Evaluating Clinical and Public Health Interventions
Evaluating clinical and public health interventions

A practical guide to study design and statistics

Mitchell H. Katz, MD
To Igael Gurin-Malous, with love
# Contents

**Preface**  

1  **Introduction**  

1.1 Why study interventions?  
1.2 How can you tell whether an intervention is effective?  

2  **Interventions**  

2.1 What type of interventions are commonly evaluated in medicine and public health?  
2.2 How do I design an intervention?  

3  **Evaluating an intervention**  

3.1 How do I engage stakeholders?  
3.2 What data should I collect for evaluating my intervention?  
3.3 How do I choose a control group for evaluating my intervention?  
3.4 How do I collect data and choose a control group for an intervention that I did not develop?  
3.5 What outcomes of an intervention should I assess?  
3.6 Why measure the exposure of subjects to an intervention?  
3.7 Should I use a longitudinal cohort or a serial cross-sectional design?  
3.8 How do I develop measurement instruments?  
3.9 How do I state the hypothesis of an intervention study?  
3.10 Can an intervention study have more than one hypothesis?  
3.11 Why should I have an analysis plan?  
3.12 How do I calculate sample size for an intervention study?  
3.13 How do I obtain an institutional review board (IRB) review for a study?
### Contents

3.14  When do I need an independent data and safety monitoring committee?  45
3.15  Why should I register my trial?  45

4  Randomized designs  47

4.1  What are randomized studies?  47
4.2  What are the advantages of randomization?  47
4.3  What are the disadvantages of randomization?  49
4.4  What are the different methods of allocating subjects?  50
4.5  What is clustered (group) randomization?  52
4.6  What can be randomized besides individuals or clusters of individuals?  54
4.7  When should subjects and researchers be masked (blinded) to treatment assignment?  55
4.8  How do I account for subjects who do not adhere to their treatment?  56

5  Nonrandomized studies  61

5.1  What are nonrandomized studies?  61
5.2  Why perform nonrandomized studies?  61
5.3  What are the disadvantages of nonrandomization?  65
5.4  How can I assemble comparable samples in the design phase without randomization?  68

6  Statistical analysis of intervention trials  73

6.1  How do I test whether my intervention has had a statistically significant effect?  73
6.2  Is the difference between the pre-intervention and the post-intervention assessment statistically significant in a longitudinal study?  74
6.3  Is the difference between the pre-intervention and the post-intervention assessment statistically significant in a serial cross-sectional study?  81
6.4  Is the difference between the pre-intervention and the post-intervention assessment of the intervention group statistically greater (lesser) than that seen in the comparison group(s) in a longitudinal cohort study?  83
6.5  Is the difference between the pre-intervention and the post-intervention assessment of the intervention group statistically greater (lesser) than that seen in the comparison group(s) in a serial cross-sectional study?  

6.6  Is the post-intervention assessment of the intervention group significantly different than the corresponding assessment of the comparison group?  

7  Methods for adjusting for baseline differences between treatment groups  

7.1  How do I adjust for baseline differences between persons who receive the intervention and those who don’t?  
7.2  What are propensity scores? How are they calculated?  
7.3  Which variables should be included in a propensity score?  
7.4  How do you assess the adequacy of a propensity score?  
7.5  How are propensity scores used to adjust for baseline differences?  
7.6  What is instrumental variable analysis? How is it used to adjust for baseline differences between groups?  
7.7  What are the underlying assumptions of instrumental variable analysis?  
7.8  How is instrumental variable analysis performed?  
7.9  What are the limitations of instrumental variable analysis?  
7.10  How do the results of multivariable adjustment alone, propensity scores, and instrumental variable analysis compare?  
7.11  What is sensitivity analysis?  

8  Time series analysis  

8.1  How do I use time series analysis to analyze how an intervention affects a series of observations over time?  
8.2  How many observations are needed for performing an interrupted time series analysis?  

9  Special topics  

9.1  What methods are available for evaluating interventions that do not occur to all subjects at the same time?  
9.2  What if I want to prove that an intervention is not inferior or is equivalent to another intervention?
Preface

My experience as a public health director has taught me that solving complex health problems requires identifying useful interventions. This book is my intervention to make it easier (and more fun) to conduct intervention studies.¹

Many different types of interventions can improve health and health care, including drugs and medical devices, laws and changes in medical and organizational practices. To enable you to evaluate a wide range of interventions, this book explains the advantages and disadvantages of many different approaches: randomized and nonrandomized studies; prospective and retrospective studies; planned clinical trials and observational studies (e.g., studies using administrative databases, disease registries); evaluations of investigator-initiated interventions as well as evaluations of interventions created by others (e.g., a law or a system change).

By covering both study design and statistical analysis in one book I hope to save you from having to go back and forth between a textbook on study design and one on statistics. Also, for many intervention studies, especially nonrandomized designs, understanding the statistical approach to analyzing the data is key to judging the robustness of the design. Within the statistics section I have organized the material the way that researchers think: what is the best test to perform to answer a particular question? For readers who wish to learn more about particular tests (e.g., McNemar test, mixed-effects model) I have provided references in the footnotes.

To illustrate the points in this book, I have included many examples from the medical and public health literature. Examining published articles builds on your knowledge of how to read the literature, and illustrates practical strategies to use in planning your study and analyzing the results. Also, you may find it helpful to model your study after one of the examples in the book. Given the number and range of examples I am certain you will find one that is similar to your work.

Preface

To minimize overlap between this book and my other two books: *Study Design and Statistical Analysis: A practical guide for clinicians* (Cambridge University Press, 2006) and *Multivariable Analysis: A practical guide for clinicians* (2nd edition: Cambridge University Press, 2006), I liberally cite them. Please don’t see this as an expression of ego, but rather as a desire to minimize overlap. Some may prefer all the material in one text. Although that would have some advantages, it would make the books longer, more expensive, and more daunting to people new to the field. If you like this book and are interested in learning more about non-intervention studies (e.g., descriptive studies, risk-factor studies) or learning more about statistics I hope you will read these other two books.


I appreciate the support of my editors Richard Marley and Katie James and the staff at Cambridge University Press.

If you have any questions or suggestions for future editions, please e-mail me at mhkatz59@yahoo.com. Readers of my other two books have e-mailed me from all around the world. Writing textbooks is an arduous and lonely business. Knowing that others benefit from the books sustains me.
Introduction

1.1 Why study interventions?

Because solving problems requires interventions but not all interventions are effective!

Let’s say you are confronting the problem of obesity in your clinic or community. A necessary first step is to understand the prevalence (frequency) of obesity, the characteristics of affected individuals (e.g., age, gender, geographic location), and how severely they are affected (e.g., presence of comorbid conditions such as diabetes). A second necessary step is to identify the risk factors for obesity, especially those that are amenable to change.

But sadly too many research agendas never move beyond this second step. Investigators conclude their manuscripts with the ubiquitous and meaningless phrase: “Interventions are needed to . . . .”, yet the intervention is never performed. A review of bibliography sources found that only 0.4 percent of academic research focused on public health interventions.\(^1\) Although intervention research is more common with pharmaceuticals, this research is often limited to efficacy trials conducted under conditions that cannot easily be replicated in the real world.

This does not have to be the case. Developing interventions can be more fulfilling than descriptive or risk-factor studies because they can directly change the world! Interventions can be drugs or medical devices, counseling or skill-building programs for individuals or groups, laws or changes in institutional practices.

Let’s return to the problem of obesity. A large body of evidence has documented a major increase in obesity rates in developed countries, with serious sequelae including type 2 diabetes. Decreased activity, larger food portions, and high caloric foods have all been shown to be risk factors for obesity.

Farley and colleagues developed a common sense intervention to increase the activity of children and thereby diminish a known risk factor for obesity: they opened a schoolyard for play during non-school hours in a low-income neighborhood in New Orleans and provided attendants to ensure children’s safety. The schoolyard was immediately popular: 710 children were observed in the schoolyard at least once during a 12-month period; 66% of the children were physically active when observed in the schoolyard.

To evaluate the impact of the schoolyard on the physical activity of children in the community, the investigators compared children in this community to children from a neighboring community. Prior to the opening of the schoolyard the number of children observed to be active in the intervention community was lower than in the comparison community. After the intervention, the number of children observed to be active was greater in the intervention community than in the comparison community, not counting the children in the schoolyard.

Besides being easy to replicate, the intervention had the advantage that it does not single out children who are obese, which may harm self-image. Rather it takes advantage of the idea that all children should be active.

To ameliorate a health problem you don’t have to be the one who develops the intervention. Many useful studies have evaluated interventions that were developed by the government (e.g., laws banning smoking at workplaces) or another organization (e.g., school-based physical education). For example, Hu and colleagues assessed the impact of taxation on cigarette sales. They estimated that the 25-cent tax that California added in 1989 to each pack of cigarettes resulted in a reduction of 514 million packs sold over an 18-month period. Evaluations such as this have been successful in motivating other states to add tobacco excise tax.

You may also find an unplanned opportunity, a “natural experiment,” to evaluate whether a change in circumstances improves a health problem. For example, Costello and colleagues studied the impact of the opening of a casino on an Indian reservation on the mental health of Native American children. The point of the casinos (the intervention) was not to improve children’s mental health. However, it is known that psychopathology is higher in children from low-income families and that casinos increase the income of families living on

---


Introduction

a reservation. Would the presence of the casinos improve the mental health of the children?

The investigators found that after the opening of the casino the psychopathology level of the previously poor children improved to the level of the children who were never poor. The study was feasible only because a cohort study was in progress prior to the opening of the casino. The beauty of the study is that it overcomes the limitations of other possible designs: it is impossible to randomize families to higher income; a longitudinal observational cohort study looking at the connection between income and children’s mental health would have the challenge of separating the impact of income gains from the impact of the factors that led to the gain in income (e.g., new job, second parent working).

1.2 How can you tell whether an intervention is effective?

It is not always easy! Of course, if you develop a new treatment for rabies (a disease that is almost uniformly fatal without prompt treatment) and your first ten patients all survive, you may have enough evidence to prove your case. But, most of the health problems that plague the twenty-first century world do not have a single cause; most of the outcomes don’t occur quickly or predictably; and no intervention is expected to be near 100% effective. Rather, problems like obesity, violence or substance abuse have multiple interrelated causes; outcomes occur over a period of years; and an intervention that reduced the prevalence of any of these conditions by 15% would be heralded as a major breakthrough.

There are a wide variety of study designs available for evaluating the effectiveness of interventions, each with their own advantages and disadvantages. Studies may be randomized or nonrandomized, prospective or retrospective, clustered or nonclustered. Investigators may use complicated analytic tools such as time-series analysis or multivariable modeling, or simply compare percentages. Regardless of what study designs and analytic tools are employed, determining whether the intervention works will be based on answering one (or more) of the following three questions:

1 Is the post-intervention assessment significantly different from the pre-intervention assessment?

5 The term evaluation encompasses a broad set of activities including assessing whether a problem exists, how well a program is functioning (e.g., number of clients being served, length of time it takes to serve a client, client satisfaction with the service), what the program costs, etc. The focus of this book is the efficacy or effectiveness of the intervention. For an excellent book on the full spectrum of evaluation activities see: Berk, R. A. and Rossi, P. H. Thinking about Program Evaluation. Thousand Oaks: Sage Publications, 1999; also Shadish, W. R. “The common threads in program evaluation.” Prev Chronic Dis 391 (2006): A03 at http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=16356356.
2 Is the change between the pre-intervention and the post-intervention assessment of the intervention group significantly different than that of the comparison group?

3 Is the outcome for the intervention group significantly different than for the comparison group?

In the next section I will review the data elements you will need to answer each of these questions. Following that I will discuss development of interventions (Chapter 2), evaluation of interventions (Chapter 3), and then compare randomized designs (Chapter 4) to nonrandomized designs (Chapter 5). In Chapter 6 I will return to how to answer these three questions from a statistical point of view.

1.2.A Is the post-intervention assessment significantly different from the pre-intervention assessment?

An intuitively simple way of determining whether an intervention works is to assess a group of people prior to it (pre-intervention) and again after it (post-intervention) (Figure 1.1). This is referred to as a one-group pre-intervention versus post-intervention design.

If you are assessing the same individuals on more than one occasion, your study is a longitudinal cohort study. If you are sampling from the same population but not necessarily the same individuals on more than one occasion, you are performing a serial cross-sectional study. By definition one-group pre-intervention versus post-intervention designs are nonrandomized designs because you need at least two groups to randomize assignment.

Regardless of whether you are assessing the same individuals over time, or samples of the same population over time, you will be testing the null hypothesis; in this case, that there is no difference between the two assessments. If the difference between the pre-intervention and post-intervention assessments is sufficiently great that it is unlikely that the difference could have occurred by chance alone, you will consider the alternative hypothesis: that the intervention worked. If the pre-intervention and post-intervention assessments are similar, you will conclude that the null hypothesis was correct: the intervention did not work.

Let’s look at a very important intervention evaluated with a one-group pre-intervention versus post-intervention design. Pronovost and colleagues designed an intervention to decrease catheter-related blood infections among patients cared for in the intensive care unit.6 The intervention included

---

strategies to increase hand-washing by medical personnel, use of barrier precautions, antibacterial cleaning of the catheter site, avoiding use of the femoral site, and removing unnecessary catheters.

Three months after the intervention there were significantly fewer infections (0 infections per 1000 catheter days) than before the intervention (2.7 infections per 1000 catheter days). The probability was small ($P \leq 0.002$) that the change in the infection rate between the pre-intervention and the post-intervention assessments occurred by chance.

To appreciate one of the weaknesses of a pre-intervention versus post-intervention evaluation such as this one, look at the hypothetical data in Figure 1.2. It appears that there has been a major drop in outcomes (e.g., infection, cancer, heart disease) following institution of an intervention.
Before concluding that the intervention worked, look at Figure 1.3. It is the continuation of the same study. You can see that outcomes were back up by year 3 and back down by year 4. Remember: two points do not make a trend.

One way to improve the strength of a single group pre-intervention versus post-intervention design is to get additional data points. For example, if the frequency of outcome had been measured several times prior to the intervention and was stable, and then changed precipitously following the intervention, and the change was sustained at several points after the intervention, as shown in Figure 1.4, you would have much greater confidence that the decrease in outcome was due to the intervention.
Table 1.1. Rate of catheter-related infection at baseline, during the intervention, and after the intervention.

<table>
<thead>
<tr>
<th></th>
<th>Rate of infection (95% confidence intervals)</th>
<th><em>P</em> value for comparison with baseline rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>2.7 (0.6–4.8)</td>
<td>–</td>
</tr>
<tr>
<td>During intervention</td>
<td>1.6 (0–4.4)</td>
<td>≤0.05</td>
</tr>
<tr>
<td>After intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–3 mo</td>
<td>0 (0–3.0)</td>
<td>≤0.002</td>
</tr>
<tr>
<td>4–6 mo</td>
<td>0 (0–2.7)</td>
<td>≤0.002</td>
</tr>
<tr>
<td>7–9 mo</td>
<td>0 (0–2.1)</td>
<td>≤0.002</td>
</tr>
<tr>
<td>10–12 mo</td>
<td>0 (0–1.9)</td>
<td>≤0.002</td>
</tr>
<tr>
<td>13–15 mo</td>
<td>0 (0–1.6)</td>
<td>≤0.002</td>
</tr>
<tr>
<td>16–18 mo</td>
<td>0 (0–2.4)</td>
<td>≤0.002</td>
</tr>
</tbody>
</table>


In the case of the study to decrease infections in the ICU, the investigators had only one measurement of infection prior to the intervention, but they had additional data points: one during the implementation period and five additional points in the post-intervention period (Table 1.1). These data points increase confidence that the intervention worked. When you have lots of consecutive data points over a period of time you can analyze your data using time series analysis (Chapter 8).

However, no matter how many data points you have prior to, during, and after an intervention, with only one group your study has a serious limitation: there is always the possibility that any observed change occurred for a reason other than the intervention. Returning to the example of the study of catheter-related infections in the ICU, perhaps infections decreased due to media attention on hospital infection rates or changes in physician prescribing practices with respect to antibiotics. To overcome this limitation we need a design that includes a comparison group. This design is discussed in the next subsection.

1.2.B Is the change between the pre-intervention and the post-intervention assessment of the intervention group significantly greater (lesser) than that of the comparison group?

Adding one or more comparison groups to a pre-intervention versus post-intervention assessment results in a stronger evaluation design than having a single group. If the subjects are assigned by random to the groups and the subjects are followed prospectively, you have a randomized controlled trial
There are also a number of ways of assembling a comparable control group without randomization (Chapter 5).

Whether the comparison group is assigned randomly or not, the major question is whether the change that occurs between the pre-intervention and the post-intervention assessment is greater (lesser) than the change over the same period of time in a comparison population (Figure 1.5).

To illustrate the benefits of adding a comparison group to a pre-intervention versus post-intervention assessment, let’s look at a study evaluating whether providing hospitals with confidential information on their performance improves the care of patients having coronary artery bypass grafting (CABG). The intervention was performed at 20 hospitals in Alabama.

In Table 1.2 you can see the impact of the intervention on four important process measures. Following the intervention the rate of internal mammary artery use, the percentage of patients discharged on aspirin, and the percentage of patients who were intubated for less than 6 hours increased, and the median time patients were intubated decreased. But as with the study of ICU catheter infections, is it possible that these changes occurred for some reason other than the intervention? Indeed, three and a half years passed between the start of the pre-intervention assessment and the end of the post-intervention assessment. Perhaps improvements in surgical technique or medical practice during these years are the actual cause of the observed improvement in practice.

The investigators addressed this concern by including CABG patients from a comparison state. As you can see in Table 1.3 the changes over time were more

---

Introduction

Table 1.2. Impact of a quality improvement program in Alabama hospitals.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre-intervention</th>
<th>Post-intervention</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal mammary artery use, %</td>
<td>73</td>
<td>84</td>
<td>( \leq 0.001 )</td>
</tr>
<tr>
<td>Aspirin therapy, %</td>
<td>88</td>
<td>92</td>
<td>( \leq 0.001 )</td>
</tr>
<tr>
<td>Intubation time, median, h</td>
<td>12</td>
<td>7</td>
<td>( \leq 0.001 )</td>
</tr>
<tr>
<td>Intubation &lt;6 h, %</td>
<td>9</td>
<td>41</td>
<td>( \leq 0.001 )</td>
</tr>
</tbody>
</table>


Table 1.3. Quality indicators for CABG surgery in Alabama compared to another state.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Alabama Pre-intervention</th>
<th>Alabama Post-intervention</th>
<th>Comparison state Baseline</th>
<th>Comparison state Follow-up</th>
<th>P value for difference between states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal mammary artery use, %</td>
<td>73</td>
<td>84</td>
<td>48</td>
<td>55</td>
<td>( = 0.001 )</td>
</tr>
<tr>
<td>Aspirin therapy, %</td>
<td>88</td>
<td>92</td>
<td>86</td>
<td>82</td>
<td>( &lt;0.001 )</td>
</tr>
<tr>
<td>Intubation time, median, h</td>
<td>12</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>( &lt;0.001 )</td>
</tr>
<tr>
<td>Intubation &lt;6 h, %</td>
<td>9</td>
<td>41</td>
<td>40</td>
<td>39</td>
<td>( &lt;0.001 )</td>
</tr>
</tbody>
</table>


favorable in the intervention state (Alabama) than in the comparison state. Temporal improvements in medical practice would have likely influenced the outcomes in the comparison state as well.

