The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 25

Clarifying the Role of Principal Stratification in the Paired Availability Design

Stuart G. Baker, *National Institutes of Health*

Karen S. Lindeman, *Johns Hopkins Medical Institutions*

Barnett S. Kramer, *National Institutes of Health*

Recommended Citation:

Baker, Stuart G.; Lindeman, Karen S.; and Kramer, Barnett S. (2011) "Clarifying the Role of Principal Stratification in the Paired Availability Design," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 25.

DOI: 10.2202/1557-4679.1338

Available at: <http://www.bepress.com/ijb/vol7/iss1/25>

©2011 Berkeley Electronic Press. All rights reserved.

Clarifying the Role of Principal Stratification in the Paired Availability Design

Stuart G. Baker, Karen S. Lindeman, and Barnett S. Kramer

Abstract

The paired availability design for historical controls postulated four classes corresponding to the treatment (old or new) a participant would receive if arrival occurred during either of two time periods associated with different availabilities of treatment. These classes were later extended to other settings and called principal strata. Judea Pearl asks if principal stratification is a goal or a tool and lists four interpretations of principal stratification. In the case of the paired availability design, principal stratification is a tool that falls squarely into Pearl's interpretation of principal stratification as "an approximation to research questions concerning population averages." We describe the paired availability design and the important role played by principal stratification in estimating the effect of receipt of treatment in a population using data on changes in availability of treatment. We discuss the assumptions and their plausibility. We also introduce the extrapolated estimate to make the generalizability assumption more plausible. By showing why the assumptions are plausible we show why the paired availability design, which includes principal stratification as a key component, is useful for estimating the effect of receipt of treatment in a population. Thus, for our application, we answer Pearl's challenge to clearly demonstrate the value of principal stratification.

KEYWORDS: principal stratification, causal inference, paired availability design

1. Introduction

Judea Pearl asks if principal stratification (Frangakis and Rubin, 2002) is a tool or a goal (Pearl, 2011). In the paired availability design for historical controls (Baker and Lindeman, 1994), which involves a type of principal stratification as a key component, principal stratification is a tool to achieve the goal of estimating the effect of receipt of treatment in a population. Because many readers may not be familiar with the paired availability design, we describe it in some detail. Special emphasis is placed on assumptions for using historical controls, estimating treatment effect in the principal strata of interest, and generalizing the estimate from the principal strata of interest to the general population. For the latter, we propose a new estimate that increases the plausibility of generalizing to the population. We also discuss the related use of principal stratification with all-or-none compliance. With this background, we address Judea Pearl's critique of the value of principal stratification as it relates to the paired availability design.

2. Overview of paired availability design

The paired availability design uses historical controls to estimate the effect of receipt of a new treatment if all persons received the new treatment instead of the old treatment. In terms of statistical inference, an ideal study would randomize subjects to receive either new or old treatment. However sometimes randomization is not feasible or desirable, such as when there are strongly held views about the merits of treatment or when blinding of patients to treatments is not feasible. In this situation the paired availability design can play an important role in estimating the effect of receipt of treatment.

The standard form of the paired availability design uses data from two time periods in each of many medical centers providing data (Baker and Lindeman, 1994, Baker et al., 2001). A modified form can also use data from more than two time periods in any particular medical center (Baker and Lindeman, 2001). By using multiple medical centers instead of a single medical center, systematic bias is reduced. The analysis is complicated because the change in availability of treatment between time periods differs among medical centers.

In order to estimate the effect of receipt of treatment, as opposed to the effect of a change in availability of treatment, Baker and Lindeman (1994) proposed a four-category potential outcomes model for receipt of treatment if arrival would have occurred in either period. Their model involved reasonable assumptions for estimation. Baker and Lindeman (1994) also proposed a likelihood formulation. This type of model and the same plausible assumptions

for estimation had also been independently proposed by various investigators in the context of all-or-none compliance in randomized trials. Permutt and Hebel (1989) proposed a version without an explicit mathematical formulation. Imbens and Angrist (1994) followed by Angrist et al. (1996) proposed a version based on instrumental variables. Cuzick et al. (1997) proposed a version in the context of cancer screening trials. Frangakis and Rubin (2002) extended this model to other settings and called it principal stratification. If reasonable assumptions hold, the principal stratification model in Baker and Lindeman (1994) yields an unbiased estimate of the effect of receipt of treatment among subjects in some principal strata. An additional assumption is needed to appropriately apply this estimate to all eligible persons.

