HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2015 November 23.

Published in final edited form as:

Stat Med. 2015 October 30; 34(24): 3214–3222. doi:10.1002/sim.6531.

Valid randomization-based *p*-values for partially post hoc subgroup analyses

Joseph J. Lee[*,†] and **Donald B. Rubin**

Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, U.S.A.

Abstract

By 'partially post-hoc' subgroup analyses, we mean analyses that compare existing data from a randomized experiment—from which a subgroup specification is derived—to new, subgroup-only experimental data. We describe a motivating example in which partially post hoc subgroup analyses instigated statistical debate about a medical device's efficacy. We clarify the source of such analyses' invalidity and then propose a randomization-based approach for generating valid posterior predictive *p*-values for such partially post hoc subgroups. Lastly, we investigate the approach's operating characteristics in a simple illustrative setting through a series of simulations, showing that it can have desirable properties under both null and alternative hypotheses.

Keywords

causal inference; Fisher randomization test; multiple comparisons; posterior predictive *p*-value; statistical significance

1. Introduction

Subgroup causal effects are often of scientific interest in randomized experiments. When subgroups are specified after observing outcomes, however, the estimated subgroup effects and *p*-values produced by traditional statistical methods do not have the typically desired repeated sampling properties. Traditional multiple comparisons (e.g., Bonferroni) adjustments tend to be overly conservative when subgroups overlap [1, 2]. Moreover, post hoc decisions often make it difficult to specify the number of comparisons being made, making such adjustments less straightforward. 'Partially post-hoc' subgroup analyses, which compare existing data—from which the subgroup specification is derived—to new, subgroup-only experimental data, are further complicated.

Here, we describe a motivating example faced by the US Food and Drug Administration (FDA) in which a partially post hoc subgroup analysis instigated statistical debate about a medical device's efficacy. We provide a statistical framework to clarify the source of statistical invalidity. We then propose a randomization-based method for generating valid posterior predictive *p*-values for such partially post hoc subgroups. Although we do not have raw data for the particular example, we investigate the method's operating characteristics

[*]Correspondence to: Joseph J. Lee, Department of Statistics, Harvard University, 1 Oxford St., Cambridge, MA 02138, U.S.A..

[†]joseph.j.lee@post.harvard.edu

through a series of simulations, showing that it exhibits both a valid type I error rate and substantial power under reasonable alternative hypotheses.

2. The Durolane trials

In March 2006, Swedish medical device company Q-Med AB (Uppsala, Sweden) submitted its medical device Durolane to the FDA for pre-market approval. Durolane is a viscous gel intended to treat osteoarthritic knee pain, administered through intra-articular (into-the-joint) injection. Prior to FDA review, the device had already been approved in a number of countries across Europe and Asia. As evidence of Durolane's efficacy, Q-Med submitted analyses of data from three randomized clinical trials.

Q-Med initially conducted two randomized, double-blind experiments attempting to demonstrate Durolane's superiority to a saline placebo (control), measuring each patient's pre-treatment and post-treatment pain scores on the Western Ontario and McMasters Universities Osteoarthritis Index (WOMAC). Higher WOMAC scores translate to higher levels of pain; patients were deemed positive responders to treatment if their WOMAC pain scores were reduced by at least 40% and at least five points on the 0–20 point scale. Comparisons of responder rates in studies 1 and 2 produced *p*-values of 0.53 and 0.49, respectively, neither of which are considered close to statistically significant by the FDA (the standard FDA cutoff for significance is 0.05).

After observing and unblinding outcome data from studies 1 and 2, Q-Med combined and filtered the data for a post hoc subgroup analysis. Patients without effusion (fluid accumulation in the knee joint) and without baseline polyarticular pain—284 (50.4%) of the 564 total patients—were included in the selected subgroup. Researchers asserted that including patients with effusion and polyarticular pain would make it '… difficult to observe a treatment effect because the presence of these conditions leads to considerable variability in the WOMAC assessment' [3]. The subgroup, however, was not specified a priori. Within the subgroup, Q-Med found a statistically significant difference in responder rates, boasting a drastically decreased *p*-value of 0.0013.

