Stuart G. Baker* and Karen S. Lindeman

Revisiting a Discrepant Result: A Propensity Score Analysis, the Paired Availability Design for Historical Controls, and a Meta-Analysis of Randomized Trials

Abstract: There is an ongoing controversy over whether epidural analgesia for women in labor increases the probability of Caesarean section. Previous research compared results from three methods for estimating the effect of epidural analgesia on the probability of Caesarean section: a propensity score analysis, the paired availability design for historical controls, and meta-analysis of randomized trials. The propensity score analysis and a paired availability design gave substantially different results with the latter in closer agreement with results of a meta-analysis of randomized trials. We updated this investigation in three ways. First, we discussed the use of causal graphs for variable selection in the propensity score analysis. Second, we introduced new extrapolation estimates to improve generalizability for the paired availability design and the meta-analysis of randomized trials with crossovers. Third, we included the results from more recent studies. This analysis provides a window into various topics in causal inference and comparative effectiveness research.

Keywords: causal graph, causal inference, Caesarean section, comparative effectiveness research, epidural analgesia

***Corresponding author: Stuart G. Baker,** National Cancer Institute, E-mail: sb16i@nih.gov
Karen S. Lindeman, Johns Hopkins Medical Institutions, E-mail: klindema@jhmi.edu

1 Introduction

About 15 years ago a major controversy arose over whether epidural analgesia for women in labor (which provides superior pain relief to other analgesia) increases the probability of Caesarean section (C/S). Three methods to estimate the effect of epidural analgesia versus other analgesia on the probability of C/S were investigated: a propensity score analysis, the paired availability design, and a meta-analysis of randomized trials [1]. A propensity score is an individual's probability of receiving treatment as function of observed baseline variables [2, 3]. A propensity score analysis is the estimation of the effect of treatment on outcome using the propensity score. The paired availability design is a before-and-after comparison of treatment effect in multiple medical centers, with an adjustment for various changes in availability of treatment over time to estimate the effect of receipt of treatment on outcome [4]. When applied to the question of whether epidural use increased the probability of C/S, a propensity score analysis and a paired availability design gave qualitatively different results with the latter in closer agreement with results of a meta-analysis of randomized trials [1]. Because the controversy over the effect of epidural analgesia on the probability of C/S is ongoing, we revisit these results in light of new data and new methodologies.

A recent advancement in the propensity score analysis is the recognition that, with an observed collider (a variable directly influenced by at least two other variables), causal graphs are needed for appropriate variable selection [5–8]. See Pearl [9] for a comprehensive discussion of causal graphs and Pearl [10] for a summary. In the presence of colliders, a causal graph allows for a more precise definition of an omitted confounder and is needed to define M-bias [9]. Both an omitted confounder and M-bias can invalidate a propensity score analysis. M-bias in a propensity score analysis is a controversial topic [5–8, 11, 12], perhaps

related to specialized terminology which can be confusing. To help clarify M-bias, this review of the topic discusses M-bias in a framework involving only probability theory.

A recent advancement in the paired availability design is the recognition that extrapolation estimates can improve generalizability [13]. The goal of the paired availability design is to estimate treatment effect (the effect of receipt of treatment on outcome) while avoiding the self-selection bias of comparing a new treatment in the later time period with an old treatment in the earlier time period. The paired availability design compares outcome in *all* eligible persons in a later time period with outcome in *all* eligible persons in an earlier time period and estimates treatment effect under plausible assumptions. The key to this estimation is a potential outcomes model involving what were later called, in a more general context, principal strata [14]. In the paired availability design, the potential outcomes model involves time periods of lower and higher availabilities of a new treatment, and the principal strata refer to four types of subjects who would receive a specified pair of treatments (old or new), if the time period of arrival were to correspond to low or high availability of the new treatment. Under reasonable assumptions, it is possible to estimate treatment effect in one or two principal strata. Sometimes an estimate of treatment effect within one principal stratum is considered an end in itself [15]. However, because the composition of principle strata can differ in a target population, a more appropriate goal is to estimate treatment effect among all persons [13, 16]. An extrapolation estimate is an estimate of treatment effect among all persons that uses information on the fraction of persons in the principal stratum. Here, new types of extrapolation estimates are introduced and compared via simulation.

The article is organized as follows. Section 2 discusses propensity scores. Section 3 discusses the paired availability design. Section 4 discusses the meta-analysis of randomized trials. Section 5 compares estimates of the effect of epidural analgesia on the probability of C/S based on a propensity score analysis, a paired availability design, and a meta-analysis of randomized trials. Section 6 is a discussion.

2 Propensity scores

A propensity score analysis with subclassification (stratification by the propensity score) is an appealing method for adjusting for baseline variables when estimating the effect of treatment on outcome in an observational study with two treatment groups [2, 3]. The propensity score method with subclassification has two advantages over regression analysis in terms of "offering initial trustworthy comparisons that are easy to communicate" [17]. First, it can easily flag a troublesome situation in which a member of one treatment group has values of baseline variables outside the range of the values of baseline variables for the other treatment group. This lack of overlap, which is harder to detect in regression models, could invalidate an analysis. Second, a propensity score method with subclassification is less dependent on the functional form of the model than a regression model. In addition, without biasing estimates of treatment effect, researchers can investigate various models for the propensity score before the final selection [18]. Rosenbaum and Rubin [2, 3] justified the use of propensity scores in terms of ignorable treatment assignment. Here, we present a justification based directly on adjusting for appropriate baseline variables, which facilitates the connection of the propensity score analysis to theory of causal graphs.

2.1 Graphical view of adjustment

A baseline variable is a variable observed before the receipt of treatment. Adjustment for baseline variables is a well-known technique for estimating the effect of intervention on outcome in an observational study with concurrent treatments. To mathematically describe adjustment, let Y denote outcome, T denote treatment, and X denote the appropriate baseline variables that when used for

adjustment yield an unbiased (causal) estimate of treatment effect. For now, suppose the appropriate baseline variables X are those variables that directly influence both the receipt of treatment and the outcome. The theory of causal graphs (to be discussed) provides a comprehensive rule for selecting the appropriate baseline variables X.

The naïve estimate of the effect of T on Y is obtained by substituting estimates into the following equation,

$$\begin{aligned}\Delta_{\text{NAIVE}} = &\int_x (\text{pr}(Y=1|T=1,X=x)\ \text{pr}(X=x|T=1)\ dx \\ &- \int_x (\text{pr}(Y=1|T=0,X=x)\ \text{pr}(X=x|T=0)\ dx.\end{aligned} \quad [1]$$

The reason the naïve estimate can yield incorrect conclusions is that distribution of X can differ in persons with $T = 0$ and persons with $T = 1$. Adjustment makes the distribution of X the same in the two groups, namely $\text{pr}(X = x|\ T = t) = \text{pr}(X = x)$, and transforms Equation (1) into

$$\begin{aligned}\Delta_{\text{CASUAL}} = &\int_x (\text{pr}(Y=1|T=1,X=x)\ \text{pr}(X=x)\ dx \\ &- \int_x (\text{pr}(Y=1|T=0,X=x)\ \text{pr}(X=x)\ dx.\end{aligned} \quad [2]$$

The causal effect is estimated by substituting estimates for the probabilities in Equation (2).

Figure 1 presents a graphical view of adjustment using a plot to graphically explain Simpson's paradox that was independently proposed by Jeon *et al.* [19] and Baker and Kramer [20]. See also Wainer [21]. The plot involves a single baseline variable X which takes values of 0 and 1, and is sufficient for adjustment. In Figure 1, the estimate of $\text{pr}(X = 1 \mid T = 1)$ is 1/3, and the estimate of $\text{pr}(X = 1 \mid T = 0)$ is 2/3. The estimates of $\text{pr}(Y = 1 \mid T = t, X = x)$ are 9/10 for $\{T = 1, X = 1\}$, 3/10 for $\{T = 1, X = 0\}$, 8/10 for $\{T = 0, X = 1\}$, and 2/10 for $\{T = 0, X = 0\}$. The naïve estimate of the effect of T on Y is obtained by substituting estimates into

$$\begin{aligned}\Delta_{\text{NAIVE}} = &\{\text{pr}(Y=1|T=1,X=1)\ \text{pr}(X=1|T=1) \\ &+ \text{pr}(Y=1|T=1,X=0)\ \text{pr}(X=0|T=1)\} \\ &- \{\text{pr}(Y=1|T=0,X=1)\ \text{pr}(X=0|T=1) \\ &+ \text{pr}(Y=1|T=0,X=0)\ \text{pr}(X=0|T=0)\}\end{aligned} \quad [3]$$

to obtain the estimate

$$\begin{aligned}D_{\text{NAIVE}} &= \{(9/10)(1/3) + (3/10)(2/3)\} - \{(8/10)(1/3) + (2/10)(2/3)\} \\ &= 5/10 - 6/10 = -1/10.\end{aligned} \quad [4]$$

In contrast, the causal estimate of the effect of T on Y is obtained by substituting estimates into the following equation involving adjustment,

$$\begin{aligned}&\Delta_{\text{CASUAL}} = \{\text{pr}(Y=1|T=1,X=1)\ \text{pr}(X=1) + \text{pr}(Y=1|T=1,X=0)\ \text{pr}(X=0)\} \\ &- \{\text{pr}(Y=1|T=0,X=1)\ \text{pr}(X=1) + \ \text{pr}(Y=0|T=0,X=0)\ \text{pr}(X=0)\}\end{aligned} \quad [5]$$

to obtain

$$\begin{aligned}D_{\text{CASUAL}} = &\{(9/10)\ \text{pr}(X=1) + (3/10)\ \text{pr}(X=0)\} \\ &- \{(8/10)\ \text{pr}(X=1) + (2/10)\ \text{pr}(X=0)\} = 1/10,\end{aligned} \quad [6]$$

In Figure 1, the naïve estimate of the effect of $T = 1$ versus $T = 0$ on Y is the vertical distance between the dashed horizontal lines, namely 5/10–6/10 = – 1/10. The casual estimate is the vertical distance between the diagonal lines, namely 3/10–2/10 = 1/10, an example of Simpson's paradox involving different signs for causal and naïve estimates.

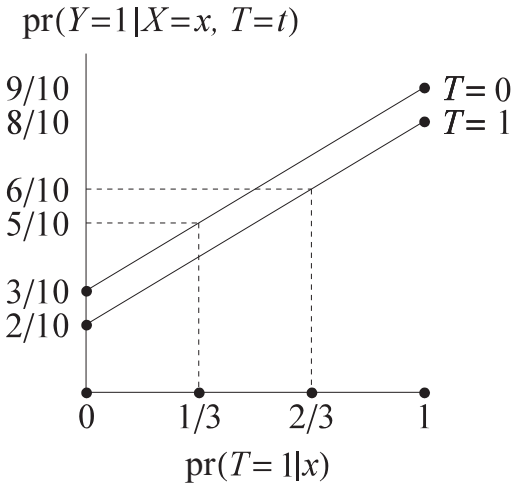

Figure 1 Graphical view of adjustment in a simple example.

2.2 Propensity scores from an adjustment perspective

When the set of baseline variables X sufficient for adjustment consists of many variables, estimating the causal effect via Equation (2) is problematic because estimates of $\text{pr}(Y = 1|t, x)$ can be unstable due to small sample sizes. To solve this problem, Rosenbaum and Rubin [2] proposed the use of a propensity score. The propensity score is the probability of receiving treatment as a function of the baseline variables. Rosenbaum and Rubin [3] justified the use of the propensity score for casual estimation by proving that if treatment assignment is ignorable on baseline variables, then treatment assignment is ignorable on the propensity score constructed from those baseline variables. Here an adjustment argument is presented to justify using propensity scores to obtain a casual estimate of treatment effect. Let $s = \text{pr}(T = 1|x)$ denote the propensity score. Without invoking any properties of the propensity score, the naïve estimated treatment effect is obtained by substituting estimates into

$$\begin{aligned}\Delta_{\text{NAIVE(PS)}} = &\int_x \int_s (\text{pr}(Y = 1|T = 1, x, s)\ \text{pr}(X = x/s, T = 1)\ \text{pr}(S = s|T = 1)ds\,dx \\ &- \int_x \int_s (\text{pr}(Y = 1|T = 0, x, s)\ \text{pr}(X = x|s, T = 0)\ \text{pr}(S = s|T = 0)ds\,dx.\end{aligned} \quad [7]$$

In Equation (7), the distributions of X and S differ by group T. As derived in Rosenbaum and Rubin [3], with more details given in Appendix A, the propensity score has the property that

$$\text{pr}(X = x|t, s) = \text{pr}(X = x|s). \quad [8]$$

Equation (8) implies that $\text{pr}(X = x|\ T = 0, s) = \text{pr}(X = x|\ T = 1, s)$, which says that the distribution of baseline variables is the same for each treatment group after conditioning on the propensity score. Substituting Equation (8) into Equation (7) gives

$$\begin{aligned}\Delta_{\text{NAIVE(PS)}} = &\int_x \int_s (\text{pr}(Y = 1|T = 1, x, s)\ \text{pr}(X = x|s)\ \text{pr}(S = s|T = 1)ds\,dx \\ &- \int_x \int_s (\text{pr}(Y = 1|T = 0, x, s)\ \text{pr}(X = x|s)\ \text{pr}(S = s|T = 0)ds\,dx.\end{aligned} \quad [9]$$

In Equation (9), only the distribution of S differs by group T. Adjustment using $\text{pr}(S = s\ |T = t) = \text{pr}(S = s)$ transforms Equation (9) into the following equation used to compute a causal estimate,

$$\begin{aligned}\Delta_{\text{CAUSAL(PS)}} = &\int_x \int_s (\text{pr}(Y=1|T=1,x,s)\,\text{pr}(X=x|s)\text{pr}(S=s)ds\,d\text{x}\\ &-\int_x \int_s (\text{pr}(Y=1\,|\,T=0,x,s)\,\text{pr}(\text{X}=x\,|\,\text{s})\,\text{pr}(S=s)ds\,dx\\ &-\int_s (\text{pr}(Y=1\,|\,T=1,s)\text{pr}(S=s)ds\\ &-\int_s (\text{pr}(Y=1\,|\,T=0,s)\,\text{pr}(S=s)ds.\end{aligned} \quad [10]$$

The method of subclassification of propensity scores [3] can be viewed as a method of estimating the causal effect in Equation (10) by substituting the appropriate estimates. The first step is to compute an estimated propensity score, denoted here as s^*. Under the method of subclassification, the estimated propensity score is split into five quintiles. Let $\text{p}(t, q)$ denote an estimate of $\text{pr}(Y = 1|t,s^*)$ for all persons with estimated propensity score s^* in the q^{th} quintile of the estimated propensity scores. Then, the propensity score subclassification estimate of causal effect is

$$D_{\text{CAUSAL}}(\text{PS}) = \Sigma_{q=1}^{5}\text{p}(1,q)/5 - \Sigma_{q=1}^{5}\text{p}(0,q)/5. \quad [11]$$

Forming subclasses based on the inverse variance of estimated treatment effect can reduce the mean squared error [22]. Other uses of the propensity score (not discussed here) are weighting based on the inverse of the estimated propensity score [23], propensity score matching [24, 25], and regression adjustment with propensity scores [24].