We discuss the paired availability design and the role of principal stratification in the context of the original example related to obstetric anesthesiology (Baker and Lindeman 1994, 2001). Participants were women in labor arriving in one or more time periods at various medical centers. The goal was to estimate the effect of receiving versus not receiving epidural analgesia on the probability of a Cesarean section (C/S). The paired availability design relies on three types of assumptions, which we discuss in turn: (1) assumptions needed to analyze data from different time periods as data from different randomization groups (2) assumptions needed to estimate the effect of treatment in some principal strata, and (3) an assumption that the estimated treatment effect in some principal strata is a good estimate of the treatment effect for all eligible persons in the entire population. We modify some of the assumptions listed in Baker et al. (2001) and Baker and Lindeman (2001).

3. Assumptions to analyze time periods as randomization groups

The paired availability design requires the following four assumptions that justify analyzing data from the two time periods as if they were data from two randomization groups.

Assumption 1. *Stable Ancillary Care.* Between the two time periods, 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 2. *Stable Disease Natural History.* Between the two time periods, 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 3. *Stable Population.* Between the two time periods, there are no changes in the characteristics of the eligible population that would affect the probability of outcome.

Assumption 4. Stable Evaluation. Eligibility criteria and definitions of outcome are constant over time.

In the application to obstetric anesthesiology, Assumption 1 says that, between the two time periods, there are no systematic changes in obstetric practice unrelated to epidural analgesia that could affect the probability of C/S. Even changes in billing or reimbursement rules can have an effect on the validity of this assumption. Assumption 1 is plausible because medical centers were situated in various geographic locations and data collection took place at various times. If data were available from additional medical centers with no change in availability of epidural analgesia, investigators could estimate the change in the probability of C/S due to changes in care unrelated to the change in availability of epidural analgesia. Then Assumption 1 would say that this estimate is sufficient to adjust for any possible bias due to systematic changes in care.

We cannot think of any examples where Assumption 2 would be seriously questioned in the application involving obstetric anesthesiology. In other settings, Assumption 2 could be violated by an increase or decrease in prevalence of resistant bacteria between the two time periods.

In the application to obstetric anesthesiology, Assumption 3 says that, between the two time periods, there are no changes in the characteristics of the eligible population that would affect the probability of C/S. Assumption 3 is plausible because medical centers were restricted to those serving a closed population, such as an army medical center or the only hospital in a geographic region. In other words, Assumption 3 is plausible because it was unlikely that a woman in labor would go to a, considerably less convenient, hospital in order to receive epidural analgesia.

In the application to obstetric anesthesiology, Assumption 4 was plausible because there was no change between the two time periods in the eligibility criterion of being in labor and the determination of the outcome of Cesarean section. In contrast, in an application in oncology where eligibility is determined by stage of cancer, the use of a new or more sensitive radiologic test to stage cancer in the second time period may artifactually improve prognosis of each stage even if the treatment in the two time periods was the same (Feinstein et al. 1985). Also in the field of oncology, the definition of an outcome of disease progression can change over time, for example with the increasing use "biochemical failure" rather than symptoms of recurrence, such as bone pain.

4. Assumptions to estimate treatment effect for some principal strata

The motivation for the principal stratification model is to estimate the effect of receipt of treatment and not the effect of availability of treatment. To this end, Baker and Lindeman (1994, 2001) proposed the following four principal strata,

- (1) always-receivers, who would receive epidural analgesia in either time period,
- (2) consistent-receivers, who would not receive epidural analgesia in the time period with less availability and would receive it in the time period with greater availability,
- (3) inconsistent-receivers, who would receive epidural analgesia in the time period with less availability and would not receive it in the time period with greater availability,
- (4) never-receivers, who would not receive epidural analgesia in either time period.

Without assumptions, it is not possible to uniquely estimate parameters associated with all outcomes and principal strata. Fortunately, by making plausible assumptions, it is possible to estimate parameters related to the effect of treatment in principal strata of interest that involve a change in receipt of treatment.

Before discussing assumptions, we need to explain the two types of changes in availability (Baker and Lindeman, 2001). For convenience we consider an increase in availability from the first to the second time period (but the reverse could also apply with re-labeling). The two types of changes in availability are

- (1) fixed availability, in which availability in the second time period subsumes the availability in the first time period,
- (2) random availability, in which availability in the second time period is greater than in the first time period, but there is a chance component as to its timing.