Per FDA's request, Q-Med conducted a third clinical trial to confirm this affirmative result for the specified patient subpopulation. All study 3 patients satisfied the covariate inclusion criteria specified in the post hoc analysis of studies 1 and 2. Instead of using a placebo control group in study 3, however, Q-Med decided to compare Durolane with the drug methylprednisolone in a non-inferiority trial. Although methylprednisolone is currently an approved standard of care for osteoarthritic knee pain, there exist concerns about its side effects, particularly its tendency to destroy cartilage over time; Durolane is purportedly able to avoid this detrimental side effect. Based on 442 patients, study 3 successfully showed Durolane's non-inferiority to methylprednisolone. But because no saline control group was included, study 3 alone did not provide direct evidence, as desired by the FDA, for Durolane's superiority to a saline placebo in reducing knee pain.

Because there was no placebo control group for study 3, Q-Med used covariate-matched saline placebo patients from studies 1 and 2 to assess Durolane's effectiveness. The selected historical placebo controls (i) met study 3 covariate inclusion criteria and (ii) provided

sufficient covariate balance in comparison with the study 3 Durolane treatment group, as determined by a propensity score model [4]. Observed outcomes were not used in the selection of historical controls. The comparison of historical controls from studies 1 and 2 with Durolane patients from study 3 favored Durolane, with a statistically significant *p*-value of 0.047. Nevertheless, in August 2009, the FDA rejected Durolane for sale in the United States.

3. A statistical framework

Here, we describe a statistical framework and explore the implications of using historical controls in the Durolane context. We argue why the combined subgroup *p*-value is invalid (in terms of type I error) and demonstrate that balancing covariate profiles (e.g., via propensity score matching) cannot fully repair its validity.

3.1. Experiment 1 and post hoc subgroup specification

Suppose that we have N_1patients in experiment 1 (representing Durolane studies 1 and 2 collectively), each defined by a single, binary covariate X (e.g., man/woman) and randomly assigned to the control or active treatment (indicated by W). Let Y be a binary experimental outcome representing whether or not a patient 'responds' to the assigned treatment (e.g., in terms of having reduced knee pain). The 'Science' table [5] for experiment 1 and its observed values under a particular assignment are shown in Table I. Each patient has two potential outcomes [6], only one of which can ever be observed [7]. This notation is sufficient under the stable unit treatment value assumption, which asserts no interference between experimental units, as well as two well-defined outcomes [8]. Unit i, j represents the jth unit from experiment i.

Suppose that after experiment 1 outcomes are observed, the researcher calculates subgroup *p*-values for each possibly scientifically relevant subgroup, defined by X values. In the setting with a single binary covariate X, there are three possible subgroups: $X = 0$, $X = 1$, and $X \in \{0, 1\}$. The researcher then presumably identifies the subgroup S for which the active treatment appears to have the most beneficial effect with respect to Y, for example, the subgroup with the most significant *p*-value in favor of the active treatment.

3.2. Experiment 2

Experiment 2 (representing Durolane study 3) consists of N_2 patients, all of whom satisfy the subgroup criteria of S and are assigned to active treatment; there are no control patients. In reality, Durolane's experiment 2 introduced a new, third treatment condition (methylprednisolone; see Section 2). However, Q-Med's ultimate goal was to satisfy FDA's request to compare Durolane with the saline placebo (the control from experiment 1). Experiment 2 patients assigned to methylprednisolone are not relevant for this purpose and are thus ignored here.

Without loss of generality, suppose that the $X = 1$, for example, woman, subgroup is chosen from experiment 1. Then experiment 2 consists entirely of female patients assigned to the active treatment; the observed values of experiment 2 Science table are shown in Table II.

3.3. Combined subgroup analysis

Because experiment 2 outcomes are realized after *S* is specified, experiment 2 data can be used to estimate the subgroup average treatment effect (SubATE) validly through traditional statistical methods. For instance, if control patients were also included in experiment 2, the data from experiment 2 alone could generate a valid subgroup *p*-value. However, comparisons using experiment 1 control patients and experiment 2 treatment patients cannot be handled in the same manner.

Consider the following two quantities: (i) $\hat{\tau}_{1,S}=\bar{Y}^{obs}_{1,j\in S}(1)-\bar{Y}^{obs}_{1,j\in S}(0)$, and (ii) $\hat{\tau}_{combined,S}=\bar{Y}^{obs}_{2,j\in S}(1)-\bar{Y}^{obs}_{1,j\in S}(0)$. The first quantity, $\hat{\tau}_{1,S}$ represents the estimated SubATE for *S* from experiment 1, from which the selected subgroup specification is derived. The second, $\hat{\tau}_{combined,S}$, represents the estimated SubATE for *S* obtained by comparing experiment 1 control patients and experiment 2 treatment patients. The combined subgroup analysis calculates $\hat{\tau}_{combined,S}$ and generates an associated subgroup *p*-value.

