Bayesian Analysis (2018) **13**, Number 3, pp. 681–701

Posterior Predictive p-Values with Fisher Randomization Tests in Noncompliance Settings: Test Statistics vs Discrepancy Measures

Laura Forastiere*, Fabrizia Mealli†, and Luke Miratrix‡

Abstract. In randomized experiments with noncompliance, one might wish to focus on compliers rather than on the overall sample. In this vein, Rubin (1998) argued that testing for the complier average causal effect and averaging permutation-based p-values over the posterior distribution of the compliance types could increase power as compared to general intent-to-treat tests. The general scheme is a repeated two-step process: impute missing compliance types and conduct a permutation test with the completed data. In this paper, we explore this idea further, comparing the use of discrepancy measures—which depend on unknown but imputed parameters—to classical test statistics and contrasting different approaches for imputing the unknown compliance types. We also examine consequences of model misspecification in the imputation step, and discuss to what extent this additional modeling undercuts the advantage of permutation tests being model independent. We find that, especially for discrepancy measures, modeling choices can impact both power and validity. In particular, imputing missing compliance types under the null can radically reduce power, but not doing so can jeopardize validity. Fortunately, using covariates predictive of compliance type in the imputation can mitigate these results. We also compare this overall approach to Bayesian model-based tests, that is, tests that are directly derived from posterior credible intervals, under both correct and incorrect model specification.

Keywords: posterior predictive p-values (PPPV), permutation testing, noncompliance, principal stratification, complier average causal effects (CACE).

1 Introduction

In randomized experiments, noncompliance arises when the actual treatment received does not correspond to the assigned treatment. With noncompliance, a simple intent-to-treat (ITT) analysis estimates the effect of the assignment and not necessarily the effect of the treatment itself. An alternative to ITT analyses is to focus on the effect of the treatment on compliers, i.e., those who would take the treatment if offered and would not if not (Imbens & Angrist, 1994; Angrist et al., 1996). In this context, researchers

*Department of Statistics, Computer Science, Applications, University of Florence, Italy, forastiere@disia.unifi.it

†Department of Statistics, Computer Science, Applications, University of Florence, Italy, mealli@disia.unifi.it

‡Harvard Graduate School of Education, luke_miratrix@gse.harvard.edu

 https://doi.org/10.1214/17-BA1062

typically estimate the average treatment effect for this subgroup, a quantity commonly called the complier average causal effect (CACE). Identification of the CACE relies on some assumptions (monotonicity and the exclusion restriction) under which a zero average ITT effect is a necessary and sufficient condition for the CACE to be zero. Therefore, under these assumptions, a valid test for the average ITT effect would also be a valid test for the CACE.

Typically, ITT tests ignore observed information on compliance behavior, which suggests we could find more powerful alternatives. Rubin (1998), in this vein, proposed to gain power by incorporating the (incompletely observed) compliance types of the units into the testing procedure. Rubin did this by imputing compliance type for those units where compliance type was unknown and then conducting a Fisher randomization test (Fisher, 1925, 1926, 1935) on the complete data. He then incorporated uncertainty in this process by averaging p-values calculated conditionally on the completed vector of compliance type over the posterior predictive distribution of this vector.

This Bayesian approach to obtaining p-values for testing has its roots in posterior predictive model-checking (Guttman, 1967; Rubin, 1981, 1984), a popular tool where one compares summary or test statistics calculated for observed data to those for synthetic data drawn from the posterior predictive distribution of a hypothesized model. Large differences between these statistics and these distributions is then taken as evidence of model misspecification. Here, if we have generated synthetic data under the null, model misspecification is evidence against the null. This approach is particularly appealing because it can naturally incorporate unknown nuisance parameters by integrating over them to obtain marginal posteriors. Posterior predictive checks can also be extended by replacing classical test statistics with discrepancy measures, that is, summary measures that can depend on the nuisance parameters themselves (Meng, 1994a; Gelman et al., 1996).

In this paper, we explore the general idea of posterior predictive Fisher randomization tests (FRT-PPs) more in depth, and conduct extensive simulation studies to show how these tests play out in practice in randomized experiments with noncompliance. Combining Fisher randomization tests with posterior predictive p-values leads to a sequence of imputation and permutation steps. At each iteration, a test statistic is computed from the data under a permuted assignment vector and an imputed compliance type vector. Rubin (1998) proposed the use of any estimator of the CACE as a classical test statistic. We investigate replacing such test statistics with discrepancy measures. These measures seem promising because they can directly estimate the CACE from the complete data.

We also closely examine the imputation step. Different methods are possible here. In particular, one might impute either under the null or under the alternative. Imposing the null seems to be a natural choice from a testing approach and should protect test validity. Unfortunately, this approach can cost in terms of power. We explore this tension and discuss how to mitigate this cost.

The imputation step, without the permutation step, is nothing more than what would typically be used for the direct estimation of the posterior distribution of the CACE. Credible intervals of this posterior distribution could themselves lead to nomi-

nal p-values. These model-based p-values are less computationally demanding than the posterior predictive p-values obtained by a Fisher randomization test within the imputation step. But perhaps they would also be more sensitive to model misspecification, in particular misspecification of the outcome model. We compare this approach to the permutation approaches under both correct and incorrect model specification.

Finally, we evaluate the benefits of incorporating predictors of compliance type. Predictive covariates help alleviate the impact of model misspecification in the imputation step, especially for the discrepancy-based approach, and thus help address many of the concerns found in our investigations.

The paper is organized as follows. In Section 2 we briefly review randomized experiments with noncompliance and set up our notation. In Section 3, which presents the core testing strategies that we examine in this work, we first introduce Fisher randomization tests using the potential outcomes framework, briefly review posterior predictive checks, and then show how to combine these ideas to calculate posterior predictive p-values in noncompliance settings using either test statistics or discrepancy measures. We describe our simulation studies in Section 4; these are targeted to compare the different choices one might make in the implementation of Section 3's ideas. Simulation results are shown and discussed in Section 5. In Section 6 we compare FRT-PPs to Bayesian model-based tests with an additional simulation. We finally discuss common patterns across our findings, and what they suggest for practice, in Section 7.

2 Randomized Experiments with Noncompliance

We examine two-arm randomized experiments where N units are assigned to treatment ($Z_i^{obs} = 1$) or control ($Z_i^{obs} = 0$), $i = 1, \ldots, N$. We use the potential outcomes framework, originally proposed by Neyman in the context of randomized experiments (Neyman, 1923) and then formalized and extended to observational studies by Rubin (1974; 1978). Let $Y_i(z)$ be the potential outcome we would observe for unit i when it is assigned to treatment level z, and let $Y_i^{obs} = Y_i(Z_i^{obs})$ be the actual observed outcome. The potential outcomes are fixed, pre-treatment quantities. This representation is ensured by the stable-unit-treatment-value assumption (SUTVA, Rubin (1980)), which states that the outcomes of any one unit are unaffected by the treatment assignments of other units and that the treatment is well-defined. The causal effect of the treatment assignment for any unit, referred to as the intent-to-treat (ITT) effect, can then be defined as a comparison between $Y_i(1)$ and $Y_i(0)$. We focus on the average ITT effect across the units, defined as

$$ITT := E[Y_i(1) - Y_i(0)]. \tag{1}$$

Finally, let $\boldsymbol{Y}$, a $n \times 2$ matrix, be the *potential outcome schedule*, represent all potential outcomes $Y_i(0), Y_i(1)$ for $i = 1, \ldots, n$.

In experiments with noncompliance, the effect of the treatment assignment, namely the ITT, differs from the effect of the treatment itself, which is often the effect of interest. Let $D_i(z)$ be an indicator of the treatment that unit i would actually receive

if assigned to z, and let $D_i^{obs} = D_i(Z_i^{obs})$ be the actual treatment received. These $D_i(z)$ are potential behaviors, analogous to the potential outcomes $Y_i(z)$. The compliance type for each unit is then defined by the joint values $D_i(0)$ and $D_i(1)$ (Angrist et al., 1996). This partition of units into compliance types is a special case of principal stratification (Frangakis & Rubin, 2002). Here, $\boldsymbol{D}$, representing $D_i(0), D_i(1)$ for $i = 1, \ldots, n$, is, just as $\boldsymbol{Y}$ above, a fixed, pre-treatment quantity. Even though compliance type is a pre-treatment characteristic it is not generally known for all units, because we can never observe both $D_i(0)$ and $D_i(1)$.

