Published in final edited form as:

Health Serv Outcomes Res Methodol. 2016 September ; 16(3): 154–171. doi:10.1007/s10742-016-0155-7.

Combining Non-randomized and Randomized Data in Clinical Trials Using Commensurate Priors

Hong Zhao^{*}, Brian P. Hobbs[†], Haijun Ma[‡], Qi Jiang[‡], and Bradley P. Carlin^{*}

^{*}Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, MN 55455 USA

[†]The University of Texas MD Anderson Cancer Center, Houston, TX 77030 USA

[‡]Amgen Inc., Thousand Oaks, CA 91320 USA

Abstract

Randomization eliminates selection bias, and attenuates imbalance among study arms with respect to prognostic factors, both known and unknown. Thus, information arising from randomized clinical trials (RCTs) is typically considered the gold standard for comparing therapeutic interventions in confirmatory studies. However, RCTs are limited in contexts wherein patients who are willing to accept a random treatment assignment represent only a subset of the patient population. By contrast, observational studies (OSs) often enroll patient cohorts that better reflect the broader patient population. However, OSs often suffer from selection bias, and may yield invalid treatment comparisons even after adjusting for known confounders. Therefore, combining information acquired from OSs with data from RCTs in research synthesis is often criticized due to the limitations of OSs. In this article, we combine randomized and non-randomized substudy data from FIRST, a recent HIV/AIDS drug trial. We develop hierarchical Bayesian approaches devised to combine data from all sources simultaneously while explicitly accounting for potential discrepancies in the sources' designs. Specifically, we describe a two-step approach combining propensity score matching and Bayesian hierarchical modeling to integrate information from non-randomized studies with data from RCTs, to an extent that depends on the estimated commensurability of the data sources. We investigate our procedure's operating characteristics via simulation. Our findings have implications for HIV/AIDS research, as well as elucidate the extent to which well-designed non-randomized studies can complement RCTs.

Keywords

Bayesian analysis; commensurate priors; Markov chain Monte Carlo (MCMC); observational studies (OSs); propensity score matching; randomized clinical trials (RCTs)

Address correspondence to Dr. Bradley P. Carlin, Division of Biostatistics, University of Minnesota School of Public Health, Mayo Mail Code 303, Minneapolis, MN 55455, USA. brad@biostat.umn.edu.

Compliance with Ethical Standards

Ethical approval: All analyses of human participants performed by the author for this work involved only secondary data analysis of pre-existing, deidentified datasets, preserving the subjects' confidentiality. As such, the work is of the kind typically considered exempt from IRB approval in the United States.

Conflict of Interest: All authors declare they have no conflicts of interest.

1 Introduction

1.1 Combining Randomized Clinical Trials (RCTs) with Observational Studies (OSs)

RCTs and OSs are the two primary approaches for evaluating the effectiveness of therapeutic interventions. In RCTs, subjects are randomly assigned to treatment groups to eliminate treatment selection bias, which can negatively impact our estimate of association between treatment and outcome. This is based on the assumption that the distribution of observed and unobserved covariates should be balanced across the treatment groups on average after randomization. Therefore, RCTs have long been recognized as the gold standard for testing the treatment efficacy or safety of an intervention, and are considered to yield the highest grade of evidence in the hierarchy of research designs (Evidence-Based Medicine Working Group, 1992; U.S. Preventive Services Task Force, 1996). Although RCTs facilitate valid treatment comparisons, they may be restrictive if the study population, those willing to undergo randomization, represent only a subset of the patient population. By contrast, OSs may enroll patient cohorts that better reflect the broader patient population because they often use less restrictive inclusion criteria. However, OSs can suffer from selection bias, and may yield invalid treatment comparisons even after adjusting for known confounders.

Combining RCTs and OSs in research synthesis is often criticized due to the limitations of OSs. Table 1 shows a “grades of evidence” ranking (U.S. Preventive Services Task Force, 1996) using internal validity as the primary criterion. The lowest grade includes anecdotal case histories and expert opinion, while OSs such as well-designed cohort or case control studies fall at intermediate levels. Recent advances in epidemiology and statistics, which have enhanced understanding of the implications of study design and analysis for causal inference (the process of analyzing a causal connection based on the conditions of the occurrence of an effect (Pearl, 2009)), challenge this evidence hierarchy (Ligthelm et al., 2007). MacLehose et al. (2000) concluded that discrepancies between RCTs and OSs estimates of effect size and outcome frequency for different groups were small for high-quality studies, but potentially large for low-quality studies. Concato et al. (2000) found that well-designed OSs provided similar estimates for treatment effects compared to those in RCTs across five clinical topics and 99 reports evaluated. More recent literature supports the opinion that systematic reviews of treatment effects should not be restricted to specific study types in all cases. Moreover, because RCTs often admit limited patient populations, conclusion obtained from RCTs may be limited in scope and thereby contradict results obtained from studies of broader patient cohorts (Ioannidis, 2005). Furthermore, if treatment effects are estimated from a single RCT, they might not be more reliable than those obtained after integrating the information acquired from several well-designed OSs. Therefore, examining the extent to which well-designed non-randomized studies can complement RCTs should further advance our understanding of how to effectuate evidence-based medicine through integration of all available sources of information.

1.2 The FIRST Study

In some cases, owing to the inherent difficulty of implementing RCTs, which require that the participating clinicians agree to forgo using their clinical expertise and randomly select

therapies for their patients (as well as additional institutional oversight), statistical methods are needed to facilitate integrative analysis based on both types of data. The Flexible Initial Retrovirus Suppressive Therapies (FIRST) trial, conducted by the Community Programs for Clinical Research on AIDS (CPCRA), offers an example. As described in MacArthur et al. (2001), highly active antiretroviral therapy-naïve, HIV-infected subjects were randomized to three strategies (nucleoside reverse transcriptase inhibitor (NRTI) was used in all three strategies): a two-class protease inhibitor (PI+NRTI), a two-class non-nucleoside reverse transcriptase inhibitor (NNRTI+NRTI), and a three-class strategy (PI+NNRTI+NRTI). Participants within the two strategies involving NNRTIs could further specify whether they wanted to be randomly assigned to a NNRTI drug (nevirapine, NVP, or efavirenz, EFV) before the randomization to strategy arms, or permit a study clinician to prescribe one of the two drugs. The three strategies were compared for long-term virological and immunological durability, drug resistance, and disease progression.