3.4. Invalidity under the null hypothesis

Suppose that the null hypothesis of zero treatment effect is true. Because the specification of *S* is a function of observed experiment 1 outcomes—and *S* is a subgroup for which the treatment appears effective— the expectation of $\hat{\tau}_{1,S}$ (over the randomization) is positive, and its associated *p*-value is skewed right, making traditional testing invalid in terms of type I error ([9]). Conceptually, $E\left(\hat{\tau}_{1,S}\right)>0$ because $\hat{\tau}_{1,S}$ is the maximum or near maximum of several estimated SubATEs from experiment 1. Examining the two terms that make up $\hat{\tau}_{1,S}$, we expect $\bar{Y}^{obs}_{1,j\in S}(1)$ to be high and $\bar{Y}^{obs}_{1,j\in S}(0)$ to be low because of the subgroup specification; in other words, we expect experiment 1 treatment patients in *S* to have artificially good outcomes and experiment 1 control patients in *S* to have artificially poor outcomes, because such outcome information was used to select *S* in the first place.

In addition, because $\hat{\tau}_{combined,S}$ shares one term with $\hat{\tau}_{1,S}$, we can see by the linearity of the expectation operator that $\hat{\tau}_{combined,S}$ also has a positive expectation and a skewed-right *p*-value under the null hypoth esis. Although $\bar{Y}^{obs}_{2,j\in S}(1)$ is realized after *S* is specified, the carry-over usage of $\bar{Y}^{obs}_{1,j\in S}(0)$ renders traditional testing of $\hat{\tau}_{combined,S}$ invalid (although, one could argue, 'less invalid' than testing of $\hat{\tau}_{1,S}$).

Finally, note in this setting that the historical controls in the combined subgroup analysis have covariate profiles (e.g., $X = 1$) that exactly match experiment 2 treatment patients. Here, the statistical problem is rooted not in any discrepancies between control and treatment covariate profiles but in the usage of observed experiment 1 outcomes under the false assumption that they are independent of the subgroup specification. Any traditionally calculated subgroup *p*-value for $\hat{\tau}_{combined,S}$ cannot be valid, regardless of any covariate balancing (e.g., propensity score matching) techniques designed to mitigate that invalidity.

4. Valid randomization-based *p*-values for post hoc subgroups in the presence of nuisance unknowns

Lee and Rubin [9] introduced a randomization-based approach for generating valid post hoc subgroup *p*-values, motivated by earlier ideas about randomization due to Fisher [10]. The fundamental insight is to specify the decision tree that led to the final test statistic value, considering what subgrouping and inferential steps would have been taken if the data had been realized under a different randomization. The approach entails (i) specifying a precise post hoc subgrouping procedure and an accompanying test statistic, (ii) calculating the test statistic on the observed data, (iii) imputing the missing potential outcomes in the study under a sharp null hypothesis, (iv) repeatedly drawing random hypothetical assignments according to the assignment mechanism and calculating test statistic values on the corresponding hypothetical datasets to construct the null randomization distribution of the test statistic, and (v) comparing the observed test statistic value against its null randomization distribution.

Calculating test statistic values on hypothetical data from a single experiment under a sharp null hypothesis is straightforward. For example, under the sharp null hypothesis of zero treatment effect, the missing potential outcomes can be imputed exactly as observed for each unit. The test statistic is then calculated using the hypothetical assignment and corresponding hypothetical observed data, given the specified subgrouping procedure. However, in settings with multiple experiments—including one or more that occur after the subgroup specification—constructing the null randomization distribution of the test statistic requires some ingenuity.

Consider our example from Section 3, in which women ($X = 1$ patients) make up the selected subgroup, S. The final test statistic, $\hat{\tau}_{combined,S}$, compares outcomes from female experiment 1 control patients with outcomes from experiment 2 treatment patients, where by construction, all of experiment 2 patients are women. But what if men had exhibited a more beneficial estimated treatment effect than women in experiment 1? Presumably, experiment 2 would then have included only men; the experimental sample for experiment 2 would have been completely different.

