Using split samples and evidence factors in an observational study of neonatal outcomes

Kai Zhang, Dylan Small, Scott Lorch, Sindhu Srinivas, Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. During a few years around the turn of the millennium, a series of local hospitals in Philadelphia closed their obstetrics units, with the consequence that many mothers-to-be arrived unexpectedly at the city's large, regional teaching hospitals whose obstetrics units remained open. Nothing comparable happened in other US cities, where there were only sporadic changes in the availability of obstetrics units. What effect did these closures have on mothers and their newborns? We study this question by comparing Philadelphia before and after the closures to a control Philadelphia constructed from elsewhere in Pennsylvania, California and Missouri, matching mothers for 59 observed covariates including year of birth. The analysis focuses on the period 1995-1996, when there were no closures, and the period 1997-1999 when five hospitals abruptly closed their obstetrics units. Using a new sensitivity analysis for difference-in-differences with binary outcomes, we examine the possibility that Philadelphia mothers differed from control mothers in terms of some covariate not measured, and perhaps the distribution of that unobserved covariate changed in a different way in Philadelphia and control-Philadelphia in the years before and after the closures. We illustrate two recently proposed techniques for the design and analysis of observational studies, namely split samples and evidence factors. To boost insensitivity to unmeasured bias, we drew a small random planning sample of about 26,000 mothers in 13,000 pairs and used them to frame hypotheses that promised to be less

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. This work was supported by grants SES-0849370 and SES-0961971 from the U.S. National Science Foundation. Kai Zhang is a PhD student, Dylan Small is associate professor, and Paul R. Rosenbaum is professor in the department of statistics at the Wharton School of the University of Pennsylvania. Scott Lorch is assistant professor of pediatrics and Sindhu Srinivas is assistant professor of obstetrics and gynecology at the University of Pennsylvania School of Medicine. 5 April 2011

sensitive to bias; then these hypotheses were tested on the large, independent complementary analysis sample of nearly 240,000 mothers in 120,000 pairs. The splitting was successful twice over: (i) it successfully identified an interesting and moderately insensitive conclusion, (ii) by comparison of the planning and analysis samples, it is clearly seen to have avoided a exaggerated claim of insensitivity to unmeasured bias that might have occurred by focusing on the least sensitive of many findings. Also, we identified two approximate evidence factors and one test for unmeasured bias: (i) factor 1 compared Philadelphia to control before and after the closures, (ii) factor 2 focused on the years 1997-1999 of abrupt closures and compared zip codes with closures to zip codes without closures, (iii) and the test for bias focused on the years 1995-1996 prior to closures and compared zip codes which would have closures in 1997-1999 to zip codes without closures in 1997-1999 — any ostensible effect found in that last comparison is surely bias from the characteristics of Philadelphia zip codes in which closures took place. Approximate evidence factors provide nearly independent tests of a null hypothesis such that the evidence in each factor would be unaffected by certain biases that would invalidate the other factor.

Key words: Design sensitivity; difference-in-differences; evidence factor; observational study; optimal matching; sensitivity analysis; split samples; test for bias.

1 Introduction: background; methodological outline

1.1 A wave of closures of hospital obstetrics units

Beginning in 1997, a series of community hospitals in Philadelphia closed their obstetrics units, so mothers who would normally have delivered at these hospitals had to seek care at the city's large regional hospitals whose obstetrics units remained open. Between 1997 and 2007, 12 of 19 hospitals in the city closed their obstetrics units. Nothing similar happened at this time in other major cities, which experienced only sporadic changes in the availability of obstetrics units. For instance, in Pittsburgh, Los Angeles, San Diego and San Francisco less than 5% of the deliveries in 1995 and 1996 were in obstetric units that subsequently closed between 1997-2005. Babies born in these and other cities will serve as controls. By contrast, in Philadelphia, over 30% of the deliveries in 1995 and 1996 occurred at obstetrics units that subsequently closed between 1997 and 2005. It is not entirely surprising that a hospital facing competitive or financial pressures would consider closing its obstetrics and neonatal units: these fields have unusually high costs associated with malpractice litigation and malpractice insurance (Kirby et al. 2006). Why closures should have concentrated in Philadelphia is less clear. In its densely urban center, Philadelphia is home to several large hospitals associated with major medical schools, but beyond its urban center, Philadelphia sprawls at considerable distance into a variety of diverse neighborhoods served by smaller community hospitals; the closures occurred here.

Of 19 Philadelphia hospitals with obstetrics units in 1995, 12 closed their obstetrics units between 1997 and 2007; see Figure 1. In part based on a split sample analysis described below, the analysis presented here focuses on five hospitals that abruptly closed in 1997-1999, before the City of Philadelphia intervened in 2000 to organize and slow the pace of subsequent closures and to offer strategies to allow for the remaining hospitals to accommodate the increased obstetric volume. It is interesting to note that four of the five closures during 1997-1999 were geographically close, suggesting a cascade in which each successive closure increased the stress on near-by units that remained open, perhaps leading to their closure. Conceivably, the geography of Philadelphia's closures explain why there was a wave of closures in Philadelphia with no similar pattern in other cities.

What was the effect of the 1997-1999 hospital closures on the health of mothers and their newborn babies? Stories were told — perhaps some were even true — of women in labor being delivered by ambulance to a hospital that had closed its obstetrics unit the previous week. Other stories were told — more likely true — of women in labor, some of them poor, travelling longer distances, perhaps in rush hour, to reach an open obstetrics unit, of overcrowding and inadequate staffing at the units that remained open. A closure in one neighborhood may force a mother who lives in that neighborhood to travel a long distance to a hospital in another neighborhood, but it may also cause overcrowding in a hospital remote from the closure, and so it may affect mothers who live near the hospital that remained open. It is easy to imagine a long trip to an overcrowded obstetrics unit is not beneficial. Then again, many of the hospitals that remained open have excellent reputations, better perhaps than the reputations of the hospitals that closed their obstetrics units. Then again, teaching hospitals are home to the most and least experienced doctors, professors of medicine and medical residents, who usually work in tandem, but who found themselves short of staff. Then again, the human race has managed to reproduce in circumstances considerably more dire than traffic and overcrowding. It is hard to know what, if anything, to expect from the five closures in 1997-1999.

1.2 Matching to build a control Philadelphia

For each birth in Philadelphia in 1995-2003, we used multivariate techniques and an optimal assignment algorithm to match a control birth from elsewhere in Pennsylvania or California or Missouri, the three states for which we had the needed data. Because there were 132,786 births in Philadelphia and 5,998,111 potential control births elsewhere, the matching was on an unusually large scale. The matching was done year-by-year, so a Philadelphia birth in 1995 was matched to a control birth in 1995, and it controlled not only characteristics of the mother and baby, but also characteristics of the mother's neighborhood, such as typical income, the frequency of poverty, and the level of education in the neighborhood. During this time period, Philadelphia mothers were quite different from the unmatched potential control group: they came from neighborhoods with lower income, more poverty, and fewer high school graduates; however, the mothers themselves (as opposed to their neighborhoods) were more likely than potential controls to have graduated high school. Philadelphia mothers were somewhat younger with less prenatal care, but their babies were, on average, slightly smaller. All of these measured differences and many other measured differences were removed year by year using matching techniques; see §2. The control mothers and infants are not only similar as individuals: as a group, they have similar temporal and measured neighborhood characteristics to births in Philadelphia in 1995-2003. Here, neighborhood characteristics are measured at the zip-code level and are indicated in Table 1.

Why build a control Philadelphia? Because of the geography of Philadelphia, the closures might be expected to affect certain neighborhoods more than others, and each neighborhood has its own demographics, income, social and health problems. A control Philadelphia permits straightforward questions about how mothers and neighborhoods in Philadelphia changed in comparison with similar mothers and neighborhoods elsewhere.

Abadie et al. (2003, 2010) developed an innovative approach to using aggregate data to synthesize a control for a region that was subjected to an intervention. Their synthetic control is a weighted combination of actual regions that were not subjected to the intervention. For example, in their study of the economic impact of terrorism in the Basque Country, Abadie and Gardeazabal (2003) use a weighted combination of two Spanish regions to approximate the economic growth that the Basque Country would have experienced in the absence of terrorism. The weighted combination is chosen to match the region subjected to the intervention in its covariates and trajectory of outcomes prior to the intervention. Abadie, Diamond and Hainmueller (2010) developed an inferential approach when using synthetic controls that is akin to permutation inference. They use placebo tests to examine whether or not the estimated effect of the actual intervention is large relative to the distribution of the effects estimated for the regions not exposed to the intervention, where the synthetic control method is also used to estimate effects for regions not exposed to the intervention. A valuable feature of Abadie et al.'s synthetic control approach is that it only requires aggregate data on regions, which are often the only type of data available. For our study of the effect of the obstetric unit closures in Philadelphia, we are fortunate to have individual data on mothers and babies, which permit, for example, comparisons of parts of Philadelphia with its control.

1.3 Splitting

Philadelphia mothers and infants may have differed from controls in ways that were not measured and hence not controlled by matching for observed covariates. After adjustment for observed covariates, the key source of uncertainty in an observational study is the possibility that differences in outcomes between treated and control subjects are not effects of the treatment but rather biases from some unmeasured way in which treated and control subjects were not comparable. Our analysis is largely directed at this possibility.

A sensitivity analysis asks how failure to control some unmeasured covariate might alter the conclusions of a study. Many issues affect the sensitivity of conclusions to unmeasured biases (Rosenbaum 2004; 2010a, Part III; 2010b), but most of these issues are difficult to appraise in the absence of data. Heller et al. (2009) made a formal argument for splitting the sample at random into a small planning sample of perhaps 10% and a large analysis sample of perhaps 90%. The planning sample is used to design the study — to frame questions and guide the analytical plan — whereupon the planning sample is discarded; then, all conclusions are based on the untouched, unexamined, untainted analysis sample. If one were to perform several or many analyses of a single data set, noting that a particular conclusion was insensitive to unmeasured biases, then one would not know whether this judgement about sensitivity to bias was distorted by capitalizing on chance in picking the most favorable of these analyses. In contrast, the use of a split sample permits exploration of unlimited scope in a planning sample, and an independent, untainted, highly focused analysis of the analysis sample. Cox (1975) evaluated splitting to control for multiple testing in randomized experiments, but Heller et al. (2009) find that splitting is even more useful in sensitivity analyses in observational studies because the biases from unmeasured covariates do not diminish as the sample size increases. If one could make decisions that would make the study less sensitive to unmeasured biases by sacrificing a small portion of the sample, then that sacrifice might be well worth making. The formal argument in Heller et al. (2009) evaluates power and design sensitivity in split samples.