Following Rubin (1998), we focus on the case of one-sided noncompliance, where $D_i(0) = 0$ for all units. With one-sided noncompliance we have two compliance types: "compliers," for whom $D_i(1) = 1$, and "never-takers," for whom $D_i(1) = 0$. In this case, compliance type is unknown only for those units in the control arm, and the typical monotonicity assumption ($D_i(1) \geq D_i(0)$) holds by design. Define a compliance type indicator C_i with $C_i = 0$ for never-takers and $C_i = 1$ for compliers. The complier average causal effect (CACE) is then

$$\tau \equiv CACE := E[Y_i(1) - Y_i(0)|C_i = 1]. \tag{2}$$

In this work, this is our estimand of interest; other estimands such as risk ratios or percent change are also possible.

Assuming the exclusion restriction for never-takers, i.e. assuming $Y_i(0) = Y_i(1)\ \forall i$: $C_i = 0$, the sharp null hypothesis we wish to test is a zero treatment effect for compliers, i.e., $H_0 : Y_i(0) = Y_i(1)\ \forall i$: $C_i = 1$.

In one-sided noncompliance settings under the exclusion restriction, the CACE can be expressed as the ratio between the ITT effect and the probability of being a complier, ITT/π_c, with $\pi_c := Pr(C_i = 1)$, assuming $\pi_c > 0$. Consequentially, a non-zero ITT implies a non-zero CACE and a rejection of a zero ITT would necessarily mean a rejection of a zero CACE.

3 Fisher Randomization Tests and Posterior Predictive *p*-Values

Fisher (1925, 1926, 1935) proposed a model-free technique to test a sharp null hypothesis of zero treatment effect at the unit level for randomized experiments. Although Fisher never used the potential outcomes framework, FRTs can be phrased in terms of potential outcomes. First, under this framework Fisher's sharp null hypothesis H_0 of no treatment effect can be formalized as $Y_i(0) = Y_i(1)\ \forall i$. Fisher's hypothesis is said to be sharp because it allows one to perfectly impute the missing potential outcomes: given the null and an observed outcome $Y_i(Z_i^{obs}) = Y_i^{obs}$, we can exactly impute the missing potential outcome $Y_i(1 - Z_i^{obs})$ as Y_i^{obs}. In other words, Fisher's null hypothesis coupled with the observed data gives us complete information on all the potential outcomes, $\boldsymbol{Y}$.

This allows us to compute Fisher p-values by directly generating the randomization distribution of the test statistic $T(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{Z})$ given our known assignment mechanism and our fully specified (due to the null) $\boldsymbol{Y}$. This distribution, called a *reference distri-*

bution, is the distribution of values of $T(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{Z})$ we could potentially have seen if the null were true. Rubin (1984) gave a Bayesian justification of these Fisher randomization tests (FRTs), based on connecting them to the posterior predictive distribution of the test statistic induced by the random assignment $\boldsymbol{Z}$. Under this formulation, Fisher p-values can be defined as

$$p := Pr\left\{T(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{Z}) \geq T(\boldsymbol{Y}^{obs}, \boldsymbol{Z}^{obs}) \mid \boldsymbol{Z}^{obs}, \boldsymbol{Y}^{obs}, H_0\right\},$$

where $\boldsymbol{Z}$ is a possible treatment vector sampled from the assignment mechanism $p(\boldsymbol{Z})$ and $\boldsymbol{Y}(\boldsymbol{Z})$ is the vector of observed outcomes one would see given $\boldsymbol{Z}$ and $\boldsymbol{Y}$.

Unfortunately, if the null depends on unknown nuisance parameters we cannot directly obtain our reference distribution, because we cannot obtain $\boldsymbol{Y}(\boldsymbol{Z})$ for an arbitrary $\boldsymbol{Z}$. In our context of non-compliance, for example, if our test statistic depends on the compliance types of all units we cannot directly implement an FRT as we have uncertainty as to which units are compliers; the compliance types are our nuisance parameters. However, we can generate posterior predictive p-values by integrating FRT p-values based on full compliance information over a posterior of possible compliance statues for our units. This idea is rooted in the literature of posterior predictive checks, which we review in the following section. We subsequently show how to extend these ideas to obtain posterior predictive p-values using FRTs in the context of non-compliance.

3.1 An Overview of Posterior Predictive Checks

Classical p-values were extended to the Bayesian framework by Guttman (1967) and Rubin (1981, 1984). Under this view, the p-value is a measure of model misfit: low p-values indicate the data are incompatible with the model. The Bayesian view of p-values is particularly appealing when the model has unknown nuisance parameters. While classical methods would typically plug-in a point estimate of the parameter and rely on known reference distributions of pivotal quantities or on asymptotic results, Bayesian tests average over the posterior distribution of the unknown parameters and use the posterior predictive distribution to simulate the reference distribution for any test statistic. We next make this general idea more precise.

Suppose we have a realization $\boldsymbol{Y}^{obs}$ of a random variable $\boldsymbol{Y}$ and we posit a *parametric null model*, $H_0 : \boldsymbol{Y} \sim f(\boldsymbol{Y} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0$. The essence of model assessment then lies in comparing the observed data with hypothetical replicates that could be observed under the assumed model. In particular, given a test statistic $T(\boldsymbol{Y}^{obs})$ where larger values contraindicate the null, the posterior predictive p-value p_T based on this test statistic is

$$\begin{aligned} p_T &= Pr\left\{T(\boldsymbol{Y}) \geq T(\boldsymbol{Y}^{obs}) \mid \boldsymbol{Y}^{obs}, H_0\right\} \\ &= \int Pr\left\{T(\boldsymbol{Y}) \geq T(\boldsymbol{Y}^{obs}) \mid \boldsymbol{Y}^{obs}, H_0, \boldsymbol{\theta}\right\} \times \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, H_0) d\boldsymbol{\theta}. \end{aligned} \tag{3}$$

The inner term, $Pr\{T(\boldsymbol{Y}) \geq T(\boldsymbol{Y}^{obs}) \mid \boldsymbol{Y}^{obs}, H_0, \boldsymbol{\theta}\}$, is the p-value testing $T(\boldsymbol{Y}^{obs})$ under the sharp null indexed by $\boldsymbol{\theta}$. The integral averages these p-values over the poste-

rior of the nuisance parameters, given the overall null hypothesis and data, to give our posterior predictive p-value.

Given a prior distribution $\pi(\boldsymbol{\theta})$, a Monte Carlo simulation-based approach to calculate p_T would draw K values of the parameters, $\{\boldsymbol{\theta}^k;\ k = 1, \ldots, K\}$, from their posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y}^{obs}, H_0)$, simulate replications of the data under the conditional distributions $f(\boldsymbol{Y} \mid \boldsymbol{\theta}^k)$, and compare the new values of the test statistic $T(\boldsymbol{Y})$ with the observed value $T(\boldsymbol{Y}^{obs})$.

Extending this framework, Meng (1994a), and later (Gelman et al., 1996), proposed replacing classical test statistics, $T(\boldsymbol{Y})$, with parameter-dependent statistics, $D(\boldsymbol{Y}, \boldsymbol{\theta})$, referred to as *discrepancy* measures. The posterior predictive p-value based on a discrepancy measure, p_D, is

$$p_D = \int Pr\left\{D(\boldsymbol{Y}, \boldsymbol{\theta}) \geq D(\boldsymbol{Y}^{obs}, \boldsymbol{\theta}) \mid \boldsymbol{Y}^{obs}, H_0, \boldsymbol{\theta}\right\} \times \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, H_0) d\boldsymbol{\theta}. \tag{4}$$

Before $T(\boldsymbol{Y}^{obs})$ was fixed. Now $D(\boldsymbol{Y}^{obs}, \theta)$ varies along with $D(\boldsymbol{Y}, \theta)$.

