Published in final edited form as:

Stat Biosci. 2011 June 21; 3(1): 63–78. doi:10.1007/s12561-011-9035-4.

Using Multiple Control Groups and Matching to Address Unobserved Biases in Comparative Effectiveness Research: An Observational Study of the Effectiveness of Mental Health Parity

Frank B. Yoon,

Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA U.S.A. 02115 Tel.: +1 617-432-0006 Fax: +1 617-432-0173

Haiden A. Huskamp,

Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA U.S.A. 02115 Tel.: +1 617-432-0838 Fax: +1 617-432-0173

Alisa B. Busch, and

Clinical Services Department, McLean Hospital, 115 Mill Street, Belmont, MA U.S.A. 02478 Tel: +1 617-855-2989

Sharon-Lise T. Normand

Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA U.S.A. 02115 Tel.: +1 617-432-3260 Fax: +1 617-432-0173

Frank B. Yoon: yoon@hcp.med.harvard.edu; Haiden A. Huskamp: huskamp@hcp.med.harvard.edu; Alisa B. Busch: abus@hcp.med.harvard.edu; Sharon-Lise T. Normand: sharon@hcp.med.harvard.edu

Abstract

Studies of large policy interventions typically do not involve randomization. Adjustments, such as matching, can remove the bias due to observed covariates, but residual confounding remains a concern. In this paper we introduce two analytical strategies to bolster inferences of the effectiveness of policy interventions based on observational data. First, we identify how study groups may differ and then select a second comparison group on this source of difference. Second, we match subjects using a strategy that finely balances the distributions of key categorical covariates and stochastically balances on other covariates. An observational study of the effect of parity on the severely ill subjects enrolled in the Federal Employees Health Benefits (FEHB) Program illustrates our methods.

Keywords

Causal Inference; Fine Balance; Quasi-Experiments; Testing in Order

1 Introduction

In 1999 President Clinton directed the Office of Personnel Management to ensure that all health plans participating in the Federal Employees Health Benefits (FEHB) Program offer parity in coverage between mental health and substance abuse (MH/SA) benefits and general medical benefits (US Office of Personnel Management, 2000). This comprehensive parity policy was implemented in January 2001. Using a difference-in-difference analysis of administrative data from seven FEHB plans and a comparison group of private plans not

subject to the parity policy, an evaluation found implementation of the policy had little effect on average total spending but statistically significant reductions in average out-of-pocket spending for enrollees in most plans (Goldman et al. 2006; Busch et al. 2006). The evaluation, however, did not examine the effects on the subgroup of enrollees with the most severe mental disorders, a group of enrollees who are the primary target of parity policies. In this paper we examine the effects on this subgroup.

We make use of a matched design to convince policy makers that the parity effect is real and not some circumstance of unobserved bias. First, we use multiple comparison groups rather than a single comparison group. The decision to use two comparisons groups provides a mechanism for testing a set of sequential hypotheses and affords a principled strategy for testing within the setting where randomization is absent. Second, we use matching of parity and non-parity beneficiaries to ensure we are comparing individuals who are similar on the observed confounders. Because of the discrete nature of the data collected for this study and for many studies like this, we use a combinatorial algorithm to exactly balance nominal variables without trying to match enrollees exactly on such variables. The algorithm enables stochastic balance on many covariates and close matching on a few important covariates.

We next provide background and motivation for the parity directive and for our choice of the matched design and analysis. In Section 2 we describe our design in fuller detail and in Section 3 our approach to inference. Section 4 provides results of the parity study, and in Section 5 we conclude with some final remarks.

1.1 Mental health parity

A well-functioning insurance market protects against the risk of large, potentially catastrophic financial losses. Historically, protection from catastrophic losses resulting from MH/SA service use has been unavailable in most private insurance plans. Typically, plans have limited the number of MH/SA inpatient days and outpatient visits covered, and used higher deductibles and cost sharing for MH/SA services than for general medical care (Barry et al. 2003). These limits are thought to harm the efficiency and fairness of insurance for MH/SA services in general, and to particularly disadvantage individuals with severe mental disorders who have high treatment costs (Frank, Goldman, and McGuire 2001).

The findings from the parity evaluation reported that the implementation of parity was associated with small changes in out-of-pocket spending but had little effect on average total spending (Goldman et al. 2006). Because removing special benefit limits for MH/SA services should disproportionately reduce the burden on individuals with high MH/SA expenditures, the small changes in average out-of-pocket spending may have been driven by much larger reductions in out-of-pocket spending among those with more severe mental disorders experiencing high cost treatments. Alternatively, small changes in average out-of-pocket spending may have been driven by decreases in out-of-pocket spending among individuals with low levels of MH/SA expenditures, while those with more severe disorders and higher out-of-pocket spending may have experienced an increase in out-of-pocket burden if they increased service use in response to the parity benefit expansions. We propose methods to assess the effect of parity on enrollees in the FEHB study who presented at baseline with a claim for a severe diagnosis such as schizophrenia (Table 1). The policy outcomes we consider are binary indicators for whether a particular type of service was utilized in the follow-up year; any type of MH/SA service; any inpatient MH/SA service; therapy; medication management; and prescriptions.

