empirical likelihood approach (Owen, 1991) is based on an easily constructed random sieve (Shen et al., 1999). In this paper, we develop a more efficient nonparametric estimator for the CACE than the standard IV estimator by using the empirical likelihood approach to construct a profile random sieve likelihood. We demonstrate substantial efficiency gains for our approach in a simulation study and apply our approach to a randomized trial of an intervention for treating depression.

estimator are shown in Section 5.

2. Notation, Assumptions and Review of Established Estimators

2.1 Notation

We consider a two-arm randomized trial with \( N \) subjects, \( n_0 \) of whom are randomly assigned to the control group. We let \( \mathbf{R} \) be the \( N \)-dimensional vector of randomization assignments for all subjects, with individual element \( R_i = r \in \{0, 1\} \) according to whether subject \( i \) is assigned active treatment \( (R_i = 1) \) or control \( (R_i = 0) \). We let \( A^r \) be the \( N \)-
dimensional vector of potential treatment receiveds under randomization assignment \( r \) with individual element \( A^r_i \), where \( A^r_i = a \in \{0, 1\} \) according to whether subject \( i \) would take the control or treatment under randomization assignment \( r \). We let \( Y^r_{i,a} \) be the vector of potential responses under randomization assignment \( r \) and treatment receiveds \( a \), with individual element \( Y^r_{i,a} \) being the potential response for subject \( i \) with the vector of randomization assignments \( r \) and the vector of treatment receiveds \( a \). The sets of \( \{ Y^r_{i,a} \mid r \in \{0, 1\}^N, a \in \{0, 1\}^N \} \) and \( \{ A^r_i \mid r \in \{0, 1\}^N \} \) are “potential” responses and treatment receiveds in the sense that we can only observe one member of each set. The observed outcome and treatment received variables for subject \( i \) are \( Y^R_{i,A^R_i} \equiv Y_i \) and \( A^R_i \equiv A_i \) respectively.

2.2. Assumptions

We make similar assumptions as Angrist et al. (1996).

**Assumption 1: Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980).**

a. If \( r = r' \), then \( A^r_i = A^{r'}_i \) for subject \( i \).

b. If \( r = r' \) and \( a = a' \), then \( Y^r_{i,a} = Y^{r'}_{i,a'} \) for subject \( i \).

The SUTVA assumption allows us to write \( Y^r_{i,a} \), \( A^r_i \) as \( Y^r, A^r \).

**Assumption 2: Random Assignment**

For all \( N \) subjects, the treatment assignment \( R \) is random: \( pr(R = c) = 1/(N^n_0) \) for all \( c \) with \( N - n_0 \) ones and \( n_0 \) zeroes.

The random assignment assumption implies independence between assignment and pre-treatment variables including potential outcomes and potential treatment receiveds.

**Assumption 3: Random Sampling**

We assume that the \( N \) subjects in the trial are i.i.d. draws from a superpopulation, i.e., \( Y^{0,0}_i, Y^{0,1}_i, Y^{1,0}_i, Y^{1,1}_i, A^0_i, A^1_i, A^0_i, A^1_i\), \( i = 1, \ldots, N \) are i.i.d. with the same distribution as the random vector \( Y^{0,0}, Y^{0,1}, Y^{1,0}, Y^{1,1}, A^0, A^1 \).

**Assumption 4: Mean Exclusion Restriction**

We assume that \( E(Y^{r,a}) = E(Y^{r',a}) \) for all \( r, r', a \), i.e., the randomization assignment affects the mean of the observed outcome only through its effect on treatment received. The
mean exclusion restriction is weaker than the unit level exclusion restriction of Angrist et al. (1996), who assume \( Y_{i}^{r,a} = Y_{i}^{r',a} \) for all \( r, r', a \). We discuss ways that our estimator can be modified to utilize the stronger unit-level exclusion restriction assumption, when it is plausible, in Section 7.

**Assumption 5: Nonzero Average Causal Effect of \( R \) on \( A \):** \( E(A_{1} - A_{0}) \neq 0 \)

This assumption requires the randomization assignment \( R \) to have some effect on the average probability of receiving treatment.

**Assumption 6: Monotonicity:** \( pr(A_{1} \geq A_{0}) = 1 \)

This assumption says that there is no one who would receive the opposite treatment of her assignment under both assignment to treatment and to control.

### 2.3 Compliance Classes

Based on a subject’s joint values of potential treatment received \((A_{0}^{i}, A_{1}^{i})\), a subject in a two-arm trial can be classified into one of four compliance classes:

\[
C_{i} = \begin{cases} 
0 \text{ (never-taker)} & \text{if } (A_{0}^{i}, A_{1}^{i}) = (0, 0) \\
1 \text{ (complier)} & \text{if } (A_{0}^{i}, A_{1}^{i}) = (0, 1) \\
2 \text{ (always-taker)} & \text{if } (A_{0}^{i}, A_{1}^{i}) = (1, 1) \\
3 \text{ (defier)} & \text{if } (A_{0}^{i}, A_{1}^{i}) = (1, 0)
\end{cases}
\]

In a two-arm trial, we can observe only one of \( A_{0}^{i} \) and \( A_{1}^{i} \), so a subject’s compliance status is not observed directly in a trial, but it can be partially identified based on treatment assignment and observed treatment-received (see Table 1). Note that the monotonicity assumption rules out the existence of defiers. For single consent design trials (Zelen, 1979), which have the property that the control group cannot get the treatment (i.e., \( pr(A_{0} = 0) = 1 \), the presence of always-takers and defiers is ruled out.

### 2.4 Major Established Estimators

Under Assumptions 1 – 6, the compliers are the only subgroup for which a randomized trial provides information on the causal effect of receiving treatment (Angrist et al., 1996). For always-takers and never-takers, assignment to treatment has no effect on treatment
received. The Complier Average Causal Effect (CACE), $E(Y^1 - Y^0 | C = 1)$, can be thought of as the causal effect of receiving treatment for the subpopulation of compliers because for compliers, assignment of treatment agrees with receipt of treatment. Angrist et al. (1996) show that under Assumptions 1-6, the CACE, $E(Y^1 - Y^0 | C = 1)$, equals

$$
\frac{E(Y | R = 1) - E(Y | R = 0)}{E(A | R = 1) - E(A | R = 0)},
$$

which is the ITT effect divided by the proportion of compliers. The standard IV estimator (the two stage least squares estimator) is the sample analogue of (1),

$$
C\hat{ACE}_{SIV} = \frac{\hat{E}(Y | R = 1) - \hat{E}(Y | R = 0)}{\hat{E}(A | R = 1) - \hat{E}(A | R = 0)},
$$

where the $\hat{E}$’s denote sample means; (2) is sometimes called the Wald estimator.

The standard IV estimator does not take full advantage of the mixture structure of the outcomes of the four observed groups in Table 1, as we will discuss in Section 3.1. Imbens and Rubin (1997a,b) present two approaches to using mixture modeling to estimate the CACE. One approach assumes a parametric distribution for the outcomes for each compliance class group under each randomization assignment, e.g. normal distributions. The CACE is then estimated via maximum likelihood for this model using the EM algorithm. This estimator provides considerable efficiency gains over the standard IV estimator when the parametric assumptions hold (see Table 4). However, when the parametric assumptions are wrong, this estimator can be inconsistent whereas the standard IV estimator is consistent (see Table 4 for finite sample results).

Imbens and Rubin’s other approach to using mixture modeling to estimate the CACE is to approximate the density of the distribution of the outcomes for each compliance class under each randomization group as a piecewise constant function, and then estimate the CACE by maximum likelihood. This approach is in principle nonparametric as the number of constant pieces in each density function can be increased with the sample size. However, Imbens and Rubin (1997b) do not provide a systematic approach for how to choose the number of pieces
and the location of the pieces. We develop a systematic easily implementable approach for doing this using empirical likelihood in the next section. Note that our approach uses only the mean exclusion restriction (Assumption 4) rather than the stronger unit level exclusion restriction that Imbens and Rubin’s (1997b) approach requires.

