BMC Medical Research Methodology

BioMed Central

Research article

The Paired Availability Design for Historical Controls

Stuart G Baker*[1], Karen S Lindeman[2] and Barnett S Kramer[3]

Address: [1]Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA, [2]Department of Anesthesiology, Johns Hopkins Medical Institutions, Baltimore, MD, USA and [3]Office of Disease Prevention and Medical Applications of Research, National Institutes of Health, Bethesda MD, USA

E-mail: Stuart G Baker* - sb16i@nih.gov; Karen S Lindeman - klindema@jhmi.edu; Barnett S Kramer - KramerB@OD.NIH

*Corresponding author

Published: 26 September 2001

BMC Medical Research Methodology 2001, 1:9

Received: 9 July 2001
Accepted: 26 September 2001

This article is available from: http://www.biomedcentral.com/1471-2288/1/9

Abstract

Background: Although a randomized trial represents the most rigorous method of evaluating a medical intervention, some interventions would be extremely difficult to evaluate using this study design. One alternative, an observational cohort study, can give biased results if it is not possible to adjust for all relevant risk factors.

Methods: A recently developed and less well-known alternative is the paired availability design for historical controls. The paired availability design requires at least 10 hospitals or medical centers in which there is a change in the availability of the medical intervention. The statistical analysis involves a weighted average of a simple "before" versus "after" comparison from each hospital or medical center that adjusts for the change in availability.

Results: We expanded requirements for the paired availability design to yield valid inference. (1) The hospitals or medical centers serve a stable population. (2) Other aspects of patient management remain constant over time. (3) Criteria for outcome evaluation are constant over time. (4) Patient preferences for the medical intervention are constant over time. (5) For hospitals where the intervention was available in the "before" group, a change in availability in the "after group" does not change the effect of the intervention on outcome.

Conclusion: The paired availability design has promise for evaluating medical versus surgical interventions, in which it is difficult to recruit patients to a randomized trial.

Background

In terms of avoiding bias, the most rigorous method for evaluating a medical intervention is the randomized controlled trial. However, many clinical investigators are unable to conduct a randomized trial because of excessive cost or required effort or difficulty overcoming strongly held beliefs among health care providers or patients. In these situations, a clinical investigator may consider a design and analysis based on observational data (Table 1).

One common method of inference from observational data is the cohort study with an adjustment for risk factors using, for example, regression models [1] and propensity scores [2]. In some situations, estimates from high-quality cohort studies have been similar to those from randomized trials [2][3][4][5]. However there are some notable exceptions, including studies of the effect of beta-carotene on cardiovascular mortality [6][7], the effect of hormone therapy on the rate of cardiovascular disease [8], the effect of epidural analgesia on the proba-

Table 1: Comparison of Several Methods of Evaluating a Medical Intervention

Method	Strengths		Weaknesses	
Randomized controlled trial	1. 2.	No temporal bias No selection bias	1. 2.	Cost and effort Recruitment
Observational cohort study	1.	No temporal bias	1. 2.	Cost of data collection Selection bias if an important risk factor is omitted or not adequately quantified
Paired availability design	1.	Lessens selection bias	1.	Assumptions in Table 2

bility of Cesarean section [9], the effect of beta-blockers on mortality [10], and the effect of aspirin on the risk of colorectal cancer [11]. Some of the discrepancy between the results of these high quality cohort studies and randomized trials may be explained by differences in the intervention, patient population, or duration of follow-up. Nevertheless, a major reason for bias with cohort studies is the failure to adjust for *all* factors related the receipt of intervention and outcome. This failure may be due to the inability to identify or collect the necessary data, or the difficulty measuring or quantifying subjective factors such as clinical judgment. For a fuller discussion of how an omitted factor related to receipt of intervention and outcome can bias results in a cohort study but not a randomized trial see Baker and Kramer [12].

Methods

An alternative and less widely known approach is the paired availability design for historical controls [9][13][14]. As we describe in more detail, the paired availability design consists of comparing outcomes in multiple hospitals or medical centers before versus after a change in availability of a medical intervention. To adjust for different changes in availability among the hospitals or medical centers, the test statistic for each hospital or medical center is the difference in outcome before and after the change in availability divided by the change in the fraction of patients who receive the intervention. These test statistics are combined in a meta-analysis, which weights the statistic from each hospital according to the reciprocal of the variance, a quantity that depends on sample size and the change in availability.