As Cox (1975) emphasized, splitting has an important advantage over most methods that address multiple testing, namely it permits human judgement to play an informed role between exploratory analysis of the planning and focused confirmatory analysis of the analysis sample. Formal or algorithmic procedures that address multiple testing, such as the Bonferroni inequality, do not leave a role for judgement; rather, their form must be prespecified. In the current study, this meant that an extensive analysis of the planning sample was discussed at a meeting of the clinicians and statisticians, and the analysis plan that emerged from that meeting reflected results from the planning sample combined with clinical and statistical judgement. For instance, before looking at any data, we thought that overcrowding in an obstetrics ward might result in an increase in Caesarean sections and birth injuries of various kinds, but the planning sample strongly suggested a focus on serious birth injuries (ICD-9 767-3), and not a focus on Caesarean sections. In part, our focus on serious birth injuries reflects what we saw in the planning sample, but in part it reflects a judgement about an effect that seems both plausible and clinically interesting. The planning split also revealed that several outcomes were simply too rare to study even with the much larger analysis sample; here, it is not the P-value but the event rate that provides information relevant to power computations for the as yet unexamined analysis sample. Although one can mechanize the evaluation of many P-values, one cannot mechanize an evaluation of many P-values that incorporates human judgement about what is plausible and interesting. Because human judgement cannot be mechanized, it is not typically possible to perform the same analysis on many repeated splits of the sample, as one might do in cross-validation.

Here, we took a small random sample of the matched pairs, 10% or 13,278 pairs in this study, and used it to plan the main analysis, which concerned the complementary 90% of pairs or 119,508 pairs. Among many outcomes examined using the planning split sample, we were led to focus on birth injuries, specifically ICD code 767.3, and on the years 1997-1999 when five hospitals abruptly closed their obstetrics units. Beginning in 2000, the City of Philadelphia intervened to slow down and organize closures. Before looking at the planning sample, it was not obvious to us whether the City's intervention had been more than a symbolic gesture, but the planning sample suggested that most of the action occurred in 1997-1999, that is, after the City's intervention there was no discernable effect of hospital closures. If this analytic focus had come about after examining many outcomes and various comparisons for those outcomes using the complete data, then there would naturally be reason for concern that the focus was distorted by capitalizing on chance events that only appear to be systematic patterns. However, this analytic focus came about by examining a random sample of 10% of the pairs, and 90% of the pairs remain to put this carefully chosen, very specific focus to a proper test. One might imagine two investigators, one who early on published a small, informal, exploratory, highly speculative and not particularly convincing study involving many comparisons, with the second investigator taking the one promising result from the first study and confirming it in a much larger independent sample. From an inferential point of view, it makes no difference whether there were two investigators or only one, that is, no difference between, on the one hand, replicating a promising but speculative finding by someone else and, on the other hand, generating both the speculative finding and the confirmation using split samples.

1.4 Evidence factors

If we are looking at a treatment effect, not a bias from unmeasured covariates, then we anticipate several patterns. First, when compared to similar births in other states, an effect of the closures should be absent in 1995-1996 and present in 1997-1999. For birth injuries, a binary outcome, this leads to a difference-in-difference analysis along the lines suggested by Gart (1969) for randomized cross-over studies; see §4 where discordant pairs become the counts in a 2×2 table that is subjected to a sensitivity analysis. Second, we identified thirteen zip codes in northern Philadelphia as close to the hospitals with closures (specifically, 19115, 19119, 19121, 19127, 19128, 19129, 19131, 19132, 19135, 19136, 19144, 19149, 19152). Of course, overcrowding occurred in the obstetrics units that remained open, and many of these were at some distance from the closures; nonetheless, it is reasonable to contrast zip codes with closures to zip codes without closures in 1997-1999, anticipating a larger effect on zip codes with closures. Finally, if the difference between the Philadelphia-versus control difference in the zip codes with closures and in zip codes without closures was already apparent in 1995-1996, before the closures, then that cannot plausibly be an effect of the closures; rather, it must indicate that our matching and difference-in-differences have failed to compare comparable mothers under different treatments. The first two comparisons are an example of evidence factors, that is, of (nearly) independent tests of the hypothesis of no treatment effect that are susceptible to different kinds of unmeasured biases (Rosenbaum 2010c), whereas the third comparison is a test for unmeasured bias (Rosenbaum 1984).

The method of difference-in-differences has a long history; see, for instance, Campbell (1957, 1969), Meyer (1985), Angrist and Krueger (2000), Shadish, Cook and Campbell (2002) and Athey and Imbens (2006). A conventional description of difference-in-differences follows, although Proposition 1 departs from this description by studying sensitivity to biases that can affect difference-in-difference studies. In a nonrandomized treatment-versus-control comparison the treatment effect is aliased with stable but unmeasured baseline differences between treated and control groups, whereas in a before-versus-after comparison, the treatment effect is aliased with trends over time. In contrast, in a difference-in-differences study, the treatment effect is aliased neither with stable unmeasured baseline differences between treated and control groups nor with trends over time that affect all groups in the same way, but it is aliased with the interaction of those two sources of bias. Proposition 1 examines sensitivity of inferences about effects to biases from such interactions. Although difference-in-differences is conventionally defined in terms of the passage of time, it is more generally relevant to situations in which a treatment effect is aliased with the interaction of two sources of bias, and this generality is exploited here in the second evidence factor, where time is replaced by Philadelphia zip codes near closures.

For a recent review of matching techniques, see Stuart (2010). For discussion of the importance of anticipated patterns in observational studies, see Campbell (1957), Trochim (1985), Shadish, Cook and Campbell (2002) and West et al. (2008). Various methods of sensitivity analysis in observational studies are discussed by Cornfield et al. (1959), Rosenbaum and Rubin (1983), Yanagawa (1984), Gastwirth et al. (1992, 1998), Rosenbaum (1995; 2002, §4), Marcus (1997), Lin et al. (1998), Robins et al. (1999), Copas and Eguchi (2001), Imbens (2003) and Deprete and Gangl (2004).

2 Matching

2.1 Philadelphia and elsewhere, before and after matching

We obtained birth certificates from all deliveries occurring in Pennsylvania, California and Missouri between 1/1/1995 and 6/30/2005. Each state's department of health linked these birth certificates to death certificates using name and date of birth, and then de-identified the records. We then linked over 98% of birth certificates to maternal and newborn hospital records. Over 80% of the remaining unlinked birth certificate records failed to identify a hospital, suggesting a birth at home or a birthing center. The unlinked records had similar gestational age and racial/ethnic distributions to the linked records. For the maternal and newborn hospital records, California, Missouri, and Pennsylvania routinely collect information on all hospital admissions within each state. Each patient record contains the UB-92 form submitted by each hospital to the state, with 15 to 25 fields for principal diagnoses and procedures occurring during the hospital stay. Birth certificates contain information on birth weight, gestational age, and patient-level demographic variables and obstetric risk factors. Sociodemographic information on the mother's zip code is obtained from the Bureau of the Census.

Each baby born in Philadelphia was matched with a baby born in other regions of Pennsylvania or California or Missouri. In each year, the match balanced 59 observed covariates. Of these, 34 covariates are listed in Table 1, which gives their means among potential controls outside Philadelphia, in Philadelphia, and in the matched controls. These covariates describe the socioeconomic status of mom's neighborhood, mom's own age, parity, prenatal care, education, race, and health insurance, and baby's birth weight and gestational age, two key measures of a newborn's health status. Because we are interested in the effects of the hospitals at the time of delivery, we adjust for quantities such as gestational age and birth weight that are essentially determined prior to admission to the hospital. These factors are associated with different risks of many neonatal outcomes (Stoll et al. 2010). A study of prenatal care, as opposed to care around the time of delivery, would not adjust for gestational age and birth weight, although in fact there is little compelling evidence that prenatal medical care has much effect on preterm delivery (American College of Gynecology 2003, Hollowell et al. 2011). Babies were also matched exactly for year of birth.

For each of the 34 covariates, Table 1 also gives the standardized absolute difference in means before and after matching, that is, Philadelphia-versus-potential controls and Philadelphia-versus-matched controls. The pooled standard deviation used in this measure is calculated as the square root of the equally weighted average of the sample variances inside and outside Philadelphia before matching, so matching changes the numerator, that is the difference in means, but it does not change the denominator, the pooled standard deviation. See Rosenbaum and Rubin (1985) for discussion of this conventional measure of covariate imbalance. In addition to the covariates in Table 1, there are 25 other covariates, $59 = 34 + 25$, which describe rare congenital anomalies or problems in the pregnancy that existed long before the start of labor.

Before matching, compared to potential controls, Philadelphia mothers were, on average, more likely to live in a low income neighborhood in which fewer people had college degrees, slightly younger with a little less prenatal care, more likely to have completed 8th grade, more often black, and gave birth to somewhat smaller babies.

Figure 2 displays all 59 absolute standardized differences in means in each of five years, 1995-1999. Before matching several covariates differed by more than 0.8 standard deviations. After matching, all $295 = 5 \times 59$ standardized differences in means after matching are less than 0.2 standard deviations. Before matching, the maximum and upper quartile of the 295 absolute standardized differences were 1.19 and 0.18, whereas after matching they were 0.19 and 0.06, respectively. For comparison, a Normal distribution has 95% of its probability on an interval that is approximately four standard deviations in length, so 0.19 and 0.06 of a standard deviation are approximately 5% and 2% of such an interval. In brief, Figure 2 shows that after matching, all of the 59 covariate means were in reasonable balance in every year; that is, Philadelphia and control-Philadelphia were similar in terms of these covariates year by year.