3. Estimation through Empirical Likelihood Approach

3.1 Motivation and Description of Empirical Likelihood Approach

We first motivate and describe our method for single consent design trials, where the control group cannot get the treatment (so the presence of always takers and defiers is ruled out). Table 2 shows the relationship between observed \((R, A)\) groups and latent compliance classes for a single consent design trial. The CACE can be re-expressed under Assumptions 1-6 as follows:

\[
CACE = \mu^{c1} - \mu^{c0} = \mu^{c1} - \left( \frac{\mu^{R=0} - (1 - \pi_c)\mu^n}{\pi_c} \right) = E(Y|R = 1, A = 1) - \frac{E(Y|R = 0) - (1 - P(A = 1|R = 1))E(Y|R = 1, A = 0)}{P(A = 1|R = 1)}
\]  

(3)

where \(\mu^{c1}, \mu^{c0}\) and \(\mu^n\) denote the mean potential outcomes of the compliers under treatment, compliers under control and never takers respectively; \(\mu^{R=0}\) denotes the mean potential outcome of the whole population of subjects when assigned to the control; and \(\pi_c\) denotes the proportion of compliers. The standard IV estimator estimates the CACE by substituting the method of moments estimates from the sample for \(E(Y|R = 1, A = 1), E(Y|R = 0), P(A = 1|R = 1)\) and \(E(Y|R = 1, A = 0)\) into (3). However, as noted by Imbens and Rubin (1997b), there are restrictions on the joint density of \((Y, R, A)\) not taken into account by the method of moments that can be useful for estimating \(E(Y|R = 0), P(A = 1|R = 1)\) and \(E(Y|R = 1, A = 0)\). Specifically, Assumptions 1-6 imply

(R1) The distribution of \(Y|R = 0\) is a mixture of the outcome distribution of the never takers under \(R = 0\) and the outcome distribution of the compliers under \(R = 0\);

(R2) The mixing proportion \(\pi_c\) for \(Y|R = 0\) equals \(P(A = 1|R = 1)\) as a consequence of the random assignment assumption (Assumption 2); and

(R3) The mean of the never takers under \(R = 0\) is equal to the mean of the never takers under
\( R = 1 \) (which equals \( E(Y|R = 1, A = 0) \)) as a consequence of the exclusion restriction assumption (Assumption 4).

The sample mean of \( Y|R = 0 \) uses only the information in those \( Y_1, \ldots, Y_N \) for which \( R_i = 0 \) to estimate \( E(Y|R = 0) \), but the restrictions (R1)-(R3) imply that there is additional information in the \( Y_1, \ldots, Y_N \) for which \( R_i = 1 \). Likewise, the sample proportion of \( A = 1|R = 1 \) uses only the information in those \( A_1, \ldots, A_N \) for which \( R_i = 1 \) to estimate \( P(A = 1|R = 1) \) but the restrictions (R1)-(R3) imply that there is additional information in the \( A_1, \ldots, A_N \) for which \( R_i = 0 \). A body of work has shown that supplementing a sample from a distribution that is a mixture of two components with samples from one or both of the components alone provides additional information for estimating aspects of the mixture distribution, e.g., Hall and Titterington (1984), Lancaster and Imbens (1996) and Qin (1999). Here, the sample of \( Y_1, \ldots, Y_N \) for which \( R_i = 1, A_i = 0 \) provides information about the never taker component of the mixture \( Y|R = 0 \) and the sample of \( A_1, \ldots, A_N \) for which \( R_i = 1 \) provides information about the mixing proportion in the mixture \( Y|R = 0 \). We now illustrate how this information is useful in a setting with a binary outcome.

Consider a setting with a binary outcome in which

\[
\pi_c = 0.5, \mu^n = 0.2, \mu^c = 0.8, \mu^0 = 0.9 \text{ and } N = 40, n_0 = 20. \tag{4}
\]

The following is a plausible sample in this setting: \#(\( Y_i = 1, A_i = 1, R_i = 1 \)) = 8, \#(\( Y_i = 0, A_i = 1, R_i = 1 \)) = 2, \#(\( Y_i = 1, A_i = 0, R_i = 1 \)) = 2, \#(\( Y_i = 0, A_i = 0, R_i = 1 \)) = 8, \#(\( Y_i = 1, A_i = 0, R_i = 0 \)) = 13 and \#(\( Y_i = 0, A_i = 0, R_i = 0 \)) = 7; the \( p \)-value for a \( \chi^2 \) test of whether this sample comes from the distribution (4) is 0.37. Note that for this sample, the method of moments estimates of the quantities in (3) – \( \hat{E}(Y|R = 1, A = 1) = 0.8, \hat{E}(Y|R = 0) = 0.65, \hat{P}(A = 1|R = 1) = 0.5, \hat{E}(Y|R = 1, A = 0) = 0.2 \) – violate the restrictions (R1)-(R3), which imply \( E(Y|R = 1, A = 0)(1 - P(A = 1|R = 1)) \leq E(Y|R = 0) \leq E(Y|R = 1, A = 0)(1 - P(A = 1|R = 1)) + P(A = 1|R = 1) \). Figure 1 plots the profile log likelihood for this sample under the probability model given by Assumptions 1-6.
with binary outcomes. The maximum likelihood estimate (MLE) of the CACE, which takes into account the mixture structure of the outcomes given by (R1)-(R3), has a noticeably higher likelihood than the standard IV estimate, which ignores some of the restrictions in (R1)-(R3). The MLE’s property of taking into full account the mixture structure leads to substantially better estimates – in 1000 simulations from model (4), the mean square error (MSE) of the MLE was 0.048 compared to 0.156 for standard IV.

To take account of the mixture structure of the outcomes given by (R1)-(R3) for more general distribution of outcomes in a nonparametric way, we use the empirical likelihood approach. The empirical likelihood for a parameter(s) such as the CACE is the nonparametric profile likelihood for the parameter. The maximum empirical likelihood estimator (MELE) of a parameter(s) has been shown to have good properties for a wide class of semiparametric problems; see Owen (2001) and Qin and Lawless (1994) for discussion.

Without loss of generality, we arrange the subjects so that \( R_1 = \cdots = R_{n_0} = 0 \) and \( R_{n_0+1} = \cdots = R_N = 1 \); thus, \((Y_1, A_1), \ldots, (Y_{n_0}, A_{n_0})\) is a random sample from the population of \( Y_{r=0,a=A^{r=0}}, A^{r=0} \) and \((Y_{n_0+1}, A_{n_0+1}), \ldots, (Y_N, A_N)\) is a random sample from the population of \( Y_{r=1,a=A^{r=1}}, A^{r=1} \). The empirical likelihood \( L_E \) of the parameters \((\pi_c, \mu^n, \mu^{c1}, \mu^{c0})\) is the maximum likelihood for multinomial distributions \((q_1, \ldots, q_{n_0})\) on \((Y_1, A_1), \ldots, (Y_{n_0}, A_{n_0})\) and \((q_{n_0+1}, \ldots, q_N)\) on \((Y_{n_0+1}, A_{n_0+1}), \ldots, (Y_N, A_N)\) that are consistent with \((\pi_c, \mu^n, \mu^{c1}, \mu^{c0})\):

\[
L_E(\pi_c, \mu^n, \mu^{c1}, \mu^{c0}) = \max \left( \prod_{i=1}^{n_0} q_i \right) \left( \prod_{i=n_0+1}^{N} q_i \right) \tag{5}
\]

subject to

\[
\sum_{i=1}^{n_0} q_i = 1, \quad \sum_{i=n_0+1}^{N} q_i = 1 \tag{6}
\]

\[
q_i \geq 0, \quad i = 1, \ldots, N \tag{7}
\]

\[
\sum_{i=n_0+1}^{N} q_i A_i = \pi_c \tag{8}
\]

\[
\sum_{i=n_0+1}^{N} q_i Y_i A_i = \mu^{c1} \pi_c \tag{9}
\]
\[ \sum_{i=n_0+1}^{N} q_i Y_i (1 - A_i) = \mu^n (1 - \pi_c) \]  
\hspace{0.5cm} (10)

There exist \( p_i^{c_0}, p_i^n, i = 1, \ldots, n_0 \) such that:

\[ \pi_c p_i^{c_0} + (1 - \pi_c) p_i^n = q_i \]  
\hspace{0.5cm} (12a)

\[ \sum_{i=1}^{n_0} p_i^{c_0} = \sum_{i=1}^{n_0} p_i^n = 1 \]  
\hspace{0.5cm} (12b)

\[ p_i^{c_0}, p_i^n \geq 0, i = 1, \ldots, n_0 \]  
\hspace{0.5cm} (12c)

\[ \sum_{i=1}^{n_0} p_i^n (Y_i - \mu^n) = 0 \]  
\hspace{0.5cm} (12d)

\[ \sum_{i=1}^{n_0} p_i^{c_0} (Y_i - \mu^{c_0}) = 0 \]  
\hspace{0.5cm} (13)

Note that throughout our paper, we will follow Owen (2001), Ch. 2.3 and regard tied data values \( Y_i, Y_j \) as representing distinct outcomes in the empirical likelihood as this simplifies calculations and does not affect inferences. The \( p_i^{c_0} \) and \( p_i^n \) in (12a)-(13) represent the population probabilities that a complier assigned to the control and a never taker assigned to the control has the same outcome as subject \( i \) respectively. The conditions (12a)-(13) involving the \( p_i^{c_0} \) and \( p_i^n \) encode the restrictions on the distribution of \( Y_i | R = 0 \) that come from it being a mixture of the compliers and never takers under assumptions 1-6, see (R1)-(R3).