This approach has two advantages. First, although not the case in this paper, a discrepancy measure often requires smaller computational effort than a test statistic when the test statistic is something like a posterior mode or an maximum likelihood estimate that has to be computed at each resampling step. Discrepancy measures, by contrast, can often be easily calculated because posterior draws of the nuisance parameters themselves are typical byproducts of Bayesian imputation procedures. Second, because of the additional integration over the nuisance parameters, the use of a parameter-dependent statistic directly checks the discrepancy between the data and the reference distribution of a measure under the null hypothesis, and not just between the data and the null hypothesis under the best fit of the model (Gelman et al., 1996; Gelman, 2013).

Meng (1994a) and Robins et al. (2000) derived several results on the frequency evaluation of discrepancy p-values under the null. If $D(\boldsymbol{Y}, \boldsymbol{\theta})$ is a pivotal quantity with known distribution $\mathcal{D}_0$ under the null, then the distribution of the p-value p_D under the null would be uniform. In the more common situation where the discrepancy is not pivotal, the distribution of p-values is no longer uniform. Meng investigated the behavior of such p-values under the prior predictive distribution conditional on the null ($p(\boldsymbol{Y}|H_0) = \int p(\boldsymbol{Y}|\boldsymbol{\theta}, H_0)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$), and found that, under this distribution, discrepancy p-values are centered around 1/2, i.e., $E\{p_D \mid H_0\} = 1/2$, and that $Pr\{p_D \leq \alpha \mid H_0\} \leq 2\alpha$. This means that there are cases in which p-values are conservative and other cases in which they are anti-conservative, but there is a bound for the Type I error of twice the nominal level. Meng's further discussion suggests, however, that in practice the error rates should rarely be this high: because the posterior p-values are stochastically less variable than $U[0, 1]$, we expect the tails to be lighter, leading to conservative tests for low values of α.

Extending this work, Robins et al. (2000) showed that discrepancy-based p-values can be seriously conservative even when the discrepancy measure has asymptotic mean 0 for all values of the nuisance parameters, whereas posterior predictive p-values based on test statistics are conservative whenever the asymptotic mean of the test statistic depends on the parameters. Arguably, a conservative test is not a bad thing per se.

Indeed, such tests are considered valid (Neyman, 1934), as the Type I error would be less than or equal to α. According to Rubin (1996a), the typical conservativeness when using discrepancies, noted by Meng (1994a) and Gelman et al. (1996), arises from the 'extra' information carried by the imputations of $\boldsymbol{\theta}$. This information can be traced to both modeling and structural assumptions used to define the posterior distributions used for the imputation; a fundamental role is played by the fact that the imputations are performed under the null model. This argument can be connected to the one for the potential conservatism of multiple imputation in Rubin (1996b), where these informative imputations are called 'superefficient'.

3.2 Posterior Predictive FRTs (FRT-PPs) with Test Statistics

Rubin (1998) first connected Fisher randomization-based tests and posterior predictive p-values in the context of noncompliance. Let $O(\boldsymbol{Z}) = [\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{D}(\boldsymbol{Z})]$, the observed data, be a function of the assignment vector $\boldsymbol{Z}$, the potential outcomes $\boldsymbol{Y}$, and the potential treatment take-ups $\boldsymbol{D}$. Let $T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$, our test statistic, be a function of the observed data and treatment assignment. In this context, we might use any estimator of the complier average causal effect (CACE). For example, we might use the posterior mean, median, or mode from a Bayesian model, a maximum likelihood estimate, or an instrumental variables estimate. We could also use a test statistic based on ranks or any other quantity that would tend to be larger when the CACE was non-zero. Even though $T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$ depends only on the data we would observe under $\boldsymbol{Z}$, we cannot directly generate a permutation-based reference distribution for $T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$ because the null does not fully specify $\boldsymbol{D}$ and so we do not know what $\boldsymbol{D}(\boldsymbol{Z})$ would be for $\boldsymbol{Z}$ other than our observed $\boldsymbol{Z}^{obs}$; the unknown parts of $\boldsymbol{D}$ are nuisance parameters. If the compliance type of all the units were known, however, then we could calculate $\boldsymbol{D}(\boldsymbol{Z})$ for all values of $\boldsymbol{Z}$ and therefore we could calculate a Fisher p-value, conditional on the compliance types $\boldsymbol{C}$. This gives a conditional p-value of

$$\begin{aligned} p_T(\boldsymbol{C}) &= Pr\left\{T(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{D}(\boldsymbol{Z}), \boldsymbol{Z}) \geq T(\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}) \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{C}, H_0\right\} \\ &= Pr\left\{T(\boldsymbol{Y}^{obs}, \boldsymbol{C}\boldsymbol{Z}, \boldsymbol{Z}) \geq T(\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}) \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{C}, H_0\right\}, \end{aligned}$$

where $\boldsymbol{C}\boldsymbol{Z}$ is the element-wise product of $\boldsymbol{C}$ and $\boldsymbol{Z}$. The last expression follows from two observations: (1) under the sharp null hypothesis and the exclusion restriction, $Y_i(Z_i)$ is always equal to the observed outcome; and (2) $D_i(Z_i) = C_i Z_i$ due to the constraints of our one-sided noncompliance setting.

However, in general we do not know the compliance types of all units. As in posterior predictive checks where p-values are averaged over the posterior distribution of nuisance parameters, Rubin (1998) proposed to average the reference distribution of the test statistic over the posterior predictive distribution of the unknown compliance types, which in turn is an average over the posterior distribution of other unknown parameters $\boldsymbol{\theta}$ of the model used to impute these compliance types. Formally, we define the posterior predictive FRT p-value of a test statistic $T(\cdot)$ as

$$p_T = \int\int p_T(\boldsymbol{C}) \times \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0, \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0) d\boldsymbol{C} d\boldsymbol{\theta}. \tag{5}$$

To estimate p_T we use a markov chain monte carlo (MCMC) approach. For this approach we, for iteration k, first draw the parameters $\boldsymbol{\theta}^k$ from their posterior distribution and then impute the missing compliance types for units in the control arm using the posterior predictive distribution conditional on this draw, i.e., using $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0, \boldsymbol{\theta}^k)$. The parameter vector $\boldsymbol{\theta}$ would typically include both $\boldsymbol{\theta}^C$, parameters for a compliance model, and $\boldsymbol{\theta}^Y$, parameters for an outcome model. We will see this separation more explicitly when discussing implementation in Section 4.3. Regardless, for the permutation step we permute the assignment vector $\boldsymbol{Z}$ and compute the test statistic based on the outcomes and treatment take-ups that would be observed under this new assignment vector, i.e., we calculate $T^k = T(\boldsymbol{Y}^{obs}, \boldsymbol{C}\boldsymbol{Z}, \boldsymbol{Z})$. Our estimate of p_T is then the proportion of iterations where the test statistic T^k is greater than or equal to the observed statistic T^{obs}.

3.3 FRT-PP with Discrepancy Measures

Following Meng (1994a) and Gelman et al. (1996), we can replace parameter-independent test statistics with parameter-dependent discrepancy measures. Rubin (1998) mentioned the possibility of using discrepancies, e.g., the difference-in-means estimate of the effect among compliers, $\overline{Y}_{c1} - \overline{Y}_{c0}$, but he only used test statistics dependent on $O(\boldsymbol{Z})$ in his examples. We extend his work by examining the behavior of discrepancy measures that directly depend on the compliance type, i.e., $D(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{C}, \boldsymbol{Z})$.