1.2 What matching can and cannot do

The major study by Goldman et al. (2006) used a difference-in-difference (DID) modeling approach that used repeated measures of FEHB Program enrollees and comparison enrollees before and after the implementation of parity; the comparison enrollees did not receive parity benefits. The DID approach has the important benefit of being robust against certain patterns of unobserved confounding. In particular, DID estimates are robust against an unobserved covariate that changes in time identically in both the intervention and comparison groups, so that, had the intervention been absent, the slopes of the two regression lines over time would have been parallel. In the parity study, this approach accounts for any secular trends in time that could explain changes in the utilization of MH/SA services. However, a DID analysis does not anticipate other patterns of unobserved confounding. For example, if the study groups evolve differently over time, such as their perceptions about seeking MH/SA care and treatment, then a selection-by-maturation interaction occurs and potentially introduces biases in the analysis (Shadish, Cook, and Campbell 2002).

In contrast to the DID modeling approach, which requires correct specification of the model terms, matching has the benefit of structuring the analysis such that its conclusions do not depend on extrapolation beyond the common support of the covariate distributions. Adjustments based on matching, however, can only balance observed covariates; unobserved covariates must be handled in a separate component of the analysis. Matching is a widely applied method for estimating causal effects of non-randomized interventions; for a comprehensive review, see Stuart (2010). In a matched design the study groups are aligned on their observed covariate distributions; under the assumption of strongly ignorable treatment assignment, the difference in outcomes between the groups in the matched sample is an unbiased estimate of the intervention effect (see Rosenbaum and Rubin 1983, 1985). In our study baseline covariates measured before the implementation of parity are included in the adjustment so that parity and comparison groups are similar on them; MH/SA service utilization outcomes are then measured in the year after baseline. Due to the severity of illnesses considered, it is important for the subjects in our study to use MH/SA services in the follow-up year; by expanding coverage, it is expected that parity will allow severely ill subjects to continually receive the services they need.

While matching can adjust for the observed covariates, what can be said about unobserved biases? An observational study always remains vulnerable to the threat of unobserved biases. Available methods allow the investigator to quantify the sensitivity of conclusions to unobserved biases in an observational study, and we briefly mention some here for reference. Rosenbaum (2002) proposes a method to place bounds on the p -values for permutation tests in a matched analysis; these bounds assess the impact of an unobserved binary covariate on the statistical significance of results from a study. Brumback et al. (2004) developed a sensitivity analysis that relies on a mean specification of the unmeasured confounder on the potential outcomes, with the added advantage that it describes the sensitivity of the analysis in units of the outcome which may be more intuitively appealing. Other methods exist that assess the impact of an unmeasured covariate on the treatment assignment mechanism and outcomes, combining the ideas of the first two; see Imbens (2003), Gastwirth, Krieger, and Rosenbaum (1998, 2000), or Copas and Li (1997) for details.

In these methods for sensitivity analysis, however, little subject-matter knowledge is incorporated. We propose in this study of FEHB parity to incorporate what we know about our observational data into a sensitivity analysis and, in doing so, to generalize an approach to make claims about unobserved biases specific to the investigation at hand. In our study of parity in the FEHB plans, we would do well to identify possible sources of unobserved

biases. For example, our data from FEHB and Medstat health insurance claims suggest that a disproportionately higher number of federal employees seek MH/SA treatment in the baseline year than their industry (Medstat) counterparts. The true underlying mechanism as to why this occurs may never be known, but the reasonable investigator can enumerate possible explanations. Federal employees perhaps have better access to MH/SA treatment or more flexible work schedules to seek it. A reasonable question to ask during a matched analysis is whether an adjustment for an observed covariate, such as the proportion of employees and dependents, is sufficient to address related sources of unobserved bias. If the investigator's impressions about these related sources are strong, how could the investigator choose a second comparison group to provide evidence against them?

2 Design, analysis, and choices

2.1 A study with two comparison groups

The use of multiple comparison groups in observational studies has been discussed for some time, although not widely implemented (Rosenbaum 1987). We used administrative data from the FEHB plans and from the group of private health plans for the years 1999–2001, and focus on those beneficiaries who had been continuously enrolled over this period. Figure 1 illustrates the longitudinal structure of the data and how the investigator can designate a second comparison group from it. Information for the private health plans arise from the Medstat MarketScan database that includes health insurance claims for large self-funded employers. These employers, which offered relatively generous coverage, were not required to implement any state parity laws because the Employee Retirement Income Security Act exempts self-insured plans from state benefit regulations. Subjects included in the Medstat data form the *first comparison* group. The US Office of Personnel Management required FEHB plans to offer MH/SA benefits in 2001, so we define the time period 1999–2000 as pre-parity and the year 2001 as post-parity. The *parity* group consists of FEHB enrollees followed from baseline in 2000 through follow-up in 2001. A *second comparison* group consists of FEHB enrollees followed from baseline in 1999 through follow-up in 2000 but prior to the implementation of parity. We refer to this group as the *second comparison* group, and like the Medstat comparison group, this comparison group did not receive parity benefits. This second comparison group is composed of enrollees from the same population as the parity group, so there should be no differences in confounders, including unobserved, between them.