The MELE of \( (\pi_c, \mu^n, \mu^{c_1}, \mu^{c_0}) \) is \( \arg \max_{\pi_c, \mu^n, \mu^{c_1}, \mu^{c_0}} L_E(\pi_c, \mu^n, \mu^{c_1}, \mu^{c_0}) \). To ease the computational burden of computing the MELE, we do not maximize over \( \mu^n \), but instead use the method of moments estimate \( \hat{\mu}^n = \frac{\sum_{i=1}^{N} Y_i R_i (1 - A_i)}{\sum_{i=1}^{N} R_i (1 - A_i)} \) and maximize \( L_E(\pi_c, \hat{\mu}^n, \mu^{c_1}, \mu^{c_0}) \) over \( (\pi_c, \mu^{c_1}, \mu^{c_0}) \). In the model (4), this approximate maximum empirical likelihood estimator (AMELE) of the CACE performed almost as well as the MELE – the MSE of the AMELE was 0.051 compared to 0.048 for the MELE. We now present an algorithm for finding the AMELE.

**3.2 Computation for Empirical Likelihood Approach**

To find the AMELE, we conduct a grid search over \( \pi_c \), finding \( \max_{\mu^{c_1}, \mu^{c_0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \mu^{c_1}, \mu^{c_0}) \) over a grid of \( \hat{\pi}_c \) from 0 to 1. As we will see below, \( \arg \max_{\mu^{c_1}} L_E(\hat{\pi}_c, \hat{\mu}^n, \mu^{c_1}, \mu^{c_0}) \) does not depend on \( \mu^{c_0} \) and \( \arg \max_{\mu^{c_0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \mu^{c_1}, \mu^{c_0}) \) does not depend on \( \mu^{c_1} \), so finding the max-
imizing $\mu^{c_1}$ and $\mu^{c_0}$ can be done separately. For finding the maximizing $\mu^{c_1}$, we note that

$$\arg\max_{\mu^{c_1}} L_E(\tilde{\pi}_c, \hat{\mu}^{n}, \mu^{c_1}, \mu^{c_0}) = \arg\max_{\mu^{c_1}} \prod_{i} q_i \text{ subject to} \ (i) \ \sum_{i:R_i=1,A_i=1} q_i = \tilde{\pi}_c, \ (ii) \ q_i \geq 0, i = 1, \ldots, N \text{ and (iii) } \sum_{i:R_i=1,A_i=1} q_i Y_i = \tilde{\pi}_c \mu^{c_1}.$$  

By multiplying the $q_i$’s by $\frac{1}{\tilde{\pi}_c}$, we see that finding $\arg\max_{\mu^{c_1}} L_E(\tilde{\pi}_c, \hat{\mu}^{n}, \mu^{c_1}, \mu^{c_0})$ is equivalent to finding the MELE of the mean of the population of $Y_{1|C=1}$ based on the random sample $Y_1, \ldots, Y_n|A_i = 1, R_i = 1$; consequently, $\arg\max_{\mu^{c_1}} L_E(\tilde{\pi}_c, \hat{\mu}^{n}, \mu^{c_1}, \mu^{c_0})$ is the mean of $Y_1, \ldots, Y_n|A_i = 1, R_i = 1$ (see Theorem 2.1, Owen, 2001). Thus, our estimate of $\mu^{c_1}$ is $\hat{\mu}^{c_1} = \left(\sum_{i=1}^{N} R_i A_i\right)^{-1} \sum_{i=1}^{N} Y_i R_i A_i$. For finding our estimate of $\mu^{c_0}$, let $(q_1^*, \ldots, q_{n_0}^*) = \arg\max_{q_1, \ldots, q_{n_0}} \prod_{i} q_i \text{ subject to} \ (6), (7), (11) \text{ and (12a)-(12d)}$ with $\mu^{n} = \hat{\mu}^{n}, \pi_c = \tilde{\pi}_c$. We have that

$$\arg\max_{\mu^{c_0}} L_E(\tilde{\pi}_c, \hat{\mu}^{n}, \mu^{c_1}, \mu^{c_0}) = \frac{\sum_{i=1}^{n_0} q_i^* Y_i - (1 - \tilde{\pi}_c)\hat{\mu}^{n}}{\tilde{\pi}_c}, \quad (14)$$

where we use the fact that for the $\mu^{c_0}$ that satisfies $\sum_{i=1}^{n_0} q_i^* Y_i = \tilde{\pi}_c \mu^{c_0} + (1 - \tilde{\pi}_c)\hat{\mu}^{n}$, the constraints (11) and (12a)-(13) are satisfied for $q_1 = q_1^*, \ldots, q_{n_0} = q_{n_0}^*$. Thus, to find $\arg\max_{\mu^{c_0}} L_E(\tilde{\pi}_c, \hat{\mu}^{n}, \mu^{c_1}, \mu^{c_0})$, we just need to find $q_1^*, \ldots, q_{n_0}^*$. To do this, we note that we can view $(q_1^*, \ldots, q_{n_0}^*)$ as the MLE of the category probabilities for the sample $Y_1, \ldots, Y_{n_0}$ from an iid multinomial model with categories $Y_1, \ldots, Y_{n_0}$, corresponding category probabilities $q_1, \ldots, q_{n_0}$ and parameter restrictions given by (6), (7), (11) and (12a)-(12d) with $\mu^{n} = \hat{\mu}^{n}, \pi_c = \tilde{\pi}_c$. Finding the MLE directly is challenging because of the complex parameter restrictions (11) and (12a)-(12d). However, consider using the EM algorithm where we regard each subject’s compliance class as “missing data.” We can reexpress the observed data likelihood $\prod_{i=1}^{n_0} q_i$ and the parameter restrictions (6), (7), (11) and (12a)-(12d) in terms of $p_i^{c_0}, p_i^{n}$ (see Appendix A for details). We can then use the EM algorithm to find the $p_i^{c_0}, p_i^{n}$ to maximize the observed data likelihood and then find the corresponding maximizing $q_i$’s by (12a). The complete data likelihood is $\prod_{i:R_i=0,C_i=1} p_i^{c_0} \prod_{i:R_i=0,C_i=0} p_i^{n}$. Because the complete data follows an exponential family distribution, the E step has a closed form expression. The $M$ step involves a calculation that is analogous to finding the empirical likelihood for the mean (Owen, 1988); convex duality enables us to avoid maximizing over $p_i^{c_0}, p_i^{n}, i = 1, \ldots, n_0$.
and instead maximize over a single variable. The tractability of both the $E$ and $M$ steps makes the EM algorithm with each subject’s compliance class as missing data easy to use for finding $q_1^*, \ldots, q_{n_0}^*$ and hence finding $\arg \max_{\mu^{c_0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \mu^{c_1}, \mu^{c_0})$ by (14). Further details on the EM algorithm are provided in Appendix A and a technical report available from the authors.

Note that given $q_i, i = 1, \ldots, n_0$, there are typically more than one set of $p_i^{c_0}, p_i^n, i = 1, \ldots, n_0$ that satisfy the constraints (12a)-(12d). This is illustrated in Figure 2, which for a sample from the model $G^2$ that is described in Section 4, shows weighted histograms (Tan, 2004) of the $p_i^{c_0}, p_i^n$ and $q_i$ that the EM algorithm converges to for two different sets of starting values of $p_i^{c_0}, p_i^n$. We can see that although the EM algorithm converges to different values of $p_i^{c_0}, p_i^n$ for the different sets of starting values, the corresponding $q_i$’s that the EM algorithm converges to are the same. This is always the case:

**Lemma 1** Regardless of the starting values for the $p_i^{c_0}, p_i^n, i = 1, \ldots, n_0$, the sequence of estimates of $q_i$ from the EM algorithm converges to the global maximum of the likelihood $\prod_{i=1}^{n_0} q_i$ subject to the restrictions (6), (7), (11) and (12a)-(12d) with $\mu^n = \hat{\mu}^n, \pi_c = \tilde{\pi}_c$.

The proof of Lemma 1 is outlined in Appendix B.

In summary, we estimate $\pi_c, \mu^n, \mu^{c_1}, \mu^{c_0}$ as follows:

1. $\hat{\mu}^n$ is the sample mean of $Y|R = 1, A = 0$.
2. $\hat{\mu}^{c_1}$ is the sample mean of $Y|R = 1, A = 1$.
3. We find for a grid of $\tilde{\pi}_c$ the MELE of $\mu^{c_0}$ given $\pi_c = \tilde{\pi}_c$, $\mu^n = \hat{\mu}^n$, $\mu^{c_1} = \hat{\mu}^{c_1}$ using the EM algorithm described above. Then $\hat{\pi}_c = \arg \max_{\tilde{\pi}_c} \max_{\mu^{c_0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \hat{\mu}^{c_1}, \mu^{c_0})$ and $\hat{\mu}^{c_0} = \arg \max_{\mu^{c_0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \hat{\mu}^{c_1}, \mu^{c_0})$.
4. Our AMELE estimate of the CACE is $\hat{\text{CACE}}_{\text{AMELE}} = \hat{\mu}^{c_1} - \hat{\mu}^{c_0}$.