The posterior predictive discrepancy-based p-value is

$$p_D = \int\int p_D(\boldsymbol{C}) \times \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0, \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0) d\boldsymbol{C} d\boldsymbol{\theta}, \tag{6}$$

with $p_D(\boldsymbol{C})$ being the compliance-dependent discrepancy-based p-value of

$$\begin{aligned} p_D(\boldsymbol{C}) &= Pr\left\{ D(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{C}, \boldsymbol{Z}) \geq D(\boldsymbol{Y}^{obs}, \boldsymbol{C}, \boldsymbol{Z}^{obs}) \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{C}, H_0 \right\} \\ &= Pr\left\{ D(\boldsymbol{Y}^{obs}, \boldsymbol{C}, \boldsymbol{Z}) \geq D(\boldsymbol{Y}^{obs}, \boldsymbol{C}, \boldsymbol{Z}^{obs}) \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{C} \right\}. \end{aligned}$$

The imputed compliance types for a particular MCMC iteration will affect both the discrepancy measure for the permuted assignment vector under the null hypothesis, $D^k = D(\boldsymbol{Y}^{obs}, \boldsymbol{C}, \boldsymbol{Z})$, and the discrepancy measure for the observed values, $D^{k,obs} = D(\boldsymbol{Y}^{obs}, \boldsymbol{C}, \boldsymbol{Z}^{obs})$, in each iteration. Our estimate of p_D is then the proportion of iterations where $D^k \geq D^{k,obs}$.

3.4 The Performance of Posterior Predictive Testing in Noncompliance Settings

The above formulation prompts several questions about how different testing approaches could perform in practice. We investigate these questions with simulations.

The first question relates to choice of imputation model. The literature on posterior predictive p-values suggests imputing the of the models allowing for such imputation) under the null hypothesis in order to ensure test validity. This would mean that, under a true alternative, the imputation of compliance types would be conducted under the wrong model, which could result in a loss of power to detect a non-zero CACE. The intuition is that if the outcome distribution for compliers and never-takers under treatment are similar, the model could erroneously impute never-taker units in the control group with outcomes close to the treatment group as compliers, resulting in small and insignificant estimates of the complier average causal effect. We investigate this phenomenon in Section 4.

This potential problem motivated us to also investigate imputing the compliance types without imposing the null. This is akin to a "plug-in" style approach in classical testing. However, this relaxation could lead to an increase in Type I error, possibly giving an invalid test. This unconstrained imputation procedure is also not fully Bayesian in that the conditioning set $\{\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}\}$ in the imputation step now differs from the one in the testing step $\{\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, H_0\}$. This is analogous to multiple imputation for missing data: the Bayesian imputation model is typically different (uncongenial) from the model used for analyzing the data, but multiple imputation inferences are valid from both a Bayesian and a frequentist point of view (Meng, 1994b).

The second question is how test statistics and discrepancy measures compare. Past literature mostly focuses on test validity, comparing p-values based on discrepancy measures to those based on classical statistics under the null hypothesis. There appears to be less work concerning power. We explore the trade-off between Type I error and power in noncompliance settings in our simulation study below. In general, we expect discrepancy measures to be relatively more powerful than test statistics when imputing compliance types under a correct imputation model, but we do not know if discrepancy measures are more sensitive to model misspecification.

The third question is how the use of observed covariates that are predictive of compliance type, when available, can improve the imputation step and thus the performance of the overall testing procedure under these different possible approaches.

4 Simulation Study

Our simulation study assesses the rejection rates of different testing procedures under a variety of scenarios in order to answer the questions outlined in the preceding section.

4.1 The Data Generating Process

We generate a population of N=500 units characterized by a single covariate $X_i \sim \mathcal{N}(0,1)$. The compliance type of each unit follows a probit model conditional on X_i:

$$C_i = \mathbb{1}\{\alpha_0 + \alpha_x X_i + \epsilon_i > 0\} \qquad \epsilon_i \sim \mathcal{N}(0,1), \tag{7}$$

where the coefficient vector $\boldsymbol{\alpha} = [\alpha_0, \alpha_x]$ is varied to result in three different levels of predictiveness: none ($\boldsymbol{\alpha} = [-0.8, 0]$), medium ($\boldsymbol{\alpha} = [-1.4, 2]$), and high ($\boldsymbol{\alpha} = [-2.8, 5]$).

The probability of being a complier is then $\Phi(\alpha_0 + \alpha_x X_i)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The values of $\boldsymbol{\alpha}$ were chosen to make the proportion of compliers π_c always 30%.

A unit's outcome follows a normal distribution with unit variance and a mean that depends on its compliance type and the simulation scenario:

$$Y_i(0) \sim \begin{cases} \mathcal{N}(\eta_n, 1) & \text{if } C_i = 0 \\ \mathcal{N}(\eta_{c0}, 1) & \text{if } C_i = 1 \end{cases} \qquad Y_i(1) = \begin{cases} Y_i(0) & \text{if } C_i = 0 \\ Y_i(0) + \tau & \text{if } C_i = 1, \end{cases} \tag{8}$$

with $\tau = 0$ under H_0 and $\tau = 0.5$ under H_1. We fix $\eta_n = 0$ and vary η_{c0} across $\{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. The observed outcome is then $Y_i^{obs} = Y_i(1)Z_i^{obs} + Y_i(0)(1-Z_i^{obs})$. In order to have a better control on the overlap between the distributions of the two potential outcomes for never-takers and compliers, we model the outcome as conditionally independent from the covariate.

Once we generate our units, we randomize $N_T = 250$ units to the treatment with $Z_i^{obs} = 1$ and $N_C = 250$ units to control with $Z_i^{obs} = 0$. The resulting $\boldsymbol{Y}^{obs}, \boldsymbol{X}, \boldsymbol{D}^{obs}$, and $\boldsymbol{Z}^{obs}$ form our simulated observed data.

4.2 Simulation Procedure

We compare test statistic-based p-values in (5) to discrepancy-based p-values in (6). For our test statistic, we use the typical instrumental variables estimator of the CACE (Imbens & Angrist, 1994; Angrist et al., 1996):

$$T(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{D}(\boldsymbol{Z}), \boldsymbol{Z}) = \frac{\widehat{ITT}_Y}{\hat{\pi}_c} = \frac{\overline{\boldsymbol{Y}}_1 - \overline{\boldsymbol{Y}}_0}{\overline{\boldsymbol{D}}_1 - \overline{\boldsymbol{D}}_0}, \tag{9}$$

with $\overline{\boldsymbol{Y}}_z$ and $\overline{\boldsymbol{D}}_z$ being the average outcome and proportion of treatment take-up of units with $Z_i = z$ (note $\overline{\boldsymbol{D}}_0 = 0$ in the case of one-sided noncompliance). For our discrepancy measure, we use the method of moments estimator of the complier average causal effect:

$$D(\boldsymbol{Y}(\boldsymbol{Z}), \boldsymbol{C}, \boldsymbol{Z}) = \overline{\boldsymbol{Y}}_{c1} - \overline{\boldsymbol{Y}}_{c0}, \tag{10}$$

where $\overline{\boldsymbol{Y}}_{cz}$ is the average outcome of the compliers with $Z_i = z$.

We examine four different methods for imputing compliance type:

1. Impute imposing the null hypothesis without including covariates; that is, use the posterior distribution $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0)$.

2. Impute without imposing the null hypothesis and without including covariates; that is, use the posterior distribution $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs})$.

3. Impute imposing the null hypothesis and including covariates; that is, use the posterior distribution $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, H_0, \boldsymbol{X})$.

4. Impute without imposing the null hypothesis and including covariates; that is, use the posterior distribution $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{X})$.

The four imputation methods combined with the use of a test statistic or a discrepancy measure give eight testing methods. These methods were assessed under $3 \times 7 \times 2 = 42$ scenarios defined by three levels of predictiveness of covariates on compliance type, seven levels of difference between the outcome mean under control for compliers and never-takers, and whether we do or do not have a complier average causal effect. We calculated the rejection rate at significance level $\alpha = 0.05$ for each method for each scenario.

