Published in final edited form as:
Nurs Res. 2009 ; 58(2): 123–127. doi:10.1097/NNR.0b013e318199b517.

Simpson's Paradox and Experimental Research

Suzanne Ameringer, PhD, RN [Assistant Professor],
Virginia Commonwealth University Richmond, Virginia

Ronald C. Serlin, PhD [Chair], and
Department of Educational Psychology University of Wisconsin-Madison Madison, Wisconsin

Sandra Ward, PhD, RN [FAAN Helen Denne Schulte Professor]
University of Wisconsin-Madison Madison, Wisconsin

Abstract

Background: Experimental research in nursing has increased considerably in recent years. To improve the quality of such research, it is critical to reduce threats to internal validity. One threat that has received inadequate attention in the nursing literature is Simpson's paradox--a case of extreme confounding that can lead to erroneous conclusions about the effects of an experimental intervention. In fact, it can lead to a conclusion about an intervention effect that is the opposite of the correct inference.

Objectives: To describe Simpson's paradox, provide a hypothetical example, and discuss approaches to avoiding the paradox.

Results: The paradox is due to the combination of an overlooked confounding variable and a disproportionate allocation of that variable among experimental groups. Different designs and analysis approaches that can be used to avoid the paradox are presented.

Discussion: Simpson's paradox can be avoided by selecting an appropriate experimental design and analysis that incorporates the confounding variable in such a way as to obtain unconfounded estimates of treatment effects, thus more accurately answering the research question.

Keywords

confounding variables; Simpson's paradox; stratification

Over the past 20 years, nursing investigators have turned increasingly toward conducting experimental tests of innovative interventions. Given that such experiments are intended to determine whether interventions should be used in practice, rigor is critical in study design. It is important to be reminded of potential threats to internal validity that can arise in experimental studies. One threat that has received inadequate attention in the nursing literature is Simpson's paradox. Simpson's paradox arises from the combination of an ignored confounding variable and a disproportionate allocation of the variable, and it can lead to a conclusion about an intervention effect that is the opposite of the correct inference (hence a paradox). Simpson (1951) demonstrated how differential analyses of contingency tables (i.e., analysis in which

Correspondence: Suzanne Ameringer, PhD, RN School of Nursing Virginia Commonwealth University 1100 East Leigh Street Richmond, VA 23219 Phone: (804) 628-7551 swameringer@vcu.edu.

the confounding variable is excluded or included) can lead to different conclusions. The impact of Simpson's paradox often has been discussed in relation to descriptive studies, but rarely has it received attention in the context of experimental studies. Given the enormous investment in time and cost in testing an intervention and the potential impact it could have on health outcomes, it is incumbent upon investigators to determine an intervention's effects as accurately as possible. The purpose of this paper is to raise awareness of the potentially devastating effects of Simpson's paradox in experimental studies. This paper comprises three sections--a description of Simpson's paradox, a hypothetical example, and a discussion of ways to avoid Simpson's paradox.

What is Simpson's Paradox?

Simpson's paradox is an extreme condition of confounding in which an apparent association between two variables is reversed when the data are analyzed within each stratum of a confounding variable. Simpson (1951) demonstrated how, when two or more 2 × 2 contingency tables are collapsed into one table, the findings from the collapsed table can be contradictory to the findings from the original tables. With Simpson's paradox, in the collapsed table the marginal correlation between cause and effect would be considered spurious. A spurious association is one that cannot be inferred to be causal because a third factor functions as a cause of the correlation among the variables. In experimental research, a spurious relationship can lead to an erroneous conclusion that an intervention is effective, when in fact it is not. This erroneous conclusion can lead to the ineffective intervention being implemented and to investigators building further studies on the erroneous conclusions, with concomitant waste of time, effort, and other resources.

For this paradox to occur, two conditions must be present: (a) an ignored or overlooked confounding variable that has a strong effect on the outcome variable; and (b) a disproportionate distribution of the confounding variable among the groups being compared (Hintzman, 1980; Hsu, 1989). The effect size of the confounding variable has to be strong enough to reverse the zero-order association between the independent and dependent variables (Cornfield et al., 1959), and the imbalance in the groups on the confounding variable has to be large (Hsu, 1989). Key to the occurrence of this paradox is the *combination* of these two conditions, because unequal sample sizes alone generally are not a problem as long as they are not coupled with other internal validity issues, such as power. (For further reading on Simpson's paradox, see Neutel, 1997 and Rücker & Schumacher, 2008.)