Figure 1 offers a pictorial representation of the study design. Here, we consider only the data from the two-class NNRTI strategy, including randomized EFV ($n = 45$) or NVP ($n = 53$), as well as patients whose clinician chose EFV ($n = 211$) or NVP ($n = 100$). Our data set excludes patients missing an 8-month plasma HIV RNA measurement. Our goal is to compare the probability of virological suppression (HIV RNA < 50 copies/ml) under EFV and NVP at 8 months, adjusting for several baseline covariates. Analysis of the small randomized substudy data alone yields insufficient power. The larger non-randomized cohort is likely more representative of the HIV/AIDS population at large, but is subject to selection bias due to patient preference, local medical practice patterns, and other factors. Thus, we need to develop a method to cautiously combine the randomized and non-randomized data, after attempting to correct the latter for treatment selection bias.

The rest of our paper is organized as follows. In Section 2, we first introduce the propensity score matching method for OSs. Then we describe a commensurate prior (Hobbs et al., 2012) approach to adaptively borrow information from the matched non-randomized (NR) cohort to complement the parameter estimation from using the randomized substudy (RS) data. In Section 3, we apply our approach to the FIRST dataset. Naïve modeling based on either the RS or NR data is described; the results are surprisingly discrepant, motivating use of our new method to combine the matched NR data with the RS data. Section 4 then evaluates our approach via simulation, showing different degrees of borrowing between the information contribute by NR and RS data depending to some extent on the model and prior distribution we choose. Finally, in Section 5 we summarize and offer directions for future research.

2 Methods

2.1 Propensity Score Matching for Observational Studies

Since treatment selection is often influenced by clinical factors, the strongest argument against using OSs for causal inferences is the potential for lurking confounders. These lurking covariates may not be equally distributed among treatment cohorts, and thereby influence the outcome variable independently of treatment. For example, in mediation

analysis, the confounders may have independent effects, or may mediate the effect of treatment Baron and Kenny (1986).

In order to adjust for confounders in non-randomized studies, many methods have been developed. Applied researchers have historically relied on regression methods to adjust for differences in baseline characteristics among groups, but this fails to account for selection bias if the covariates are associated with treatment assignment. *Propensity score* (PS) analysis has emerged as a widely used method to reduce the effects of confounding, and enable improved treatment effect estimates in OSs. This approach, introduced by Rosenbaum and Rubin (1983), utilizes the estimated conditional probability of being assigned to a certain treatment given a group of observed covariates, under the key assumption of no unmeasured confounders.

To understand how randomized trials and PS analysis using OSs compare for causal inference, we first need to consider attributes of the design and model that determine causal effects. Using the notation in Rosenbaum and Rubin (1983), suppose we have subjects $i = 1, \dots, n$ and z_i is the potential treatment assignment for the i^{th} subject, where $z_i = 0$ denotes the control group and $z_i = 1$ the active treatment group. Therefore, in the two-sample setting, we will have two potential outcomes (patient endpoints) $r(z_i)$ for the i^{th} subject: $r(0)$ under the control regimen and $r(1)$ under active treatment, though usually only one of them is observed. The one not observed is called the *counterfactual alternative* for the observed outcome. For each subject, the treatment effect then is defined as $r(1) - r(0)$ (Rosenbaum and Rubin, 1983; Imbens, 2004), whence the *average treatment effect* (ATE) in the population is defined as $E[r(1) - r(0)]$. A related treatment effect measurement is *average treatment effect for the treated* (ATT) (Imbens, 2004), which is defined as $E[r(1) - r(0)|Z = 1]$.

In RCTs, since treatments are assigned at random, we assume that, on average, treated subjects have similar characteristics to those in the competing study arms. As such, these two measures of treatment effects should coincide, and enable valid estimation of the causal relation between treatment and outcome. In OSs, however, usually it is necessary to assume that $E[r(1)] = E[r(1)|Z = 1]$ (and similarly for the control group) due to selection bias. In such cases, given covariates X , we can compute the PS as $e(X) = P(Z = 1|X)$. Rosenbaum and Rubin (1983) proved that $X \perp Z|e(X)$ (where “ $\perp$ ” denotes “independent”), and that $[r(1) - r(0)] \perp Z|e(X)$ given the assumption of “strongly ignorable treatment assignment” (i.e., $[r(1) - r(0)] \perp Z|X$). This implies that treatments may be viewed as being randomly assigned to subjects with roughly the same propensity score, yielding unbiased estimates of the ATE (Austin, 2011). Therefore, the causal relationship can be investigated in OSs with the critical assumption that there are no unknown confounders.

There are many methods for PS estimation, the most common of which regresses treatment assignment on baseline covariates through logistic regression (D’Agostino, 1998; Kurth et al., 2006; Rubin, 2008). Adjustment using these estimated propensity scores is then accomplished using one or a combination of four main methods: stratification (Rosenbaum and Rubin, 1984), matching (Rosenbaum and Rubin, 1985), inverse probability weighting (Lunceford and Davidian, 2004), or regression (covariate) adjustment (Austin, 2011).

Previous research has suggested that matching on PSs results in a greater reduction in selection bias than the other methods (Austin and Mamdani, 2006; Austin et al., 2007). Therefore, PS matching will be used here, but the other approaches could be used to effectuate matched pairs among the non-randomized patients. Rosenbaum and Rubin (1985) demonstrate three techniques using PSs to construct matched subjects: (i) nearest available matching on the estimated propensity score, (ii) Mahalanobis metric matching including the propensity score, and (iii) nearest available Mahalanobis metric matching within calipers defined by the propensity score. Additional algorithms were compared by Austin (2014) via simulation.

2.2 Combining RS and PS-matched NR Study Data

After accounting for measurable imbalances in attributes among the NR patients using the propensity-score based methods described in the previous subsection, we can use Bayesian hierarchical modeling to combine the NR and RS data for integrative analysis. Because randomization precludes systematic selection bias, the matching step only pertains to the NR cohort. This subsection describes models for integrating the NR and RS data.

Several authors have considered methods for incorporating historical data using Bayesian hierarchical modeling. Pocock (1976) advocated incorporating historical control data into clinical trial analysis under certain “acceptability” conditions. Ibrahim and Chen (2000) introduced *power prior* (PP) methods to down-weight the historical data relative to the current data. The PP approach assumes identical model parameters in the historical and current data and uses a weight (likelihood exponent) between 0 and 1 to control the extent to which the historical information influences the posterior. However, the weight is often difficult to estimate from the data, especially for non-Gaussian models.

