Statistica Sinica Preprint No: SS-2016-0116R1	
Title	More Powerful Multiple Testing in Randomized Experiments with Non-Compliance
Manuscript ID	SS-2016-0116R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2020016.0116
Complete List of Authors	Joseph J Lee Laura Forastiere Luke Miratrix and Natesh S Pillai
Corresponding Author	Joseph J Lee
E-mail	joseph.j.lee@post.harvard.edu

principal strata (Frangakis and Rubin (2002)) based on compliance behavior aids inference for the desired causal effect. We use these tools in our approach but adapt them for simultaneous testing of multiple outcomes and subgroups.

Multiple testing issues are common in randomized experiments because multiple outcomes and subgroups of interest are often measured and analyzed for possible effects. Traditionally, practitioners have applied Bonferroni corrections to sets of p-values in order to control their familywise error rate (FWER), i.e., the rate at which at least one type I error is made, in a straightforward manner. Bonferroni corrections, however, tend to be overly conservative, especially when those p-values are correlated (Westfall and Young (1989)). This fact has led many applied researchers to avoid Bonferroni corrections and abandon multiple comparisons adjustments altogether (Cabin and Mitchell (2000); Nakagawa (2004); Perneger (1998); Rothman (1990)). Other avenues exist; randomization-based procedures can provide greater power while maintaining the FWER by accounting for correlated tests. Brown and Fears (1981) and Westfall and Young (1989) first introduced permutation-based multiple testing adjustments, though they did not explicitly motivate them using randomized assignment mechanisms. Randomization-based procedures are additionally appealing because they do not require any assumptions about the underlying distribution (here, joint) of the data. Furthermore, recent increases in computational power have helped such procedures become more tractable and gain popularity (Good (2005)).

In this article, we connect methodological ideas to appropriately handle both non-compliance and multiple testing in randomized experiments. We build up to this combined approach in stages. In Section 2, we elucidate the method proposed by Rubin (1998) for evaluating meaningful causal effects in the presence of non-compliance. In Section 3, we extend the ideas of Westfall and Young (1989) to fully randomization-based multiple comparisons adjustments and propose such adjustments as a straightforward yet more powerful alternative to Bonferroni corrections. In Section 4, we merge the notions of non-compliance and multiple testing, and outline a combined method of analysis that demonstrates power advantages from both perspectives. In each of Sections 2–4, we empirically show the benefits of the described methods through a series of simulated experiments. In Section 5, we apply traditional methods and our combined method to JTPA data to evaluate the program's effects on employment rate by time period. We illustrate how the methods lead to different conclusions regarding the significance of estimated JTPA effects. Section 6 concludes.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0116

2 Experiments with Non-compliance

2.1 Non-compliance as a missing data problem

Suppose we have a randomized experiment with N units, indexed by i, with observed covariates X_i, randomly assigned to control or active treatment. Let Z_i be a binary indicator for assignment to active treatment, and let $D_i(z)$ be a binary indicator for receipt of active treatment under assignment z. A unit's compliance behavior C_i is defined by the pair of potential outcomes (Neyman (1923); Rubin (1974)) $(D_i(0), D_i(1))$; this notation is adequate under the stable unit treatment value assumption (Rubin (1980, 1986)), which asserts no interference between experimental units, as well as two well-defined outcomes. Each unit then belongs to one of four possible compliance strata:

- Compliers ($C_i = c$), who receive their treatment assignment: $(D_i(0), D_i(1)) = (0, 1)$.
- Never-takers ($C_i = nt$), who never receive the active treatment: $(D_i(0), D_i(1)) = (0, 0)$.
- Always-takers ($C_i = at$), who always receive the active treatment: $(D_i(0), D_i(1)) = (1, 1)$.
- Defiers ($C_i = d$), who receive the opposite of their treatment assignment: $(D_i(0), D_i(1)) = (1, 0)$.

If non-compliance is one-sided — i.e., units assigned to control are prohibited from receiving the active treatment — then $D_i(0) = 0$ for all i. In such settings, always-takers and defiers do not exist, and two possible strata are left: compliers and never-takers. Real-world scenarios involving one-sided non-compliance include many clinical trials, in which new drugs are unavailable to control patients, and some job training experiments, in which training programs and additional services are unavailable to the control group.