In particular, for each of the 42 scenarios, we repeated the following 2000 times:

1. Given the specific values of $\boldsymbol{\alpha}$ and η_{c0} and assuming the null ($\tau = 0$) or the alternative hypothesis ($\tau = 0.5$), generate and randomly assign a sample of N units to treatment and control, resulting in $(\boldsymbol{Y}^{obs}, \boldsymbol{X}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs})$.
2. For each of the four imputation methods, calculate posterior predictive p-values using the test statistic and the discrepancy measure via repeating the following steps $K = 2000$ times and averaging the outputs of the last 1000 iterations (1000 discarded as burn-in):
   (a) Impute the compliance type $\tilde{C}_i$ for units in the control group using the specified imputation method.
   (b) Compute the observed test statistic $T(\boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs})$ and the "observed" discrepancy measure $D(\boldsymbol{Y}^{obs}, \tilde{\boldsymbol{C}}, \boldsymbol{Z}^{obs})$.
   (c) Take a random sample from the set of all possible assignment vectors $\boldsymbol{Z}$.
   (d) Compute reference $T(\boldsymbol{Y}^{obs}, \tilde{\boldsymbol{C}}\boldsymbol{Z}, \boldsymbol{Z})$ and $D(\boldsymbol{Y}^{obs}, \tilde{\boldsymbol{C}}, \boldsymbol{Z})$, based on the imputed compliance types and the sampled assignment vector.
   (e) Compare the reference test statistic and discrepancy measure to their observed values and output 1 if the new value is larger than the observed and 0 otherwise.

Multiple permutations in the inner loop is not required because averaging step (e) over the multiple trials gives the desired overall expectation.

4.3 Implementing the Imputation of Compliance Types

We compute posterior predictive p-values using an MCMC approach. Each iteration is comprised of an imputation and a permutation step. During the imputation step unknown compliance types are drawn from the predictive posterior distribution $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs})$, with additional conditioning on H_0 and covariates $\boldsymbol{X}$ depending on the method used. Because the parameters of this distribution are unknown, the predictive posterior distribution of interest has to be averaged over the posterior distribu-

tion of the parameters. We do this with data augmentation (Tanner & Wong, 1987), a two-stage Gibbs-sampler that samples the model parameters from their full conditional distribution $\mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{C})$ and then samples the vector of compliance types from $\mathrm{p}(\boldsymbol{C} \mid \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{\theta})$.

For imputation we use a joint model formed by combining the model for the compliance type from (7) and the following model for the potential outcomes:

$$\begin{aligned} Y_i(0)|C_i = 1 &\sim \mathcal{N}(\eta_{c0}, \sigma_c) \qquad Y_i(1)|C_i = 1 \quad \sim \mathcal{N}(\eta_{c1}, \sigma_c) \\ Y_i(0)|C_i = 0 &\sim \mathcal{N}(\eta_n, \sigma_n) \qquad Y_i(1)|C_i = 0 \quad \sim \mathcal{N}(\eta_n, \sigma_n). \end{aligned} \tag{11}$$

These means and variances follow from the exclusion restriction. By definition, the difference $\eta_{c1} - \eta_{c0}$ corresponds to the causal estimand τ.

At each iteration, after the parameters are drawn from their posterior, the unknown compliance types for units in the control group are imputed as Bernoulli draws with probabilities

$$Pr(C_i = 1|Y_i^{obs}, Z_i = 0, X_i, \boldsymbol{\theta}) = \frac{\phi(Y_i^{obs}; \eta_{c0}, \sigma_c)p_i}{\phi(Y_i^{obs}; \eta_{c0}, \sigma_c)p_i + \phi(Y_i^{obs}; \eta_n, \sigma_n)(1 - p_i)}, \tag{12}$$

where $p_i = \Phi(\alpha_0 + \alpha_x X_i)$, the prior probability of unit i being a complier, and $\phi(\cdot)$ is the standard normal probability density function. We used conjugate prior distributions on our parameters, that is, we use normal distributions for the means and inverse gamma distributions for the variances:

$$\alpha_0, \alpha_x \sim \mathcal{N}(0, 5) \qquad \eta_{c0}, \eta_{c1}, \eta_n \sim \mathcal{N}(0, 10) \qquad \sigma_c, \sigma_n \sim \mathcal{IG}(0.1, 0.1).$$

For this overall model our parameters are $\theta^C = (\alpha_0, \alpha_x)$ and $\theta^Y = (\eta_n, \eta_{c0}, \eta_{c1}, \sigma_c, \sigma_n)$. For imputation methods 1 and 3, we impose the null hypothesis when imputing the compliance type by assuming $\eta_{c0} = \eta_{c1} = \eta_c$. Similarly, we set $\alpha_x = 0$ for imputation methods 1 and 2, where we do not take covariates into account.

5 Results

To compare validity and power of the 8 Bayesian FRT methods, we computed their rejection rates with level $\alpha = 0.05$ for simulations conducted under the null hypothesis and under the alternative hypothesis, respectively.

5.1 No Predictive Covariate Case

Figure 1 shows results of the simulated scenarios when compliance type does not depend on covariates. For each imputation method and for both the test statistic and the discrepancy measure, we plot the rejection rates against $\eta_{c0} - \eta_n$, the difference between

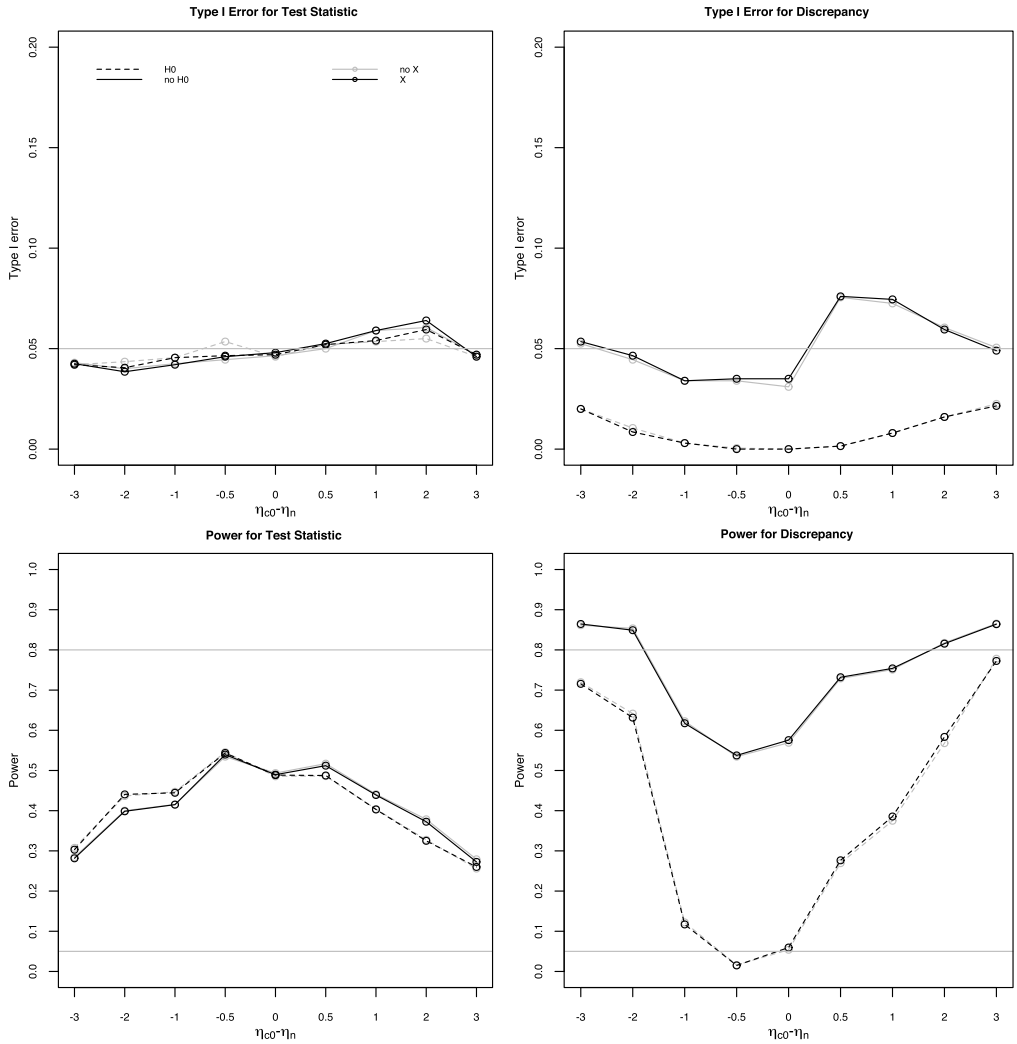

Figure 1: Type I error (top) and power (bottom) at $\alpha = 0.05$ vs. $\eta_{c0} - \eta_n$, for both test statistics (left) and discrepancy variables (right) with zero predictiveness of the covariates. Dashed lines correspond to imputation under the null, solid lines to unconstrained imputation. Black indicates imputing with covariates, grey without.

complier and never-taker means. Generally, as shown by the grey and black lines being essentially the same, imputation with covariates has negligible impact compared to without. This is as expected as the covariate is useless in this setting.