Hobbs et al. (2011) proposed *commensurate priors* for linear and generalized linear mixed models to facilitate dynamic partial pooling of between-source information, wherein the extent of borrowing is estimated flexibly. This Bayesian hierarchical model assumes that the parameter vector θ for the current data follows a normal distribution with mean θ_0 estimated from historical data, and a precision or *commensurability* parameter τ . When evidence for commensurability is very strong between the two sources of data, τ will be sufficiently large that θ is constrained to be very close to θ_0 . On the other hand, when τ is close to 0, the variance of conditional prior for θ will be inflated, indicating weak commensurability and thus less borrowing between θ_0 and θ . In previous work, the goal was to develop hierarchical priors that facilitate a model where the degree of borrowing was driven by the similarity of the historical and current data, and thus the authors considered posterior inference using various types of priors for the commensurability parameters.

Similarly, we can link the modeling of matched NR data and unmatched RS data using commensurate prior framework. First, we can fit a Bayesian hierarchical model to the matched NR data as described by Agresti and Min (2004). Our clinical outcome can be a member of the exponential family (McCulloch and Searle, 2001), and here we assume it is binary and follows a Bernoulli distribution (logit link). Let the observations for pair m be (y_{m1}, y_{m2}) , where m refers to the matched pair index, and the second index $k = 1$ or 2 refers

to the control and treatment groups, respectively. Then we can impose a standard fixed treatment effects model for pair (y_{m1}, y_{m2}) as follows:

$$\log \left(\frac{\pi_{m1}}{1 - \pi_{m1}} \right) = \alpha_m \text{ and } \log \left(\frac{\pi_{m2}}{1 - \pi_{m2}} \right) = \alpha_m + \lambda_0, \quad (1)$$

where π_{mk} denotes the probability that $y_{mk} = 1$ and α_m the baseline effect for the control arm in the m^{th} pair. Here by allowing for pairwise specific intercepts, we implicitly assume that the matches themselves characterize varying magnitudes of prognostic effects, while the extent to which one treatment should be preferred with respect to another, λ_0 , is global. Therefore, we expect that the log-odds of obtaining a response is heterogeneous among the pairs, and we can account for this using a hierarchical model to estimate the extent of inter-pair heterogeneity. We do so by letting the α_m follow a normal distribution with mean μ_0 and precision parameter τ_1 , and assume λ_0 is a common treatment effect.

If we additionally let the treatment effects vary between pairs, we can switch to a random treatment effects model,

$$\log \left(\frac{\pi_{m1}}{1 - \pi_{m1}} \right) = \alpha_m \text{ and } \log \left(\frac{\pi_{m2}}{1 - \pi_{m2}} \right) = \alpha_m + \lambda_m, \quad (2)$$

where λ_m denotes the now pair-specific treatment effect for m^{th} pair relative to the baseline. This model acknowledges that the extent to which one treatment might be favored versus another varies by pair, suggesting an interaction between treatment and the covariates used for PS matching. For estimation, we can assume that the random effects λ_m are exchangeable with a common mean λ_0 and precision parameter τ_2 to characterize inter-pair heterogeneity in treatment effectiveness as a variance component. Without loss of generality, weakly informative priors can be used for these parameters. Specifically, the fixed effects (μ_0 and λ_0) are assumed to follow independent $N(0, 1000)$ distributions since these parameters are well informed by the data in this model, and we adopt conventional Inverse Gamma(0.1, 0.1) priors for the precisions τ_1 and τ_2 . While odds ratios are not collapsible across the pair index m (Greenland et al., 1999), this does not preclude us from assigning

them an exchangeable prior, $\lambda_m \stackrel{iid}{\sim} N(\lambda_0, \frac{1}{\tau_2})$. Since we assign τ_2 a weakly informative prior, we do not force the λ_m (or even their posterior mean) to be equal to λ_0 ; we merely encourage shrinkage of the λ_m toward their own grand mean.

For the unmatched RS data, we can incorporate the supplemental information from the matched NR study by fitting another general linear model with logistic link function

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mu + \lambda z_i. \quad (3)$$

Here i indexes subjects, μ is the baseline effect, λ is the coefficient for the treatment effect, and z is the 0 – 1 indicator for treatment arms. In order to incorporate the matched NR data,

a commensurate prior can be structured for λ so that $\lambda \sim N(\lambda_0, \frac{1}{\tau_3})$, where the commensurability parameter τ_3 (a precision) controls the degree of borrowing from the NR data. Here, we specify the prior for μ as a $N(0, 1000)$ distribution. However, as discussed in Hobbs et al. (2012), estimation of τ_3 is inherently difficult. Therefore, a “spike and slab” prior (Mitchell and Beauchamp, 1988) for τ_3 is more appropriate here, since it induces sparsity for estimating hierarchical variance components that are difficult to estimate from the data, and yields a dynamic borrowing procedures that has desirable bias-variance tradeoffs when integrating information that is potentially biased. In essence, this distribution is a mixture of a uniform distribution and a probability mass concentrated at a point $a > S_u$:

$$P(\tau_3 < S_l) = 0,$$

$$P(\tau_3 < v) = (1 - p_0) \frac{v - S_l}{S_u - S_l}, \quad S_l \leq v < S_u, \quad (4)$$

$$\text{and } P(\tau_3 \geq S_u) = P(\tau_3 = a) = p_0,$$

where S_u and S_l are the upper and lower bounds for a uniform distribution, and p_0 denotes the prior probability that τ_3 attains the value of the spike. Usually, we calibrate the values for a (the spike), the boundaries of the uniform distribution (the slab), and p_0 to represent different degrees of informative context (weakly informative to very informative). Sensitivity of the spike and slab prior to hyperparameter selection has been investigated in previous research. Murray et al. (2015) found that posterior inference for λ was not sensitive to modest shifts in these hyperparameters except for p_0 . These authors suggest further simplifying this prior, which we do by modifying it to a “two-spike” scale mixture prior $\lambda \sim p_0 N(\lambda_0, 1/R) + (1 - p_0) N(\lambda_0, 1/r)$, by introducing a second spike at a small value r . With this specification, one can either choose a value for p_0 or specify a hyperprior for it.