2.2 How the matching was done

There were 132,786 births in Philadelphia and 5,998,111 potential control births to choose from in building the matched comparison. In matching, a large sample size should be a luxury, but if inappropriate methods are used, it can appear to be a hindrance. A 132786×5998111 distance matrix would contain approximately 7.96×10^{11} numbers, and this is well beyond what can be handled with current combinatorial optimization techniques on current computers. There is a simple solution, however: match exactly for some important covariates, thereby reducing one large problem to a series of smaller problems; see Rosenbaum (2010a, §9.3).

We ordered the covariates by priority, year of birth being first because of the structure of the study, followed by gestational age in weeks $(0, 33]$, $(33, 36]$, $(36, 38]$, $(38, 40]$ and $(40, \infty)$, categories based on an estimated propensity score for the propensity to be born in Philadelphia, mother's age in years $(0, 18]$, $(18, 34]$, $(34, \infty)$, mother's education in four groups by degree. The algorithm first looked at the size of the distance matrix within a given year; if that was too large, it looked at the size of the distance matrix within a given year and gestational age; if that was too large, it looked within a given year, gestational age and propensity score group, and so on. Once the size of the distance matrix was manageable, the distance matrix was computed using a rank-based Mahalanobis distance within calipers for an estimated propensity score (Rosenbaum and Rubin 1985; Rosenbaum 2010a, §8), and an optimal match was determined to minimize the total distance within matched pairs (Rosenbaum 1989; 2010a, §8). Calipers on the propensity score ensure a close match on a unidimensional summary sufficient to remove bias from imbalances in observed covariates; see Rosenbaum and Rubin (1985) and Abadie and Imbens (2011) for discussion of calipers and unidimensionality in matching. The computations used Hansen's (2007) `optmatch` package in `R`; see also Hansen and Klopfer (2006).

3 Splitting

3.1 A 10%-90% random split for design and analysis

As proposed by Heller et al. (2009), within each year, the planning sample was a 10% sample of pairs drawn at random without replacement. The analysis sample was the complementary 90% of pairs. As noted in §1.2, the base period, 1995-1996 had no closures of obstetrics units, 1997-1999 had five abrupt closures, whereas beginning in 2000 the City of Philadelphia intervened to prevent abrupt closures so that closures followed some delay and reorganization among open hospitals. The planning sample looked at 38 outcomes in each of two time periods defined by the City's intervention in the process of closure, 1997-1999 and 2000-2003, for all zip codes, for zip codes close to closures and for zip codes remote from closures, so a total of $38 \times 2 \times 3 = 228$ significance levels were computed. Consistent with the discussion by Cox (1975) and Heller et al. (2009), sample splitting served as a substitute for a correction for multiple testing.

The planning sample suggested several interesting hypotheses, and here we focus on one of these, namely birth injury ICD-9 767.3. Unlike some of the other 767 codes, code 767.3 is a serious injury, such as fracture of long bones or the skull, not a routine abrasion of a normal birth. The planning sample suggested an increase in such birth injuries in Philadelphia in 1997-1999 with a return to normal in 2000-2003, with some indication that the increase was more pronounced for mothers who lived in zip codes affected by closures.

The planning sample is used informally to suggest interesting hypotheses and appropriate analyses. To motivate and clarify the theoretical discussion in §4, we present an analysis of birth injury for the 10% planning sample in the same form that will be used in the final analysis of the complementary 90% sample. Actually, we did quite a bit of analysis of the planning sample before settling upon this form. Having selected this form, the analysis of the complementary 90% sample simply used this one form on this outcome. The analysis of the 90% sample incorporates a sensitivity analysis developed in §4.

3.2 Birth injury in the planning sample: the largest difference, two nearly independent tests for effect and a test for unmeasured bias

Table 2 is the analysis of birth injury for the 10% planning sample. It has four panels labeled "a comparison focused on the most affected groups," "factor 1," "factor 2," "bias test." Factor 1 is the simplest comparison, so it is described first; then the other parallel comparisons are described briefly. Table 2 counts Philadelphia-control pairs discordant for birth injury, that is, pairs in which exactly one baby experienced a birth injury. Factor 1 compares Philadelphia to control in 1997-1999 versus 1995-1996. In 1995-1996, there were 85 pairs containing one birth injury, and in 43 pairs it was the Philadelphia baby who was injured and in 42 pairs it was the control baby who was injured. In contrast, during the period of closures, 1997-1999, there were 184 pairs with birth injuries, and in 141 of the 184 pairs it was the Philadelphia baby who experienced the injury. The odds ratio in this 2×2 table is 3.19, so it looks as if there was an increase in the risk of birth injury in Philadelphia during the period of hospital closures. Because of this observation in the planning sample, the analysis in the complementary 90% sample will look for an increase in risk for this same outcome. Our data do not locate the birth injury as occurring either in the hospital or prior to reaching the hospital, say in an ambulance. The most affected group contrasts Philadelphia zip codes near closures to matched controls in 1995-1996 and in 1997-1999; both a priori and as indicated in this planning split sample, it seems reasonable to think that if a strong effect is to be found, it will be found here.

Gart (1969) proposed an analysis for a randomized, two-period cross-over experiment with a binary outcome which we generalize for use here. His analysis is suggested by a logit model with additive pair and time effects plus a treatment effect. In such a model, the nuisance parameters are eliminated by conditioning on sufficient statistics, so that the treatment effect is tested by comparing two sets of discordant matched pairs to the hypergeometric distribution in a 2×2 table analysis. In Table 2, we perform this analysis several times, and in §4 we examine the analysis in the context of a non-randomized observational study and generalize it to permit a sensitivity analysis. Happily, after a few steps, the sensitivity analysis for binary difference-in-differences turns out to be an almost standard sensitivity analysis for a 2×2 table, so the situation in observational studies develops in parallel with Gart's (1969) analysis for a randomized cross-over study. There is, however, a curious transformation of the magnitude of the sensitivity parameter; see Proposition 1.

Judged by Gart's test, the increase in risk of birth injury in "factor 1" in the planning sample is significantly different from an odds ratio of 1, with one-sided significance level 0.000023 and one-sided 95% confidence interval $[1.95, \infty)$. In the planning sample alone, if one did a Bonferroni correction for 228 two-sided tests, the significance level would be approximately 0.01.

In Table 2, factor 2 looks just at the years of closures, 1997-1999, and contrasts zip codes near closures in 1997-1999 to zip codes remote from closures. As mentioned in §1.2, the overcrowding did not occur at the closed obstetrics units but at the ones that remained open, so mothers in zip codes remote from closures may have been affected by sharing an overcrowded obstetrics unit with mothers who came from zip codes with closures. On the other hand, mothers in zip codes with closures faced a newly lengthened trip to the obstetrics unit and may have been unexpected there. In any event, factor 2 is another difference-in-difference analysis in the manner of Gart (1969) but now contrasting Philadelphia-control pairs for zip codes near closures to pairs for zip codes remote from closures. The odds ratio is 1.51, consistent with increased risk, but it does not differ significantly from 1 in this 10% planning sample. The panel labeled “bias test" in Table 2 is the same comparison but done in the years before closures: any systematic difference here could not be an effect of the closures and must reflect some uncontrolled bias. The odds ratio is 0.65 and is not significantly different from 1 in this 10% planning sample.

The analysis for the most affected group in Table 2 looks just at zip codes near closures, comparing 1997-1999 to 1995-1996. It is in this comparison that we might anticipate the largest effect. The odds ratio is 5.8 with a one-sided 95% confidence interval of $[2.03, \infty)$. Because this is one of the largest of hundreds of estimated odds ratios in the 10% planning sample, we have reason to suspect that it is biased upwards; nonetheless, this seems like a promising comparison to make in the independent 90% analysis sample which will be examined in §5.

There is an important difference between, on the one hand, factors 1 and 2 and, on the other hand, the analysis of the most affected groups. Factors 1 and 2 are not redundant; indeed, they are nearly independent tests when the hypothesis of no treatment effect is true, that is, they are approximate evidence factors. If the null hypothesis of no effect were true, then exact evidence factors would be statistically independent (Rosenbaum 2010c) and, strictly speaking, factors 1 and 2 in Table 2 do not qualify; however, they are nearly independent and so are approximate evidence factors (Rosenbaum 2011, Lemma 4 and §7). Moreover, the unmeasured biases that affect these two comparisons are different — in factor 1, unmeasured ways Philadelphia changed over time differently than control-Philadelphia, in factor 2 unmeasured ways that the difference between Philadelphia moms and controls in zip codes with closures in 1997-1999 differed from the pairs for zip codes without closures. In this sense, the two factors are providing separate, not redundant, information about birth injuries possibly caused by abrupt hospital closures. In contrast, the most affected analysis in Table 2 is heavily redundant with the other two analyses; it expresses the same evidence in a different way.

What does it mean to say that two evidence factors are "nearly independent"? It means that under the null hypothesis, the two P-values for the two factors are stochastically larger than the uniform distribution on the unit square, so viewing them as independent P-values would not lead to inflation of the type-1 error rate. For example, in a 2×3 contingency table, the null hypothesis of independence may be tested by computing a chi-square for independence with one degree of freedom comparing column one to the total of columns two and three, and another chi-square for independence comparing columns two and three (Lancaster 1949, expression 18). These two P-values are not independent, because the second column of the first table is the marginal row total of the second table; however, the pair of resulting P-values are stochastically larger than uniform under the null hypothesis of independence. For detailed discussion of approximate evidence factors together with associated sensitivity analyses, see Rosenbaum (2011).

It was a given that we would look at infant mortality, so that decision was made without reference to the planning sample, and the entire data set was used. Although we do not present that analysis here, it is worth mentioning that for death there were no significant differences in the four analyses that parallel Table 2 and the point estimates suggest that nothing dramatic had occurred.