Fixed availability would occur if epidural analgesia is available at additional times in the second time period, for example daytime and evening in the second time period versus daytime only in the first time period. Fixed availability would also occur if epidural analgesia in the second time period is available to patients meeting certain criteria in the first time period as well as additional patients. Random availability would arise if more anesthesiologists (applying the same criteria for use of epidural analgesia) are available to provide epidural analgesia but are working elsewhere in the hospital at random times.

The following two assumptions ensure statistical identifiability of estimates obtained from the principal stratification model.

Assumption 5. Stable Effects of the Treatments of Interest. The effect of each treatment on the probability of outcome does not change over the time periods.

Assumption 6. Stable Preferences. Preference for treatment does not change over the time periods.

Assumption 5 implies that the probability of outcome in always-receives or never-receives does not vary with time period. In the application to obstetric anesthesiology, Assumption 5 would be violated for always-receives if both (i) the availability of epidural analgesia affects the time of initiation of epidural analgesia after arrival and (ii) the time of initiation of epidural analgesia affects the probability of C/S. Because a study by Chestnut (1997) found no evidence for (ii), Assumption 5 is plausible for always-receives. Assumption 5 is plausible for never-takers, because the fraction of women with rapid deliveries (who do not receive epidural analgesia) does not depend on availability of epidural analgesia.

In the application to obstetric anesthesiology, Assumption 6 is plausible because there was no new information, such as widespread reports of risk of epidural analgesia or direct to consumer advertising campaigns, that would have changed preferences. If availability is fixed, Assumption 6 implies that there are no inconsistent receivers. If availability is random, Assumption 6 implies that the effect of epidural analgesia on consistent-receives is the same as on inconsistent-receives because their preferences are constant but availability changes.

Under fixed availability, Assumptions 1 to 6 make it possible to estimate treatment effect among consistent receivers. Baker and Lindeman (1994) proposed maximum likelihood estimation. If all maximum likelihood parameter estimates lie in the interior of the parameter space, the maximum likelihood estimate of the effect of receipt of treatment among consistent receivers is the ratio of the following two quantities: (i) the difference in probabilities of outcomes in the two time periods, and (ii) the difference in the fraction who received treatment. This special maximum likelihood estimate is sometimes called a perfect fit estimate because it is derived by setting observed counts equal to their expected values (Baker and Lindeman, 1994, Baker, 2011). If some maximum likelihood estimates lie on the boundary of the parameter space, the EM algorithm can be used to compute the maximum likelihood estimates (Baker, 2011).

Under random availability, Assumptions 1 to 6 make it possible to estimate the common treatment effect among consistent-receives and inconsistent-receives. The substitution estimate equals the standard instrumental variables estimate (Baker and Lindeman, 2001).

5. Assumption of generalizability from some principal strata to the population

So far we have discussed estimation of the effect of receipt of treatment in the principal strata of interest, namely consistent-receivers in the case of fixed availability or both consistent-receivers and inconsistent receivers in the case of random availability. However, the goal is to estimate the effect of receipt of treatment in all eligible persons, not only those in the principal strata of interest. To achieve this goal, we make the following assumption.

Assumption 7. Generalizability from Some Principal Strata. The final estimated effect of receipt of treatment on the probability of outcome in the principal strata of interest is a good estimate of the effect of receipt of treatment on the probability of outcome among all eligible persons in the general population.

In the original formulation of the paired availability design, the final estimated effect of receipt of treatment in the principal strata of interest is a weighted average, over medical centers, of the estimated effect of receipt of treatment in the principal strata of interest in each medical center. The weights are proportional to the reciprocal of the estimated variances of the estimated effect of receipt of treatment. In the application to obstetric anesthesiology, the final estimated effect of the receipt of epidural analgesia on the probability C/S in the principal strata of interest was 0.00 with 95% confidence interval of (-0.06, 0.05). Under Assumption 7, this final estimate applies to all eligible persons.

Here, we propose a new final estimate of the effect of receipt of treatment on the probability of outcome for which Assumption 7 is more plausible than with the aforementioned weighted average final estimate. This new final estimate, called the extrapolated estimate, is presented in Figure 1, which is based on data from Baker and Lindeman (2001). Discussing Figure 1 generally, the y-axis is the estimated effect of receipt of treatment on the probability of outcome in the principal strata of interest. The x-axis is the change in the fraction receiving treatment. We fit a regression line to these data (using weights equal to the reciprocal of the estimated variances of the estimated effect of receipt of treatment). The extrapolated estimate is the y-value on the regression line corresponding to an x-value of change of 1.00 in the fraction receiving treatment, which corresponds to all eligible persons receiving treatment. In other words, the extrapolated estimate uses changes in the fraction receiving treatment in the various medical centers to project the estimated effect of receipt of treatment if all subjects received treatment, making Assumption 7 more likely to hold.