A program for computing $\hat{\text{CACE}}_{\text{AMELE}}$ is available from the authors.

**3.3 Estimation in Trials in Which the Assigned to Control Group Can Access the Treatment**
We have illustrated our method for single consent design trials, but our method can be directly applied to more general trials under Assumptions 1-6 in which the control group can access the treatment. For such trials, we have one more compliance class, the always takers, in addition to the compliers and never takers (see Table 3); we denote the proportion of always takers and the mean of always takers’ potential outcomes by $\pi_a$ and $\mu^a$ respectively. The empirical likelihood $L_E$ of the parameters $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{cl}, \mu^{co})$ is the maximum likelihood for multinomial distributions $(q_1, \ldots, q_{n_0})$ on $(Y_1, A_1), \ldots, (Y_{n_0}, A_{n_0})$ and $(q_{n_0+1}, \ldots, q_N)$ on $(Y_{n_0+1}, A_{n_0+1}), \ldots, (Y_N, A_N)$ that are consistent with $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{cl}, \mu^{co})$ and the restrictions on the parameter space specified by Assumptions 1-6, namely $L_E(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{cl}, \mu^{co}) = \max(\prod_{i=1}^{n_0} q_i \left(\prod_{i=n_0+1}^{N} q_i\right))$ subject to (i) $\sum_{i=1}^{n_0} q_i = 1; \sum_{i=n_0+1}^{N} q_i = 1$; (ii) $q_i \geq 0, i = 1, \ldots, N$; (iii) $\sum_{i: R_i=1, A_i=0} q_i = 1 - \pi_a - \pi_c$; (iv) $\sum_{i: R_i=0, A_i=1} q_i = \pi_a$; (v) $\sum_{i: R_i=1, A_i=0} q_i Y_i = \mu^n (1 - \pi_a - \pi_c)$; (vi) $\sum_{i: R_i=0, A_i=1} q_i Y_i = \mu^a \pi_a$; (vii) There exist $p_i^0, p_i^n$ for the $i$ with $R_i = 0, A_i = 0$ such that (viia) $\frac{\pi_c}{1-\pi_c} p_i^0 + \frac{\pi_a}{1-\pi_a} p_i^n = q_i$, (viib) $\sum p_i^0 = \sum p_i^n = 1$, (viic) $p_i^0, p_i^n \geq 0$, (viid) $\sum p_i^n (y_i - \mu^n) = 0$ and (viie) $\sum p_i^0 (Y_i - \mu^{co}) = 0$; and (viii) There exist $p_i^{cl}, p_i^a$ for the $i$ with $R_i = 1, A_i = 1$ such that (viia) $\frac{\pi_c}{\pi_c + \pi_a} p_i^{cl} + \frac{\pi_a}{\pi_c + \pi_a} p_i^a = q_i$, (viib) $\sum p_i^{cl} = \sum p_i^a = 1$, (viic) $p_i^{cl}, p_i^a \geq 0$, (viid) $\sum p_i^a (Y_i - \mu^a) = 0$ and (viie) $\sum p_i^{cl} (Y_i - \mu^{cl}) = 0$.

As with the single consent design, rather than finding the MELE of $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{cl}, \mu^{co})$, we find the approximate MELE by setting $\mu^n$ equal to the sample mean of $Y|R = 1, A = 0$ (these are the known never takers in the sample) and $\mu^a$ equal to the sample mean of $Y|R = 0, A = 1$ (these are the known always takers in the sample), and then maximizing the empirical likelihood over $(\pi_c, \pi_a, \mu^{cl}, \mu^{co})$. This can be done by using the EM algorithm for estimating $\mu^{co}$ in the $Y|R = 0, A = 0$ sample as in Section 3.2 and an analogous EM algorithm for estimating $\mu^{cl}$ in the $Y|R = 1, A = 1$ sample. The details are provided in the technical report available from the authors.

4 Simulation Studies

In Section 3, we introduced our approximate maximum empirical likelihood estimator (AMELE) as an alternative to the standard IV estimator and Imbens and Rubin’s parametric estimator. To understand the differences among these three estimators, we conduct a
simulation study. We consider single consent design trials as discussed in Section 3.1. We set \( \pi_c = 0.5 \) and compare the three estimators under different outcome distributions and under sample sizes of \( N = 100 \) and \( N = 500 \) with \( pr(R = 1) = 0.5 \). The outcome distributions we consider are Normal (N), Gamma (G), and Log Normal (LN) distributions. For each outcome distribution, we set \( \mu^c = 2, \mu^d = 1 \), so the CACE = \( \mu^c - \mu^d = 1 \). The variances are fixed at 1.

Before explaining our settings of \( \mu^n \), we discuss the impact of the distance between \( \mu^n \) and \( \mu^d \) on the efficiency of the AMELE relative to standard IV. The distance between \( \mu^n \) and \( \mu^d \) is a measure of the separation between the distributions of the compliers and never takers under the control. To see the impact of the distance between \( \mu^n \) and \( \mu^d \), we consider under what conditions are the AMELE and standard IV estimates equal. Standard IV estimates the CACE by substituting method of moments estimates into (3). The AMELE estimates the CACE by substituting maximum empirical likelihood estimates into (3) conditional on \( E(Y|R = 1, A = 0) \) being set equal to its method of moments estimate. The AMELE equals the standard IV estimate if the method of moments estimates of \( \hat{P}(A = 1|R = 1) \) and \( \hat{E}(Y|R = 1, A = 0) \), denoted by \( \hat{P}(A = 1|R = 1) \) and \( \hat{E}(Y|R = 1, A = 0) \) respectively, satisfy (11) and (12a)-(12d) with \( q_i = \frac{1}{n_0} \) for \( i = 1, \ldots , n_0 \). This will happen if and only if \( \hat{\mu}^n \) is between the trimmed mean of \( Y|R = 0 \) over the 0 to \( (1 - \hat{P}(A = 1|R = 1)) \) quantiles and the trimmed mean of \( Y|R = 0 \) over the \( \hat{P}(A = 1|R = 1) \) to 1 quantiles. It is more likely that \( \hat{\mu}^n \) will escape these bounds when the distributions of the compliers and the never takers are more separated. When \( \hat{\mu}^n \) does escape these bounds, we expect that the AMELE will provide a better estimate than standard IV because the AMELE is taking better account of the mixture structure of outcomes implied by Assumptions 1-6. Thus, we expect that the AMELE will gain more efficiency over standard IV when the distance between \( \mu^n \) and \( \mu^d \) is greater, because then the distributions of the compliers and never takers under the control are more separated.

To see the effect of the separation between the compliers and never takers under the control, we chose two sets of values for \( \mu^d \) and \( \mu^n \) such that the distributions of the compliers
and never takers under the control are well separated under one set of values but are close to each other under another set of values. In the first setting, \( N^1, G^1 \) and \( LN^1 \), the distributions of never-takers and compliers under control are well separated, that is,

\[
\begin{pmatrix}
Y_{i}^{1,1}|C_{i} = 1 \\
Y_{i}^{0,0}|C_{i} = 1 \\
Y_{i}^{0,0} = Y_{i}^{1,0}|C_{i} = 0
\end{pmatrix}
\sim N, G, \text{ or } LN \text{ with (mean, variance) }
\begin{pmatrix}
(2, 1) \\
(1, 1) \\
(3, 1)
\end{pmatrix}
\text{, respectively.}
\]

In the second setting, \( N^2, G^2 \) and \( LN^2 \), the distributions of never-takers and compliers under control are close to each other, that is,

\[
\begin{pmatrix}
Y_{i}^{1,1}|C_{i} = 1 \\
Y_{i}^{0,0}|C_{i} = 1 \\
Y_{i}^{0,0} = Y_{i}^{1,0}|C_{i} = 0
\end{pmatrix}
\sim N, G, \text{ or } LN \text{ with (mean, variance) }
\begin{pmatrix}
(2, 1) \\
(1, 1) \\
(1.5, 1)
\end{pmatrix}
\text{, respectively.}
\]

For each setting, we present summary results over 1000 replications with sample sizes of 100 and 500. Table 4 shows the bias and MSE from the three different estimators for the CACE for the different settings considered.