The paired availability design avoids many of the biases of analyses based on traditional historical controls. With traditional historical controls, investigators compare outcome among subjects who receive a new intervention with outcome among a previous group of subjects who received the standard intervention. Selection bias often arises because subjects who receive the new intervention are typically not comparable to subjects who received the standard intervention [15]. The paired availability design reduces selection bias because the comparison is between *all* subjects (those who received the intervention and those who did not) before the change in availability and *all* subjects (those who received the intervention and those who did not) after the change in availability. Thus the intervention is the availability of treatment, instead of the receipt of treatment. If the sample of all subjects eligible for intervention is comparable before and after the change in availability, one can obtain an unbiased estimate of the effect of a change in availability by comparing outcome among all subjects before the change in availability with the outcome among all subjects after the change in availability. This is analogous to obtaining an unbiased estimate of the effect of intent-to-treat by comparing outcomes among all subjects in each arm of a randomized trial subject to noncompliance.

For both the paired availability design and randomized trials subject to noncompliance, the ideal goal is an unbiased estimate of the effect of receipt of treatment. If certain requirements hold, which we discuss, a simple adjustment gives an unbiased estimate of the effect of receipt of treatment in the paired availability design. Similarly, in certain situations involving randomized trials with noncompliance, such as switching interventions immediately after randomization, a similar adjustment also yields an unbiased estimate of the effect of receipt of treatment [16][17]. Readers interested in a formal mathematical statement of these requirements and how they

Table 2: Requirements for Paired Availablity Design

Requirement	Specific Criteria	
Stable population	1.	Hospital serves one geographic area or is military medical center
	2.	No in- or out- migration
	3.	Eligibility criteria constant over time
	4.	No underlying change in prognosis over time
Stable treatment	1.	Other patient management constant over time
Stable evaluation	1.	Evaluation criteria constant over time
Stable preference	1.	No publicized credible reports
	2.	No direct-to-consumer advertising
Effect of the intervention on outcome does not change with a change in availability (only applicable when some in "before" group receive intervention)	1.	Effect of intervention does not depend on when the intervention was given during the course of disease
	2.	No learning curve for the intervention

give rise to the simple estimate should consult references [9][13][14][16][17].

Results

To assist investigators who are contemplating a paired availability design, we provide an expanded list of requirements for valid inference as well as a simpler method of data analysis than previously discussed in the literature.

Design

The paired availability design uses data collected in either a prospective or retrospective manner, or a combination of the two. Although implementing a multi-center study may initially appear burdensome, two mitigating factors lessen the burden: (1) randomization is not required and (2) investigators need not collect data on risk factors if the requirements hold. The requirements (to follow) are most likely to hold when the time period for the entire study is not too long. We recommend limiting the total study duration to not more than two years, recognizing there may be exceptions due to patient accrual rate, intervention, and outcome. If availability changes gradually, it is often sufficient to split the data halfway between the start of the "before" period and the end of the "after" period; although more sophisticated statistical techniques can be employed [9]

The change in availability between the "before" and "after" periods can take different forms which do not affect the design or analysis. With fixed availability, the intervention is available to all subjects who arrive during a certain time of day or day of the week. With random availability, the intervention is available only if the necessary personnel or equipment is available, which occurs at random. In either case, subjects can decide whether or not to undergo the intervention.

The study design has five requirements for making appropriate inference: stable population, stable treatment, stable evaluation, stable preference, and no effect of availability on the effect of intervention (Table 2). Stable population, treatment, and evaluation, are required for appropriate inference in any medical study involving comparisons over time. Stable preference and no effect of availability on the effect of intervention, are needed to adjust for differences in availability among hospitals or medical centers.