Here, we propose an extension of the aforementioned method that generates valid posterior predictive *p*-values [11, 12] in the presence of nuisance unknowns, for example, male experiment 2 outcomes. We expand experiment 2 Science table, conceptualizing it as an augmented experiment with N_2' patients $\left(\left(N_2' > N_2\right)\right.$, containing patients with the same mix of X values as experiment 1 and filled with missing data. For experiment 2 patients that exist in reality (i.e., women), potential outcomes under active treatment are observed, but potential outcomes under control are missing (unobserved). For experiment 2 patients that exist only in our augmented framework (i.e., men), both potential outcomes are missing. The observed and unobserved values of the augmented experiment 2 Science table are shown in Table III.

Because the male treatment potential outcomes are not observed in experiment 2, values of $\hat{\tau}_{combined,S}$ under hypothetical randomizations can be considered random variables, with uncertainty resulting from these missing potential outcomes. Given a set of imputed values, however, we can construct the randomization distribution of $\hat{\tau}_{combined,S}$ and calculate a randomization-based *p*-value. Thus, by multiply imputing [13] the missing male potential outcomes according to a distributional model that assumes the null hypothesis, they can be 'integrated out' to produce a posterior predictive *p*-value; the posterior predictive *p*-value is the average *p*-value over the multiple imputations of the missing treatment potential outcomes.

Under the null hypothesis, the posterior predictive distribution of the missing treatment potential outcomes is informed by the observed experiment 1 potential outcomes for both control and treatment patients. In the framework with binary outcomes and two independent covariate subgroups (men versus women), there are two parameters to model for imputation: θ_m, the probability of a successful male outcome, and θ_f, the probability of a successful female outcome. A typical, non-controversial Bayesian model involves two independent Beta priors and Binomial likelihoods. (In the simulations as follows, we specify diffuse, independent Beta(1,1) priors.) Because of conjugacy, the posterior distributions of θ_m and θ_f remain Beta, leading to a Beta-Binomial posterior predictive distribution for the missing treatment potential outcomes. Here, modeling is further simplified by the fact that the treatment potential outcomes in experiment 2 need to be imputed only for the male patients; θ_f can be ignored because it is not needed. For a post hoc subgroup test statistic *T*, the full procedure for obtaining a posterior predictive *p*-value in the two-experiment setting is as follows:

Specify precisely the post hoc subgrouping procedure and the subgroup test statistic of interest, *T* (e.g., $\hat{\tau}_{combined,S}$).

Perform the post hoc subgrouping procedure on the observed dataset to obtain the observed subgroup test statistic, T^{obs}.

Using the augmented experiment 2 framework, multiply impute the missing treatment potential outcomes (e.g., *M* times), using their posterior predictive distributions according to a distributional model that assumes the null hypothesis. *M* is a large number (e.g., 10,000) that controls the Monte Carlo integration error.

For each of the *M* imputed datasets, calculate a randomization-based *p*-value for *T* according to the actual assignment mechanism(s) and specified subgrouping procedure ([9]), treating the imputed values as true. The randomization-based *p*-value is the proportion of hypothetical randomizations for which the corresponding test statistic value, T^{hyp}, is as extreme as or more extreme than T^{obs}.

The posterior predictive *p*-value for the null hypothesis with respect to *T* equals the average of the *M* randomization-based *p*-values.

The method outlined previously can be viewed as a form of data augmentation [14], with the expanded experiment 2 population making it possible to obtain a posterior predictive *p*-value for *T*. Rubin [15] described a similar procedure in a different setting, in order to obtain

a posterior predictive *p*-value for the complier average causal effect in a single experiment with non-compliance. In that setting, the nuisance unknowns were the missing compliance statuses of the patients in the experiment who were assigned to the control treatment. In the same paper, a computational shortcut was identified: for each of the *M* imputed datasets, only one hypothetical assignment needs to be drawn in step 4. The individual randomization-based *p*-values then equal either 0 or 1, and the posterior predictive *p*-value is the average of the indicators.

As mentioned by Lee and Rubin [9], specifying a post hoc procedure exactly as it occurred may be difficult. In such cases, the randomization-based posterior predictive *p*-values can still place helpful bounds on the significance level of estimated subgroup effects by using reasonable approximations that place limits on the post hoc procedure.

