

# **Making Informed Decisions: Assessing the Strengths and Weaknesses of Study Designs and Analytic Methods for Comparative Effectiveness Research**

*A Briefing Document for Stakeholders*

---

February 2012

## **Senior Editors**

Priscilla Velengtas, PhD  
Outcome

Penny Mohr, MA  
Center for Medical Technology Policy

Donna A. Messner, PhD  
Center for Medical Technology Policy

## **ACKNOWLEDGMENTS**

The editors would like to acknowledge the following individuals who authored content for this document:

April Duddy, MS  
Outcome

Kristina Franke, MPH  
Outcome

Danielle M. Whicher, MHS  
Center for Medical Technology Policy

Rachael M. Moloney, MHS  
Center for Medical Technology Policy

Swapna U. Karkare, MS  
Center for Medical Technology Policy

The editors also would like to thank Michelle B. Leavy, MPH, of Outcome, who served as the managing editor for this document, and Jennifer S. Graff, PharmD, of the National Pharmaceutical Council, who reviewed and contributed to the document.

## TABLE OF CONTENTS

<b>1.</b>	<b>Introduction</b>	<b>2</b>
<b>2.</b>	<b>Experimental Study Designs</b>	<b>4</b>
<b>2.1</b>	Pragmatic Clinical Trials (PCTs)	<b>4</b>
<b>2.2</b>	Crossover Designs	<b>6</b>
<b>2.3</b>	N of 1 Randomized Controlled Trials (RCTs)	<b>8</b>
<b>2.4</b>	Cluster Randomized Controlled Trials (RCTs)	<b>10</b>
<b>2.5</b>	Delayed-start Designs	<b>12</b>
<b>3.</b>	<b>Experimental Methods</b>	<b>15</b>
<b>3.1</b>	Adaptive Designs and Bayesian Methods	<b>15</b>
<b>4.</b>	<b>Nonexperimental Study Designs</b>	<b>17</b>
<b>4.1</b>	Cohort and Case-Control Studies	<b>17</b>
<b>5.</b>	<b>Nonexperimental Methods</b>	<b>19</b>
<b>5.1</b>	New-User Designs	<b>19</b>
<b>5.2</b>	Restriction	<b>21</b>
<b>5.3</b>	Subgroup Analysis	<b>22</b>
<b>5.4</b>	Propensity Score Methods	<b>23</b>
<b>5.5</b>	Instrumental Variable Methods	<b>25</b>
<b>5.6</b>	Sensitivity Analyses	<b>27</b>
<b>5.7</b>	External Information	<b>28</b>
<b>6.</b>	<b>Glossary of Terms</b>	<b>30</b>

## 1. INTRODUCTION

Comparative effectiveness research (CER) is defined as “the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in “real-world’ settings.”<sup>1</sup> The goal of CER is to improve patient outcomes by providing decision-makers, such as patients, providers, policy-makers, and payers, with information as to which interventions are most effective for specific types of patients. As the number of treatment options for many conditions has increased, decision-makers have begun seeking comparative information to support informed treatment choices. Comparative effectiveness information is often not available, however, either due to lack of funding, or because clinical research focuses on demonstrating efficacy. Efficacy measures how well interventions or services work under ideal circumstances, while effectiveness examines how well interventions or services work in real-world settings, where patients may have more complex conditions. The Institute of Medicine has estimated that less than half of all medical care in the United States is supported by adequate effectiveness evidence.<sup>2</sup>

CER aims to close this evidence gap by producing information that decision-makers can use to make informed treatment and coverage decisions. CER therefore must be designed to meet the real-world needs of decision-makers. This practical focus of CER introduces unique requirements for the design and implementation of studies. For example, tradeoffs of validity, relevance, feasibility, and timeliness must be considered in the context of the specific decision-makers and decisions. These unique considerations lead to questions concerning which study designs and methods are appropriate for CER questions.

Understanding which approach to conducting a CER study is best to use under which circumstances is a question of significant debate. Generally, CER approaches fall into two broad categories: experimental study designs and methods, and nonexperimental study designs and methods. In experimental designs, patients are randomized (assigned by chance, not by a physician’s decision) to a particular therapy based on the study protocol. In nonexperimental designs, patients and physicians make real-world treatment decisions, and patterns of care and outcomes are observed. Some argue that experimental study designs are needed to answer most CER questions, because randomization eliminates concerns regarding channeling bias (ie, the tendency of clinicians to prescribe specific treatments based on a patient’s prognosis), and achieves balance with regard to measured and unmeasured confounders (ie, extraneous variables that may play a role in the outcomes of interest). Others believe that nonexperimental studies, incorporating ways to address channeling bias and other confounding in the design and/or analysis, represent important alternatives to randomized studies. In practice, each research approach has advantages and disadvantages, and the research approach for a CER question should be selected based upon the specific features or characteristics of the study question.

The purpose of this document is to provide brief descriptions of both experimental and nonexperimental study designs and methods that may be used to address CER study questions. Each design or analytic topic is described, along with the strengths and limitations associated with the approach. Examples are provided to demonstrate the use of the described methods in the literature. While this document does not prescribe methodologies for specific CER research questions, it is part of a larger effort to develop a systematic approach to determining which methods are best able to address given CER questions. In its current form, this document provides information needed for researchers and consumers of the literature to understand the relative strengths and limitations of various CER design and analytic approaches, and how their use may affect study results and interpretation.

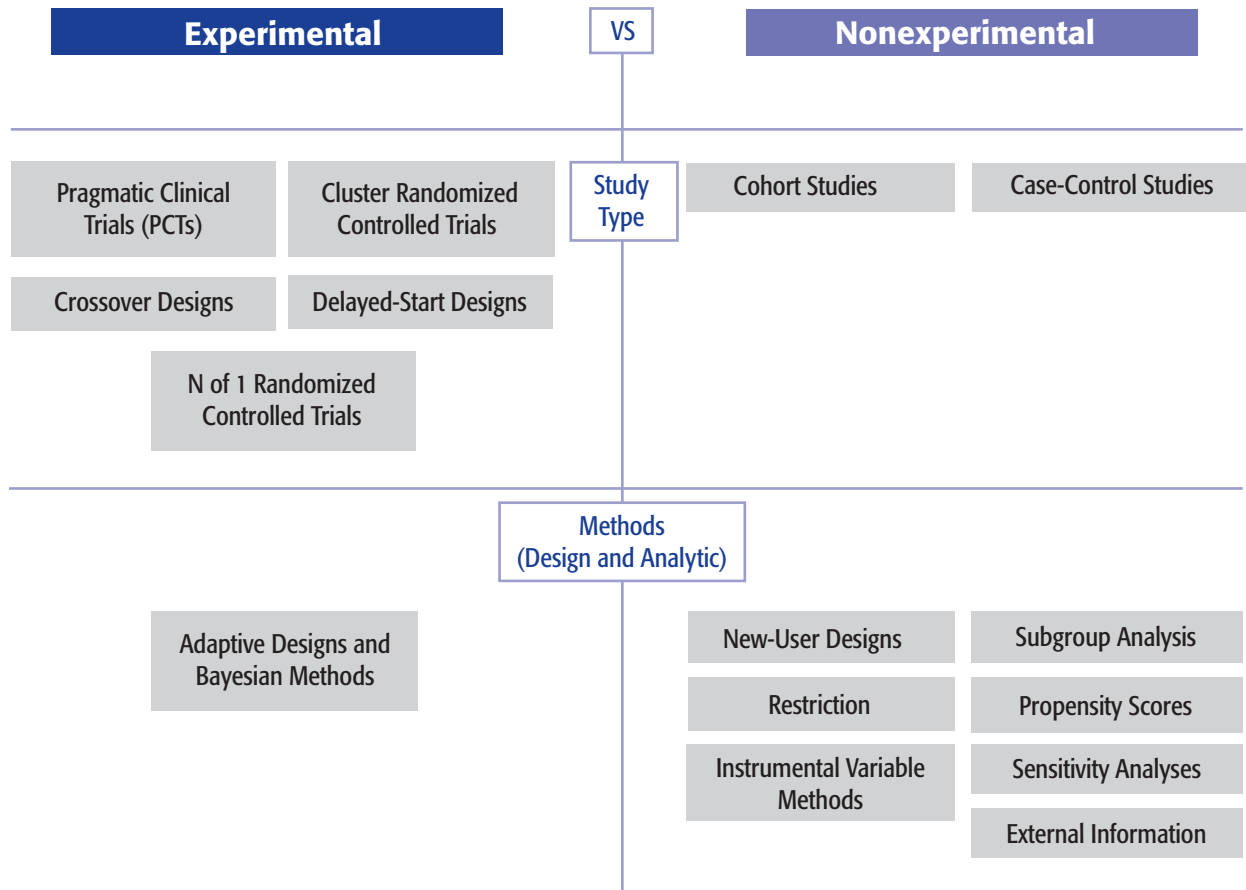
This document is organized into four sections: experimental study designs; experimental methods; nonexperimental study designs; and nonexperimental methods. The organization of the document is depicted in Figure 1. A glossary of frequently used terms may be found at the end of the document.

---

<sup>1</sup> Federal Coordinating Council for Comparative Effectiveness Research. *Report to the President and the Congress*. US Department of Health and Human Services; June 30, 2009.

<sup>2</sup> Institute of Medicine (IOM). *Knowing What Works in Health Care: A Roadmap for the Nation*. Washington, DC: The National Academies Press; 2008.

**Figure 1. Experimental and nonexperimental study types and methods**



## 2. EXPERIMENTAL STUDY DESIGNS

### 2.1 Pragmatic Clinical Trials (PCTs)

#### Introduction

Pragmatic clinical trials (PCTs) are randomized controlled trials (RCTs) that are designed to determine the risks, benefits, and costs of an intervention as they would occur in routine clinical practice.<sup>1-2</sup> They are often contrasted with explanatory trials, which aim to determine whether a clinical intervention is effective under optimal circumstances (eg, in a carefully chosen, treatment-adherent patient population).<sup>1</sup>

#### Pragmatic versus Explanatory Trials

In contrast to explanatory trials, PCTs generally include a broader range of patients (by reducing the number of inclusion/exclusion criteria), a broader range of study sites (by including community-based, non-academic sites), and outcomes that are aligned with the evidence needs of decision-makers (eg, patients, clinicians, payers, and policy-makers).<sup>3-4</sup>

It is important to note that pragmatic and explanatory trials are not distinct concepts. Rather, trials may incorporate differing degrees of pragmatic and explanatory components. For example, a trial may have strict eligibility criteria, including only high-risk, compliant, and responsive patients (explanatory side of the spectrum), but have minimal to no monitoring of practitioner adherence to the study protocol and no formal follow-up visits (pragmatic side of the spectrum). The degree to which a trial is “pragmatic” versus “explanatory” depends upon an evaluation of all components of the trial.<sup>5</sup>

#### Recommended Uses

While explanatory trials are best for determining efficacy, PCTs are well-suited for understanding effectiveness, or whether a clinical intervention works in the real world.<sup>1</sup> It is possible to incorporate pragmatic features into efficacy trials in order to provide evidence that addresses the real-world use. Features that could be incorporated include enrolling patients who more closely reflect the range of real-world patients likely to receive the treatment post-approval, and incorporating a broader range of outcomes, with greater emphasis on functional status, quality of life, and longer-term impacts.<sup>6</sup>

#### Potential Issues

There may be more heterogeneity (lack of uniformity) of treatment effect in PCTs compared to explanatory trials because of broadening of eligibility criteria and inclusion of more practitioners who have differing levels of expertise. This variation in the patient population and practice settings needs to be accounted for in the trial design and analysis. Also, many PCTs aim to include community health centers and clinics that have not traditionally participated in clinical research. Therefore, researchers conducting PCTs will likely have to spend a significant amount of time and resources training individuals at these facilities so that they are capable of participating in these trials.

#### Strength

- PCTs evaluate effects of treatment in real-world settings; a positive result in a PCT can inform practice because it provides evidence that the treatment/intervention is effective in usual practice.<sup>7</sup>

#### Limitations

- A negative result in a PCT cannot provide information on whether the treatment/intervention is effective under optimal conditions.<sup>7</sup>
- Investigators may need to invest more time and resources to conduct PCTs compared to explanatory trials.
- PCTs offer more design and analytic challenges than RCTs due to the potential heterogeneity of treatment effect.

#### Selected Examples

- National Emphysema Treatment Trial (NETT).<sup>4,8-9</sup> This large RCT compared lung volume reduction surgery with standard care for patients with emphysema. The trial had minimal exclusion criteria, included 17 different clinical sites, and included outcomes measures that were relevant to patients, clinicians, and providers: mortality and maximum exercise capacity. Due to the variability in practitioner expertise and the variability in standard of care, a more pragmatic design was used.
- Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) Trial.<sup>10</sup> This large RCT compares different noninvasive diagnostic tests for coronary artery disease. In particular, the trial examines whether for low-intermediate coronary artery risk patients with chest pain, an initial “anatomic” testing strategy (using computed tomographic angiography) is clinically superior to usual care or an initial “functional” stress testing strategy. This trial continues to enroll patients, with a target of reaching 10,000 patients from 150 different clinical sites. Trial endpoints include death, myocardial infarction, major peri-procedural complications, and hospitalization for unstable angina. The trial assesses quality of life, resource use and cost effectiveness, enabling assessment of the real-world impact. In order to ensure generalizability of results to a range of patients, a broad array of practice settings, practitioner specialties, and test types relevant to clinical decision-making are allowed in the study.