2.3 Review of causal graphs

Before the advent of causal graphs and the recognition of colliders (to be discussed), the recommendation for correct causal inference based on multivariate observational data was to adjust for all observed variables associated with both treatment and outcome [26]. However with the advent of causal graphs [9], this recommendation has been superseded by a recommendation (to be discussed) that allows for colliders. Below, is a brief summary of the relevant aspects of causal graphs for this discussion.

A causal graph is a diagram showing the direct influence of variables on each other in an observational study. In a causal graph, the direct influence of variable A on variable X is written as $A \rightarrow X$. Writing $A \rightarrow X \rightarrow B$ says that X directly influences (or directly causes) B without additional information from A. Causal graphs are usually said to be "directed" meaning there are no double arrows, and "acyclic" meaning there are no loops specified by arrows.

Causal graphs are needed for variable selection when there is an observed collider. A collider is a variable directly influenced by at least two other variables so that graphically there are at least two arrows pointing to it. Using the terminology in Morgan and Winship [27], the basic patterns for causal relations are

Collider: $A \rightarrow X \leftarrow B$ (mutual causation),
Non-collider: $A \rightarrow X \rightarrow B$ (mediation),
$A \leftarrow X \leftarrow B$ (mediation),
$A \leftarrow X \rightarrow B$ (mutual dependence).

Colliders play a key role in causal inference because conditioning on a collider has a very different implication than conditioning on a non-collider [9]. See Appendix A for a formal proof.

If X is a collider between A and B, then A and B are unconditionally independent, but A and B are dependent conditional on X. To illustrate this property of a collider, suppose you are waiting for a bus. Here, A is the occurrence or not of traffic accident that would delay the bus, B is an indicator of whether a slow or fast driver is behind the wheel of the bus (as determined by scheduling done weeks in advance), and X is an indicator of whether the bus is on time or late. If there is a traffic accident or the slow driver is behind the wheel, the bus will be late. Prior to the scheduled bus arrival time, the occurrence or not of a traffic accident

(*A*) and the presence of a slow or fast driver (*B*) are independent. If, prior to the scheduled bus arrival time, you hear a traffic report that does not mention any accident on the bus route (*A*), you have no additional information about which driver is behind the wheel (*B*), so events *A* and *B* are independent. However, if besides hearing this traffic report (*A*), you also find the bus is late (conditioning on *X*), your most likely conclusion is that the slow driver is behind the wheel of the bus (*B*). Thus *A* and *B* are independent events, but conditioning on *X* makes *A* and *B* dependent events.

If *X* is a non-collider between *A* and *B*, then *A* and *B* are unconditionally dependent, but *A* and *B* are independent conditional on *X*. To illustrate this property of a non-collider, suppose you are going to a bus stop to catch a bus that will take you to an appointment. You are concerned the bus might be delayed due to a traffic accident. Here, *A* is a traffic accident or not, *X* is an indicator if the bus is on time or late, and *B* is an indicator of being on-time or late for your appointment. Suppose, that prior to the scheduled bus arrival time you hear on the radio that there is an accident that could delay the bus. In this case, information on the traffic accident (*A*) increases the probability of your being late for your appointment (*B*), so events *A* and *B* are dependent. Once the bus arrives late (conditioning on *X*), information about a prior traffic accident (A), provides no additional information about whether or not you will be late for your appointment (*X*), assuming any later traffic accident is independent of an earlier traffic accident. Thus, *A* and *B* are dependent, but conditioning on *X* makes *A* and *B* independent.

A path in a causal graph is a sequence of variables connected by arrows. There are two fundamental types of paths connecting treatment *T* and outcome *Y*, back door and front door. A back door path is any path linking *T* and *Y* which ends in an arrow pointing to *T*. A front door path is a path connecting *T* and *Y* in which an arrow points away from *T*. The causal effect is the effect of *T* on *Y* on the front door path. A back door path from *T* to *Y* in which *T* and *Y* are dependent can bias estimates of causal effects obtained from the front door path. The fundamental requirement for causal inference from *T* to *Y* after conditioning on a set of baseline variables *X* is that *T* and *Y* are independent on *all* back door paths after conditioning on *X*, which implies that a change in *T* systematically affects *Y* only through the front door path.

The d-separation criterion (Pearl 2009b) is a rule for determining whether or not *T* and *Y* are independent on all back door paths after conditioning on a set of baseline variables *X*. A set of baseline variables *X* is said to block a single back door path P from *T* to *Y* if conditioning on *X* makes *T* and *Y* independent on that path. In other words, "blocking" of a back door path can be thought of as blocking dependence between *T* and *Y* on the back door path. Operationally, a set of baseline variables *X* blocks a single back door path P from *T* to *Y* if either

(i) path P contains at least one non-collider in *X* (so *T* and *Y* are independent due to conditioning on a non-collider) or
(ii) path P contains at least one collider outside *X* with no descendants in *X* (so that *T* and *Y* are independent due to not conditioning on a collider).

D-separation of *T* and *Y* by conditioning on *X* requires that *X* block *all* back door paths. If any back door path is not blocked after conditioning on *X*, there is no d-separation of *T* and *Y* by conditioning on *X*, so adjustment on *X* will yield biased estimates of casual effects.

2.4 Causal graphs and propensity scores

Given the possible presence of colliders, a propensity score analysis can be improved by using causal graphs to select variables to include in the propensity score. Based on the theory of causal graphs, the following two assumptions are needed for unbiased (causal) estimation of treatment effect when using propensity scores,

Assumption PS-1 (no omitted confounder),
Assumption PS-2 (no latent M-bias collider).

These two assumptions will be discussed in turn.

2.4.1 Omitted confounder

An omitted confounder with respect to a propensity score involving baseline variables X is a non-collider *outside* of X that is on a back door path from treatment T to outcome Y that is *not* blocked by X, so T and Y are dependent on this back door path, and T and Y are not d-separated, preventing correct causal inference.

Figure 2 illustrates an omitted confounder. The propensity score excludes X_1 so that back door path $T \leftarrow C \rightarrow A \leftarrow X_1 \rightarrow B \rightarrow D \rightarrow Y$ is not blocked by the variables X_2 in the propensity score, and thus T and Y are dependent on this path. Hence X_1 is an omitted confounder, there is no d-separation of T and Y, and there is bias in estimating the causal effect on the front door path.

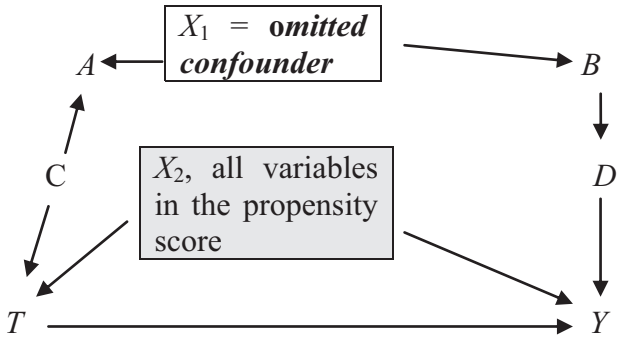

Figure 2 An omitted confounder in propensity score analysis. The shaded box indicates variables included in the propensity score.

2.4.2 Latent M-bias collider

An M-diagram has the form $T \leftarrow A \rightarrow X \leftarrow B \rightarrow Y$, which can be plotted in the shape of the letter "M" where the upper points of the letter are A and B and the middle point of the letter is X, which is a collider (Figure 3).

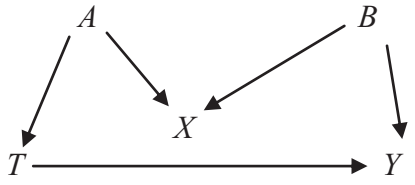

Figure 3 An M-diagram.

A latent M-bias collider (a name coined here) with respect to a propensity score involving adjustment for a set of baseline variables X is a collider *included in* the set X that is on a back door path from treatment T to outcome Y which is *not* blocked by any variables in X; in this case, T and Y are dependent (not d-separated) on the aforementioned back door path preventing correct causal inference.

Figure 4 illustrates a latent M-bias collider. The propensity score (which depends on only X_1 and X_2) includes the X_1 collider but not the non-colliders C, A, B, and D, so the back door path $T \leftarrow C \rightarrow A \rightarrow X_1 \leftarrow B \rightarrow D \rightarrow Y$ is *not* blocked by conditioning on the propensity score, and thus T and Y are dependent on this path. Hence, X_1 is a latent M-bias collider, there is no d-separation of T and Y, and there is bias in estimating the causal effect on the front door path.

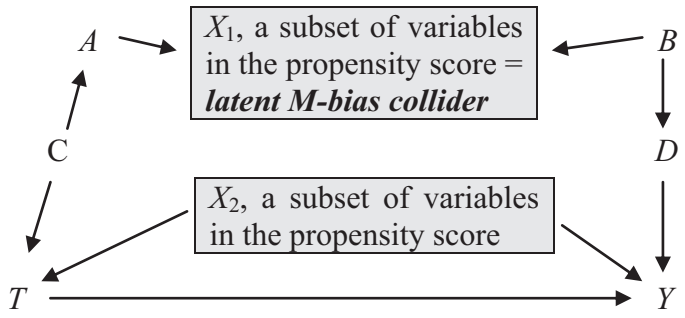

Figure 4 A latent M-bias collider. Shaded boxes indicate variables included in the propensity score.

2.5 Application to epidural analgesia and C/S

Based on the theory of causal graphs, the two key assumptions for a valid propensity score analysis are (i) no omitted confounder and (ii) no latent M-bias collider. To investigate these assumptions in the application to epidural analgesia and C/S, we formulated the causal graph in Figure 5.

Our primary discussion involves the propensity score analysis in Lieberman *et al.* [28]. Author KSL, an obstetric anesthesiologist, formulated the causal graph in Figure 5 using the variables considered in Lieberman *et al.* [28] as well as other variables thought to be appropriate. Some aspects of the causal graph might be debatable as there are limited data that confirm or deny its structure. Nevertheless, there is support for the following key considerations.

(a) *Dystocia (abnormal labor) directly influences cervical dilation*. Impey *et al.* [29, 30] found that women with less cervical dilation at admission had a greater duration of labor. A likely explanation is that dystocia failed to dilate the cervix to a similar degree as in women with normal labor.

(b) *Hypertension directly influences cervical dilation*. Obstetricians frequently admit patients with gestational hypertension to the hospital for induction or augmentation of labor with very little cervical dilation in order to deliver the baby in a manner to avoid potential problems from prolonged hypertension.

(c) *Dystocia directly influences the probability of C/S*. Dystocia is an accepted indication for C/S.

(d) *Chronic hypertension directly influences the type of analgesia selected*. Obstetricians and anesthesiologists encourage patients with hypertension to select labor epidural analgesia for its safety over other labor analgesia techniques and for its potential use during C/S, which is more likely in women with hypertension.

(e) *Obstetrical group directly influences both type of analgesia received and probability of C/S*. Different obstetricians may favor different types of analgesia and have different rates of C/S. Also, according to an observational study by Beilin *et al.* [31], obstetrical group is a confounder for epidural analgesia and C/S.

(f) *Intense pain in labor directly influences both type of analgesia and probability of C/S*. Women in intense pain are more likely to request and receive epidural analgesia with its superior pain relief than other analgesia. Intense pain may also lead to C/S by impairing the labor process.

Inspection of the causal graph in Figure 5 allows evaluation of the assumptions needed for unbiased (causal) estimation of treatment effect.

Assumption PS-1 (no omitted confounder) does not likely hold because intense pain and obstetric group are omitted confounders. Graphically, intense pain and obstetric group are non-colliders outside the variables in the propensity score and the path T (epidural) $\leftarrow X_0$ (intense pain, obstetric group) $\rightarrow B$ (dystocia) $\rightarrow Y$ (C/S) is not blocked. Lieberman *et al.* [28] did not report collecting data on pain levels.