While adjustments such as matching can control for observed covariates, the first comparison group of Medstat enrollees may be quite different from the parity enrollees on unobserved baseline covariates. For example, federal employees and industry employees may have different spending habits or stigma associated with mental health services that are not captured in the observed data. Industry employees may have less flexibility with regard to work schedules that would otherwise permit them to be treated for and to deal with severe illness. In fact we can infer this information from the observed covariate distributions; for the study groups from the FEHB and Medstat plans, there are a higher proportion of Federal employees than Medstat employees seeking treatment for MH/SA disorders (Figure 2). Matching takes care of this observed difference, but there very well may be unobserved factors that influence the seeking of treatment. As such, an observed post-parity difference in utilization of MH/SA services between the parity and first comparison groups may be due to an unobserved bias, due to the intervention, or a combination of both. This type of ambiguity is a central concern in any observational study.

Using the second comparison group could partly address this ambiguity. The second comparison group is composed of enrollees from the same population as the parity group, and thus this group should not differ from the parity group on their observed or unobserved

characteristics. The major difference between the parity and second comparison groups is that the second comparison group is followed before the implementation of parity in 2001. A contrast between these two groups could show an effect of parity, without the same concerns about unobserved selection biases that the contrast with the Medstat comparison group would have. A disadvantage with the second comparison group, however, is that its enrollees are not followed at the same time as the parity group, so any temporal changes such as changes in treatment for severe illness or accessibility of care may be confounded in its comparison with the parity group. This disadvantage, though possibly minor, is partially addressed by the first comparison group, whose enrollees from Medstat are followed at the same time as the parity group. This intuition is developed further in §2.2.

Two contrasts have been described: one between the parity and first comparison groups, and the other between the parity and second comparison groups. The first contrast could suffer from selection bias but not a temporal trend; the second contrast does not suffer selection bias but may suffer a temporal trend. Taken together, these two contrasts build a more convincing body of evidence than an analysis with a single contrast could provide. With very little cost, the researcher can easily address these concerns of hidden biases by careful selection of a second comparison group.

First, in order to construct the parity and second comparison groups, we randomly split the sample of FEHB enrollees into two equally sized groups, denoted *parity* and *second comparison*. We then select individuals from the parity, first comparison, and second comparison groups on the basis of having at least one claim in the baseline year for the severe diagnoses listed in Table 1. For the parity and first comparison groups the baseline year is 2000, and for the second comparison group it is 1999. Matched sets of triplets consisting of one parity enrollee, one first comparison enrollee, and one second comparison enrollee were created. We examined distributions of observed baseline covariates to ensure balance among the triplets (Table 2).

There are a few key points to be made here. The sampling procedure above takes advantage of the relatively large ratio between the sizes of total FEHB group and of the Medstat comparison group (approximately 5:1). Without this favorable ratio, splitting the FEHB group might incur costs with respect to finding good matched samples; for example, if the overall ratio were instead 2:1, then splitting the FEHB enrollees into the two groups would likely not permit good matches because the available pool would be severely limited. A secondary issue is that randomly splitting the FEHB group introduces variation in the analysis – the estimates and conclusions from one iteration of the sampling procedure may differ from those from another. We do not address this issue here; for a theoretical discussion of split sample designs in observational studies see Heller, Rosenbaum, and Small (2009). Our primary goal is to address the first-order concerns about bias that do not diminish with increasing sample size.

2.2 Appealing to logic about trends in bias

In a standard observational study, for example, comparing the FEHB parity against the Medstat comparison enrollees, a significant estimated effect could very well be due to unobserved bias. In fact a comparison of these groups suggests that the odds of using any MH/SA services in the follow-up year was significantly greater in the parity versus the first Medstat comparison group. However, we might reasonably believe that different unobserved mechanisms are producing this observed effect. As a check against this concern, we compare the parity enrollees against enrollees from the second comparison group, who were also from FEHB Program plans but followed before parity's implementation. Using this contrast could be also viewed as incorporating a pretest-posttest design in the analysis procedure; see Laird (1983). Likewise, the second comparison suggests that the odds of

using any MH/SA services in the follow-up year was greater in the parity group versus the second comparison group, both comprised of FEHB enrollees. As a final check, the two comparison groups who did not receive parity benefits could be contrasted to check the similarity of their utilization outcomes. For example, if we should find an insignificant difference between the two control groups, then this suggests that we can be less concerned about the unobserved bias due to selection.

The trend in outcomes could also be a combination of unobserved selection bias and a temporal shift in utilization of MH/SA services. However, the temporal shift would have needed to occur within a relatively short period and cover up the effect of the unobserved selection. While not implausible, the effect due to this combination is much less likely than either source of bias alone. This rationale, along with other arguments for and against observed patterns of outcomes, can be used by the investigator to form a body of evidence suggesting that the intervention effect is real and not due to ambiguities such as unobserved biases.