Table 3: Example of calculations from data in Baker and Lindeman [Reference 9]

hospital	before" group data			after group data			estimate	std error	weight
	n1	e1	P1	n2	e2	p2	y	s	w
1	116	.586	.172	103	.223	.184	-.033	.143	44
2	180	.290	.080	180	.440	.090	.067	.196	24
3	373	.131	.110	421	.587	.100	-.022	.048	208
4	1000	.100	.040	1000	.450	.050	.029	.026	313
5	1298	.000	.074	1084	.480	.065	-.019	.022	333
6	1919	.000	.275	2073	.316	.229	-.146	.044	225
7	3195	.010	.030	3733	.290	.031	.004	.015	365
8	4778	.008	.194	4859	.586	.190	-.006	.014	369
9	4685	.187	.149	6170	.551	.125	-.046	.015	352
10	8108	.467	.248	9918	.678	.280	.152	.031	288
11	11159	.328	.209	11869	.499	.209	.000	.031	288

n1 (n2) = number of subjects in "before" ("after") group. e1 (e2) = fraction of subjects in "before" ("after") group that had epdiural analgesia, p1 (p2) = fraction of subjects in "before" ("after") group that had a Cesarean section, y= estimated effect of epidural analgesia on the probability of Cesarean section = (p2-p1)/(e2-e1), s= standard error of y= square root of (p2 (1-p2))/n2 + p1 (1-p1)/n1) /(e2-e1)^2, w^* = weight used in random effects meta-analysis. We computed the weights as follows. Let i index studies, so y_i and s_i are the values of y and s for study i. It is convenient to define $w_1 = 1/s_i^2$. Following DerSimonian and Laird [Reference 19], to compute v, the variance of the true effect among the k studies, we set v equal to the larger of $(Q-(k-1)) / (\Sigma w_i - \Sigma w_i^2/\Sigma w_i)$ and 0, where $Q = \Sigma w_i (y_i - m)^2$, $m = \Sigma y_i w_i/\Sigma w_i$. The random-effects weights are $w^*_i = 1/(s_i^2 + v)$, and the summary statistic is $y^* = \Sigma y_i w^*_i/\Sigma w^*_i$, with standard error s^* = square root of $1/\Sigma w^*_i$. Following Proschan and Follman [reference 20], the 95% confidence interval is $(y^* - t_{k-} s^*, y^* + t_{k-1} s^*)$, where t_{k-1} is the value of the 97 $^1/_2$ percentile of a t-distribution with k-1 degrees of freedom. In this example, k = 11, Q = 50.1, v = .0025, m =-.007, s^* = .019 y^* = -.005, t_{10} = 2.23, y^* = -.005 and the 95% confidence interval is (-.047, .037).

The first requirement, stable population, is that the composition of subjects eligible for the intervention should not change from the "before" to the "after" period in ways that would affect outcome. This requirement would be violated if subjects seek treatment because of the availability of the treatment under study. The assumption is therefore violated if hospitals advertise the availability of a new diagnostic test or medical intervention. In addition, each hospital or medical center should serve a well-defined population with little in- or out- migration. Examples include the only hospital in a geographic region or a military medical center. The presence of two or more hospitals in a region could introduce bias if the new intervention were available in only one hospital and it were not possible to exclude from the analysis patients who switched hospitals to undergo the new intervention. The stable population requirement would also be violated by changes in eligibility criteria over time. If eligibility is determined by a medical diagnosis, the method of diagnosis must not change over time. Lastly the stable population requirement would be violated if the underlying prognosis of patients changed over time. For example in a study of treatment for a viral infection which is spreading through a population, the most susceptible subjects would likely enter the trial first, which would violate the stable population requirement if they have the worse prognosis after infection.

The second requirement, stable treatment, is that the patient management unrelated to the intervention is identical in the "before" and "after" groups. Thus, in studying the effect of epidural analgesia on the probability of Caesarian section, other forms of obstetric management should be constant over time. Similarly, in studying the effect of an intense chemotherapeutic regimen for cancer on survival, the type of antibiotic should not change of over time, as new and more effective antibiotics could lower treatment-related mortality irrespective of the efficacy of the anticancer regimen.

The third requirement, stable evaluation, is that the method of evaluation is identical in the "before" and "after" groups. For example, the use of a new radiologic test to stage cancer in the "after" group may artifactually improve prognosis of each stage, independent of the therapy [18].