The extent to which Simpson's paradox is likely to occur in experimental research is difficult to determine because what has not been tested and reported in a publication can not be detected easily by a reader. One way to investigate this matter is to examine findings across studies. If there is inconsistency in the relationship between an outcome and treatment across studies, then it may be that confounding has occurred in at least some of those studies. A number of examples of Simpson's paradox have been described in the literature (Appleton, French, & Vanderpump, 1996; Julious & Mullee, 1994; Perera, 2006; Reintjes, de Boer, van Pelt, & Mintjes-de Groot, 2000; Wagner, 1982). Four of these articles address studies from the health care literature, two of which are described here. Although these were not experimental studies, the deleterious consequences of overlooking confounding variables are demonstrated. Reintjes et al. (2000) presented an example of Simpson's paradox from a study of urinary tract infections (UTIs) in which the association between antibiotic prophylaxis and UTIs had a relative risk (RR) <1, but when the data were stratified by hospitals with low and high incidence of UTIs the results were reversed and the RR was >1 within each stratum. That is, once the confounding variable (hospital's incidence of UTI) was included in the analysis, it became clear that antibiotic prophylaxis was associated with a higher incidence of UTIs.

Appleton et al. (1996) drew an example from a study of smoking and long-term survival status in women. The percentage of women who smoked and lived was significantly higher than the percentage of women who smoked and died. But when the data were stratified by age, the odds shifted, and within each of the age groups women who smoked were less likely to survive than women who did not smoke. In this example, age was the confounding variable that was overlooked in the initial analysis. Age was related to the outcome, survival status. In addition, the relative number of women who evidenced long-term survival greatly differed among the age categories, with the proportion of women not surviving increasing with age. In both of these studies, the two conditions necessary for Simpson's paradox to arise were present--an overlooked and unequally distributed confounding variable. The reason the effects are reversed is not intuitively clear. A hypothetical example of an experimental study is provided to aid in understanding this phenomenon.

A Hypothetical Example of an Experimental Study

Demonstrated in this hypothetical example is how results can differ in an experimental study, depending on whether or not the confounding variable is taken into account. This is an example of a test of the effectiveness of an intervention on the improvement of patients. The results for this example are shown in Tables 1 and 2. The two experimental groups are intervention and control (standard care). The confounding variable is Seriousness (the seriousness of the illness), with two strata: High and Low. The outcome is Improvement. Improvement scores can range from 0 to 25, with higher scores indicating greater improvement. The results for the intervention and control groups (main effect) are shown in Table 1. Note that the confounding variable was excluded in this table. The control group's Improvement scores were significantly higher than the intervention group's scores, $t(98) = -2.00$, $p = .048$, indicating that the intervention was not as effective as the standard care treatment.

Examining the results shown in Table 2 when the confounding variable is included, the mean Improvement scores for the two groups are stratified by High and Low Seriousness. When the confounding variable is taken into account, the results are reversed. For each stratum of Seriousness, the intervention group did significantly better than the control group. Note that the total number of subjects in both the intervention ($n = 50$) and control groups ($n = 50$) is equal, but the proportions of subjects in the High and Low Seriousness strata differ between the groups (intervention = 34:16 vs. control = 16:34). Contradictory findings between main effects (Table 1) and strata effects (Table 2) result from both the disproportionate allocation of the confounding variable among the groups and the strong association between the confounding variable (Seriousness) and the outcome variable (Improvement).

Avoiding Simpson's Paradox

Avoiding Simpson's paradox centers on first identifying potential confounding variables during the planning phase of a study, and then controlling for these variables in both the design and analysis. The goal, of course, is to determine the true relationship between the independent and outcome variables; that is, to obtain unconfounded estimates of treatment effects.

During the planning phase of a study, relevant confounding variables must be identified and the likely strata frequencies in the treatment groups determined. Identifying potential confounding variables can be accomplished by reviewing the literature and by conducting pilot studies. For example, in the pediatric pain literature, gender has been associated with pain intensity, in that females have reported higher levels of pain intensity than males (Keogh & Eccleston, 2006). Gender, then, has the potential to be a confounding variable. Further, if a particular pain condition under study is more prevalent in males than females, then it would be expected that fewer females would be recruited. Unfortunately, despite extensive review of the literature, there is no guarantee that every potential confounding variable will be known.

The next step is to determine which effects will be examined, or which effects (main or strata) accurately reflect the hypothesis being tested. Consider the hypothetical example previously presented. The hypothesis should state whether the main effects (Intervention versus Control) or the strata effects (Intervention effects within High and Low Seriousness strata) will be tested.