In many practical settings, researchers are most interested in the compliers because the effect of treatment assignment is synonymous with the effect of treatment receipt for those units. Strata membership, however, can never be fully determined for all units because they depend on the two potential outcomes of D, one of which is unobserved. Membership can, on the other hand, be partially determined based on the observed potential outcome, D_i^{obs}. Table 1 outlines the possible compliance strata based on units' observed treatment assignment and receipt. An example "Science" table (Rubin (2005)) under one-sided non-compliance and its observed values under a particular assignment are shown in Table 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0116

Assignment	Receipt	Possible C_i Values	
Z_i	D_i^{obs}	One-sided Non-compliance	Two-sided Non-compliance
0	0	c, nt	c, nt
0	1	–	at, d
1	0	nt	nt, d
1	1	c	c, at

Table 1: Units' possible compliance strata based on observed treatment assignment and receipt.

		$D(z)$		Compliance	$Y(z)$		Assignment	$D(z)$		Compliance	$Y(z)$	
Unit	X_i	$D_i(0)$	$D_i(1)$	C_i	$Y(0)$	$Y(1)$	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_i(0)$	$Y_i(1)$
1	X_1	0	0	nt	$Y_1(0)$	$Y_1(1)$	0	0	?	?	Y_1^{obs}	?
2	X_2	0	1	c	$Y_2(0)$	$Y_2(1)$	1	0	1	c	?	Y_2^{obs}
3	X_3	0	1	c	$Y_3(0)$	$Y_3(1)$	1	0	1	c	?	Y_3^{obs}
4	X_4	0	0	nt	$Y_4(0)$	$Y_4(1)$	1	0	0	nt	?	Y_4^{obs}
...				...						...		
N	X_N	0	1	c	$Y_N(0)$	$Y_N(1)$	0	0	?	?	Y_N^{obs}	?

Table 2: An example Science table under one-sided non-compliance (left) and its corresponding observed and unobserved values under a particular assignment (right).

Because strata memberships are not fully observed, uncertainty with respect to complier-specific effects stems from the missing compliance statuses (i.e., D potential outcomes) in addition to the missing Y potential outcomes. One approach to handling the additional uncertainty is to, in a Bayesian framework, view the missing compliance statuses as random variables. By multiply imputing the missing compliance statuses, e.g., according to a distributional model, they can be integrated out, and we can make inference specific to the compliers.

2.2 Randomization-based posterior predictive p-values

As described by Meng (1994), a posterior predictive p-value can be viewed as the posterior mean of a classical p-value, averaging over the posterior distribution of nuisance factors (e.g., missing compliance statuses) under the null hypothesis. Rubin (1998) introduced a randomization-based procedure, which we expound on here, for obtaining posterior predictive p-values for estimated complier-only effects. One posterior predictive p-value is the average of many p-values calculated from multiple "compliance-complete" datasets with imputed compliance statuses; for each compliance-complete dataset, the p-value is obtained through a randomization test (Fisher (1925, 1935)).

Within one randomization test, however, calculations of the test statistic do not use all of the compliance information from the compliance-complete data; rather, they use only the compliance

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0116

information that would have actually been observed under particular hypothetical randomizations. Though implied, this step of re-observing the data is not explicitly stated by Rubin (1998); we place it in Step 5 of our procedure for emphasis because it is an important prerequisite for conducting a proper test. Unlike discrepancy variables (Meng (1994)), which may depend on unobserved factors (e.g., missing compliance statuses), test statistics must be functions of only the observed data. In order to conduct a proper test, the true observed test statistic value must be measured against the correct distribution, i.e., the distribution of that same test statistic.

In this section, we assume a single outcome for simplicity. The procedure for obtaining a randomization-based posterior predictive p-value is as follows.

1. **Choose a test statistic and calculate its observed value.**

   Choose a test statistic, T, to estimate an effect on the outcome, Y. Calculate T on the observed data to obtain T^{obs}.

   Examples include the maximum-likelihood estimate (MLE) of CACE or the posterior median of CACE, given the observed compliance statuses and potential outcomes, under the exclusion restriction (see Angrist, Imbens, and Rubin (1996); Imbens and Rubin (1997)).

for $m : 1$ to M **do**

2. **Impute missing compliance statuses.**

   Impute the missing compliance statuses, drawing once from their posterior predictive distribution according to a compliance model that assumes the null hypothesis (e.g., of zero effect).

3. **Impute missing potential outcomes.**

   Impute the missing Y potential outcomes under the sharp null hypothesis. Under the typical sharp null hypothesis of zero treatment effect, the missing potential outcome for unit i is imputed exactly as Y_i^{obs}.

4. **Draw a random hypothetical assignment.**

   Draw a random hypothetical assignment vector according to the assignment mechanism used in the original experiment.

5. **Re-observe the data.**

   Treating the imputed compliance statuses, imputed potential outcomes, and hypothetical