Because the paired availability design involves multiple hospitals or medical centers, random violations of the stable population, treatment, and evaluation requirements will tend to average out, and not affect the conclusion. The main concern is with systematic violations. To minimize systematic violations, if possible, a wide variety of hospitals or medical centers should be studied.

The fourth requirement, stable preference, is strengthened in the absence of new information in the "after" period that would change a subject's preference for the medical intervention. This requirement could be violated by a widely publicized report of a harmful side effect of the new treatment, or direct-to-consumer advertising of the intervention to consumers. To the best of our knowledge, in the paired availability design to study the effect of epidural analgesia on the probability of Cesarean section, there were no credible reports of either detrimental or beneficial side effects to the mother or fetus from epidural analgesia and no relevant direct-to-consumer advertising. In contrast, if the media reported preliminary results that radioactive seed implants had fewer side effects than previous approaches for treating prostate cancer, healthier subjects who care most about the side effects may be more likely to request the new therapy than less healthy subjects who only care if treatment reduces the risk of mortality.

The fifth requirement is that the effect of the intervention on outcome does not change with a change in availability. Importantly, it applies only when there are some subjects in the "before" group who undergo the intervention. Mathematically the following two assumptions are required to estimate method effectiveness [9][13][14]. (1) Subjects in the "before" group who undergo intervention have comparable counterparts in the "after" group who undergo intervention (which is justified by the other requirements) *and the* effect of intervention is the same in both groups, i.e. it does not depend on a change in availability. (2) Subjects in the "after" group who do *not* undergo the intervention have comparable counterparts in the "before" group who do *not* undergo the intervention (as justified by the other requirements) *and* the effect of no intervention is the same in both groups. By definition no intervention is the same in the "before" and "after" groups, so a change in availability of intervention would have no bearing on (2). Thus the effect of a change in availability on the effect of the intervention only pertains to (1), where subjects in the "before" group undergo intervention.

The fifth requirement would be violated if increased availability caused some subjects to undergo the intervention sooner in the course of the disease, changing prognosis. The fifth requirement would also be violated if there were a learning curve with new intervention, such as a surgical technique that improves with the number of procedures. If such violations are likely, the design should be restricted to hospitals or medical centers where no subjects in the "before" group received the intervention.

Baker and Lindeman provided a formula to calculate the required the number of hospitals or medical centers to achieve sufficient power for hypothesis testing [13]. However, the formula may be difficult to use if the required information on the likely variability of an effect over hospitals or medical centers is not readily available. In such situations, as a rule of thumb, we recommend a minimum of 10 hospitals or medical centers, with 15 preferable, and 20 ideal.

Analysis

The purpose of the analysis is to estimate the effect of the receipt of the medical intervention, which is also called method-effectiveness [16]. As derived by Baker and Lindeman [9][13], if the aforementioned requirements hold, for each hospital or medical center, the estimated effect of receipt of treatment is

D/F, where

D= difference in outcome before and after change in availability

F = fraction that received intervention *after* change in availability – fraction that received intervention *before* change in availability

If the outcome measure is a continuous variable such as blood measure, D is a difference in the average outcomes between the "before" and "after" groups. If the outcome measure is binary, such as success or failure, D is a difference in the fraction who fail or succeed in the "before" and "after" groups.

The above estimate, D/F, has an analog in the analysis of randomized trials when some subjects switch treatments soon after randomization. With an intent-to-treat analysis, one can estimate use-effectiveness, D*, which is the effect of random assignment of treatment on outcome. Similarly, one can estimate F*, the fraction of subjects in the study group that received the new treatment minus the fraction of subjects in the control group that received the new treatment. Invoking an assumption analogous to the fifth requirement for the paired availability design, the estimated method-effectiveness is D*/F* [17].