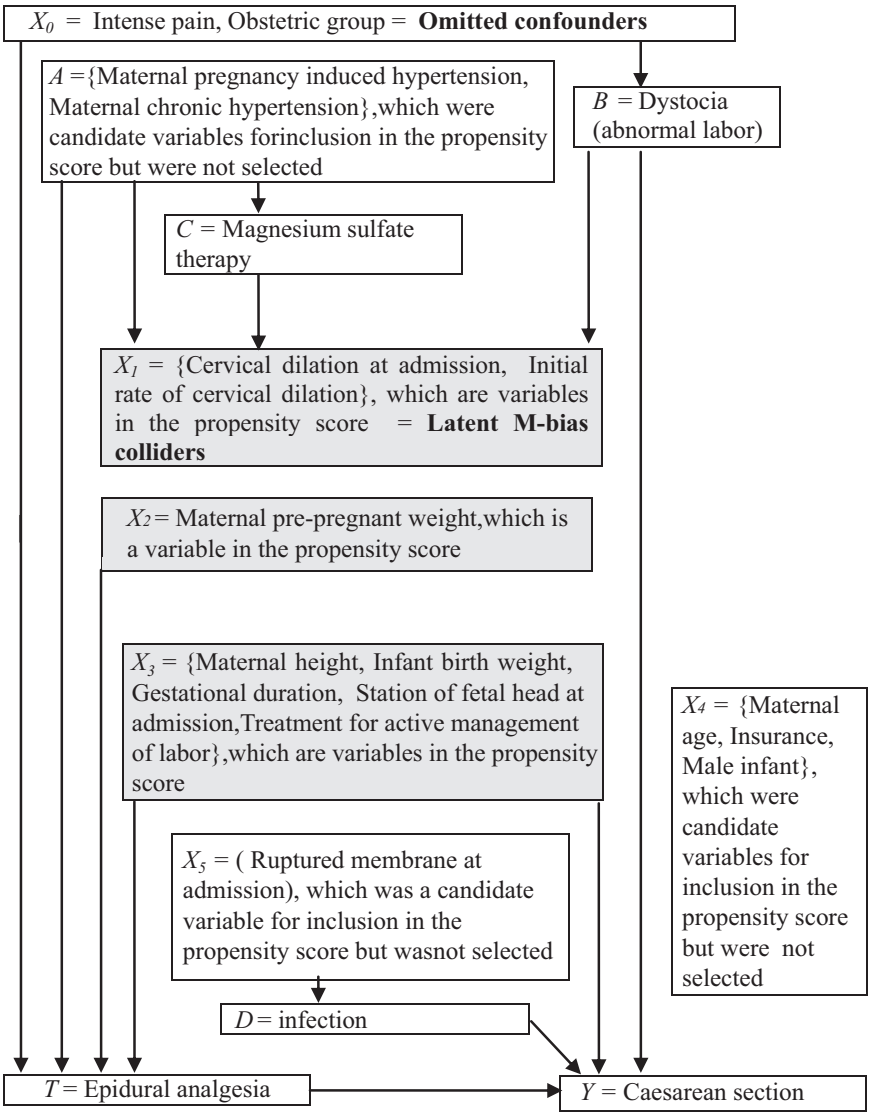

Figure 5 Causal graph for observational study of effect of epidural analgesia on the probability of Caesarean section. Shaded boxes indicate variables included in the propensity score.

Although Lieberman *et al.* [28] obtained data from 17 prenatal care sites [32], they did not report obstetrical group.

Assumption PS-2 (no latent M-bias collider) does not likely hold because cervical dilation at admission and initial rate of cervical dilation are latent M-bias colliders. Graphically, the path T (epidural) ← A (maternal pregnancy induced hypertension, maternal chronic hypertension) → X_1 (cervical dilation at admission and initial rate of cervical dilation) ← B (dystocia) → Y (C/S) is not blocked by the propensity score which includes X_1 but neither A nor B.

The omitted confounder and the latent M-bias colliders provide a plausible explanation as to why the result from the propensity score analysis of Lieberman *et al.* (1996) differed substantially from the result from meta-analysis of randomized trials in Baker and Lindeman [1].

Nguyen *et al.* [33] also used a propensity score analysis to estimate the effect of epidural analgesia on the probability of C/S. Their baseline variables were type of care, age, race, marital status, height, weight,

educational level, country of birth, language spoken, payment method, prior medical history, major ante partum complications, cervical dilation at admission, rupture of membranes at presentation, complications at presentation. As with Lieberman *et al.*, they found a strong effect of epidural analgesia on the probability of C/S. Also, as with Lieberman *et al.* [28], bias could have arisen due to the same omitted confounders of intense pain and obstetrical group, and the same latent M-bias collider of cervical dilation at admission.

Lieberman *et al.* [28] only studied women who were nulliparous (no previous live births). In contrast, Nguyen *et al.* [33] performed separate analyses by parity (number of previous successful live births) and found that the estimated effect of epidural on the probability of C/S was larger for women who were nulliparous than for women who were multiparous (at least one previous live birth). The estimates and confidence intervals based on these propensity score analyses are summarized in Section 5.

2.6 A note on variable selection

Lieberman *et al.* [28] used stepwise logistic regression to select baseline variables for the propensity score analysis. Let X_{TY} denote variables that directly cause both treatment selection and outcome, X_T denote variables that directly cause only treatment selection, and X_Y denote variables that directly cause only outcome. Based on simulations for propensity score subclassification involving no unobserved confounders and no observed colliders, Brookhart *et al.* [34] and Austin *et al.* [25] found that the mean squared error was smaller when adjusting for a combination of X_{TY} and X_Y versus a combination of X_{TY} and X_T. If there are no unobserved confounders and there are observed colliders, Pearl [9, 35] noted that adjusting for either (i) all observed direct causes of outcome or (ii) all observed direct causes of treatment selection gives the same unbiased estimate of the effect of treatment on outcome (as either strategy blocks all back-door paths). However, if there are unobserved confounders; Pearl [35] prefers the former to avoid bias amplification. VanderWeele and Shpitser [36] showed that if an adjustment using any subset of the observed baseline variables (including colliders) controls for confounding, confounding could also be controlled by adjusting for baseline variables that are direct causes of treatment selection or outcome or both. These variable selection methods will yield biased results if there is an omitted confounder or a latent M-bias collider.

3 The paired availability design

The paired availability design is a method for estimating treatment effect (the effect of receipt of treatment on outcome) using historical controls [1, 4, 13, 37–39). There are two problems with using historical controls to estimate treatment effect. First, a naïve comparison of outcome among recipients of new treatment at a later time with outcome of recipients of old treatment at an earlier time is generally biased. Second, a comparison of all outcomes at a later time with all outcomes an earlier time gives a diluted estimate of treatment effect even if there is no systematic error over time.

The paired availability design circumvents these problems by comparing all outcomes at a later time with all outcomes at an earlier time, and then using an adjustment to obtain an unbiased estimate (if assumptions hold) of the effect of receipt of treatment in a subgroup called a principal stratum. Each principal stratum is a baseline variable determined by the actual or hypothetical receipt of different treatments at different time periods (or, in other applications, different randomization groups). Principal stratification is the name given to this type of model when applied more generally [14].

The principal stratification model with two plausible assumptions for identifiability that is discussed here was independently formulated by Permutt and Hebel [40], Baker and Lindeman [4], Imbens and Angrist [41] followed by Angrist *et al.* [42], and Cuzick *et al.* [43]. Baker *et al.* [13] introduced the extrapolation estimate to increase the plausibility of generalizing the treatment effect from a principal stratum to the entire population. Here, new extrapolation estimates are introduced, evaluated by simulation, and applied

to data on epidural analgesia and probability of C/S. Following Baker and Lindeman [4], Cuzick *et al.* [43], and Cox [44], our formulation of the principal stratification model uses standard probability notation instead of the potential outcomes notation of Angrist *et al.* [42].

3.1 Review of the paired availability design

Let T0 denote a standard treatment which is available to all eligible patients, and let T1 denote a new treatment with limited availability. Let time period $Z = 0$ denote lower availability of T1 and time period $Z = 1$ denote higher availability of T1 at a particular medical center. The goal is to estimate the average causal effect over medical centers of the receipt of T1 instead of T0 on the probability of outcome Y.

3.1.1 Time periods similar to randomization groups

The paired availability design requires that estimates of the effect of time period on outcome are unbiased, as if the time periods were similar to randomization groups. To this end, the following four assumptions [13] are invoked:

Assumption PAD-1 (stable population): from one time period to the next, there are no changes in the characteristics of the eligible population that would affect the probability of outcome;

Assumption PAD-2 (stable ancillary care): from one time period to the next, there are no systematic changes in patient management unrelated to the treatment of interest that would affect the probability of outcome after any adjustment;

Assumption PAD-3 (stable disease progression): from one time period to the next, there are no systematic changes in the timing of disease-related events or the spectrum of manifestations of disease in the absence of treatment;

Assumption PAD-4 (stable evaluation): from one time period to the next, there are no changes in eligibility criteria and definitions of outcome.

Some of these assumptions can be made more plausible by design. To support **Assumption PAD-1 (stable population)**, investigators can choose medical centers with little in- or out- migration, such as geographically isolated medical centers or army medical centers. Also, if data are available from medical centers in which treatment does not change over time, investigators may be able to adequately estimate the background effect of time period on the probability of outcome and use that estimate to adjust for bias due to changes over time [37]. To support **Assumption PAD-2 (stable ancillary care)**, investigators can follow protocols and minimize staff changes. If these assumptions hold, the change in probability of outcome as a function of time period is

$$\Delta_{\text{overall}} = \text{pr}(Y = 1 \,|\, Z = 1) - \text{pr}(Y = 1 \,|\, Z = 0). \quad [12]$$

3.1.2 Principal stratification and identifiability

To estimate the effect of receipt of treatment from a comparison of outcomes in different time periods, the paired availability design uses a principal stratification model with two assumptions for identifiability.

A participant is defined as a person who arrives at the designated medical center during either of the time periods and receives either treatment T0 or T1. Using this definition of a participant and the convention that availability of treatment T1 is greater in time period $Z = 1$ than in time period $Z = 0$, the principal strata, denoted $R = r$, are as follows:

Never-receiver, $R = n$, if a participant would receive treatment T0 if the time period of arrival corresponded to an availability of treatment at either the level in time period $Z = 0$ or the level in time period $Z = 1$,

Consistent-receiver, $R = c$, if a participant would receive T0, if the time period of arrival corresponded to an availability of treatment at the level in time period $Z = 0$ and would receive T1 if the time period of arrival corresponded to an availability of treatment at the level in time period $Z = 1$,

Inconsistent receiver, $R = i$, if a participant would receive T1 if the time period of arrival corresponded to availability of treatment at the level in time period $Z = 0$, and would receive T0 if the time period corresponded to an availability of treatment at the level in time period $Z = 1$,

Always-receiver, $R = a$, if a participant would receive treatment T1 if the time period of arrival corresponded to availability of treatment at either the level in time period $Z = 0$ or the level in time period $Z = 1$.

Let $T = 0$ denote receipt of treatment T0, and let $T = 1$ denote receipt of treatment T1. The definitions of the principal strata imply

$$\text{pr}(T = 1 \,|\, Z = 0) = \text{pr}(R = i) + \text{pr}(R = n), \qquad [13]$$

because only participants in principal strata $R = i$ and $R = n$ would receive T1 in time period $T = 0$. Similarly,

$$\text{pr}(T = 1 \,|\, Z = 1) = \text{pr}(R = c) + \text{pr}(R = a) \qquad [14]$$

because only participants in principal strata $R = c$ and $R = a$ would receive T1 in time period $T = 1$. The treatment effect in principal stratum r is written as

$$\Delta_{\text{stratum(r)}} = \text{pr}(Y = 1 \,|\, Z = 1, r) - \text{pr}(Y = 1 \,|\, Z = 0, r). \qquad [15]$$

This treatment effect in Equation (15) is called a causal effect because it represents the effect of treatment in a subset of participants defined by a baseline variable with the two time periods taking the role of randomization groups. Using Equation (15), Equation (12) can be written as

$$\Delta_{\text{overall}} = \Delta_{\text{stratum}(a)}\text{pr}(R = a) + \Delta_{\text{stratum(c)}}\text{pr}(R = c) - \Delta_{\text{stratum}(i)}\text{pr}(R = i) - \Delta_{\text{stratum}(n)}\text{pr}(R = n) \qquad [16]$$

To obtain the causal effect of receipt of treatment, the following two additional assumptions are invoked:

Assumption PAD-5 (stable treatment effect): the effect of treatment on the probability of outcome does not change over the time periods among always-receivers and never-receivers;

Assumption PAD-6 (stable preferences): preference for treatment does not change over the time periods.

These assumptions have important implications for formulating the causal effect. **Assumption PAD-5 (stable treatment effect)** implies the probability of outcome in always-receives and never-receivers does not vary with time period, namely

$$\text{pr}(Y = 1 \,|\, Z = z, R = n) = \text{pr}(Y = 1 \,|\, \text{R} = n), \qquad [17]$$

$$\text{pr}(Y = 1 \,|\, Z = z, R = a) = \text{pr}(Y = 1 \,|\, \text{R} = a). \qquad [18]$$

Substituting Equations (17) and (18) into Equation (16) gives

$$\Delta_{\text{overall}} = \Delta_{\text{stratum}(c)}\text{pr}(R = c) - \Delta_{\text{stratum}(i)}\text{pr}(R{=}i). \qquad [19]$$

The implications of **Assumption PAD-6 (stable preferences)** depend on whether availability is fixed or random. Under fixed availability, the availability of treatment T1 in the second time period subsumes the availability of T1 in the first time period, for example availability in evenings and daytime subsumes availability only in the daytime. Under random availability, availability of treatment T1 in the second

time period is greater than in the first time period, but there is a chance component related to its timing. Under fixed availability, **Assumption PAD-6 (stable preferences)** implies

$$\mathrm{pr}(R = i) = 0, \quad [20]$$

because no participant would switch from T1 to T0 when T1 becomes more available. Under random availability, **Assumption PAD-6 (stable preferences)** implies

$$\mathrm{pr}(Y = 1 \,|\, Z = z, R = c) = \mathrm{pr}(Y = 1 \,|\, Z = 1 - z, R = i), \quad [21]$$

because receipt of T1 or T0 occurs by chance among principal strata $R = c$ and $R = i$. Based on Equations (13) and (14), the effect of time period on the probability of receiving treatment T1 is

$$\Delta_{\text{treated}} = \mathrm{pr}(T = 1 \,|\, Z = 1) - \mathrm{pr}(T = 1 \,|\, Z = 0) = \mathrm{pr}(R = c)^{\circ} - \mathrm{pr}(R = i). \quad [22]$$

Equation (21) implies $\Delta_{\text{stratum}(c)} = \Delta_{\text{stratum}(i)}$. Combining either Equation (20) or (21) with Equations (19) and (22), and simplifying gives the causal effect of treatment in the consistent-receiver principal stratum (under fixed availability) or the combination of consistent-receiver and inconsistent-receiver principal strata (under random-availability), which is written in both cases as

$$\Delta_{\text{stratum}(c)} = \Delta_{\text{overall}} / \Delta_{\text{treated}}, \quad [23]$$

Look-see proofs, which have been derived for the Pythagorean theory [45] and for the sum of an infinite geometric series [46], provide insight not available with algebraic proofs. Figures 6 and 7 provide look-see proofs for the causal effect in Equation (23) under fixed and random availability, respectively.