Note also from Table 1.3 the importance of having two measurements from both of the states. If you had only the second measurements for the median intubation time and for the percentage of patients who were intubated for less than six hours, you might conclude that there was no difference between the two states. If you had only the second measurements for internal mammary artery use you would think that the intervention was a much greater success that it was – even before the intervention, the use of the internal mammary artery as part of the CABG procedure was substantially higher in Alabama hospitals than in the comparison state.

Although the inclusion of a comparison group (in this case concurrent controls) greatly strengthens the model, there remains an important potential weakness. Is the comparison group comparable to the intervention group? As you can see from Table 1.4, patients from the intervention group and the comparison group are similar on most characteristics at both baseline and follow-up assessments. Nonetheless there are differences (for example, the percentage with left main disease is lower in Alabama than in the comparison state).
Could some other difference between Alabama hospitals and those in the comparison state explain the differences that we are attributing to Alabama’s intervention? Yes. The only way to be certain that there are no important differences between the intervention and the comparison group is to randomize assignment to the groups (Chapter 4).

Still, it is important to note that this intervention, which had an important impact, would have been very difficult to perform using a randomized design. Randomization of individual patients would have been impossible because the intervention was being performed on the level of physicians and hospitals.

A clustered randomization design (Section 4.5), where the hospitals were randomized to be in the intervention, would have been superior from a design point of view. However, this study was possible because of a statewide initiative in Alabama to improve the care of patients receiving CABG in all hospitals and had the cooperation of the state peer review organization. The comparison state only had to agree to release medical record information.

Randomizing hospitals may have created another problem: it may have increased the diffusion of the intervention to non-intervention hospitals because physicians often work at more than one hospital and share information through local and state professional associations. Diffusion of the intervention would have been good for care of patients receiving CABG but would

---

**Table 1.4. Characteristics of patients receiving a CABG in Alabama compared to patients receiving a CABG in another state**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Alabama Pre-intervention</th>
<th>Alabama Post-intervention</th>
<th>Comparison state Baseline</th>
<th>Comparison state Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>4090</td>
<td>1694</td>
<td>2288</td>
<td>926</td>
</tr>
<tr>
<td>Mean age, y</td>
<td>69.9</td>
<td>70.7</td>
<td>70.6</td>
<td>71.4</td>
</tr>
<tr>
<td>Male, %</td>
<td>65</td>
<td>55</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>White, %</td>
<td>91</td>
<td>91</td>
<td>94</td>
<td>93</td>
</tr>
<tr>
<td>CAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left main, %</td>
<td>16</td>
<td>19</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td>MI within 3 days, %</td>
<td>9</td>
<td>14</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>MI within 6 months, %</td>
<td>22</td>
<td>26</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>CHF</td>
<td>16</td>
<td>22</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Poor LVF, %</td>
<td>26</td>
<td>29</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Cardiogenic shock, %</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>COPD, %</td>
<td>25</td>
<td>30</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>Diabetes mellitus, %</td>
<td>29</td>
<td>32</td>
<td>27</td>
<td>32</td>
</tr>
<tr>
<td>Dialysis, %</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

have made it difficult to demonstrate that the intervention worked. Having the intervention performed in one state and compared to performance in another state minimized contamination of the comparison community with the intervention (Section 3.6).

A final note on this study: you may have been surprised when you saw from the $N$ of patients listed in Table 1.4 that the same patients are not in both the baseline and follow-up. That’s because this is a serial cross-sectional study. The advantages and disadvantages of serial cross-sectional designs are discussed in Section 3.7 and the statistics for analyzing serial cross-sectional data are reviewed in Section 6.3 and 6.5.

1.2.C Is the outcome for the intervention group significantly different than for the comparison group?

Sometimes we need to evaluate an intervention for which no pre-intervention measurement has been made. For example, laws and organizational practices are often changed without anyone planning an evaluation. With ingenuity it may be possible to identify some pre-intervention measures, perhaps by identifying a prospective cohort that was operating before the intervention, or assembling a retrospective cohort by using existing records, such as an electronic health record (see Section 3.4 for suggestions). But what if you cannot identify any pre-intervention measurements? In such circumstances it may be possible to compare a post-intervention assessment to an assessment of a comparison group (Figure 1.6).

For example, beginning in 1979, a number of local laws were passed restricting tobacco smoking in public workplaces (e.g., restaurants, stores) in California. Moskowitz and colleagues evaluated whether these ordinances resulted in

![Figure 1.6](image-url)
people quitting smoking. In 1990 they performed a cross-sectional survey of adult indoor workers. Among persons working in areas with strong no-smoking ordinances 26% of smokers quit in the prior six months, while only 19% of smokers quit in areas with no ordinances. The probability that this difference occurred by chance was small ($P < 0.001$).  

A major weakness of this design is that without a pre-intervention assessment, we cannot test for or adjust for pre-existing differences between the groups. Perhaps persons living in the areas where tobacco smoking was restricted were more likely to quit smoking anyway, for some reason other than the workplace ordinances. For example, maybe there is stronger community pressure to not smoke and that is the reason the bans were passed and people quit. This is a fair criticism, and one not easily refuted. This is why intervention studies without a pre-intervention measurement are much weaker. Nonetheless, the study added to a growing body of research indicating that smoking bans have positive impacts on community health.  

Another way to evaluate an intervention without a pre-intervention assessment is to perform a case-control study. With a case-control study, you identify people with a disease or condition and a control group of persons without the disease or the condition, and then assess what percent of each group had the intervention (Figure 1.7). If the percentage of persons having the intervention is significantly different in the two groups, you have suggestive evidence that the intervention works.

---


10 Case-control methods may be more familiar to you from studies of risk factors/exposures of disease, but the method works for evaluating interventions as well.
For example, Loomis and colleagues used a case-control study to evaluate the effectiveness of safety measures for preventing homicides at workplaces. Using a statewide database from the medical examiner system they identified 105 homicides that occurred at a work site. Next they identified 210 workplaces from a business phone listing where no homicide had occurred. To determine whether the safety measures were effective in preventing homicides, they asked informants at the workplaces whether a variety of security measures were in place during the month the homicide had occurred (for cases) and for the same

---

Table 1.5. Effect of environmental control measures on the risk of homicide in North Carolina workplaces.

<table>
<thead>
<tr>
<th>Control Measures</th>
<th>Cases (N = 105)</th>
<th>Controls (N = 210)</th>
<th>Odds Ratio (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lighting and visibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bright lighting outside building</td>
<td>28%</td>
<td>33%</td>
<td>0.8 (0.5–1.3) 0.5 (0.3–1.0)</td>
</tr>
<tr>
<td>Bright lighting inside during work hours</td>
<td>57%</td>
<td>64%</td>
<td>0.7 (0.5–1.2) 0.7 (0.3–1.3)</td>
</tr>
<tr>
<td>Bright lighting inside before/after hours</td>
<td>37%</td>
<td>41%</td>
<td>0.8 (0.5–1.4) 0.9 (0.4–1.7)</td>
</tr>
<tr>
<td>Workers visible from outside</td>
<td>61%</td>
<td>60%</td>
<td>1.1 (0.7–1.7) 1.0 (0.5–1.9)</td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barrier between workers and public</td>
<td>83%</td>
<td>88%</td>
<td>0.6 (0.3–1.3) 0.6 (0.2–1.8)</td>
</tr>
<tr>
<td>Security and surveillance devices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video cameras</td>
<td>25%</td>
<td>16%</td>
<td>1.8 (1.0–3.2) 1.3 (0.6–2.8)</td>
</tr>
<tr>
<td>Alarms</td>
<td>19%</td>
<td>26%</td>
<td>0.7 (0.4–1.2) 0.5 (0.2–1.0)</td>
</tr>
<tr>
<td>Mirrors</td>
<td>30%</td>
<td>21%</td>
<td>1.6 (0.9–2.8) 0.8 (0.4–1.8)</td>
</tr>
<tr>
<td>Other devices</td>
<td>24%</td>
<td>25%</td>
<td>1.0 (0.6–1.7) 0.8 (0.4–1.7)</td>
</tr>
<tr>
<td>Cash storage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drop box</td>
<td>21%</td>
<td>21%</td>
<td>1.0 (0.5–2.0) 0.5 (0.2–1.3)</td>
</tr>
<tr>
<td>Safe with worker access</td>
<td>46%</td>
<td>48%</td>
<td>0.9 (0.5–1.6) 1.3 (0.6–2.9)</td>
</tr>
<tr>
<td>&gt; Five environmental control measures</td>
<td>34%</td>
<td>40%</td>
<td>0.8 (0.5–1.3) 0.5 (0.2–1.0)</td>
</tr>
</tbody>
</table>

* Adjusted for a priori high-risk industry, residential or industrial location, open Saturdays, workforce race and sex, relocated or opened in last 2 years. Ns vary for the different measures; see original article for actual Ns.

month for the controls. Cases and controls were then compared on whether the safety measure was in place. As you can see in Table 1.5, having bright light outside the building, having alarms, and having greater than five environmental control measures were associated with a lower likelihood of a homicide (last column) after adjustment for a variety of factors.

A major limitation of case-control studies such as this is that you are relying on reporting of the past by cases and controls, and people’s memories can be affected by events that have occurred to them. For example, informants at workplaces that experienced a homicide might feel that their workplace was unsafe (since a homicide had occurred) and under-report workplace safety measures. This problem is eliminated in nested case-control studies (case-control studies that are performed within an ongoing cohort) because cases and controls are interviewed prior to the development of the disease or condition. But in the real world there are circumstances where no pre-intervention measurement exists and you have to do your best to evaluate the intervention.
Interventions

2.1 What type of interventions are commonly evaluated in medicine and public health?

A wide variety of interventions can be evaluated. Table 2.1 is meant to be illustrative, perhaps even inspirational, but not exhaustive.

I will discuss more about how to design interventions in the next section and then in Chapter 3 take up the question of how to evaluate interventions.

2.2 How do I design an intervention?

The first step in designing an intervention is to have a thorough understanding of the problem you are trying to solve. Read the published literature. Identifying published articles has been made easy and free by PubMed (http://www.ncbi.nlm.nih.gov/Pubmed). Perform a Google search as well because sometimes interesting work is unpublished but available on the net. Talk to others in your field and on the ground. Involving community experts and stakeholders is a great way of learning more about a problem and potential solutions (Section 3.1).

The process of researching the problem will likely also lead you to solutions that have already been tried. When possible, build on what has already been learned. Although the universe of possible interventions is broad, I offer the following considerations that I believe will increase the chance that your intervention will be a success (Table 2.2).

2.2.A Base intervention on theory and/or pathophysiology

An intervention is more likely to be effective if it is grounded in theory. For example, Hawkins and colleagues developed an intervention to decrease adolescent high-risk behaviors such as violence, crime, substance abuse, and sexual activity. The intervention was based on the social development

Evaluating Clinical and Public Health Interventions

Table 2.1. Types of interventions that can be evaluated.

<table>
<thead>
<tr>
<th>Type of intervention</th>
<th>Example</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigator-developed interventions</td>
<td>Develop an intervention to increase exercise among young people and evaluate whether youth exposed to the intervention exercise more than those who are not exposed.</td>
<td>Can ensure that design of intervention and evaluation fit well together.</td>
<td>Most expensive, need to consent people for the intervention as well as for the evaluation.</td>
</tr>
<tr>
<td>Quality improvement program</td>
<td>Evaluate infection rates after placing antimicrobial gel dispensers outside patient rooms.</td>
<td>No need to create or pay for intervention; may not need to consent participants (depending on the outcome measure used).</td>
<td>Not possible to randomize participants; may be difficult to identify a comparable group that has not experienced the intervention.</td>
</tr>
<tr>
<td>Introduction of a new medication, procedure, or device into practice</td>
<td>Evaluate whether drug-eluting stents are superior to regular stents.</td>
<td>No need to create or pay for intervention.</td>
<td>The first group to receive an intervention may differ from others (e.g., may be sicker); also subject to survival bias (i.e., subjects have to live long enough to get treatment).</td>
</tr>
<tr>
<td>Change in style or practice</td>
<td>Evaluate whether adding menu labeling improves diets.</td>
<td>No need to create or pay for intervention.</td>
<td>May only have historic controls available.</td>
</tr>
<tr>
<td>Evaluating a law or government regulation</td>
<td>Evaluate smoking rates following passage of smoking bans.</td>
<td>No need to create or pay for intervention; laws and regulations often have very broad implications and can be implemented in other localities.</td>
<td>Generally have to rely on administrative data for pre-law/regulation data.</td>
</tr>
<tr>
<td>Natural or man-made occurrence</td>
<td>Gambling casinos open on an Indian reservation.</td>
<td>Natural experiments can produce conditions that would be unethical to create as part of a research project.</td>
<td>Cannot plan for natural experiments.</td>
</tr>
</tbody>
</table>

An intervention is more likely to be effective if it is grounded in theory.

model. According to the model, specific social factors (e.g., parental involvement, social participation skills) influence children’s bonds to their school; strong bonds to school, in turn, protect youth against engaging in problematic behaviors.

Based on this theory, the intervention targeted teachers, parents, and the children themselves during early childhood years (grades one to six). Teachers
Interventions were instructed on methods of increasing the involvement of children in the class and on helping children learn better ways of interacting with each other. Children received social instruction in grade six. Parents were offered classes in child behavior appropriate to the age of their children.

The investigators found that children who had received the intervention were less likely to report violent behavior and heavy drinking by the age of 18 years than children who had not participated in the intervention.

Grounding an intervention in a theory increases our confidence that the result is not spurious. Why? Because our expectation prior to the study (in Bayesian terms, the prior probability of success) is higher that the intervention will work if it is based on an accurate theory. Conversely, if an intervention is not based on a theory or prior experience, we worry more that a statistically significant finding might be spurious.

Grounding an intervention in a theory also increases the generalizability of a study result. Generalizability refers to the likelihood that the intervention would have a similar affect among other persons in other settings. How does theory improve generalizability? If a theory is true then changes in populations and settings should not change the overall impact of the intervention. (Think of the theory of relativity – it applies throughout the universe.)

As has been pointed out by others, the major limitation to basing interventions on theory is how few theories there are that can explain complex human behavior.

---

Table 2.2. Considerations for developing interventions.

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base intervention on theory and/or</td>
<td>Increases chances of intervention succeeding, increases chance that positive findings are real, improves generalizability.</td>
</tr>
<tr>
<td>pathophysiology</td>
<td></td>
</tr>
<tr>
<td>Build on prior experience</td>
<td>Allows for sample size calculation, identifies potential weaknesses that can be addressed prior to initiation of study, increases chances of intervention succeeding, increases chance that positive findings are real.</td>
</tr>
<tr>
<td>Make your intervention practical</td>
<td>Determines likelihood that intervention will be sustained or translated into standard practice.</td>
</tr>
<tr>
<td>Cost</td>
<td></td>
</tr>
<tr>
<td>Availability of expertise</td>
<td></td>
</tr>
<tr>
<td>Willingness of persons to participate</td>
<td></td>
</tr>
</tbody>
</table>

---

For most medical studies, the “theory” is the pathophysiology of a disease. To the extent that the impact of the intervention on the physiology of the disease can be predicted ahead of the study, the intervention is more likely to be successful and the findings of your study (assuming they are in the predicted direction) are more likely to be believed.

For example, the ADVANCE trial was designed to test whether a combination of an ACE-inhibitor and a thiazide diuretic could reduce vascular complications among type 2 diabetics. Although most studies aiming at preventing the vascular complications of diabetes focus on reducing glucose levels, the investigators knew that reducing blood pressure also reduces vascular complications. The average systolic blood pressure drop was 5.6 mmHg among patients randomized to receive an ACE-inhibitor. Of greater significance, the investigators found that major macrovascular and microvascular complications were reduced by 9% among those randomized to receive an ACE-inhibitor and a thiazide diuretic compared to those randomized to placebo. The effect of ACE-inhibitors on vascular complications among diabetics in the ADVANCE trial was so strong that it has become standard practice to treat all diabetics with evidence of renal disease with an ACE-inhibitor even if they are not hypertensive.

Of course, just because you think that you understand how an intervention would improve a disease is no guarantee that you are right. For example, there were detailed physiologic explanations of why estrogen would be expected to decrease the incidence of heart disease in women, including that estrogen is known to decrease low-density lipoproteins (the bad kind of cholesterol), increase high-density lipoproteins (the good kind of cholesterol), improve endothelial vascular function, and lower fibrinogen and plasminogen levels. These explanations were in turn supported by a number of observational studies showing that estrogen decreased heart disease in women. Unfortunately, when a large blinded multicenter randomized placebo-controlled trial was conducted, estrogen plus progestin increased cardiovascular disease.

2.2.B Build on prior experience

An intervention that has succeeded previously is more likely to be successful. Of course, we wouldn’t have major breakthroughs if we only extended previously successful interventions.

If your intervention is novel, conduct a pilot study before performing a full-scale implementation. There are several reasons for this. First, as a practical matter, it is impossible to conduct a sample size calculation (Section 3.12) unless you can estimate the effectiveness of an intervention. Second, pilot studies are particularly helpful in pretesting measurement instruments and recruitment strategies. Third, a pilot study is likely to help you recognize potential weaknesses in your design. For example, you may find that you have an unacceptably high drop-out rate in the intervention arm. This may lead you to modify the intervention. Or perhaps you will find that your outcome measurement does not fully capture the effects of the intervention.

2.2.C Make your intervention practical

Ultimately the best reason for developing and evaluating an intervention is to identify those that should be broadly implemented.

The traditional paradigm for developing, evaluating, implementing, and translating interventions is shown in Figure 2.1. In the first stage an intervention is tested to see whether it works in a research setting (efficacy trial). The teaching has been that in an efficacy trial the intervention should be delivered in a uniform fashion by trained research personnel. Subjects are enrolled, closely followed, and may receive stipends for their participation.