We can also reduce our hyperparameter specification burden by setting $S_l = 0$ and $S_u = a$ (a “two-parameter spike and slab” prior) when we have a smaller sample size. We used simulation to check the sensitivity of the two-parameter spike and slab prior to choice of hyperparameters a and p_0 , by describing each commensurate prior’s informative content by posterior *effective supplemental sample size* (ESSS), similar to the effective historical sample size (EHSS) in equation (3) of Hobbs et al. (2013). In our setting, let $P(y_0, y, \tau_3)$ be the posterior precision using the combined data, let $P(y)$ be the posterior precision using only the randomized data, and n be the sample size of the randomized data. Then the ESSS

when using the commensurate prior can be approximated as $n \left(\frac{P(y_0, y, \tau_3)}{P(y)} - 1 \right)$, which is approximately the effective sample size of the joint posterior minus the randomized sample

size. So the ESSS characterizes the effective number of additional randomized patients we gain by using the Bayesian hierarchical model. If there is little gain in precision after incorporating the non-randomized data, the ESSS is small. Our results show that when we fix p_0 and let a increase, the ESSS increases sharply until a is about 50, and levels off after that. By contrast, the ESSS is insensitive to the choice of p_0 under different spike values a .

3 Results from the FIRST Trial

3.1 Baseline Characteristics and Naive Modeling

In this section, we demonstrate our method with an application using the FIRST dataset from Section 1.2. Baseline covariates were compared between treatment groups using the standard t test for continuous variables and the chi-squared test for categorical variables, respectively. Table 2 gives summaries of available covariates, including age, race (white vs. others), progression of disease before randomization or not (“podbl”, 1: yes), baseline average cd4 count (“cd4bl”), log baseline HIV RNA level (“lnabb”), gender (1: male), malesex indicating homosexual activity for men (1: yes), and injection drug use (“idu”, 1: yes) for NVP-treated or EFV-treated participants in both the RS and NR groups. Results show that the baseline covariates were acceptably balanced between treatment groups in the randomized cohort, while one covariate (baseline average CD4 count) differed significantly between the NVP or EFV participants in the NR cohort (unadjusted p-value = 0.03). This could be the result of an early study (Leth et al., 2005) that suggested EFV was more efficacious for low CD4 patients. NVP was also suspected of causing liver problems (Stern et al., 2003), whereas EFV’s side effects (dreams, night-mares) were thought to be less severe, possibly further explaining the physicians’ overall preference for EFV (211 vs 100).

Let $Y_i = 1$ if patient i experiences virological suppression (VS) at the 8-month visit. We assume $Y_i \sim \text{Bernoulli}(\pi_i)$ where π_i denotes the probability of VS for $i=1, \dots, N_c$, $c = RS, NR$, where N_c represents sample size for either the RS or NR cohort. Now let $x_1, \dots, x_p$ denote p baseline covariates, and z an indicator variable for the intervention group (such that $z_i = 1$ is NVP; $z_i = 0$ is EFV). Then we can compare the odds of VS between treatments using a generalized version of model (3),

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = (\mathbf{X}\mathbf{b})_i + \lambda z_i, \quad (5)$$

where $\mathbf{b} = (\beta_1, \dots, \beta_8)'$, $\mathbf{X}$ is a $N_c \times 8$ design matrix, and λ is the log-odds ratio of VS for NVP versus EFV. If we perform a naive frequentist analysis for the randomized and NR data respectively without any PS adjustment in the NR cohort, we find that the treatment effects between NVP and EFV groups were significantly different in the NR cohort (p-value = 0.02), but showed no significant difference in the randomized cohort (p-value = 0.24). More surprisingly, the results differed in the direction of the effect. Specifically, the log-odds ratio λ using EFV as the reference group was -0.68 with a 95% confidence interval (CI) $[-1.25, -0.11]$ in the NR cohort, indicating an increase in the odds of VS for patients receiving EFV relative to those receiving NVP, but was 0.79 in the RS cohort with a 95% CI $[-0.54, 2.13]$,

indicating a relative *decrease* in odds of VS for patients receiving EFV. The smaller sample size in the RS cohort leads to the decrease in precision (wider interval estimate).

Motivated by this discrepancy, we seek to combine the RS and NR data using the methods described in Section 2.2. Specifically, we will use PS matching within the NR cohort, and then combine the information with the RS data using Bayesian hierarchical commensurate prior models, which gives privilege to the RS data.

3.2 Combining RS and PS-matched NR Study Data

As mentioned in Section 2.1, we first estimate PS in the NR cohort using logistic regression, i.e., we regress the probability of being assigned to the NVP group on the 8 covariates described in Table 2. For the FIRST dataset, after we calculate the PSs using the NR cohort, we must match participants between its NVP and EFV arms. After randomizing the order of the observations within each arm, we match one treatment observation with a control observation, subject to a maximum allowable difference between their propensity scores. This maximum allowable difference is called the *caliper width*, and is usually designated by simply assigning a reasonable but subjective value that determines the extent of allowable dissimilarity among matched pairs; for example, 0.1. A common criticism of 1:1 matching is that it may discard a large number of observations and lead to reduced power. But according to Stuart (2010), 1:1 matching may have advantages compared to k:1 matching due to several reasons; e.g., if the treatment group stays the same size and only the control group decreases in size, power may not be markedly reduced (Ho et al., 2007). After each pair is matched, it is removed from the pool. This procedure is repeated until all NVP patients are matched to the EFV patients, or until no further EFV observations fulfill the matching criteria (in our case, resulting in 89 matched pairs). After matching, balance diagnostics, e.g., those proposed by Austin (2008), can be performed to assess whether the propensity score model has been adequately specified. At this point, all the covariates should be balanced between NVP and EFV groups. In our data, we obtained a p-value of 0.81 for the baseline average CD4 count in the NR study using a paired *t* test, suggesting that the procedure achieved reasonable balance among the matched pairs. The log-odds ratio λ was re-estimated using the PS-matched NR data, resulting in a point estimate of -0.72 with a 95% CI $[-1.43, -0.01]$ (p-value = 0.04).