4 Observational studies with binary outcome and difference-in-differences

4.1 Notation: base and intervention periods; exposed and unexposed regions

There are I pairs, $i = 1, \ldots, I$, of two mothers, $k = 1, 2$, who gave birth in the same year, one giving birth in Philadelphia, denoted $Z_{ik} = 1$, the other giving birth elsewhere, denoted $Z_{ik} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. The mothers have been matched for an observed covariate $\mathbf{x}_{ik}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but there is concern also about an unobserved covariate u_{ik} that was not matched, so possibly $u_{i1} \neq u_{i2}$. Because we match for year of birth, year is included in $\mathbf{x}_{ik}$.

In using mothers outside Philadelphia as controls for mothers inside Philadelphia, we are contemplating what would have happened to paired mothers had they interchanged roles, the Philadelphia mother living and delivering in Pittsburgh, say, and the Pittsburgh mother with whom she is paired delivering in Philadelphia. That is to say, each mother (or her newborn baby) has two potential binary responses, r_{Tik} if mother ik delivered in Philadelphia or r_{Cik} if mother ik delivered elsewhere; see Neyman (1923) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tik} = r_{Cik}$ for $i = 1, \ldots, I$, $k = 1, 2$. In Table 2, (r_{Tik}, r_{Cik}) refers to birth injury of type ICD-9 767.3, and $(r_{Tik}, r_{Cik}) = (1, 0)$ indicates that baby ik would have experienced a birth injury in Philadelphia but not in, say, Pittsburgh. Under Fisher's H_0, $(r_{Tik}, r_{Cik}) = (0, 0)$ or $(r_{Tik}, r_{Cik}) = (1, 1)$, so some babies had birth injuries and others did not, but changing where mother ik delivered would not change whether a birth injury occurred.

Write $R_{ik} = Z_{ik}\, r_{Tik} + (1 - Z_{ik})\, r_{Cik}$ for the observed response of mother ik. Also, write $\mathcal{F} = \{(r_{Tik}, r_{Cik}, \mathbf{x}_{ik}, u_{ik}),\, i = 1, \ldots, I,\, k = 1, 2\}$.

4.2 Model for sensitivity analysis

Even if Fisher's null hypothesis H_0 were true, birth outcomes might be different in Philadelphia and elsewhere because mothers in Philadelphia differ from mothers elsewhere. This may be expressed in terms of a model that relates delivery in Philadelphia to characteristics of mothers and their neighborhoods in $\mathcal{F}$. This model begins by describing the situation prior to matching. The model says that prior to matching, the Z_{ik} were conditionally independent given $\mathcal{F}$ with

$$\Pr\left(Z_{ik} = 1 \middle| \mathcal{F} \right) = \frac{\exp\left\{\kappa\left(\mathbf{x}_{ik}\right) + \gamma u_{ik} + \varrho r_{Cik}\right\}}{1 + \exp\left\{\kappa\left(\mathbf{x}_{ik}\right) + \gamma u_{ik} + \varrho r_{Cik}\right\}}, \;\; 0 \leq u_{ik} \leq 1 \tag{1}$$

where $\kappa\left(\cdot\right)$ is an unknown function. In (1), by Bayes theorem, the term $\kappa\left(\mathbf{x}_{ik}\right)$ permits the distribution of observed covariates $\mathbf{x}_{ik}$ in Philadelphia to differ from the distribution among potential controls before matching, as indeed is seen to be the case in Table 1; moreover, because year is in $\mathbf{x}_{ik}$, (1) permits this difference in observed covariates to be different in different years.

In (1), if $\varrho \neq 0$ then the response r_{Cik} the mother or baby would exhibit outside Philadelphia is related to whether the mother delivers in Philadelphia; that is, by Bayes theorem under (1), birth injuries may be more or less common in Philadelphia than elsewhere. A bias of the form $\varrho \neq 0$ would be the worst type of bias if one were comparing Philadelphia to matched control, but the study compares Philadelphia in two time periods to control in two time periods, and for this comparison $\varrho \neq 0$ is less of a problem. Of course, we cannot estimate ϱ because we observe R_{ik} not r_{Cik}; in particular, we never observe r_{Cik} when $Z_{ik} = 1$, so we could not fit (1) even if we somehow knew that $\gamma = 0$.

If $\gamma \neq 0$ in (1), then the unobserved (and hence unmatched) covariate u_{ik} is related to whether a mother delivers in Philadelphia. Because $0 \leq u_{ik} \leq 1$ in (1), two mothers ik and

ik' with $(\mathbf{x}_{ik}, r_{Cik}) = (\mathbf{x}_{ik'}, r_{Cik'})$ may differ in their odds of delivering in Philadelphia by a factor of at most $\Gamma = \exp(\gamma)$ because u_{ik} and $u_{ik'}$ differ. Because u_{ij} is otherwise unconstrained, it may be different in Philadelphia and control in a different way before and after hospital closures. The term γu_{ik} with $0 \leq u_{ik} \leq 1$ introduces a bias of entirely unspecified form but of a magnitude determined by the magnitude of the sensitivity parameter Γ.

To aid interpretation, it is sometimes convenient to unpack the single parameter Γ into two parameters (Δ, Λ) as $\Gamma = (1 + \Delta\Lambda) / (\Delta + \Lambda)$ where Λ controls the relationship between $u_{i1} - u_{i2}$ and $Z_{i1} - Z_{i2}$ and Δ controls the relationship between $u_{i1} - u_{i2}$ and $Y_{Ci} = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$. Here, Y_{Ci} is 1 if the Philadelphia baby would have had a birth injury if delivery had occurred outside Philadelphia but the control would not, $Y_i = -1$ if the situation were reversed, and $Y_i = 0$ if both babies would have had the same outcome outside Philadelphia. If $\varrho = 0$ so that McNemar's test may be used in a sensitivity analysis comparing Philadelphia babies to controls, a value of $\Gamma = 1.25$ unpacks into the curve $1.25 = (1 + \Delta\Lambda) / (\Delta + \Lambda)$, which includes, for example, $(\Delta, \Lambda) = (2, 2)$ for a u_{ik} that doubles the odds of delivering in Philadelphia and doubles the odds of a birth injury, but it also includes $(\Delta, \Lambda) = (1.4, 5)$ and $(\Delta, \Lambda) = (5, 1.4)$. Analogously, $\Gamma = 2$ unpacks into $(\Delta, \Lambda) = (3, 5)$ and $(\Delta, \Lambda) = (5, 3)$ and other values on the curve $\Gamma = (1 + \Delta\Lambda) / (\Delta + \Lambda)$. For discussion of various aspects of this interpretation of the magnitude of Γ, see Gastwirth, Krieger and Rosenbaum (1998, §2) and Rosenbaum and Silber (2009a).

Our analysis eliminates ϱ in (1) as a nuisance parameter; see Proposition 1. In one sense the value of ϱ does matter because it affects the patterns of data we see, but in another sense it does not matter because no matter what value ϱ takes on, the difference-in-differences analysis will fully account for it. Because of this and because (1) is linear in u_{ik} and r_{Cik} on the logit scale, we may assume without loss of generality that the unobserved covariate, u_{ik}, is uncorrelated with birth injuries in the absence of closures, r_{Cik}, because if this were not the case, we could replace u_{ik} by its least squares residual $\breve{u}_{ik} = u_{ik} - (\vartheta + \eta r_{Cik})$, so $\breve{u}_{ik}$ and r_{Cik} are uncorrelated, and $\kappa(\mathbf{x}_{ik}) + \gamma u_{ik} + \varrho r_{Cik}$ in (1) equals $\{\kappa(\mathbf{x}_{ik}) + \vartheta\} + \gamma \breve{u}_{ik} + (\varrho + \eta) r_{Cik}$. In other words, an unobserved covariate u_{ik} cannot bias the analysis by virtue of being related to birth injuries; it must instead in Factor 1 be related to birth injuries in a different way in different years, or in Factor 2 it must be related to birth injuries in a different way in different zip codes. Although this appears to be an attractive feature of the difference-in-differences analysis, there is a nontrivial price to be paid for it. If ϱ were known to be zero, then Philadelphia and control-Philadelphia could be compared directly, say using McNemar's test for binary responses in matched pairs, and the bias from u_{ik} would be of magnitude γ on the logit scale or $\Gamma = \exp(\gamma)$ in terms of odds; see Rosenbaum (2002, §4.3.2). In contrast, although the difference-in-differences analysis may take u_{ik} to be uncorrelated with r_{Cik}, the analysis faces a bias from u_{ik} of magnitude 2γ on the logit scale or $\Theta = \Gamma^2 = \exp(2\gamma)$ in terms of odds; again, see Proposition 1. In brief, the difference-in-difference analysis is completely unaffected by certain unmeasured biases perfectly correlated with r_{Cik}, but is twice as sensitive to certain other unmeasured biases uncorrelated with r_{Cik}. A mathematically distinct yet conceptually related phenomenon has been noted previously, with difference-in-differences studies being more severely affected by errors-of-measurement (Freeman 1984, Griliches and Hausman1986).

After matching for $\mathbf{x}_{ik}$, so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ and $Z_{i1} + Z_{i2} = 1$, the model (1) implies

$$\Pr(Z_{i1} = 1 | \mathcal{F}, Z_{i1} + Z_{i2} = 1) = \frac{\exp(\gamma u_{i1} + \varrho r_{Ci1})}{\exp(\gamma u_{i1} + \varrho r_{Ci1}) + \exp(\gamma u_{i2} + \varrho r_{Ci2})}. \quad (2)$$

In particular, (2) is $\frac{1}{2}$ if $\gamma = \varrho = 0$, but otherwise treatment assignment is biased.

An alternative but nearly equivalent formulation of the model would omit reference to the population prior to matching — that is, omit reference to (1) — and take (2) as the starting point, that is, take (2) as a model for treatment assignment Z_{ik} within a given matched pair i. Our sense is that the step from (1) to (2) is useful in making it clear what matching for $\mathbf{x}_{ik}$ does and what it fails to do. There is, however, one advantage in beginning with (2). Once a matched pair is formed, there is one Philadelphia zip code attached to that pair, and by including that zip code in $\mathcal{F}$ as an attribute of the pair i (not the mother k), we may understand (2) as a model for the identity k of the Philadelphia mother in pair i. That is, in this formulation, (2) asks: Given that pair i contains two mothers, one from Philadelphia zip-code xxxxx and the other from a zip code with similar attributes elsewhere in Pennsylvania, California or Missouri, and given specific values of (u_{i1}, r_{Ci1}) and (u_{i2}, r_{Ci2}) for these two mothers, what is the chance that mother $i1$ is the Philadelphia mother and $i2$ is the mother from elsewhere? This distinction between starting with (1) and starting with (2) is relevant only to comparisons of pairs with a zip code near a hospital closure versus pairs with a zip code remote from closures — in such comparisons, zip code is treated as a fixed attribute of the pair, as year is treated as a fixed attribute of the pair in temporal comparisons.