5. Operating characteristics

5.1. Simulation setup

To evaluate the operating characteristics of the proposed method, we simulate random datasets under various treatment effect hypotheses. We first randomly sample $N_1 = 500$ patients for experiment 1, drawing from a population of 50% women and 50% men. We randomly assign $N_1/2$ of these patients to control and the other $N_1/2$ to active treatment. We then draw random Bernoulli outcomes according to the probabilities in Table IV.

After observing experiment 1 outcomes, we specify the subgroup *S* for further study. There are three possible choices for *S*: men, women, or all patients; *S* is the subgroup exhibiting the smallest *p*-value based on control versus treatment experiment 1 responder rates. (If multiple subgroups share the smallest *p*-value, the one in that pool with the largest number of included units is selected. If multiple subgroups share the smallest *p*-value and the largest sample size within that pool, one of them is selected at random.)

Experiment 2 is conducted with new patients, all of whom satisfy the criterion of *S* and are assigned to active treatment. The combined subgroup *p*-value is calculated for $\hat{\tau}_{combined,S}$, comparing experiment 2 treatment units in *S* to experiment 1 control units in *S*. Randomization-based posterior predictive *p*-values are then generated according to the procedure described in Section 4.

5.2. Simulation results

Under the null hypotheses described in Table IV, both experiment 1 and combined (experiments 1 and 2) subgroup *p*-values are invalid in terms of type I error, that is, both subgroup *p*-values incorrectly reject the null hypothesis more often than desired by the nominal significance level. Figure 1 shows the histograms of these *p*-values under the null hypotheses. As expected, all of the histograms are heavily skewed right, indicating the *p*-values' invalidity. The combined subgroup *p*-values are slightly less skewed, suggesting that they are, in some sense, 'less invalid' than experiment 1 subgroup *p*-values.

On the other hand, the posterior predictive *p*-value appears valid—in fact, conservative—in terms of type I error (Figure 2). In other words, when the null hypothesis is true, the

posterior predictive *p*-value rejects it less often than indicated by the nominal significance level. Such conservatism often arises when multiply imputing missing data under a null hypothesis ([15]) and seems to become more extreme when the proportion of missing data is large.

Table V displays simulation results, comparing the type I error rates based on experiment 1, combined, and posterior predictive subgroup *p*-values.

Conservatism under a null hypothesis is often welcome, especially in FDA contexts, provided that the method exhibits sufficient power under reasonable alternative hypotheses; simulations show this to indeed be the case. Under the alternative hypotheses described in Table IV, the method has substantial power, rejecting the null hypothesis at $\alpha = 0.05$ in 21% and 69% of replications under alternative hypotheses A (5% treatment effect) and B (10% treatment effect), respectively. Figure 3 displays the histograms of posterior predictive *p*-values under these alternative hypotheses. We also note that in the motivating example from Section 2, pre-experiment sample size calculations aimed to capture 80% power at $\alpha = 0.05$ assuming larger treatment effects of 15–20%; according to simulations, our method achieves over 95% power when generating data with such large effects.

6. Discussion

We have presented a randomization-based approach for generating valid posterior predictive *p*-values for partially post hoc subgroups. We have also demonstrated that the resulting *p*-values can have substantial power under reasonable alternative hypotheses. In the simple illustrative example provided, the multiple imputation of the missing experiment 2 treatment potential outcomes is facilitated by a conjugate model assuming independence between the subgroups. Such independence may not always be plausible, especially as the number of relevant covariates increases. For instance, only two covariates—effusion and polyarticular pain—defined the Durolane subgroup; thus, subgroup patients could and did share values of other additional covariates with non-subgroup patients. The augmented framework then requires an imputation model that relates covariates to the outcomes, for example, Bayesian logistic regression, under the null hypothesis. The missing experiment 2 treatment potential outcomes are again multiply imputed according to their posterior predictive distribution, which can be empirically constructed using Markov chain Monte Carlo techniques.

Having a single binary covariate in our illustration provides a straightforward imputation model that allows us to highlight our methodological contributions. When dealing with more complicated imputation models—as with all parametric models—it may be useful to investigate the sensitivity of the resulting inference to the model specification and the choice of priors. More generally, our randomization-based approach applies to randomized experiments with nuisance unknowns; examples of such unknowns include missing compliance statuses ([2]) and missing outcome data from patients who exist in reality (as opposed to existing only in the augmented framework).