Efficacy trials typically enroll homogeneous populations because subjects who are atypical (e.g., over 90 years of age) may respond differently to the treatment. Related to this, many efficacy trials exclude persons who are very ill or thought to be at greater risk of harm from the intervention or who have a deficit (e.g., dementia) that results in their being unable to comply with all aspects of...
Evaluating Clinical and Public Health Interventions

the study. Because the focus of an efficacy trial is to eliminate confounding, it is expected that the controls will be randomized.

Following the traditional paradigm, if an intervention works in an efficacy trial, it should then be tested in “real-world” conditions (effectiveness trial). In an effectiveness trial subjects may be enrolled or passively followed. Effectiveness trials should include a broad range of subjects and the practitioners who deliver the intervention should be similar in skills to the people who will be implementing the intervention in the real world.

Effectiveness trials may randomize subjects to an intervention and a control group, or may have a nonrandomized comparison group, or may have no comparison group at all. The reason you may not need a control group is that once you have proven that an intervention is more efficacious than standard treatment in a randomized trial, you can focus your efforts in the effectiveness trial on simply demonstrating that your intervention works as well in real-world conditions as it did in the trial.

If an intervention is found to work in an effectiveness trial, then it can be translated into the standard of practice.

There are several problems with this paradigm. First, those interventions that are most likely to succeed in efficacy trials are not necessarily the same interventions that are likely to succeed in real-world conditions. For example, interventions that are expensive, time-consuming, or require highly trained staff may be very successful in an efficacy study but unsuccessful in an effectiveness study. Glasgow and colleagues have pointed out that this has resulted in a disconnect in what was supposed to be a linear transition between efficacy and effectiveness studies. The authors suggest a number of ways to close the gap. Without detailing their thoughtful argument, which I hope you will read, much of what they are calling for is performing intervention studies at both the efficacy and effectiveness stages that are likely to be implementable.

To be certain that you do not develop and evaluate an intervention that will never be adopted consider: expense, expertise, and interest in participation among those needing the intervention. Further suggestions on how to translate your research findings into standard practice are included in Chapter 10.

Interventions that are extremely expensive or that require expertise not generally available are difficult to translate from research settings to the real world. For example, expensive medications to treat common diseases in the underdeveloped world, such as malaria, are unlikely to receive broad use, especially

---

Interventions

If a large proportion of eligible subjects are excluded from an intervention trial, it is probably difficult to implement it in a real-world setting.

if they require technologically sophisticated laboratory studies to monitor for toxicity.

Some interventions are difficult to translate because a large number of persons who need the intervention are ineligible to participate. People may be ineligible to participate because it requires them to be in a particular state of health (e.g., not be housebound) or have certain life circumstances (e.g., telephones). For example, chronic disease management interventions involving mail or phone reminders would not be practical for caring for the homeless.

One helpful indication of the practicality of an intervention is the proportion of potential subjects who are excluded from the study. Interventions that exclude a large percentage of persons who need the intervention are not very practical, even when they are effective. For example, Wolchik and colleagues developed two theory-grounded interventions to decrease mental health and behavioral problems among adolescents with divorced parents. One intervention entailed mothers participating in 11 group sessions and two individual sessions; the other intervention involved these same 13 sessions for the mothers and another 11 group sessions for the children; the control group received books on postdivorce adjustment.

The investigators found that compared to adolescents in the control group, those in the two intervention groups had fewer symptoms of mental disorders, and less alcohol and drug use six years after the intervention. These are important benefits and it is heartening that the effect was long-lasting. But will the intervention be widely implemented? I think not.

Here’s the reason. Of 1816 families randomly selected from divorce-court records or identified via referral or response to the media, 1331 were contacted by phone. Of these only 671 met the inclusion criteria of the study (e.g., neither the mother nor any child in the family needed treatment for a mental health problem; mother had no live-in boyfriend or plans to remarry during the study; mother and child were fluent in English). Of these 671 families, 453 completed the recruitment visit, 341 agreed to participate, and 315 completed the pretest. At the pretest 49 more were found to be ineligible and 26 refused to participate after the pretest, resulting in 240 families being randomized to one of the three groups; an additional 26 families withdrew after being assigned. Therefore, only 214 (12%) of divorced families with children who were originally identified actually completed the trial.

For this trial, families were paid for their participation (although not for the actual counseling sessions); in the real world people are not offered financial

---

Evaluating Clinical and Public Health Interventions

incentives, which may result in lower participation. So while the evaluation demonstrates compelling benefits for those who participate, it is an open question how many divorced families would participate if it were available. Even if we had a strategy to convince all divorced families to participate, the results of the study cannot be generalized to all divorced families because only a small percentage of identified families participated.

Although the cost of the program is not listed, it is likely to be very expensive because it included a large number of sessions and extensively trained master’s degree-level clinicians led the sessions. Even if enough funding were provided, it may not be possible to identify a sufficient number of master’s degree clinicians to deliver the intervention.

My comments about this study may seem critical, but I specifically chose it to illustrate the challenges of translating research findings because I thought it was a good intervention for a common problem with a strong outcome. Although it can be difficult to know how easily an intervention will translate into everyday practice, consider issues of translation right from the start and try to create an intervention that can be broadly implemented.
Evaluating an intervention

3.1 How do I engage stakeholders?

If you have developed an intervention, you have ideally involved stakeholders in the creation of it. But whether you are evaluating an intervention you developed or whether you are evaluating an intervention developed by someone else, it is critical to engage stakeholders.

Depending on the program you are evaluating, the stakeholders may be community members, businesses, advocacy groups, patients, teachers, parents, or elected officials. Stakeholders are an invaluable part of any evaluation. They can lead you to the important questions, problems, and pitfalls that will otherwise ambush the unsuspecting researcher; they can help you to develop your intervention, pretest it, enroll subjects, and obtain cooperation from necessary institutions, including writing support letters for funders; and they can help you to interpret your data and disseminate your results.

It is often helpful to establish a community advisory committee for your project, as well as having open forums so that persons who are interested but don’t have the time to participate on a regular basis can be involved. Monthly meetings with food are a good way to keep ideas and information flowing back and forth. If you receive funding for your study, provide stipends for committee members; it demonstrates that you value their time equally to your own.

3.2 What data should I collect for evaluating my intervention?

If you are designing an intervention you have a tremendous advantage over those who attempt to evaluate interventions they have not created (Section 3.4). You have the opportunity to embed the evaluation within the intervention.

At a minimum, collect data prior to the intervention, during the intervention, and after the intervention (Table 3.1). Often it is better to collect multiple

---

Evaluating Clinical and Public Health Interventions

Table 3.1. Important data collection points for evaluating interventions.

<table>
<thead>
<tr>
<th>Needed data points</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple measurements of the outcome prior to the intervention.</td>
<td>Determine whether there is a trend prior to the intervention being implemented.</td>
</tr>
<tr>
<td>Multiple measurements of the outcome during the intervention.</td>
<td>Determine a dose response to the intervention.</td>
</tr>
<tr>
<td>Multiple measurements of the outcome after the intervention.</td>
<td>Determine whether the impact of the intervention wanes, stays the same, or grows.</td>
</tr>
</tbody>
</table>

...data points at these three critical points. Multiple data points prior to the intervention determine whether there is a temporal trend prior to the intervention (e.g., perhaps hospital infections were already decreasing prior to your intervention); multiple data points during the intervention provide a measure of dose response (e.g., as participants have greater exposure to the intervention is there greater change?). Multiple data points after the intervention tell you whether the impact of the intervention diminishes, increases, or stays the same over time.

3.3 How do I choose a control group for evaluating my intervention?

Besides planning your data collection points, another major planning issue is to determine whether you will have a control group. Although important evaluations can be conducted without a control group using a one-group pre-intervention versus post-intervention design (Section 1.2.A), having a control group strengthens all designs.

The different types of control group with their advantages and disadvantages are listed in Table 3.2.

The best control group is created by randomizing all eligible subjects and observing the controls concurrently. This ensures that group membership is unbiased. However, at times it may not be ethical or feasible to withhold the intervention from some of the subjects (e.g., parents may not consent to their children being randomized to receive or not receive an intervention designed to prevent substance use). Even if it is ethical, it may be hard to motivate people to remain in a trial where they know they are not receiving an intervention (this latter problem can be minimized in medication studies where a placebo can be provided – placebos are harder to construct for other types of interventions).

Waiting-list controls are a practical compromise when it is not possible to randomize subjects to receive or not receive an intervention. Subjects may be randomized to receive an intervention or to be a waiting-list control with the promise that the controls will receive the intervention as soon as the first group...
Evaluating an intervention has received it. For example, Stein and colleagues conducted a randomized study of cognitive behavioral treatment for students who reported exposure to violence and had clinical symptoms of posttraumatic stress disorder (PTSD).\(^2\) Given that all the prospective subjects had PTSD it would be hard to not offer them treatment. Therefore, the investigators randomized subjects to the intervention or to a waiting-list group. The investigators found that students who received the intervention had significantly fewer symptoms of PTSD, less depression, and better psychosocial function than those in the waiting-list group. After the waiting-list controls received the same intervention, the two groups were similar with respect to all three measures.\(^3\) Unfortunately, use of waiting-list controls does not allow for blinding or placebos.


\(^3\) If instead of randomizing subjects to an intervention versus waiting-list control, you randomize them to one of two different interventions, and then switch them to the other intervention after a wash-out period, you are performing a crossover design. For more see: Katz, M. H. \textit{Study Design and Statistical Analysis: A Practical Guide for Clinicians}. Cambridge: Cambridge University Press, 2006: pp. 18–19.

<table>
<thead>
<tr>
<th>Type of control</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized control</td>
<td>Groups are comparable on measured and unmeasured characteristics. Allows the use of placebos and blinding.</td>
<td>May not be ethical or practical to randomize.</td>
</tr>
<tr>
<td>Waiting-list control (randomized)</td>
<td>All the advantages of randomized controls. Subjects know they will receive the intervention.</td>
<td>Cannot blind subjects or investigators to whether the subject is receiving the intervention.</td>
</tr>
<tr>
<td>Waiting-list control (non-randomized)</td>
<td>Fits clinical practice (when there is limited supply people have to wait). Only people eligible for the intervention and desirous of it will be included.</td>
<td>Same as randomized waiting list plus people who receive the intervention right away may be different from those waiting for it.</td>
</tr>
<tr>
<td>Concurrent control, enrolled in study</td>
<td>Controls do not need to receive the intervention, which may make the study easier and cheaper</td>
<td>Subjects who receive the intervention may be different from those who don’t. Subjects may not feel motivated to participate in a study for which they derive no benefit.</td>
</tr>
<tr>
<td>Concurrent control, passively followed (e.g., medical record review)</td>
<td>No effort required by controls.</td>
<td>Subjects who receive the intervention may be different from those who don’t. With passive follow-up, may lose track of subjects and not know what happened to them.</td>
</tr>
<tr>
<td>Historic control</td>
<td>Data may have already been collected.</td>
<td>Subjects who receive the intervention may be different from those who don’t. Temporal changes add another source of bias.</td>
</tr>
</tbody>
</table>
At times it may not be possible to randomize subjects to be in a waiting-list group (e.g., government-sponsored services). However, if there is enough room only for a certain number of subjects to receive the intervention at any one time (e.g., an after-school program for children) subjects placed on the waiting list may be willing to be observed as part of an evaluation of the intervention.

For example, studies of the effectiveness of organ transplantation often compare individuals who receive a transplant to those who are on a waiting list for a transplant. Venstrom and colleagues used a transplant waiting list to assess the effectiveness of pancreas transplantation in patients with diabetes and preserved kidney function.\(^4\) Using a time-dependent covariate (see Section 9.1), they demonstrated that persons receiving a solitary pancreas transplant had higher mortality (RR = 1.57; 95% CI 0.98–2.53) than persons who were waiting for a solitary pancreas transplant.

The potential bias when using waiting-list controls that are not randomized is the possibility that persons who receive the intervention are different than those who do not. In the case of the pancreas transplantation study, those who received a pancreas transplant and those left on the waiting list were indistinguishable with respect to a number of prognostic factors including age, sex, and comorbid illness. But non-white patients were less likely to receive transplants than white patients. Although it is possible to statistically adjust for baseline differences using multivariable analysis (Chapter 7), in the absence of randomization we can never exclude the possibility that our results are biased by unmeasured characteristics.

On the other hand, even when a waiting list is nonrandomized, using patients from a waiting list has an advantage over other nonrandomized controls. Specifically, patients drawn from a waiting list are by definition eligible for the intervention and desirous of it. Returning to the example of pancreas transplant, there are patients with a similar need for pancreas transplant who would not be eligible for transplant (e.g., actively drinking) or do not want a transplant. A control group drawn from administrative records of patients with diabetes and preserved kidney function would include such persons, but such a group would not necessarily be comparable to those patients receiving transplants.

When waiting-list controls are not possible the next best alternative is to identify a group of subjects who are willing to be observed concurrently. Unfortunately, identifying concurrent controls can be a challenge for several reasons.

---

Evaluating an intervention

First, if they are not receiving any intervention, it may be hard to motivate subjects to participate, especially if they must devote a lot of energy to the study (e.g., completing complicated interviews or questionnaires, obtaining laboratory tests, etc.). This problem can sometimes be resolved by financially compensating subjects for their time with a stipend or by appealing to the desire to help others by advancing science.

What is harder to deal with is the likelihood that the subjects will inherently be different from those who agree to receive the intervention. In some cases, measuring the control and intervention subjects before the intervention helps to establish a baseline that can be used to adjust for baseline differences. Also, a variety of multivariable techniques can be used to adjust for baseline differences, but none of these techniques can completely eliminate bias.

An advantage of concurrent controls is that they may be less expensive than waiting-list controls if the intervention is expensive to conduct.

The challenges of concurrent controls may be less if they can be passively observed through review of a record, such as a medical record. For example, Two Feathers and colleagues evaluated a lifestyle intervention for African-Americans and Latinos with type 2 diabetes. In designing the intervention the investigators worked closely with the affected community. Because of historic distrust of research, it was decided to perform the study without a randomized control group. Instead they evaluated the efficacy of the intervention by comparing glycosylated hemoglobin (Hb\textsubscript{A1c}) values of participants to the values abstracted from a random sample of diabetics receiving care in the same health system. This design was possible because Hb\textsubscript{A1c} assessments are uniformly ordered for diabetic patients. However, they did not have a control group for the other outcomes of their study (dietary knowledge, following a healthy eating plan, vegetable consumption). For these outcomes the investigators used a one-group pre-intervention versus post-intervention design.

Historical controls are likely to be different from your intervention subjects both because they are not part of the intervention group and because of temporal changes (even if their characteristics are exactly the same as the intervention group they were living in an earlier period). On the other hand, historic controls have one potential advantage. If a group is observed prior to an intervention being available, they could not have received the intervention. Therefore, there is less confounding due to willingness or ability to receive the intervention (there is still some confounding because were the intervention offered some of the historical controls may have said no).

---

For example, Golden and colleagues used historic controls to evaluate the effectiveness of inhaled pentamidine in preventing *Pneumocystis carinii* (jiroveci) pneumonia (PCP) in patients with AIDS and a prior episode of PCP. At the time (1987), there were no other drugs licensed to prevent PCP and in the absence of treatment subjects had a very high risk of developing a second episode. Because recurrent PCP had a high case-fatality rate it was felt that it was unethical to have a placebo control group. Instead the authors demonstrated that subjects who received inhaled pentamidine had a much lower risk of developing a second episode than patients with a prior episode of PCP who had been diagnosed prior to the availability of inhaled pentamidine (Figure 3.1). Although historic controls are generally the weakest controls, this study helped to establish inhaled pentamidine as an effective method of prophylaxis.

### Figure 3.1


Historic controls are the weakest of controls.

3.4 How do I collect data and choose a control group for an intervention that I did not develop?

Ideally, all interventions would be developed with an evaluation plan that included collecting data points prior to, during, and after the intervention as well as identifying an appropriate control group.

Unfortunately, many times an intervention will be launched without an evaluation plan. This is true of many laws, regulations, and other government initiatives.

---

governmental initiatives, which are often passed on the assumption that they will be beneficial.

Although challenging, evaluating interventions for which no evaluation was planned can be fruitful. For example, many of the smoking bans (e.g., no sales of tobacco to youth, no smoking in restaurants) were implemented locally, found to be effective, and were then instituted in other localities.

Even if you have not developed an intervention, if you have been closely following the development of a law or practice you may be able to launch a data collection before the intervention. More often investigators need to be creative to identify existing data and controls that can be used to evaluate interventions. Several potential sources of data and/or comparison groups are shown in Table 3.3 along with examples of intervention studies that have made use of these sources.

With the advent of electronic medical records, patient cohorts increasingly serve as a rich source of observations prior to and following an intervention as well as a source of controls. For example, the Veterans Affairs medical centers in the United States have an extensive patient database. Concato and colleagues used it to evaluate the effectiveness of screening for prostate cancer. They identified 71,661 men, 50 years and older, who had an ambulatory visit to any of 10 Veterans Affairs medical Centers in New England between 1989 and 1990 and who did not have a diagnosis of prostate cancer before 1991. From this group they identified 501 men diagnosed with prostate cancer between 1991 and 1995 who died by 1999. From that same cohort they identified 501 controls who did not have prostate cancer and who matched the cases on age and site of diagnosis.

After adjustment for race and comorbidity, the investigators found that screening for prostate cancer was unassociated with mortality (OR = 1.08; 95% CI 0.71–1.64; \( P = 0.72 \)); in other words, screening did not appear to decrease mortality. Note that at the time this study was conducted it would not have been considered ethical to randomize men to prostate screening or not because of the presumption that prostate screening saved lives.

A limitation of patient records as a data source for assessing interventions is that the records (whether electronic or paper) may not have all of the data that you need. Also subjects may be lost to follow-up without any indication of what happened to them.