If the hypothesis specifies testing main effects, then the strata effects of the confounding variable would not be tested. However, the main effect means and proportions will be affected by both the unequal cell sizes and the association between the confounding variable and the outcome. Therefore, it would be essential to control for the effects of the confounding variable in the design and analysis, even though these effects would not be tested. Consider the hypothetical example. If the researcher's aim is to determine treatment effects (intervention versus control), then results about the effectiveness of the intervention would be sought, regardless of where the study is conducted (setting) or how seriously ill the patients are. One consideration, though, is that the researcher will need to decide how to interpret the results if there is a treatment by stratum interaction. An interaction would indicate that differences between treatments exist across strata, which would not be revealed in the test of the main effects. In this study, the goal would be to obtain unconfounded main effects by controlling for the seriousness of the illness in the design and analysis.

On the other hand, the hypothesis may call for the strata effects to be tested. For example, the effects of interest may be the treatment effects between the experimental groups for each stratum of Seriousness. Because *separate* strata effects are of interest, the researcher will not test whether the results within separate strata differ from each other. As with main effects, an appropriate design and analysis to best estimate strata effects should be chosen. In estimating the within-strata effects, results are not confounded by unequal cell sizes and the association between the treatment and the confounding variable. In summary, it is the hypothesis that determines the level at which to draw a conclusion, main or strata effects.

Once the researcher has identified relevant confounds and has determined which effects are of interest, the next step is to choose an appropriate design and analysis plan that best controls for the confounding variable in the context of the research question. A key point to consider in designing a study is that Simpson's paradox could not arise if the groups are equivalent on the confounding variable. To avoid Simpson's paradox, a design that is most effective for generating balanced group sample sizes would be selected and appropriate statistical control procedures would be applied to account for potential confounding factors. A number of designs can be used to achieve balance among treatment groups. Three designs that can be used in experimental studies for producing groups balanced on confounding variables are presented: simple randomization, randomized block design, and minimization. The researcher must decide which design is most suitable for both controlling the confounding variable and answering the research question.

Simple randomization is a method of assigning subjects to groups in a random manner. Groups produced by simple randomization are expected to be equivalent because each subject in the sample has an equal chance of being assigned to any group. Differences between groups are attributed to chance only (Polit & Hungler, 1999). An advantage of randomization is that it controls for both known and unknown extraneous variables, but there is one caveat. With a large sample size, generally greater than 200 (Lachin, 1988; Shadish, Cook, & Campbell, 2002), simple randomization is likely to generate equivalent groups. But with a small sample size, simple randomization may be less effective in achieving proportional distributions of confounding variables (Hsu, 1989). When the relative frequencies of a known confound are low for one or more of the strata of the confounding variable and the researcher wants to use simple randomization, it may be necessary to increase the sample size to achieve close equivalence between groups. When information about confounds is lacking, Hsu (1989) suggests using simple randomization, but the sample size should be *sufficient*. Hsu provides a chart of estimated probabilities associated with nonequivalence of confounding variables in randomized, equal-sized groups that is a useful guide to researchers for determining a sufficient sample size.

To better ensure a proportional distribution of subjects on a known confounding variable, an alternative to using simple randomization is to use blocking. A randomized block design is one “in which subjects are paired or grouped into subsets on the basis of common characteristics” (Marascuilo & Serlin, 1988, p. 563), thus reducing the effects of the confounding variable. Using the example of pain in children and gender, the blocking variable would be gender. Male and female subjects would be randomized separately to the treatment conditions. In this way, if the effect of an intervention on a pain condition that occurs more frequently in males was being examined, there would be a proportional distribution of males and females between the treatment conditions. A disadvantage of the randomized block design is that, although it is generally practical with just a few blocking variables, it is completely unwieldy when there are many blocking variables.