## Bibliography

1. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis*. 1967;20:637-648.
2. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*. 2003;290:1624-1632.
3. Luce B, Kramer J, Goodman S, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med*. 2009;151(3):206-209.
4. Mullins CD, Whicher DW, Reese ES, and Tunis S. Generating evidence for comparative effectiveness research using more pragmatic randomized controlled trials. *Pharmacoeconomics*. 2010;28(10):969-976.
5. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol*. 2009;62:464-475.
6. Center for Medical Technology Policy. Methodological guidance for the design of more informative (or pragmatic) pharmaceutical clinical trials: Expert Working Group meeting summary. 21 May 2009; Baltimore, MD. Available at: <http://cmtpnet.org/cmtp-research/guidance-documents/PCT%20Meeting%20Summary%20100609%20no%20watermark.pdf>. Accessed March 28, 2011.
7. Karanicolas PJ, Montori VM, Devereaux PJ, et al. The practicalists' response. *J Clin Epidemiol*. 2009;62:489-494.
8. National Heart, Lung, and Blood Institute. National Emphysema Treatment Trial (NETT): evaluation of lung volume reduction surgery for emphysema. 20 May 2003 [online]. Available at: <http://www.nhlbi.nih.gov/health/prof/lung/nett/lvrsweb.htm>. Accessed March 28, 2011.
9. National Emphysema Treatment Trial Research Group. A randomized trial comparing lung-volume-reduction surgery with medical treatment for severe emphysema. *N Engl J Med*. 2003;348(21):2059-2073.
10. Project information - NIH RePORTER - NIH research portfolio online reporting tools expenditures and results [homepage on the Internet]. Available from: [http://projectreporter.nih.gov/project\\_description.cfm?projectnumber=1R01HL098237-01](http://projectreporter.nih.gov/project_description.cfm?projectnumber=1R01HL098237-01). Accessed March 28, 2011.

## 2.2 Crossover Designs

### Introduction

A crossover design allows patients to act as their own controls, enabling comparisons between and within groups.<sup>1,2</sup> Patients receive a sequence of treatments over successive periods of time, crossing over to an alternative therapy as part of the sequence.<sup>3</sup> At the start of the study, every patient is assigned to a sequence (eg, AB vs. BA), with successive treatments typically separated by a washout period (a specified period of nonuse prior to initiation of therapy).<sup>4</sup> Because all participants receive the same number of treatments, usually in different sequences, it is possible to determine how patient characteristics influence response to treatment.<sup>4</sup> There are several types of crossover designs with different numbers of periods and sequences, and all these designs have their own sets of considerations.<sup>5</sup>

### Considerations for Crossover Designs

When choosing between parallel (traditional) design and crossover design, the following factors that determine the effectiveness of the crossover design should be considered:

*Carryover and period effects on treatment outcomes:* There is a possibility that the effect of a treatment in one period may carry over into the next period.<sup>1,3</sup> These are known as “carryover effects.” Also, during the period of investigation, the disease may naturally progress, regress, or fluctuate in severity.<sup>4</sup> These changes are known as “period effects.” Unless both carryover and period effects are known to be negligible, a crossover design loses its advantages.<sup>4</sup> In order to ensure negligible carryover effects, there is a need to have sufficiently long washout periods between active treatment periods.<sup>1,3</sup>

*Treatment sequencing and patient assignment:* The sequence in which treatments are administered should ideally be assigned randomly. This protects against conscious and unconscious bias by ensuring that there are no systematic differences between patients receiving A/B versus B/A.<sup>4</sup>

*Crossover rules and timing of measurements:* Two types of crossover rules are most commonly used: one in which the treatment switch takes place after a specified length of time (time-dependent), and one in which the treatment switch is determined by the clinical characteristics of the patient (disease-state dependent). These crossover points should be concealed from both patients and observers in order to reduce the influence of carryover effects and period effects.<sup>4</sup>

*Dropouts, faulty data, and other data problems:*<sup>2</sup> Each patient serves as his or her own control; thus, the single patient contributes a large proportion of the total information. For this reason, high dropout rates are a major issue with this type of design.<sup>1</sup> Additionally, dropout rates tend to be higher in crossover designs than in parallel ones, because patients must receive at least two treatments to provide a complete data point. The initial sample size should be sufficiently large to compensate for this effect.

*Statistical analysis and sample size:* The basic unit for statistical analysis is the patient, not an individual measurement.<sup>2</sup> Analyses should be based on paired data (eg, from the same patient), and sample size calculations should consider within-patient variability in outcomes.<sup>1</sup>

### Recommended Uses

Crossover designs are most appropriate to study treatments for stable and chronic diseases.<sup>1</sup> The designs are frequently used to compare new and developmental treatments. They are especially useful when only small differences exist between the new treatment and the standard one, and the effects are very similar.<sup>6</sup> These designs should be used for studies in which the effects of treatments are brief and reversible.<sup>2,7</sup>

### Potential Issues

In order to be effective and valid, crossover designs must be employed only in situations where carryover effects are expected to be minimal or absent, dropout rates are expected to be low, and the disease process is stable.<sup>4</sup> Inappropriate use of a crossover design can lead to biased effect estimates, and therefore to incorrect conclusions about the comparative effectiveness of therapies.

### Strengths

- Crossover design removes between-patient variation.<sup>3</sup>
- It requires fewer patients than a parallel study for an equal number of treatment comparisons, because each experimental unit (ie, patient) can be used several times.<sup>5</sup> This is an economical use of resources.<sup>3</sup>
- Patients can indicate preferences for one treatment versus another, because patients receive multiple treatments in a single crossover study.<sup>1</sup>



## Limitations

- There is no guarantee that washout periods will completely control for carryover effects.<sup>5</sup>
- Long washout periods might unavoidably increase the duration of the experiment.<sup>3</sup>
- Ethical concerns (how long can a patient be refused treatment during a washout period) and incomplete knowledge (what washout period length is sufficient) may sometimes lead to inadequate washout periods.<sup>3</sup>
- Within each unit, or patient, responses to therapy are likely to be correlated (eg, a single patient's response to treatment A is correlated with that patient's response to treatment B; the responses are not independent). This causes complexities in both the design and the analysis.<sup>5</sup>

## Selected Examples

- Johansson, et al. *Cancer Treat Rev.* (1982).<sup>8</sup> The authors conducted a double-blinded, two-period crossover trial to compare the therapeutic and adverse effects of two oral anti-emetics, nabilone versus the then-conventional prochlorperazine, in 27 adult cancer patients undergoing concurrent chemotherapy, and suffering from uncontrolled vomiting and nausea, despite use of traditional anti-emetics. The sequence of treatment periods (A/B or B/A) was assigned at random, and the crossover rule was based on the underlying pattern of chemotherapy: patients would receive the same dosage of the same chemotherapy drugs for two consecutive cycles before crossover. Similarly, since the investigational anti-emetics were administered in coordination with chemotherapy, the washout period corresponded to the break time between treatments in the chemotherapy regimen. The researchers identified a significant period effect (ie, fluctuations in disease severity or progression), with more severe vomiting scores recorded during the second treatment period. Nevertheless, the analysis clearly led to the conclusion that nabilone was significantly more effective than prochlorperazine at reducing chemotherapy-induced nausea in patients refractory to anti-emetic therapy. While the authors were not explicit regarding why a crossover design was chosen, the study demonstrates the efficiency of the crossover design—the authors were able to reach a statistically significant conclusion utilizing a small number of patients.
- Koke, et al. *Pain.* (2004).<sup>9</sup> The authors conducted a single-blinded, randomized, crossover trial to compare hypoalgesic effects of three different methods of transcutaneous electrical nerve stimulation (TENS), varying in frequency and intensity, among 180 adult chronic pain patients not taking any other treatment for pain. TENS treatments were administered twice daily for two weeks and were separated by a two-week washout period. Outcomes were assessed at the beginning and end of each two-week period, with severity of pain measured using a validated visual analogue scale (VAS). The crossover design was chosen due to the heterogeneity (lack of uniformity) among the chronic pain population. Each patient served as his or her own control, reducing heterogeneity and increasing study power.
- Maggiore, et al. *Am J Kidney Dis.* (2002).<sup>10</sup> While many reports note that the use of cool dialysate has a protective effect on blood pressure during hemodialysis treatments, formal clinical trials are lacking. To account for within-patient differences in hemodialysis outcomes, a randomized, open trial with a crossover design was used to compare the effect of two different procedures for thermal balance on the frequency of hemodialysis sessions complicated by symptomatic hypotension. The investigational procedure, isothermic dialysis, was compared against the control, thermoneutral dialysis, in 116 adults who had been on standard hemodialysis treatment for three months or longer with demonstrated risk of symptomatic hypotensive episodes. Patients were kept on standard hemodialysis for a run-in period of one week and then randomized to one of two treatment period sequences, A/B or B/A, each treatment period lasting four weeks. The authors assumed no carryover effect from one treatment period to the next.

## Bibliography

1. Mills EJ, Chan A, Wu P, et al. Design, analysis and presentation of crossover trials. *Trials.* 2009;10(27):1-6.
2. Elbourne DR, Altman DG, Higgins JPT, et al. Meta-analyses involving cross-over trials: methodological issues. *Int J Epidemiol.* 2002;31:140-149.
3. Yang M, Stufken J. Optimal and efficient crossover designs for comparing test treatments to a control treatment under various models. *J Stat Plan Inference.* 2008;138(1):278-285.
4. Louis TA, Lavori PW, Bailar JC, et al. Crossover and self-controlled designs in clinical research. *N Engl J Med.* 1984;310:24-31.
5. Matthews JNS. Recent developments in crossover designs. *Int Stat Rev.* 1988;56(2):117-127.
6. Cleophas TJ, de Vogel EM. Crossover studies are a better format for comparing equivalent treatments than parallel group studies. *Pharm World Sci.* 1998;20:113-117.
7. Cleophas TJ. A simple method for the estimation of interaction bias in crossover studies. *J Clin Pharmacol.* 1990;30:1036-1040.
8. Johansson R, Kijku P, Groenroos M. A double-blind, controlled trial of nabilone vs. prochlorperazine for refractory emesis induced by cancer chemotherapy. *Cancer Treat Rev.* 1982;9(Suppl B):25-33.
9. Koke A, Schouten J, Lamerichs-Geelen M, et al. Pain reducing effect of three types of transcutaneous electrical nerve stimulation in patients with chronic pain: a randomized crossover trial. *Pain.* 2004;108:36-42.
10. Maggiore Q, Pizzarelli F, Santoro A, et al. The effects of control of thermal balance on vascular stability in hemodialysis patients: results of the European randomized clinical trial. *Am J Kidney Dis.* 2002;40(2):280-290.

## 2.3 N of 1 Randomized Controlled Trials (RCTs)

### Introduction

N of 1 trials are individual randomized, double-blinded crossover trials that compare two or more treatment options for a single patient.<sup>1</sup> In such a study, the patient undergoes multiple sequential treatment periods during which an active intervention is paired with a matched placebo or an alternative therapy. For each period, the order of administration of the active therapy or comparator is assigned randomly, such as by a coin toss, and ideally both the patient and clinician are blind to the assignment. Appropriate outcomes (those that are of interest to and readily reported by the patient) are often measured through the use of a diary or questionnaire.<sup>1-5</sup>

### Characteristics of N of 1 Trials

Several essential characteristics required for the conduct of N of 1 trials have been identified:<sup>5</sup>

- The condition for which the intervention is being used is chronic and relatively stable.
- The half-life of the intervention is relatively short, or the modification is reversible.
- There is a rapid onset and offset of action for the intervention.
- The effects of the intervention can be measured using a validated outcome measure.
- The intervention does not alter the underlying cause of the condition.

Clinicians have been slow to adopt this method within their everyday clinical practice. This could be due to a number of reasons, including time constraints, costs of design and implementation, and a general lack of awareness of this type of research.

### Data Analysis and Interpretation

The simplest method for interpreting the resulting data in N of 1 trials is to plot it on a chart and visually inspect it. While this method is subject to bias, it can be convincing in cases where the difference in effect between the active intervention and the comparator is pronounced. Simple statistical methods can also be used to interpret results.

It is also possible to combine the results from multiple N of 1 trials if the studies are investigating the same sets of interventions. Some studies have pooled data from multiple N of 1 trial analyses, and combined results using a variety of statistical modeling techniques. These types of analyses can allow researchers to apply the individually beneficial N of 1 trial results to population-based research, increasing their generalizability.<sup>4</sup>

### Recommended Uses

N of 1 trials are often used by clinicians when they are faced with therapeutic uncertainty for a particular patient. This approach offers an alternative for making individualized treatment decisions based on objective data, patient values, and patient-centered outcomes.<sup>6</sup>

### Potential Issues

Similar issues arise as those associated with crossover designs (see section 2.2). N of 1 trials should only be used where the disease state is considered stable and carryover effects are assumed to be absent or negligible. Inappropriate use of this design may compromise the accuracy of results. Further, an N of 1 trial becomes increasingly less feasible to conduct as the time required to observe treatment effects increases.