As illustrated in the calculations accompanying Table 3, we use a standard approach for a random effects meta-analysis [19] to summarize the estimated effect of receipt of treatment over all studies. The summary statistic is a weighted average of the estimated effect of receipt of treatment for each hospital or medical center, where the weight for each hospital or medical center is the reciprocal of the sum of the sampling variance and the variance of the true effect over the studies. The sampling variance

of the estimated effect of receipt of treatment, the variability due to taking a sample from a hypothetical larger population, approximately equals the sampling variance of the numerator, which is a standard calculation, divided by denominator squared. The variability of the true effect, which arises because the medical intervention is not exactly identical among all hospitals, is computed using the formula in DerSimonian and Laird [19]. The standard error of the summary statistic is the square root of the reciprocal of the sum of the weights. An approximate 95% confidence interval is computed as the summary statistic plus or minus the standard error multiplied by the 97 1/2 percentile of a t-distribution with degrees of freedom equal to the number of hospitals or medical centers minus one [20]. This value can be found in tables in many statistics books; for example for 11 hospitals, there are 10 degrees of freedom and the value is 2.23.

Example

Baker and Lindeman applied the paired availability design to study the effect of epidural analgesia on the probability of Cesarean section [9]. They identified 11 hospitals or medical centers where epidural analgesia was introduced, expanded, or discontinued. Stable population and stable treatment requirements were supported by the reports of the investigators. Stable evaluation held because of the unambiguous nature of the outcome. The stable preference requirement likely held, as there were no widely published reports concerning risks or benefits of epidural analgesia and no direct-to-consumer advertising of the procedure. Because increased availability would likely cause some subjects to receive epidural analgesia earlier in the course of labor, there was concern about violating the fifth requirement of no effect of availability on the effect of intervention. However, because a randomized trial had previously shown that the effect of epidural analgesia on the probability of Cesarean section did not differ whether epidural analgesia was initiated early or late in labor, the requirement was thought to hold. A slightly simplified version of the data from Baker and Lindeman [9] is given in Table 3. In particular, to simplify the calculations for the hospital designated as number 11, we regrouped data from multiple time periods into two time periods.

Using the aforementioned method of analysis, with more details in the notes for Table 3, the estimated increase in the probability of Cesarean section due to epidural analgesia was -0.005 with a 95% confidence interval of (-.047, .037). This is fairly close to the more exact calculations based on a permutation distribution in [9]. Importantly, these results were similar to those from a meta-analysis of randomized trials adjusted for switching of treatments that yielded an estimate of method effectiveness of .02 with 95% confidence interval of (-.02, .08) [9]. In contrast, a high quality propensity score analysis of cohort data gave a much larger estimate of. 10 with a 95% confidence interval of (.07, .13). The bias may be due to the omission of a risk factor for intense pain early in labor [9].

Discussion

The paired availability design has promise for evaluating medical versus surgical interventions. For such an evaluation, it would be difficult to recruit patients to a randomized trial because few patients want to be randomized to those options. Also, many physicians feel uncomfortable assigning their patients to invasive versus non-invasive interventions. Thus a validated alternative method of evaluation would be of considerable value. We think that, in some cases, the paired availability design would be well suited for this type of evaluation. The key to the stable population requirement is having clear and constant eligibility criteria. For stable treatment, ancillary care and the method of evaluation must be the same over time. For the stable preference assumption to hold, there should be no advertising of the medical intervention. For the requirement of no effect of availability on the effect of intervention, either the surgical technique should have stabilized or the design should only include hospitals with no previous surgeries.

A possible example would be an analysis of surgical removal of liver metastases in patients with colorectal cancer. Although liver metastatectomy has been associated with favorable outcomes, a more rigorous evaluation is needed. An analysis of prospective cohort data is likely to be biased because of the difficulty observing or quantifying important risk factors such as patient performance status, tumor doubling times, and meticulous staging.

Several conditions listed in Table 2 are favorable to a paired availability design. The surgical approach has been relatively stable for years. The use of CT scans and CEA blood testing in follow-up of patients after resection of the primary tumor has been popular for at least two decades. Although systemic therapy has changed, efficacy of chemotherapy for metastatic disease has reached a plateau. Metastatectomy is not one of the procedures heavily advertised by hospitals or medical centers. An ideal circumstance would be to apply the paired availability design to hospitals before and after the arrival of a surgical oncologist who brings the procedure into common practice for the first time at that hospital.