4.3 Sensitivity analysis with binary outcomes in difference-in-differences

We wish to focus on a set $\mathcal{S} \subseteq \{1, \ldots, I\}$ of the pairs, and to contrast two subsets of the pairs in $\mathcal{S}$, denoted by $v_i = 1$ and $v_i = 0$. In the first evidence factor in Table 2, all pairs are used, $\mathcal{S} = \{1, \ldots, I\}$, and $v_i = 1$ for birth pairs in years 1997-1999 and $v_i = 0$ for pairs in 1995-1996. In the second evidence factor in Table 2, $\mathcal{S} \subset \{1, \ldots, I\}$ are the pairs in 1997-1999, and $v_i = 1$ for pairs with a Philadelphia mother in a zip code near a closure and $v_i = 0$ for pairs with a Philadelphia mother not near a closure.

Consider testing Fisher's null hypothesis $H_0 : r_{Tik} = r_{Cik}$ using the conditional distribution of $T' = \sum_{i \in \mathcal{S}} \sum_{k=1}^{2} v_i \, Z_{ik} \, R_{ik}$ given $W' = \sum_{i \in \mathcal{S}} \sum_{k=1}^{2} Z_{ik} \, R_{ik}$. In the first evidence factor in Table 2, this is the conditional distribution of T', the number of birth injuries in Philadelphia during the years 1997-1999 of abrupt closures, given the total W' of birth injuries in Philadelphia in all years 1995-1999. If H_0 is true, then $r_{Tik} = r_{Cik} = R_{ik}$, and T' and W' receive only constant contributions from concordant pairs with $0 = R_{i1} - R_{i2} = r_{Ci1} - r_{Ci2}$. Renumber the pairs so that pairs $j = 1, \ldots, J$ are both in $\mathcal{S}$ and are discordant pairs in the sense that $R_{j1} \neq R_{j2}$, and pairs $j+1, \ldots, I$ are either not in $\mathcal{S}$ or are concordant pairs with $R_{j1} = R_{j2}$. Let $T = \sum_{j=1}^{J} \sum_{k=1}^{2} v_j \, Z_{jk} \, R_{jk}$ and $W = \sum_{j=1}^{J} \sum_{k=1}^{2} Z_{jk} \, R_{jk}$ and notice that, given $\mathcal{F}$ and $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$, they differ from T' and W' by a constant when H_0 is true. Write $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{J2})^T$ and $\mathbf{r}_C = (r_{C11}, r_{C12}, \ldots, r_{CJ2})^T$ for the $2J$-dimensional vectors, and write $\mathcal{Z}$ for the set containing the 2^J vectors $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{J2})^T$ with each $z_{jk} = 0$ or $z_{jk} = 1$ and $z_{j1} + z_{j2} = 1$. With a slight abuse of notation, conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ will be abbreviated to conditioning on $\mathcal{Z}$. Write $v_+ = \sum_{j=1}^{J} v_j$.

In Proposition 1, the case (5) of $\Gamma = 1$ is essentially due to Gart (1969). In (4) conditioning on W has eliminated the potential bias in (2) from ϱr_{Ci1}, leaving only the potential bias from γu_{i1}.

Proposition 1 *Let $\Theta = \Gamma^2$. Under H_0 and the sensitivity model (1),*

$$\Upsilon\left(J,\, w,\, v_+,\, t,\, \frac{1}{\Theta}\right) \leq \Pr\left(T \geq t \middle|\, \mathcal{F},\, \mathcal{Z},\, W = w\right) \leq \Upsilon\left(J,\, w,\, v_+,\, t,\, \Theta\right) \tag{3}$$

where

$$\Upsilon\left(J,\, w,\, v_+,\, t,\, \Theta\right) = \frac{\sum_{k=\max(t, w+v_+-J)}^{\min(w, v_+)} \binom{v_+}{k} \binom{J-v_+}{w-k} \Theta^k}{\sum_{k=\max(0, w+v_+-J)}^{\min(w, v_+)} \binom{v_+}{k} \binom{J-v_+}{w-k} \Theta^k} \tag{4}$$

is the extended hypergeometric distribution. In particular, if $\gamma = 0$ in (1), so that $\Gamma = 1$, then

$$\Pr\left(T \geq t \middle| \mathcal{F},\, \mathcal{Z},\, W = w\right) = \sum_{k=\max(t, w+v_+-J)}^{\min(w,v_+)} \frac{\binom{v_+}{k}\binom{J-v_+}{w-k}}{\binom{J}{w}} \tag{5}$$

is the hypergeometric distribution.

Proof. The proof consists in transforming a sensitivity analysis for 2×2 tables counting discordant pairs, such as the 2×2 tables in Table 2, into a sensitivity analysis for unrelated events in 2×2 tables, and then applying standard methods for the latter situation. Throughout the proof, assume H_0 is true for the purpose of testing it, so $r_{Tik} = r_{Cik} = R_{ik}$. Using (2), we have

$$\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}\right) = \frac{\exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk} + \varrho \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} r_{Cjk}\right)}{\prod_{j=1}^{J} \left\{\exp\left(\gamma u_{j1} + \varrho r_{Cj1}\right) + \exp\left(\gamma u_{j2} + \varrho r_{Cj2}\right)\right\}}. \tag{6}$$

Let $\mathcal{Z}_w = \left\{\mathbf{z} \in \mathcal{Z} : w = \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} r_{Cjk}\right\}$. Then $|\mathcal{Z}_w| = \binom{J}{w}$. Conditioning on $W = w$ or equivalently on $\mathbf{Z} \in \mathcal{Z}_w$ yields

$$\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}_w\right) = \frac{\exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk}\right)}{\sum_{\mathbf{b} \in \mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} b_{jk} u_{jk}\right)}$$

which no longer depends upon ϱ. Because the J pairs are discordant, $1 = |r_{Cj1} - r_{Cj2}|$ for every j, we may without loss of generality renumber the two subjects in each pair j so that $r_{Cj1} = 1$ and $r_{Cj2} = 0$; then $v_j \sum_{k=1}^{2} z_{jk} r_{Cjk} = v_j z_{j1}$ and $T = \sum_{j=1}^{J} v_j z_{j1}$ and $W = \sum_{j=1}^{J} z_{j1}$; see Table 3. Also, write $\widetilde{u}_j = u_{j1} - u_{j2}$, so that $-1 \leq \widetilde{u}_j \leq 1$. Define the J-dimensional vectors $\widetilde{\mathbf{u}} = (\widetilde{u}_1, \ldots, \widetilde{u}_J)^T$, $\mathbf{v} = (v_1, \ldots, v_J)^T$ and $\mathbf{1} = (1, \ldots, 1)^T$. Let $\chi(A) = 1$ if event A occurs and $\chi(A) = 0$ otherwise. Then using $\sum_{k=1}^{2} z_{jk} u_{jk} =$

$u_{j2} + z_{j1}\left(u_{j1} - u_{j2}\right)$ and simplifying

$$
\begin{aligned}
&\Pr\left(T \geq t \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}_w\right) \\
= \quad & \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j \sum_{k=1}^{2} z_{jk} r_{Cjk} \geq t\right) \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk}\right)}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} b_{jk} u_{jk}\right)}
\end{aligned}
$$

$$
= \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left(\gamma \sum_{j=1}^{J} z_{j1} \widetilde{u}_j\right)}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} b_{j1} \widetilde{u}_j\right)} = \lambda_t\left(\widetilde{\mathbf{u}}\right), \text{ say.} \tag{7}
$$

Then to prove (3) it suffices to show

$$
\lambda_t\left(\mathbf{1} - 2\mathbf{v}\right) \leq \lambda_t\left(\widetilde{\mathbf{u}}\right) \leq \lambda_t\left(2\mathbf{v} - \mathbf{1}\right), \tag{8}
$$

because $w = \sum_{j=1}^{J} z_{j1}$ is fixed for $\mathbf{z} \in \mathcal{Z}_w$, so that, for example,

$$
\begin{aligned}
\lambda_t\left(2\mathbf{v} - \mathbf{1}\right) &= \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left\{\gamma \sum_{j=1}^{J} z_{j1}\left(2v_j - 1\right)\right\}}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left\{\gamma \sum_{j=1}^{J} b_{j1}\left(2v_j - 1\right)\right\}} \\
= \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left(2\gamma \sum_{j=1}^{J} z_{j1} v_j\right)}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left(2\gamma \sum_{j=1}^{J} b_{j1} v_j\right)} &= \Upsilon\left(J,\, w,\, v_+,\, t,\, \Gamma^2\right).
\end{aligned}
$$

The proof of (8) is identical to the proof of Proposition 1 in Rosenbaum (1995), except in that proof, $0 \leq u_j \leq 1$ whereas here $-1 \leq \widetilde{u}_j \leq 1$, so the upper bound in (3) is attained with $\widetilde{u}_j = 2v_j - 1$ rather than with $u_j = v_j$ (or with $u_j = r_j$ in the notation of that proof).

■

5 Confirmatory analysis using the 90% sample

Table 4 is for the analysis sample of 90% of pairs but is otherwise parallel to Table 2 for the 10% planning sample. The initial impression of Table 4 is that it exhibits many of the same patterns as Table 2, albeit sometimes in a more muted form. For instance, in Table 2, the odds ratio for the most affected groups was 5.80, whereas in Table 4 it is 2.19. This is not surprising given that Table 2 was selected as the most promising of many possible analyses, while Table 4 is an independent replication of that one most promising analysis.