### Strengths

- N of 1 trials allow physicians to individualize treatments in clinical practice.
- Physicians and patients engage in shared decision-making.
- Some evidence exists that patients participating in N of 1 trials have enhanced knowledge and awareness of their condition, which may lead to greater adherence to treatment and better disease management.<sup>1</sup>
- N of 1 trials may enhance researcher understanding of the connection between population-based studies, individual characteristics, and their respective treatment responses.
- Costs of N of 1 trials are considerably less than for traditional RCTs.<sup>2</sup>

### Limitations

- N of 1 trials may be time-consuming and potentially expensive for an individual physician, particularly within the primary care setting where there may be a lack of administrative experience.
- The trial requires collaboration with experienced pharmacy colleagues for the preparation of the matching placebos of the trial.<sup>1,3,5</sup>

### Selected Examples

- Zucker, et al. *J Rheumatol.* (2006).<sup>4</sup> The authors conducted a study in which the results of multiple N of 1 trials were combined in order to compare different fibromyalgia therapies, and to assess the feasibility and outcomes of this type of study design for practice-based

effectiveness research. The primary outcome across the multiple N of 1 trials was the score on the Fibromyalgia Impact Questionnaire (FIQ), a validated tool to assess quality of life in fibromyalgia patients. Eight rheumatologists enrolled 58 patients in individual randomized, double-blind, multi-crossover, N of 1 trials comparing a single drug and a combination therapy. A central pharmacy was used to prepare the random-order treatment kits. As noted above, the FIQ scores were pooled and compared to analogous outcomes from a traditional crossover fibromyalgia trial using a two-sided t-test.

- Louly, et al. *Clin Ther.* (2009).<sup>7</sup> The authors investigated the effectiveness of Tramadol 50 mg compared with a placebo; a double-blind N of 1 RCT was conducted on a patient with rheumatoid arthritis and interstitial lung disease who developed a chronic dry cough. The drug and placebo were prepared as identical capsules by an external researcher who had no contact with the patient and no knowledge of the study results.

## Bibliography

1. Madhok V, Fahey T. N-of-1 trials: an opportunity to tailor treatment in individual patients. *Br J Gen Pract.* 2005;55:171-172.
2. Guyatt GH, Keller JL, Jaeschke R, et al. The N of 1 randomized controlled trial: clinical usefulness. Our three-year experience. *Ann Intern Med.* 1990;112:293-299.
3. Guyatt G, Sackett D, Adachi J, et al. A clinician's guide for conducting randomized trials in individual patients. *Can Med Assoc J.* 1988;139:497-503.
4. Zucker DR, Ruthazer R, Schmid CH, et al. Lessons learned combining N-of-1 trials to assess fibromyalgia therapies. *J Rheumatol.* 2006;33:2069-2077.
5. Scuffham PA, Nikles J, Mitchell GK, et al. Using N-of-1 trials to improve patient management and save costs. *JGIM.* 2010;25:906-913.
6. Tsapas A, Matthews DR. Using N-of-1 trials in evidence-based clinical practice. *JAMA.* 2009;301:1022-1023; author reply 3.
7. Louly PG, Medeiros-Souza P, Santos-Neto L. N-of-1 double-blind, randomized controlled trial of tramadol to treat chronic cough. *Clin Ther.* 2009;31:1007-1013.

## 2.4 Cluster Randomized Controlled Trials (RCTs)

### Introduction

Cluster RCTs are studies in which patients are grouped (clustered) on the basis of geography of practice (eg, caregivers, hospitals, communities), and then randomized as a group to either the intervention or control arm.<sup>1-4</sup> Community-based cluster RCTs are generally characterized by small numbers of clusters (eg, hospitals) with a large number of patients in each cluster.<sup>5</sup> They are viewed as a pragmatic methodology to measure the effectiveness of an intervention on a large scale.<sup>4</sup>

### Recommended Uses

Cluster RCTs have commonly been used to evaluate the delivery of healthcare services, the effects of educational interventions, or the effects of organizational changes.<sup>1,6</sup> More recently, researchers have suggested that cluster RCTs can be used to determine the comparative effectiveness of two or more therapies where there exists a natural variation in practice patterns.<sup>7</sup> Longitudinal studies sometimes use cluster techniques to study the impact of interventions on the same group of patients over time.<sup>3</sup>

One reason to use cluster RCTs is to avoid “contamination” between those patients receiving the intervention and those who are not. Such contamination may weaken the estimate of treatment effect.<sup>3,6</sup> For example, individuals who are part of a study of behavioral intervention to reduce smoking or prevent coronary heart disease may share advice, and skew the effects of the intervention.<sup>2,6</sup>

Cluster RCTs are also recommended when randomization at the level of the patient is not practical or desirable. For example, in a study in which the goal is to evaluate the impact of new clinical practice guidelines, the investigational “intervention” takes place at the level of the caregiver or hospital, not at the patient level. Nevertheless, patient outcomes can be affected by the implementation of new guidelines. In this case, randomization at the patient level is inappropriate for the research question. Instead, the new practice guidelines are implemented in entire hospitals or clinics, and each of these study sites is randomized as one group to the trial. This example provides an illustration of why cluster designs are often used for interventions that involve education of healthcare professionals.<sup>5</sup>

### Potential Issues

Design and analysis of cluster RCTs is far more complex than for individually randomized trials. Data points from patients within a cluster tend to be correlated, and this correlation must be accounted for in both study design and analysis.<sup>8</sup> There are also differences in the randomization process between cluster RCTs and traditional RCTs. As individuals are consented after randomization in cluster RCTs, there is the potential for selection bias if those that subsequently consent are different than those that refuse to consent.<sup>8</sup>

### Strengths

- The cluster RCT evaluates the real-world effectiveness of an intervention as opposed to efficacy. It may be useful for comparative effectiveness research because the focus is on understanding the effects of an intervention in a pragmatic, real-world setting.
- It provides an alternative methodology for assessing the effectiveness of therapies in settings where randomization at the individual level is inappropriate or impossible.

### Limitations

- The complex design of the cluster RCT requires adequate understanding, implementation, and proper reporting by researchers. In order to ensure accurate interpretation of cluster RCTs by readers, the Consolidated Standards of Reporting Trials (CONSORT) statement for reporting individually randomized trials was extended to include guidelines for reporting of cluster RCTs in 2004.<sup>1,8</sup>
- The cluster RCT design requires a greater number of patients compared to individual RCT designs because of intracluster correlation.<sup>1</sup> In order to obtain statistical power equivalent to that of individual randomization designs, nonstandard sample size approaches must be applied to avoid statistically underpowered studies (eg, inflation using a calculated intracluster correlation coefficient).<sup>2,9</sup>
- Autonomous patient decision-making can be jeopardized in cluster designs, as patients may not have access to treatments or procedures available only at other locations. Informed consent, then, represents an especially complex question in cluster RCTs. Cluster designs should not be used in circumstances where the same information could be reliably obtained through individual randomization.<sup>7,10-11</sup>

### Selected Examples

- Bass, et al. *Can Med Assoc J* (1986).<sup>12</sup> To evaluate an intervention to enhance the effectiveness of hypertension screening and management in general practice, the authors randomized 34 physicians' private practices to each of two intervention groups. The study objective was to determine the effectiveness of a new system of care in which a medical assistant oversees the screening and attends to the education, compliance, and follow-up of hypertensive patients, compared to a system in which patients are simply tested and prescribed the appropriate medication with no follow-up care or management. Seventeen physician practices, comprising 15,659 patients, were matched with a control group of 17 physician practices, with 16,465 patients. Given the nature of the intervention under study (a new system of care) randomization at the patient level was impractical and a cluster design was chosen.

- World Health Organisation European Collaborative Group. *Lancet* (1986).<sup>13</sup> The collaborative designed a trial to evaluate the effects of a multi-factorial prevention strategy for coronary heart disease (CHD). This study included 80 factories consisting of a sample population of 60,881 men. The unit of randomization was the factory. Outcomes measured included the reduction in total and fatal CHD, as well as non-fatal myocardial infarction and total deaths. A cluster RCT was chosen due to the educational component of the intervention; the design avoided contamination (eg, if randomization at the patient level had been done, patients participating in the trial might have shared educational materials with those not participating, and skewed the effects).
- Zucker, et al. *Control Clin Trials*. (1995).<sup>14</sup> The authors designed a large scale community health trial for the prevention of cardiovascular disease among children and adolescents; interventions were implemented on a school-wide basis, so that each student in a school was assigned the same intervention. The study question was whether a behavior-focused cardiovascular health-education program could produce positive results in elementary school children. The trial involved 96 elementary schools, 40 of which were part of the control arm, 28 of which had only a school-based intervention program, and 28 of which implemented both a school-based and family-based intervention. The cluster RCT design was chosen because of the educational nature of the intervention.
- Platt, et al. *Med Care*. (2010).<sup>15</sup> The authors reported from an ongoing cluster RCT comparing the effectiveness of three screening and preventive measures to reduce methicillin-resistant *Staphylococcus aureus* (MRSA) infection in intensive care units. The unit of randomization chosen in this trial was the hospital. This was the practical way to control for contamination of the trial across the treatment arms, as the screening and preventive measures were implemented hospital wide.

## Bibliography

1. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ*. 1998;317:1171-1172.
2. Donner A. Some aspects of the design and analysis of cluster randomization trials. *J Roy Stat Soc C-App*. 1998;47:95-113.
3. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. *J Eval Clin Pract*. 2005;11:479-483.
4. Eldridge S, Ashby D, Bennett C, et al. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336:876-880.
5. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med*. 2007;26:2-19.
6. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. *BMJ*. 1999;318:1407-1409.
7. Sabin JE, Mazor K, Meterko V, et al. Comparing drug effectiveness at health plans: the ethics of cluster randomized trials. *Hastings Cent Rep*. 2008;38:39-48.
8. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ*. 2004;328:702-708.
9. Kerry S, Bland M. Cluster randomization. *Br J Gen Pract*. 1998;48(431):1342.
10. Mazor KM, Sabin JE, Boudreau D, et al. Cluster randomized trials: opportunities and barriers identified by leaders of eight health plans. *Med Care*. 2007;45:S29-37.
11. Taljaard M, McRae AD, Weijer C, et al. Inadequate reporting of research ethics review and informed consent in cluster randomized trials: review of random sample of published trials. *BMJ*. 2011;342:d2496.
12. Bass MJ, McWhinney IR, Donner A. Do family physicians need medical assistants to detect and manage hypertension? *Can Med Assoc J*. 1986;134:1247-1255.
13. World Health Organisation European Collaborative Group. European Collaborative Trial of Multifactorial Prevention of Coronary Heart Disease: final report on the 6-year results. *Lancet*. 1986;1:869-872.
14. Zucker DM, Lakatos E, Webber LS, et al. Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization. *Control Clin Trials*. 1995;16:96-118.
15. Platt R, Takvorian S, Septimus E, et al. Cluster randomized trials in comparative effectiveness research: randomizing hospitals to test methods for prevention of healthcare-associated infections. *Med Care*. 2010; 48(6 Suppl): S52-57.

## 2.5 Delayed-start Designs

### Introduction

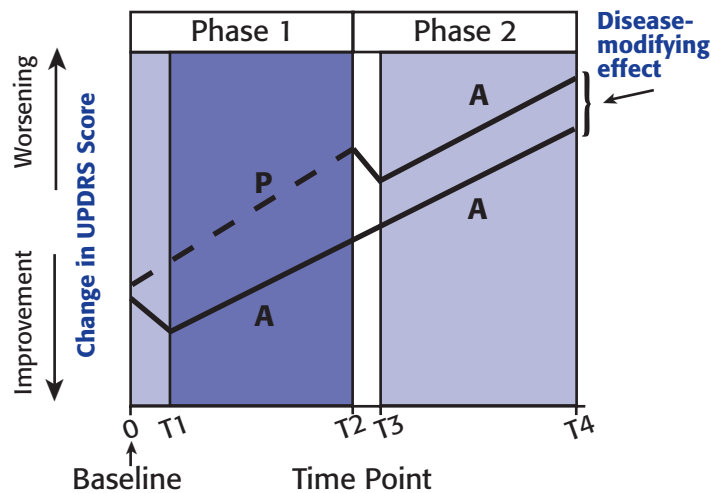
Delayed-start designs, or randomized-start designs, allow studies to determine whether an intervention acts by reducing symptoms or by modifying disease, an important distinction for slowly progressing diseases such as Parkinson's disease.<sup>1,2</sup>

### Study Design

A delayed-start study is conducted in two phases. In Phase I, patients are randomized to receive either the active treatment or a control treatment. Phase I should be long enough in duration to allow the effects of the treatment on the disease symptoms to be fully captured. At the end of Phase I, any significant differences between the two groups (if present) reflect differences in symptomatic effects, disease modifying effects, or both. In Phase II, the patients in the control group are switched to receive a "delayed-start" of the active treatment, so that every patient in the study now receives the active treatment.<sup>1,2</sup> Patients are then followed for a predetermined length of time. Any beneficial symptomatic effects should be equal in the two groups, and any remaining differences must be attributable to the disease-modifying effect of the treatment.

Figure 1 is the diagrammatic representation of a delayed-start study design.<sup>1</sup> Patients are randomized to receive the allotted active or control treatment in Phase I, and followed from point 0 to T2. Data are collected from the time period T1 to T2 and are used to estimate the effect of the agent on symptoms, and for indications of possible disease-modifying effect. All patients receive the active treatment in Phase II (T2 to T4). Data collected from the time period T3 to T4 are used to estimate the disease-modifying effect of the agent. Data from the time periods 0 to T1 and T2 to T3 may reflect transitory responses and are usually not used in the analysis (UPDRS is the Unified Parkinson's Disease Rating Scale; A represents active treatment, and P represents placebo).

**Figure 1: Delayed-start Design\***



\*Figure 1. From: D'Agostino RB. The delayed-start study design. *N Engl J Med.* 2009;361(13):1304-1306.