Informally, we find a significant effect between the parity and first comparison groups, a significant effect between the parity and the second comparison groups, and finally no significant difference between the two comparison groups. Thus we can be less concerned about the specific unobserved biases discussed in §2.1 than in an analysis with the first comparison group alone; while we may have been previously concerned about access to MH/SA care, we are less concerned about it after analyzing the second comparison group. However, the basic analysis described is not suitable for several reasons. First, finding a significant difference between the parity and comparison groups does not mean that the difference is meaningful for policy. Second, failure to find a significant difference between the two comparison groups does not mean that they are the same; an equivalence test is needed. Finally, without any organizing structure the probability of rejecting a true null hypothesis is well above the nominal level due to multiple testing.

An intuitive way to resolve this last issue is to test the hypotheses in a logical order of priority (Rosenbaum 2008). We order the above contrasts into three steps. In Step 1, we test for a meaningful parity effect between the parity and first comparison groups. If one is found, then we proceed to Step 2, in which a contrast is made between the parity and second comparison groups. If a significant effect is found between them, then in Step 3, we characterize the similarity between the two comparison groups by an equivalence interval. The basic procedure just described comes without cost to the researcher, in the sense that the power and type I error rate of the study are essentially unchanged by the introduction of the second comparison group. We can say something important about unobserved confounding, based on logical choices, without altering the conclusions of the standard observational study. The procedure is developed formally in §3.3.

3 Inference

3.1 Matching in the study

There are $n = 356$ Medstat enrollees who were treated for a severe illness in their baseline year; this first comparison group is the smallest group, so we match along its covariate distribution. While not conventionally estimating the effect of treatment on the treated, we aim to answer to question, “What would be parity’s effect on those who did *not* receive its benefits?” (see Stuart 2010 for discussion of matching estimands). In our study we match sample from the parity and second comparison groups to form 356 triplets that consist of a parity, a first comparison, and a second comparison enrollee. The covariates available to us are demographic in type, which include age, employee/dependent status, sex, and clinical in type, which include total MH/SA spending, and indicator variables for use of inpatient

service and for use of any services to treat other co-occurring MH/SA diagnoses, such as acute adjustment disorder, based on insurance claims filed in the baseline year. A summary of covariate distributions is given in Table 2, and balance plots in Figure 2; in general, the distributions in the matched sample look well balanced.

The type of matching includes recent technologies that allow us to exert finer control over the observed clinical covariates, in the following sense. Notably, our study population consists of enrollees who were treated for a severe mental health disorder in the baseline year, based on claims that were filed in that year. These are the enrollees thought to most likely to benefit from parity, as they should be in need of treatment even in the absence of parity benefits. An important feature is that use of health services for these severe illnesses likely is confounded with the diagnosis. For example, bipolar patients should always receive medication under their treatment plan, whereas patients diagnosed with major depressive disorder may require intensive cognitive therapy as part of their regimen. Patients with psychotic episodes most likely require inpatient stays, which can be prohibitively expensive without insurance coverage. The health services outcomes, such as their utilization and composition, can differ substantially across these clinical groups so it is critical in the mind of the investigator that the rates of these diagnoses in our study sample be as similar as possible. We are, in fact, able to achieve zero sample bias on the prevalence of these diagnoses in the insurance claims through a technique called *fine balance* matching, so that the proportion of enrollees who filed respective claims were exactly the same across these diagnoses.

Matching with fine balance is described in Rosenbaum, Ross, and Silber (2007); for a textbook discussion of matching techniques, including fine balance, see Rosenbaum (2010). Briefly, matching with fine balance achieves zero sample bias on the important categorical covariates, such as the clinical diagnoses described above, as well as their interactions, so the joint distributions of these covariates are exactly the same between the matched groups, without needing to stratify on them. Stratification on a covariate generally poses a strong constraint on the matching algorithm, in the sense that the resulting matched distances under stratification are not necessarily the minimum. Fine balance takes as its argument an augmented distance matrix which selects exactly the number of controls needed to acquire a matched sample whose joint distribution of the important categorical covariates are the same. An example of this matrix appears in Table 3; in it, the discrepancies d_{ij} characterize the differences between treated subject i and control j , with larger values indicating poor matches. For example, d_{ij} might be the absolute difference in propensity scores between subjects i and j ; in our application, we use the Mahalanobis distance with propensity score calipers based on the covariates listed in Table 2 and Figure 2. With respect to the categorical covariate to be finely balanced, the matching algorithm uses the sinks to pull away one control subject at covariate level A and one at level B so that the matched control sample contains one subject at level A and two at level B . Subject to this constraint, the algorithm minimizes the objective function $\sum_{(i,j)} d_{ij}$ over the possible treated-control pairings (i, j) so that stochastic balance on other covariates is achieved. For example in Table 3, matching with fine balance might return the matched control sample $\{\gamma_1^A, \gamma_3^B, \gamma_4^B\}$.