Missing data and losses of subjects to follow-up are less common with ongoing research cohorts than with cohorts derived from patient records because

---

### Table 3.3. Potential sources of existing data with examples of evaluations of interventions.

<table>
<thead>
<tr>
<th>Source of prior observations</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic medical records</td>
<td>The effectiveness of prostate screening on mortality among patients seen at one of 10 Veterans Affairs medical centers.a</td>
</tr>
<tr>
<td>Ongoing research cohort</td>
<td>The impact of postmenopausal hormone use was assessed among participants of the Nurses’ Health Study.b</td>
</tr>
<tr>
<td>Periodic surveys</td>
<td>The effect of Medicaid managed care on the access to care of uninsured persons was assessed by comparing responses over time to questions from the National Health Interview Survey.c</td>
</tr>
<tr>
<td>Administrative databases</td>
<td>The impact of graduated driver licensing on car crashes among 16-year-old drivers was assessed using 3 years of crash data from Michigan State Police records.d</td>
</tr>
<tr>
<td>Insurance or pharmacy databases</td>
<td>The effect of prior authorization on decreasing the use of long-acting oxycodone was studied by comparing Medicaid prescription drug records of states with prior authorization requirements to those states that did not have such requirements.e</td>
</tr>
<tr>
<td>Disease or procedure databases</td>
<td>Drug-eluting stents were compared to coronary artery bypass grafting by identifying patients from two registries: Cardiac Surgery Reporting System and Percutaneous Coronary Intervention Reporting System of New York.f</td>
</tr>
<tr>
<td>Birth and death registries</td>
<td>The role of national health insurance in improving life expectancy was assessed using population and death registries in Taiwan.g</td>
</tr>
</tbody>
</table>

---


Evaluating an intervention

Subjects are asked to make regular visits and complete standard forms, and investigators make active efforts to stay in contact with study members. For example, one of the largest and most successful cohort studies is the Nurses’ Health Study. Grodstein assessed the effect of postmenopausal hormone use on the risk of colorectal cancer. The investigators took advantage of self-reported data on hormone use and colorectal cancer from biennial questionnaires completed over a 14-year period among 59,002 women. After exclusion of women who reported screening sigmoidoscopy (to avoid bias due to more intensive screening among hormone users), there was a decreased risk of colorectal cancer among current hormone users (RR = 0.64; 95% CI 0.49–0.82). Nested case-control studies such as this one are particularly efficient in terms of time and cost since the data are already collected and usually in a format that makes analysis easy.

Periodic surveys ask the same questions each time of a sample of a population; this allows comparisons of answers prior to and after the implementation of an intervention. For example, the National Health and Nutrition Examination Survey (NHANES) interviews and performs physical examinations on a representative sample of 5000 Americans each year.

Periodic surveys can also be used to assemble a comparison group. For example, the change in an outcome before and after an intervention performed in one state can be compared to changes in the outcome at the same time points in another state where the intervention was not performed.

Municipalities have administrative databases that are a rich and often overlooked source of data. They exist on local, state, and federal levels. Examples of useful municipal databases include reports of traffic accidents, police records, jail/prison records, drug-overdose records, suicide records, medical examiner records, and supplemental income records.

Health services research has exploded with the availability of insurance and pharmacy databases. The most useful of these databases are those that include all persons in a particular group and/or all of the care that those persons receive. For example, Medicare databases include virtually all persons in the United States over 65 years and all of the hospital care they receive regardless of what hospital they were admitted to. Medicaid records (a public insurance program for low-income persons in the United States) can be used to compare policies that vary from state to state. Private health insurance providers have data sets that include all of the care that their members have received. Countries that

have a universal insurance system have claims for the entire country for all of the care their residents receive.

Disease- or intervention-specific databases can be particularly helpful when you need more information than would typically be available in insurance records. For example, insurance records can be effective in identifying people with certain diseases (usually through International Classification of Diseases (ICD) codes) and/or people who received certain procedures. But they do not generally include information about when the person developed the disease or, in the case of procedures, details about them (e.g., how long the patient waited for the procedure, complications associated with the procedure).

Birth and death registries are essential for calculating mortality and life expectancy. Birth registries often have additional data about perinatal outcomes such as birth weight and death registries generally include cause of death, as well as factors contributing to death.

To evaluate some interventions it may be necessary to link several of the databases listed in Table 3.3. For example, Lipscombe and colleagues sought to determine whether the use of thiazolidinediones (TZD) in diabetics was associated with cardiovascular disease. To do so they linked a prescription record database (use of hypoglycemic agents), an ambulatory care database (emergency room records), a hospital database (hospital admission records), a health insurance plan database (physician records), a vital status database (demographics and mortality), and a disease-specific database (diabetes records). They found that use of TZD treatment was associated with increased risk of congestive heart failure, acute myocardial infarction, and death.

All of the databases listed in Table 3.3 can be used to identify controls. However, if you review the different types of controls listed in Table 3.2 you realize that randomized controls will rarely be an option. An exception would be a government intervention that was distributed on a random basis; for example, affordable housing is often distributed based on a lottery of persons who meet the eligibility criteria and submit their application within a window period. A registry of persons who did and did not receive housing by a particular date based on randomization would provide the basis for a very robust evaluation of a housing initiative.

More commonly, nonrandomized controls will be used (such as in the studies of prostate and colon cancer discussed above). For studies of laws or institutional practices, samples of persons from other geographic areas of a town, city, state, or country are a good source of controls.

Evaluating an intervention

I hope that I have convinced you that you can successfully perform an evaluation of an intervention for which no evaluation was planned. Now let me mention two advantages of this type of study. First, they are often relatively inexpensive to perform because you do not have to pay for the intervention. Second, to the extent that the intervention has been widely implemented, a negative study (one that shows no impact of the intervention) has as much impact as a positive study. Why? Because by identifying an existing implementation that is not working, you may save costs for the system by eliminating an ineffective intervention. You also may spur development of new methods where there was previously complacency due to the belief that the existing intervention worked. We will take up the issue of studies with negative results again in Section 9.8.

3.5 What outcomes of an intervention should I assess?

It is common to hear policy and legislative leaders speak of the need to evaluate a particular program to find out whether it works. But if you asked them what would convince them that the program was a success you may be surprised at the variety of responses.

I was meeting with a group of stakeholders – health, business, civic, and union leaders – to develop a universal health access program for my county. Throughout the process many participants spoke of the importance of evaluating whether the program worked. At one of the sessions, I asked them what outcome, five years in the future, would make them believe that the program had worked. Each one had a different answer: more people coming forward for care, higher satisfaction with care, more preventive care, fewer emergency department visits, improved health status, lower costs. Each of these outcomes is potentially important, but each requires a somewhat different approach.

Therefore, the most important step in planning an evaluation of a program is to decide what outcome(s) you are going to assess. In some cases, the ideal outcome may not be feasible. For example, for many medical interventions increasing lifespan is the most important outcome, but using death as a study outcome may not be feasible if the study period is relatively short.

In situations where the outcome of interest takes a long time to occur, you may need to use a proximal marker. A proximal marker is highly predictive of the outcome you wish to study but occurs earlier and therefore results in more outcomes in shorter follow-up times. For example, glycosylated hemoglobin (Hb\text{a1c}) is a marker of diabetes control that predicts complications of diabetes.
3.6 Why measure the exposure of subjects to an intervention?

Those among you who have not yet performed an intervention trial may assume that people who participate in a study and are assigned to an intervention receive it and those not assigned to an intervention do not. If only this were the case, trial results would be so much easier to interpret! More commonly, some subjects in the intervention group don’t receive the intervention and some in the control group do. It is critical to the interpretation of the results of a study to measure exposure to the intervention. Why?

Let’s start with the case of an intervention that appears to work. Showing that people who were assigned to the intervention group actually received the intervention (and those who were not assigned to the intervention did not receive it) strengthens the validity of the study.

Exposure should be measured in terms of the intensity of the exposure (e.g., percent of doses taken, number of visits made, etc.) not just as whether the intervention was received – yes or no. In this way you may test for dose–response relationships (persons who receive a more intense exposure to the intervention should show greater changes). Dose–response relationships strengthen causal inference – the assumption that the intervention produces the demonstrated result.

Alternatively, if most people in the intervention group had no or minimal exposure to the intervention, then the apparent effect of the intervention might be due to some other factor.

In the case of an intervention that does not appear to work, it is important to determine whether it didn’t work because the intervention is ineffective, or because the participants were not sufficiently exposed to the intervention, or because too many people in the control group were exposed to the intervention. Understanding the reason that an intervention did not work is critical to the design of future studies.

For example, O’Loughlin and colleagues evaluated a community-based cardiovascular program performed in a low-income community of Montreal, Quebec.$^{10}$ A nearby community served as the control. At the three-year follow-up there were no declines in smoking or in eating high-fat foods in either the intervention or the control community; physical inactivity actually increased in both communities. However, participation in the program was very low. Participation in the walking club, workshops in healthy eating and weight reduction, videos, and contests were each less than 1.5% of the population.

---

Evaluating an intervention

Given this level of participation, it would not be realistic to expect the intervention to have worked.

It is also not realistic to expect the intervention group to do better than the control group if a lot of the control patients are exposed to the intervention. For example, in a randomized study of an educational intervention to decrease hospital-wide infections (e.g., increased hand-washing, avoiding inappropriate antibiotic use), doctors randomized to no intervention may change their behavior, just due to observing their colleagues. (I’m using the example of hospital-wide infections to help you remember how contamination can occur in a research setting!) To decrease contamination, studies of observable behavior often perform randomization by group, such as by hospital or by school (Section 4.5).

3.7 Should I use a longitudinal cohort or a serial cross-sectional design?

As discussed in Section 1.2.A, there are two ways of performing pre-intervention versus post-intervention assessments: assessing the same people over time (longitudinal cohort) and serial (repeated) sampling from the same population (serial cross-sectional design). The two designs have different advantages and disadvantages (Table 3.4).

The major advantage of longitudinal studies (the more commonly used design) is that each person can serve as his or her own control. For example, if the outcome of interest is weight loss at 6 months, weight at baseline can serve as the starting point for each subject. This minimizes the importance of differences in baseline weight between subjects in the different groups.

Another advantage of longitudinal studies is that all or most of the subjects in the intervention group will have received the intervention, making it easier to demonstrate an effect of the intervention. This is not true when you perform repeated cross-sectional samples of a community. Depending on how extensively the intervention has been disseminated in the population, you may find a relatively small proportion of persons who have actually been exposed to it.

However, longitudinal cohort studies also have problems when used to evaluate interventions. If you ask subjects the same questions repeatedly, your questioning is likely to affect their answers. In particular, in longitudinal studies participants generally interact repeatedly with study staff and often develop strong loyalty to them – a good thing when you need people to attend repeated visits. However, the bond between subjects and study personnel can be detrimental if participants respond to questions the way they believe the investigators want them to (rather than telling the truth). (If you have any doubts about this, how honest were you the last time your dentist asked you how often you flossed your teeth?)
Social desirability responses can be particularly an issue with intervention trials because subjects will know the goal of the study is to change their behavior and they will not want to displease the study staff by admitting to behavior they know is wrong (e.g., admitting to drug use when in a study of drug treatment).

The best way to avoid the problem of social desirability in responses is to have an objective marker of behavior (e.g., saliva test for cigarette smoking). However, this is not always possible. With serial cross-sectional studies subjects have not been asked the same question repeatedly and have not interacted extensively with study staff; therefore their answers are less likely to be influenced by their participation in the study.

Besides being less likely to elicit socially desirable responses, serial cross-sectional designs are superior when you are trying to demonstrate that an intervention has affected a whole community. For example, the Minnesota Heart Health Program was a large community intervention involving 400,000 persons in six communities. The goal of the intervention was to reduce morbidity and mortality from cardiovascular disease. Three communities received individual, group, and community-level programming designed to decrease blood pressure, improve healthy eating, decrease smoking, and increase activity. The program was evaluated by conducting cross-sectional surveys of 300–500 persons in each community seven to eight times over a

---

Evaluating an intervention
decade. A serial cross-sectional design made sense because the goal was to change the behavior of the entire community, not just persons enrolled in an intervention.

To evaluate changes in medical practice, in most cases you will have no choice but to use serial cross-sections. To see why look back at Table 1.2. The data are from a study of whether providing hospitals with confidential information on their performance improves their care of patients having coronary artery bypass grafting (CABG). The investigators looked at four measures of quality of CABG. A longitudinal study would not work because once patients have had their surgery, it’s not possible to evaluate whether surgery techniques improved. (Even if a patient had a repeat CABG it would not be the same. Patients who require a repeat surgery are different than first-comers – the outcome of the first surgery would affect the second one.) Instead, you must look at serial cross-sections of cardiac patients to see whether surgery practice improved.

Serial cross-sectional studies may be cheaper than longitudinal cohort studies because there is no need to stay in contact with subjects. Even with extensive follow-up some subjects will be lost to follow-up in longitudinal cohort studies; this cannot occur with serial cross-sectional studies.

Because they have different advantages and disadvantages, investigators of population-level interventions may perform both. In fact, the Minnesota Heart Health Program supplemented the cross-sectional samples by taking a sample of participants from the pre-intervention cross-sectional surveys and following them over time (longitudinal cohort). Results were similar: with both the cross-sectional samples and the cohort sample, the intervention did not produce substantially greater improvements in cardiac risk factors than those seen in the comparison groups.

Serial cross-sectional studies require a different statistical approach than longitudinal studies. These are covered in Sections 6.3 and 6.5, respectively.

3.8 How do I develop measurement instruments?
The type of measurement instruments you will need to develop will depend on what you are studying and how you will be collecting your data (e.g., questionnaires, interviews, observations, etc.). This is beyond the scope of this book.12

To perform your analysis you will also need to manage your data (enter it, clean it, recode, transform, and derive variables) and export it into a software package that can perform the needed analyses.13

**3.9 How do I state the hypothesis of an intervention study?**

Every intervention study should have a hypothesis – what it is you are trying to prove. To facilitate statistical testing, the hypothesis should be stated in the null form. The null hypotheses for the three major questions of intervention studies (Section 1.2) are:

1. There is no change between the pre-intervention and the post-intervention assessment.
2. The change between the pre-intervention and the post-intervention assessment is no greater in the intervention group than the change in the comparison group.
3. The outcome for the intervention group is no different from that of the comparison group.

We use statistics to determine the probability that the null hypothesis is correct. If the probability is very low, we consider the alternative hypothesis: that the intervention is associated with a significant change. The three null hypotheses can be restated as alternative hypotheses, as follows:

1. There is a change between the pre-intervention and the post-intervention assessment.
2. The change between the pre-intervention and the post-intervention assessment in the intervention group is greater/less than the change in the comparison group.
3. The outcome for the intervention group is different than that of the comparison group.

Note that the alternative hypothesis is stated in a neutral direction (e.g., alternative hypothesis #1 doesn’t say whether the change between the pre-intervention and the post-intervention assessment is up or down). This is referred to as a two-sided hypothesis.

Even though the alternative hypothesis is best stated in a neutral way, you can have an opinion; for example, you may believe that the intervention group will do better than the non-intervention group. In fact, if you didn’t have an

---

Evaluating an intervention

opinion, in many cases, you wouldn’t do the study. Still, it is generally best to state and test the hypothesis in a neutral way.

In rare cases it may be appropriate to state and test a one-sided alternative hypothesis. The rationale for stating a one-sided alternative hypothesis is that only one side of the hypothesis is possible or of interest. For example, Nadelman and colleagues studied the efficacy of doxycycline in preventing Lyme disease in an area where the disease was hyperendemic. Subjects who had removed an attached *I. scapularis* tick from their body within the prior 72 hours were randomized to either a single dose of doxycycline or placebo. It is very unlikely that a single dose of doxycycline could increase the risk of developing Lyme disease and the investigators therefore considered their hypothesis to be one-sided.

The benefit of testing a one-sided hypothesis is that it is easier to detect a statistical association: a smaller sample size is required to demonstrate a given effect size. For example, a given effect size and sample size that results in a *P* value of 0.08 with a two-sided test (hypotheses have “sides” and tests have “tails”), would have a *P* value of 0.04 with a one-sided test. Given the slavish attention paid to the arbitrary cut-off of 0.05, you can see why some investigators prefer a one-tailed test.

The problem with one-sided hypotheses is being certain that only one direction of the alternative hypothesis is possible. For example, Pletcher and colleagues developed an intervention to improve treatment of alcohol withdrawal in hospitalized patients. The intervention was to promulgate quality guidelines for the care of patients in their hospital based on the published literature and recommendations of a prominent specialty society. The problem was when they evaluated the impact of the new guidelines using a pre-intervention versus post-intervention design (since this was a typical hospital quality improvement program no control group was used), they found a significant increase of in-hospital deaths after the guidelines were implemented.

Therefore, even if one side of the hypothesis seems very unlikely, you will do better to use a two-tailed test. Two-sided hypotheses and tests are more rigorous and they are what most journal editors expect.

Going back to the doxycycline treatment example, even though the authors specified a single-sided hypothesis, they ultimately performed a two-tailed test. They found that erythema migrans (rash diagnostic of Lyme disease) developed at the site of the tick bite in 8 of the 247 subjects who received placebo.

---


Evaluating Clinical and Public Health Interventions

compared to 1 of the 235 subjects in the doxycycline group ($P < 0.04$). Because the two-tailed test was under the statistical threshold of $P = 0.05$ the decision to use a two-tailed test had no impact on whether the result was considered statistically significant. Would they have reported their results in the same way if their two-tailed $P$ value had been $P = 0.06$? I would hope so, but I know for sure that whether the results were $P$ equals 0.06 or equals 0.04, I would take doxycycline following an *I. scapularis* tick bite.

### 3.10 Can an intervention study have more than one hypothesis?

Yes. It is often better to specify more than one hypothesis (to assess whether the intervention is associated with more than one outcome). This can strengthen a study by demonstrating that the intervention has multiple positive effects. For example, an exercise intervention might decrease weight, improve muscle mass, decrease blood pressure, decrease fasting blood glucose, and improve mood.

Demonstrating that an intervention affects multiple outcomes on a causal pathway also increases the validity of your results. For example, it would be more believable that an intervention decreases diabetic retinopathy if it also decreases fasting blood glucose.

However, when you test for multiple differences between groups, the likelihood increases that at least one of these comparisons will be statistically significant. In fact, if you compare an intervention to a placebo on 20 different outcomes you would expect that one of the comparisons would be significant at the $P < 0.05$ threshold by chance alone.

To avoid this problem, distinguish the primary outcome from the secondary outcomes in the analysis plan. You will then need to adjust your significance level for the multiple comparisons. Methods for adjustment for multiple comparisons are discussed in Section 9.3.

### 3.11 Why should I have an analysis plan?

Every study should have an analysis plan. In the case of a randomized study it should be written prior to randomizing subjects (ideally), but definitely prior to unblinding group assignment and/or performing any data analysis. In the case of a nonrandomized intervention, it should be written prior to data collection (ideally), but definitely prior to data analysis.

An analysis plan details how the data will be analyzed once they are collected. It should include the major outcome of interest (e.g., weight loss) as
Evaluating an intervention

well as secondary outcomes (e.g., fasting glucose level). The study hypotheses should be stated in both the null form and the alternative form (Section 3.9).

The plan should state what statistics will be used to test the hypotheses. For example: the null hypothesis is that there will be no difference in weight loss between subjects randomized to the intervention and those randomized to usual care; this will be tested by comparing the weight change in the intervention group to the weight change in the usual care group using a two-tailed \( t \)-test (Section 6.2). If you have multiple outcomes (Section 3.10), state how you will adjust for multiple comparisons (Section 9.3). If you are planning to do any subgroup analysis (Section 9.5) this should be described.