As in Table 2, Table 4 provides several pieces of information consistent with an increase in birth injuries caused by abrupt hospital closures. First, in Factor 1, there is an increase from 1995-1996 to 1997-1999 in the relative frequency of birth injuries in Philadelphia when contrasted with control-Philadelphia. Second, in Factor 2, in the years 1997-1999, there is a greater excess of birth injuries in zip codes near hospital closures than in zip codes remote from hospital closures when contrasted with matched pairs in control-Philadelphia. The test for bias looks at these same zip code groups but in the years before closures, yielding an odds ratio of 1.08 which does not differ significantly from 1. That is to say, zip codes with closures look different after the closures but did not look different before the closures. These pieces of information are not greatly redundant with each other; that is, the first two pieces are approximate evidence factors. The most affected group contrasts zip codes near closures in 1995-1996 to 1997-1999 to matched controls in control-Philadelphia; this yields the largest estimated odds ratio of 2.19. In the absence of bias from unmeasured covariates, this would suggest roughly a doubling of the odds of birth injuries in the affected regions of Philadelphia during the period of abrupt closures.

Unlike Factor 1 in Table 2, in Table 4 there is strong evidence that birth injuries were more common in Philadelphia than in control-Philadelphia in 1995-1996 when there were no closures. Specifically, if McNemar's test is applied to the $844 = 505 + 339$ pairs discordant for birth injury in 1995-1996, the two-sided P-value is 1.2×10^{-8}. Expressed in terms of (1), it appears that $\varrho \neq 0$, so the elimination of ϱ by conditioning is essential. We could not reasonably apply McNemar's test to the $1745 = 1231 + 514$ discordant pairs in 1997-1999, because the comparison in 1995-1996 suggests that at least part of the difference in birth injuries in 1997-1999 was already present in 1995-1996 when there were no closures.

Table 5 is the sensitivity analysis based on Table 4 using Proposition 1. Table 5 eliminates ϱ by conditioning and worries about an unobserved covariate u_{ik} uncorrelated with birth injuries in the absence of closures, r_{Cik}, but possibly related to changes or differences in the frequencies of birth injuries. In Table 5, the analysis is reported in terms of Γ, but from Proposition 1 the sensitivity bound is calculated using the extended hypergeometric distribution with parameter $\Theta = \Gamma^2$.

Birth injuries were more common in Philadelphia than among matched controls even before Philadelphia hospitals began to close their obstetrics units; however, there was a substantial increase in the relative frequency of birth injuries during the years 1997-1999 of abrupt closures, and this increase was substantially more pronounced in zip codes served by hospitals that closed. Moreover, zip codes served by hospitals that closed did not exhibit any relative excess of birth injuries in the years 1995-1996 prior to closures. A moderate bias from an unobserved covariate u_{ik} of magnitude $\Gamma = 1.3$ (or $\Lambda = 2$ and $\Delta = 2.3$ in §4.2) could produce any one of these associations, but this u_{ik} would need to be somewhat unusual: it would need to be uncorrelated with birth injuries r_{Cik} (see §4.2) yet strongly correlated with the change in birth injuries over time and with the post-closure difference in zip codes with closures. Such unobserved covariate is logically possible, but is rendered somewhat less plausible by the need to explain the results in factor 1, factor 2 and the bias test, no one of which is redundant with another.

Table 4 permits two other informative analyses. Although one expects an effect of closures in zip codes with closures, as discussed earlier it is less clear what one should expect for mothers living in zip codes without closures. Comparing pairs discordant for birth injuries in zip codes without closures in 1997-1999 and 1995-1996, the point estimate of the odds ratio is 1.35 with 95% confidence interval [1.11, 1.63], suggesting a small increase in birth injuries for mothers in zip codes without closures. In addition, in the $2 \times 2 \times 2$ table in Table 4 recording pairs discordant for birth injuries, time interval, and with or without closures, the three factor interaction in a log-linear model is not plausibly zero, with likelihood ratio chi-square of 6.27 on 1 degree of freedom, P-value $= 0.012$, so the increase in birth injuries appears to have been larger in zip codes with closures than in zip codes without closures. This pattern of results is not inconsistent with overcrowding at the hospitals that remained open, with mothers remote from the closures being nonetheless affected by the influx of mothers from zip codes with closures.

6 Discussion

Between 1997 and 2007, 12 of 19 hospitals in Philadelphia closed their obstetrics units. Our study built a control Philadelphia with some of the temporal and sociodemographic structure of Philadelphia thereby framing and simplifying questions about how Philadelphia might have changed in the absence of widespread closures of obstetrics units.

Because this series of hospital closures is a unique event, it will never be possible to replicate this study using a new independent sample. Motivated by considerations of improved design sensitivity (Heller et al. 2009), we created an internal replication, a small planning sample of about 13,000 pairs of mothers, and an independent confirmatory analysis sample of about 120,000 pairs. The planning sample suggested a focus on serious birth injuries (ICD-9 767.3), with a relative increase in injuries in the years 1997-1999 of abrupt closures, especially in zip codes served by obstetrics units that abruptly closed.

This led to two evidence factors, one test for bias from unmeasured covariates, and a sensitivity analysis.

In a scientific report, what is the appropriate way to report a split sample analysis? In our methodological discussion here, we have focused on one confirmatory analysis. Our sense is that both exploratory and confirmatory analyses should be presented (Tukey 1980), but that these two types of analyses should be distinguished based on their different histories. That is, a table might present parallel analyses for many interesting outcomes with a bright red line separating confirmatory from exploratory analyses. Above the red line are a few analyses suggested by the planning sample, with independent confirmation or not from the much larger analysis sample. Below the line are exploratory analyses of many outcomes, perhaps aided by some interpretive guidance from multiple testing procedures, such as the Bonferroni inequality, and their associated sensitivity analyses (e.g., Heller et al. 2009, §3.3; Rosenbaum and Silber 2009b, §4.5). Though perhaps interesting and worthy of further study, hypotheses that are first suggested by the analysis sample or the complete data would inevitably be regarded as speculative unless confirmed by multiple testing procedures.

References

Aakvik A. (2001), "Bounding a matching estimator: the case of a Norwegian training program," *Oxford Bulletin of Economics and Statistics*, 63, 115-43.

Abadie, A., Diamond, A., Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493-505.

Abadie, A. and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 112-132.

Abadie, A. and Imbens, G. W. (2011), "Bias-corrected matching estimators for average treatment effects," *Journal of Business and Economic Statistics*, 29, 1-11.

American College of Gynecology (2003), "Practice bulletin: Management of preterm labor," *Obstetrics and Gynecology*, 101, 1039-1047.

Angrist, J., and Krueger, A. (2000), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, eds. O. Ashenfelter, D. Card, Amsterdam: Elsevier, 1277-1366.

Athey, S. and Imbens, G.W. (2006), "Identification and Inference in Nonlinear Difference-in- Differences Models", *Econometrica*, 74, 431-497.

Campbell, D.T. (1957), "Factors relevant to the validity of experiments in social settings," *Psychological Bulletin*, 54, 297-312.

Campbell, D. T. (1969), "Reforms as experiments," *American Psychologist*, 24, 409-429.

Copas, J. and Eguchi, S. (2001), "Local sensitivity approximations for selectivity bias," *Journal of the Royal Statistical Society* B 63, 871-96.

Cox, D. R. (1975), "A note on data-splitting for the evaluation of significance levels," *Biometrika*, 62, 441-4.

Cornfield, J., Haenszel, W., Hammond, E. et al. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*," 22, 173–203.

Diprete, T. A. and Gangl, M. (2004), "Assessing bias in the estimation of causal effects," *Sociological Methodology*," 34, 271-310.

Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Freeman, R.B. (1984), "Longitudinal analyses of the effect of trade unions," *Journal of Labor Economics*, 2, 1-26.

Gart, J. J. (1969), "An exact test for comparing matched proportions in crossover designs," *Biometrika*, 56, 75-80.

Gastwirth, J. L. (1992), "Methods for assessing the sensitivity of comparisons in Title VII cases to omitted variables," *Jurimetrics Journal*, 33, 19–34.

Gastwirth, J. L., Krieger, A. M. and Rosenbaum, P. R. (1998), "Dual and simultaneous sensitivity analysis for matched pairs," *Biometrika*, 85, 907-920.

Griliches, Z. and Hausman, J.A. (1986), "Errors in Variables in Panel Data," *Journal of Econometrics*, 31, 93-118.

Hansen, B. B. (2007), "Optmatch: flexible, optimal matching for observational studies," *R News* **7**, 18-24.

Hansen, B. B., Klopfer, S. O. (2006), "Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, 609–627.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2009), "Split samples and design sensitivity in observational studies," *Journal of the American Statistical Association*, 104, 1090-1101.

Hollowell, J., Oakley, L., Kurinczuk, J. J., Brocklehurst, P., and Gray, R. (2011), "The effectiveness of antenatal care programmes to reduce infant mortality and preterm birth in socially disadvantaged and vulnerable women in high-income countries: a systematic review," *BMC Pregnancy Childbirth*, 11, 13.

Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review*, 93, 126-132.

Kirby, P.B., Spetz, J., Maiuro, L., et al. (2006), "Changes in service availability in california hospitals, 1995 to 2002," *Journal of Healthcare Management*, 51, 26-38.

Lancaster, H. O. (1949), "The derivation and partition of χ^2 in certain discrete distributions," *Biometrika*, 36, 117-129.

Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998), "Assessing sensitivity of regression to unmeasured confounders in observational studies," *Biometrics* 54, 948-63.

Marcus, S. M. (1997), "Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect," *Journal of Educational Statistics*, 22, 193-201.

Meyer, B.D. (1995), "Natural and Quasi-Natural Experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151-161.

Neyman, J. (1923), "On the application of probability theory to agricultural experiments, Reprinted in *Statistical Science*, 1990, 5, 463-80.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. (1999), "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, pp. 1-94, New York: Springer.

Rosenbaum, P. R. (1984), "From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment," *Journal of the American Statistical Association*, 79, 41-48.

Rosenbaum, P.R. (1989), "Optimal matching in observational studies," *Journal of the American Statistical Association*, 84, 1024-1032.