In our parity example, the key diagnosis categories we control for with fine balance are bipolar affective disorder and specified psychosis diagnoses, and their interaction; the proportion of enrollees with these illnesses are exactly the same in the matched samples. We performed two matchings in order to construct the matched parity and second comparison groups, so there were two Mahalanobis distance matrices with propensity score calipers. The two propensity score models included those covariates listed in Table 2, and calipers were imposed between pairs whose estimated propensity scores deviated by more than 0.2 times the standard deviation of the estimated scores. Balance plots in Figure 2 and a summary of

the covariate distributions in Table 2 indicate that the matched samples are similar on their covariate distributions; see Imai, King, and Stuart (2008) and Hansen and Bowers (2008) for details about assessing covariate balance in non-randomized studies.

3.2 Model for binary outcomes in matched pairs

The primary outcomes relate to utilization and spending for MH/SA services in the follow-up year of our study, after controlling for observed baseline characteristics. We denote utilization by a binary indicator $U_{ij} = 1$ if a service was used and $U_{ij} = 0$ otherwise, for the j th member $j \in \{1, 2\}$ in matched pair $i \in \{1, \dots, n\}$. For matched-pair binary data a logistic-normal model treats the baseline log odds in each pair as a mean-zero random effect:

$$\text{logit}\{P(U_{ij}=1|Z_{ij})\}=\alpha_i+\beta \cdot Z_{ij}, \quad (1)$$

where $\alpha_i \sim N(0, \sigma^2)$ and Z_{ij} is a binary indicator of intervention status. Estimation by maximum likelihood is performed by first integrating over the random effect then maximizing the marginal likelihood; see Agresti (2002, Chapter 9) for a textbook discussion. In this setting $\exp(\beta)$ is the marginal odds ratio for the binary outcome U ; for example, if U represents use of any MH/SA services in the follow-up year then $\exp(\beta)$ is the population odds ratio of use between groups $Z = 1$ and $Z = 0$.

There are three parameters of interest based on the contrasts between: (1) the parity and the first comparison groups, denoted by β_{pc} ; (2) the parity and the second comparison groups, denoted by β_{pd} ; and (3) the two comparison groups, denoted by β_{cd} . The quantities of interest are the odds ratio of using a MH/SA service in the follow-up year. We examined whether a patient used any MH/SA service, including inpatient stays, therapy visits, prescriptions, and medication management.

The procedure in §3.3 tests whether there is a substantial difference between the parity group and the first comparison group. If there is a difference, then it tests whether there is a substantial difference between the parity group and second comparison groups. If successful, it characterizes the size of the difference between the two comparison groups. In the process, the chance that the procedure in §3.3 makes at least one false statement is at most α .

3.3 Testing procedure

The testing procedure for multiple comparison groups that we implement in this study follows closely the procedure described in Yoon (2009); the theory of testing in order was developed by Rosenbaum (2008). As described in §3.2 there are three parameters of interest relating to the contrasts between the parity and each of the two comparison groups, to show the intervention effect, and between the two comparison groups, to characterize their equivalence; two hypotheses are required for this last step for equivalence, so there are a total of four null hypotheses. A meaningful effect of parity can be characterized by a number $\Delta \geq 0$ that is chosen by the researcher. For example, in choosing $\exp(\Delta) = 2$ the investigator aims to show that parity doubled the odds of using a MH/SA service in the follow-up year; if the investigator chooses $\exp(\Delta) = 1$, then any difference in the odds of service utilization poses a meaningful result. For all $\delta \in [0, \Delta]$ and $\kappa \in (0, \infty]$ four hypotheses are considered. They are $H_{\delta,1}: \beta_{pc} \leq \delta$, $H_{\delta,2}: \beta_{pd} \leq \delta$, $H_{\kappa,3}: \beta_{cd} \leq -\kappa$ and $H_{\kappa,4}: \beta_{ce} \geq \kappa$. The last two hypotheses $H_{\kappa,3}$ and $H_{\kappa,4}$ preclude one another in the sense that at most one can be true; the equivalence interval characterizing the similarity of the comparison groups will be calculated by testing these two hypotheses simultaneously; see Bauer and Kieser (1996), Berger and Hsu (1996).