Results obtained from combining the NR and RS data using the Bayesian hierarchical model described in Section 2.2 are summarized here for the FIRST dataset. For computation, we used OpenBUGS (Lunn et al., 2009) to draw two Markov chain Monte Carlo (MCMC) chains from the posterior for 30,000 iterations each, after 20,000 iterations of burn-in. The final estimate of the log-odds ratio λ was -0.30 with a 95% Bayesian credible interval (BCI) of $[-0.88, 0.32]$ for the fixed effects model using the two parameter spike and slab prior. Here, we chose $a = 40$ and $p_0 = 0.3$, a moderately informative prior justified by the relatively high quality of the NR data (all subjects, both NR and RS, met the FIRST entry criteria). We call this prior “moderately informative” since from simulation the effective supplemental sample size equals 53 when the sample size of the randomized group is 100, and equals 154 when the sample size of the randomized group is 425, both modest relative to the total sample sizes. The random effects model (2) was also fitted, but not reported here since its results

were similar. The pooled treatment effect of -0.30 lies in between those obtained using either the RS or matched NR data alone, indicating an increase in the odds of VS for EFV patients compared to NVP subjects, though the increase is not statistically significant based on its BCI. The BCI width (1.20) for the log-odds ratio λ is smaller than those obtained using either the RS data alone (2.67) or the matched NR data alone (1.42). The posterior mean of τ_3 was estimated as 25.85 using the fixed effects model, showing a moderate degree of borrowing from the NR data.

Our results are consistent with the findings in van den Berg-Wolf et al. (2008), showing a lower rate of virological failure in persons taking EFV as compared to those taking NVP, although our combined data result is not statistically significant. This might be because we cautiously borrowed from the matched subset of NR data, while they combined all the NR and RS data from both the NNRTI and 3-class strategies, making the strong assumption that these data sources are exchangeable and equally free of bias. They also found an *increased* risk of disease progression or death for those randomized to EFV when compared to those randomized to NVP. These findings indicate pros and cons of using either medication in terms of different clinical endpoints. Our results suggest physicians may rely on their clinical judgment when prescribing these drugs, keeping in mind the expected risk-benefit trade-off for their patients.

4 Simulation Studies

We use a series of simulations to examine the performance of our models for combining NR data with randomized data using commensurate priors after propensity score matching. As in our FIRST example, here we consider only binary outcomes.

4.1 Simulation Settings

We simulated data with 4 baseline covariates (x_1 to x_4) for both the NR or RS data. These covariates were simulated from independent $N(0,1)$ distributions. Among these 4 covariates, we designated two of them as affecting treatment selection (x_1 and x_2) for the NR data, while two other covariates (x_2 and x_3) affect the binary outcome for both datasets. Furthermore, these covariates were allowed to have a weak, moderate, strong, or very strong effect on treatment selection or outcome corresponding to β values of $\log(1.25)$, $\log(1.5)$, $\log(1.75)$ and $\log(2)$, respectively. For each subject, the true probability of treatment selection (propensity score) was determined from the following logistic model:

$$\log \left(\frac{PS_i}{1 - PS_i} \right) = \beta_{00} + \beta_{01}x_{1i} + \beta_{02}x_{2i}, \quad (6)$$

the treatment selection model. Therefore, the true treatment status for each patient i was generated from a Bernoulli distribution with a subject-specific parameter PS_i denoting the probability of being assigned to the treatment group: $Z_i \sim \text{Bernoulli}(PS_i)$. We then generated a binary outcome using the formula:

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \lambda Z_i, \quad (7)$$

the outcome model. Similarly, the outcome Y_i was then generated from a Bernoulli distribution using $Y_i \sim \text{Bernoulli}(\pi_i)$.

Several scenarios were simulated to test the degree of borrowing when we used different spike and slab commensurate priors for τ_3 . In Scenario 1, we assumed that the same model applied to the outcomes for both the NR and RS data, where β_2 and β_3 were set to $\log(1.25)$, $\log(1.5)$ respectively. The true treatment effect λ was set to -0.7 when the total sample size was 200 for two cohorts (100 patients in each cohort), indicating an odds ratio of 0.5 comparing treatment to placebo. We also set the true treatment effect to -0.4 for a sample size 850 (425 patients in each cohort), giving an odds ratio of 0.67. In Scenario 2, we kept the same true treatment effects λ for the NR and RS data; however, different sets of β coefficients (see (7)) for the covariates were used in the outcome models. Specifically, for the NR data, we used $\log(1.75)$ and $\log(2)$ for β_2 and β_3 , while for the RS data, $\log(1.25)$ was used for both β_2 and β_3 . In Scenario 3, we assumed that the treatment effects were very different between the NR and RS data (0.7 for the RS cohort but -0.7 for the NR cohort), while β_2 and β_3 in the outcome model stayed the same for both cohorts as in Scenario 1. In Scenario 4, we assumed not only that the treatment effects λ were very different in the NR and RS data (0.7 for the RS cohort but -0.7 for the NR cohort), but also used the same β_2 and β_3 in the outcome models as in Scenario 2. Besides the above scenarios, we also tested scenarios in which we predicted the PSs from the treatment selection model (see (6)) using only x_1 and x_2 , only x_3 and x_4 , or all four covariates (x_1 to x_4). Results showed that these models perform similarly as long as we predict the PSs using covariates in the true model (x_1 and x_2 in our case); therefore, we will use x_1 to x_4 to predict PS in all settings.

We checked three different commensurate priors under each scenario as follows: (1) a two parameter spike and slab prior with $a = 40$ and $p_0 = 0.3$ ("CP1"), (2) a two parameter spike and slab prior with $a = 1000$ and $p_0 = 0.01$ ("CP2"), and (3) a two-spike prior with $R = 2000$, $r = 0.01$ and $p_0 = 0.1$ ("CP3"). The results using all three commensurate priors were also compared to the model where we used only the RS data ("No CP"). In each of these settings, we generated 1000 simulated datasets (each of sample size 200 or 850) using R, where we used the Brugs package to call OpenBUGS, once for each simulated dataset. Using each of the 1000 simulated datasets, we performed the propensity score matching based on *all* the baseline covariates x_1 to x_4 in the generated NR data. We then used the fixed effects model for paired data as shown in (1) on the propensity-score-matched sample. Finally, model (3) was used to estimate the treatment effect in the RS data using each commensurate prior for each simulated sample. The average posterior bias, 95% BCI width, and mean squared errors (MSE) for all the parameters (β_2 , β_3 and λ) were calculated across the 1000 simulated datasets. Empirical coverages of the 95% BCIs for parameter λ were calculated, and the power for finding a significant treatment effect was estimated as the empirical proportion of these BCIs that did not contain zero out of 1000 simulations under the alternative hypothesis (true treatment effect was different from 0).