We see from Table 4 that

1) The parametric estimator based on the normality assumption is unbiased and efficient under the true normal distributions, but is biased (23% – 40% bias) under non-normal distributions. The parametric estimator is more efficient than standard IV and AMELE for the normal distribution settings, but is less efficient than AMELE for all the non-normal settings considered and is less efficient than standard IV for most of the non-normal settings (in particular, all of the non-normal settings with a sample size of 500).

2) Both the AMELE and standard IV have low bias for all settings considered. The AMELE has < 5% bias when the distributions of the never takers and the compliers under the control are close to each other. When the distributions of the never takers and compliers under the control are well separated and the sample size is 100, the AMELE has around 10% bias but this bias drops to < 5% when the sample size increases to 500.
3) The AMELE is more efficient than standard IV for all settings considered. The gain in MSE from the AMELE estimator is more substantial when the distributions of never-takers and compliers under the control are well separated (as expected from the discussion above). The gain in MSE is as large as 56%. The gain is generally smaller with a sample size of 500 than 100. In additional simulations not presented in Table 4, we found that there is still a gain in MSE from the AMELE over standard IV with a sample size of 1000, e.g., the AMELE still has about a 20% gain in MSE in $G^1$ with $N = 1000$.

We also did a simulation study for the setting of Section 3.2 in which the assigned to control group can access the treatment. The results are not presented, but are available from the authors. The pattern of results is similar as that for the single consent design trials. In particular, (a) the parametric estimator shows some bias when the true distributions are not normal; (b) the AMELE is more efficient than the standard IV estimator, and the gain in efficiency is more substantial when the never takers and compliers’ distributions under the control are well separated and the always takers and compliers’ distributions under the treatment are well separated.

5 Asymptotic Properties

In Section 4, we showed that the AMELE estimator gains over standard IV in a range of finite sample situations, with larger gains when the compliers and never takers’ outcome distributions under the control are more separated. The standard IV estimator is based on estimating the distribution of $(Y, A, R)$ by the empirical distribution of $(Y, A, R)$; the method of moments estimates that standard IV is based on are the moments of the empirical distribution. The source of the AMELE’s gain over standard IV is that the empirical distribution of $(Y, A, R)$ might not satisfy the restrictions given by Assumptions 1-6. The AMELE takes into account these restrictions to provide a better estimate of the distribution of $(Y, A, R)$ than the empirical distribution. However, unless the distribution of $(Y, A, R)$ is “at the boundary” of the restrictions given by Assumptions 1-6, the empirical distribution of $(Y, A, R)$ should satisfy the restrictions with probability converging to 1 as the sample size $N \to \infty$. Consequently, the AMELE will be asymptotically equivalent to the standard IV.
We establish this result in Theorem 1 under condition (15) below. Condition (15) specifies that the distribution of \((Y, A, R)\) is not “at the boundary” of the restriction that the \(Y|R = 0\) is a mixture of the compliers and never takers under the control in the sense that the distributions of the compliers and never takers under the control overlap at least minimally. In the condition (15) below, we let \(F^c_0\) and \(F^n_0\) denote the CDFs of potential outcomes under the control for compliers and never takers respectively and \(G = (\pi_c)F^c_0 + (1 - \pi_c)F^n_0\) denote the CDF of potential outcomes under the control. The condition is:

\[
\frac{1}{1 - \pi_c} \int_{-\infty}^{G^{-1}(1 - \pi_c)} zdG(z) < \int_{-\infty}^{\infty} zdF^n_0(z) = \mu^n,
\]

\[
\mu^n = \int_{-\infty}^{\infty} zdF^n_0(z) < \frac{1}{1 - \pi_c} \int_{G^{-1}(\pi_c)}^{\infty} zdG(z).
\]  

(15)

Condition (15) says that the trimmed mean of the \(\pi_n\) smallest part of the mixture of never takers and compliers is strictly less than the mean of the never takers and that the trimmed mean of the \(\pi_n\) largest part of the mixture of never takers and compliers is strictly greater than the mean of the never takers. Under condition (15), we have

**Theorem 1** Consider a single consent design. Suppose (i) (15) holds, (ii) \(0 < \pi_c < 1\) and (iii) \(\frac{n_n}{n} = d, 0 < d < 1\). Then \(P(\text{CACE}_{AMELE} = \text{CACE}_{SIV}) \to 1\) as \(N \to \infty\).

The proof of Theorem 1 is in Appendix C.

Although Theorem 1 says that the AMELE and standard IV are equivalent asymptotically under condition (15), the simulation study in Section 4 showed that the AMELE can provide substantial gains in practical situations. The gains provided by the AMELE are analogous to the gains provided in estimating a population mean by knowing restrictions on the range of the mean. For example, consider estimating the mean \(\mu\) of a normal distribution \(N(\mu, \sigma^2)\) based on an iid sample \(Y_1, \ldots, Y_N\) when it is known that \(\mu\) is less than or equal to an upper bound \(\mu_U\). If \(\mu\) is reasonably close to \(\mu_U\), then the MLE will gain substantially over the sample mean (the MLE if \(\mu\) was unrestricted) for many sample sizes. But as long as \(\mu\) is less than \(\mu_U\) by any amount, the estimators are equivalent asymptotically because for large enough \(N\), the sample mean is less than \(\mu_U\) with high probability.
6 Application to Depression Study

In this section, we apply our method to analyze a randomized trial of an intervention to improve treatment of depression among depressed elderly patients in primary care practices (Bruce et al., 2004). The encouragement intervention was that a depression care specialist collaborated with the patient’s primary care physician to facilitate adherence to a depression treatment strategy and provide education and assessment to the patient. The control was usual care. This study, called the PROSPECT study, involved 539 depressed patients in 20 primary care practices at three sites followed for six visits: baseline, 4, 8, 12, 18, and 24 months. Each practice was randomized to either intervention (treatment) or usual care (control). For illustrative purposes, we ignore the fact that the trial was a group randomized trial and treat it as a completely randomized trial; for analyses that account for the group randomization, see Small et al. (2007). Compliance with the intervention was categorized as a binary variable, whether or not a patient had seen a depression care specialist in the prior four months of follow-up. Patients in practices randomized to the usual care group (control) did not have access to the depression specialist, so there are only compliers and never-takers in this trial. To see the effects of estimators under different situations, we analyze two outcomes. One is the patients’ Hamilton depression scores measured at 4 months, which take integer values between 0 and 50. A lower value of the outcome means less depression. Another outcome of analysis is the composite anti-depression (CAD) scores among males at one site (Cornell) measured at 12 months. The CAD is a score from 0 to 4 that indicates how much the patient is being treated for depression. A score of 3 or 4 is considered adequate treatment for depression while 1 or 2 means the patient is being treated in some way, but it is not considered an adequate dose.

Table 5 shows the three estimates of the CACE for the Hamilton and CAD outcomes described above. The percentile bootstrap with 1000 resamplings was used to compute approximate 95% confidence intervals. We first consider the Hamilton score at 4 months (second column of Table 5). The scores were observed for 517 subjects and 92.7% of these subjects that were assigned to treatment complied with the treatment. All the CACE es-
timates are negative and the 95% CIs do not include zero, indicating that the intervention has a significant beneficial effect on depression compared to usual care. Comparing the three estimation methods, we first note from the histograms of the outcome in Figure 3 that the outcomes for the never takers and compliers under the treatment are far from normally distributed, suggesting that the parametric estimator based on the normality assumption is probably a biased estimator. The standard IV estimator and the AMELE have very close point estimates and similar 95% CIs (see below for more discussion of the reasons for this similarity). We now consider the outcome of the CAD scores among males at the Cornell site at 12 months (third column of Table 5). The scores were observed for 37 subjects and 75% of these subjects that were assigned to treatment complied with the treatment. The AMELE and standard IV CACE estimates show a significant beneficial effect of the intervention on treating depression while the parametric normal estimate does not show a significant effect. Like for the Hamilton score, the histograms of the outcomes in Figure 4 show that the outcomes from the never takers and compliers under the control are far from normally distributed, suggesting that the parametric estimator based on the normality assumption is a biased estimator. Unlike for the Hamilton score, for the CACE of the intervention on the CAD score, the AMELE has a substantially narrower 95% CI than standard IV.

The greater gain in efficiency of the AMELE compared to standard IV for the CAD study than the Hamilton study is related to three factors. First, the sample size in the $R = 0$ group is smaller for the CAD study, making it more likely that the empirical distribution of $(Y, A, R)$ will not conform to (and deviate more substantially from) the restrictions implied by Assumptions 1-6. Second, the compliance rate among the subjects assigned to treatment is higher for the Hamilton study (93%) than the CAD study (75%), providing less scope in the Hamilton study for the extra information about Assumptions 1-6 used by the AMELE to have an impact. Third, the separation between the never takers’ and compliers’ outcome distributions in the control group is greater for the CAD than the Hamilton – using the estimates of $\mu^n$ and $\mu^c_0$ from substituting method of moments estimates into the population expressions for these quantities in (3), the estimated absolute standardized difference between
the never takers and compliers’ mean in the control group, $|\hat{\mu}^a - \hat{\mu}^c_0|/\hat{S}\hat{D}(Y|R = 0)$, is 2.34 for the CAD compared to 0.72 for the Hamilton. As we have shown in our simulation studies, the AMELE will have a larger gain in efficiency over standard IV when the distributions of the never takers and compliers in the control group are more separated.