Acknowledgements

The first author is supported by the US Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

1. Westfall PH, Young SS. *p* value adjustments for multiple tests in multivariate binomial models. Journal of the American Statistical Association. 1989; 84(407):780–786.
2. Lee JJ, Miratrix L, Pillai NS. More powerful multiple testing in randomized experiments with non-compliance, submitted.
3. Q-Med, AB. Durolane knee premarket approval application (p060013), sponsor executive summary. Presented to the U.S. Food and Drug Administration; 2009.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.
5. Rubin DB. Causal inference using potential outcomes. Journal of the American Statistical Association. 2005; 100(469):22–331.
6. Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science. 1923; 1990; 5(4):465–472. Translated by.
7. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66(5):688–701.
8. Rubin DB. Comment on randomization analysis of experimental data: the Fisher randomization test. Journal of the American Statistical Association. 1980; 75(371):591–593.
9. Lee JJ, Rubin DB. Evaluating the validity of post-hoc subgroup inferences: a case study. The American Statistician.
10. Fisher, RA. The Design of Experiments. Oliver & Boyd; Oxford: 1935.
11. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics. 1984; 12(4):1151–1172.
12. Meng XL. Posterior predictive p-values. The Annals of Statistics. 1994; 22:1142–1160.
13. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.; New York: 1987.
14. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association. 1987; 82(398):528–540.
15. Rubin DB. More powerful randomization-based p-values in double-blind trials with non-compliance. Statistics in Medicine. 1998; 17(3):371–385. [PubMed: 9493260]

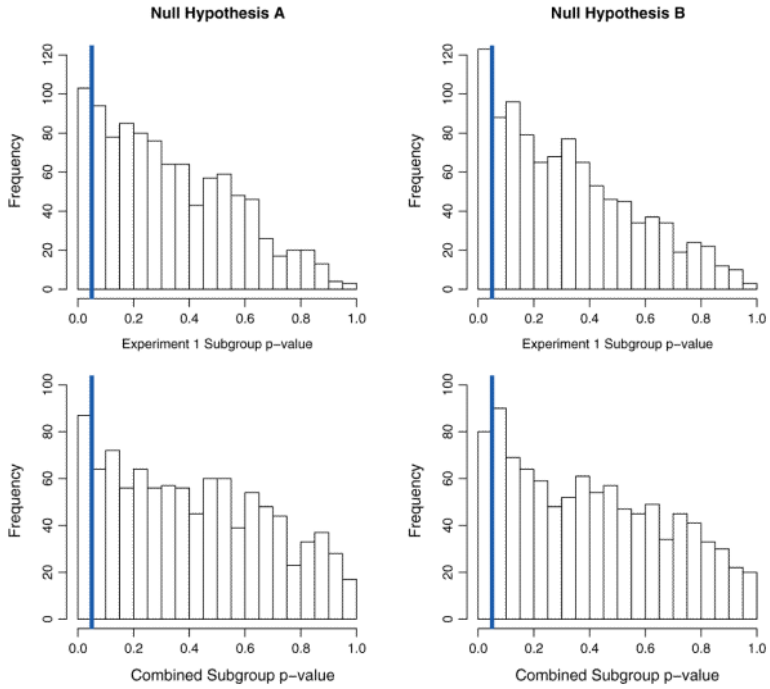

Figure 1.

Empirical distributions of experiment 1 (top) and combined (experiments 1 and 2; bottom) subgroup *p*-values under the null hypotheses described in Table IV based on 1000 simulated datasets.

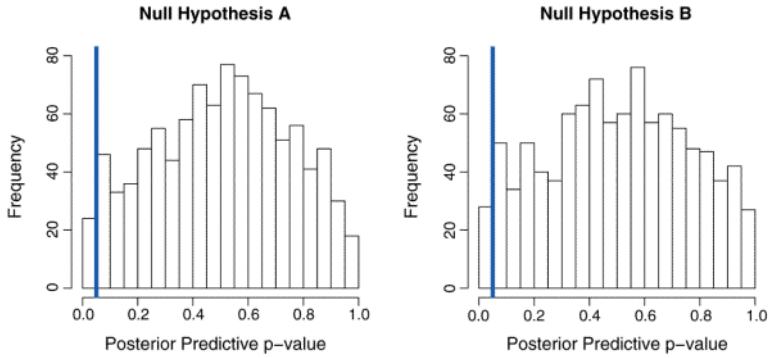

Figure 2.