4.2 Simulation Results

In the construction of the true models above, we are imposing increasing heterogeneity between the NR and RS studies from Scenario 1 to Scenario 4. Therefore, we expect our models to capture this trend using commensurate priors for λ , borrowing less from the NR data when heterogeneity is high (i.e., we expect τ_3 to be small in models with CP1 and CP2, or ρ_0 to be small in models with CP3).

Tables 3 and 4 display the posterior estimates for the main parameters λ , τ_3 (or ρ_0 under CP3), MSEs for λ , β_2 and β_3 , as well as 95% empirical coverages for λ and power averaged over the 1000 simulated datasets under each scenario described in Subsection 4.1. In Table 3, under Scenarios 1 and 2 where it is beneficial to use the NR studies (since the true treatment effects λ were the same across different sources of data), models with CP1 and CP2 (moderately informative to informative priors) performed better since they gave narrower 95% BCIs for λ that did not cover 0 (high power), indicating a statistically significant treatment effect. Moreover, the posterior estimates for τ_3 were large using models with CP1 and CP2, showing a substantial degree of borrowing from the NR data. These models yielded better (10% to 20% higher) power compared to models with No CP. We also obtained about 35% reduction in MSE for the CP1-2 approach and about 20% for CP3 compared to models with No CP, where less MSE reduction was observed in the absence of bias as the trade-off. In the situation when we only used the RS data, we see the widest 95% BCI, covering 0 even when we used a larger sample size, indicating a non-significant treatment effect. The model with CP3 showed moderate borrowing from the NR data under first two scenarios, since this prior is weakly informative compared to CP1 and CP2.

In Table 4 under Scenarios 3 and 4 where we gave opposite signs for the treatment effects for the NR and RS data, the above results are essentially reversed, since now the true model assumes high heterogeneity between different sources of data. The posterior biases for λ are much larger using models with CP1 and CP2 compared to those from models with CP3 or No CP, especially when the sample size is small. Furthermore, although the posterior mean estimates for τ_3 decreased compared to those under Scenario 1 for models with CP1 and CP2, these priors are informative in the sense that they represent a strong preference for borrowing strength across cohorts. This is revealed by the power estimates in Table 4, where models with CP1 and CP2 resulted in very low power (0.09 and 0.07 for Scenario 3; 0.11 and 0.09 for Scenario 4) for studies with sample size 200, and similar decreases relative to No CP and CP3 for sample size 850. By contrast, the model with CP3 resulted in more adaptive borrowing from the NR data, e.g., the posterior mean estimate for ρ_0 decreased from 0.65 in Scenario 1 to 0.03 in Scenario 3 for sample size 850. Moreover, the power from model with CP3 under Scenario 3 is 0.3 and 0.83 for sample sizes 200 and 850, respectively (similar results as No CP), indicating good performance under the no-borrowing scenario using this weakly informative prior. Results from Scenario 4 were similar to those from Scenario 3, showing similar degree of heterogeneity between different sources of data under these scenarios.

Moreover, results shows that Type-I errors for different models under various settings were well controlled under 0.05, showing no inflation problems (results omitted here).

5 Discussion and Future Work

In this paper, we developed a practical Bayesian statistical tool to combine OSs and RCTs. This was achieved using hierarchical models with priors that facilitate adaptive borrowing from matched NR data when justified by its commensurability with the corresponding RS data. Although our FIRST dataset is somewhat unique given that the same trial produced both the NR and RS cohorts, our model can be applied more broadly to integrate more and different kinds of OS data. Our approach also has the benefit of increasing the external validity of our results, because in practice only a select subpopulation is willing to be randomized. Borrowing strength is most appropriate when the two data sources are homogeneous, as in our FIRST dataset. However, in other settings, where we have more broadly collected data sources which include lower-quality OSs, models with weakly informative two-spike priors like CP3 would appear to be more appropriate, given their flexibility for adjusting to evidence of heterogeneity across studies. This is also true when we do not have any prior information regarding the NR studies. As such, our method is very much in the spirit of the 21st Century Cures Act to streamline drug approvals, currently working its way through the U.S. House of Representatives (Upton et al., 2015; Avorn and Aaron, 2015).

Our proposed method better utilizes all available data, but it has certain limitations. First, its improvement in precision may not outweigh its potential for bias arising from “low-quality” PS matches. This is because “highest-quality” matches would arise only from identifying pairs of patients who are likely targets for the same treatment, but receive opposite treatments. As a somewhat controversial alternative to our approach in Section 2.1, where we first perform PS estimation in only the NR cohort, we could instead use all the data (both RS and NR) in our PS matching. Specifically, for the randomized data, we no longer assume their PSs are fixed at 0.5, but rather estimate the probability of treatment assignment using data from *all* subjects. That is, we “predict” the PS for randomized subjects from the fitted partial regression coefficients obtained from the NR data alone. These PSs then have the interpretation of a measure of the probability that the patient would have been assigned to the treatment arm using the intrinsic “treatment selection”. After PSs have been calculated for all patients, matching *between* the RS and NR data (e.g., matching one randomized control patient with high PS to one non-randomized treated patient with similarly high PS) enables estimation of the treatment selection counterfactuals. Similarly, these counterfactuals can then be used to estimate the average treatment effect using Bayesian hierarchical models for paired binary data as described in Section 2.2. The advantage of this method is that it is more likely to use high-quality matches (counterfactuals), but at the cost of disassembling the “gold standard” RS data, and possibly also discarding many observations in the (much larger) NR study due to the matching across cohorts.

Future methodological research in this area looks to extending our Section 2.2 approach to permit more types of auxiliary data to be incorporated into network meta-analysis. Other possible research directions include seeking non-propensity score methods to mitigate treatment selection bias. The instrumental variables (IV) method is widely used in economics because of the difficulty of doing controlled trials (Newhouse and McClellan, 1998). Compared to PS methods, it is preferred by some researchers due to unobservable

confounding in OSs. The main idea of the IV method is to define variables called *instruments* that have two properties. First, they should highly correlate with the treatment choice, and second, they must not be directly related to the outcome measure. Given such variables, one can estimate how much the instrument induces variation in the treatment assignment, which later affects the outcome. A key assumption of the IV method is that there is no direct association between the IV and the outcome except through the treatment variable, since one can then think of the IV as a device to achieve a pseudo-randomization, akin to the coin flipping. The difficulty in IV method is how to find good instrumental variables and how to validate their selection. This would be especially challenging in our FIRST dataset, since we will need two IVs: one for the patient's choice to accept or decline the substudy randomization, and a second for the clinician's choice of drug (NVP or EFV) for those declining. Still, we hope to compare results from IV methods to those from PS matching methods using our FIRST study and through simulation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Dr. Thomas Murray for helpful discussions regarding the two-spike prior distribution and other aspects of commensurate prior modeling, and to Drs. James Neaton and Kathy Huppler-Hullsiek for sharing the FIRST data and providing valuable insights regarding its proper interpretation.