The analysis plan should state whether there will be any interim analyses to assess whether the study should be terminated early (Section 9.4). It should state whether the data will be analyzed using intention-to-treat, per-protocol analyses, or both (Section 4.8). It should also state how you will handle missing data (e.g., excluding missing values, imputing values) (Section 9.6).

There are several reasons why it is important to write an analysis plan prior to analyzing your data. First, pre-specifying the analysis will decrease bias and prevent false positive results. To understand why, consider an extreme example. As you analyze your data, you construct a composite outcome that maximizes the apparent effectiveness of the intervention. You do this by grouping outcomes that are related to the intervention (this would lead to a type 1 error – falsely concluding that the intervention works when the null hypothesis is true).

When analyzing study data there are many decisions to be made along the way (how to code a variable, how to deal with missing data) and if each of these decisions is not decided ahead, the natural tendency will be to choose the method that produces the effect you are trying to demonstrate, thus biasing your results. By pre-specifying how you will analyze your data you are protecting the integrity of your analysis.

Besides these reasons, it is not possible to conduct a sample size calculation (Section 3.12) without knowing how you will analyze your data. You will also find that considering ahead of time how you will conduct your analysis will help ensure that you collect all the data you will need to answer your question (e.g., covariates).

Finally, having a preset analysis plan should prevent the selective reporting of results (e.g., reporting only those results that supported the study hypothesis or changing outcomes midstream). Unfortunately, even an analysis plan does not always prevent investigators from straying: a study comparing published randomized reports with their study protocols found that reporting was often
Evaluating Clinical and Public Health Interventions

biased and inconsistent with the protocol; specifically 62% of trials had at least one primary outcome that was changed, introduced, or omitted.¹⁶

Having an analysis plan should not prevent pursuing analyses that arise during the analysis. Such analyses are called post-hoc analyses and can be very valuable. However, the $P$ values associated with these results substantially underestimate the likelihood that the findings are due to chance. For this reason, results from post-hoc analyses should be viewed more tentatively than those from pre-specified analyses. This is especially true if the analysis is suggested by your data analysis (as opposed to a new question that arose after the pre-analysis plan was formulated).

### 3.12 How do I calculate sample size for an intervention study?

Calculation of sample size for intervention studies is no different mathematically than the calculation of sample size for other types of studies. The number of subjects needed depends on the estimated size of the effect of the intervention, the number of groups studied, the type of outcome measure (e.g., interval, ordinal), the variability of the measurements (e.g., standard error), the number of covariates in the model (assuming multivariable analysis will be performed), the duration of follow-up (for studies of time), and the number of subjects lost to follow-up.

Because it is not mathematically different, I will not review the methods, but instead direct you to the references, which include free software for calculating sample size.¹⁷

What is critically important with intervention studies is to ensure that the study is sufficiently powered to rule out a meaningful effect if there is none.¹⁸

---


Evaluating an intervention

I will take up this subject again later in the section on publishing negative studies (Section 9.8).

3.13 How do I obtain an institutional review board (IRB) review for a study?

All studies involving human subjects require a review by an institutional review board (IRB). These boards, also known as human subjects committees, review the goals and procedures of your study to be certain that the rights of human subjects are protected.

A key determination is whether the investigators must obtain informed consent of the subjects and how the informed consent will be obtained (e.g., signed, verbal assent). The usual reason for not requiring informed consent is that the risk to the individual is very minimal (e.g., medical record review) or consent is impossible (e.g., a study of the best method to resuscitate patients found down in the field).

Many funders require IRB approval prior to even funding a study. Most journals will not publish studies that have not been appropriately reviewed by an IRB.

As implied by the name, IRBs are usually associated with an institution such as a university or governmental entity. The members are drawn from a variety of disciplines and should ideally include clinicians, ethicists, researchers, lawyers, and research subjects or their advocates.

Although all studies involving human subjects should be reviewed by an IRB, an IRB may make the decision to grant the study an exemption or waiver, meaning it does not require the approval process of the committee. Waivers are granted to studies that have no or little potential for harm of human subjects, such as a hospital quality improvement program, an analysis of an administrative database with no identifiers, a descriptive review of a novel clinical program where all the data were collected as part of the clinical work, or a research study conducted as part of governmental public health efforts (e.g., a disease-outbreak investigation).

To illustrate the complexity of issues concerning IRBs, informed consent, and waivers, let me tell you what happened with the study on reducing catheter-related infections by Pronovost and colleagues (Section 1.2.A). It caused a huge stir in the IRB world, played out on the pages of the *New York Times*.

Here are the basic facts. Provonost and colleagues sought to implement a quality improvement project to reduce catheter-related infections in 67

---

Michigan hospitals. The study protocol was submitted to the IRB at Johns Hopkins University. The IRB judged that the intervention was exempt from review because it was a quality improvement program. Therefore, informed consent of patients was waived.

The Federal Office for Human Research Protections (OHRP) received a written complaint that the research was conducted without the review of an IRB and without the informed consent of the subjects. As a result of the investigation the study was voluntarily halted. The OHRP ruled that the Johns Hopkins IRB should have reviewed the study (not made it exempt) and that informed consent should have been obtained from the subjects (from both the patients and the physicians who participated in sessions on how to reduce infections).

Both parts of this decision have been hotly debated. Hospital leaders, who have been performing quality improvement efforts for decades without IRB approval or informed consent of patients, were concerned that requiring IRB review and informed consent would be onerous and discourage attempts to improve quality. To calm these protests, OHRP offered a new pathway that struck many of us in the research community as illogical. The OHRP ruled that hospitals could conduct quality improvement programs without IRB review so long as the sole aim was to improve patient care – in other words, as long as the hospital did not assess whether the practice worked! Many of us feel that it is unethical to engage in quality improvement programs without assessing whether they work. (If you have any doubt of this look back at the example of the quality improvement program for the treatment of alcohol withdrawal in Section 3.9; the quality improvement effort was associated with increased mortality.) Besides making sure that the intervention really works, isn’t it unethical not to tell others what they can do to improve patient care?

Nonetheless, if you receive a research grant to do a study, and/or think there is a chance that you will publish the results of the study (two criteria that OHRP considers when deciding whether you are doing more than implementing an intervention to improve quality), submit the protocol to the IRB even if you think it should be waived. This applies to any study of human subjects (e.g., review of administrative data, descriptive review of a clinical program). They can perform an expedited review of it.

Obtaining informed consent is a much more challenging issue. It greatly complicates a study, results in some people not participating, and in the case of a quality improvement program has the complication of it being unclear what to do if the person does not want to participate. Given that the protocol of Provonost and colleagues included only interventions that were considered
standard of care (e.g., hand-washing), it is unclear what the alternative to participation would be. Would you tell the patient that if he or she did not wish to be in the study then the nurse would not remind the doctor to wash his or her hands? I would hope not. I suppose you could tell them that you wouldn't include them in the recording of results.

Fortunately OHRP appears to have softened their stance; they are allowing informed consent to be waived in certain quality improvement programs. (Indeed, when the Johns Hopkins IRB reviewed the study in response to the OHRP investigation, they granted the investigators a waiver for patient consent.) By submitting your protocol for IRB review, you will learn from the experts whether or not you can waive informed consent.

The events surrounding the study of decreasing ICU infections also illustrate another important point: there can be disagreements in the interpretation of the rules governing human subject protection. If you ever conduct a multi-center trial where you are required to obtain approval of the IRB associated with each institution you will see this point played out in subtle ways. Each IRB will try to edit your consents in different ways to fit their sense of how to best obtain informed consent.

3.14 When do I need an independent data and safety monitoring committee?

If you are conducting a prospective randomized study where participants and/or their doctors do not know the treatment they are receiving (double-blinded), then you need to set up an independent data and safety monitoring committee. The committee reviews events that occur during the study that could require stopping the study or withdrawing individual patients. Unlike the investigators, the independent data and safety monitoring committee may unblind the assignment if it is for the benefit of the subjects without introducing bias into the study. If your study has pre-specified interim analyses (Section 3.11) with early stopping rules (Section 9.4), the data and safety monitoring committee should perform the analysis.

3.15 Why should I register my trial?

Trial registration is an important method of ensuring that all of the evidence on a topic is available to clinicians, consumers, and researchers. Prior to the establishment of registries, unpublished trials were hard to find. This is a problem because negative studies (those where the intervention is not found to be
Evaluating Clinical and Public Health Interventions

effective) are less likely to be published.$^{20}$ Without the ability to identify unpublished studies, reviews of prior research will probably overestimate the effectiveness of the interventions. Making it easy to find all the available studies is of even greater significance with the growing reliance on systemic reviews (meta-analyses) for making clinical and policy decisions.

Which intervention studies should be registered? According to the International Committee of Medical Journal Editors: “Any research project that prospectively assigns human subjects to intervention and comparison groups to study the cause-and-effect relationship between a medical intervention and a health outcome.”$^{21}$ Both randomized and nonrandomized prospective studies should be registered. Retrospective studies, pharmokinetic studies, and phase 1 trials of toxicity do not require registration.

If I haven’t convinced you of the value of registering your trial, consider this: All major journals require it! There are several available trial registries:

- http://www.anzctr.org.au
- http://www.clinicaltrials.gov
- http://isrctn.org
- http://www.umin.ac.jp/ctr

All these trial registries include the minimum necessary data set including information about the methods and results.

---


Randomized designs

4.1 What are randomized studies?

In a randomized study, subjects are assigned to each group by chance, such as by the flip of a coin. This is known as random assignment and should not be confused with random selection, which is the selection of cases from a population at random. To illustrate the difference: if you have a nonrandom sample, such as a convenience sample of patients in a clinic, you can still randomize them to different groups.

Although an unbiased coin would work perfectly well for randomizing subjects to two groups, randomization is generally performed using a computerized random number generator: you specify the number of groups you want to create and what proportion you want in each group (for example, in studies of equal allocation you would specify that you wanted half the subjects in each group), and the computer will spit out numbers that can be used to assign subjects randomly to the groups.¹

4.2 What are the advantages of randomization?

The major advantage of randomization is that it produces unbiased groups; that is, because group assignment is determined randomly it cannot be influenced by factors that may otherwise confound the results of the study.

Confounders are factors that are associated with group assignment and are causally related to outcome. Because there is no relationship between potential confounders and group assignment in a randomized controlled trial, we can be confident that any differences that develop between the groups are not due to confounders (Figure 4.1).

In comparison, when assignments are not randomized, a number of factors (e.g., patient preferences, family or social group preferences, physician

¹ A free, easy-to-use random number generator is available at www.randomizer.org. Random numbers produced by a computer algorithm are technically not random, but pseudorandom (an insight I owe to Josephine Wolf); nonetheless, they work well for study purposes.
preferences, investigator preferences) may influence whether or not a subject takes a particular medication, follows a diet, obeys a law, etc. These factors may also be causally related to outcome (e.g., persons compliant with medications may live longer because they engage in a variety of health-promoting behaviors such as visiting the doctor regularly, following medical advice) and thereby confound the results of your study.

Because assignment is unbiased when randomized, randomization tends to produce groups that are similar on baseline characteristics (i.e., each group has a similar proportion of men and women, younger and older people, healthier and sicker persons). Comparability of groups at baseline is essential to ascribing differences between the intervention group and the comparison group to the intervention, rather than to inherent characteristics of the group (e.g., healthier people are more likely to survive surgery than sicker people).

Comparability of groups is so important that the results section of almost every published intervention study begins with a “Table 1” comparing the baseline characteristics of the study groups. In the case of randomized studies, the groups are expected to be similar.

Unfortunately, randomization does not guarantee that the groups will be similar on every characteristic in every randomized study. Sometimes by chance (bad luck!) randomization will produce groups that have important differences on baseline characteristics. For example, DiClemente and Wingood randomized 128 women to one of three groups: a five-session social skills HIV-prevention program; a one-session education group; or a delayed education control. Despite the randomization, at baseline women randomized to the social skills intervention were in longer-term relationships (45 months) than women in the HIV education group (21 months) and the delay control group (33 months) \( (P = 0.06) \). Also participants in the delayed group were more likely (83%) to be very poor (income less than $400 per month) than those in the intervention group (72%) and those in the education control (60%) \( (P = 0.10) \).

---

Randomized designs

Note that neither of these baseline differences reached the conventional significance level of $P < 0.05$. However, baseline difference between study groups where the variable is associated with the outcome can be a confounder even if the difference between the groups does not reach statistical significance. Also, you can have several variables that are each a little different between the groups, with all variables favoring one group (subjects in one group are somewhat younger, somewhat less likely to be smokers, and somewhat less likely to have an underlying disease). While none of these differences may even approach statistical significance together they may result in one group being more likely to have a particular outcome regardless of the intervention.

Did the differences in the baseline characteristics of the women in the HIV-prevention study of DiClemente and Wingood invalidate the study? No. However, the differences between the groups required the investigators to adjust their analysis for income and length of current relationship using multivariable analysis.

Because even small differences in baseline characteristics can potentially bias the results of studies, investigators performing randomized studies are increasingly using multivariable analysis just as you would do to adjust for baseline differences with nonrandomized studies.

Whether or not you adjust for baseline differences, the true power of randomization is that it produces groups that are similar on unmeasured characteristics. This is critical because it is impossible to adjust for unmeasured or unknown characteristics. Therefore, in a nonrandomized design, even if the groups are balanced on all measured characteristics known to affect outcome, there remains a question as to whether differences that occur during the course of the study are due to the intervention or to some unknown characteristic.

A final advantage of randomization is that it ensures that the significance levels of the statistical techniques used to analyze data are valid.\(^3\)

4.3 What are the disadvantages of randomization?

Although randomization is the best way of determining whether an intervention works, it has several disadvantages.

Foremost, in many cases randomization is impossible. For example, it is important to understand whether changes in laws (e.g., requirements that children be placed in car safety seats, taxes on tobacco) are having a beneficial effect. Since it is not possible to randomize people to be subject to the law other designs are needed.

To randomize subjects, there must be a state of equipoise, meaning that the arms of the study are believed to be equal. This does not mean that everyone believes that they are equal; in fact, a perfect situation for randomization is one where half the doctors feel treatment A is clearly superior and the other half believe that treatment B is clearly superior.

However, if one treatment is thought to be superior then it may be considered unethical to randomize people. For example, it was thought to be unethical to randomize critically ill patients to receive or not receive a pulmonary artery catheter because the catheter was thought essential to managing medications and fluids for these patients. However, when a well-done nonrandomized study showed that patients were more likely to die if they had a catheter placed, enough doubt was introduced that a randomized controlled trial became possible; it showed that patients randomized to receive a pulmonary artery catheter had greater mortality, and this has resulted in a substantial decrease in the use of pulmonary artery catheters.

Even when randomization is feasible and ethical, these designs have limited generalizability. Randomized controlled trials are conducted under strict conditions: subjects must consent themselves to be randomized; they must make frequent study visits; and they must agree to intrusive questions and often laboratory tests. Therefore, subjects who are willing to enter and remain in randomized controlled trials are different from other subjects. Also, the medical care delivered within these trials is different from the medical care delivered in normal clinical practice.

Not only are the subjects in randomized controlled trials different, and treated differently, but their behavior may actually change simply due to being observed. This is referred to as the Hawthorne effect. The Hawthorne effect is not unique to randomized controlled trials. However, the more intense observation in these trials compared to most observational trials makes the Hawthorne effect potentially more potent.

### 4.4 What are the different methods of allocating subjects?

There are four major ways to randomly allocate subjects:

1. Randomization with equal allocation.
2. Blocked randomization.

---


Randomized designs

3 Stratified randomization.
4 Randomization with unequal allocation.

Most studies are performed with equal allocation (simple randomization), i.e., an equal number of subjects are randomized to each group. Equal allocation is simple and provides maximum power. Although randomization with equal allocation will result in approximately equal numbers of subjects in each group, you could get by chance more subjects in one group than another. This can be an issue with small studies (< 20 subjects per group) and can be avoided with blocked randomization.

With blocked randomization, subjects are randomized in small blocks (e.g., four or six subjects), thereby guaranteeing an equal number of subjects in each group. With two groups (for example, A and B) randomization in blocks of four could result in the following patterns: AABB, ABAB, BABA, ABBA, BAAB, BBAA. In all of these patterns, if you randomize four subjects, you would have two subjects in group A and two subjects in group B. In contrast, with randomization of four subjects without blocking you could have a pattern of AAAA or BBBB.

As with unblocked randomization, blocked randomization does not guarantee that important prognostic factors will be balanced. Blocked randomization only makes sense in small studies when an imbalance in the number of subjects per group would decrease power. For example, if you were randomizing 24 patients, you would not want to risk having 10 in one group and 14 in another.

Both equal allocation and blocked randomization may result in imbalances in the prognostic characteristics across study groups (Section 4.2). This problem can be avoided with stratified randomization.

As implied by the name, stratified randomization means randomizing subjects within certain strata so as to be certain to balance groups across important prognostic factors. For example, if it were important to have an equal number of men and women in your study because gender has a strong influence on the outcome or the effectiveness of the intervention, you could stratify your randomization by gender.

Although stratification may sound sensible, it requires knowledge of the important baseline characteristics prior to the study, is unworkable for more than a few baseline characteristics, and requires you to incorporate the stratification in your analysis, thus complicating data analysis. In practice stratified randomization is used only in (1) small studies with strong prognostic factors, (2) trials where there will be interim analyses (Section 9.4) with small numbers of patients, or (3) equivalence and non-inferiority studies (Section 9.2). 7

---

Randomization with unequal allocation is essentially a simple randomization except that instead of randomizing patients one-to-one to the different groups, you randomize them in an unequal fashion such as two-to-one favoring one group. Generally, unequal randomization is chosen so that a higher proportion of subjects receive the treatment (in a treatment versus placebo study) or a new treatment (in a new treatment versus conventional treatment study).

The advantage of unequal randomization is that it might provide greater incentive for some subjects to enter a treatment trial if they feel that they have a better than equal chance of receiving the treatment (say 66%). Also, by having more subjects in the treatment arm you will learn more about side effects associated with a novel treatment. The problem with unequal randomization is that it decreases the power of your study. Also, unequal randomization suggests that the investigator believes that one arm is superior to the other arms. This is inconsistent with the principle of equipoise.

Overall, unless you have a strong reason to do otherwise, perform a simple randomization with equal allocation.

### 4.5 What is clustered (group) randomization?

It is possible to conduct a randomized study by randomizing groups of participants, rather than individuals. This type of randomization is referred to as group or cluster randomization.

Commonly randomized groups include schools, medical practices, clinics, hospitals, businesses, and communities. (Table 4.1).