7 Conclusions and Discussion

When noncompliance is present in randomized trials, under the IV assumptions 1 – 6, the complier average causal effect (CACE) is an identified parameter which measures the causal effect of receiving treatment for the subgroup of compliers. In this paper, we propose a new empirical likelihood based estimator for the CACE, the AMELE, that takes the full implications of Assumptions 1-6 into better account than the standard IV estimator. In a simulation study and an empirical example of a randomized trial for treating depression, we showed that the AMELE can provide substantial efficiency gains over standard IV for practical finite sample settings. Like standard IV, the AMELE is robust under non-normal distributions. Asymptotically, the AMELE equals standard IV with probability converging to 1 under condition (15).

In this paper, we focus on estimating the CACE in randomized trials with noncompliance. Our method can be extended to observational studies in which a variable $R$ which encourages ($R = 1$) or does not encourage ($R = 0$) a subject to take the treatment is not randomly assigned but is “as good as randomly assigned” (ignorable) conditional on some covariates; such studies are discussed in Abadie (2003) and examples are given in Table 1 of Angrist and Krueger (2001). Specifically, suppose we replace Assumption 2 (Random Assignment) with Assumption 2’ that the encouragement variable $R$ is independent of $Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0}, A_0, A_1$ conditional on a subject’s covariate vector $X$ and the encouragement variables of different subjects are independent. Also, suppose we expand Assumption 3 to Assumption 3’ that $X_i, Y_{i,1}, Y_{i,1}, Y_{i,0}, Y_{i,0}, A_{i,0}, A_{i,1}$ are i.i.d. draws from a superpopulation and expand Assumption 4 (the mean exclusion restriction) to condition on covariates, i.e, let Assumption 4’ be $E(Y^{r,a}|X) = E(Y^{r',a}|X)$ for all $r, r', a, X$. Furthermore, for a single consent design, suppose we consider linear models for the expected
potential outcomes in a compliance class given the covariates and a logistic model for compliance given the covariates, i.e., \( E(Y^{1,1}|C = 1, X) = X'\beta^{c1}, \) \( E(Y^{0,0}|C = 1, X) = X'\beta^{c0}, \) \( E(Y^{1,0}|C = 0, X) = E(Y^{0,0}|C = 0, X) = X'\beta^n \) and \( P(C = 1|X) = \expit(X'\alpha) \) where \( \expit(z) = \frac{e^z}{1+e^z}. \) We include an intercept in the covariate vector \( X. \) Under this model, the CACE for compliers with covariate vector \( X \) is \( X'\beta^{c1} - X'\beta^{c0}. \) Under Assumptions 1, 2', 3', 4', 5 and 6 and the above models for the outcomes and compliance probabilities, we have that the empirical likelihood of \( \alpha, \beta^{c1}, \beta^{c0} \) and \( \beta^n \) is: \( L_E(\alpha, \beta^n, \beta^{c1}, \beta^{c0}) = \max_{q_1, \ldots, q_N} \prod_{i=1}^{N} q_i \) subject to (i) \( \sum_{i=1}^{n_0} q_i = 1, \sum_{i=n_0+1}^{N} q_i = 1; \) (ii) \( q_i \geq 0, i = 1, \ldots, N; \) (iii) \( \sum_{i=n_0+1}^{N} q_i. X_{ij}(A_i - \expit(X_j'\alpha)) = 0, j = 1, \ldots, p; \) (iv) \( \sum_{i=n_0+1}^{N} q_i. A_i. X_{ij}(Y_i - X_j'\beta^{c1}) = 0, j = 1, \ldots, p; \) (v) \( \sum_{i=n_0+1}^{N} q_i(1 - A_i). X_{ij}(Y_i - X_j'\beta^n), j = 1, \ldots, p; \) (vi) There exist \( t_{i1}^{c0}, t_{i1}^{c1}, \) \( i = 1, \ldots, n_0 \) such that (via) \( t_{i1}^{c0} + t_{i1}^{c1} = q_i; \) (vib) \( t_{i1}^{c0}, t_{i1}^{c1} \geq 0; \) (vic) \( \sum_{i=1}^{n_0} t_{i1}^{c0} + \sum_{i=1}^{n_0} t_{i1}^{c1} = 1; \) (vid) \( \sum_{i=1}^{n_0} t_{i1}^{c0}. X_{ij}(1 - \expit(X_j'\alpha)) + \sum_{i=1}^{n_0} t_{i1}^{c1}. X_{ij}(\expit(X_j'\alpha)) = 0; \) (vie) \( \sum_{i=1}^{n_0} t_{i1}^{c0}. X_{ij}(Y_i - X_j'\beta^n) = 0, j = 1, \ldots, p; \) and (vif) \( \sum_{i=1}^{n_0} t_{i1}^{c0}. X_{ij}(Y_i - X_j'\beta^{c0}) = 0, j = 1, \ldots, p. \) Here the \( t_{i1}^{c0} (t_{i1}^{c1}) \) represent the population probabilities that a subject assigned to the control has the same outcome and covariates as subject \( i \) and is a complier (never taker). The above expression for the empirical likelihood builds on Owen’s (2001, Ch. 4) discussion of empirical likelihood for regression models. As in our method of Section 3, we can compute the AMELE by estimating \( \beta^n \) using the \( R = 1, A = 0 \) sample and maximizing the empirical likelihood over \( \alpha, \beta^{c1}, \beta^{c0} \) given \( \beta^n = \hat{\beta}^n. \)

When deriving the AMELE in our paper, we have assumed that the never takers’ (always takers’) mean is the same under assignment to treatment and control (weak exclusion restriction), rather than that the never takers’ (always takers’) entire outcome distribution is the same under assignment to treatment and control (strong exclusion restriction). This makes our estimator appropriate for certain cases where the assignment to treatment changes the variance but not the mean of the never takers’ and or always taker’s outcomes, but may make the gain in efficiency from the empirical likelihood approach not as large as possible when the strong exclusion restriction holds. We are working on adapting our approach to situations in which it is reasonable to assume more equality constraints for the aspect of the.
never-takers or always-takers under \( R = 0 \) distributions and the never-takers or always-takers under \( R = 1 \) distributions than just equality of means.

**APPENDIX**

**Appendix A: Details of the EM Algorithm**

Reexpressing the observed data likelihood \( \prod_{i=1}^{n_0} q_i \) and the parameter restrictions (6), (7), (11) and (12a)-(12d) in terms of \( p_i^0, p_i^n \), we have that the observed data likelihood (with \( \pi_c = \tilde{\pi}_c, \mu^n = \hat{\mu}^n \)) is \( \prod_{i=1}^{n_0} [\tilde{\pi}_c p_i^0 + (1 - \tilde{\pi}_c) p_i^n] \) with parameter restrictions

\[
\begin{align*}
\sum_{i=1}^{n_0} p_i^0 &= \sum_{i=1}^{n_0} p_i^n = 1 \quad (16a) \\
p_i^0, p_i^n &\geq 0, i = 1, \ldots, n_0 \quad (16b) \\
\sum_{i=1}^{n_0} p_i^n (Y_i - \hat{\mu}^n) &= 0, \quad (16c)
\end{align*}
\]

where \( q_i = \tilde{\pi}_c p_i^0 + (1 - \tilde{\pi}_c) p_i^n \) and \( \mu^0 = \sum_{i=1}^{n_0} p_i^0 Y_i \). Note that if \( \hat{\mu}^n \) is such that there is no \( p_i^0, p_i^n \) that satisfies restrictions (16a)-(16c), then our AMELE estimator does not exist; in this case we can modify the AMELE to use \( \hat{\mu}^n \) as the closest point to \( \sum_{i: R_i = 1, A_i = 0} Y_i / \# \{R_i = 1, A_i = 0\} \) (the usual estimate of \( \mu^n \) for the AMELE) such that there exists \( p_i^0, p_i^n, \ i = 1, \ldots, n_0 \) that satisfy restrictions (16a)-(16c). Viewing each subject’s compliance class as missing data, the complete data likelihood is \( \prod_{i: R_i = 0, c_i = 1} p_i^0 \prod_{i: R_i = 0, c_i = 0} p_i^n \).