Funding: The work of the first and last authors was supported in part by a grant from the Amgen Research Grant Program. The work of the second and last authors was supported in part by National Cancer Institute Grant 1-R01-CA157458-01A1. Finally, the work of the second author was supported in part by National Cancer Institute M.D. Anderson Cancer Center Support Grant P30-CA016672.

References

- Agresti A, Min Y. Effects and noneffects of paired identical observations in comparing proportions with binary matched pairs data. *Statistics in Medicine*. 2004; 23:65–75. [PubMed: 14695640]
- Austin P. A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*. 2008; 27:2037–2049. [PubMed: 18038446]
- Austin P. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*. 2011; 46:399–424. [PubMed: 21818162]
- Austin P. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*. 2014; 33:1057–1069. [PubMed: 24123228]
- Austin P, Grootendorst P, Anderson G. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*. 2007; 26:734–753. [PubMed: 16708349]
- Austin P, Mamdani M. A comparison of propensity score methods: A case study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*. 2006; 25:2084–2106. [PubMed: 16220490]
- Avorn J, Aaron S. The 21st Century Cures Act - will it take us back in time? *New England Journal of Medicine*. 2015; 372:2473–2475. [PubMed: 26039522]
- Baron R, Kenny D. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51:1173–1182. [PubMed: 3806354]

- Concato J, Shah N, Horwitz R. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*. 2000; 342:1887–1892. [PubMed: 10861325]
- D'Agostino R. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–2281. [PubMed: 9802183]
- Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA: the Journal of the American Medical Association*. 1992; 268:2420–2425. [PubMed: 1404801]
- Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science*. 1999; 14:29–46.
- Ho D, Imal K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*. 2007; 15:199–236.
- Hobbs B, Carlin B, Mandrekar S, Sargent D. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011; 67:1047–1056. [PubMed: 21361892]
- Hobbs B, Carlin B, Sargent D. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*. 2013; 10:430–440. [PubMed: 23690095]
- Hobbs B, Sargent D, Carlin B. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*. 2012; 7:639–674. [PubMed: 24795786]
- Ibrahim J, Chen M. Power prior distributions for regression models. *Statistical Science*. 2000; 15:46–60.
- Imbens G. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*. 2004; 86:4–29.
- Ioannidis J. Contradicted and initially stronger effects in highly cited clinical research. *JAMA: the Journal of the American Medical Association*. 2005; 294:218–228. [PubMed: 16014596]
- Kurth T, Walker A, Glynn R, Chan K, Gaziano J, Berger K, Robins J. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*. 2006; 163:262–270. [PubMed: 16371515]
- Leth FV, Andrews S, Grinsztejn B, Wilkins E, Lazanas M, Lange J, Montaner J. The effect of baseline CD4 cell count and HIV-1 viral load on the efficacy and safety of nevirapine or efavirenz-based first-line HAART. *AIDS*. 2005; 19:463–471. [PubMed: 15764851]
- Ligthelm R, Borzi V, Gumprecht J, Kawamori R, Wenying Y, Valensi P. Importance of observational studies in clinical practice. *Clinical Therapeutics*. 2007; 29:1284–1292.
- Lunceford J, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 2004; 23:2937–2960. [PubMed: 15351954]
- Lunn D, Spiegelhalter DJ, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*. 2009; 28:3049–3067. [PubMed: 19630097]
- MacArthur R, Chen L, Mayers D, Besch C, Novak R, van den Berg-Wolf M, Yurik T, Peng C, Schmetter B, Brizz B, Abrams D. The rationale and design of the CPCRA (Terry Bein Community Programs for Clinical Research on AIDS) 058 FIRST (Flexible Initial Retrovirus Suppressive Therapies) trial. *Controlled Clinical Trials*. 2001; 22:176–190. [PubMed: 11306155]
- MacLehose R, Reeves B, Harvey I, Sheldon T, Russell I, Black A. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment*. 2000; 4:1–154.
- McCulloch, CE., Searle, SR. *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons; 2001.
- Mitchell T, Beauchamp J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*. 1988; 83:1023–1032.

- Murray T, Hobbs B, Carlin B. Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *Annals of Applied Statistics*. 2015; 9:1549–1570. [PubMed: 26557211]
- Newhouse J, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*. 1998; 19:17–34.
- Pearl J. Causal inference in statistics: An overview. *Statistics Surveys*. 2009; 3:96–146.
- Pocock S. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*. 1976; 29:175–188. [PubMed: 770493]
- Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rosenbaum P, Rubin D. Reducing bias in observational studies using sub-classification on the propensity score. *Journal of the American Statistical Association*. 1984; 79:516–524.
- Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985; 39:33–38.
- Rubin D. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*. 2008; 2:808–840.
- Stern J, Robinson P, Love J, Lanes S, Imperiale M, Mayers D. A comprehensive hepatic safety analysis of nevirapine in different populations of hiv infected patients. *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2003; 34:S21–S33. [PubMed: 14562855]
- Stuart E. Matching methods for causal inference: a review and a look forward. *Statistical Science*. 2010; 25:1–21. [PubMed: 20871802]
- Upton F, DeGette D, Pitts J, Pallone F, Green G. 21st Century Cures Act. 2015
- U.S. Preventive Services Task Force. Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force. Baltimore: Williams & Wilkins; 1996.
- van den Berg-Wolf M, Hullsiek K, Peng G, Kozal M, Novak R, Chen L, Crane L, MacArthur R. Virologic, immunologic, clinical, safety, and resistance outcomes from a long-term comparison of Efavirenz-based versus Nevirapine-based antiretroviral regimens as initial therapy in HIV-1-infected persons. *HIV Clinical Trials*. 2008; 9:324–336. [PubMed: 18977721]

Figure 1.

Outline of FIRST design and randomization for eligible subjects.

Table 1

Grades of evidence: hierarchical rankings for different study designs (U.S. Preventive Services Task Force, 1996).