The hypotheses are testing in a logical order of priority, which was loosely described in §2.2. We describe the procedure in full detail here. In Step 1, the parity group and the first comparison group are contrasted on their utilization outcomes by testing $H_{\delta,1}$ for all $\delta \in [0, \Delta]$, first starting with the smallest $\delta = 0$ and continuing with larger values of δ , until a statistically insignificant result is shown, or until $H_{\Delta,1}$ is ultimately rejected. The overall procedure stops, and no further inferences are made when an insignificant result is reached, that is, when $H_{\delta,1}$ fails to be rejected for some $\delta < \Delta$. If $H_{\Delta,1}$ is ultimately rejected, then the researcher concludes that the odds ratio of a using MH/SA service under parity is at least $\exp(\Delta)$ and continues to the next step. In Step 2, the procedure follows as before, forming a contrast between the parity and the second comparison groups by testing $H_{\delta,2}$ for all $\delta \in [0, \Delta]$ starting with the smallest δ 's, until the first insignificant result is reached or $H_{\Delta,2}$ is rejected. If all these tests lead to rejection, then the odds ratio of using MH/SA services is claimed to be at least $\exp(\Delta)$, so that an effect of this magnitude is found when comparing the parity versus each of the two comparison groups. In Step 3, the similarity in outcomes between the two comparison groups is characterized by an equivalence interval, which is formed by simultaneously testing $H_{\kappa,3}$ and $H_{\kappa,4}$ for all $\kappa > 0$, starting with the largest κ . Testing in Step 3 continues until one of these hypotheses fails to be rejected at some finite $\kappa^* > 0$, the value which characterizes the equivalence interval between the two comparison groups.

By Proposition 3 in Rosenbaum (2008), the chance that this procedure tests and rejects at least one true hypothesis is at most α . Hypotheses $H_{\delta,1}$ and $H_{\delta,2}$ are tested in logical order, at two levels. In the first level the effect of parity between the parity and first comparison groups is evaluated before looking at the second comparison group, that is, $H_{\delta,1}$ for all $\delta \in [0, \Delta]$ is tested before $H_{\delta,2}$. Once the researcher has found a significant and meaningful effect of parity in the first level, the researcher can then evaluate the effect of parity with the second comparison group, which does not suffer from selection bias; note that this second level is not reached unless the first level finds a significant effect. If an effect of parity is found in the two contrasts, then hypotheses $H_{\kappa,3}$ and $H_{\kappa,4}$ are tested, to make a statement about the equivalence of the two comparison groups.

Because the contrast of the parity group and the first comparison group is completed in the first step before possibly continuing to the second, any hypothesis $H_{\delta,1}$ with $\delta \in [0, \Delta]$ that would have been rejected had there been only one comparison group is also rejected by the procedure. The presence of the second comparison group does not inhibit the rejection of any hypothesis $H_{\delta,1}$ with $\delta \in [0, \Delta]$. More precisely, the power to reject these hypotheses is not reduced by the presence of the second control group, nor by the need to control the probability of falsely rejecting several additional hypotheses that involve the second comparison group.

4 Illustration: the effect of mental health and substance abuse parity on the severely ill

We focus on a smaller population of enrollees who received care for at least one severe MH/SA diagnosis in their baseline year (Table 1). Because we chose eligible enrollees on the basis of diagnoses received in their baseline year, all subjects have positive, non-zero MH/SA spending at baseline. An important outcome in our analysis is whether the rates of any MH/SA service use in the follow-up year differed significantly between the parity group and groups that did not receive parity benefits. It is hypothesized that provision of MH/SA benefits under parity may result in increased use of any MH/SA services. Additionally, the composition of such services may likely change in response to parity under managed behavioral health care (Huskamp 1999; Huskamp et al. 1998).

Table 4 shows the results of our analysis based on the procedure in §3.3 on five binary outcomes. Each outcome indicates whether a particular type of MH/SA service was used in the follow-up year. The results indicate that, with simultaneous 95% confidence, the odds of using any MH/SA service in the parity group are 1.3 times greater than the odds in each of the two comparison groups, and the two comparison groups have odds that differ in the interval [0.49, 2.0]. The equality of the odds ratio of 1.3 for using any MH/SA services in the follow-up year is a feature of our testing procedure; we chose Δ such that an odds ratio greater than 1.3 characterized a meaningful parity effect. The equivalence interval between the two comparison groups is disappointingly large and raises concerns about residual hidden biases, although the inferences in Steps 1 and 2 indicate to some degree, that the two comparison groups are suitable for the observational study. In contrast to overall MH/SA service usage, the odds of using inpatient MH/SA services in the parity group are found to be less than the odds in each of the two comparison groups by a factor smaller than 0.67, while the two comparison groups look more equivalent on inpatient use. The decrease in inpatient use in the parity group is consistent with the hypothesis that managed care organizations might have sought more outpatient services in treating their severely ill enrollees. On this note, the analysis suggests that the odds of accessing a therapist in the follow-up year were higher for the parity group, though by small degree. Additionally, because the parity and second comparison groups show no difference in the rates of accessing therapy, the previous inference may be subject to hidden confounding on this outcome.

Prescription usage under parity appears to have increased. The parity group in the follow-up year had odds of using any prescription that were at least 1.4 times larger than the odds for each of the two comparison groups. The degree of equivalence for the two comparison groups lies in the interval [0.71, 1.4]. In connection with prescription use, the odds of using any medication management in the parity groups were lower than those of the first comparison group. Again, hidden biases may have produced this result. The parity and second comparison groups show no difference in whether any medication management visits were made in the follow-up year.