**E-step:** The expectation of the complete data log-likelihood conditional on observed data and the parameter estimates \( p_i^{0(k-1)} \) and \( p_i^{n(k-1)} \) at the \((k-1)th\) step is

\[
Q^{(k)} = E\left( \sum_{i=1}^{n_0} [C_i (log p_i^0 + log \tilde{\pi}_c) + (1 - C_i) \{log p_i^n + log (1 - \tilde{\pi}_c)\}] | Y_1, \ldots, Y_{n_0}, p_i^{0(k-1)}, p_i^{n(k-1)} \right)
\]

\[
= \sum_{i=1}^{n_0} \left[ W_i^{(k)} (log p_i^0 + log \tilde{\pi}_c) + (1 - W_i^{(k)}) \{log p_i^n + log (1 - \tilde{\pi}_c)\} \right]
\]

where \( W_i^{(k)} = pr^{(k-1)}(C_i = 1|Y_i, R_i = 0, A_i = 0) = \tilde{\pi}_c p_i^{0(k-1)}/[\tilde{\pi}_c p_i^{0(k-1)} + (1 - \tilde{\pi}_c) p_i^{n(k-1)}] \).

**M-step:** We want to maximize \( Q^{(k)} \) over \( p_i^0, p_i^n \) subject to (16a)-(16c) with \( \mu^n = \hat{\mu}^n, \pi_c = \tilde{\pi}_c \).

We do this by conducting a grid search over \( \mu^0 = \sum_{i=1}^{n_0} p_i^0 Y_i \). We now discuss maximizing
Step 1: We show that maximizing $\prod_{i=1}^{n_0} q_i$ subject to \((6), (7), (11)\) and \((12a)-(12d)\) with $\mu^0 = \tilde{\mu}^0$ gives $Q^{(k)}$. We will denote the maximizing values of $\tilde{p}^{(k)}_i, \tilde{p}^n_i$ for $\mu^0 = \tilde{\mu}^0$ by $\tilde{p}^{(0)}_i, \tilde{p}^n_i$. Note that $\tilde{\mu}^0$ is a possible value of $\mu^0$ if and only if

$$\{p^{(0)}_i, i = 1, \ldots, n_0 | \sum_i p^{(0)}_i = 1, p^{(0)}_i \geq 0, \sum_i \tilde{p}^{(0)}_i (Y_i - \mu^0) = 0 \}$$

For such a $\tilde{\mu}^0$, maximizing $Q^{(k)}$ via Lagrange multipliers subject to \((16a)-(16c)\) and $\mu^0 = \tilde{\mu}^0$ gives

$$\tilde{p}^{(0)}_i = \frac{W^{(k)}_i}{(\sum_i W^{(k)}_i) \{1 + \tilde{c}(Y_i - \tilde{\mu}^0)\}}$$

$$\tilde{p}^n_i = \frac{1 - W^{(k)}_i}{(\sum_i (1 - W^{(k)}_i)) \{1 + \tilde{n}(Y_i - \tilde{\mu}^n)\}}$$

where $\tilde{c}$ and $\tilde{n}$ can be determined in terms of $\tilde{\mu}^0$ and $\tilde{\mu}^n$ by

$$0 = \sum_i \tilde{p}^{(0)}_i (Y_i - \tilde{\mu}^0) = \sum_i \frac{W^{(k)}_i (Y_i - \tilde{\mu}^0)}{(\sum_i W^{(k)}_i) \{1 + \tilde{c}(Y_i - \tilde{\mu}^0)\}}$$

$$0 = \sum_i \tilde{p}^n_i (Y_i - \tilde{\mu}^n) = \sum_i \frac{(1 - W^{(k)}_i) (Y_i - \tilde{\mu}^n)}{(\sum_i (1 - W^{(k)}_i)) \{1 + \tilde{n}(Y_i - \tilde{\mu}^n)\}}$$

The rightmost expressions in \((20)\) and \((21)\) are monotonically decreasing in $\tilde{c}$ and $\tilde{n}$ respectively so that a safeguarded zero finding algorithm (e.g., Brent’s method) can be used. Starting points for the zero finding algorithm can be found by noting that since $0 \leq \tilde{p}^{(0)}_i, \tilde{p}^n_i \leq 1$, $\tilde{c} \in (\frac{1 - W^{(k)}_i}{\mu^0 - Y(n_0)}, \frac{1 - W^{(k)}_i}{\mu^0 - Y(1)})$ and $\tilde{n} \in (\frac{1 - (1 - W^{(k)}_i)}{\mu^0 - Y(n_0)}, \frac{1 - (1 - W^{(k)}_i)}{\mu^0 - Y(1)})$, where $Y(n_0) = \max(Y_i | R_i = 0)$ and $Y(1) = \min(Y_i | R_i = 0)$. The $k$th step parameter estimates $\tilde{p}^{(0)}_i, \tilde{p}^n_i$, $i = 1, \ldots, n_0$ are the $\tilde{p}^{(0)}_i, \tilde{p}^n_i$ that correspond to the $\tilde{\mu}^{(0)}$ that maximizes $Q^{(k)}$ over the grid of $\tilde{\mu}^{(0)}$ considered. Note that we can avoid the need to consider the constraint \((17)\) by replacing the logarithm function with the pseudo-logarithm function of Owen (2001, pg. 62) in the definition of $Q^{(k)}$.

Appendix B: Outline of Proof of Lemma 1

The complete proof is provided in a technical report available from the authors. Here we outline the steps in the proof.
\( \hat{\mu}^n, \pi_c = \tilde{\pi}_c \) is a convex optimization problem so that there is a unique global maximum.

Step 2: Our problem involves maximization over a constrained parameter space. Nettleton (1999) shows that under regularity assumptions, the EM algorithm converges to either (a) a stationary point or (b) boundary points in the constrained parameter space at which the likelihood function can be increased only by moving in a direction outside the parameter space. For an unconstrained parameter space, under regularity assumptions, the EM algorithm converges only to points of type (a) (Wu, 1983). We show that even though our parameter space is constrained, under regularity assumptions, the EM algorithm converges only to points of type (a) for our problem.

Step 3: We combine the results in Steps 1 and 2 with results on EM for unconstrained problems of Wu (1983) and Dempster et al. (1977) to prove the lemma.

Appendix C: Proof of Theorem 1

Let \( Z_1, \ldots, Z_{n_0} \) denote the \( Y|R = 0 \) sample. Let \( \hat{\pi}_c^{R=1} \) equal the method of moments estimate of \( \pi_c \) based on the \( R = 1 \) sample, \( \hat{\pi}_c^{R=1} = \#\{R_i = 1, A_i = 1\} / (N - n_0) \). Note that if there exist \( p_t^0, p_t^n \) that satisfy (i) \( \hat{\pi}_c^{R=1} p_t^0 + (1 - \hat{\pi}_c^{R=1})p_t^n = \frac{1}{n_0} \), (ii) \( \sum_{i=1}^{n_0} p_t^n(Z_i - \hat{\mu}^n) = 0 \), (iii) \( \sum_{i=1}^{n_0} p_t^0 = \sum_{i=1}^{n_0} p_t^n = 1 \) and (iv) \( p_t^0, p_t^n \geq 0 \), then the AMELE equals the standard IV estimate and the maximizing values of \( q_i \) are \( q_i = \frac{1}{n_0}, i = 1, \ldots, n_0 \). By considering the minimum and maximum values of \( \sum_{i=1}^{n_0} p_t^n Z_i \) subject to (i), (iii) and (iv) above, we have that there exist \( p_t^0, p_t^n \) that satisfy (i)-(iv) if and only if \( \hat{\mu}^n \in [\mu_l(N), \mu_u(N)] \) where

\[
\mu_l(N) = \sum_{i=1}^{\lfloor k_{n_0} \rfloor} Z(i) \frac{1}{k_{n_0}} + Z(\lfloor k_{n_0} \rfloor+1) \frac{k_{n_0} - \lfloor k_{n_0} \rfloor}{k_{n_0}}
\]

\[
\mu_u(N) = \sum_{i=n_0-\lfloor k_{n_0} \rfloor}^{n_0} Z(i) \frac{1}{k_{n_0}} + Z(n_0 - \lfloor k_{n_0} \rfloor) \frac{k_{n_0} - \lfloor k_{n_0} \rfloor}{k_{n_0}}
\]

\( k_{n_0} = n_0(1 - \hat{\pi}_c^{R=1}) \) and \( \lfloor k \rfloor \) is the greatest integer less than or equal to \( k \). Let \( \tilde{\mu}_l(N) \) and \( \tilde{\mu}_u(N) \) be the trimmed sample means of \( Z_1, \ldots, Z_{n_0} \) trimmed to the \([0, 1 - \pi_c]\) quantiles and \([\pi_c, 1]\) quantiles respectively, i.e., \( \tilde{\mu}_l(N) = \sum_{i=1}^{\lfloor n_0(1-\pi_c) \rfloor} Z(i) \frac{1}{1-\pi_c} + Z(\lfloor n_0(1-\pi_c) \rfloor+1) \frac{n_0(1-\pi_c) - \lfloor n_0(1-\pi_c) \rfloor}{n_0(1-\pi_c)} \).