Figure 1. Computation of the extrapolated estimate denoted by the letter “E.” The circle size is proportional to the reciprocal of the estimated variance of the estimated effect of receipt of treatment. The line is the fitted linear model; dashed lines are 95% confidence intervals for the fitted model.

In the application to obstetric anesthesiology depicted in Figure 1 the extrapolated estimate of the effect of the receipt of epidural analgesia on the probability C/S among all eligible is 0.03 with 95% confidence interval of (-0.03, 0.09). The extrapolated estimate is qualitatively similar to the original final estimate, so does not affect the original conclusions. However, in terms of Assumption 7, it has a stronger justification than the original final estimate.

6. Relation to principal stratification in randomized trials with all-or-none compliance

It is instructive to compare principal stratification in the paired availability design with principal stratification in a randomized trial with all-or-none compliance. All-or-none compliance means that persons are randomly assigned to either treatment T0 or T1, and persons in each randomization arm immediately receive either T0 or T1. Using the stratum names coined by Angrist et al. (1996), the principal strata are

- (1) always-takers, who receive T1 regardless of assigned arm,
- (2) compliers, who receive T1 if assigned to T1 and T0 if assigned to T0,
- (3) defiers, who received T0 if assigned T1 and T1 if assigned T0,
- (4) never-takers, who receive T0 regardless of assigned arm.

Based on the review in Baker (1997), we list the following three types of randomized controlled trials involving all-or-none compliance.

(1) Subjects are randomized to either T0 or T1, but some randomized to T1 cannot receive it and so immediately receive T0, and vice versa. An example is when T1 is cervical epidural injection which sometimes cannot be performed because it is technically too difficult for the specific patient (Newcombe, 1988).

(2) Subjects are randomized to either an offer of T0 or an offer of T1; some subjects randomized to the offer of T0 immediately receive T1 and vice versa. An example is a cancer screening trial in which some subjects randomized to an offer of screening refuse screening, and some subjects randomized to no screening, immediately obtain screening outside of the trial.

(3) Subjects who can choose either T0 or T1 are randomized to no encouragement or encouragement to choose T1. This type of trial is sometimes called a randomized encouragement design. Brown et al. (2009) list various types of encouragements in these designs including reminders, incentives, additional training, and coaching.

In work independent of the paired availability design, Angrist et al. (1996) introduced two assumptions for their principal stratification model for all-or-none compliance in randomized trials: (i) the exclusion restriction which is an analog to Assumption 5 and (ii) monotonicity which is an analog to Assumption 6 for fixed availability. The exclusion restriction generally requires all-or-none compliance and would likely be violated with partial compliance involving delayed or intermittent receipt of treatment T0 or T1. Using the exclusion restriction and monotonicity assumptions, Angrist et al. (1996) estimated the effect of receipt of treatment on a continuous outcome among compliers as the ratio of the following two quantities: (i) the difference in mean outcomes between randomization groups and (ii) the difference in the fraction who receive treatment. Angrist et al. (1996) called this ratio a Local Average Treatment Effect (LATE) estimate. This is analogous to the ratio estimate among consistent receivers in the paired availability design, which can therefore be called a LATE estimate.

With respect to the use of principal stratification to draw conclusions, the main difference between the paired availability design and randomized trials with all-or-none compliance comes with generalizing the LATE estimate to all eligible persons. The paired availability design has increased plausibility for this generalization for two reasons. First the paired availability design includes the case of random availability which has no analog in randomized trials with all-or-none compliance. With random availability, as opposed to fixed availability, preferences that differ among principal strata may not play as large a role in determining the probability of outcome. Second, the paired availability design

involves an extrapolated estimate based on multiple studies with different changes in availability. While an extrapolated estimate could also be computed from multiple randomized trials with all-or-none compliance, the range of changes in fraction receiving treatment is likely to be larger with changes in availability than changes in compliance. A larger range of changes in the fraction who receive treatment increases the plausibility that Assumption 7 applies to the extrapolated estimate.