5 Discussion

Comparative effectiveness research that relies on observational data is always vulnerable to residual confounding – the baseline covariates may have missed an important hidden covariate that renders the study groups incomparable. The use of matching strategies to ensure comparability of subjects on *observed* confounders will be an increasingly important strategy. Very often the observed confounders are categorical, with several such variables playing key roles. We capitalized on a matching strategy that balanced on the important clinical diagnoses that were initially imbalanced between the parity and comparison groups, and thus were likely to have confounded the estimate of parity's effect. An important feature of the matching technology implemented here is that matching with fine balance does not require stratification on the clinical diagnoses, so that the optimal matching algorithm is not constrained by it. While the choice to implement fine balance is left to the investigator, it is nonetheless a valuable option should the objective of zero sample bias on the clinical diagnoses be motivated by the policy context.

The use of multiple comparison groups incorporates important subject-matter information, at the disposal of the researcher, to build a body of evidence for a causal claim by reducing ambiguities due to specific sources of hidden bias. This advantage comes at little cost, so long as data are available for a suitable second comparison group. In particular, a quasi-experimental design with multiple comparison groups may be especially useful for

longitudinal data in which the intervention is implemented at a certain time point, but had not been available prior to this point.

In our illustration we found that mental health and substance use parity significantly increased the odds of using MH/SA services for severely ill enrollees in the FEHB plans, although the composition of the services appeared to have changed under parity. One limitation of the analysis in §4 is that no multiplicity adjustments were made for testing multiple outcomes; further work needs to be done in the area of joint tests, which can possibly capitalize on the obvious correlations between outcomes in order to increase the power to detect true differences while maintaining the proper type I error rates.

Acknowledgments

Funded by Grants R01-MH054693 (Normand and Yoon), R01-MH080797 (Huskamp, Busch, and Normand), and K01-MH071714 (Busch) from the National Institute of Mental Health, Bethesda, MD USA.

We are grateful to Hocine Azeni from Harvard Medical School for valuable programming assistance; and to Vanessa Azzone, Colleen Barry, and Howard Goldman for meaningful feedback and discussions.

References

- Abadie A. Semiparametric difference-in-differences estimators. *Review of Economic Studies*. 2005; 71:1–19.
- Agresti, A. *Categorical Data Analysis*. 2. John Wiley & Sons, Inc; Hoboken, NJ: 2002.
- Barry CL, Gabel JR, Frank RG, Whitmore HH, et al. Design of mental health insurance coverage: Still unequal after all these years. *Health Affairs*. 2003; 22:127–37. [PubMed: 14515888]
- Bauer P, Kieser M. A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*. 1996; 83:934–7.
- Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*. 1996; 11:283–319.
- Brumback BA, Hernan MA, Haneuse SJPA, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*. 2004; 23:749–67. [PubMed: 14981673]
- Busch AB, Huskamp HA, Normand SL-T, Young AS, et al. The impact of parity on major depression treatment quality in the Federal Employees Health Benefits Program after parity implementation. *Medical Care*. 2006; 44:506–12. [PubMed: 16707998]
- Copas JB, Li HG. Inference for non-random samples” (with discussion). *Journal of the Royal Statistical Society, Series B*. 1997; 59:55–95.
- Frank RG, Goldman HH, McGuire TG. Will parity coverage result in better mental health care? *New England Journal of Medicine*. 2001; 345:1701–4. [PubMed: 11759651]
- Gastwirth JL, Krieger AB, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*. 1998; 85:907–20.
- Gastwirth JL, Krieger AB, Rosenbaum PR. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society, Series B*. 2000; 62:545–55.
- Goldman HH, Frank RG, Burnam MA, Huskamp HA, et al. Behavioral health insurance parity for federal employees. *New England Journal of Medicine*. 2006; 354:1378–86. [PubMed: 16571881]
- Hansen BB, Bowers J. Covariate balance in simple, stratified, and clustered comparative studies. *Statistical Science*. 2008; 23:219–236.
- Heller R, Rosenbaum PR, Small DS. Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*. 2009; 104:1090–101.
- Huskamp HA. Episodes of mental health and substance abuse treatment under a managed behavioral health care carve-out. *Inquiry*. 1999; 36:147–61. [PubMed: 10459370]
- Huskamp HA. How a managed behavioral health care carve-out and benefit expansion affected spending on treatment episodes. *Psychiatric Services*. 1998; 49:1559–62. [PubMed: 9856616]

- Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*. 2008; 171:481–502.
- Imbens GW. Sensitivity to exogeneity to assumptions in program evaluation. *American Economic Review*. 2003; 93:126–32.
- Laird NM. Further comparative analyses of pretest-posttest research designs. *The American Statistician*. 1983; 37:329–30.
- Lehmann, EL. *Testing Statistical Hypotheses*. John Wiley; New York: 1959.
- Rosenbaum PR. The role of a second control group in an observational study. *Statistical Science*. 1987; 2:292–316.
- Rosenbaum, PR. *Observational Studies*. 2. Springer; New York: 2002.
- Rosenbaum PR. Testing hypotheses in order. *Biometrika*. 2008; 95:248–52.
- Rosenbaum, PR. *Design of Observational Studies*. Springer; New York: 2010.
- Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*. 2007; 102:75–83.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*. 1985; 39:33–8.
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company; Boston: 2002.
- Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science*. 2010; 25:1–21. [PubMed: 20871802]
- Yoon, FB. New methods for the design and analysis of observational studies. *Dissertations*. 2009. available from ProQuest, Paper AAI3381886, <http://repository.upenn.edu/dissertations/AAI3381886>

Fig. 1. Longitudinal structure of the data allows for a natural second comparison group of FEHB enrollees observed pre-parity.