Then, letting \( G \) denote the CDF of the potential outcomes under the control, we have that
as $N \to \infty$, 

$$
\tilde{\mu}_t(N) \xrightarrow{p} \frac{1}{1 - \pi_c} \int_{-\infty}^{G^{-1}(1 - \pi_c)} zdG(z) = \mu_t^\infty, \\
\tilde{\mu}_u(N) \xrightarrow{p} \int_{G^{-1}(\pi_c)}^{\infty} zdG(z) = \mu_u^\infty
$$

by the properties of trimmed means (J. Shao, *Mathematical Statistics*, Chapter 5). Now we show that $\mu_t(N) \xrightarrow{p} \mu_t^\infty$ and $\mu_u(N) \xrightarrow{p} \mu_u^\infty$ by showing $|\mu_t(N) - \tilde{\mu}_t(N)| \xrightarrow{p} 0$ and $|\mu_u(N) - \tilde{\mu}_u(N)| \xrightarrow{p} 0$ as $N \to \infty$. We have

$$
|\mu_t(N) - \tilde{\mu}_t(N)| \leq \left| \frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c} \right| \max\left( |Z([n_0(1 - \pi_c) + n_0]\frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c}[+1])|, |Z([n_0(1 - \pi_c) - n_0]\frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c}[-1])| \right)
$$

where $[k]$ is the least integer greater than or equal to $k$. The first term on the right hand side of (22) converges in probability to 0 as $N \to \infty$ and the second term converges in probability to a number less than or equal to $\max(|G^{-1}(1 - \pi_c + a)|, |G^{-1}(1 - \pi_c - a)|)$ for any number $a > 0$ (for this, note that $n_0 = dN \to \infty$ as $N \to \infty$ since $d > 0$). This shows that the right hand side (and hence the left hand side) of (22) converges in probability to 0 as $N \to \infty$. Similarly,

$$
|\mu_u(N) - \tilde{\mu}_u(N)| \leq \left| \frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c} \right| \max\left( |Z([n_0\pi_c + n_0]\frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c}[+1])|, |Z([n_0\pi_c - n_0]\frac{1}{1 - \pi_c} - \frac{1}{1 - \pi_c}[-1])| \right)
$$

$$
\xrightarrow{p} 0.
$$

Thus, we conclude that $\mu_t(N) \xrightarrow{p} \mu_t^\infty$ and $\mu_u(N) \xrightarrow{p} \mu_u^\infty$. By the assumption (15) that the distribution of compliers and never takers overlaps, we have $\mu_t^\infty < \mu^u$ and $\mu_u^\infty > \mu^u$. Combining the facts that $\mu_t^\infty < \mu^u < \mu_u^\infty$, $\mu_t(N) \xrightarrow{p} \mu_t^\infty$ and $\mu_u(N) \xrightarrow{p} \mu_u^\infty$ with the fact that $\hat{\mu}^n \xrightarrow{p} \mu^n$ (by the law of large numbers, using that $N - n_0 = (1 - d)N \to \infty$ as $N \to \infty$), we conclude that $P(\mu_t(N) < \hat{\mu}^n < \mu_u(N)) \to 1$ as $N \to \infty$. Thus, $P(C\hat{A}C_{AMELE} = C\hat{A}C_{SIV}) \to 1$ as $N \to \infty$. 

24


Table 1: The relation of observed groups and latent compliance classes

<table>
<thead>
<tr>
<th>$R_i$</th>
<th>$A_i$</th>
<th>$C_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1 (Complier) or 2 (Always-taker)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 (Never-taker) or 3 (Defier)</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0 (Never-taker) or 1 (Complier)</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2 (Always-taker) or 3 (Defier)</td>
</tr>
</tbody>
</table>

Table 2: The relation of observed groups and latent compliance classes in single consent design trials

<table>
<thead>
<tr>
<th>$R_i$</th>
<th>$A_i$</th>
<th>$C_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1 (Complier)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 (Never-taker)</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1 (Complier), 0 (Never-taker)</td>
</tr>
</tbody>
</table>

Table 3: The relation of observed groups and latent compliance classes under Assumptions 1 – 6

<table>
<thead>
<tr>
<th>$R_i$</th>
<th>$A_i$</th>
<th>$C_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1 (Complier) or 2 (Always-taker)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 (Never-taker)</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0 (Never-taker) or 1 (Complier)</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2 (Always-taker)</td>
</tr>
</tbody>
</table>
Table 4: Estimates of the complier average causal effect (true value is 1) in single consent treatment trials

<table>
<thead>
<tr>
<th>Dist’n</th>
<th>N</th>
<th>Bias</th>
<th>MSE</th>
<th>Std. IV</th>
<th>AMELE</th>
<th>Parametric</th>
<th>Std. IV</th>
<th>AMELE</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N¹</td>
<td>100</td>
<td>0.0178</td>
<td>−0.1141</td>
<td>−0.0240</td>
<td>0.3482</td>
<td>0.2003</td>
<td>0.1649</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.0202</td>
<td>−0.0016</td>
<td>−0.0054</td>
<td>0.0679</td>
<td>0.0515</td>
<td>0.0294</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N²</td>
<td>100</td>
<td>0.0150</td>
<td>0.0105</td>
<td>−0.0053</td>
<td>0.1682</td>
<td>0.1604</td>
<td>0.1311</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.0019</td>
<td>0.0020</td>
<td>−0.0062</td>
<td>0.0214</td>
<td>0.0211</td>
<td>0.0186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G¹</td>
<td>100</td>
<td>0.0429</td>
<td>−0.0981</td>
<td>0.2851</td>
<td>0.3697</td>
<td>0.1945</td>
<td>0.2424</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>−0.0060</td>
<td>−0.0212</td>
<td>0.3963</td>
<td>0.0637</td>
<td>0.0529</td>
<td>0.1907</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G²</td>
<td>100</td>
<td>0.0088</td>
<td>−0.0048</td>
<td>0.3390</td>
<td>0.1957</td>
<td>0.1726</td>
<td>0.2311</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.0235</td>
<td>0.0232</td>
<td>0.3765</td>
<td>0.0454</td>
<td>0.0450</td>
<td>0.1561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN¹</td>
<td>100</td>
<td>0.0173</td>
<td>−0.1364</td>
<td>0.2299</td>
<td>0.2277</td>
<td>0.1008</td>
<td>0.1897</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.0177</td>
<td>−0.0266</td>
<td>0.3666</td>
<td>0.0411</td>
<td>0.0235</td>
<td>0.1568</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN²</td>
<td>100</td>
<td>−0.0007</td>
<td>−0.0137</td>
<td>0.2813</td>
<td>0.0670</td>
<td>0.0563</td>
<td>0.1593</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.0126</td>
<td>0.0129</td>
<td>0.2627</td>
<td>0.0120</td>
<td>0.0117</td>
<td>0.0814</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Results from the PROSPECT study

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Hamilton score estimate (95% CI)</th>
<th>Composite anti-depression score estimate (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard IV</td>
<td>−2.55(−4.13, −0.97)</td>
<td>1.86(0.76, 3.14)</td>
</tr>
<tr>
<td>AMELE</td>
<td>−2.54(−4.12, −0.97)</td>
<td>1.60(0.73, 2.40)</td>
</tr>
<tr>
<td>Parametric</td>
<td>−2.82(−4.39, −1.16)</td>
<td>1.41(−0.66, 2.47)</td>
</tr>
</tbody>
</table>
Figure 1: Profile log likelihood for the MLE and standard IV estimates of the CACE for the sample described in Section 3.1
Figure 2: True and weighted histograms of the outcome in the control group in $G^2$.

a1,b1,c1. Weighted histograms with $p_i^{c0}$ as the weight with two sets of different initial values;
b1,b2,c1. Weighted histograms with $p_i^c$ as the weight with two sets of different initial values;
a2,b2,c2. Weighted histograms with $q_i$ as the weight with two sets of different initial values.

Figure 3: Histograms of the Hamilton score in different groups in the PROSPECT study.
a. Histogram in $R = 1, A = 1$ group;
b. Histogram in $R = 1, A = 0$ group;
c. Histogram in $R = 0$ group.
Figure 4: Histograms of the Composite anti-depression score in different groups in the PROSPECT study.

a. Histogram in $R = 1, A = 1$ group;
b. Histogram in $R = 1, A = 0$ group;
c. Histogram in $R = 0$ group.