Minimization, a dynamic allocation procedure, is a third option for generating equivalent groups on confounding variables in intervention studies. Minimization is a “method of randomly assigning subjects to comparable groups in which levels of selected potentially confounding covariates are evenly distributed” (Zeller, Good, Anderson, & Zeller, 1997, p. 345). Unlike the randomized block design, minimization is particularly useful for studies with numerous confounds. In the case of experimental studies, subjects are recruited, assessed for selected covariates or confounds, and assigned to one of the treatment groups. Assignment to condition can be determined in several ways, but it is basically determined by “calculating for each treatment group the comparative degree of imbalance that would occur if the subject were assigned to that group” (McEntegart, 2003, p. 298). The subject would be assigned to the treatment group in which the imbalance between groups on the selected confounding variables is minimized by assigning the subject into that group. For example, if a preselected confounding variable is gender and the next subject enrolled (e.g., female) would increase the imbalance (of females) in one treatment group, then she would be assigned to the other treatment group to minimize the imbalance. Several disadvantages have been noted about minimization. Tu, Shalay, and Pater (2000) reported that when there were covariate interactions, (interactions between confounding variables) minimization was less effective in balancing groups than blocking. McEntegart (2003) noted two important considerations when using minimization. First, minimization is most effective when each of the levels of a factor has approximately equal numbers of subjects. Second, if a factor has small numbers of subjects, then assigning a higher weight to this factor would be required to achieve balance, though giving more weight to this factor could affect the balance deleteriously on the other factors. Another disadvantage of minimization is that the allocation scheme can, but commonly does not, include a random element, and therefore an equal distribution of unknown variables is less likely. Use of a random element is suggested when using a dynamic allocation procedure such as minimization (Altman et al., 2001; International Conference on Harmonisation Harmonised Tripartite Guideline, 1999).

A priori identification of potential confounding variables is key to avoiding Simpson's paradox. However, if an investigator performs an analysis and suspects that a confounding variable is having an effect on the outcome, an exploratory analysis can be conducted, though findings should be reported as such. In fact, when reporting randomized, controlled trials, the Consolidate Standards of Reporting Trials (CONSORT) guidelines strongly recommend describing whether adjusted variables were chosen a priori (prespecified) or exploratory (based on the data; Altman et al., 2001) due to the validity of hypotheses-driven versus data-driven findings. Confounding variables identified on an exploratory basis can then be built into the design of and tested in future studies.

Whether the focus is on main or on strata effects, the researcher should control statistically for confounding variables. Again, to avoid Simpson's paradox, the goal is to obtain unconfounded treatment effects. Statistical control reduces variability associated with the confounding variables, thus reducing the potential confounding of those variables (Tu et al., 2000). The purpose of employing statistical control procedures is to make group effects less affected by the confounding variable.

Regardless of the statistical procedure employed, when the confounding variable is a blocking variable, as in the randomized block and minimization designs, the blocking variable should be included as a factor in the analysis; otherwise, the *p*-value for the difference between the experimental groups may be overestimated (Tu et al., 2000; Weir & Lees, 2003). With the confounding variable and the treatment variable included in the analysis, there are a number of statistical models that can be applied when analyzing the data. The default models in most statistical packages such as SPSS and SAS result in the marginal means being estimated as simple arithmetic averages of the strata means. This is a reasonable approach in experimental research, because there are no extant treatment and control populations, and so hypothetically the confounding variable would be expected to be represented proportionally in the treatment and control populations. In the example above, the estimate of the marginal treatment mean would equal the average of 19 and 13.91, or 16.46 (compared to 15.54 in Table 1, which is the usual weighted arithmetic mean), and the estimate of the marginal control mean would equal the average of 18.26 and 13.13, or 15.70 (compared to 16.62 in Table 1). This results in a treatment effect of (16.46−15.70) = 0.76, which is the average of the treatment effects in the two strata, (19−18.26) = 0.74 and (13.91−13.13) = 0.78. By weighting the strata means equally in the estimation of the marginal means, the differences in the marginal treatment means (in the collapsed table, Table 2) is consistent with the treatment differences in the strata, and Simpson's paradox is avoided.

Multiple regression analysis and analysis of covariance (ANCOVA; a special case of multiple regression) are very effective for obtaining unconfounded treatment effects, because confounding variables are controlled for statistically by providing estimates of treatment effects that have the variance that is explained by the confounding variable removed (Cohen, 1982; Marascuilo & Serlin, 1988). Although a nominal variable was used as a confounding variable in the hypothetical example, ordinal, interval, or ratio variables also have the potential to confound (and reverse) a relationship. Any variable can be treated as a covariate in its original scaling or it can be dichotomized. However, dichotomization should be used with caution because it has the potential to cause a loss of information, effect size, and power (MacCallum, Zhang, Preacher, & Rucker, 2002), as well as to cause spurious results (Maxwell & Delaney, 1993). Simple randomization combined with multiple regression or ANCOVA, using the confounding variable as a categorical independent factor, in the case of a nominal variable, or as the covariate, can eliminate the risk of obtaining confounded treatment effects.