Cluster randomization has several advantages over randomization of individuals. Often, it is easier to randomize groups than individuals. For example, to test an intervention to promote healthful eating and exercise among school-age children, consider how much easier it would be to randomize 20 schools to either an educational intervention or to no intervention than to individually randomize the students within the school. Remember that to randomize the students you would need to obtain their informed consent prior to their randomization. If you randomize the schools, you can obtain informed consent from the students and their families after randomization. Knowing which arm of the study they are in may increase participation.

Continuing with the same example, consider how much easier it would be to implement an intervention in 10 schools than to create an intervention in

---

Randomized designs

Table 4.1. Examples of clustered randomization.

<table>
<thead>
<tr>
<th>Types of clusters that can be randomized</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools</td>
<td>Eighteen schools were randomized to one of three groups: classroom-based intervention to improve cardiac health for all 3rd and 4th graders, a risk-based intervention only for students with cardiovascular risk factors, or no intervention.</td>
</tr>
<tr>
<td>Medical practices or clinics or hospitals or nursing homes</td>
<td>Thirty-one hospitals were randomized either to an intervention designed to increase breastfeeding or to usual practice.</td>
</tr>
<tr>
<td>Businesses</td>
<td>Thirty-six pharmacies were randomized to one of three study groups: pharmaceutical care, peak flow monitoring only, or usual care for patients with asthma.</td>
</tr>
<tr>
<td>Communities</td>
<td>Twelve rural communities were randomized either to a community intervention to reduce inappropriate antibiotic use plus clinical support systems for clinicians in that area or to just the community intervention.</td>
</tr>
</tbody>
</table>

---


---

all 20 schools and have only the randomized half of the students be exposed to the intervention.

Besides being easier to perform, clustered randomization makes it less likely that the control group will be exposed to (contaminated by) the intervention. Using the school example, if you randomized students within schools imagine how hard it would be to keep the children randomized to a control group from hearing the messages of the intervention from their fellow schoolmates.

Perhaps the best reason to do a clustered randomization is because the intervention is operating on a group (not individual) level. For example, to perform an intervention to improve hospital or clinic performance, it wouldn’t make sense to randomize individual patients or doctors because the intervention is aimed at improving care within the entire hospital or clinic.
Evaluating Clinical and Public Health Interventions

For example, Shafer and colleagues sought to increase the screening rate for chlamydia among sexually active adolescent girls. Five clinics were randomized to perform the intervention and five were randomized to provide usual care. In the intervention clinics, administrators, medical assistants, nurses, and practitioners were engaged in redesigning clinic flow to increase screening rates; the group decided that the best way to increase screening would be to ask all adolescents to leave a urine specimen at registration, and let the practitioners decide later which ones to send. The investigators found that the screening rate for sexually active adolescent girls was higher in the intervention clinics than in the usual care clinics.

Remember that whether you are randomizing groups or individuals, to achieve an equal balance of baseline characteristics there must be a sufficient number of units to randomize. If, for example, you randomize only two cities, your sample is too small for randomization to balance baseline characteristics. To illustrate: consider randomizing one of two cities to an intervention or a control group. The two cities that have agreed to participate and accept randomization are: New York City and New Delhi. No matter which city is randomized to the intervention, the two cities and the participants within them will be different. Conversely, if you randomize two similar cities, the cities and the participants in them will be similar on measured characteristics but not because of the randomization.

Data from a clustered randomized study, or any clustered study (nonclustered studies are discussed in Section 5.4.B), must be analyzed differently than data from a nonclustered study. Specifically, the correlation between clustered individuals must be incorporated into the analysis.  

4.6 What can be randomized besides individuals or clusters of individuals?

Because of the unique ability of randomization to produce groups that are comparable with respect to both measured and unmeasured characteristics, investigators who cannot randomize individuals or clusters of individuals will sometimes identify some other feature to randomize. For example, O’Donnell and colleagues were interested in determining the impact of video-based education on patients seen at an STD clinic. They compared three treatment

---


10 Two commonly used techniques for analyzing clustered data are generalized estimating equations and mixed-effects models. For more on these techniques see: Chapter 6, footnote 22.

Randomized designs

conditions: video education, video education plus group discussion, and control. Because it was logistically difficult to randomize patients within a clinic setting, treatment conditions were assigned on random days of clinic operation. In other words, on any given day subjects were enrolled into only one condition and that condition was chosen randomly. Although the investigators could have prevented bias by randomizing STD clinics (clustered randomization), that would have required a much more complicated study with multiple sites and investigators. Randomizing days was an economical and practical solution.

One criticism of this type of design is that study staff would know prior to the enrollment of any individual what treatment assignment they would get. This is not true when you randomize individuals.

Another feature that may be randomized is duplicative body parts. For example, Kiel and colleagues randomly assigned nursing home residents to wear a one-sided hip protector on the left hip or the right hip. The randomization of hip was done at the level of the nursing home (clustered randomization) to make it easier for the staff to remember which side of the body to place the pad. Unfortunately, the incidence rate of hip fracture was similar for the protected and unprotected hips.

4.7 When should subjects and researchers be masked (blinded) to treatment assignment?

Whenever possible!

As we have discussed, randomization prevents bias in the treatment assignment (Section 4.2). However, if the subjects or the researchers know the assignment (open label) bias enters! Why? Because subjects and researchers will have expectations – whether true or not – about the different treatments being studied.

For example, in an active drug trial, the person who knows they have received the active drug may feel better immediately – even before taking the first dose – just knowing that they will be getting the active drug. The person randomized to the higher-dose arm of a low-dose/high-dose trial may experience side effects because they know that they are receiving a high dose of the medication.

Investigators may be equally influenced. Those who believe that a particular intervention is efficacious may unconsciously perceive those who are

Evaluating Clinical and Public Health Interventions

randomized to that intervention to be doing better. This can cause measure-
ment error, especially if the measurements can be affected by how a question is
worded, or how an answer is interpreted, or the perception of the observers.

These problems can be avoided by masking subjects and researchers to treat-
ment assignment. This is called a double-masked (blinded) study.

Given the advantages of masking, why not mask subjects in all trials? The
answer is that in many trials it is impossible. For example, it is not possible
to mask subjects to whether they receive a behavioral counseling session or a
housing voucher or a telephone follow-up. Some interventions, such as surgery,
can theoretically be blinded, but raise difficult ethical issues. In other cases,
subjects may only be willing to participate if the study is open label.

If you can’t mask treatment assignment, consider what strategies you can
employ to decrease bias due to subjects or investigators knowing the treatment
assignment. For subjects, it is important to emphasize to them that it is not
known which treatment is better and that’s why it is being studied.

To decrease bias from knowledge of treatment by researchers use objective
markers, whenever possible. For example, use an automatic blood pressure
monitor to measure blood pressure. Any materials that can be assessed without
the subjects’ presence (e.g., X-ray) should be read by practitioners who do not
know the treatment assignment or even the hypothesis of the study.

4.8 How do I account for subjects who do not adhere to their treatment?

In a randomized clinical trial there will invariably be subjects who are assigned
to a treatment but never take a dose, or who stop taking the treatment because
of side effects or because they don’t like it, or who are assigned to the placebo
group but on their own initiative cross over and take the active treatment. (If
you are thinking this can’t happen if subjects are blinded to treatment assign-
ment: think again. Subjects are known to send their study medications for drug
testing and then start the active treatment if they find out they were rand-
omized to placebo.)

The problems of stopping or switching treatment are not unique to rand-
omized trials. In fact, it is even more common in nonrandomized observa-
tional trials because in everyday life patients often stop taking their medicines
or switch to a new one. We expect this in observational trials. But the conse-
quences are greater in randomized controlled trials because those who stop
taking their randomized treatment, or cross over into one of the study arms,
dermine the advantages of randomization.

Randomized designs

Remember that the reason we randomize subjects is to eliminate bias; this enables us to ascribe differences that occur between the groups during the study to the intervention. However, the factors that lead to not complying with treatment, or crossing over to the other treatment arm, are anything but random. Therefore, if you exclude persons who do not take their assigned treatment or reclassify such persons based on their actual behavior you are introducing bias into the randomized groups.

Rather than excluding or reclassifying subjects who do not adhere to their treatment, analyze your data as if the subject received their assigned treatment (or placebo) to the end of the study. This is called an intention-to-treat analysis.\textsuperscript{14}

Intention-to-treat analysis is necessary to maintain the advantages of randomization. In addition intention-to-treat analysis protects against treatments appearing more favorable than they are. Why? Because dropouts and crossovers occur for a reason – typically because subjects are not happy with how things are going. Therefore we would expect that people who drop out would not do as well as subjects who stay in the study.

Besides maintaining the randomization, the results of an intention-to-treat analysis are more informative to clinicians starting patients on a regimen because at the outset we do not know whether a patient will comply or not.

That being said, it is sometimes important to see the results from those subjects who actually adhered to the study. This is called a per-protocol analysis. Per-protocol analyses typically exclude persons who were never exposed to the intervention or were insufficiently exposed (e.g., took less than 85% of the drug in a medication trial). This has the advantage that if people were not sufficiently exposed to the intervention to benefit, including them in the intervention arm dilutes the overall impact of the intervention.

One type of per-protocol analysis is an as-treated analysis. An as-treated analysis groups subjects based on the treatment they received, rather than the group to which they were randomized. As you can imagine, if you have a large number of subjects who cross over into the other arm, an intention-to-treat analysis will not well characterize the benefits and the risks of the different arms (because the groups as randomized are a mixture of the arms).

This potential role of an as-treated analysis is well illustrated by a trial of surgical versus non-operative treatment of lumbar disk herniation.\textsuperscript{15} In this clinical trial, subjects were randomized to surgery or to non-operative care


Evaluating Clinical and Public Health Interventions

At two years only 140 (57%) of the 245 subjects randomized to surgery actually had surgery; meanwhile 107 (42%) of the 256 randomized to non-operative care had crossed over and had surgery!

As you can see in Table 4.2, the crossovers were not random. Among those assigned to surgery, those who actually had the surgery were significantly younger, of lower income, more likely to have pain with straight leg raise, more likely to have disk herniation at a lower level, in greater pain, in worse physical function, in greater disability, more dissatisfied with their symptoms, and more likely to assess themselves as getting worse. There were also many statistically significant differences within the non-operative treatment group between those who did and did not have surgery.

Although having so many subjects crossing over to the other treatment arm may seem extreme, this is the real world of randomized trials comparing surgical treatments to medical treatments (i.e., it is natural that people might decide – even

Table 4.2. Comparisons of subjects who received their assigned treatment to those who did not, for both those assigned to surgery and those assigned to non-operative treatment.

<table>
<thead>
<tr>
<th></th>
<th>Assigned to surgery No. (%)</th>
<th>Assigned to non-operative treatment No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surgery (n = 140)</td>
<td>No surgery (n = 92)</td>
</tr>
<tr>
<td>Age, mean, y</td>
<td>40</td>
<td>44</td>
</tr>
<tr>
<td>Annual income &lt;$50 000, %</td>
<td>47</td>
<td>29</td>
</tr>
<tr>
<td>Pain with straight leg raise (ipsilateral), %</td>
<td>67</td>
<td>53</td>
</tr>
<tr>
<td>Herniation level,%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2–3/L3–4</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>L4–5</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td>L5–S1</td>
<td>61</td>
<td>54</td>
</tr>
<tr>
<td>Bodily pain, mean*</td>
<td>24</td>
<td>32</td>
</tr>
<tr>
<td>Physical function, mean*</td>
<td>36</td>
<td>46</td>
</tr>
<tr>
<td>Disability index, mean**</td>
<td>52</td>
<td>41</td>
</tr>
<tr>
<td>Sciatica indices, mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Bothersomeness</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Satisfaction with symptoms: very dissatisfied, %</td>
<td>89</td>
<td>64</td>
</tr>
<tr>
<td>Self-assessed health trend: getting worse, %</td>
<td>41</td>
<td>26</td>
</tr>
</tbody>
</table>

* Higher score is better.
** Lower score is better.

Randomized designs

Table 4.3. Comparison of outcomes in patients randomized to surgery to those randomized to non-operative care, intention-to-treat analysis.

<table>
<thead>
<tr>
<th></th>
<th>Surgery ((n = 186))</th>
<th>Non-operative ((n = 187))</th>
<th>Treatment effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily pain, mean*</td>
<td>40.3</td>
<td>37.1</td>
<td>3.2 (–2.0 to 8.4)</td>
</tr>
<tr>
<td>Physical function, mean*</td>
<td>35.9</td>
<td>35.9</td>
<td>0 (–5.4 to 5.5)</td>
</tr>
<tr>
<td>Disability index, mean**</td>
<td>–31.4</td>
<td>–28.7</td>
<td>–2.7 (–7.4 to 1.9)</td>
</tr>
</tbody>
</table>

* Higher score is better.
** Lower score is better.


Table 4.4. Comparison of intent-to-treat and as-treated analysis.

<table>
<thead>
<tr>
<th></th>
<th>Intention-to-treat (surgery vs. non-operative care) treatment effect (95% CI)</th>
<th>As-treated (surgery vs. non-operative care) treatment effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily pain, mean*</td>
<td>3.2 (–2.0 to 8.4)</td>
<td>15.0 (10.9 to 19.2)</td>
</tr>
<tr>
<td>Physical function, mean*</td>
<td>0 (–5.4 to 5.5)</td>
<td>17.5 (13.6 to 21.5)</td>
</tr>
<tr>
<td>Disability index, mean**</td>
<td>–2.7 (–7.4 to 1.9)</td>
<td>–15.0 (–18.3 to –11.7)</td>
</tr>
</tbody>
</table>

* Higher score is better.
** Lower score is better.


if they were unsure what would be best at the start of a trial – that they really did or did not want surgery; or their condition may have changed after randomization in a way that made them strongly favor one treatment over the other).

Nonetheless, the authors were true to the principle that the primary analysis should be intention-to-treat. As you can see in Table 4.3, using an intention-to-treat analysis, those subjects initially assigned to surgery had less pain and less disability. Unfortunately, although two of the three primary outcomes favored the surgery group (the third physical function showed no change) the confidence intervals all included zero, consistent with no difference between the groups.

You may be wondering, since this is an intention-to-treat analysis, why the sample size for the surgery group is 186 instead of 245 and the sample size for the non-operative care is 187 instead of 256. Good question! Remember that intention-to-treat analysis cannot incorporate subjects who were not evaluated for the outcome in question because they withdrew, were lost to follow-up or died (for this you need to censor observations; see Section 6.4.G).

Although it was not the primary analysis, the authors reported the as-treated analysis results. As you can see in Table 4.4 the as-treated analysis shows a
much stronger benefit for surgery than the intention-to-treat. In fact, in the as-treated analysis the differences between the surgery and the non-operative group are statistically significant. This is consistent with the possibility that surgery is more effective than non-operative care but the effect is diluted by all the people in the surgical group who did not have surgery and all the people in the non-operative group who had surgery. Nonetheless, the study cannot be considered to conclusively favor surgery.
Nonrandomized studies

5.1 What are nonrandomized studies?

Nonrandomized intervention studies refer to a diverse group of studies that evaluate interventions without randomizing subjects to receive or not receive the intervention. They are often referred to as quasi-experimental studies (quasi- meaning resembling). I prefer the term nonrandomized intervention study because it allows inclusion of studies that do not resemble experiments.

Unlike randomized trials, which must be planned prospectively, nonrandomized studies may be planned, or they may be designed after the intervention has occurred. Nonrandomized studies may or may not have a control group.

5.2 Why perform nonrandomized studies?

Several common scenarios in which nonrandomized designs should be performed along with examples are shown in Table 5.1.

What the first four scenarios in Table 5.1 have in common is that they are circumstances in which randomization is not ethical or practical.

For example, Raschke and colleagues developed a computerized system to alert physicians when a patient of theirs was prescribed a medication that could result in an adverse drug event (e.g., a patient receiving digoxin with a low serum potassium level).\(^1\) Theoretically, the investigators could have randomized the doctors to receive or not receive the alerts, or have randomized patients to have their doctors receive alerts or not, or have randomized events (e.g., each time the computer recognized a potential drug event it could be programmed to send or not send a warning on a random basis). However, the investigators felt it would be unethical to not send a warning to physicians each time the computer identified a potential adverse event.

Instead, the investigators evaluated the effect of their system by assessing whether clinicians took action after receiving an alert. Although the study did

---

Table 5.1. Common scenarios for performing nonrandomized studies.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison of patients before and after an intervention in settings where it is unethical, difficult, or unnecessary to obtain a control group.</td>
<td>Impact of a computer alert system for detecting adverse drug events.(^a)</td>
</tr>
<tr>
<td>Comparison of patients receiving an intervention to those who receive a different intervention or no intervention, when it is unethical or unfeasible to randomize patients.</td>
<td>Effect of a community mobilization to decrease cardiovascular risk factors.(^b)</td>
</tr>
<tr>
<td>Changes occurring after passage of a law or creation of a new program.</td>
<td>Association of an indoor smoking ban with incidence of acute coronary syndrome.(^c)</td>
</tr>
<tr>
<td>Detection of a side effect that is rare or occurs long after the intervention.</td>
<td>Relationship of use of antipsychotic medicines with sudden death in a study linking prescription records to death certificates.(^d)</td>
</tr>
<tr>
<td>Generalization of a finding that has been established in a randomized trial.</td>
<td>Warfarin is as effective for non-valvular atrial fibrillation when used in clinical settings as in randomized clinic settings.(^e)</td>
</tr>
</tbody>
</table>


not have a control group this was not a major limitation because it is reasonable to assume that if physicians made changes based on the alerts, the alerts were helpful.

In other situations, it is critical to have a control group even if the control group is not randomized because of the possibility that changes seen in the intervention group are due to factors other than the intervention. For example, Carleton and colleagues conducted a community education study to decrease cardiovascular disease.\(^2\) The intervention involved mobilizing the community

Nonrandomized studies

of Pawtucket, Rhode Island, including individuals, community organizations, public and private schools, large work sites, supermarkets, restaurants, and city government, in a variety of educational programs.

Community mobilizations can only be randomized if you have the resources to perform the mobilization in a large enough number of communities (and observe an equally large number of communities where the intervention is not being performed) such that your sample size of communities (not individuals) will be large enough to produce comparable groups. This is rarely feasible and most community mobilizations do not randomize communities but instead choose comparison communities. In the case of the Pawtucket study they used a single comparison city. They found that following the intervention there were decreases in cholesterol, blood pressure, and smoking prevalence in Pawtucket residents, but similar decreases occurred in the comparison community. In this case, having a control group was critical in demonstrating that the changes were not due to the intervention.