Rosenbaum, P. R. (1995), "Quantiles in nonrandom samples and observational studies," *Journal of the American Statistical Association*, 90, 1424-1431.

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-64.

Rosenbaum, P. R. (2010a), *Design of Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2010b), "Design sensitivity and efficiency in observational studies," *Journal of the American Statistical Association*, 105, 692-702.

Rosenbaum, P. R. (2010c), "Evidence factors in observational studies," *Biometrika*, 97, 333-345.

Rosenbaum, P. R. (2011), "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association*, 106, to appear.

Rosenbaum, P. R. and Rubin, D. B. (1983), “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome,” *Journal of the Royal Statistical Society* B, 45, 212–8.

Rosenbaum, P.R., Rubin, D.B. (1985), “Constructing a control group by multivariate matched sampling methods that incorporate the propensity score,” *American Statistician*, 39, 33-38.

Rosenbaum, P. R. and Silber, J. H. (2009a), “Amplification of sensitivity analysis in observational studies,” *Journal of the American Statistical Association*, 104, 1398-1405.

Rosenbaum, P. R. and Silber, J. H. (2009b), “Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units,” *Journal of the American Statistical Association*, 104, 501-511.

Rubin, D. B. (1974), “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66, 688-701.

Shadish, W. R., Cook, T. D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inferenc,* Boston: Houghton-Mifflin.

Stoll, B. J., Hansen, N. I., Bell, E. F., Shankaran, S., Laptook, A. R., Walsh, M. C., Hale, E. C. et al. (2010), “Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network,” *Pediatrics*, 126, 443-456.

Stuart, E.A. (2010), “Matching Methods for Causal Inference: A review and a look forward,” *Statistical Science*, 25: 1-21.

Trochim, W. M. K. (1985), “Pattern matching, validity and conceptualization in program evaluation,” *Evaluation Review*, 9, 575-604.

Tukey, J. W. (1980), “We need both exploratory and confirmatory,” *American Statistician*, 34, 23-25.

West, S.G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D.C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., Mullen, P.D. (2008), "Alternatives to the randomized controlled trial," *American Journal of Public Health*, 98, 1359-1366.

Yanagawa, T. (1984), "Case-control studies: assessing the effect of a confounding factor," *Biometrika*, 71, 191-194.

Table 1: Covariate balance before and after matching. For Zip Code data, zip-fr means the fraction of the Zip Code with this attribute. An absolute standardized difference in mean of 0.2 or greater is in **bold**.

Sample Size	5,998,111	132,786	132,786	Absolute	
	Potential	Philadelphia	Matched	Standardized	
	Controls	Births	Controls	Difference	
Covariate	Covariate Mean or Proportion			Before	After
	Mom's Neighborhood (Zip code)				
Income (K$)	46	30	30	**1.16**	0.04
Income Missing	0.00	0.00	0.00	0.06	0.00
Poverty (zip-fr)	0.15	0.25	0.23	**0.91**	0.13
Poverty Missing	0.00	0.00	0.00	0.06	0.00
High School (zip-fr)	0.74	0.68	0.69	**0.37**	0.07
HS Missing	0.00	0.00	0.00	0.06	0.00
College (zip-fr)	0.22	0.15	0.15	**0.51**	0.01
College Missing	0.00	0.00	0.00	0.06	0.00
	Mom				
Mom's Age	28	26	26	**0.21**	0.01
Parity	2.10	2.20	2.20	0.07	0.03
Parity Missing	0.00	0.01	0.01	0.09	0.04
Prenatal Care (Month Started)	2.40	2.70	2.60	**0.22**	0.04
PC Missing	0.02	0.11	0.08	**0.37**	0.11
	Mom's Education				
Below 8th Grade	0.10	0.02	0.02	**0.32**	0.02
Some High School	0.17	0.21	0.20	0.11	0.04
HS Graduate	0.30	0.38	0.40	0.17	0.05
Some College	0.20	0.19	0.19	0.02	0.01
College Graduate	0.13	0.09	0.10	0.11	0.01
More than College	0.09	0.06	0.06	0.11	0.00
Missing	0.01	0.04	0.04	0.17	0.04
	Mom's Race				
White	0.71	0.31	0.32	**0.87**	0.03
Black	0.07	0.42	0.46	**0.88**	0.11
Asian	0.07	0.03	0.03	0.18	0.03
Other	0.12	0.06	0.05	**0.20**	0.05
Missing	0.02	0.17	0.14	**0.52**	0.13
	Mom's Health Insurance				
Government	0.40	0.40	0.39	0.01	0.02
Other Insurance	0.57	0.58	0.60	0.02	0.04
Uninsured	0.03	0.01	0.01	0.11	0.04
Missing	0.00	0.01	0.00	0.11	0.06
	Baby				
Birth Weight, (grams)	3345	3179	3189	**0.26**	0.02
Birth Weight Missing	0.00	0.00	0.00	0.04	0.03
Gestational Age (Weeks)	39	38	38	0.14	0.01
Gestational Age Missing	0.05	0.01	0.01	**0.22**	0.02
Small at Gestational Age	0.09	0.14	0.12	0.16	0.05

Table 2: Results for birth injury in the planning component of the split sample. The table counts discordant pairs in which exactly one baby in the pair was injured. Factor 1 contrasts the affected years (1997-1999) with hospital closures in Philadelphia to the base years (1995-1996) without closures. Factor 2 looks within the affected years (1997-1999) and contrasts zip codes with (W) closures to zip codes without (W/O) closures. The bias test contrasts the same zip codes, but in the years (1995-1996) prior to closures, so a difference there cannot be an effect caused by hospital closures, and would instead indicate a failure to control some unmeasured bias. The *P*-values and odds ratios are from Gart's (1969) procedure.

A comparsion focused on the most affected groups				
Birth Outcomes in Discordant Pairs		Zip Codes With Closures 1995-1999		
Philadelphia Baby	Control Baby	Affected 1997-1999	Base 1995-1996	Total (+)
Injured	Not Injured	52	8	60
Not Injured	Injured	12	11	23
	Total (+)	64	19	83
Odds Ratio		5.80		
Alternative		1-sided		
P-value		0.0016		
95% Interval		$[2.03, \infty)$		

		Factor 1			Factor 2			Bias Test		
Birth Outcome		1995-1999			1997-1999			1995-1996		
Discordant Pairs		Time Period			Zip Code Closures			Zip Code Closures		
Philadelphia Baby	Control Baby	Affected 97-99	Base 95-96	+	W	W/O	+	W	W/O	+
Injured	Not Injured	141	43	184	52	89	141	8	35	43
Not Injured	Injured	43	42	85	12	31	43	11	31	42
	Total (+)	184	85	269	64	120	184	19	66	85
Odds Ratio		3.19			1.51			0.65		
Alternative		1-sided			1-sided			2-sided		
P-value		0.000023			0.19			0.44		
95% Interval		$[1.95, \infty)$			$[0.76, \infty)$			[0.20, 2.03]		

Table 3: General form of the table under H_0 after renumbering within the J discordant pairs so that $r_{Cj1} = 1$ and $r_{Cj2} = 0$ for each j.

	$v_j = 1$	$v_j = 0$	Total
$z_{j1} = 1$	$\sum_{j=1}^{J} v_j z_{j1}$	$\sum_{j=1}^{J} (1 - v_j) z_{j1}$	w
$z_{j1} = 0$	$\sum_{j=1}^{J} v_j (1 - z_{j1})$	$\sum_{j=1}^{J} (1 - v_j)(1 - z_{j1})$	$J - w$
Total	v_+	$J - v_+$	J

Table 4: Results for birth injury in the analysis component of the split sample. This table, which is the basis for conclusions rather than hypothesis generation, has the same structure as Table 2 but is based on an independent sample of pairs that is approximately nine times larger.

A comparsion focused on the most affected groups				
Birth Outcomes in Discordant Pairs		Zip Codes With Closures 1995-1999		
Philadelphia Baby	Control Baby	Affected 1997-1999	Base 1995-1996	Total (+)
Injured	Not Injured	475	131	606
Not Injured	Injured	137	83	220
	Total (+)	612	214	826
Odds Ratio		2.19		
Alternative		1-sided		
P-value		3.71×10^{-6}		
95% Interval		[1.63, ∞)		

		Factor 1			Factor 2			Bias Test		
Birth Outcome		1995-1999			1997-1999			1995-1996		
Discordant Pairs		Time Period			Zip Code			Zip Code		
Philadelphia	Control	Affected	Base		Closures			Closures		
Baby	Baby	97-99	95-96	+	W	W/O	+	W	W/O	+
Injured	Not Injured	1231	505	1736	475	756	1231	131	374	505
Not Injured	Injured	514	339	853	137	377	514	83	256	339
	Total (+)	1745	844	2589	612	1133	1745	214	630	844
Odds Ratio		1.61			1.73			1.08		
Alternative		1-sided			1-sided			2-sided		
P-value		4.37×10^{-8}			9.33×10^{-7}			0.69		
95% Interval		[1.39, ∞)			[1.42, ∞)			[0.78, 1.51]		

Table 5: Sensitivity analysis in the 90% analysis sample. The table gives the upper bound on the one-sided P-value testing the null hypothesis of no effect of closures on birth injuries for the three effect comparisons in Table 4 for departures from random assignment of various magnitudes Γ.