In Figures 6 and 7, the area of the box (outlined by the black perimeter) is proportional to the probability of being in a principle stratum. Also, the area of the colored bar as a fraction of the area of the box is the probability of outcome in the principal stratum. Treatment received (which is a function of principal stratum and time period) is indicated above each box for the principal stratum. Under **Assumption PAD-5 (stable treatment effect),** the area of the colored bars in both time periods are the same among never-receivers and always-receivers. Under **Assumption PAD-6 (stable preferences)** with fixed availability, there are no inconsistent-receivers (Figure 6). Under **Assumption PAD-6 (stable preferences)** with random availability, the area of the colored bar corresponding to a given treatment is the same for a consistent-receiver and an inconsistent- receiver (Figure 7). Graphically, the causal effect $\Delta_{\text{overall}}/\Delta_{\text{treated}}$ is the difference (between time periods) in the sum of the bar areas divided by the difference (between time periods) in box areas corresponding to T1. The right sides of Figures 6 and 7 demonstrate that this causal effect is a difference in probabilities of outcome among either (i) consistent-receivers under fixed availability (Figure 6) or (ii) a combination of consistent-receivers and inconsistent-receivers under random availability (Figure 7).

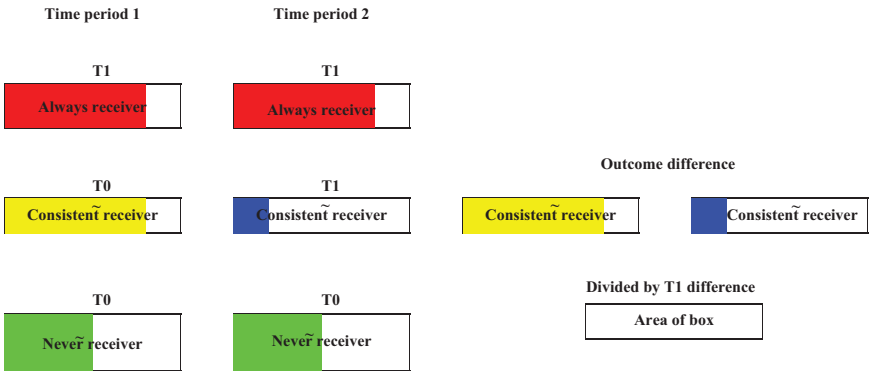

Figure 6 Look-see proof of causal effect under fixed availability.

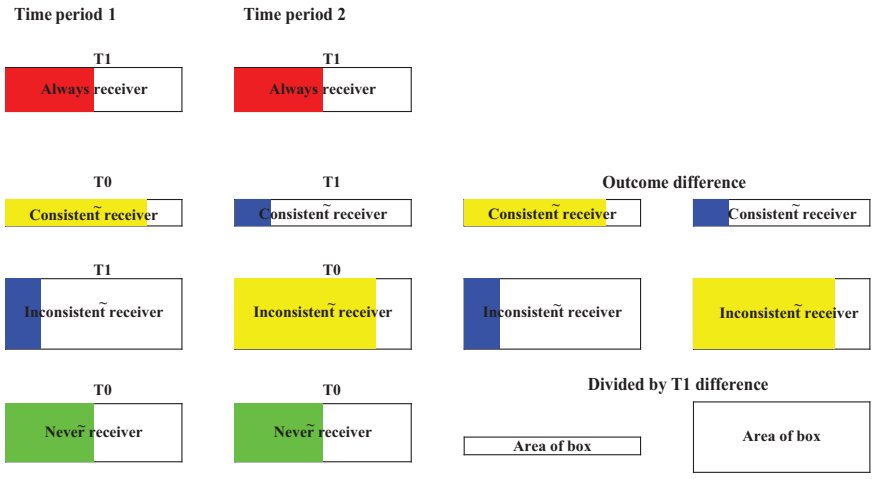

Figure 7 Look-see proof of causal effect under random availability.

In terms of estimates, Equation (23) can be written as

$$d_{\text{stratum}(c)} = d_{\text{overall}} / d_{\text{treated}}. \quad [24]$$

where d_{overall} is the difference in estimated probabilities of outcome between the two time periods, and d_{treated} is the differences in estimated probabilities of receiving the treatment of interest in the two time periods.

The estimate $d_{\text{stratum}(c)}$ in Equation (24) is analogous to the estimated local average treatment effect in principal stratification models for randomized trials with all-or-none compliance [42]; it is maximum likelihood if all estimates of parameters yield a perfect fit to the categorical data [4]. If there is no perfect fit solution, the maximum likelihood estimate of treatment effect can be computed using an iterative algorithm [47, 13].

3.1.3 Generalizability

The assumptions presented so far ensure an unbiased (causal) estimate of treatment effect within principal stratum $R = c$, under fixed availability, or within both principal strata $R = c$ and $R = i$ under random availability. A limitation of these estimates is that the composition of principal strata can change in the future, making the estimates invalid. For example, a favorable estimate of treatment effect may encourage some never-receivers to become consistent-receivers. Therefore, the goal is to estimate the effect of receipt of treatment among all persons, not just among persons in some principal strata. This goal requires the following assumption:

Assumption PAD-7 (generalizability): the estimated effect of receipt of treatment in principal strata $R = c$ (under fixed availability) or $R = c$ and $R = i$ (under random availability) is a good estimate of the effect of receipt of treatment among all eligible persons.

3.2 Extrapolation estimates

The plausibility of **Assumption PAD-7 (generalizability)** can be increased by using an extrapolation estimate [13]. This discussion focuses on principal stratum $R = c$ under fixed availability, but the results apply directly to principal strata $R = c$ and $R = i$ under random availability.

There are three key ideas for extrapolation. First, if the change in the fraction of consistent-receivers equals 1, then everyone is receiving T1 instead of T0. (In practice, the maximum change in fraction of

consistent-receivers may be a little less than one, as for example when epidural analgesia cannot be administered due to rapid delivery). Second, there is an estimate of the fraction of women who are consistent-receivers at each medical center. Third, the estimates from these fractions may be used to extrapolate to the treatment effect when the fraction of women who are consistent-receivers equals 1.

Let j index medical center. Let d_j denote the value of $d_{\text{stratum}(c)}$ for medical center j. Let v_j denote the estimated variance of d_j. Let f_j denote the estimate of d_{treated} for medical center j. We discuss the following four estimates: the original random effects estimate without extrapolation and three extrapolation estimates.

3.2.1 Random effects (RE)

DerSimonian and Laird [48] introduced well-known random effects estimate for a meta-analysis of random trials, which has been applied to the paired availability design. In the DerSimonian and Laird [48] framework, the estimated variance under a random effects model is $v_{\text{RAN}j} = v_j + \tau^2$, where $\tau^2 = \max[0,\{Q - (k-1)\}/\{\Sigma_j w_j - \Sigma_j (w_j)^2/(\Sigma_j w_j)\}$, $Q = \Sigma_j w_j \{d_j - \Sigma_j w_j d_j\}^2$, $w_j = 1/v_j$ and k is the number of studies. Let $w_{\text{RAN}} = 1/v_{\text{RAN}j}$. The estimated treatment effect and its variance are

$$\begin{aligned} d_{\text{RE}} &= d_j h_{REj}, \\ v_{\text{RE}} &= \Sigma_j v_{\text{RAN}j} (h_{\text{RE}j})^2, \\ \text{where } h_{\text{RE}j} &= w_{\text{RAN}j}/\Sigma_j w_{\text{RAN}j}. \end{aligned} \quad [25]$$

To compute confidence intervals for the random effects estimate, Baker and Lindeman [1] used a computationally intensive paired permutation approach of Follman and Proschan [49]. A much simpler approximate confidence interval [49, 50] is $CI_{\text{RE}} = [d_{\text{RE}} - t_{(k-1)} se_{\text{RE}}, d_{\text{RE}} - t_{(k-1)} se_{\text{RE}}]$, where $se_{\text{RE}} = (v_{\text{RE}})^{1/2}$ and $t_{(k-1)}$ is the upper 0.025 quantile of a t-distribution with k–1 degrees of freedom.

3.2.2 Random effects using fraction treated (REF)

The REF extrapolation estimate takes the RE estimate and multiplies the weight by the fraction receiving treatment, giving the following estimated treatment effect and variance,

$$\begin{aligned} d_{\text{REF}} &= \Sigma\Sigma_j d_j h_{\text{REF}j}, \\ v_{\text{REF}} &= \Sigma_j v_{\text{RAN}j} (h_{\text{REF}j})^2, \\ \text{where } h_{\text{REF}j} &= f_i w_{\text{RAN}j}\Sigma_j f_i w_{\text{RAN}j}. \end{aligned} \quad [26]$$

The approximate 95% confidence interval is $\text{CI}_{\text{REF}} = [d_{\text{REF}} - t_{(k-1)} se_{\text{REF}}, d_{\text{REF}} - t_{(k-1)} se_{\text{REF}}]$, where $se_{\text{REF}} = (v_{\text{REF}})^{1/2}$.

3.2.3 Random effects using fraction treated squared (REF2)

The REF2 extrapolation estimate takes the RE estimate and multiplies the weight by the square of the fraction receiving treatment, giving the following estimated treatment effect and variance,

$$\begin{aligned} d_{\text{REF2}} &= \Sigma_j d_j h_{\text{REF2}j}, \\ v_{\text{REF2}} &= \Sigma_j v_{\text{RAN}j} (h_{\text{REF2}j})^2, \\ \text{where } h_{\text{REF2}j} &= (f_j)^2 w_{\text{RAN}j}/\Sigma_j (f_j)^2 w_{\text{RAN}j}. \end{aligned} \quad [27]$$

The approximate 95% confidence interval is $\text{CI}_{\text{REF2}} = [d_{\text{REF2}} - t_{(k-1)} se_{\text{REF2}}, d_{\text{REF2}} - t_{(k-1)} se_{\text{REF2}}]$, where $se_{\text{REF2}} = (v_{\text{REF2}})^{1/2}$.

3.2.4 Flat/linear/quadratic/sigmoid model (FLQS)

The FLQS estimate involves fitting the following models to the data

$$\begin{aligned} &\textit{Flat}\text{: } d_j = \alpha_0 + \varepsilon_j, \\ &\textit{Linear}\text{: } d_j = \beta_0 + \beta_1 f_j + \varepsilon_j, \\ &\textit{Quadratic}\text{: } d_i = \gamma_0 + \gamma_1 f_j + \gamma_2 (f_j)^2 + \varepsilon_j, \\ &\textit{Sigmoid}\text{: } d_j = \phi_0 + (_1 - \phi_0) \exp(f_j - \phi_2)/\{1 + \exp(f_j - \phi_2)\} + \varepsilon_j, \end{aligned} \quad [28]$$

where ε_j denotes a random error which is normally distributed with mean 0. Model selection is based on the smallest value of the Akaike information criterion [51]. If the flat model is selected, the RE estimates are reported. Otherwise, the extrapolation estimate is computed at $f_j = f_{\max}$, where $f_{\max}$ is the largest anticipated change in the fraction who would receive treatment in a population.

3.2.5 Simulation

We investigated the properties of the estimates RE, REF, REF2, and FLQS (with $f_{\max} = 1$) using simulation. Because the number of possible data patterns is too large for a comprehensive investigation, we selected eight informative patterns for investigation: four shapes of flat, linear, quadratic, and sigmoid, each with either a small and large variance for the random generation of points (Figure 8).

Mean squared errors (Figure 9) and coverage probabilities for nominal 95% confidence intervals (Figure 10) were computed using 2000 simulations of the patterns in Figure 8.

For the flat shape, RE had the smallest mean squared error, and all estimates had good coverage probabilities. For the linear, quadratic, and sigmoid data, REF2 and FLQS performed best in terms of mean squared error and coverage probabilities. Because no single estimate performed best for all patterns, a sensitivity analysis using all the estimates is recommended.

This simulation included some medical centers with a large fraction of consistent receivers. The extrapolation estimate will be less informative if there are no medical centers with a large fraction who are consistent receivers.

3.3 Application to epidural analgesia and C/S

The paired availability design was applied to the study of the effect of epidural analgesia on the probability of C/S. To reduce bias from including only published studies, data were collected from abstracts, articles, and a personal note (Table 1). In this application, treatment T0 is not receiving epidural analgesia (which includes receiving opioid analgesia or not receiving any analgesia) and treatment T1 is receiving epidural analgesia. Based on the information reported in some studies and the clinical expertise of author KSL, who is an obstetric anesthesiologist, the first six assumptions for the paired availability design were thought to be plausible.

Assumption PAD-1 (stable population) was plausible because some medical centers were geographically or institutionally isolated (Table 1), and it is unlikely that a woman in labor would go to an inconvenient hospital in order to receive epidural analgesia.

Assumption PAD-2 (stable ancillary care) was plausible because most studies reported no changes in obstetric practice other than the increase in availability of epidural analgesia (Table 1).

Assumption PAD-3 (stable disease progression) was plausible because there was no known risk factor altering the time course of C/S for women in labor. However, there is a concern that women not in labor who received C/S, for example following a previous C/S, could have a probability of C/S that changed over time [52].