In conclusion, experimental research is critical to advancing nursing knowledge because of its value in testing the effects of interventions. Simpson's paradox, a case of extreme confounding, is a threat to internal validity in experimental research that can be avoided by identifying relevant confounding variables, determining which effects address the hypothesis under question, choosing an appropriate design for balancing groups, and using statistical control procedures. A meticulous approach to planning and designing an experimental study can reduce the risk of Simpson's paradox.

References

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Annals of Internal Medicine 2001;134(8):663–694. [PubMed: 11304107]

Appleton DR, French JM, Vanderpump MPJ. Ignoring a covariate: An example of Simpson's paradox. The American Statistician 1996;50(44):340–341.

Cohen, J. “New-look” multiple regression/correlation analysis and the analysis of variance/covariance. In: Keren, G., editor. Statistical and methodological issues in psychology and social sciences research. Erlbaum Associates; Hillsdale, NJ: 1982. p. 41-69.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. Journal of the National Cancer Institute 1959;22(1):173–203. [PubMed: 13621204]

Hintzman DL. Simpson's paradox and the analysis of memory retrieval. Psychological Review 1980;87 (4):398–410.

Hsu LM. Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. Journal of Consulting and Clinical Psychology 1989;57(1):131–137. [PubMed: 2647799]

International Conference on Harmonisation Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group; Statistics in Medicine; 1999. p. 1905-1942.

Julious SA, Mullee MA. Confounding and Simpson's paradox. British Journal of Medicine 1994;309 (6967):1480–1481.

Keogh E, Eccleston C. Sex differences in adolescent chronic pain and pain-related coping. Pain 2006;123 (3):275–284. [PubMed: 16644131]

Lachin JM. Properties of simple randomization in clinical trials. Controlled Clinical Trials 1988;9(4): 312–326. [PubMed: 3203523]

Marascuilo, LA.; Serlin, RC. Statistical methods for the social and behavioral sciences. W. H. Freeman and Company; New York: 1988. p. 562-591.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods 2002;7(1):19–40. [PubMed: 11928888]

Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psychological Bulletin 1993;113:181–190.

McEntegart DJ. The pursuit of balance using stratified and dynamic randomization techniques: An overview. Drug Information Journal 2003;37(3):293–308.

Neutel CI. The potential for Simpson's paradox in drug utilization studies. Annals of Epidemiology 1997;7 (7):517–521. [PubMed: 9349920]

Perera R. Statistics and death from meningococcal disease in children. British Medical Journal 2006;332 (7553):1297–1298. [PubMed: 16740556]

Polit, DF.; Hungler, BP. Nursing research. Principles and methods. 6th ed.. Lippincott; Philadelphia: 1999. p. 175-218.

Reintjes R, de Boer A, van Pelt W, Mintjes-de Groot J. Simpson's paradox: An example from hospital epidemiology. Epidemiology 2000;11(1):81–83. [PubMed: 10615849]

Rücker G, Schumacher M. Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. BMC Medical Research Methodology 2008;8:34. [PubMed: 18513392]

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin; Boston: 2002. p. 279-313.

Simpson EH. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society. Series B (Methodological) 1951;13(2):238–241.

Tu D, Shalay K, Pater J. Adjustment of treatment effect for covariates in clinical trials: Statistical and regulatory issues. Drug Information Journal 2000;34:511–523.

Wagner CH. Simpson's paradox in real life. The American Statistician 1982;36(1):46–48.

Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. Statistics in Medicine 2003;22:705–726. [PubMed: 12587101]

Zeller R, Good M, Anderson GC, Zeller DL. Strengthening experimental design by balancing potentially confounding variables across treatment groups. Nursing Research 1997;46(6):345–349. [PubMed: 9422055]

Table 1

Mean Improvement Scores of Patients by Group (n = 100)

Intervention (*n* = 50)		Control (*n* = 50)			
Mean	***SD***	**Mean**	***SD***	***t* (98)**	***p***
15.54	2.71	16.62	2.69	−2.00	.048

Table 2

Mean Improvement Scores of Patients by Groups Stratified by Seriousness of Illness (n = 100)

	Intervention (*n* = 50)			Control (*n* = 50)				
Seriousness	***n***	**Mean**	***SD***	***n***	**Mean**	***SD***	***t* (48)**	***p***
High	34	13.91	1.26	16	13.13	1.26	2.06	.045
Low	16	19.00	1.32	34	18.26	1.14	2.03	.048