Fig. 2. Covariate balance before and after matching for the two comparison groups. The dashed vertical lines display bounds such that observations within these bounds are considered small.

Table 1

Eligibility criteria: baseline diagnoses by ICD9-CM codes for the severely ill sub-population.

Schizophrenic disorders	
295.0	Simple type
295.1	Disorganized type
295.2	Catatonic type
295.3	Paranoid type
295.4	Acute schizophrenic episode
295.5	Latent schizophrenia
295.6	Residual schizophrenia
295.7	Schizo-affective type
295.8	Other specified types of schizophrenia
295.9	Unspecified schizophrenia
Major depression and unspecified affective psychoses	
296.24	Major depressive disorder psychosis, single episode
296.3	Major depressive disorder recurrent episode
296.34	Major depressive disorder psychosis, recurrent episode
296.90	Unspecified affective psychosis
296.99	Other specified affective psychoses
Bipolar affective disorders	
296.0	Manic disorder, single episode
296.1	Manic disorder, recurrent episode
296.4	Bipolar affective disorder, manic
296.5	Bipolar affective disorder, depressed
296.6	Bipolar affective disorder, mixed
296.7	Bipolar affective disorder, unspecified
296.80	Manic-depressive psychosis, unspecified
296.81	Atypical manic disorder
296.82	Atypical depressive disorder
296.89	Other
301.1	Affective personality disorder
301.11	Chronic hypomanic personality disorder
301.13	Cyclothymic disorder
Paranoid states	
297.0	Paranoid state, simple
297.1	Paranoia
297.2	Paraphrenia
297.3	Shared paranoid disorder
297.8	Other specified paranoid states
297.9	Unspecified paranoid state
Other nonorganic psychoses	
298.0	Depressive type psychosis

- 298.1 Excitative type psychosis
 - 298.3 Acute paranoid reaction
 - 298.4 Psychogenic paranoid psychosis
 - 298.8 Other and unspecified reactive psychosis
 - 298.9 Unspecified psychosis
-

Table 2

Distribution of baseline covariates in the matched sample. $n = 356$ matched triplets formed. A binary covariate for a specific diagnosis indicates whether a MH/SA claim was filed in the baseline year for that diagnosis. The important diagnostic categories for severe illness (bipolar disorder, affective psychosis) were finely balanced; other co-occurring MH/SA diagnoses were balanced by matching on the Mahalanobis distance with propensity score calipers.

		Parity (FEHB)	Comparison 1 (Medstat)	Comparison 2 (Pre-Parity)
Log(total MH spending), \$	mean (sd)	7.5 (1.1)	7.5 (1.2)	7.5 (1.1)
Age, years	mean (sd)	47 (9)	48 (10)	47 (9)
Employee	%	46	45	46
Female		72	72	71
Baseline inpatient use		8	10	10
Bipolar affective disorder with psychosis		3	3	3
Bipolar affective disorder, no psychosis		28	28	28
Affective psychosis, no bipolar		11	11	11
No bipolar nor psychosis		58	58	58
Major depressive disorder		72	71	72
Acute adjustment disorder		4	4	4
Substance use disorder		2	2	2
Anxiety		10	9	9
Other mental health		6	6	6

Table 3

Augmented distance matrix for optimal matching with fine balance

Available controls										
Treated	γ_1^A	γ_2^A	γ_3^B	γ_4^B	γ_5^B	Rowsum constraint				
τ_1^A	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	1				
τ_2^B	d_{21}	d_{22}	d_{23}	d_{24}	d_{25}	1				
τ_3^B	d_{31}	d_{32}	d_{33}	d_{34}	d_{35}	1				
$Sink_1$	0	0	∞	∞	∞	1				
$Sink_2$	∞	∞	0	0	0	1				

Table 4

Oddsratios (OR) for service use. Steps 1 and 2 ORs are the lower bounds of the one-sided confidence interval for the OR of using the type of MH/SA service. In Step 3 the 95% equivalence interval characterizes the similarity of the two comparison groups on the basis of using the type of MH/SA service.

Outcome	Step 1	Step 2	Step 3
	Parity vs first comparison	Parity vs second comparison	Equivalence of two comparison groups
Any MH services	> 1.3	> 1.3	[0.49, 2.0]
Inpatient MH services	< 0.67	< 0.67	[0.77, 1.3]
Therapy	> 1.0	inconclusive, > 1	
Medication management	< 0.8	inconclusive, < 1	
Any prescription	> 1.4	> 1.4	[0.71, 1.4]