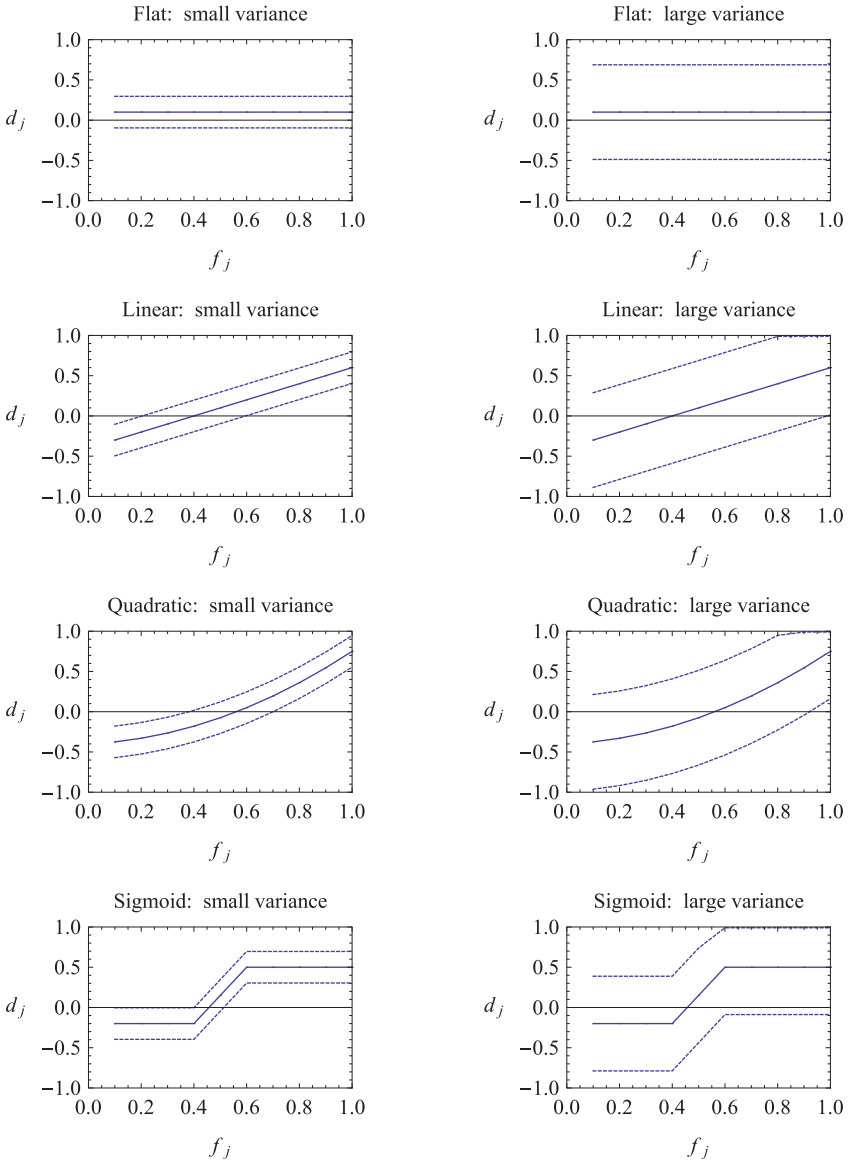

Figure 8 Hypothetical patterns for simulation. Dashed lines are 95% confidence intervals. Each point represents a study.

This concern was mitigated by restricting the population studied to women in labor, women with anticipated vaginal delivery, women with elected C/S excluded, or women without previous C/S (Tables 1 and 2).

Assumption PAD-4 (stable evaluation) was plausible because the eligibility criterion of labor did not change over time and the determination of the C/S outcome is unambiguous.

Assumption PAD-5 (stable treatment effect) says that an always-receiver has the same probability of C/S in both time periods regardless of (i) receiving epidural analgesia after other analgesia in the time period of less availability and (ii) receiving epidural analgesia from the start in the time period of greater availability of epidural analgesia. Support for this assumption comes from randomized trials showing that the timing of epidural initiation does not affect the probability of C/S [53–56].

Assumption PAD-6 (stable preferences) was plausible because there was no new information that would have changed the preference for epidural analgesia.

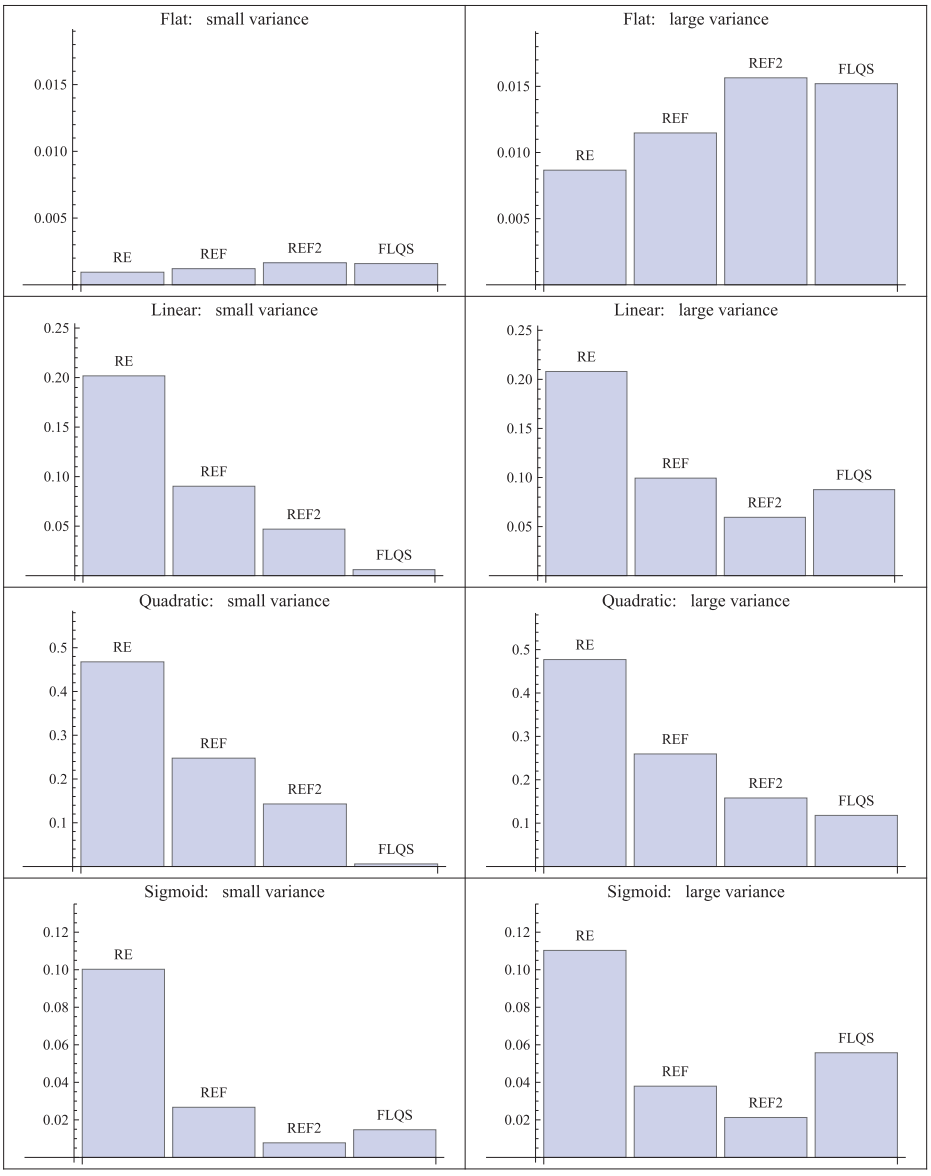

Figure 9 Mean squared errors based on simulation.

To increase the plausibility of **AssumptionPAD-1(stable population)** and **Assumption PAD-2** (**stable ancillary care)**, the analysis was restricted to a total time change over two periods of 6 or fewer years, which meant splitting the data from Dailey (2000) into two groups and dropping one study of duration of 11 years that had been included in Baker and Lindeman [1]. We now turn to the final assumption.

A**ssumption PAD-7 (generalizability)** was made more plausible by using extrapolation estimates. For the FLQS estimate, the maximum fraction who could receive epidural analgesia was set to 0.89 to account for rapid deliveries. The value of 0.89 was based on studies reporting numbers of rapid deliveries [66–68].

Figure 11 shows the estimates of treatment effect and 95% confidence intervals for each study in the paired availability design. Summary estimates and confidence intervals are presented in Section 5, which compares all the estimates.

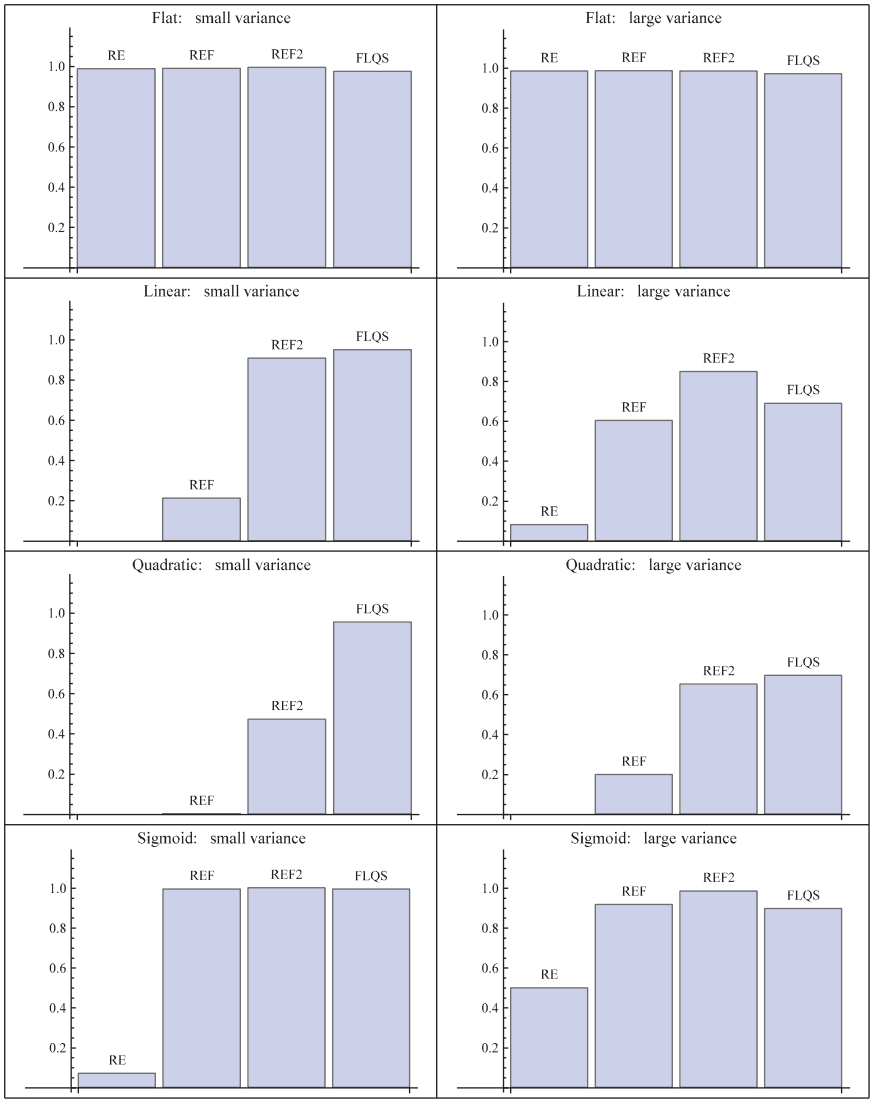

Figure 10 Coverage of 95% confidence intervals based on simulation.

4 Meta-analysis of randomized trials

Randomized trials were also used to study the effect of epidural analgesia on the probability of C/S. A randomized trial has the advantage over the propensity score analysis and the paired availability design of balancing the distribution of both observed and unobserved variables in the two groups [69], thereby obviating some assumptions in the other approaches. Nevertheless, randomized trials are not a perfect gold standard in this application for reasons discussed below.

One concern with randomized trials in this application is a high fraction of crossovers (epidural to non-epidural and vice versa) that makes intent-to-treat estimates less relevant because the effect of receipt of epidural analgesia on the probability of C/S is substantially diluted [52]. A second concern is that an estimate of treatment effect based on only those who comply with treatment protocol in the randomization group can be biased. These concerns can be circumvented by using a similar principal stratification model with two plausible assumptions as that discussed with the paired availability design.

Table 1 Information about studies used in the paired availability design to investigate the effect of epidural analgesia on the probability of C/S.

Study	Source	Type of medical center	Parity	Support for Assumption PAD-3 (stable disease progression)	Support for Assumption PAD-1 (stable population) and Assumption PAD-2 (stable ancillary care)
Gribble and Meier [57]	Article	regional center	mixed with nulliparous subset	"in labor"	"our patient population base has not changed … the same eight obstetricians were involved … no apparent changes in their approaches"
Larsen [58]	Meeting abstract	city hospital	mixed		"same five obstetricians … patient population was similar during each year"
Mancuso [59]	Meeting abstract	army medical center	mixed	"elective Cesarean deliveries excluded"	
Johnson and Rosenfeld [60]	Article	family practice	mixed	no previous C/S	"no other changes in physician care, nursing staff, hospital, physicians involved, or demographics of population"
Newman *et al.* [61]	Meeting abstract	city hospital	mixed		
Lyon *et al.* [62]	Article	Army medical center	nulliparous	"anticipated vaginal delivery"	"no change in patient population demographics"
					"small practice variations could not be excluded because of personnel changes"
Fogel *et al.* [63]	Article	City-county hospital	Mixed (31% nulliparous)		"same group of attending obstetricians and anesthesiologists"
Yancey *et al.* [64]	Article	Army medical center	Mixed (48% nulliparous)	data available from women with no previous C/S	"no dramatic personnel changes"
Dailey (personal communication, 1999)	Personal note	City hospital	Mixed		
Impey *et al.* [29, 30]	Article	City hospital	Nulliparous	"in spontaneous labor"	"consistency of obstetric practice in this group that is almost unparalleled". "confounding variables, such as electronic fetal monitoring were constant"
Zhang *et al.* [65]	Article	Army medical center	Nulliparous	"spontaneous onset of labor"	"no significant personnel change nor any new obstetric protocol implemented"

The following additional notation is introduced. Let $G = 0$ denote the non-epidural analgesia randomization group, and let $G = 1$ denote the epidural analgesia randomization group. Let T0 denote not receiving epidural analgesia and T1 denote receiving epidural analgesia. T0 includes receiving study non-epidural drug, receiving a non-study drug, refusal to receive any pain relief, and rapid delivery. Using the terminology of Angrist *et al.* [42], the principal strata are

Never-taker, if a participant would receive treatment T0 if randomized to either $G = 0$ or $G = 1$.
Complier, if a participant would receive T0 if randomized to $G = 0$ and T1 if randomized to $G = 1$.
Defier, if a participant would receive T1 if randomized to $G = 0$ and T0 if randomized to $G = 1$.
Always-taker, if a participant would receive treatment T1 randomized to $G = 0$ or $G = 1$.

The following assumptions are invoked, where the terms "exclusion restriction" and "monotonicity" come from Angrist *et al.* [42].

Table 2 Basic estimates computed for paired availability design to study the effect of epidural analgesia on the probability of C/S.