Posterior predictive p-values based on the test statistic are largely unaffected by the choice of imputation method. The test statistic depends on the imputed compliance types through the corresponding proportion of compliers $\hat{\pi}_c$. Because this estimate is

robust to model misspecification,[1] Bayesian tests based on this test statistic appear to have a size around the nominal level for any imputation method and for any distance between compliers and never-takers in the control group. The power of these tests, however, decreases as $|\eta_{c0} - \eta_n|$ gets larger. This occurs because the variance of $\widehat{ITT}_Y$ depends on the difference in outcomes between compliers and never-takers.

Unlike test statistic p-values, discrepancy-based p-values are very sensitive to the compliance imputation method, given that within the permutation test the discrepancy measure assumes the imputed compliance types for every unit are correct.

As expected, when imputing conditional on the null hypothesis, the test is conservative with a Type I error around 0.01. The conservativeness seen in this case is presumably due to the fact that the complete data carries information on the null hypothesis, given that this method imputes compliance types under the correct imputation model (consisting of a correct model specification, the exclusion restriction assumption, and the null hypothesis) (Rubin, 1996a).

When outcomes for compliers and never-takers in the control group are close there is a massive loss of power if we assume the null when imputing. The difficulty in disentangling the mixture between never-takers and compliers when these have similar outcomes, in combination with the exclusion restriction assumption (implying $E\{Y_i(0)|C_i = 0\} = E\{Y_i(1)|C_i = 0\}$) and the null hypothesis (implying $E\{Y_i(0)|C_i = 1\} = E\{Y_i(1)|C_i = 1\}$), leads to an overestimation of the mean η_{c0} for control compliers and, as a consequence, units in the control group imputed as compliers tend to have outcomes close to those for treated compliers.

Imputing without assuming the null partially corrects this phenomenon. Under this alternate imputation strategy, the discrepancy-based FRT-PP yields greater power, outperforming the test-statistic-based approach with an average gain in power of 30%. Unfortunately, this gain in power comes with a substantial price. There is a range of values for $\eta_{c0} - \eta_n$ where the test is invalid.

5.2 Predictive Covariate Cases

We next examine whether improved prediction of compliance type changes these patterns. Figure 2 shows results for the scenario where covariates affect compliance type with a medium (left) and high (right) level of predictiveness. Since our test statistic does not depend on imputed compliance type, performance of statistic-based tests was similar to the previous case (Figure 1) and is therefore not shown. For discrepancy-based tests, incorporating predictive covariates substantially reduces invalidity and increases power. Across the three power subplots (bottom row of Figures 1 and 2) we see the covariate-based approaches' lines steadily moving up as predictiveness increases, while the no-covariate approaches' lines remain essentially the same. In particular, with strongly predictive covariates we can impute under the null, maintaining validity, without substantial sacrifice of power.

[1] Because of randomization, $Pr(C_i = 1) = Pr(D_i = 1|Z_i = 1)$.

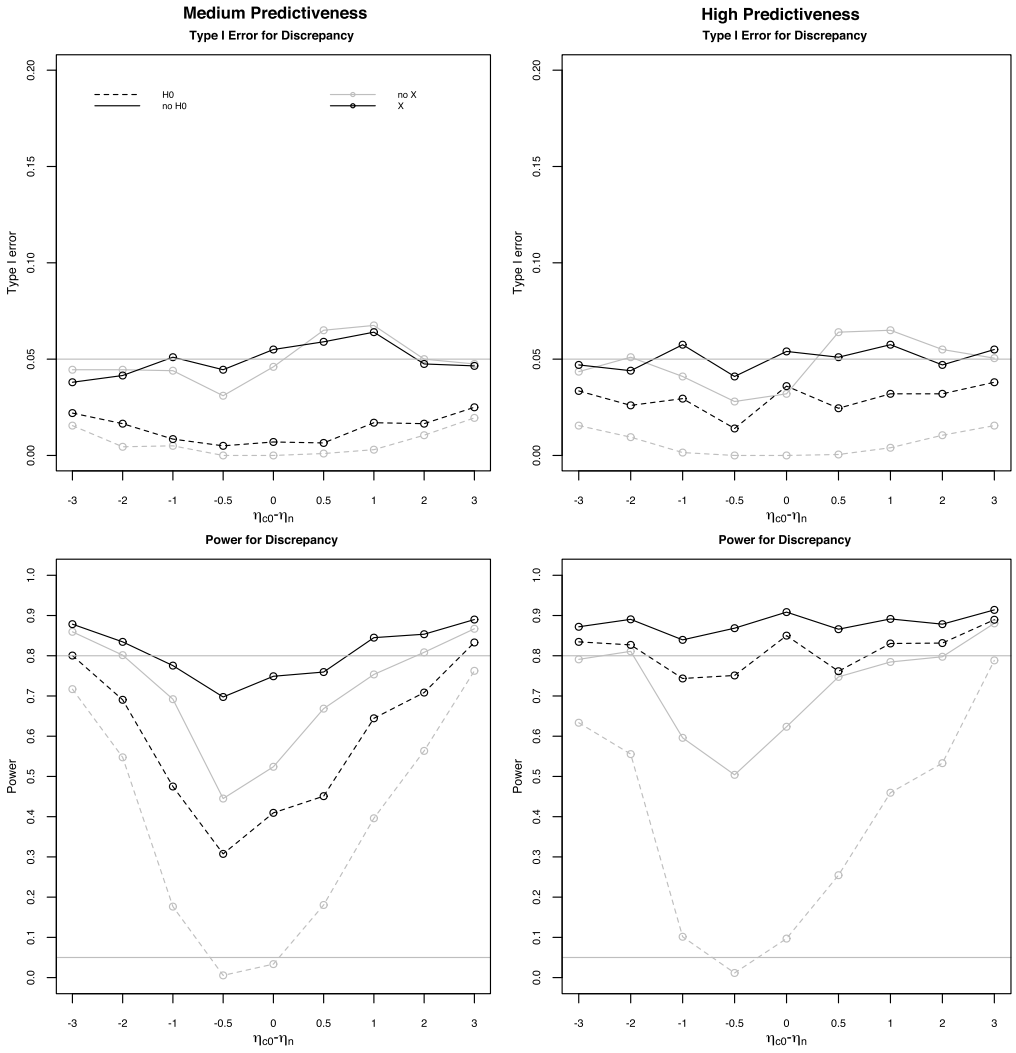

Figure 2: Type I error and power vs $\eta_{c0} - \eta_n$ for medium (left) and high (right) predictiveness scenario. See caption of Figure 1 for explanation of legend.