A common situation for performing a nonrandomized study is evaluation of a new law or government program. In these cases, randomization is rarely an option. Instead, investigators often look for an area of the country that has not passed a similar law or developed a similar program. For example, in 2006 all indoor smoking in public places was banned in Scotland. Pell and colleagues evaluated the impact of the legislation by assessing trends in hospitalizations for acute coronary syndrome. In the period following passage of the legislation, hospital admissions for acute coronary syndrome decreased by 17% compared to the period before the legislation. To assess the likelihood that this drop was due to factors other than the smoking ban, the investigators performed a similar review of hospitalizations in England. No similar ban was in place in England at that time and the decrease in acute coronary syndrome was only 4%.

Nonrandomized studies may be necessary to detect side effects that are rare or occur long after treatment (long latency). In such cases randomized studies may not be feasible because of the number of people who would have to be enrolled or because of the length of time that the trial would have to run. For example, sudden death in association with the use of antipsychotics has been reported. However, sudden death occurs in the absence of antipsychotic drug use, raising the question of whether the rate of sudden death is higher among persons taking antipsychotics than among those not on these medications. If one were to try to answer this question with a randomized controlled trial,

---

Evaluating Clinical and Public Health Interventions

Table 5.2. Association between use of antipsychotics and sudden death.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Moderate dosage</th>
<th>Low dose</th>
<th>Past year only</th>
<th>Non-user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person-years of follow-up</td>
<td>26,749</td>
<td>31,864</td>
<td>37,881</td>
<td>1,186,501</td>
</tr>
<tr>
<td>Sudden cardiac deaths</td>
<td>46</td>
<td>51</td>
<td>53</td>
<td>1,337</td>
</tr>
<tr>
<td>Rate per 10,000 person-years</td>
<td>26.9</td>
<td>14.4</td>
<td>15.7</td>
<td>11.3</td>
</tr>
<tr>
<td>Adjusted rate ratio</td>
<td>2.39</td>
<td>1.30</td>
<td>1.20</td>
<td>1</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>1.77–3.22</td>
<td>0.98–1.72</td>
<td>0.91–1.58</td>
<td>(reference)</td>
</tr>
</tbody>
</table>


tens of thousands of patients would need to be enrolled because sudden death is rare.

Instead Ray and colleagues answered the question using a nonrandomized design. The investigators linked Medicaid files (which included prescription records) to death certificates to compare the frequency of sudden death among persons taking moderate-dose antipsychotics, persons taking low-dose antipsychotics, those who previously took antipsychotics, and non-users of antipsychotics.

Note in the first line of Table 5.2 the large number of person-years in each of the categories. This is a major asset for the study because of the small number of events (line 2) and the low rate of events per 10,000 person-years (line 3). With a small sample size differences between the groups would be hard to detect. In fact, prior randomized studies comparing antipsychotics to placebo had too few patients to identify this rare side effect. With this large observational sample the investigators demonstrated that sudden death was significantly higher among moderate users of antipsychotics than among non-users of antipsychotics.

The fifth indication for performing a nonrandomized study is to increase generalization of a finding. Sometimes an intervention has been proven to work in a randomized clinical trial (efficacy trial), but there are concerns as to whether it would work in a real-world setting (effectiveness trial) (Section 2.2.C). This is especially true of interventions that are complicated or require a lot of cooperation on the part of subjects.

For example, several randomized controlled trials have demonstrated that persons with atrial fibrillation and nonvalvular heart disease should receive anticoagulation with warfarin to prevent ischemic strokes. Although the prevalence of bleeding is higher in persons taking warfarin the extra risk has

---

Nonrandomized studies

been shown to be worth the benefit in these tightly controlled research studies. However, in clinical practice warfarin is a difficult drug to manage. It has a narrow therapeutic window: incorrectly taking an extra dose or missing a dose, or changes in diet, can result in wide swings in the level of anticoagulation. Also, few elderly persons were included in the clinical trials, despite the fact that atrial fibrillation is common in the elderly and it is elders in whom bleeds can be particularly a problem due to the higher incidence of falls.

To determine whether the benefits and side effects of warfarin in clinical practice were similar to those seen in research settings, Go and colleagues reviewed medical, pharmaceutical, and laboratory records of 11 526 patients with non-valvular atrial fibrillation and no contraindication to anticoagulation who were enrolled in a staff-model health maintenance organization (Kaiser Permanente). They found that the benefits of warfarin in preventing ischemic stroke, as well as the increased risk of intracranial hemorrhage, were similar in magnitude to those seen in randomized studies. This gives a strong rationale for anticoagulating patients with nonvalvular atrial fibrillation who have no contraindication to it.

5.3 What are the disadvantages of nonrandomization?

The major disadvantage of nonrandomized designs is that the groups may differ at baseline in ways that will confound your results. Confounding occurs in an intervention study when a factor is associated with whether or not the participant receives the intervention and is causally related to outcome (Figure 5.1).

To illustrate the impact of confounding in a nonrandomized study let’s look at a study assessing the efficacy of influenza vaccination for elderly persons living in the community. Using administrative databases of persons enrolled in a health maintenance organization, the researchers identified persons who had received vaccination and compared them to those who had not. As you can see in Table 5.3, the mean number of hospitalizations due to respiratory infections during two influenza seasons is higher in the vaccination group. From these data we would conclude that the vaccine is not effective in decreasing hospitalizations, and may actually increase hospitalizations due to respiratory infections.

However, as is usually the case in nonrandomized studies, the two groups differed on a number of baseline characteristics, including frequency of coronary heart disease, chronic lung disease, diabetes, vasculitis or rheumatologic

Evaluating Clinical and Public Health Interventions

Table 5.3. Hospitalizations per 1000 elderly persons for acute and chronic respiratory conditions among vaccine recipients and non-recipients during two seasons of influenza.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes (n = 15 288)</td>
<td>Yes (n = 14 647)</td>
</tr>
<tr>
<td></td>
<td>No (n = 11 081)</td>
<td>No (n = 11 979)</td>
</tr>
<tr>
<td>Mean no. of hospitalizations per 1000 persons for acute and chronic respiratory conditions.</td>
<td>31.8</td>
<td>30.7</td>
</tr>
<tr>
<td></td>
<td>31.0</td>
<td>26.5</td>
</tr>
</tbody>
</table>


disease, and dementia or stroke (Table 5.4). And in most cases the differences favored the no-vaccination group; in other words, the persons receiving the vaccination were sicker. This type of confounding is referred to as confounding by indication. Confounding by indication occurs when a factor is related to the likelihood of a subject receiving an intervention and is causally related to outcome.  

In the case of this study, patients who were chronically sicker were more likely to receive influenza vaccine because they and their physicians were more concerned about them becoming seriously ill if they were to develop influenza. (You can also have confounding by contraindication if the treatment is one that is avoided for sicker people, such as surgery.)

---

7 Confounding by indication should be distinguished from selection bias. Selection bias occurs when a variable influences whether a subject receives an intervention (e.g., region where the subject lives) but is not independently associated with the outcome (generally where you live does not affect your likelihood of disease, unless the disease is due to an environmental exposure). For more on this issue see: Salas, M., Hofman, A., and Stricker B.H. “Confounding by indication: an example of variation in the use of epidemiologic terminology.” *Am. J. Epidemiol.* **149** (1999): 981–3.

As you can see in Table 5.5, when the investigators used multivariable analysis to adjust for baseline differences between the two groups, the rate of hospitalization was significantly higher for those who did not receive the vaccine.

Table 5.4. Baseline characteristics of patients by vaccination status during two influenza seasons.  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes (n = 15 288)</td>
<td>No (n = 11 081)</td>
</tr>
<tr>
<td>Outpatient diagnoses during prior 12 mo. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>15.5</td>
<td>8.9</td>
</tr>
<tr>
<td>Chronic lung disease</td>
<td>9.9</td>
<td>5.2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>10.8</td>
<td>6.4</td>
</tr>
<tr>
<td>Vasculitis or rheumatologic disease</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Dementia or stroke</td>
<td>2.2</td>
<td>3.9</td>
</tr>
<tr>
<td>Inpatient diagnoses during prior 12 mo. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>4.8</td>
<td>2.9</td>
</tr>
<tr>
<td>Chronic lung disease</td>
<td>2.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Pneumonia prior 12 mo. (%)</td>
<td>4.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>


Table 5.5. Hospitalizations per 1000 elderly persons for acute and chronic respiratory conditions among vaccine recipients and nonrecipients during two influenza seasons.  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes (n = 15 288)</td>
<td>No (n = 11 081)</td>
<td>Yes (n = 14 647)</td>
</tr>
<tr>
<td>Acute and chronic respiratory conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>31.8</td>
<td>31.0</td>
<td>30.7</td>
</tr>
<tr>
<td>Adjusted</td>
<td>23.7</td>
<td>39.0</td>
<td>24.1</td>
</tr>
<tr>
<td>Difference (95% CI)</td>
<td>−15.3 (−22.4 to −8.2)</td>
<td>−9.0 (−16.0 to −2.0)</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

The 95% confidence intervals for the difference between the hospitalization rates excludes zero. The multivariable analysis tells us a different – and more valid – story about the effectiveness of influenza immunization in preventing hospitalizations due to respiratory disease.

Multivariable analysis is the most commonly used analytic method for dealing with confounding in nonrandomized designs. Stratification – looking for differences within groups – can be used when there are only one or two important differences but this was not the case in this study, where the vaccinated group and the nonvaccinated group differed on a large number of characteristics. And, in fact, it is rare for groups to differ by only one or two factors, and so stratification is not often used to adjust for baseline differences. Other advanced methods of dealing with confounding in nonrandomized studies include propensity analysis, instrumental analysis, and sensitivity analysis (Chapter 7).

There are also ways of achieving comparable baseline groups through the design phase of your study. These are considered in the next section.

**5.4 How can I assemble comparable samples in the design phase without randomization?**

Two methods for assembling comparable samples in the design phase are matching and identification of similar groups (clusters) of participants.

**5.4. A Matching**

With individual matching, persons who receive the intervention are matched to persons not receiving the intervention on those variables that are thought to be related to outcome.

For example, Sjostrom and colleagues used an individual matching design to evaluate the effectiveness of bariatric surgery for obesity.\(^9\) Obese subjects were

---

## Table 5.6. Comparison of matched patients self-selecting for surgery or for conventional treatment.

<table>
<thead>
<tr>
<th></th>
<th>Surgical treatment ((n = 2010))</th>
<th>Conventional treatment ((n = 2037))</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex (no.)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>590</td>
<td>590</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1420</td>
<td>1447</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Women who were post-menopausal (%)</strong></td>
<td>32</td>
<td>36</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Age at examination (yr)</strong></td>
<td>46</td>
<td>47</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Daily smoking (%)</strong></td>
<td>28</td>
<td>20</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Diabetes (%)</strong></td>
<td>7</td>
<td>6</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Weight (kg)</strong></td>
<td>119</td>
<td>117</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Height (m)</strong></td>
<td>1.7</td>
<td>1.7</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Waist circumference (cm)</strong></td>
<td>124</td>
<td>122</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Hip circumference (cm)</strong></td>
<td>126</td>
<td>124</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Systolic blood pressure</strong></td>
<td>141</td>
<td>140</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Triglycerides (mmol/liter)</strong></td>
<td>2.23</td>
<td>2.15</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Cholesterol (mmol/liter)</strong></td>
<td>5.84</td>
<td>5.75</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>Current health score</strong></td>
<td>21</td>
<td>23</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Monotony avoidance score</strong></td>
<td>23</td>
<td>23</td>
<td>0.525</td>
</tr>
<tr>
<td><strong>Psychasthenia score</strong></td>
<td>23.9</td>
<td>23.2</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Quantity of social support</strong></td>
<td>6</td>
<td>6</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>Quality of social support</strong></td>
<td>4</td>
<td>4</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>Stressful life events</strong></td>
<td>3</td>
<td>2</td>
<td>0.09</td>
</tr>
</tbody>
</table>


Nonrandomized studies recruited through the mass media and through primary health care centers, and were asked to complete questionnaires, including stating their preference for surgical treatment or for medical treatment. Subjects who wanted surgery and who were eligible for it were scheduled for the procedure. Those who did not want surgery formed the pool of potential controls.

Using a computer matching algorithm, cases were matched to controls on 18 variables (a case could have more than one match). As you would imagine, if you required that the cases match the controls exactly on 18 variables, such that a case weighing 120.3 kg could only match with a case weighing exactly 120.3 kg, you would have very few matches. Instead, you specify reasonable ranges for each variable; you might, for example, specify that the weight must be within 2 kg. Because the matches are not exact, the groups may still differ significantly from one another on certain variables after matching. This can be seen in Table 5.6.

After the investigators performed their matching process, the surgery group and the control group were similar on seven variables: sex, prevalence of diabetes, height, systolic blood pressure, monotony avoidance score, quantity of social support, and quality of social support. On the other eleven variables, there remained statistically significant differences, or near significant differences, but many of these differences were not very large. For example, the difference in the average weight between the two groups was only 2.3 kg, and the group choosing the bariatric surgery were heavier, a difference that would be expected to bias against the intervention appearing to be effective.

Nonetheless, because there were significant differences after matching, the investigators used multivariable analysis to adjust for baseline differences between the two matched groups. Adjusting for a number of potential confounders, the investigators found that persons who had bariatric surgery had a lower risk of death (relative hazard 0.73; 95% confidence intervals 0.56–0.95). An important indication of the success of their matching procedure is that the unadjusted risk of death for persons who had undergone bariatric surgery (relative hazard = 0.76; 95% confidence intervals 0.59–0.99) was almost identical to the multivariable result. In other words, although there were significant differences in the two groups after matching, the matching procedure appears to have eliminated confounding, at least for those variables the investigators measured.

This study also illustrates another important point about individual matching as a strategy for achieving comparable groups. The cases and controls should be drawn from the same population. In the case of this study, cases and controls were drawn from the population of obese persons responding to the media or being referred by their providers. Drawing cases and controls from different populations (e.g., cases from hospital wards and controls from random-digit dialing) introduces bias.

There are two different strategies for analyzing matched data. If matching is used only to create comparable groups, then standard statistical techniques can be used to analyze the data – this was the case with the study of bariatric surgery. Alternatively, if the individual one-to-one or one-to-x matching is maintained, then the data must be analyzed with methods specific for matched data.\(^\text{10}\)

Although matching can be helpful in assembling comparable groups it has disadvantages. There may be cases for which there is no match – requiring the case to be dropped from the sample. Once you match for a variable, you cannot study its impact on the outcome. Finally, if the matched variables are associated

Nonrandomized studies

with group membership, matching can cause selection bias. For these reasons, most investigators prefer not to match and instead use analytic methods of achieving comparable groups.

5.4.B Identification of comparable clusters (groups)

In Section 4.5, I reviewed clustered (grouped) randomization. However, even without randomization, identifying comparable clusters can be helpful in assembling comparable groups of subjects.

Commonly used clusters for nonrandomized designs with examples are shown in Table 5.7.

You may be wondering why once you have assembled comparable clusters you wouldn’t go ahead and randomize them. This would be a superior design from a statistical point of view. There are three reasons for performing a nonrandomized cluster study: (1) randomization is too difficult to accomplish or will not be accepted by the target population; (2) your study is taking advantage of a natural experiment and you can’t choose the group receiving the intervention; or (3) the number of clusters is too small for randomization to result in comparable groups (Section 4.5).

To illustrate the real-world difficulties of randomizing even within clusters, consider a study designed to assess the effectiveness of vaccination in preventing influenza among children. The investigators identified 11 demographically similar clusters of schools. In the case of seven of the 11 clusters, the intervention school (where the children received the influenza vaccination) was chosen randomly. But in the case of four of the clusters, the school administrators chose which school within the cluster would receive the vaccination. The schools would not allow the use of placebos because they felt that there would be no benefit to the children in receiving placebo. Although a fully randomized, placebo-controlled study would have been a stronger design for proving efficacy of the vaccination, the study would not have happened had the investigators held rigidly to these criteria. As it happened, the study demonstrated that children who received the influenza vaccination were less likely to have a febrile influenza-like illness during the peak influenza week. When the results were limited to those children in schools chosen randomly, vaccination remained effective.

11 The term “clusters” can be confusing, in part, because there can be different levels of clusters. For example, a school has a cluster of students; a district has a cluster of schools; a state has a cluster of districts, etc. In the case of this study, the choice of which school within the cluster of eligible schools received the intervention was made either randomly or by the administrators.

As occurs with individual matching, there may be differences between the individuals within matched clusters that require statistical adjustment with multivariable analysis. These methods are discussed in Chapter 7.

Table 5.7. Nonrandomized designs using clusters.

<table>
<thead>
<tr>
<th>Potential cluster</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools</td>
<td>An intervention to improve diet and increase physical activity of students was implemented at six schools; the comparison group was students from eight matched schools.(^a)</td>
</tr>
<tr>
<td>Towns or cities</td>
<td>An education intervention for adults on cardiovascular disease prevention was implemented in three communities; the comparison group was adults in three communities matched on size and type.(^b)</td>
</tr>
<tr>
<td>Hospitals or clinics or doctors</td>
<td>An intervention to reduce antibiotic prescribing for acute bronchitis in adults was implemented in one primary care practice; the comparison groups were one clinic that received only a limited intervention, and two clinics that provided usual care.(^c)</td>
</tr>
<tr>
<td>Businesses</td>
<td>An intervention to improve nutrition of shoppers was implemented in 20 supermarkets; the comparison group was 20 supermarkets matched with regard to neighborhood racial composition, age, occupation, income, proportion of homeowners, and presence of pharmacy in the store.(^d)</td>
</tr>
<tr>
<td>Neighborhoods, police beats</td>
<td>A multifaceted violence program shows decreased shootings in police beats that received intervention compared to those that did not.(^e)</td>
</tr>
</tbody>
</table>


6.1 How do I test whether my intervention has had a statistically significant effect?

Up until this section, I have not explicitly addressed how to test whether the effect of an intervention is statistically significant. We have, however, covered the basic ingredients.

First, the study hypothesis should be stated in the null. To review, the null hypotheses for the three major questions of intervention studies are:

1. There is no change between the pre-intervention and the post-intervention assessment.
2. The change between the pre-intervention and the post-intervention assessment in the intervention group is no greater than the change in the comparison group.
3. The outcome for the intervention group is no different than that of the comparison group.

Next we use statistics to determine the probability that the null hypothesis is correct. If the probability is very low, we consider the alternative hypothesis: that the intervention is associated with a significant change.

To choose the correct statistic you need to identify:

(a) the nature of your outcome variable,
(b) the number of times you have measured the outcome,
(c) the number of study groups,
(d) whether you need a bivariate test or a multivariable test, and
(e) whether your data are longitudinal cohort or serial cross-sectional.