Study	Time period	Number	Fraction receiving epidural	Fraction with C/S
Gribble *et al.* [57]	1986–87	1,298	0.000	0.090
	1989–91	1,084	0.480	0.082
Nulliparous subset	1986–87	526	0.000	0.167
Nulliparous subset	1989–91	425	0.610	0.160
Larsen [58]	1989–90	1,919	0.270	0.280
	1990–91	2,073	0.380	0.230
Mancuso [59]	1990–91	4,685	0.190	0.150
	1991–92	4,087	0.670	0.120
Johnson and Rosenfeld [60]	1993	103	0.220	0.180
	1994	116	0.590	0.170
Newman *et al.* [61]	1998	2,628	0.400	0.240
	1989	2,808	0.460	0.240
Time period ↑	1990	2,672	0.550	0.260
Time period ↓	1991	2,486	0.610	0.270
	1992	2,520	0.660	0.290
	1993	2,492	0.710	0.280
	1994	2,420	0.740	0.280
Lyon *et al.* [62]	1992–93	373	0.130	0.110
	1993–94	421	0.590	0.100
Fogel *et al.* [63]	1992–93	3,195	0.012	0.091
	1993–94	3,733	0.290	0.097
Yancey *et al.* [64]	1992–93	4,778	0.008	0.130[a]
	1995–96	4,859	0.590	0.130[a]
Dailey (1999, personal communication)	1989	2,175	0.180	0.210
Time period ↑	1990	2,239	0.270	0.200
Time period ↓	1991	2,115	0.350	0.230
	1992	2,226	0.430	0.220
	1993	2,404	0.400	0.200
	1994	2,476	0.470	0.210
Time period ↑	1995	2,450	0.490	0.210
Time period ↓	1996	2,334	0.520	0.210
	1997	2,320	0.480	0.200
	1998	2,289	0.540	0.220
Impey *et al.* [29, 30] Time period ↑	1987	1,000	0.099	0.038
Time period ↓	1992	1,000	0.450	0.050
	1994	1,000	0.570	0.040
Zhang *et al.* [65]	1993	507	0.100	0.140
	1996	581	0.840	0.120

Note: [a]Excludes C/S after a previous C/S.

Assumption RCT-1 (exclusion restriction): the effect of treatment on the probability of outcome does not change with randomization group among always-takers and never-takers,

Assumption RCT-2 (monotonicity): no participant would receive T1 if randomized to T0 and receive T0 if randomized to T1.

Assumption RCT-3 (generalizability): the estimated effect of receipt of treatment based on compliers is a good estimate of the effect of receipt of treatment among all eligible persons.

Assumption RCT-4 (blinding equivalence): knowledge of treatment received did not affect the probability of outcome, as if the investigators and patients were blinded to the treatment received.

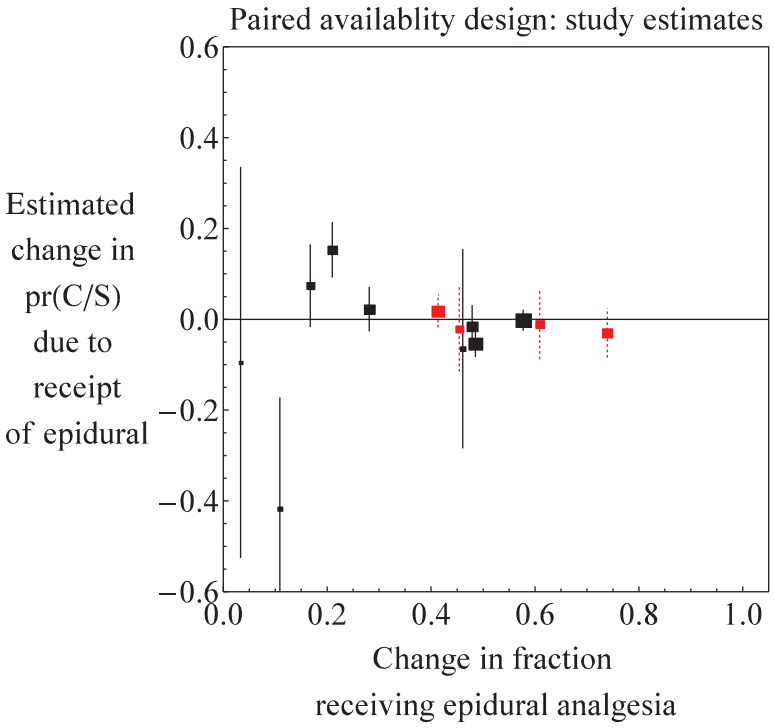

Figure 11 Estimates of effect of receipt of epidural analgesia on probability of C/S for each study in the paired availability design. The rectangle size is proportional to the reciprocal of the estimated variance. Vertical lines correspond to 95% confidence intervals. Red dashed lines correspond to studies involving only nulliparous women; black solid lines correspond to studies involving women of mixed parity.

We now consider the plausibility of these assumptions in the meta-analysis of randomized trials involving the effect of epidural analgesia on the probability of C/S.

Assumption RCT-1(exclusion restriction) says that women who receive epidural analgesia after receiving non-epidural analgesia (and thus classified as receiving epidural analgesia) have the same probability of C/S as women who received only epidural analgesia. Support comes from randomized trials showing that the timing of epidural analgesia does not affect the probability of C/S [53–56].

Assumption RCT-2(monotonicity) is plausible because preference for epidural analgesia does not change with assignment to randomization group.

The combination of **Assumption RCT-1(exclusion restriction)** and **Assumption RCT-2 (monotonicity)** allows estimation of the effect of epidural analgesia on the probability of C/S among compliers, which is analogous to Equation (24) for the paired availability design; this estimate is called the local average treatment effect [42] or the complier average causal effect [70].

Assumption RCT-3 (generalizability) is needed because the composition of principal strata can change in the future. To make this assumption more plausible, extrapolation estimates are used. However, there is also a concern that persons enrolled in a randomized trial are not representative of the general population [52]. The use of multiple randomized trials to broaden the study population mitigates this concern. The estimates for each trial that are used to compute the extrapolation estimate are displayed in Figure 12.

Assumption RCT-4 (blinding equivalence) was plausible because of the use of written protocols for performing a C/S (Tables 3 and 4).

The estimates and confidence intervals based on the meta-analysis of the randomized trials are summarized in Section 5.

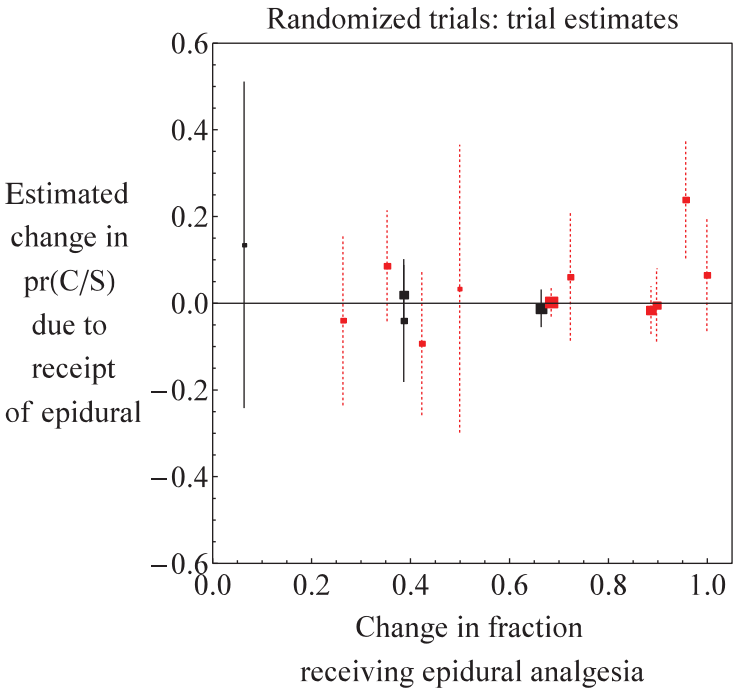

Figure 12 Estimates of the effect of receipt of epidural analgesia on probability of C/S for each trial in the meta-analysis of randomized trials. The rectangle size is proportional to the reciprocal of the estimated variance. Vertical lines correspond to 95% confidence intervals. Red dashed lines correspond to trials involving only or almost only nulliparous women; black solid lines correspond to trials involving women of mixed parity.

Table 3 Information about randomized trials to study the effect of epidural analgesia on the probability of C/S.

Study	Source	Parity	Support for Assumption RCT-4 (blinding equivalence)
Philipsen and Jensen [71]	Article	Nulliparous (93%)	
Thorp *et al.* [72]	Article	Nulliparous	indications for C/S
Muir *et al.* [73]	Meeting abstract	Nulliparous	
Ramin *et al.* [74][a]	Article	Mixed (56% nulliparous)	"all staff followed a written procedural manual"
Bofill *et al.* [75]	Article	Nulliparous	"strict guidelines regarding labor management were in force throughout the duration of the study"
Sharma *et al.* [66]	Article	Mixed (54% nulliparous)	"all staff followed procedures recording in a written manual"
Clark *et al.* [76]	Article	Nulliparous	indications for C/S
Gambling *et al.* [67]	Article	Mixed (53% nulliparous)	"all women were treated using standardized written protocols"
Loughnan *et al.* [77]	Article	Nulliparous	"labour was managed according to a protocol"
Norris *et al.* [78]	Article	Mixed (42% nulliparous)	"obstetric residents and labor nurses unaware of the anesthetic administered, managed labors according to standardized protocols"
Howell *et al.* [79]	Article	Nulliparous	"apart from the choice of analgesia, no attempt was made to influence. practice relating to other aspects of management of labor"
Dickinson *et al.* [80]	Article	Nulliparous	"those monitoring the adherence to the obstetrical protocol were blinded to analgesic treatment group"
Sharma *et al.* [81]	Article	Nulliparous	"all pregnancies were managed … following a written protocol"
Halpern *et al.* [82]	article	Nulliparous	indications for C/S
Sharma *et al.* [68]	Article	Nulliparous	"all pregnancies were managed … following a written protocol"

Note: [a]the reported outcome was operative delivery for dystocia, which is usually C/S.

Table 4 Basic estimates computed from randomized trials to study the effect of epidural analgesia on the probability of C/S.

Study	Randomization group	Number	Fraction receiving epidural	Fraction with C/S
Philipsen and Jensen [71]	Control	54	0.000	0.110
	Epidural	57	1.000	0.180
Thorp *et al.* [72]	Control	45	0.022	0.022
	Epidural	48	0.980	0.250
Muir *et al.* (1996)	Control	22	0.500	0.091
	Epidural	28	0.000	0.110
Ramin *et al.* [74]	Control	666	0.160	0.050
	Epidural	665	0.090	0.059
Bofill *et al.* [75]	Control	51	0.240	0.059
	Epidural	49	0.960	0.100
Sharma *et al.* [66]	Control	357	0.014	0.045
	Epidural	358	0.680	0.036
Clark *et al.* [76]	Control	162	0.520	0.140
	Epidural	156	0.940	0.096
Gambling *et al.* [67]	Control	607	0.260	0.056
	Epidural	616	0.650	0.063
Loughnan *et al.* [77]	Control	310	0.570	0.130
	Epidural	304	0.830	0.120
Howell *et al.* [79]	Control	185	0.280	0.086
	Epidural	184	0.670	0.071
Norris *et al.* [78]	Control	1,071	0.099	0.150
	Epidural	112	9.400	0.130
Dickinson *et al.* [80]	Control	499	0.370	0.140
	Epidural	493	0.720	0.170
Sharma *et al.* [81]	Control	233	0.060	0.086
	Epidural	226	0.950	0.071
Halpern *et al.* [82]	Control	118	0.100	0.100
	Epidural	124	1.000	0.097
Sharma *et al.* [68]	Control	1,364	0.140	0.100
	Epidural	1,339	0.820	0.110

5 Comparison of estimates

Two sets of analyses were conducted: one for nulliparous women and one for women of mixed parity (Figure 13). Generally, no separate data were available for multiparous women. In studies involving women of mixed parity, the reported fraction of nulliparous women ranged from 31 to 56%. With data from only four studies, the FLQS estimate was thought to be too unreliable and not included. For data on nulliparous women in Nguyen *et al.* [33], the estimated effect of epidural analgesia on the probability of C/S was based on Equation (11). For data on women of mixed parity (43% nulliparous) in Nguyen *et al.* [33], the estimated effect of epidural analgesia on the probability of C/S was computed using the equation in Appendix C.

The main result is that confidence intervals for the effect of epidural analgesia on the probability of C/S from the paired availability design and the randomized trials both included zero, in contrast to the confidence intervals from the propensity score analysis which excluded zero.

6 Discussion

In these investigations of the effect of epidural analgesia on the probability of C/S, randomized trials are viewed as the gold standard but with additional assumptions related to crossovers and lack of blinding.

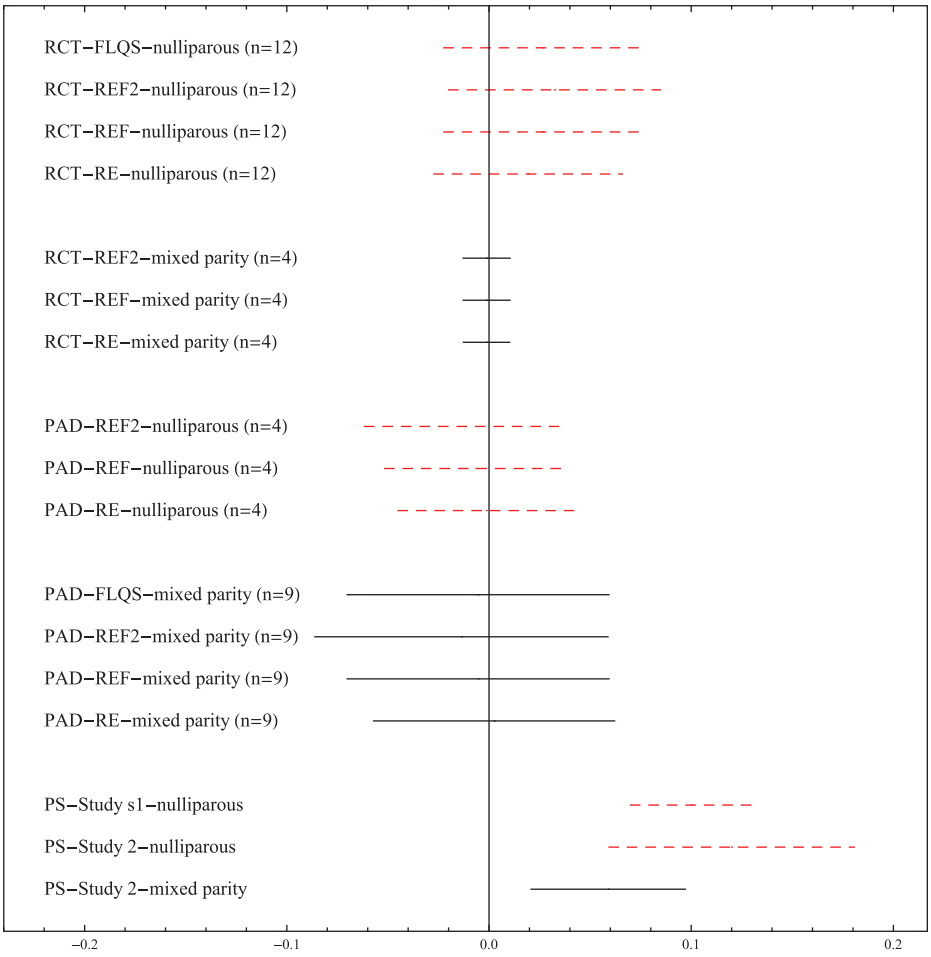

Figure 13 Estimates of the effect of epidural analgesia on the probability of C/S for propensity score (PS), paired availability design (PAD), and randomized clinical trial (RCT). The number of studies is in parentheses. The red (dotted line) denotes studies involving only (or almost only) nulliparous women. For propensity score analyses, Study 1 is Lieberman *et al.* (1995) and Study 2 is Nguyen *et al.* [33].