6 Comparison with Model-Based Tests

The FRT-PP methods based on discrepancies rely heavily on the imputation model. The advantage of Fisher randomization tests as model-free tests is then somewhat lost. Given this, one might think to just rely on the posterior distribution of the parameters. In particular, for the model in (11), an estimate of the parameter difference $\eta_{c1} - \eta_{c0}$ is an estimate of the complier average causal effect τ. Therefore, the posterior distribution of $\eta_{c1} - \eta_{c0}$ could be used directly to test the null hypothesis with a one-sided, model-based p-value of $Pr\{\eta_{c1} - \eta_{c0} \geq 0 | \boldsymbol{Y}^{obs}, \boldsymbol{D}^{obs}, \boldsymbol{Z}^{obs}, \boldsymbol{X}\}$.

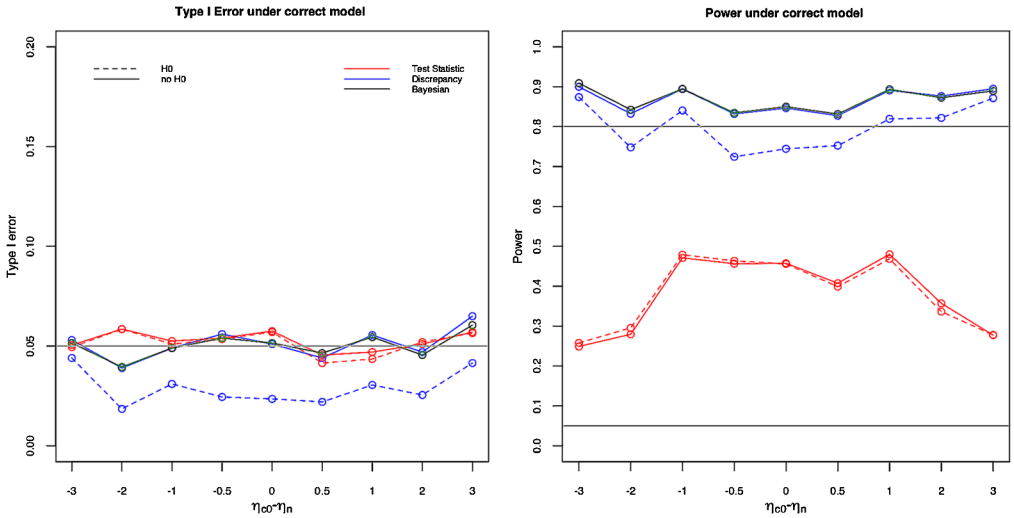

Figure 3: Type I error and power for level of significance $\alpha = 0.05$ against $\eta_{c0} - \eta_n$ with highly predictive covariates, for the Bayesian model-based approach (black), the FRT-PP based on test statistics (red), and the FRT-PP based on discrepancies (blue), all imputing with covariates.

Using these Bayesian model-based p-values to test the null hypothesis is model-dependent, but so is the FRT-PP based on discrepancies. The direct Bayesian approach thus seems superior as it does not require a permutation step, making it less computationally demanding as well as more transparent.

On the other hand, the fully model-based test could be relying more heavily on model assumptions than the FRT-PP approach, given that it depends on good estimation of both η_{c1} and η_{c0}. By contrast, our FRT-PP does not directly depend on the estimation of η_{c1}, and it only uses η_{c0} and η_n indirectly to impute compliance types in the control group. Consequently, we might expect model misspecification to affect model-based p-values more severely. We explore this by conducting several simulation studies under varying degrees of misspecification.

6.1 Correct Model Specification

As an initial investigation, we assessed the performance of the Bayesian model-based approach when the model is correctly specified. We replicated the simulation study in Section 4.1 with high predictiveness of covariates, comparing the direct Bayesian model-based approach to test statistic and discrepancy-based FRT-PP, with covariates included in the imputation model (Methods 3 and 4). Figure 3 shows the rejection rates and power of these different approaches. The characteristics of the Bayesian modeling approach closely track those of the unconstrained discrepancy-based FRT-PP.

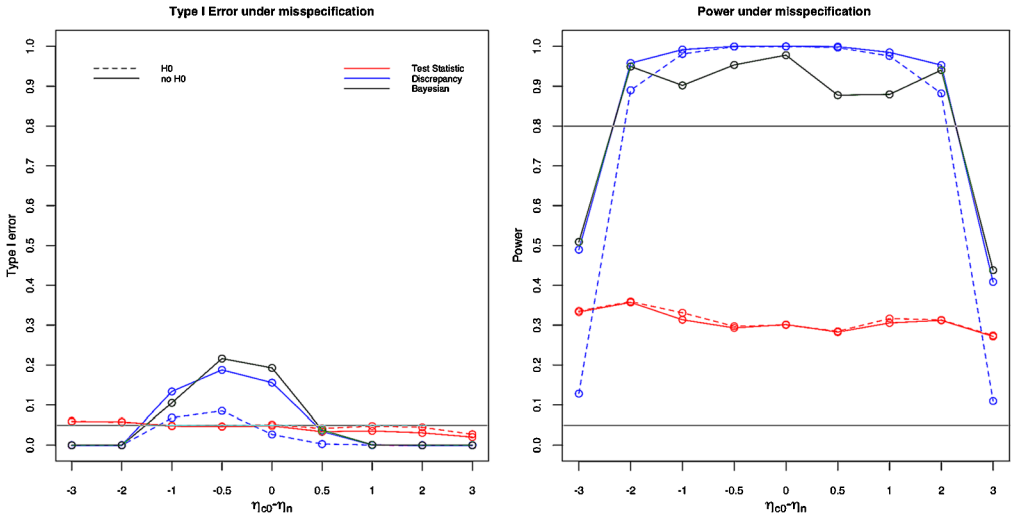

Figure 4: Type I error and power for level of significance $\alpha = 0.05$ against $\eta_{c0} - \eta_n$ for a misspecified model with a skewed outcome distribution. See caption of Figure 3 for explanation of legend.

6.2 Model Misspecification

We next keep the same compliance and outcome models as before, but generate outcome data from non-normal distributions. We then analyze using methods 1 through 4 and the full Bayesian model. In all cases we include our covariate in our estimation procedure. For example, Figure 4 shows Type I error and power in the presence of a skewed outcome generated as

$$Y_i(0) = \sqrt{R_i} \text{ with } R_i \sim \exp\left(\frac{1}{\eta_{C_i}^2}\right),$$
$$Y_i(1) = Y_i(0) + C_i\tau \ ,$$

where $\eta_{C_i} = \eta_{c0}$ if $C_i = 1$ and η_n if $C_i = 0$.

In this skewed outcome case, both the Bayesian model and the unconstrained discrepancy-based FRT-PP approach give large Type I error rates of above 20% when η_{c0} is close to η_n. Imposing the null for the discrepancy-based FRT-PP helps, although the rejection rates are still elevated. That being said, it also has reasonable power, which is encouraging. The FRT-PP using test statistics maintain nominal rates but have low power.

We see similar trends (Figure 5) where, leaving the other aspects of the data generating process the same, we used a mixture of normals as our baseline outcome distribution:

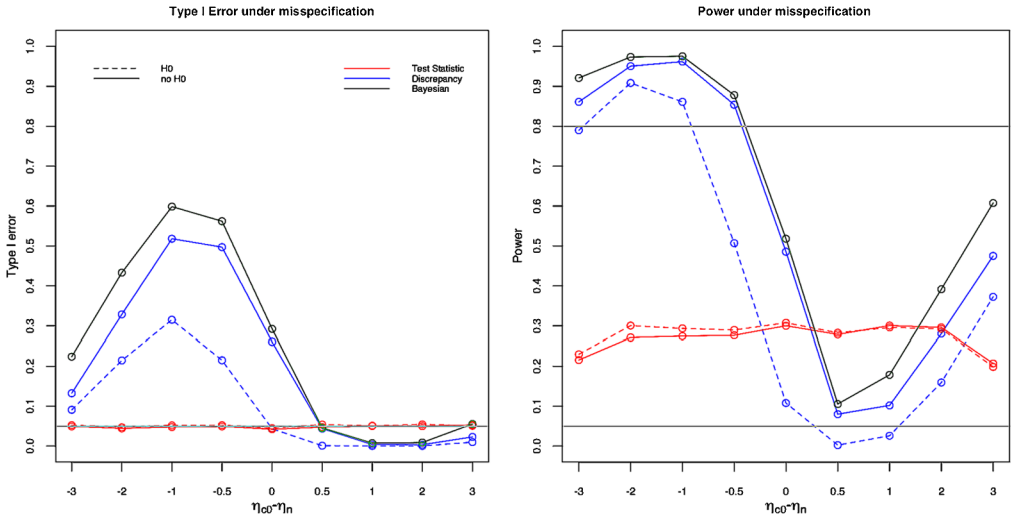

Figure 5: Type I error and power for level of significance $\alpha = 0.05$ against $\eta_{c0} - \eta_n$ for a misspecified model with a normal mixture for the outcome distribution. See caption of Figure 3 for explanation of legend.