### Recommended Uses

Delayed-start study designs are useful when there is a need for a clinical end-point that reliably measures disease progression and is not confounded by the study intervention's effects on symptoms.<sup>2</sup> The main advantage of delayed-start study design is that it helps separate the disease-modifying effects of an administered treatment from its short-term beneficial effects on symptoms. In short, this study design controls for confounding by the symptomatic effect of the study intervention.<sup>1-3</sup>

Delayed-start designs are also useful when there is a need for prolonged study duration.<sup>2</sup> For example, the neuroprotective effects of a treatment cannot be expected to manifest rapidly in slowly progressive diseases like Parkinson's or Alzheimer's.<sup>2</sup> This design can be used to study treatments for diseases that progress slowly, but are ultimately debilitating, like Alzheimer's disease, Parkinson's disease, rheumatoid arthritis, and chronic obstructive pulmonary disease (COPD).<sup>1</sup>

Delayed-start study designs have also been proposed as a useful strategy to evaluate staggered implementation of certain policy changes, such as outcomes associated with restrictions to treatment coverage.<sup>4</sup>

## Potential Issues

Planning and designing a delayed-start design study requires a strong statistical background and knowledge of the disease mechanism. Careful determination must be made of the duration of Phase I, the duration of the “data-not-used-zones” (reflected in Figure 1 as the transition time periods of 0 to T1, and T2 to T3), the number of repeated measurements to be taken during phases, the appropriate analytic method, and the appropriate statistical tests.<sup>1,2</sup>

Missing data can also be a significant issue. Participants in the control group are more likely to drop out of the study due to lack of treatment effect, leading to a differential dropout rate between the treatment group and the control group. The protocol should specify appropriate statistical methods that can control for this issue. Covariate analysis and propensity scores (see section 5.4) can also be used to check for an imbalance between treatment and control groups. In all cases sensitivity analysis (see section 5.6) should be performed.

## Strengths

- Delayed-start study design separates the disease-modifying effects of administered treatment from short term beneficial effects on symptoms.<sup>1-3</sup>
- The study design also addresses ethical concerns raised with respect to RCTs. More patients receive the active intervention as compared to those in a traditional trial.<sup>3</sup> All participants eventually receive the potentially beneficial medical intervention, while a control group is maintained in the initial phase.<sup>4</sup>

## Limitations

- Delayed-start design requires sufficient understanding of the study design and clinical progression of the disease to define adequate Phase I and Phase II durations, and statistical methodology to address analytical considerations.
- Only the first half of the study is considered double-blind; the second half is open-label, a limitation that may introduce bias through unblinding.<sup>5</sup>
- The delayed-start design study may encounter enrollment issues; it needs to recruit patients who are willing to be off the symptomatic therapy for the first half of the study if they are randomized to the control arm. In the case of Parkinson's disease, the majority of patients require symptomatic treatment at the time of diagnosis, and thus they are not suitable for participation in delayed-start studies.
- Only patients with mild, early, and more slowly progressive disease may be eligible for this type of study; study findings may not be generalizable to patients with more advanced stages of disease.<sup>2-3,5</sup>
- The studies are susceptible to high dropout rates and patient discontinuation in the Phase I placebo group, because these patients do not experience any treatment effects. Differential baseline characteristics between patients in Phase II and discontinued patients may introduce confounding, and compromise results.

## Selected Examples

- Parkinson Study Group. *Arch Neurol.* (2004).<sup>6</sup> The TVP-1012 in Early Monotherapy for Parkinson's Disease Outpatients (TEMPO) Study compared effectiveness of early versus later initiation of rasagiline on the progression of disability in patients with Parkinson's disease. A double-blind, parallel-group, randomized, delayed-start clinical trial was implemented in which 404 individuals with early Parkinson's disease, not requiring dopaminergic therapy, were enrolled at 32 sites in the United States and Canada. Participants were randomized to receive 1 mg or 2 mg per day of rasagiline for one year or placebo for six months, followed by 2 mg per day of rasagiline for six months.
- Hauser, et al. *Movement Disorders.* (2009).<sup>7</sup> TEMPO Open-Label Extension Study: The study compared the long-term clinical outcome of early versus delayed rasagiline treatment in early Parkinson's disease. The study was a long-term extension of the TVP-1012 study described above. Patients were randomly assigned to initial treatment with rasagiline (early-start group) or placebo for six months, followed by rasagiline (delayed-start group), in the TVP-1012 in Early Monotherapy for Parkinson's Disease Outpatients (TEMPO) Study. The Tempo Open-Label Extension Study described the results of the long-term open-label extension of the TEMPO study, in which patients were treated with rasagiline for up to 6.5 years.
- Olanow, et al. [Several publications]. (2008-2011).<sup>8-10</sup> Attenuation of Disease Progression with Azilect Given Once-Daily (ADAGIO) Study: The study examined the potential disease-modifying effects of rasagiline in Parkinson's Disease. This was a double-blind trial, where a total of 1,176 patients with untreated Parkinson's disease were randomly assigned to receive rasagiline (at a dose of either 1 mg or 2 mg per day) for 72 weeks (the early-start group); or a placebo for 36 weeks, followed by rasagiline (at a dose of either 1 mg or 2 mg per day) for 36 weeks (the delayed-start group). To determine a positive result with either dose, the early-start treatment group had to meet each of the three hierarchical end points of the primary analysis, based on the UPDRS: superiority of early-start treatment to placebo in the rate of change in the UPDRS score between weeks 12 and 36, superiority to delayed-start treatment in the change in the UPDRS score between the baseline and week 72, and non-inferiority to delayed-start treatment in the rate of change in the score between weeks 48 and 72.

## Bibliography

1. D'Agostino RB. The delayed-start study design. *N Engl J Med*. 2009;361:1304-1306.
2. Delayed-Start Trials (Part 1): Design and Interpretation. Available from: [http://www.neura.net/images/pdf/Neura\\_V10I1\\_Delayed.pdf](http://www.neura.net/images/pdf/Neura_V10I1_Delayed.pdf). Accessed March 22, 2011.
3. Clarke CE. Are delayed-start design trials to show neuroprotection in Parkinson's disease fundamentally flawed? *Movement Disorders*. 2008;23(6):784-789.
4. Maclure M, Carleton B, Schneeweiss S. Designed delays versus rigorous pragmatic trials. Lower carat gold standards can produce relevant drug evaluations. *Med Care*. 2007;45:S44-S49.
5. Ahlskog JE, Uitti RJ. Rasagiline, Parkinson neuroprotection, and delayed-start trials: still no satisfaction? *Neurology*. 2010;74:1143-1148.
6. Parkinson Study Group. A controlled, randomized, delayed-start study of rasagiline in early Parkinson disease. *Arch Neurol*. 2004;61:561-566.
7. Hauser RA, Lew MF, Hurtig HI. Long-term outcomes of early versus delayed rasagiline treatment in early Parkinson's disease. *Movement Disorders*. 2009;24(4):564-573.
8. Olanow CW, Hauser RA, Jankovic J, et al. A randomized, double-blind, placebo-controlled, delayed start study to assess rasagiline as a disease modifying therapy in Parkinson's disease (the adagio study): rationale, design, and baseline characteristics. *Movement Disorders*. 2008;23(15):2194-2201.
9. Olanow CW, Rascol O, Hauser R. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. *N Engl J Med*. 2009;361:1268-1278.
10. Olanow CW, Rascol O, Hauser R. Correction to "A double-blind, delayed-start trial of rasagiline in Parkinson's disease." *N Engl J Med*. 2011;364:1882.



## 3. EXPERIMENTAL METHODS

### 3.1 Adaptive Designs and Bayesian Methods

#### Introduction

An adaptive design “uses accumulating data to decide how to modify aspects of the trial as it continues, without undermining the validity and integrity of the study.”<sup>1-3</sup> The ways in which the trial or statistical procedures may be modified is typically defined prospectively, but can also occur during the trial.<sup>4</sup> Adaptive designs include pre-specified interim points or phases in the protocol where investigators can reevaluate hypotheses or planned statistical procedures. At every planned phase, data analysis is performed and the updated information is utilized to make certain adaptations.<sup>5</sup>

#### Categories of Adaptation

Adaptations are classified into three major categories:<sup>4</sup>

*Prospective adaptations:* These design adaptations are pre-specified in the protocol before the trial begins. Some examples include adaptive randomization; stopping a trial early due to safety, futility, or efficacy at interim analysis; or dropping inferior treatment groups.

*Concurrent adaptations:* These are changes that are incorporated as the trial continues. Concurrent adaptations are not specified a priori, but are incorporated when the need arises, such as modifying inclusion/exclusion criteria, evaluability criteria (defining what it means to be “treated”), dose regimen, or treatment duration; or changing hypothesis or study end points.

*Retrospective adaptations:* These adaptations are typically made to the statistical analysis plan. They can occur after data collection, but prior to unblinding of treatment modalities.

The US Food and Drug Administration guidance document on Adaptive Trial Design recommends that all adaptations should be specified during the trial design phase, before the trial is implemented, so that the Type I error rate (the probability of rejecting the null hypothesis when it is true) can be appropriately calculated.<sup>2-3</sup> A priori planning addresses the concern that the use of adaptive designs may alter the trial until it no longer addresses the question under consideration.<sup>4</sup> Nevertheless, when ad hoc adaptations occur in practice, they proceed through amendments to the protocol.<sup>4</sup>

#### Bayesian Methods in Adaptive Design

While adaptive studies can use either non-Bayesian or Bayesian analytic techniques, Bayesian statistics are especially suited to adaptive studies. Adaptive designs utilizing Bayesian approaches take into consideration all the available information for the scientific question under consideration.<sup>6</sup> As opposed to traditional approaches, the Bayesian approach uses both the evidence that accumulates over time (eg, interim clinical trial data)<sup>7</sup> and prior information (eg, literature or previous relevant studies).<sup>8</sup> Prior information is combined with the current data to obtain a “posterior distribution” (a term for the probability of an event that is conditional on the prior information and current data) which is analyzed and used to draw conclusions.

#### Recommended Uses

This method can be particularly useful because it can incorporate newly available, high-quality evidence into the study design, allowing for reduction in the sample size, time, and cost needed to acquire the information most relevant to decision-makers.<sup>9</sup>

Since the design allows for adjustments as one learns new information during the trial, it can be useful in studying rare diseases or pediatrics, where there is limited knowledge in the field.<sup>10</sup> For example, when the efficacy of various treatment groups is unknown, adaptive designs (eg, drop-the-loser design) can use interim results to inform as to when to “drop” an inferior treatment arm, as predetermined in the protocol. Patients can then be randomized to the remaining “superior” groups.<sup>4</sup>

The most commonly acceptable adaptive designs include response adaptive randomization in Phase II product trials, blinded sample size re-estimation, and early stoppage of the trial for safety, efficacy, or futility.

#### Potential Issues

One should avoid building multiple adaptations into a single trial, as this may increase the complexity of the trial and introduce difficulty in interpreting the results.<sup>11</sup> When designing such a trial, premature release of interim results to the general public should be avoided, as this will endanger the integrity of the trial.<sup>11</sup>

Further, Bayesian methods, if employed, are complicated and require substantial effort with respect to both design and conduct of trial, and clear explanation to avoid misinterpretation of results.<sup>12</sup>

## Strengths

- Adaptive study design allows flexibility to redesign clinical trials at interim stages,<sup>13</sup> to decrease the patient exposure to harm, and to reduce trial duration and costs.<sup>14</sup>
- Investigators are enabled to identify and rectify inappropriate assumptions about certain planning parameters, such as the treatment-effect size, that were not well understood at the design stage of the trial.<sup>15</sup>

## Limitations

- Adaptations may introduce bias and make analysis and interpretation challenging.<sup>4</sup> Design may also introduce bias if blinding is compromised during interim analyses.<sup>11</sup> To avoid introducing bias, investigators should take precaution by keeping treatment allocation blinded, or by having independent third parties conduct the analyses.<sup>15</sup>
- Multiple interim examinations of the data during the trial increase the probability of Type I error, a false positive, which reduces the power of statistical comparisons.<sup>11</sup> Methods to control the Type I error rate should be pre-specified and controlled.<sup>15</sup>
- Statistical measures to assess confidence in the study results (eg, p-values and the corresponding confidence intervals) for treatment effects may not be reliable because of the inability to preserve the Type I error rate.<sup>4</sup>
- The logistics of adaptive trials are more complicated than those of standard trials.<sup>16</sup>

## Selected Example

- Nelson. *J Natl Cancer Inst.* (2010).<sup>17</sup> In an adaptive Phase II trial of interventions for advanced non-small cell lung cancer (NSCLC), researchers tested differential responses to anti-tumor agents in relation to a series of tumor biomarker subtypes. In the first phase of the study, the researchers took fresh core needle biopsy samples of tumors and analyzed them for four biomarker subtypes. They then randomly assigned 40% of the study population (N=255) to four different treatment groups, each one receiving a drug having a mechanism of action known to be related to one of the biomarkers. (For example, one of the biomarkers was an epidermal growth factor receptor [EGFR] genetic marker called *KRAS/BRAF*, while one of the therapies used in the study was erlotinib, an EGFR inhibitor.) In the second phase, which was the adaptive segment of the trial, the patients' response to treatment was correlated with their tumor biomarker status. This information was used for decision-making on treatment assignments for the remaining 60% of the patient population. These treatment decisions were implemented in the third phase of the study. Overall, using the biomarker-guided treatment allocation determined by the adaptive phase of the study, 46% of the patients had achieved stable disease after eight weeks. Only 30% of advanced NSCLC patients receiving traditional chemotherapy achieved stable disease after eight weeks.