Γ	Upper bound on 1-sided P-value		
	Most Affected	Factor 1	Factor 2
1.0	0.0000	0.0000	0.0000
1.1	0.0003	0.0007	0.0012
1.15	0.0019	0.0145	0.0118
1.2	0.0083	0.1126	0.0636
1.25	0.0277	0.3892	0.2066
1.3	0.0730	0.7301	0.4445

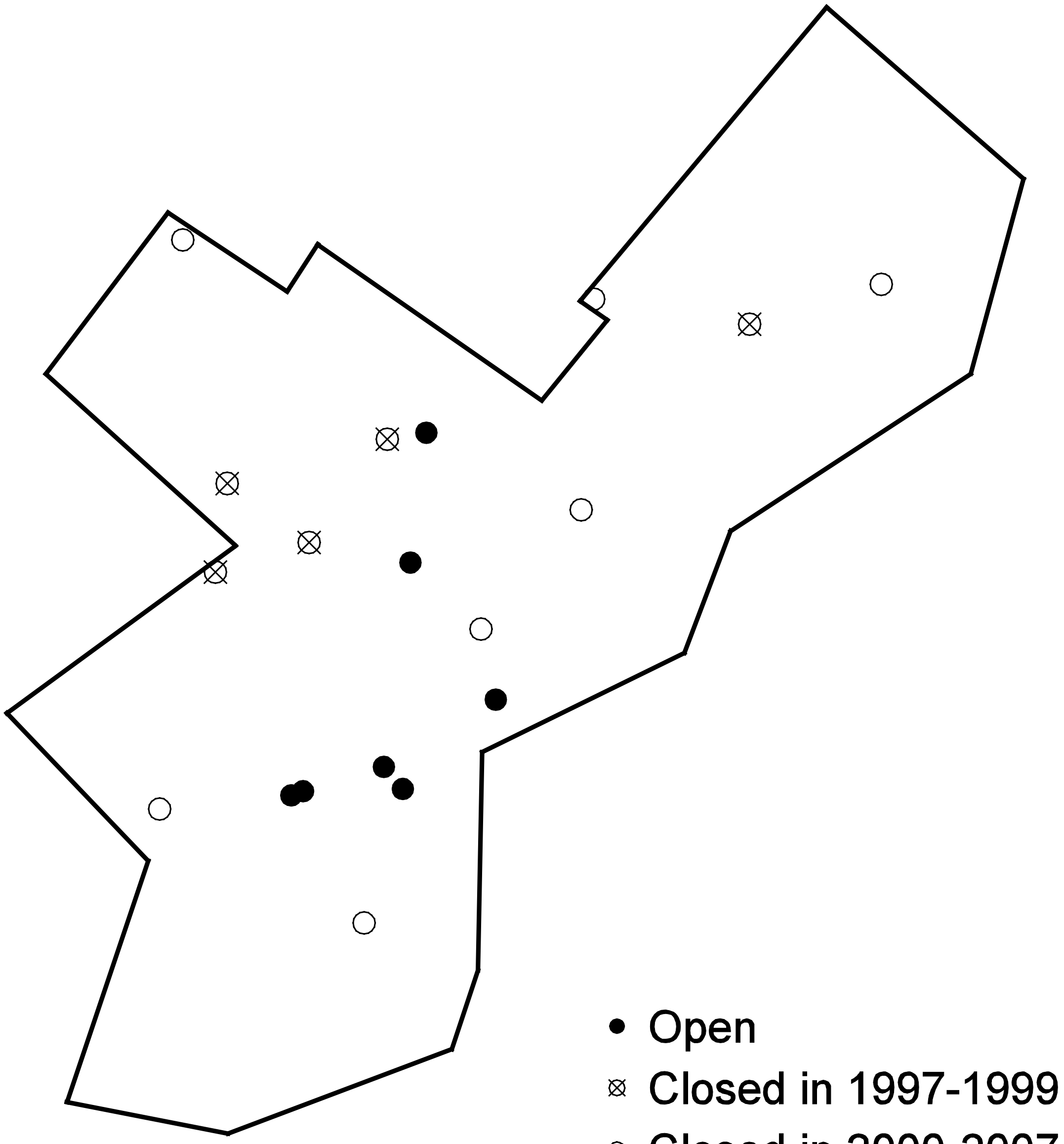

Figure 1: Map of the City of Philadelphia showing hospitals that closed their obstetrics units. The analysis in the current paper focuses on closures in 1997-1999, before the City intervened to pace and organize the process of closure.

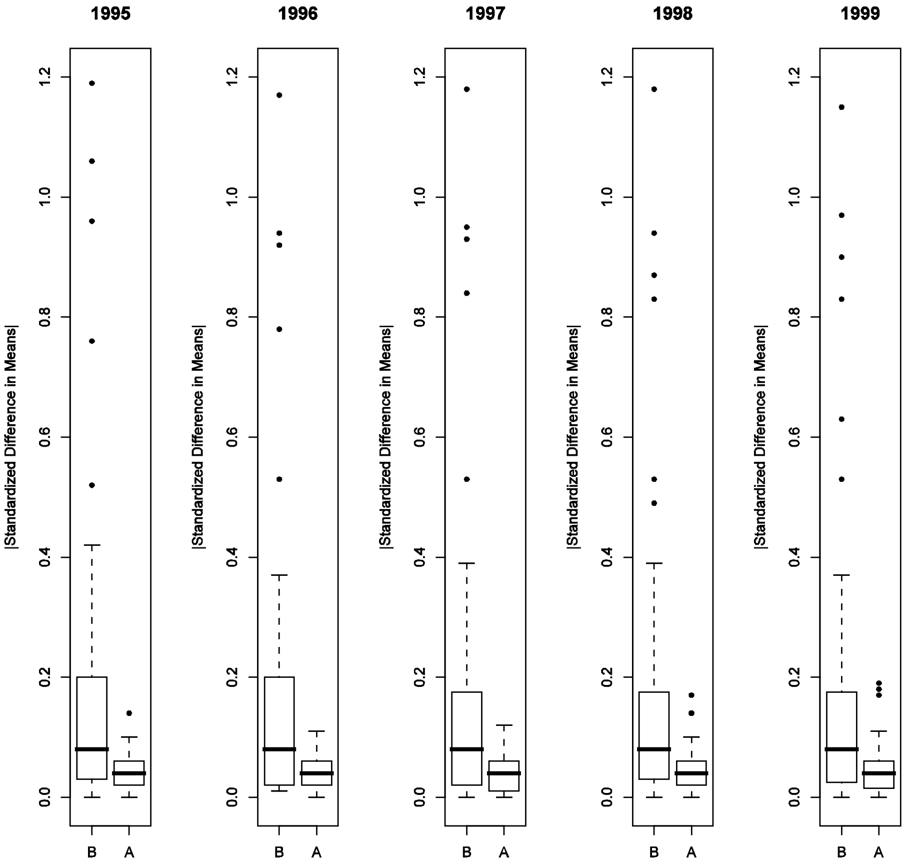

Figure 2: Covariate balance before (B) and after (A) matching for 59 covariates in each of five years, measured as the absolute difference in means in units of a pooled standard deviation.

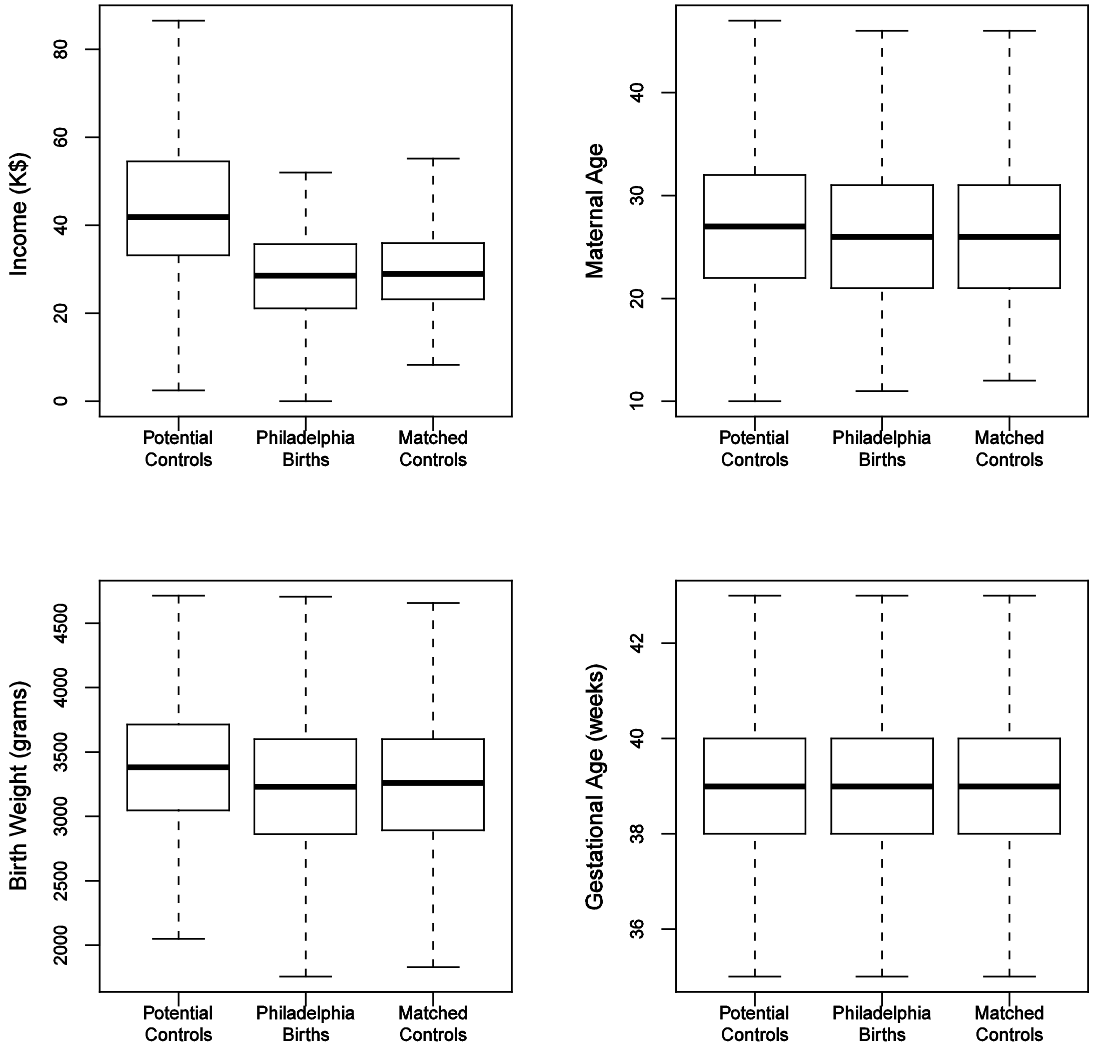

Figure 3: Covariate imbalance before and after matching for four continuous covariates, namely income, maternal age, birth weight, and gestational age.