$$Y_i(0) \sim \begin{cases} Q_i\mathcal{N}(\eta_n - 1, 1) + (1 - Q_i)\mathcal{N}(\eta_n + 1, 1) & \text{if } C_i = 0 \\ Q_i\mathcal{N}(\eta_{c0} - 1, 1) + (1 - Q_i)\mathcal{N}(\eta_{c0} + 1, 1) & \text{if } C_i = 1 \end{cases} \quad \text{with } Q_i \sim Ber(0.5). \tag{13}$$

For this distribution, the Type I error rate for the *constrained*, discrepancy-based FRT-PP reaches 30%, well above the 5% level. Only the test statistic approach is valid. By comparing the rejection rates for the null and alternative, we see how rejection is largely driven by the structure of the mean outcomes on the control side: When the compliers have a lower mean under control than the never-takers, the rejection rate is higher regardless of effect.

Similar patterns were repeated across many different data generating processes. Overall, inference using the Bayesian model closely tracks that of the discrepancy-based FRT-PP with the null not imposed. The discrepancy-based FRT-PP with the null imposed often has some improvement in terms of validity, but it also tends to have lower power. The test statistic approaches are quite robust, keeping nominal Type I error under all scenarios, but this comes at a substantial loss in power.

Overall, our results suggest that a joint strategy of model-based imputation followed by a permutation test on the completed data can provide more robust alternatives to a full Bayesian approach. While the unconstrained, discrepancy-based FRT-PP generally performs similarly to the full Bayesian approach, the other FRT-PP methods are generally more valid. They tend to have substantially lower power under many alternatives, however, including those with correct model specification. We leave the impact of more complex outcome models to future work.

7 Conclusion

Fisher randomization tests (FRTs) compare the reference distribution of a specified test statistic given a null hypothesis to its observed value. In noncompliance settings, one can impute the unknown compliance types with a Bayesian approach and then conduct such randomization tests with the complete data, but we still need to account for the uncertainty of the imputation. The FRT-PP does this by averaging permutation p-values over the posterior distribution of the unknown compliance types.

We compared the use of discrepancy measures to classical test statistics within this framework. We also examined the behavior of these methods when the null is or is not used in the imputation step. We found that discrepancy-based FRT-PPs that use an imputation method under the null are generally valid testing procedures. Unfortunately, however, this validity can be costly: We found many true alternatives where the systematic misclassification of compliers and never-takers, due to imposing the null, led to a severe reduction in power. One might hope to overcome this limitation by using an unconstrained imputation method that does not impose the null. Unfortunately, this unconstrained imputation can give quite invalid tests, even in scenarios where the model is correctly specified.

By contrast, FRT-PPs with test statistics, while low power, were generally valid: the size of the test stayed close to the nominal levels under the null for all scenarios. Overall, FRT-PPs with test statistics were less affected by the choice of the imputation method when compared to FRT-PPs with discrepancy measures. This is likely because the test statistic only depends on the imputation model through the estimated proportion of compliers, and so model misspecification is irrelevant as long as the proportion of imputed compliers is not significantly compromised.

We also compared the FRT-PP methods to full Bayesian model-based testing. We found that model-based tests had very similar characteristics, both under correct specification and misspecification, to the discrepancy-based tests that did not impose the null hypothesis. However, for FRT-PP one can, by imposing the null in the imputation step or by using test statistics, protect against model misspecification. This, however, results in a substantial loss of power under a wide variety of alternatives.

Overall, as a recommendation for practitioners, if the model is potentially misspecified (e.g., important covariates are missing) one should avoid a full Bayesian model and use FRT-PP based on test statistics, not discrepancy measures, to maintain validity. If, however, we are confident in our posited model, the use of discrepancy-based tests could in principle improve power, especially if the imputation is not performed under the null. Under this case, however, the gains of this more complex approach over a full Bayesian approach are unclear. Regardless, incorporating predictive covariates, if present, can substantially reduce the risk of invalidity as well as give increased power for all the imputation methods. Therefore, if one has at least moderately predictive covariates, and a reasonably well-fitting model on the outcomes, we recommend leaving the imputation method unconstrained.

The paper has focused on the analysis of randomized experiments with noncompliance, a special case of principal stratification. In principle, all the testing approaches

that we have investigated could be modified and extended to other causal settings with intermediate variables as well as to more general settings with partially observed mixtures. All these extensions are feasible but not straightforward, given that the underlying assumptions (e.g, exclusion restriction in noncompliance) are context-specific.

References

Angrist, J. D., Imbens,G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91, 444–472. 681, 684, 690

Fisher, R. A. (1925). Statistical Methods for Research Workers. *1st ed. Oliver and Boyd, Edinburgh*. 682, 684

Fisher, R. A. (1925). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. 682, 684

Fisher, R. A. (1925). The design of experiments. *Edinburgh: Oliver and Boyd*. 682, 684

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58, 21–29. MR1891039. doi: https://doi.org/10.1111/j.0006-341X.2002.00021.x. 684

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733–807. MR1422404. 682, 686, 687, 688

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7, 2595–2602. MR3121624. doi: https://doi.org/10.1214/13-EJS854. 686

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society* B, 29(1) 83–100. MR0216699. 682, 685

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–476. 681, 690

Meng, X. L. (1994a). Posterior predictive p-values. *Annals of Statistics*, 22, 1142–1160. 682, 686, 687, 688

Meng, X. L. (1994b). Multiple-imputation inferences under uncongeniality. *Statistical Science*, 4, 538–573. 689

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Rolniczych Tom X [in Polish]; translated in Statistical Science*, 5, 465–480. MR1092986. 683

Neyman, J. (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection with discussion. *Journal of the Royal Statistical Society*, 97, 558–625. 687

Robins, J. M., Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95, 1143–1156. MR1804240. doi: https://doi.org/10.2307/2669750. 686

Rubin, B. D. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology* 66, 688–701. 683

Rubin, B. D. (1978). Bayesian inference for causal effects. *Annals of Statistics*, 6, 34–58. MR0472152. 683

Rubin, D. B. (1980). Comment on "Randomization Analysis of Experimental Data in the Fisher Randomization Tes" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593. MR0590687. 683

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4), 377–401. 682, 685

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4), 1151–1172. MR0760681. doi: https://doi.org/10.1214/aos/1176346785. 682, 685

Rubin, D. B. (1996a). Discussion of "Posterior predictive *p*-values?" by Gelman, A., Meng, X. L. and Stern, H.. *Statistica Sinica*, 6, 787–792. MR1311969. doi: https://doi.org/10.1214/aos/1176325622. 687, 694

Rubin, D. B. (1996b). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–520. MR2750144. doi: https://doi.org/10.1177/0008068320080305. 687

Rubin, D. B. (1998). More powerful randomization-based *p*-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17(3), 371–85. 681, 682, 684, 687, 688

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussions). *Journal of the American Statistical Association*, 82, 528–550. MR0898357. 692

Acknowledgments

The authors thank Peng Ding, Joseph Lee, and Natesh Pillai for helpful comments, conversations, and suggestions. This work is partially funded by PRIN 2012 grant.