Empirical distributions of posterior predictive *p*-values under the null hypotheses described in Table IV based on 1000 simulated datasets.

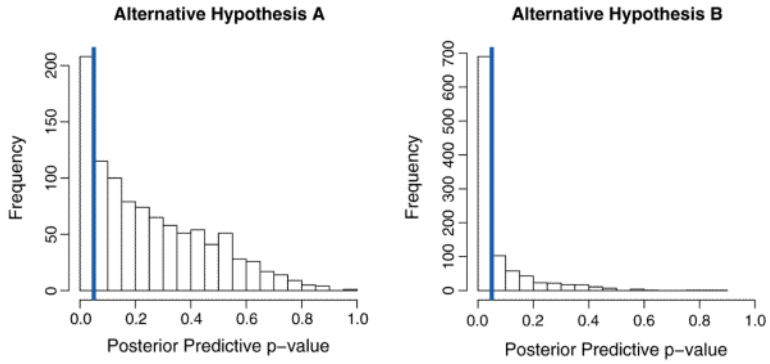

Figure 3.

Empirical distributions of posterior predictive *p*-values under the alternative hypotheses described in Table IV based on 1000 simulated datasets.

Table I

The Science table for experiment 1 (left) and its corresponding observed values under a particular assignment (right).

Experiment 1						
	Covariate	Potential outcomes		Assignment	Observed outcomes	
Unit (i,j)	$X_{i,j}(0)$	$Y_{i,j}(0)$	$Y_{i,j}(1)$	$W_{i,j}^{obs}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
1,1	0	$Y_{1,1}(0)$	$Y_{1,1}(1)$	0	$Y_{1,1}^{obs}$	?
1,2	1	$Y_{1,2}(0)$	$Y_{1,2}(1)$	0	$Y_{1,2}^{obs}$	?
1,3	0	$Y_{1,3}(0)$	$Y_{1,3}(1)$	1	?	$Y_{1,3}^{obs}$
…		…			…	
$1,N_1$	1	$Y_{1,N_1}(0)$	$Y_{1,N_1}(1)$	1	?	Y_{1,N_1}^{obs}

Table II

The observed values of the Science table for experiment 2.

Experiment 2				
	Covariate	Assignment	Observed outcomes	
Unit (i, j)	$X_{i,j}$	$W_{i,j}^{obs}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
2,1	1	1	?	$Y_{2,1}^{obs}$
2,2	1	1	?	$Y_{2,2}^{obs}$
2,3	1	1	?	$Y_{2,3}^{obs}$
…		…		
2, N_2	1	1	?	Y_{2,N_2}^{obs}

Table III

The observed and unobserved values of the augmented experiment 2 Science table.

Experiment 2				
	Covariate	Assignment	Potential outcomes	
Unit (i, j)	$X_{i,j}$	$W_{i,j}^{obs}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
2,1	1	1	?	$Y_{2,1}^{obs}$
2,2	1	1	?	$Y_{2,2}^{obs}$
2,3	1	1	?	$Y_{2,3}^{obs}$
…		…		
$2,N_2$	1	1	?	Y_{2,N_2}^{obs}
$2,N_2+1$	0	1	?	?
$2,N_2+2$	0	1	?	?
…		…		
$2,N_2^{'}$	0	1	?	?

Table IV

Simulated outcome success (‘response’) probabilities under various treatment effect hypotheses.

	Men		Women	
	Control	**Active treatment**	**Control**	**Active treatment**
Null hypothesis A	0.50	0.50	0.50	0.50
Null hypothesis B	0.20	0.20	0.80	0.80
Alternative hypothesis A	0.50	0.55	0.50	0.55
Alternative hypothesis B	0.20	0.30	0.80	0.90

Table V

Type I error rates (at $\alpha = 0.05$) of experiment 1, combined, and posterior predictive subgroup *p*-values based on 1000 simulated datasets.

	Type I error rate at $\alpha = 0.05$	
Subgroup *p*-value	**Null hypothesis A (%)**	**Null hypothesis B (%)**
Experiment 1	10.3	12.3
Combined (experiments 1 and 2)	8.7	8.0
Posterior predictive	2.0	2.3