The meta-analysis of randomized trials and the paired availability design involves similar assumptions related to principal stratification, **Assumption RCT-1 (exclusion restriction)**, A**ssumption RCT-2 (monotonicity)** and their analogs **Assumption PAD-5 (stable treatment effect)** and **Assumption PAD-6 (stable preferences)**. However, it does not follow that the paired availability design will yield similar estimates as the meta-analysis of randomized trials because the paired availability design requires additional assumptions in lieu of randomization. Both the propensity score method and the paired availability design will yield correct estimates if their assumptions hold and likely incorrect estimates if their assumptions do not hold. Therefore, when performing these analyses, it is a critical look at the plausibility of the assumptions.

Based on the theory of causal graphs, the propensity score method requires two assumptions, **Assumption PS-1 (no omitted confounder)** and **Assumption PS-2 (no latent M-bias collider)**. Unfortunately, it is easy to be overconfident that these assumptions hold. For example in the application discussed here, Lieberman *et al.* [28] wrote "for some other factor not controlled in our analysis to be responsible for the association we have noted, it would have to be very strongly associated with epidural analgesia use and cesarean delivery. There are no obvious candidates apart from the factors we have measured." We believe that omitted confounders of intense pain and obstetrical practice and a latent M-bias

colliders of cervical dilation at admission and initial rate of cervical dilation could have biased results from the propensity score analysis.

The paired availability design requires seven assumptions. Four assumptions ensure that time periods can be treated like arms of a randomized trial for drawing conclusions: **Assumption PAD-1 (stable population)**, **Assumption PAD-2 (stable ancillary care)**, **Assumption PAD-3 (stable disease progression)**, **Assumption PAD-4 (stable evaluation)**. Some design choices can make these assumptions more plausible: short time periods, institutionally or geographically isolated clinics, restrictions to only health care providers present in both time periods, and eligibility criteria. The two assumptions needed for model identifiability, **Assumption PAD-5 (stable treatment effect)** and **Assumption PAD-6 (stable preferences)**, are often reasonable. Perhaps, the least appreciated assumption is **Assumption PAD-7 (generalizability)**. The extrapolation estimate increases the plausibility of this assumption; it is most informative when the estimated fraction of women who are consistent-receivers (under fixed availability) or consistent-receivers and inconsistent-receivers (under random availability) is large for some medical centers; otherwise there could be considerable uncertainty in extrapolation. In this regard, a sensitivity analysis is recommended.

Whenever possible, randomized trials instead of observational studies should be used to estimate the causal effect of treatment on outcome because fewer assumptions are required. However, when randomized trials cannot be implemented due to ethical considerations or expense, propensity scores informed by causal graphs and the paired availability design with the extrapolation estimate should be considered along with a critical appraisal of the plausibility of their assumptions.

Acknowledgements: This work was supported by the National Institutes of Health. SGB thanks Judea Pearl for help with causal graphs and Ulrich Abel for an invitation to a 1997 conference on nonrandomized studies at the German Cancer Research Center, where an excellent talk by Donald Rubin sparked an interest in propensity scores. This article can be viewed as an update of the propensity score analysis and paired availability design discussed at that conference. The authors also thank reviewers for helpful comments.

Appendix A

This Appendix reviews the proof in Rosenbaum and Rubin [3] for an important property of propensity scores. Recall that the propensity score is $s = \text{pr}(T = 1\,|\text{x})$. Let $E()$ denote the expected value, so

$$E(t\,|\text{x}) \;=\; 1 \times \text{pr}(T \;=\; 1\,|\text{x}) + 0 \times \text{pr}(T \;=\; 1\,|\text{x}) = \text{s}. \qquad [\text{A.0}]$$

Also,

$$\begin{aligned} \text{pr}(T \;=\; 1\,|s) &= 1 \times \text{pr}(T \;=\; 1|s) + 0 \times \text{pr}(T \;=\; 1\,|s) = E(t|s) \\ &= \; E\{E(t\,|\text{x})\,|s\}, \text{by a mathematical identity}, \\ &= E\{s\,|s\}, \text{from (A.0)}, \\ &= \; s \;=\; \text{pr}(T \;=\; 1\,|\text{x}). \end{aligned} \qquad [\text{A.1}]$$

Because T is binary, from (A.1),

$$\text{pr}(T \;=\; t\,|s) = \text{pr}(T \;=\; t\,|\text{x}). \qquad [\text{A.2}]$$

In addition,

$$\begin{aligned} &\text{pr}(X \;=\; x,\; T \;=\; t\,|s) \\ &= \text{pr}(T \;=\; t\,|\text{x}, \text{s})\,\text{pr}(X \;=\; x\,|s), \text{bya mathematical identity}, \\ &= \text{pr}(T \;=\; t\,|\text{x})\,\text{pr}(X \;=\; x\,|s), \text{because } x \text{ includes } s, \\ &= \text{pr}(\text{T} \;=\; t\,|s)\,\text{pr}(X \;=\; \text{x}\,|\text{s}), \text{from (A.2)}. \end{aligned} \qquad [\text{A.3}]$$

Dividing both sides of (A.3) by pr($T = t$ |s) gives pr($X = x \mid t, s$) = pr($X = x$ |s), which is the key property from Equation (8).

Appendix B

This Appendix proves the fundamental results concerning independence and dependence of variables directly connected to non-colliders and colliders.

B.1 Conditioning on non-collider *X* where *A* ← Z ← *B*

Let pr($A = a|x$) = f($a|x$), pr($X = x|b$) = g($x|b$), pr($B = b$) = h(b). Therefore pr($A = a$,$x = x$, $B = b$) = f($a|x$) g($x|b$) h(b). Variables A and B are unconditionally dependent namely,

$$\begin{aligned}&\mathrm{pr}(A = a, B = b) \neq \mathrm{pr}(A = a)\ \mathrm{pr}(B = b), \text{because}\\ &\mathrm{pr}(A = a, B = b)\ \Sigma_x\ \mathrm{f}(a|x)\ g(x|b)\ h(b),\\ &\mathrm{pr}(A = a) = \Sigma_x\ \mathrm{f}(a|x)\Sigma_b \mathrm{g}(x|b)\ h(b),\\ &\mathrm{pr}(B = b) = \Sigma_x\ g(x|b)\ h(p).\end{aligned} \tag{B.1}$$

Also variables A and B are independent conditional on x, namely pr($A = a$, $B = b$|x) = pr($A = a\ x$) pr($B = b|x$), because

$$\begin{aligned}&\mathrm{pr}(A = a, B = b|x) = \mathrm{f}(a|x)\ \mathrm{g}(x|b)\ \mathrm{h}(b)/\{\Sigma_a\Sigma_b\ \mathrm{f}(a|x)\ \mathrm{g}(x|b)\ \mathrm{h}(b)\}\\ &= \mathrm{f}(a|x)\ \mathrm{g}(x|b)/\{\Sigma_a\ \mathrm{f}(a|x)\Sigma_\mathrm{b}\ \mathrm{g}(x|b)\},\\ &\mathrm{pr}(A = a|x) = \Sigma_b\ \mathrm{f}(a|x)\ \mathrm{g}(x|b)\ \mathrm{h}(b)/\{\Sigma_a\Sigma_b\ \mathrm{f}(a|x)(x|b)\ \mathrm{h}(b)\}\\ &= \mathrm{f}(a|x)/\Sigma_a\ \mathrm{f}(a|x),\\ &\mathrm{pr}(B = b|x) = \Sigma_a\ \mathrm{f}(a|x)\ \mathrm{g}(x|b)\ h(b)/\{\Sigma_a\Sigma_b\ \mathrm{f}(a|x)\ \mathrm{g}(x|b)\ \mathrm{h}(b)\}\\ &= \mathrm{g}(x|b)/\Sigma_b\ \mathrm{g}(x|b).\end{aligned} \tag{B.2}$$

B.2 Conditioning on non-collider *X* where *A* → Z → *B*

Let pr($A = a$) = f(a), pr($X = x|a$) = g($x|a$), and pr($B = b|x$) = h(b|x), so pr($A = a$,$x = x$, $B = b$) = f(a) g($x|a$) h(b|x). Variables A and B are unconditionally dependent namely,

$$\begin{aligned}&\mathrm{pr}(A = a, B = b) \neq \mathrm{pr}(A = a)\ \mathrm{pr}(B = b)\ \text{because}\\ &\mathrm{pr}(\mathrm{A} = \mathrm{a}, \mathrm{B} = \mathrm{b}) = \Sigma_x\ \mathrm{f}(a)\ \mathrm{g}(x|a)\ \mathrm{h}(b|\mathrm{x}),\\ &\mathrm{pr}(A = a) = \mathrm{f}(a),\\ &\mathrm{pr}(B = b) = \Sigma_x\Sigma_b\ \mathrm{f}(a)\ \mathrm{g}(x|a)\ \mathrm{h}(b|\mathrm{x}).\end{aligned} \tag{B.3}$$

Also variables A and B are independent conditional on x, namely, pr($A = a$, $B = b \mid x$) = pr($A = a|\ x$) pr($B = b|x$), because

$$\begin{aligned}&\mathrm{pr}(A = a, B = b|x) = \mathrm{f}(a)\ \mathrm{g}(x|a)\ \mathrm{h}(b|x)/\{\Sigma_a\Sigma_b\ f(a)\ g(x|a)\ h(b|x)\},\\ &\mathrm{pr}(A = a|x) = \Sigma_b\ \mathrm{f}(a)\ g(x|a)\ \mathrm{h}(b|x)/\{\Sigma_a\Sigma_b\ \mathrm{f}(a)\ \mathrm{g}(x|a)\ \mathrm{h}(b|x)\}\\ &= \mathrm{f}(a)\ \mathrm{g}(a, x) = \{\Sigma_a\ \mathrm{f}(a)\ \mathrm{g}(a, x)\},\\ &\mathrm{pr}(B = b|\mathrm{x}) = \Sigma_a\ \mathrm{f}(a)\ g(x|a)\ \mathrm{h}(b|x)/\{\Sigma_a\Sigma_b\ \mathrm{f}(a)\ \mathrm{g}(x|a)\ \mathrm{h}(b|x)\}\\ &= \mathrm{h}(b|x)/\Sigma_a\ \mathrm{h}(b|\mathrm{x}).\end{aligned} \tag{B.4}$$

B.3 Conditioning on non-collider *X* where *A* ← *X*→ *B*

Let pr($A|X$) = f($a|x$), pr($X = x$) = g(x), and pr($B = b|x$) = h($b|x$), so pr($A = a$, $X = x$, $B = b$) = f($a|x$) g(x) h($b|x$). Variables A and B are unconditionally dependent, namely,

$$\begin{aligned}&\text{pr}(A = a, B = b) \neq \text{pr}(A = a)\ \text{pr}(B = b)\text{because}\\&\text{pr}(A = a, B = b) = \Sigma_x\ \text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x),\\&\text{pr}(A = a) = \Sigma_x\ \text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x),\\&\text{pr}(B = b) = \Sigma_x\ \text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x).\end{aligned} \quad [B.5]$$

Also variables A and B are independent conditional on x, namely

$$\begin{aligned}&\text{pr}(A = a, B = b|x) = \text{pr}(A = a|x)\ \text{pr}(B = b|x), \text{because}\\&\text{pr}(A = a, B = b|x) = \text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x)/\{\Sigma_a\Sigma_c\, f(a|x)\ \text{g}(x)\ \text{h}(b|x)\}\\&= \text{f}(a/x)\ \text{h}(b|x)/\{\Sigma_a\ \text{f}(a|x)\Sigma_b\ \text{h}(b|x)\},\\&\text{pr}(A = a|x) = \Sigma_b\ \text{f}(a, x)\ \text{g}(x)\ \text{h}(b|x)/\{\Sigma_a\ \Sigma_b\ \text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x)\}\\&= \text{f}(a|x)/\Sigma_a\ \text{f}(a|x),\\&\text{pr}(B = b|x) = \Sigma_a\ \text{f}\,\text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x)/\{\Sigma_a\Sigma_b\,\text{f}(a|x)\ \text{g}(x)\ \text{h}(b|x)\}\\&= \text{h}(b|x)/\Sigma_b\ \text{h}(b|x).\end{aligned} \quad [B.6]$$

B.4 Conditioning on collider *X* where *A* → *Z* ← *B*

Let pr($X = x|\ a,\ b$) = f($x|a,b$), pr($A = a$) = g(a), and pr($B = b$) = h(b), so pr($A = a$, $X = x$, $B = b$) = f($x|a,b$) g(a) h(b). Variables A and B are unconditionally in dependent, namely,

$$\begin{aligned}&\text{pr}(A = a, B = b) = \text{pr}(A = a)\ \text{pr}(B = b)\text{because}\\&\text{pr}(A = a, B = b) = \Sigma_x\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b) = g(a)\ \text{h}(b).\\&\text{pr}(A = a) = \text{g}(a),\\&\text{pr}(B = b) = \text{h}(b).\end{aligned} \quad [B.7]$$