7. Discussion: response to Judea Pearl's critique

Judea Pearl (2011) noted that the LATE estimate of Baker and Lindeman (1994) and Imbens and Angrist (1994) focuses on a “specific stratum or a subset of strata for which the causal effect could be identified under various combinations of assumptions and design.” Pearl (2011) also wrote “however, most authors in this category do not state explicitly whether their focus on specific stratum is motivated by mathematical convenience, mathematical necessity (to achieve identification) or a genuine interest in the stratum under analysis.” With the paired availability design, the focus on the principal strata is motivated by genuine interest, as the principal strata of interest are only those strata that involve a change in the type of treatment received. Fortuitously, there are plausible assumptions (Assumptions 5 and 6) that ensure identification for these strata. Mathematical convenience is a by-product when perfect fit maximum likelihood estimates are available.

Pearl (2011) is also concerned that principal stratification is “an intellectual restriction that confines its analysis to the assessment of strata-specific effects”. As noted previously the goal in the paired availability design is to estimate the effect of receipt of treatment in all subjects not just those in the principal strata of interest. For this reason we believe our new extrapolated estimate is important to increase the plausibility of generalizability from the principal strata of interest to all eligible subjects.

A skeptical reader might nevertheless question the plausibility of Assumption 7, which states generalizability, and Assumptions 5 and 6, which are needed for identifiability with principal stratification. But the key point is that, to our knowledge, there is no better alternative analysis with these data. Without principal stratification and Assumption 5, 6, and 7, one could use the same data to estimate the effect of a change in availability of treatment; this estimate is simply the average of the differences in probabilities of outcome between the two time periods. However, with the possible exception of some healthcare policy questions, the goal is usually to estimate the effect of receipt of treatment, not the effect of a change in availability of treatment, which is motivation for using the tool of principal stratification.

The paired availability design fits into Pearl's interpretation of principal stratification as "an approximation to research questions concerning population averages." The operative word in Pearl's category is "approximation." The extrapolated estimate in the paired availability design is an approximate answer to the goal of estimating the effect of receipt of treatment using data from historical trials. It is approximate because of the necessary assumptions. In the application involving obstetric anesthesiology we have discussed why the assumptions are plausible. As the statistician John Tukey famously said (Meier, 1975) "an approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." Principal stratification in the context of paired availability design is valuable because it gives a plausible approximation to estimate the research quantity of interest.

References

- Angrist, J., Imbens, G., and Rubin, D. (1996): "Identification of causal effects using instrumental variables (with comments)," *Journal of the American Statistical Association*, 91, 444–472.
- Baker, S. G. (1997): "Compliance, all-or-none," in Kotz, S., Read, C. R., and Banks, D. L. (eds.). *The Encyclopedia of Statistical Science, Update Volume I*. New York: John Wiley and Sons, Inc., 134-138.
- Baker, S. G. (2011): "Estimation and inference for the causal effect of receiving treatment on a multinomial outcome: An alternative approach," *Biometrics*, 67, 319-325.
- Baker, S. G. and Lindeman, K. S. (1994): "The paired availability design: A proposal for evaluating epidural analgesia during labor," *Statistics in Medicine*, 13, 2269–2278.
- Baker, S. G. and Lindeman, K. S. (2001): "Rethinking historical controls," *Biostatistics*, 2, 383- 396.
- Baker, S.G., Lindeman, K.L., and Kramer, B.S. (2001): "The paired availability design for historical controls," *BMC Medical Research Methodology*, 1:9.
- Brown, C. H. Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthén, B., and Gibbons, R. D. (2009): "Adaptive designs for randomized trials in public health," *Annual Review of Public Health*, 30, 1-25.
- Chestnut, D. H. (1997): "Epidural analgesia and the incidence of Cesarean section. Time for another close look," *Anesthesiology*, 87, 472-6.
- Cuzick, J., Edwards, R., and Segnan, N. (1997): "Adjusting for non-compliance and contamination in randomized clinical trials," *Statistics in Medicine*, 16, 1017-1029.

- Feinstein, A. R., Sosin, D. M., and Wells, C. K. (1985): "The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer," *The New England Journal of Medicine*, 312, 1604-1608.
- Frangakis, C. E. and Rubin, D. B. (2002): "Principal stratification in causal inference," *Biometrics*, 1, 21-29.
- Imbens, G. W. and Angrist, J. D. (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467-475.
- Meier, P. (1974): "Statistics and medical experimentation," *Biometrics*, 31, 511-529.
- Newcombe, R. G. (1988): "Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur," *Statistics in Medicine*, 7, 1179-1186.
- Pearl, J. (2011): "Principal stratification — a goal or a tool?" *The International Journal of Biostatistics*, 7, 20.
- Permutt, T. and Hebel, R. (1989): "Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight," *Biometrics*, 45, 619-622.