A well-designed randomized study of liver metastatectomy would still give a more statistically valid assessment of the procedure than the paired availability design. However, such a randomized study has never been done, despite the use of metastatectomy for many years. A

paired availability design would likely be subject to fewer biases than a cohort study comparing outcomes of patients who did versus those who did not undergo the surgical procedure.

To decide if a method for analyzing observational data is generally reliable, one should have experience comparing the results to those obtained from a randomized trial. In the only application of the paired availability design to date. Baker and Lindeman obtained similar results from the paired availability design as from a meta-analysis of randomized trials. These results differed substantially from a multivariate adjustment for concurrent controls, which likely omitted an important risk factor[2]. Hopefully this article will spur new studies using the paired availability design, including some comparing the results to those from randomized trials.

Conclusion

We wish to emphasize that the randomized trial represents the strongest form of evaluation and should be implemented if possible. However we recognize that there are situations where the randomized trial is difficult to implement, such as comparing medical versus surgical interventions. If the requirements for the paired availability design are met, we recommend it as an alternative with advantages over the usual analyses from observational studies.

Competing Interests

None Declared

References

1. Concato J, Feinstein AR, Holford TR: **The risk of determining risk with multivariable models.** *Ann Intern Med* 1993, **118**:201-210
2. Rubin DB: **Estimating causal effects from large data sets using propensity scores.** *Ann Intern Med* 1997, **27**:757-763
3. Concato J, Shah H, Horwitz RI: **Randomized controlled trials, observational studies, and the hierarchy of research designs.** *N Engl J Med* 2000, **342**:1887-92
4. Benson K, Hartz AJ: **A comparison of observational and randomized, controlled trials.** *N Engl J Med* 2000, **342**:1878-86
5. Baker SG, Lindeman KS: **Randomized and nonrandomized studies. Statistical considerations.** *Anesthesiology* 2000, **92**:928-930
6. Egger M, Schneider M, Smith GD: **Spurious precision? Meta-analysis of observational studies.** *Br Med J* 1998, **316**:140-144
7. Jha P, Flather M, Lonn E, Farkouh M, Yusuf S: **The antioxidant vitamins and cardiovascular disease. A critical review of epidemiologic and clinical trials data.** *Ann Intern Med* 1995, **123**:860-872
8. Grady D, Hulley SB: **Hormones to prevent coronary disease in women: when are observational studies adequate evidence?** *Ann Intern Med* 2000, **133**:999-1001
9. Baker SG, Lindeman KS: *Rethinking historical controls. Biostatistics,* 2001
10. Smith RP, Meier P: **Observational studies and randomized trials.** *New Engl J Med* 2000, **343**:1196
11. Sandler RS: **Aspirin and other nonsteroidal anti-inflamatory agents in the prevention of colorectal cancer.** *PPO Updates* 1997, **11**:1-14
12. 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.** *Journal of Women's Health & Gender-Based Medicine,* 2001
13. Baker SG, Lindeman KS: **The paired availability design: A proposal for evaluating epidural analgesia during labor.** *Statistics in Medicine* 1994, **13**:2269-2278
14. Baker SG: **The paired availability design: an update.** *In Abel U, Koch A, Nonrandomized Comparative Clinical Studies, Dusseldorf: Medinform-Verlag;* 199879-84
15. Pocock SJ: **Clinical trials. A practical approach.** *Chichester: John Wiley and Sons,* 1982
16. Sheiner LB, Rubin DB: **Intention-to-treat analysis and the goals of clinical-trials.** *Clin Pharmacol Ther* 1995, **57**:6-15
17. Angrist JD, Imbens GW, Rubin DR: **Identification of causal effects using instrumental variables.** *Journal of the American Statistical Association* 1996, **92**:444-455
18. Feinstein AR, Sosin DM, Wells CK: **The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer.** *N Engl J Med* 1985, **312**:1604-8
19. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Controlled Clinical Trials* 1986, **7**:177-188
20. Follman DA, Proschan MA: **Inference in random effects meta-analysis.** *Biometrics* 1999, **55**:732-737