I	Evidence obtained from at least one properly randomized, controlled trial.
II-1	Evidence obtained from well-designed controlled trials without randomization.
II-2	Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.
II-3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III	Opinions of respected authorities, based on clinical experience; descriptive studies and case reports; or reports of expert committees.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Comparison of baseline characteristics between FIRST treatment groups in the randomized and non-randomized cohorts prior to matching. Continuous variables are reported as mean $\pm$ standard deviation. Dichotomous variables are reported as N (Proportion)

	Randomized cohort			Non-randomized cohort		
	EFV (N=45)	NVP (N=53)	p-value	EFV (N=211)	NVP (N=100)	p-value
AGE	39.0 $\pm$ 7.6	36.7 $\pm$ 8.5	0.17	38.6 $\pm$ 9.9	38.9 $\pm$ 8.4	0.78
RACE (RACE=3)	12 (26.7)	14 (26.4)	0.98	61 (28.9)	26 (26.0)	0.59
podbl (=1)	16 (35.6)	22 (41.5)	0.55	83 (39.3)	38 (38.0)	0.82
cd4bl	227.7 $\pm$ 207.3	209.5 $\pm$ 193.0	0.66	190.2 $\pm$ 189.0	242.9 $\pm$ 227.3	0.03
lnrabb (log10(rnabl))	5.1 $\pm$ 0.9	5.2 $\pm$ 0.8	0.56	5.1 $\pm$ 0.8	4.9 $\pm$ 0.8	0.08
GENDER (=male)	35 (77.8)	40 (75.5)	0.79	167 (79.2)	77 (77.0)	0.67
malesex (=1)	21 (46.7)	22 (41.5)	0.61	100 (47.4)	41 (41.0)	0.29
idu (=1)	5 (11.1)	11 (21.2)	0.18	25 (11.9)	17 (17.0)	0.22

Table 3

Posterior estimates and MSEs for key parameters, 95% empirical coverages for λ , and power using simulated datasets under Scenarios 1 and 2 (borrowing warranted). Each cell represents an average over 1000 simulations.

scenario	sample size	true λ		priors for τ_3	posterior estimates for λ		τ_3 or p_0 mean	MSE			95% coverage	power
		NR	RS		bias	95% BCI width		β_2	β_3	λ		
1	200	No CP		No CP	-0.07	1.84	-	0.13	0.15	0.48	0.93	0.40
		CP 1		CP 1	-0.08	1.49	26.07	0.12	0.14	0.31	0.94	0.54
		CP 2		CP 2	-0.08	1.42	504.68	0.13	0.14	0.30	0.93	0.59
		CP 3		CP 3	-0.07	1.73	0.47	0.13	0.15	0.41	0.94	0.44
	850	No CP		No CP	-0.01	0.82	-	0.02	0.03	0.09	0.95	0.50
		CP 1		CP 1	0.00	0.71	26.89	0.02	0.03	0.06	0.96	0.62
		CP 2		CP 2	0.00	0.66	505.31	0.02	0.03	0.06	0.95	0.69
		CP 3		CP 3	0.00	0.75	0.65	0.02	0.03	0.07	0.96	0.59
	200	No CP		No CP	-0.08	1.81	-	0.13	0.14	0.48	0.93	0.41
		CP 1		CP 1	-0.04	1.48	26.04	0.12	0.13	0.31	0.94	0.51
		CP 2		CP 2	-0.03	1.41	504.70	0.12	0.13	0.29	0.93	0.54
		CP 3		CP 3	-0.07	1.71	0.46	0.13	0.13	0.41	0.94	0.44
	850	No CP		No CP	-0.01	0.81	-	0.02	0.02	0.09	0.96	0.50
		CP 1		CP 1	0.01	0.70	26.95	0.02	0.02	0.06	0.96	0.61
		CP 2		CP 2	0.01	0.65	505.81	0.02	0.02	0.06	0.95	0.67
		CP 3		CP 3	0.00	0.74	0.66	0.02	0.02	0.07	0.96	0.58

Table 4

Posterior estimates and MSEs for key parameters, 95% empirical coverages for λ , and power using simulated datasets under Scenarios 3 and 4 (borrowing not warranted). Each cell represents an average over 1000 simulations.

scenario	sample size	true λ		priors for τ_3	posterior estimates for λ			τ_3 or p_0 mean			MSE			95% coverage	power
		NR	RS		bias	95% BCI width					β_2	β_3	λ		
3	200	-0.7	0.7	No CP	0.16	1.93	-	-	-	-	0.16	0.18	0.65	0.93	0.36
				CP 1	-0.40	1.80	21.41	0.15	0.17	0.60	0.83	0.09	0.07	0.07	0.30
				CP 2	-0.55	1.59	479.21	0.15	0.17	0.66	0.69	0.07	0.07	0.07	0.30
				CP 3	0.08	2.12	0.20	0.16	0.18	0.67	0.93	0.93	0.93	0.93	0.30
	850	-0.7	0.7	No CP	0.02	0.91	-	0.03	0.03	0.11	0.95	0.87	0.60	0.60	0.40
				CP 1	-0.16	0.95	8.31	0.03	0.03	0.16	0.86	0.60	0.40	0.40	0.83
				CP 2	-0.26	0.97	201.45	0.03	0.03	0.23	0.74	0.40	0.83	0.83	0.83
				CP 3	0.01	0.93	0.03	0.03	0.03	0.12	0.94	0.83	0.83	0.83	0.83
	200	-0.7	0.7	No CP	0.16	2.08	-	0.16	0.16	0.64	0.93	0.38	0.11	0.11	0.09
				CP 1	-0.37	1.75	22.25	0.15	0.16	0.56	0.84	0.11	0.09	0.09	0.32
				CP 2	-0.49	1.57	484.00	0.15	0.15	0.59	0.74	0.09	0.32	0.32	0.32
				CP 3	0.07	2.09	0.22	0.16	0.16	0.65	0.92	0.32	0.32	0.32	0.32
4	850	-0.7	0.7	No CP	0.02	0.91	-	0.03	0.03	0.11	0.95	0.88	0.61	0.61	0.37
				CP 1	-0.16	0.94	9.45	0.03	0.03	0.16	0.87	0.61	0.37	0.37	0.82
				CP 2	-0.28	0.95	240.64	0.03	0.03	0.24	0.71	0.37	0.82	0.82	0.82
				CP 3	0.01	0.93	0.04	0.03	0.03	0.12	0.95	0.82	0.82	0.82	0.82