Also variables A and B are dependent conditional on x, namely

$$\begin{aligned}&\text{pr}(A = a, B = b|x) \neq \text{pr}(A = a|x)\ \text{pr}(B = b|x), \text{ because}\\&\text{pr}(A = a, B = b|x) = \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)/\{\Sigma_a\Sigma_{\text{b}}\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)\},\\&\text{pr}(A = a|x) = \Sigma_b\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)/\{\Sigma_a\Sigma_b\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)\},\\&\text{pr}(B = b|x) = \Sigma_a\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)/\{\Sigma_a\Sigma_b\ \text{f}(x|a, b)\ \text{g}(a)\ \text{h}(b)\}.\end{aligned} \quad [B.8]$$

Appendix C

This Appendix describes estimation of effect of epidural analgesia on the probability of C/S using data from propensity score quantiles in Nguyen *et al.* [33] which was separately provided for nulliparous women ($w = 0$) and multiparous women ($w = 1$). Let p(t, q, w) denote an estimate of pr($Y = 1|t,s^*, w$) where s^* is the estimated propensity score s^* in the qth quintile of the estimated propensity scores. Let m denote the fraction of women who are multiparous. Then the propensity score subclassification estimate of causal effect for women of mixed parity is

$$\begin{aligned}D_{\text{CAUSAL(PS)}} = &\{\ \Sigma_{q=1}^5\,\text{p}(1,\ q,\ 0)/5\ -\ \Sigma_{q=1}^5\ \text{p}(0,\ q,\ 0)/5\}\ (1 - m)\\&+\{\Sigma_{q=1}^5\ \text{p}(1,\ q,\ 1)/5\ -\ \Sigma_{q=1}^5\ \text{p}(0\,q,\ 1)/5\}\,m.\end{aligned}$$

References

1. Baker SG, Lindeman KS. Rethinking historical controls. Biostatistics 2001;2:383–96.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–25.
4. Baker SG, Lindeman KS. The paired availability design: a proposal for evaluating epidural analgesia during labor. Stat Med 1994;13:2269–78.
5. Pearl J. Letter to the editor: remarks on the method of propensity scores. Stat Med 2009;28:1420–3.
6. Shrier D. Letter to the editor. Stat Med 2008;27:2740–1.
7. Shrier D. Letter to the editor: propensity scores. Stat Med 2009;28:1317–8.
8. Sjolander A. Letter to the editor: propensity scores and m-structures. Stat Med 2009;28:1416–20.
9. Pearl J. Causality: models reasoning and inference. Cambridge: Cambridge University Press, 2009b.
10. Pearl J. An introduction to causal inference. Int J Biostat 2010;6:7.
11. Rubin DB. Author's reply (to Ian Shrier's Letter to the Editor). Stat Med 2008;27:2741–2.
12. Rubin DB. Author's reply: should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat Med 2009;28:1420–3.
13. Baker SG. Estimation and inference for the causal effect of receiving treatment on a multinomial outcome: an alternative approach. Biometrics 2011;67:319–25.
14. Frangakis CE, Rubin DB. Principle stratification in causal inference. Biometrics 2002;58:21–9.
15. VanderWeele TJ. Principal stratification – uses and limitations. Int J Biostat 2011;7:28.
16. Pearl J. Principal stratification – a goal or a tool? Int J Biost 2011;7:20.
17. Rubin DB. Estimation from nonrandomized treatment comparisons using subclassification on propensity scores. In: Abel U, Koch A, editors. Nonrandomized comparative clinical studies. Dusseldorf: Symposion Publishing, 1998:85–100.
18. Rubin DB. Propensity score methods. Am J Ophthalmol 2010;149:7–9.
19. Jeon JW, Chung HY, Bae JS. Chances of Simpson's paradox. J Korean Stat Soc 1987;16:117–25.
20. Baker SG, Kramer BS. Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. J Women's Health Gender-Based Med 2001;10:867–72.
21. Wainer H. The BK-plot: making Simpson's paradox clear to the masses. Chance 2002;15:60–2.
22. Hullsiek KH, LouisTA. Propensity score modeling strategies for the causal analysis of observational data. Biometrics 2002;2:179–93.
23. Lunceford JK, Davidan M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937–60.
24. D'Agostino RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
25. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734–53.
26. Breslow NE, Day NE. Statistical methods in cancer research, Vol. 1: the analysis of case-control studies. Lyon: The International Agency for Research on Cancer, 1980.
27. Morgan SL, Winship C. Counterfactuals and causal inference: Methods and principles for social research. Cambridge: Cambridge University Press, 2007.
28. Lieberman E, Lang J, Cohen A, D'Agostino R, Datta S, Frigoletto F. Association of epidural analgesia with cesarean delivery in nulliparas. Obstet Gynecol 1996;88:993–1000.
29. Impey L, Hobson J, O'Herlihy C. Graphic analysis of actively managed labor: prospective computation of labor progress in 500 consecutive nulliparous women in spontaneous labor at term. Am J Obstet Gynecol 2000;183:438–43.
30. Impey L, MacQuillian K, Robson M. Epidural analgesia need not increase operative delivery rates. Am J Obstet Gynecol 2000;182:358–63.
31. Beilin Y, Freidman F Jr, Andres LA, Hossain S, Bodian CA. The effect of obstetrician group and epidural analgesia on the risk for cesarean delivery in nulliparous women. Acta Anaesthesiol Scand 2000;44:959–64.
32. Frigoletto FD Jr, Lieberman E, Lang JM, Cohen A, Barss V, Ringer S, et al. A clinical trial of active management of labor. New Engl J Med 1995;333:745–50.
33. Nguyen UDT, Rothman KJ, Demissie S, Jackson DJ, Lang JM, Ecker JL. Epidural analgesia and risks of cesarean and operative vaginal deliveries in nulliparous and multiparous women. Matern Child Health J 2010;14:705–12.
34. Brookhart MA, Schneeweis S, Rothman K. Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.
35. Pearl J. On a class of bias-amplifying covariates that endanger effect estimates. In: Grunwald P, Spirtes P, editors. Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence. Corvallis, OR: AUAI, 2010:417–24.

36. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics 2011;67:1406–13.
37. Baker SG. The paired availability design: an update. In: Abel U, Koch A, editors. Nonrandomized comparative clinical studies. Dusseldorf: Medinform-Verlag, 1998:79–84.
38. Baker SG, Kramer BS, Lindeman KS. The paired availability design: if you can't randomize, perhaps this applies. Chance 2006;19:57–60.
39. Baker SG, Lindeman KS, Kramer BS. The paired availability design for historical controls. BMC Med Res Methodol 2001;1:9.
40. Permutt T, Hebel R. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. Biometrics 1989;45:619–22.
41. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica 1994;62:467–75.
42. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;92:444–55.
43. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. Stat Med 1997;16:1017–29.
44. Cox DR. Discussion. Stat Med 1998;17:387–9.
45. Gardner M. Martin Gardner's sixth book of mathematical games from Scientific American. San Francisco, CA: W.H. Freeman and Company, 1971:154.
46. Maor E. Trigonometric delights. Princeton, NJ: Princeton University Press, 1998:122–3.
47. Cheng J. Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. Biometrics 2009;65:96–103.
48. DerSimonian R, Laird N. Meta-analysis in clinical trials. Cont Clin Trials 1986;7:177–88.
49. Follman DA, Proschan MA. Valid inference in random effects meta-analysis. Biometrics 1999;55:732–7.
50. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Stat Med 2002;31:3153–9.
51. Akaike H. Likelihood of a model and information criteria. J Econometrics 1981;16:3–14.
52. Lieberman E. Epidemiology of epidural analgesia and cesarean delivery. Clin Obstet Gynecol 2004;47:317–31.
53. Chestnut DH, Mcgrath JM, Vincent RD, Penning DH, Choi WW, Bates JN, et al. Does early administration of epidural analgesia affect obstetric outcome in nulliparous women who are in spontaneous labor? Anesthesiology 1994;80:1201–8.
54. Ohel G, Gonen R, Vaida S, Barak S, Gaitini L. Early versus late initiation of epidural analgesia in labor: does it increase the risk of cesarean section? A randomized trial. Am J Obstet Gynecol 2006;194:600–5.
55. Wong CA, Scavone BM, Peaceman AM, McCarthy RJ, Sullivan JT, Diaz NT, et al. The risk of cesarean delivery with neuraxial analgesia given early versus late in labor. New Engl J Med 2005;352:655–65.
56. Wang F, Shen X, Guo X, Peng Y, Gu X, The Labor Examining Group. Epidural analgesia in the latent phase of labor and the risk of cesarean delivery. Anesthesiology 2009;111:871–80.
57. Gribble RK, Meier PR. Effect of epidural analgesia on the primary Cesarean rate. Obstet Gynecol 1991;78:231–4.
58. Larsen DD. The effect of initiating an obstetric anesthesiology service on rate of Cesarean section and rate of forceps delivery. Abstract presented at the annual meeting of the Society of Obstetric Anesthesia and Perinatology, 1992.
59. Mancuso JJ. Epidural analgesia in an army medical center: impact on Cesareans and instrumental deliveries. Abstract 13, presented at the annual meeting of the Society for Obstetric Anesthesiology and Perinatology, Palm Springs, 1993.
60. Johnson S, Rosenfeld JA. The effect of epidural anesthesia on the length of labor. J Fam Pract 1995;40:244–7.
61. Newman LM, Perez EC, Krolick TJ, Ivankovich AD. Labor analgesia, Cesarean anesthesia, and cesarean delivery rates for 18,000 deliveries from 1988 through 1994. Anesthesiology 1995;83:3A (Abstract A967).
62. Lyon D, Knuckles G, Whitaker E, Salgado S. The effect of instituting an elective labor epidural program on the operative delivery rate. Obstet Gynecol 1997;90:135–41.
63. Fogel S, Shyken JM, Leighton BL, Mormol JS, Smeltzer J. Epidural labor analgesia and the incidence of cesarean delivery for dystocia. Anesth Analg 1998;87:119–23.
64. Yancey MK, Pierce B, Schweitzer D, Daniels D. Observations on labor epidural analgesia and operative delivery rates. Am J Obstet Gynecol 1999;180:353–9.
65. Zhang J, Yancey MK, Klebanoff JS, Schweitzer D. Am J Obstet Gynecol 2001;185:128–34.
66. Sharma SK, Sidawi JE, Ramin SM, Lucas MJ, Levno KJ, Cunningham G. Cesarean delivery: a randomized trial of epidural versus patient-controlled meperidine analgesia during labor. Anesthesiology 1997;87:487–94.
67. Gambling DR, Sharma SK, Ramin SM, Lucas MJ, Leveno KJ, WileyJ, et al. A randomized study of combined spinal-epidural analgesia versus intravenous meperidine during labor. Anesthesiology 1998;89:1336–44.
68. Sharma SK, McIntire DD, Wiley J, Leveno KJ. An individual patient meta-analysis of nulliparous women. Anesthesiology 2004;100:142–8.
69. Meier P. Statistics and medical experimentation. Biometrics 1975;31:511–52.
70. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments. Ann Stat 1997;25:305–27.
71. Philipsen T, Jensen N. Epidural block or parenteral pethidine as analgesia in labour; a randomized study concerning progress in labour and instrumental deliveries. Euro J Obstet Gynecol Reprod Biol 1989;30:27–33.
72. Thorp JA, Hu DH, Albin RM, Mcnitt J, Meyer BA, Cohen G, et al. The effect of intrapartum epidural analgesia on nulliparous labor: a randomized, controlled, prospective trial. Am J Obstet Gynecol 1993;169:851–8.

73. Muir HA, Shulkla R, Liston R, Writer D. Randomized trial of labor analgesia: a pilot study to compare patient-controlled intravenous analgesia with patient-controlled epidural analgesia to determine if analgesic method affects delivery outcome. Can J Anaesth 1996;43:A60 (Abstract).
74. Ramin SM, Gambling DR, Lucas MJ, Sharma SK, Sidawi E, Leveno KJ. Randomized trial of epidural versus intravenous analgesia during labor. Obstet Gynecol 1995;86:783–9.
75. Bofill JA, Vincent RD, Ross EL, Martin RW, Norman PF, Werhan CF, et al. Nulliparous active labor, epidural analgesia, and cesarean delivery for dystocia. Am J Obstet Gynecol 1997;177:1462–70.
76. Clark A, Carr D, Loyd G, Cook V, Spinnato J. The influence of epidural analgesia on cesarean delivery rates: a randomized, prospective clinical trial. Am J Obstet Gynecol 1998;179:1527–33.
77. Loughnan RA, Carli F, Romney M, Dore CJ, Gordon H. Randomized controlled comparison of epidural bupivicaine versus pethidine for analgesia in labor. Br J Anesth 2000;84:715–9.
78. Norris MC, Fogel ST, Conway-Long C. Combined spinal-epidural versus epidural labor analgesia. Anesthesiology 2001;95:913–20.
79. Howell CJ, Kidd C, Robers W, Upton P, Lucking L, Jones PW, et al. A randomized controlled trial of epidural compared with non-epidural analgesia in labour. Br J Obstet Gynecol 2001;108:27–33.
80. Dickinson E, Paech MJ, McDonald SJ, Evans SF. The impact of intrapartum analgesia on labour and delivery outcomes in nulliparous women. Aust N Z J Obstet Gynecol 2002;42:65–72.
81. Sharma SK, Alexancer JM, Messick G, Bloom S L, McIntire DD, Wiley J, et al. A randomized trial of epidural versus intravenous meperidine analgesia during labor in nulliparous women. Anesthesiology 2002;96:546–51.
82. Halpern SH, Muir H, Breen TW, Campbell DC, Barrett J, Liston R, et al. A multicenter randomized controlled trial comparing patient-controlled epidural with intravenous analgesia for pain relief in labor. Obstet Anesth 2004;99:1532–8.
83. Frolich MA, Orth V, Knitza R, Finsterer U, Hepp H, Peter K. Does epidural analgesia reduce the incidence of operative delivery? Abstract presented at the annual meeting of the Society of Obstetric Anesthesiology and Perinatology in Hamilton, Bermuda, 1997.