## Bibliography

1. Gallo P, Chuang-Stein C, Dragalin V, et al. Adaptive design in clinical drug development - an executive summary of the PhRMA Working Group. *J Biopharm Stat.* 2006;16(3):275-283.
2. US Department of Health and Human Services. Food and Drug Administration. Guidance for Industry. *Adaptive Design Clinical Trials for Drugs and Biologicals.* Washington, DC: US Department of Health and Human Services. Food and Drug Administration; February 2010.
3. Berry DA. Adaptive clinical trials: the promise and the caution. *J Clin Oncol.* 2011;29(6):606-609.
4. Chow SC, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis.* 2008;3:11.
5. Chang M, Chow S, Pong A. Adaptive design in clinical research: issues, opportunities and recommendations. *J Biopharm Stat.* 2006;16:299-309.
6. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov.* 2006;5:27-36.
7. Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med.* 2009;151:206-209.
8. Tunis SR. *Implementing Comparative Effectiveness Research: Priorities, Methods, and Impact - Strategies to Improve Comparative Effectiveness Research Methods and Data Infrastructure.* Washington, DC: Engelberg Center for Health Care Reform at Brookings; June 2009:35-54.
9. Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med.* 2009;151:206-209.
10. Wang M, Wu YC, Tsai GF. A regulatory view of adaptive trial design. *J Formos Med Assoc.* 2008;107(12 Suppl):S3-S8.
11. Coffey CS, Kairalla JA. Adaptive clinical trials: progress and challenges. *Drugs RD.* 2008;9(4):229-242.
12. Thall PF, Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *Eur J Cancer.* 2007; 43:859-866.
13. Baiardi P, Giaquinto C, Giroto S. Innovative study design for pediatric clinical trials. *Eur J Clin Pharmacol.* 2011;67(Suppl 1):109-115.
14. Sietsema W, Sennewald E. An introduction to adaptive clinical trial designs. *Regulatory Rapporteur.* 2010; 7(10):4-6. Available from: [http://www.topra.org/sites/default/files/focus1\\_17.pdf](http://www.topra.org/sites/default/files/focus1_17.pdf). Accessed August 4, 2011.
15. Vandemeulebroecke M. Group sequential and adaptive designs – a review of basic concepts and points of discussion. *Biom J.* 2008;4:541-557.
16. Gaydos B, Anderson K, Berry D, et al. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Inf J.* 2009;43:539-556.
17. Nelson, NJ. Adaptive clinical trial design: has its time come? *J Natl Cancer Inst.* 2010;102(16):1217-1218.

## 4. NONEXPERIMENTAL STUDY DESIGNS

### 4.1 Cohort and Case-Control Studies

#### Introduction

Nonexperimental studies are studies in which patients are not randomized to a treatment or intervention; patients and physicians make real world treatment decisions, and patterns of care and outcomes are observed. A nonexperimental study design may be used for many types of studies, such as descriptive, effectiveness, comparative effectiveness, or safety studies.

#### Types of Nonexperimental Studies

Cohort and case-control are two types of nonexperimental study designs that can be used for comparative effectiveness research.

##### *Cohort Studies*

Cohort studies follow a group of patients with a common condition(s) or exposure(s) over time. Cohort studies may include only patients with the condition or exposure of interest, or they may include a comparison group of patients who do not have the condition or exposure of interest. In a typical comparative effectiveness cohort study, patients who are exposed to the interventions of interest and who meet inclusion and exclusion criteria are enrolled in the study. Patients are followed for a period of time, and data on outcomes of interest are collected. For example, a comparative effectiveness cohort study may enroll patients who are using two types of cholesterol-lowering medications and follow them for a period of one year to assess reduction in cholesterol and other outcomes of interest, such as acute myocardial infarction (AMI).

##### *Case-Control Studies*

Case-control studies enroll patients who have experienced a particular outcome of interest (“cases”), and patients who have not experienced the outcome of interest but who are representative of the same population as the cases (“controls”).<sup>2</sup> For each case enrolled in the study, one or more controls are identified and enrolled as well. For example, a case-control study may examine the relationship between oral contraceptive use and AMI. Women of child-bearing age with AMI are enrolled as cases, and women of child-bearing age without AMI are enrolled as controls. The association between oral contraceptive use and AMI can then be examined.<sup>3</sup> For comparative effectiveness research, case-control designs can be used to compare the frequency of outcomes, particularly rare outcomes, among patients exposed to different therapies.

#### Recommended Uses

Nonexperimental studies are particularly useful in situations where randomization is not possible because of ethical or logistical issues. For example, randomization may be infeasible for ethical reasons in sensitive populations (eg, pregnant women) and impractical for logistical reasons in studies that require a large enrolled population or long-term follow-up data. Nonexperimental studies typically have broader inclusion criteria than randomized trials, and therefore may be able to enroll a more representative patient population (eg, patients with multiple co-morbidities, elderly patients). Nonexperimental studies can also provide information on patterns of use. Case-control studies are particularly useful for studying rare outcomes, where it would be difficult to enroll a sufficient sample size in a cohort study.<sup>4</sup> In the oral contraceptive example discussed above, the outcome of interest (AMI) is relatively rare in the population of interest (women of child-bearing age). Therefore, a cohort study would need to be quite large to have sufficient patients to examine this question, making a case-control design an attractive option.

Data for a nonexperimental study may be collected prospectively or retrospectively. In studies with prospective data collection, exposures (eg, exposure to therapies of interest) are measured before the outcomes of interest occur. For example, in the cholesterol-lowering study described above, patients who are taking the medications of interest are enrolled in the study and followed going forward to see if they develop the outcome of interest (eg, AMI). In a study with retrospective data collection, data are abstracted from existing data sources (eg, administrative databases, medical records, etc). As a result, exposures are measured after the outcomes of interest occur. In the cholesterol-lowering study example, all patients taking the medications of interest in an existing data set would be identified. The data from these patients would then be analyzed to see if the outcome of interest was also present.

#### Potential Issues

The primary issue with nonexperimental study designs is the potential for confounding. Because patients are not randomized to treatment groups, it is possible that some patient characteristics are not evenly distributed between the two groups. Confounding occurs when these characteristics are linked to the outcome of interest. In these cases, the characteristics are referred to as confounding variables, and the study analysis must account for these variables. For example, a study comparing the effectiveness of two cholesterol-lowering medications would need to collect and account for dietary interventions. Confounding by indication is also possible in nonexperimental designs. In confounding by indication, a patient characteristic that is related to the outcome of interest influences treatment choice. For example, a study might compare two products for asthma—a newly released product and an existing product. Patients with more severe disease, who were not responding well to the existing product, might be preferentially prescribed the new product, creating uneven treatment groups. Additionally, in nonexperimental studies there is the potential for unmeasured confounding, when a factor that is unknown and unmeasured in the study is affecting the results. For example, in the cholesterol-lowering medications study, an outcome of interest might be AMI. Diabetes is a risk factor for AMI and is associated with the use of cholesterol-lowering medications. If the study does not collect information on diabetes as a risk factor, the study results could be biased. Multiple analytical techniques may be used to address the potential for confounding (see section 5).

Prospective nonexperimental studies can define and collect the specific data elements of interest for the study questions; however, prospective studies require time for data collection and are more expensive than retrospective studies. Retrospective studies, on the other hand, can be completed relatively quickly and are typically cost-effective. However, retrospective studies are limited by the availability of the existing data. Data elements that were not captured in the data source are not available for analysis, and information on how data elements were defined may be lacking.

### Strengths

- Cohort and case-control design can be used in situations where randomization is not ethical or feasible.
- Broad inclusion criteria and limited exclusion criteria can produce a study population that is more representative of the target population, with results that are therefore more generalizable.
- Retrospective studies may be completed relatively quickly and cost-effectively, compared to other types of studies.
- Case-control studies may be useful for examining rare outcomes.

### Limitations

- Confounding is a major issue. Confounding may be addressed through analytical techniques, but unmeasured confounding is always a concern for cohort studies.
- Some data elements of interest may not be available in retrospective studies.

### Selected Examples

- Kerlikowske, et al. *Ann Intern Med.* (2011).<sup>5</sup> The authors conducted a prospective cohort study to examine the comparative effectiveness of digital versus film-screen mammography in community practices in the United States. The study enrolled 329,261 women ages 40 to 79 who underwent either digital or film-screen mammograms. The primary outcomes of interest were mammography sensitivity, specificity, cancer detection rates, and tumor outcomes. A cohort study was chosen in order to evaluate the comparative effectiveness across multiple outcomes, which cannot be done in case-control studies.
- Massarweh, et al. *Ann Surg Oncol.* (2011).<sup>6</sup> The authors conducted a retrospective cohort study to examine the safety and effectiveness of radiofrequency ablation (RFA) for treatment of hepatocellular carcinoma. The study used data from the Surveillance, Epidemiology, and End Results (SEER) database linked to Medicare data, and compared patients who received RFA, resection, or no treatment. The outcomes of interest were mortality and readmission, potentially requiring a long period of follow-up outside of a retrospective study.
- WHO Collaborative Study. *Lancet.* (1997).<sup>3</sup> A case-control study was conducted to assess the association between oral contraceptive use and AMI, a rare event. Women ages 20-44 years who were admitted to the hospital for a definite or possible AMI were enrolled as cases. Each case was matched with up to three controls, drawn from women who were hospitalized for reasons other than AMI. Participants were interviewed while in the hospital, and odds ratios were calculated to compare the risk of AMI in current users of oral contraceptives to the risk in non-users.

### Bibliography

1. Rothman K, Greenland S. *Modern Epidemiology*. 3<sup>rd</sup> ed. Philadelphia, PA: Lippincott Williams & Wilkins; 1998.
2. Sackett DL, Haynes RB, Tugwell P. *Clinical Epidemiology*. Boston, MA: Little, Brown and Company; 1985:228.
3. WHO Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Acute myocardial infarction and combined oral contraceptives: results of an international multicentre case-control study. *Lancet.* 1997;26:349(9060):1202-1209.
4. Hennekens CH, Buring JE. *Epidemiology in Medicine*. Boston, MA: Little, Brown and Company; 1987.
5. Kerlikowske K, Hubbard RA, Miglioretti DL, et al; for the Breast Cancer Surveillance Consortium. Comparative effectiveness of digital versus film-screen mammography in community practice in the United States: a cohort study. *Ann Intern Med.* 2011;155(8):493-502.
6. Massarweh NN, Park JO, Yeung RS, Flum DR. Comparative assessment of the safety and effectiveness of radiofrequency ablation among elderly Medicare beneficiaries with hepatocellular carcinoma [published on line ahead of print]. *Ann Surg Oncol.* 27 September 2011.

## 5. NONEXPERIMENTAL METHODS

### 5.1 New-User Designs

#### Introduction

A new-user design is a type of nonexperimental study that identifies all patients in a defined population at treatment initiation. This type of design eliminates certain biases by excluding prevalent users, retaining only those patients with a minimum period of nonuse and/or no use prior to treatment.<sup>1</sup>

#### Study Design and Methods

*New users:* Given the difficulties in identifying new users, especially in pharmacoepidemiologic studies, patients can be identified as “new” users if they have a specified period of nonuse prior to initiation of therapy (a “washout period”). The chance of an initiator being a true new user can be increased by requiring longer washout periods before the start of follow-up.<sup>2,3</sup>

*Comparison groups:* Choosing a comparison group can be a complex and subjective issue, with the ideal comparison being with patients having identical distributions of measured and unmeasured risk factors for the study outcome. New-user designs can be implemented as a cohort study or as a nested case-control study (a case-control study performed within an existing cohort), with the comparison group ideally consisting of alternative treatment users or non-users.<sup>3</sup>

*Alternative treatment users:* Patients that use an alternative treatment that has the same medical indication.

*Non-users:* Patients that do not use an alternative treatment. They can be identified and assigned a random date for the start of follow-up.

#### Recommended Uses

A new-user design should be considered when the rate at which treatment-related outcomes occur varies with the time elapsed since start of treatment. Often, the period of early use of pharmacotherapy is associated with an elevated or reduced risk of the outcome, which means that bias can be introduced if prevalent users are included and the risk of outcome does in fact vary with time.<sup>1</sup> Consider studies from the 1990s that reported an excess risk of venous thromboembolism in users of third-generation oral contraceptives as compared to users of earlier agents. Since the risk of venous thromboembolism was greater earlier in therapy, and the users of third generation oral contraceptives had started use more recently than the users of earlier therapies, the excess risk identified may have been attributable to the difference in duration of therapy between comparison groups.<sup>1,4</sup> If users of earlier therapies had been observed from the time of treatment initiation, additional events might have been ascertained. New-user designs eliminate this type of under-ascertainment bias because analysis begins with the start of the current course of treatment for every patient.<sup>1</sup>

A new-user design should also be considered when disease risk factors may be altered by the intervention. If studies include prevalent users, it is a challenge to control for disease risk factors that may be altered by the intervention. In a new-user design, these disease risk factors can be measured prior to the start of treatment to eliminate their influence on treatment. This allows for the adjustment of confounders prior to treatment initiation and delineates a clear temporal sequence of assessment and confounder adjustment before treatment initiation.<sup>1,5</sup>

#### Potential Issues

The logistical difficulty of identifying the time that treatment began and collecting information on potential confounders prior to start of treatment is an important limitation of new-user designs.<sup>1</sup> The use of large automated databases and record linkage allows for new users to be identified efficiently, and provides detailed information that can be used to define potential confounders and medication use.

By eliminating prevalent users in a study design, the sample size, and therefore the study power, is often reduced. If automated databases have longitudinal data that includes sufficient history on patients, the power of the study will not decrease, because the time of first use can be identified for each user. If sufficient history on patients is not available, investigators can assess the magnitude of potential biases related to including prevalent users. If no evidence of bias is found, then prevalent users may be included in the analysis.<sup>1,5</sup>

#### Strengths

- New-user study design eliminates under-ascertainment of early events by excluding prevalent users.
- The new-user design takes into account disease risk factors as confounders.

#### Limitations

- It may be difficult to determine the time of treatment initiation.
- New-user study design often leads to reduced sample size because of the exclusion of prevalent users.

## Selected Examples

- Schneeweiss, et al. *Medical Care*. (2007).<sup>2</sup> The authors conducted a study of statin users and mortality within one year in a dual eligible Medicare population. Because the average duration of use may underemphasize the effects related to initiation and overemphasize longer term use, one of the first study cohort restrictions was to include only incident, or new, statin users.
- Ray WA, et al. *Lancet*. (2002).<sup>6</sup> This study used multisite retrospective data from the Tennessee Medicaid program, Canada, and the United Kingdom to look at the risk of myocardial infarction and fatal coronary heart disease in users of generally prescribed NSAIDs among patients with recent hospitalization for coronary heart disease. The cardiovascular safety of NSAIDs is highly controversial, and the risk of future cardiovascular events is associated with prior history, potentially confounding the results. A cohort of new NSAID users and an equal number of non-users were matched based on age, sex, and the date medication use began. The study endpoint was hospital admission for AMI, or death from coronary heart disease.
- Ray WA, et al. *Inj Prev*. (2002).<sup>7</sup> The study used data from the Tennessee Medicaid program, retrospectively comparing the risk of hip fractures among users of statins with that among comparable users of other lipid-lowering agents. Selected new users of all lipid-lowering drugs and randomly selected non-user controls were at least 50 years of age, and did not have life threatening illness, nursing home residence, or diagnosis of dementia or osteoporosis. The main outcome measure was fracture of the proximal femur, excluding pathological fractures or those resulting from severe trauma. The study required that only new users of lipid-lowering agents be included, to ensure detection of all fractures that occurred following lipid-lowering agent use. If a prevalent cohort had been used, early events occurring after treatment initiation might have been missed.

## Bibliography

1. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9):915-920.
2. Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Medical Care*. 2007;45(10):S131-S142.
3. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force report-part II. *Value in Health*. 2009;12(8):1053-1061.
4. Suissa S, Spitzer WO, Rainville B, et al. Recurrent use of newer oral contraceptives and the risk of venous thromboembolism. *Hum Reprod*. 2000;15(4):817-821.
5. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf*. 2010;19(8):858-868.
6. Ray WA, Stein CM, Hall K, et al. Non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease: an observational cohort study. *Lancet*. 2002;359:118-123.
7. Ray WA, Daugherty JR, Griffin MR. Lipid-lowering agents and the risk of hip fracture in a Medicaid population. *Inj Prev*. 2002;8:276-279.

## 5.2 Restriction

### Introduction

Restriction is a method employed to reduce bias by making comparison groups as similar as possible; the goal is to make the groups alike in terms of potential confounding factors and treatment effects. Patients are “restricted” to defined levels of measurable patient characteristics (eg, gender, age group, risk for disease, indication for treatment, etc). If a characteristic is used as a restriction variable, it cannot act as a confounder.<sup>1</sup>

### Recommended Uses

In the real world, therapeutics are not used randomly, but rather for specific indications and in specific populations. A patient’s condition determines the initiation and choice of a specific treatment. When patient cohorts do not have similar indications, treatment can be correlated with disease severity and duration, leading to confounding by indication.<sup>2-3</sup> Confounding by indication weakens the internal validity of study findings by artificially increasing, decreasing, or reversing the estimate of association.<sup>4</sup> Restriction can be used to eliminate confounding by indication.

Consider an example comparing anti-hypertensive agents,  $\beta$ -blockers vs diuretics, on the outcome of myocardial infarction (MI). Since  $\beta$ -blockers are also indicated for treatment of angina, a risk factor for MI, the comparison is subject to bias because of confounding by indication. Restricting the analysis in this example to those patients with no clinical history of cardiovascular disease (CVD) will result in cohorts of patients whose prognosis is unaffected by the presence of clinical CVD.<sup>4</sup>

Restricting study cohorts to patients who are homogeneous in terms of their indication for the study drug will also improve the balance of patient predictors of the study outcome among exposure groups, thus reducing confounding. In addition, restricting study cohorts can increase the likelihood that all included patients will have a similar response to therapy, and therefore reduces the likelihood of effect modification.<sup>1-2</sup>

### Potential Issues

Generalizability of study results can vary depending on the type of restriction applied to the analysis. For example, including only new users and nonusers in a cohort avoids under-representation of treatment effects that occur shortly after treatment initiation, and thus does not limit generalizability. However, when a study restricts high- or low-risk subgroups based upon disease severity or comorbidities, the generalizability of study results is compromised, because the patient population to which physicians can apply the results is limited.<sup>1</sup>

Investigators should also carefully consider variables used to restrict the study population, to assure that restriction does not inadvertently introduce bias.

### Strengths

- Restriction addresses confounding by indication.
- The method can be used to achieve balance in patient characteristics.

### Limitation

- Generalizability of results may be limited dependent upon the criteria for restriction of the population.

### Selected Example

- Schneeweiss, et al. *Medical Care*. (2007).<sup>1</sup> The authors conducted a study of statin users and mortality within one year, in a dual eligible Medicare population. The authors used increasing levels of restriction to reduce bias in the study including: (1) including only incident drug users (defined as having had no statin use in 12 months prior to the index date); (2) choosing a comparison group most similar to the intervention group (by restricting the comparison group to incident users of another preventative therapy); (3) excluding patients with contraindications (by estimating a propensity score model, and trimming the first percentile of the propensity score distribution); (4) excluding patients with low adherence (defined as not filling a second and a third prescription within 180 days after initiation); and (5) assessing subgroups of risk found in trials (by restricting to those patients who would have been eligible to participate in the PROSPER trial). The initial adjusted risk rate for the population was 0.62 for statin use, but after applying the first four restrictions it was changed to 0.80. The level 5 restriction, which restricted patients to those who would have been eligible for the PROSPER trial, yielded an adjusted risk rate of 0.79. This allowed the authors to demonstrate the value of restriction for adjusting for confounding in nonexperimental studies; restricting on the first four levels yielded a comparable risk rate to the one observed in the clinical trial population.

### Bibliography

1. Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Medical Care*. 2007;45(10):S131-S142.
2. Gerhard T. Bias: considerations for research practice. *Am J Health-Syst Pharm*. 2008;65:2159-2168.
3. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force report-part II. *Value in Health*. 2009;12(8):1053-1061.
4. Psaty BM, Siscovick DS. Minimizing bias due to confounding by indication in comparative effectiveness research. *JAMA*. 2010;304(8):897-898.

## 5.3 Subgroup Analysis

### Introduction

Nonexperimental studies include a broad range of patients receiving treatment in routine clinical practice. Because patients within a study population may differ substantially from one another (ie, the patient population is “heterogeneous”), patients may also vary in their response to treatment. Subgroup analysis is a common method used to evaluate whether treatment effects differ between defined subgroups of patients in a nonexperimental study.

### Recommended Uses

Subgroup analysis should be performed in cases where it is suspected that treatment effects may differ across subsets of patients in a study.<sup>1</sup>

To determine whether treatment effects differ, a single statistical test (ie, a test for “interaction”) is first conducted to determine if there is any statistically significant difference in treatment effects across levels of a patient characteristic (eg, age groupings).<sup>2</sup> If a statistically significant difference is found, further testing is required to identify which levels of response (eg, which age groups) are experiencing different treatment effects.

### Potential Issues

Many nonexperimental studies are designed to evaluate the difference in treatment effectiveness between two main treatment groups, and are not initially designed with subgroup analysis in mind. As a result, most studies have only sufficient statistical power to detect the main effect differences overall among all treatment groups in the study.<sup>3</sup> Subgroup analyses may not be able to detect a statistically significant difference in one or more subgroups when, in fact, there actually is such a difference. If a subgroup effect does exist, it may go undetected because the study simply is not large enough.<sup>1-3</sup>

Given the large number of baseline variables, many subgroup analyses in nonexperimental studies arise from post hoc data exploration, as opposed to a predefined subgroup analysis plan.<sup>3</sup> These analyses should be considered exploratory, and only in exceptional circumstances should the analyses affect the main conclusions drawn from the study.<sup>3</sup>

### Strength

- Subgroup analysis assesses whether subgroups of patients may differentially benefit or experience harm due to a treatment, informing clinical decision-making.<sup>3</sup>

### Limitation

- Multiple subgroup analyses are commonly performed. With multiple subgroup analyses (ie, multiple interaction tests), the probability of observing a false positive (finding a significant interaction when one does not exist) is inflated. This could lead to an erroneous conclusion that treatment effect differs across subgroups when it does not.<sup>4</sup>

### Selected Example

- Maraschini, et al. *Interact Cardiovasc Thorac Surg.* (2010).<sup>5</sup> Previously reported data suggested both gender and age might play a role in 30-day in-hospital mortality, however many studies were insufficiently powered, or reported conflicting results. This nonexperimental study evaluated the effect of gender and age on 30-day in-hospital mortality following coronary surgery among 74,577 patients. The patients were stratified by gender and age to test for the effect modification of both variables. The study found that females and elder patients had a significant increased risk for a 30-day in-hospital mortality.<sup>5</sup>

### Bibliography

1. Wang R, Lagakos S, Ware J, et al. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med.* 2007;357(21):2189-2194.
2. Brookes S, Whitely E, Egger M, et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol.* 2004;57:229-236.
3. Pocock S, Assmann S, Enos L, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine.* 2002;21:2917-2930.
4. Lagakos S. The challenge of subgroup analyses—reporting without distorting. *N Engl J Med.* 2006; 354(16):1667-1669.
5. Maraschini A, Seccareccia F, Errigo P, et al. Role of gender and age on early mortality after coronary artery bypass graft in different hospitals: data from a national administrative database. *Interact Cardiovasc Thorac Surg.* 2010;11:537-542.



## 5.4 Propensity Score Methods

### Introduction

Propensity scores are used to address confounding in nonexperimental studies. A propensity score is a summary variable that is computed by collapsing a number of measured variables (eg, age, sex, race, etc) into a single variable, the propensity score. The propensity score can be interpreted as the predicted probability of treatment (ie, the *propensity* for treatment), and is based on the values of the observed variables used to compute the score.

While the propensity score can be used to adjust for multiple measured confounders, it cannot be used to address unmeasured or unknown variables which may affect the outcome (eg, various levels of health literacy in poly-pharmacy patients).<sup>1-2</sup>

### Recommended Uses

Overall, a propensity score is useful because of its ability to balance covariates between treatment groups (ie, make the treatment groups as similar as possible in terms of observed characteristics). Propensity-score methodology can be advantageous when there are a large number of variables that must be accounted for relative to the number of outcomes in a study.<sup>3</sup> Once created, a propensity score can be used for matching or restriction in study design, or for stratification, modeling, or weighting in study analysis.<sup>3-4</sup> The various uses of propensity scores are described briefly below:

*Propensity-Score Restriction:* The population is restricted so that patients in the comparison groups have overlapping propensity score values, making the groups more comparable.

*Propensity score matching:* The propensity score is used to match patients in each comparison group to each other, and treatment effects are estimated using the matched comparisons. For example, patients in treatment group A with propensity scores between 0.6-0.7 can be matched to patients in treatment group B with propensity scores between 0.6-0.7. This matching process occurs across all ranges of the propensity score for the entire study population. Thus, comparisons will only be made between treatment A and treatment B patients that are similar to each other (have the same propensity score), increasing the validity of the treatment effect estimate.

*Propensity score stratification:* The propensity score is divided into categories, and data are examined within the categories of the propensity score. For example, a researcher may decide to divide the study population into three categories—those with low, mid-range, and high propensity scores. Examining the treatment effects within each category of the propensity score helps reduce confounding by those variables used to create the propensity score (eg, age, sex, race, etc).

*Propensity score modeling:* The propensity score is used as a covariate in a statistical model evaluating the relationship between the treatment and the outcome of interest. The researchers essentially treat the propensity score as they would any other independent variable in the model. This accounts for any potential confounding by the variables used to create the propensity score.

*Propensity score weighting:* The propensity score is used to reweight the exposed group, unexposed group, or both groups so that both groups have a similar propensity score distribution. This is often called inverse probability weighted estimation (IPW), and it ensures that the treatment groups are comparable.

### Potential Issues

Several uses of the propensity score (eg, restriction, matching) restrict the study population. Restriction does not retain patients who do not have overlapping propensity scores; matching does not retain patients who cannot be matched. This exclusion of patients leads to a reduced study size and consequently reduced statistical power.

Further, while some uses allow simple, tabular comparison of patients in each comparison group after the propensity score is created, in more complex methodologies (eg, weighting, modeling) data cannot easily be broken down into tables to demonstrate that covariates are balanced between comparison groups. In simpler terms, these more complex methodologies make it difficult to visually represent how the comparison groups are similar following the use of the propensity score.

### Strength

- Use of propensity scores allows adjustment for multiple observed confounders using a single summary measure.

### Limitations

- Some uses of the propensity score can lead to reduced statistical power if all observations are not used in the analysis.
- Some uses of the propensity score do not allow for simple visual inspection of covariate balance following creation of the propensity score.

## Selected Examples

### Restriction

- Schneeweiss et al. *Arch Gen Psychiatry*. (2010).<sup>2</sup> The authors assessed the relationship between use of different anti-depressants and risk for suicide attempts in a claims database. To enable analysis of comparable populations across various types of mental health disorders and medication classes, the authors created a propensity score. This score was developed using all potential confounders (eg, age, sex, comorbidities, medications, etc), and restricted the study population based on the overlap in the propensity scores to ensure that the treatment groups were comparable.

### Matching

- Johannes et al. *Pharmacoepidemiol Drug Saf*. (2007).<sup>5</sup> The authors assessed whether risk of coronary heart disease (CHD) differed between diabetic patients treated with thiazolidinediones and patients treated with combined oral metformin and sulfonylurea therapy. A propensity score was computed based on predetermined covariates (cardiac disease risk factors, indicators of underlying diabetes severity, etc). Patients in each comparison group were then matched on their propensity scores. The matching reduced the extent of confounding by the variables used to compute the propensity score.

### Stratification

- Jaar et al. *Ann Intern Med*. (2005).<sup>6</sup> The authors compared peritoneal dialysis versus hemodialysis, and risk of death. They found that patients receiving peritoneal dialysis had a better case-mix profile at baseline compared to patients receiving hemodialysis. A propensity score was calculated using baseline variables that were expected confounders (eg, age, sex, race, employment status, etc). The authors then stratified the population into three categories of the propensity score to reduce confounding by the variables used to compute the propensity score.

### Modeling

- Applegate et al. *Catheter Cardiovasc Interv*. (2006).<sup>7</sup> The authors assessed vascular closure devices (VCDs) versus manual compression, and the risk of vascular complications. The authors computed a propensity score based on potential confounders (eg, demographics, clinical presentation, body surface area, etc) and then included the propensity score as a variable in a regression model to reduce confounding by the variables used to compute the propensity score.

### Weighting

- Scott et al. *Ann Surg Innov Res*. (2010).<sup>8</sup> The authors assessed Video-Assisted Thoracic Surgical (VATS) lobectomy versus open lobectomy, and risk for perioperative outcomes (eg, mortality, complication rate, and length of stay). A propensity score was computed based on measured confounders (age, gender, preoperative FEV1, etc). The propensity score was used to conduct IPW, in order to reduce confounding by the variables used to compute the propensity score.

## Bibliography

1. Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force report - part III. *Value in Health*. 2009;12(8):1062-1073.
2. Schneeweiss S, Patrick AR, Solomon DH, et al. Variation in the risk of suicide attempts and completed suicides by antidepressant agent in adults: a propensity score-adjusted analysis of 9 years' data. *Arch Gen Psychiatry*. 2010;67(5):497-506.
3. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding [published online ahead of print]. *Stat Methods Med Res*. 24 January 2011.
4. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
5. Johannes CB, Koro CE, Quinn SG, et al. The risk of coronary heart disease in type 2 diabetic patients exposed to thiazolidinediones compared to metformin and sulfonylurea therapy. *Pharmacoepidemiol Drug Saf*. 2007;16(5):504-512.
6. Jaar BG, Coresh J, Plantinga LC, et al. Comparing the risk for death with peritoneal dialysis and hemodialysis in a national cohort of patients with chronic kidney disease. *Ann Intern Med*. 2005;143(3):174-183.
7. Applegate RJ, Sacrinty MT, Kutcher MA, et al. Propensity score analysis of vascular complications after diagnostic cardiac catheterization and percutaneous coronary intervention 1998-2003. *Catheter Cardiovasc Interv*. 2006;67(4):556-562.
8. Scott WJ, Matteotti RS, Egleston BL, et al. A comparison of perioperative outcomes of video-assisted thoracic surgical (VATS) lobectomy with open thoracotomy and lobectomy: results of an analysis using propensity score based weighting. *Ann Surg Innov Res*. 2010;4(1):1.

## 5.5 Instrumental Variable Methods

### Introduction

Instrumental variable (IV) methods are designed to allow for the accurate estimation of treatment effects in the presence of unmeasured confounding (where confounders are either unknown or not measured in the data being analyzed).<sup>1</sup> This is done by finding a variable associated with the variation in treatment, and using it to achieve balance on unmeasured confounders across treatment groups.<sup>2</sup>

An IV is a variable that is 1) related to treatment, and 2) neither directly nor indirectly related to the outcome of interest.<sup>2</sup> For example, a study examining the effects of intensive therapy of AMI patients and mortality risk could use proximity to a tertiary care facility as an IV.<sup>3</sup> Proximity to a tertiary care facility is strongly associated with whether a patient receives intensive therapy for AMI, and it can be reasonably assumed that proximity to a tertiary care facility is not directly or indirectly related to mortality.<sup>3</sup>

### Recommended Uses

IV methods are useful in situations where a significant amount of unmeasured confounding is suspected, and an IV that is strongly associated with treatment is available for use.

### Potential Issues

For nonexperimental designs, the identification and evaluation of potential IVs can be very difficult. Investigators must use expert knowledge to justify that the IV is valid; there is no way to empirically assess whether the IV is directly or indirectly associated with the outcome.<sup>2</sup> If the IV is not strongly associated with treatment, or is associated with the outcome, results may also be misleading.

IV methods are also inefficient, requiring a large sample size for precise treatment effect estimates.<sup>1-2,4</sup> IV methods are used as a primary analytic method when investigators suspect a significant amount of unmeasured confounding. In other cases, IV methods are used as a secondary analysis.<sup>2</sup>

### Strength

- IV methods offer potential to significantly reduce bias due to unmeasured confounding.

### Limitations

- The method is limited in practice by the availability of valid IVs (those that are strongly associated with the treatment and not directly or indirectly associated with the outcome).<sup>2,5</sup>
- Treatment effect estimates will be biased if the IV is directly or indirectly associated with outcome (and this cannot be verified empirically).<sup>4</sup>
- Biases are amplified if the association between the IV and treatment is weak.<sup>4</sup>
- Traditional IV analyses do not account for patients who may switch or discontinue a treatment over time; in these situations, more complex analytic methods are required.<sup>4</sup>
- Additional, unverifiable assumptions are required if treatment effect is heterogeneous.<sup>4,6-7</sup>
- The IV method is less efficient than other adjustment methods (wider confidence intervals).<sup>1-2</sup>

### Selected Examples

- Brooks et al. *Health Serv Res.* (2003).<sup>8</sup> The study examined breast conserving surgery plus irradiation (BCSI) versus mastectomy for average survival. The study used two IVs: BCSI percentage in the patient's area (grouped into lower and greater; if less than 20% of early stage breast cancer surgeries in the 50 mile radius around the patient's residence were BCSI, the classification was lower) and distance to a radiation treatment center (grouped into near and far; if the distance to a radiation treatment center in year of diagnosis was less than 19 miles away, the classification was near). These IVs were chosen because proximity to facilities that perform the treatments of interest is strongly associated with receiving the treatments of interest. It was also reasonably assumed that proximity to facilities was not directly or indirectly related to average survival.
- Schneeweiss et al. *N Eng J Med.* (2008).<sup>9</sup> The study examined aprotinin, an injection used during complex surgery, to evaluate the use during coronary artery bypass graft (CABG) and the risk of death. The IV was surgeon "preference" for aprotinin. If a surgeon administered the drug to 90% or more of their patients they preferred aprotinin; if a surgeon administered the drug to 10% or fewer of their patients, they did not prefer aprotinin. The authors also looked at a 100% vs 0% split. (If a surgeon administered the drug to 100% of their patients they preferred aprotinin; if a surgeon administered the drug to 0% of their patients, they did not prefer aprotinin.) Surgeon preference was chosen as an IV because it is strongly associated with treatment choice. (As this was a nonexperimental study, treatment choice was at the discretion of the surgeon.) It was also reasonably assumed that surgeon preference for aprotinin was not directly or indirectly associated with risk of death.

## Bibliography

1. Bennett DA. An introduction to instrumental variables analysis: part 1 [published online ahead of print September 23, 2010]. *Neuroepidemiology*. 2010;35(3):237-240.
2. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554.
3. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. 1994;272(11):859-866.
4. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360-372.
5. Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force report-part III. *Value in Health*. 2009;12(8):1062-1073.
6. Rassen JA, Brookhart MA, Glynn RJ, et al. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships [published on line ahead of print April 8, 2009]. *J Clin Epidemiol*. 2009;62(12):1226-1232.
7. Brookhart MA, Rassen JA, Wang PS, et al. Evaluating the validity of an instrumental variable study of neuroleptics: can between-physician differences in prescribing patterns be used to estimate treatment effects? *Med Care*. 2007;45(10)(suppl 2):S116-122.
8. Brooks JM, Chrischilles EA, Scott SD, Chen-Hardee SS. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. *Health Serv Res*. 2003;38(6)(pt 1):1385-1402.
9. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. *N Engl J Med*. 2008;358(8):771-783.

## 5.6 Sensitivity Analyses

### Introduction

Sensitivity analyses assess the potential impact of unmeasured confounders on study results.<sup>1</sup> Sensitivity analyses are the “last line of defense against biases after every effort has been made to eliminate, reduce, or control them in study design, data collection, and data analysis.”<sup>2</sup> To examine the potential for unmeasured confounding, informed assumptions are made in order to quantify the effect of an unmeasured confounder on the treatment effect estimate.<sup>3</sup>

### Types of Sensitivity Analyses

There are multiple approaches to conducting a sensitivity analysis, each with its own limitations and application to nonexperimental studies. It is important to understand each model's assumptions, carry out the calculations, and pay attention to interpretations of results. Spreadsheets and statistical packages often expedite the analysis and produce graphical illustrations that are useful for understanding the results.<sup>4</sup>

Two widely used approaches to sensitivity analysis are the array approach and the rule-out approach.<sup>3,5</sup> The array approach is used to understand how the strength of an unmeasured confounder (and its imbalance among treatment groups) affects the observed treatment-effect estimate. The end result is an array of different risk estimates over a wide range of parameter values. For example, if smoking prevalence in a population is unknown, it may be varied from 20%-90% of the population to observe the associated changes in the effect estimate.

The rule-out approach is used to assess the extent of confounding from a single variable that would be necessary to explain the observed treatment-effect estimate. Confounders that are not strong enough to eliminate the observed treatment effect can be ruled out.

### Recommended Uses

Sensitivity analyses should be conducted in cases where unmeasured confounding is suspected, in order to determine the extent of the bias. While not covered in detail in this brief, sensitivity analyses may also be used to assess the sensitivity of study findings to changes in exposure and outcome definitions, and to other assumptions made during conduct of the study.

### Potential Issues

While there are several quantitative approaches for assessing sensitivity of study results to potential unmeasured confounders, assessing whether an analysis is in fact insensitive to unmeasured confounding is still a matter of judgment.<sup>1</sup>

#### Strength:

- Sensitivity analyses can assess the potential effect of unmeasured confounding on study results.

#### Limitation:

- Various approaches each have their own limitations. The rule-out approach is limited to one binary confounder and does not assess the magnitude of confounding; several additional approaches not described require extensive technical understanding and programming skills to conduct. Investigators must understand the limitations of each approach, and choose the appropriate analysis to conduct.

### Selected Example

- Groenwold et al. *Intern J of Epidemiol.* (2010).<sup>1</sup> The authors discuss different methods of sensitivity analysis and apply them to a clinical example of influenza vaccination and mortality risk. In one example, the authors simulate what happens to the treatment-effect estimate as several associations are varied: (1) the strength of the association of the unmeasured confounder and vaccination status; (2) the strength of the association of the unmeasured confounder and mortality risk; and (3) the prevalence of the confounder. The authors conclude that in order to eliminate the treatment effect (ie, to arrive at a conclusion of no estimated treatment effect) the prevalence of the confounder would need to be 20%-40%, and the relationship between the unmeasured confounder and vaccination and mortality would have to be very strong (an odds ratio [OR] of 3.0).

### Bibliography

1. Groenwold RHH, Nelson DB, Nichol KL, et al. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *Int Jour of Epidemiol.* 2010;39:107-117.
2. West SL, Strom BL, Poole C. Validity of pharmacoepidemiology drug and diagnosis data. In: Strom BL, ed. *Pharmacoepidemiology*. 3<sup>rd</sup> ed. Chichester: Wiley; 2000:668.
3. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf.* 2006;15:291-303.
4. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf.* 2010;19(8):858-868.
5. Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force report - part III. *Value in Health.* 2009;12(8):1062-1073.

## 5.7 External Information

### Introduction

In nonexperimental studies, important information such as patient clinical characteristics in claims-based studies and lab data in questionnaire-based studies may not be available.<sup>1</sup> For studies examining an outcome of thromboembolism, for example, both smoking status and obesity, risk factors for the outcome and potential confounders, may not be recorded in the data source. While simple sensitivity analyses (see section 5.6) can assess this potential unmeasured confounding, the approach treats each variable (either smoking or obesity) as independent variables. The combined effect of the unmeasured variables (both smoking status and obesity) is not considered.<sup>1,2</sup> This can be a concern when it is anticipated that accounting for each covariate separately may yield a different result compared to accounting for the variables jointly. External information can be used to adjust for multiple unmeasured confounders and their joint effects.

There are two types of external information that can be used: data on individuals within the main study population (internal validation studies) or data on individuals outside of the main study population (external validation studies).

### Internal Validation Studies

Internal validation studies use additional data obtained from a subset of the participants in the main study population.<sup>2,3</sup> This additional sample can be a random subsample of the cohort, or a nonrandom sample selected based on information on exposure, outcome, or both.<sup>3</sup> Nonrandom sampling is referred to as a “2-stage design” and requires analyses that take into account the sampling mechanism employed.<sup>3</sup>

### External Validation Studies

External validation studies use additional data collected outside of the main study population. External data are usually cross sectional survey data and not specific to a particular hypothesis. These data are often relatively inexpensive to acquire.<sup>3</sup>

### Recommended Uses

Internal and external validation studies are useful when an existing database does not contain important information on risk factors or potential confounders of interest. The use of external information allows the investigator to adjust for multiple unmeasured confounders, as well as to address potential joint confounding by unmeasured covariates (eg, the joint effect of smoking status and obesity in the example above).<sup>2,3</sup>

### Potential Issues

Different techniques must be employed in analyses depending on the availability of information, the type of validation study (eg, internal or external) and the sampling mechanism (eg, random or non-random) utilized. Methods such as multiple imputation, maximum likelihood and estimating equations, and propensity score calibration may be used, and require a detailed understanding of methodology and associated assumptions of each analytic technique in order to be applied correctly.

### Strengths

- External information can be analyzed to allow adjustment for multiple unmeasured confounders.
- External validation studies may be used for multiple main studies because they are not specific to any particular research question.<sup>3</sup>

### Limitations

- Internal validation studies can be conducted only when it is feasible to collect additional information from patients in the data source.<sup>3</sup>
- External validation studies often do not have the exact measures required or do not fully represent the main study population.<sup>3</sup>

### Selected Examples

#### Internal validation studies

- Eng, et al. *Pharmacoepidemiol Drug Saf.* (2008).<sup>2</sup> The authors conducted a study of oral contraceptive use and risk for thromboembolism using an administrative claims database. Because certain thromboembolic risk factors (eg, smoking and obesity) that are also potential confounders were not readily available in the administrative claims database, a case cohort design was employed to assess residual confounding by unmeasured variables in the main study cohort. Patients were randomly sampled from the main study cohort to create the sub-cohort. Additional information from the sub-cohort was collected and used to impute missing values in the main cohort.
- Walker, et al. *Lancet.* (1981).<sup>4</sup> The authors conducted a study of vasectomy and risk for non-fatal myocardial infarction in an HMO population. In the initial sample the incidence of non-fatal myocardial infarction was slightly lower in men without vasectomy, after controlling for birth year and length of observation. To adjust for cardiac risk factors that may confound these results, medical records were accessed for select members of the population. Nonrandom sampling identified a subset of the population with medical information.

#### External validation study

- Stürmer, et al. *Am J Epidemiol.* (2005).<sup>3</sup> The authors conducted a study of NSAID use and one-year mortality in the New Jersey Medicaid population. Initial estimation using a propensity score to adjust for measured confounders (age, race, sex, other diagnoses, medications, etc) indicated that NSAID use was associated with a decrease in short-term mortality in elderly hospital patients. However, the authors suspected that there were potentially other confounders biasing the treatment-effect estimate. Data from the Medicare Current Beneficiary Survey (MCBS), an ambulatory survey, were used to collect data on unmeasured confounders (eg, smoking, body mass index, education, income, etc) among a population with similar age and sex distribution as in the original study. Effect estimates in the main study were then adjusted using statistical techniques, and there was no longer an association observed between NSAID use and one-year mortality.

#### Bibliography

1. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol.* 2005;162(3):279-289.
2. Eng PM, Seeger JD, Loughlin J, et al. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiol Drug Saf.* 2008;17(3):297-305.
3. Stürmer T, Glynn RJ, Rothman KJ, et al. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care.* 2007;45(10)(suppl 2):S158-165.
4. Walker AM, Jick H, Hunter JR, et al. Vasectomy and non-fatal myocardial infarction. *Lancet.* 1981;1(8210):13-15.

## 6. Glossary of Terms

Term	Definition
Applicability	See “external validity.”
Association	A relationship between two variables, such that as one changes, the other changes in a predictable way. A positive association occurs when one variable increases as another one increases. A negative association occurs when one variable increases as the other variable decreases. Association does not imply causation. Also called correlation.
Bias	A systematic error in study design that results in a distorted assessment of the intervention’s impact on the measured outcomes. In clinical trials, the main types of bias arise from systematic differences in study groups that are compared (selection bias), exposure to factors apart from the intervention of interest (performance bias), participant withdrawal or exclusion (attrition bias), or assessment of outcomes (detection bias). Reviews of studies may also be particularly affected by reporting bias, where a biased subset of all the relevant data is available.
Blinding	A randomized trial is “blind” if the participant is unaware of which arm of the trial he is in. Double blind means that both participants and investigators do not know which treatment the participants receive.
Case-control study	A nonexperimental study that compares individuals with a specific disease or outcome of interest (cases) to individuals from the same population without that disease or outcome (controls) and seeks to find associations between the outcome and prior exposure to particular risk factors. This design is particularly useful where the outcome is rare and past exposure can be reliably measured. Case-control studies can be retrospective or prospective.
Causality	An association between two characteristics that is demonstrated to be due to cause and effect (ie, a change in one causes change in the other).
Causation	See “causality.”
Cohort study	A nonexperimental study with a defined group of participants (the cohort) that is followed over time. Outcomes are compared between subsets of this cohort who were exposed or not exposed (or exposed at different levels) to a particular intervention or other factors of interest. Cohort studies can either be retrospective or prospective.
Comorbidity	A medical condition that exists simultaneously with another medical condition.
Comparative Effectiveness Research (CER)	The generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.
Comparison group	See “control group.”
Confidence interval	A range of values for a variable of interest, eg, a rate, constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits.
Confounder	See “confounding variable.”
Confounding by indication	Occurs when the underlying diagnosis or other clinical features that trigger the use of certain treatment are also related to patient outcome.



Term	Definition
Confounding variable	A variable (or characteristic) more likely to be present in one group of participants than another that is related to the outcome of interest and may potentially confuse (confound) the results. For example, if individuals in the experimental group of a controlled trial are younger than those in the control group, it will be difficult to determine whether a lower risk of death in one group is due to the intervention or the difference in ages (age is the confounding variable). Confounding is a major concern in non-randomized studies. Also called confounder.
Control group	Participants in the control arm of a study.
Correlation	See "association."
Covariate	An independent variable not manipulated by the study that affects the response.
Cox proportional hazard model	A statistical model which is used to analyze survival data.
Effect modification	A situation in which the measure of effect is different across values of another variable (eg, the measure of effect is different across race, age, etc.).
Efficacy	(Of a drug or treatment). The maximum ability of a drug or treatment to produce a result regardless of dosage. A drug passes efficacy trials if it is effective at the dose tested and against the illness for which it is prescribed.
Endpoint	See "outcome."
Experimental study	A study in which the investigators actively intervene to test a hypothesis. It is called a trial or clinical trial when human participants are involved.
Explanatory trials	A controlled trial that seeks to measure the benefits of an intervention in an ideal setting (efficacy) by testing a causal research hypotheses. Trials of healthcare interventions are often described as either explanatory or pragmatic. See also "pragmatic trial."
External validity	The extent to which results provide a correct basis for generalizations to other populations or settings. Also called generalizability, applicability. See also "applicability."
Generalizability	See "external validity."
Homogeneous	Having similarity of participants, interventions, and measurement of outcomes across a set of studies.
Hypothesis testing	An objective framework to determine the probability that a hypothesis is true.
Incidence	The number of new cases of an event that develop within a given time period in a defined population at risk, expressed as a proportion.
Incidence rate	A measure of the frequency with which an event, such as a new case of illness, occurs in a population over a period of time. The denominator is the population at risk; the numerator is the number of new cases occurring during a given time period.
Internal validity	The extent that the design and conduct of a study are likely to have prevented bias. More rigorously designed (better quality) trials are more likely to yield results that are closer to the truth. See "bias."
Intra-class correlation coefficient (ICC)	Used to assess correlation between classes of data (eg, inter-rater reliability).
Logistic regression	A statistical technique that predicts the probability of a dichotomous dependent variable (eg, dead or alive) using, typically, a combination of continuous and categorical independent variables.
Masked	See "blinding."

Term	Definition
Matching	When individual cases are “matched” with controls that have similar confounding factors (age, sex, BMI, etc) to reduce the effect of the confounding factors on the association being investigated.
Multiple imputation	A method to predict and fill in the missing values of a study based on the observed data and the missing-data pattern.
Multivariate analysis	Involves the construction of a mathematical model that describes the association between the exposure, disease, and confounders.
Nested case-control study	A study where cases and controls are selected from patients in a cohort study (a case-control study “nested” within a cohort study).
New-user design	A type of study that restricts the analysis to persons under observation at the start of the current course of treatment.
Nonexperimental study design	A study in which investigators observe the course of events and do not assign participants to the intervention. Also called an observational study.
Observational study	See “Nonexperimental study design.”
Outcome	The result of an experimental study that is used to assess the efficacy of an intervention. Also called endpoint.
Patient registry	An organized system that uses nonexperimental study methods to collect uniform data (clinical or other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes.
Pharmacoepidemiology	Study of the use, effects, and outcomes of drug treatment from an epidemiological (population) perspective.
Placebo	A placebo is an inactive drug, therapy or procedure that has no treatment value. In clinical trials, experimental treatments are often compared with placebos to assess the treatment’s effectiveness.
Power	The probability of rejecting the null hypothesis when a specific alternative hypothesis is true. The power of a hypothesis test is one minus the probability of Type II error. In clinical trials, power is the probability that a trial will detect an intervention effect of a specified size that is statistically significant. Studies with a given number of participants have more power to detect large effects than small effects. In general, power is set at 80% or greater when calculating sample size. Also called statistical power.
Precision	In statistics, precision is the degree of certainty surrounding an effect estimate for a given outcome. The greater the precision, the less the measurement error. Confidence intervals around the estimate of effect are one way of expressing precision, with a narrower confidence interval defining a higher precision.
Prevalence	The proportion of a population that is affected by a given disease or condition at a specified point in time. It is not truly a rate, although it is often incorrectly called prevalence rate.
Probability	The probability (likelihood) of an event is the relative frequency of the event over an infinite number of trials.
Prospective study	A study in which exposures are measured by the investigator before the outcome events occur.
P-value	The probability (ranging from zero to one) that the results observed in a study (or more extreme results) could have occurred by chance.

Term	Definition
Quasi-experimental	A study that is similar to a true experiment except that it lacks random assignment of participants to treatment and control groups. A quasi-experimental design may be used to reveal a causal relationship in situations where the researcher is unable to control all factors that might affect the outcome. Because full experimental control is lacking, the researcher must thoroughly consider threats to validity and uncontrolled variables that may account for the results.
Randomization	The process of randomly assigning participants to one of the arms of a controlled trial. Ensures that participants have an equal and independent chance of being in each arm of the study. There are two components to randomization: generation of a random sequence and its implementation, ideally in such a way that those enrolling participants into the study are not aware of the sequence (concealment of allocation).
Randomized controlled trial (RCT)	An experimental study (controlled trial) in which participants are randomly assigned to treatment groups (experimental and control groups).
Regression analysis	A statistical modeling technique used to estimate or predict the influence of one or more independent variables on a dependent variable.
Residual confounding	Confounding by unmeasured variables in a study.
Restriction	Limiting of cohort entry to individuals with a certain range of values for a confounding factor (eg, age, race, etc) to reduce the effect of the confounding factor.
Retrospective study	A study in which exposures are measured by the investigator after the outcome events have occurred.
Standard error	The standard deviation of a theoretical distribution of sample means about the true population mean.
Structural equation modeling	Structural equation modeling includes a broad range of multivariate analysis methods aimed at finding interrelations among the variables in linear models by examining variances and covariances of the variables.
T-test	A statistical examination of two population means. A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size.
Type I error	The error that results if a true null hypothesis is rejected or if a difference is concluded when no difference exists. Also called alpha error, false alarm and false positive.
Type II error	The error that results if a false null hypothesis is not rejected or if a difference is not detected when a difference exists. Also called beta error, missed opportunity, and false negative.
Variance	A measure of the dispersion shown by a set of observations, defined by the sum of the squares of deviations from the mean, divided by the number of degrees of freedom in the set of observations.

## Glossary References:

Aschengrau A, Seage G. *Essentials of Epidemiology in Public Health*. 2nd ed. Sudbury, MA: Jones and Bartlett; 2008.

Centers for Disease Control. Reproductive Health: Glossary. <http://www.cdc.gov/reproductivehealth/epiglossary/glossary.htm#C>

Eurordis Rare Diseases Europe. Clinical trials glossary. June 2007. [http://www.eurordis.org/IMG/pdf/CT\\_GLOSSARY\\_FINAL.pdf](http://www.eurordis.org/IMG/pdf/CT_GLOSSARY_FINAL.pdf). Accessed October 17, 2011.

Hines LE. Glossary of comparative effectiveness research terms. University of Arizona Comparative Effectiveness Research Group. September 2011. <http://cer.pharmacy.arizona.edu/images/CER.Glossary.v9.pdf>. Accessed October 17, 2011.

Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9):915-920.

Rosner, Bernard. *Fundamentals of Biostatistics*. 6th ed. Belmont, CA: Thomson-Brooks/Cole; 2006.

Rothman KJ, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

Strom BL. *Pharmacoepidemiology*. 4<sup>th</sup> ed. Chichester: J Wiley; 2005. Print.

The Institute for Statistics Education. Glossary of statistical terms. 2011. <http://www.statistics.com/index.php?page=glossary>. Accessed October 17, 2011.

US National Institutes of Health. Glossary of clinical trials terms. 2008. <http://clinicaltrials.gov/ct2/info/glossary>. Accessed October 17, 2011.