To help you identify the type of outcome variable you have, I have listed them with examples in Table 6.1.¹

¹ I have not included nominal variables in Table 6.1 because they are infrequently used as outcome variables. For a detailed review of variable types, including how to determine whether an interval variable has a normal distribution see Katz, M. H. Study Design and Statistical Analysis: A Practical Guide for Clinicians. Cambridge: Cambridge University Press, 2006: pp. 35–6, 52–9.
Evaluating Clinical and Public Health Interventions

In terms of number of measurements, the choice of statistic generally depends on whether you have two data points, multiple data points, or a series of data points (e.g., 20 or more); the latter circumstance will be considered in the section on time series (Chapter 8). For choice of statistics, whether you have two study groups or more than two will affect some decisions. Finally, the statistics for longitudinal cohort studies are different than those for cross-sectional studies because with the former the observations are not independent of one another.

To make it easier for you to find the right test, I have organized the statistical sections by these different components: hypothesis 1 is Sections 6.2 and 6.3, hypothesis 2 is Sections 6.4 and 6.5, and hypothesis 3 is Section 6.6. Within each of these sections, I will cover the statistics for the different types of outcome, different numbers of observations, different numbers of groups, and bivariate and multivariable statistics. Because the statistics for longitudinal studies are very different than those for serial cross-sectional studies, I discuss these separately (Sections 6.2 and 6.4 are for longitudinal studies and Sections 6.3 and 6.5 are for serial cross-sectional studies).

### 6.2 Is the difference between the pre-intervention and the post-intervention assessment statistically significant in a longitudinal study?

The bivariate tests for comparing the difference between the pre-intervention and the post-intervention assessments in longitudinal studies without a control group (one group) are shown in Table 6.2.

<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interval (continuous), normal distribution</td>
<td>Equal-sized intervals on all parts of the scale are equal. Values on the variable have a bell-shaped distribution.</td>
<td>Blood pressure in a probability sample of the population.</td>
</tr>
<tr>
<td>Interval (continuous), non-normal distribution</td>
<td>Equal-sized intervals on all parts of the scale are equal. Values on the variable are not bell-shaped, but may be skewed to the right, to the left, or have a bimodal distribution or a threshold distribution.</td>
<td>Length of hospital stay (typically, the variable will be skewed to the right; the shortest stay is usually less than a day, the mean stay may be 3–4 days, and some patients stay in the hospital for months).</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Multiple categories that can be ordered.</td>
<td>NYHA classification of heart failure.</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>Two categories.</td>
<td>Alive/dead, heart attack (yes/no).</td>
</tr>
<tr>
<td>Time to outcome</td>
<td>Time from start of observation to dichotomous outcome.</td>
<td>Time to death, time to cancer remission.</td>
</tr>
</tbody>
</table>

In terms of number of measurements, the choice of statistic generally depends on whether you have two data points, multiple data points, or a series of data points (e.g., 20 or more); the latter circumstance will be considered in the section on time series (Chapter 8). For choice of statistics, whether you have two study groups or more than two will affect some decisions. Finally, the statistics for longitudinal cohort studies are different than those for cross-sectional studies because with the former the observations are not independent of one another.

To make it easier for you to find the right test, I have organized the statistical sections by these different components: hypothesis 1 is Sections 6.2 and 6.3, hypothesis 2 is Sections 6.4 and 6.5, and hypothesis 3 is Section 6.6. Within each of these sections, I will cover the statistics for the different types of outcome, different numbers of observations, different numbers of groups, and bivariate and multivariable statistics. Because the statistics for longitudinal studies are very different than those for serial cross-sectional studies, I discuss these separately (Sections 6.2 and 6.4 are for longitudinal studies and Sections 6.3 and 6.5 are for serial cross-sectional studies).

The bivariate tests for comparing the difference between the pre-intervention and the post-intervention assessments in longitudinal studies without a control group (one group) are shown in Table 6.2.

I will review the tests for each category of outcome variable in the next four sections.
6.2. A Interval outcome with a normal distribution

When you have a pre-intervention versus post-intervention design, and an interval variable with a normal distribution, the easiest way to demonstrate the difference between the two measurements is to calculate the mean difference (post-intervention mean – pre-intervention mean) and the 95% confidence intervals of that difference. This is an especially satisfying method when the outcome is in clinically meaningful units, such as kilograms of weight, mm of Hg of blood pressure, etc. With variables such as these, the reader quickly gains a sense of whether the difference is clinically meaningful in addition to whether it is statistically significant.

For example, Savoye and colleagues developed a weight management program, Bright Bodies, for overweight children aged 8 to 16 years.\(^2\) The family-based intervention included exercise, nutrition, and behavioral modification. At 6 months, the average weight loss was 2.6 kg (Table 6.3). This gives you a good sense of the average impact of the intervention. The fact that the 95% confidence intervals exclude 1 tells you that the reduction was unlikely to be due to chance. Unfortunately by 12 months the average weight change from baseline was a slight increase. The confidence intervals do not exclude 1, telling you that the change in weight is not statistically significant.

Another way to test whether the difference between a pre- and post-intervention measurement of a normally distributed interval variable is greater than you would expect by chance is to use a paired \(t\)-test. For example, Garfinkel and colleagues evaluated whether yoga would relieve the symptoms of carpal tunnel syndrome.\(^3\) They found that there was a significant improvement from


Evaluating Clinical and Public Health Interventions

Table 6.3. Weight change among 105 overweight children enrolled in Bright Bodies.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>6 months</th>
<th>12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean weight (kg)</td>
<td>87.0</td>
<td>84.4</td>
<td>87.3</td>
</tr>
<tr>
<td>Mean change from baseline (kg)</td>
<td>−2.6</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>95% confidence intervals</td>
<td>−4.2 to −0.9</td>
<td>−1.4 to 2.0</td>
<td></td>
</tr>
</tbody>
</table>


Table 6.4. Comparison before and after a yoga course (twice a week for 8 weeks).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean (SD)</th>
<th>Pre-course</th>
<th>Post-course</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip strength, mmHg</td>
<td></td>
<td>161.6 (70.4)</td>
<td>187.4 (68.8)</td>
<td>0.009</td>
</tr>
<tr>
<td>Pain, visual analog scale (1–10)</td>
<td>5.0 (2.8)</td>
<td>2.9 (2.2)</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Median nerve sensory conduction, ms</td>
<td>4.40 (1.5)</td>
<td>3.97 (1.5)</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Median nerve motor conduction, ms</td>
<td>4.79 (1.3)</td>
<td>4.27 (1.4)</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>


To compare multiple measurements of a normally distributed interval variable use repeated measures analysis of variance.

pretest to posttest on several measures; specifically, grip strength increased and the amount of pain decreased; there was also a trend toward improvement in the sensory conduction times (faster times are better) (Table 6.4).

The statistical significance of the differences in Table 6.4 are assessed using a paired $t$-test. A paired $t$-test is calculated as the mean change between the paired observations divided by the standard deviation of that change. Large $t$ values are associated with small $P$ values.

When you have more than two measures of a normally distributed interval variable you cannot calculate a mean difference or use a paired $t$-test. Instead use repeated measures analysis of variance.

For example, Taubert and colleagues evaluated the impact of dark chocolate on blood pressure. The changes in blood pressure at 6 weeks, 12 weeks, and 18 weeks from baseline are shown in Table 6.5.

If the $P$ value of the repeated measures analysis of variance is statistically significant you can make pairwise comparisons. However, the $P$ values from these comparisons must be adjusted for multiple comparisons. The authors of

Statistical analysis of intervention trials

Table 6.5. Effect of consumption of dark chocolate on blood pressure.

<table>
<thead>
<tr>
<th></th>
<th>Mean change from baseline (mmHg)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 weeks</td>
<td>12 weeks</td>
</tr>
<tr>
<td>Systolic</td>
<td>−0.6</td>
<td>−2.4</td>
</tr>
<tr>
<td>Diastolic</td>
<td>−0.3</td>
<td>−1.3</td>
</tr>
</tbody>
</table>


this paper used the Holm method. For example, the decrease in blood pressure between baseline and 6 weeks was not statistically significant ($P = 0.16$) but the decrease between baseline and 12 weeks and the decrease between baseline and 18 weeks were both statistically significant ($P < 0.001$).

The most important aspect of the study is that it provides a handy excuse to eat more chocolate!

6.2.B Non-normally distributed interval variable or ordinal variable

When you have a non-normally distributed interval variable or an ordinal variable and you need to test the difference between the pre-intervention and the post-intervention measurement, use the Wilcoxon signed-rank test. For example, Belfort and colleagues evaluated the effectiveness of the medication pioglitazone for the treatment of steatohepatitis (fatty liver) in diabetics. The investigators found there was significant improvement in all four pathological measures of liver disease following treatment (Table 6.6). Because these measures were ordinal, statistical significance was based on the Wilcoxon signed-rank test.

When you have a non-normally distributed interval or ordinal outcome at three or more points in time, use the Friedman test. For example, Back and colleagues evaluated the impact of tonsillectomy on erythrocyte sedimentation rate (a measure of inflammation). The medians of the rate are shown in Table 6.7.

The Friedman test was significant at $P < 0.001$, suggesting that the procedure causes an inflammatory host response. As with a repeated measures analysis of

6 Wilcoxon signed-rank test is not the same as the Wilcoxon rank sum test, which is also known as the Mann–Whitney test. The Wilcoxon signed-rank test, as with other statistics recommended for using with variables that do not have a normal distribution, is referred to as a non-parametric test.
Evaluating Clinical and Public Health Interventions

Table 6.6. Histologic findings of liver biopsy specimens before and after treatment with pioglitazone.

<table>
<thead>
<tr>
<th></th>
<th>Before treatment n</th>
<th>After treatment n</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steatosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (&lt;5%)</td>
<td>0</td>
<td>6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1 (5–33%)</td>
<td>5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2 (&gt;33–66%)</td>
<td>8</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3 (&gt;66%)</td>
<td>13</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Ballooning necrosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (None)</td>
<td>4</td>
<td>13</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1 (Few balloon cells)</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>2 (Many balloon cells)</td>
<td>16</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Lobular inflammation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0 or 1 focus)</td>
<td>0</td>
<td>6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1 (2–4 foci)</td>
<td>5</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>2 (&gt;4 foci)</td>
<td>21</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Fibrosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (None)</td>
<td>2</td>
<td>5</td>
<td>0.002</td>
</tr>
<tr>
<td>1 (Perisinusoidal or periportal)</td>
<td>12</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>2 (Perisinusoidal and portal/periportal)</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3 (Bridging fibrosis)</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4 (Cirrhosis)</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>


Table 6.7. Median of erythrocyte sedimentation rate following tonsillectomy.

<table>
<thead>
<tr>
<th></th>
<th>Pre-operative</th>
<th>Post-operative day 1</th>
<th>Post-operative day 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erythrocyte</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sedimentation rate</td>
<td>4 mm/hr</td>
<td>8 mm/hr</td>
<td>17.5 mm/hr</td>
</tr>
</tbody>
</table>


variance, when you have a significant value on the Friedman test, you can make pairwise comparisons by using a method that adjusts for multiple pairwise comparisons. In the case of this study, the investigators used the Dunnet method\(^9\) to show that the change between the pre-operative rate and the rate on the first post-operative day was statistically significant (\(P < 0.05\)), as was the change between the pre-operative rate and the rate on the 14th post-operative day.

\(^9\) The Dunnet test is used for multiple comparisons when you are comparing different groups to a control. In the case of this study the pre-operative rate serves as the control.
When you have a longitudinal study, a dichotomous outcome, and two data points use McNemar’s test to assess whether a significant change has occurred between the pre-intervention and the post-intervention assessments. For example, Slifkin and colleagues evaluated the impact of an insurance program on access to care for children and adolescents. Using McNemar’s test the investigators showed that children and adolescents were significantly more likely to have a provider for checkups and to have had a checkup in the last year, and were less likely to have an unmet medical need after the establishment of the insurance program (Table 6.8).

With three or more measures of a dichotomous outcome over time, use Cochran’s Q. For example, Fagard and colleagues performed a trial of treatment interruptions for human immunodeficiency virus infection. The theory underlying their study was that increased replication of HIV during treatment interruptions would stimulate the immune response and lead to low viral loads. If this theory were correct then you would expect that over time with an increased number of interruptions, the probability of the viral load being detectable would decrease.

The investigators followed 86 patients who interrupted their treatment at five points in time: weeks 2, 12, 22, 32, and 42. Using a cut-off of viral load > 200 copies, the percent of patients having a detectable viral load was stable,

<table>
<thead>
<tr>
<th></th>
<th>At the start of insurance %</th>
<th>After insurance %</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>School-aged children (6–11 years) n = 58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has provider for checkups</td>
<td>68</td>
<td>93</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Had a checkup in last year</td>
<td>16</td>
<td>41</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Had unmet medical need</td>
<td>62</td>
<td>2</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Adolescents (12–18 years) n = 63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has provider for checkups</td>
<td>59</td>
<td>93</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Had a checkup in last year</td>
<td>25</td>
<td>54</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Had unmet medical need</td>
<td>51</td>
<td>2</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

ranging from 60% to 66%. The Cochran $Q$ test had a $P$ value of 0.7, indicating no significant change over time in proportion of patients with detectable viral load. Indeed, later studies confirmed that treatment interruptions are not therapeutically beneficial for HIV-infected patients.

6.2.D Multivariable tests for determining whether there is a statistically significant difference between the pre-intervention and the post-intervention assessment in a longitudinal study

In general, most analyses of pre-intervention versus post-intervention assessments in a single group do not require multivariable analysis. Because there is only one group there is no need to adjust for differences between groups: subjects are essentially acting as their own control.

However, at times single-group pre-intervention versus post-intervention studies will use multivariable analysis. The usual reason is to identify additional predictors of the post-intervention assessment besides the fact that the intervention has occurred.

For example, in Section 3.3 I cited a study by Two Feathers and colleagues assessing an intervention to improve diabetes-related outcomes among African-Americans and Latinos.\(^{12}\) For most of the outcomes of the study there was no control group and the investigators used a pre-intervention versus post-intervention design. Consistent with the above discussion and Table 6.2, pre-intervention changes were evaluated with the Wilcoxon signed-rank test for ordinal variables and interval variables\(^{13}\) and McNemar’s test for dichotomous variables.

To learn more about factors that led to the improved outcomes of their subjects other than the intervention, the investigators conducted two types of multivariable analysis: multiple logistic regression for the two outcomes that were dichotomous (dietary knowledge and following a healthy eating plan) and multiple linear regression for the two outcomes that were linear (number of vegetable servings a day and the glycosylated hemoglobin (Hb$_{a1c}$) level) (Table 6.9).

The first row shows the relationship between the baseline assessment and the post-intervention assessment for each of the four outcomes. Not surprisingly, the baseline assessments are strongly related to the outcomes. But after adjusting for

---


\(^{13}\) The authors don’t state whether their interval variables had a non-normal distribution, but even if they were normally distributed, when you have both ordinal and interval variables it is often easier to use non-parametric statistics with both; there is no harm in using non-parametric statistics with interval variables with a normal distribution, except that non-parametric tests have less power.
the pre-intervention assessment we can appreciate the other associations with the outcomes. For example, after the intervention African-Americans had significantly less dietary knowledge than Latinos and those who attended the third class had greater dietary knowledge than those who missed the third class. Glycosylated hemoglobin (Hb\textsubscript{a1c}) levels were lower among men, persons who tested their blood glucose frequently, and those with higher quality of life.

### Table 6.9. Multivariable analysis of predictors of post-intervention assessment scores for four outcomes.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Dietary knowledge OR (95% CI)</th>
<th>Healthy eating plan OR (95% CI)</th>
<th>Vegetable servings/day OR (95% CI)</th>
<th>Hb\textsubscript{a1c} Regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline response</td>
<td>4.37 (1.2, 12.8)**</td>
<td>3.48 (1.2, 10.6)*</td>
<td>0.584***</td>
<td>0.691***</td>
</tr>
<tr>
<td>Age</td>
<td>0.976 (0.94, 1.01)</td>
<td>0.963 (0.92, 1.01)</td>
<td>0.095</td>
<td>0.059</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.1 (0.28, 4.4)</td>
<td>0.96 (0.25, 3.7)</td>
<td>−0.028</td>
<td>0.187*</td>
</tr>
<tr>
<td>Female</td>
<td>Referent</td>
<td>Referent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>0.16 (0.03, 0.77)*</td>
<td>0.11 (0.02, 0.59)**</td>
<td>−0.099</td>
<td>0.010</td>
</tr>
<tr>
<td>Latino</td>
<td>Referent</td>
<td>Referent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 3 attendance</td>
<td>4.2 (1.4, 12.7)**</td>
<td>1.1 (0.19, 6.1)</td>
<td>0.082</td>
<td>...</td>
</tr>
<tr>
<td>Dietary knowledge</td>
<td>...</td>
<td>4.2 (1.3, 14.1)*</td>
<td>0.186*</td>
<td>...</td>
</tr>
<tr>
<td>Monitoring blood glucose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–3 days/wk</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>Referent</td>
</tr>
<tr>
<td>4–7 days/wk</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>−0.266**</td>
</tr>
<tr>
<td>Quality of life</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>−0.254**</td>
</tr>
</tbody>
</table>

* P<0.05; ** P<0.01; *** P<0.001.


### 6.3 Is the difference between the pre-intervention and the post-intervention assessment statistically significant in a serial cross-sectional study?

The statistical analysis of serial cross-sectional data is both simpler and more complicated than the analysis of longitudinal data. It is simpler because with serial cross-sectional data there is no need to take into account the paired nature of the data as must be done with repeated observations of the same persons. It is more complicated because with serial cross-sectional data multivariable analysis is almost always needed to account for differences in the characteristics of the serial samples.

The statistics for assessing whether differences between pre-intervention and post-intervention assessments are statistically significant in a serial cross-sectional study without a control group are shown in Table 6.10.
Jones and colleagues used a serial cross-sectional design to assess whether a community-level intervention was effective at reducing HIV risk behavior in Black men who have sex with men (MSM). The intervention was conducted in three cities; there were no control cities. The intervention consisted of group discussions with opinion leaders and social marketing materials. To assess the impact of the intervention serial samples of Black MSM were drawn prior to the intervention (wave 1) and at three post-intervention points (waves 2–4).

As you can see in column 1 of Table 6.11, the frequency of unprotected intercourse decreased over the four waves of the study. The chi-square was significant ($P = 0.01$). A significant chi-squared simply tells you that the difference between the expected and observed number of subjects in each cell is unlikely to have occurred by chance. It does not tell you where the difference lies. Looking at the four waves it appears that the largest difference is between the first wave and the fourth wave. The chi-squared $P$ value is equal to 0.001 for this comparison.

The odds ratios in the second column basically provide the same information as the first column but in a format that facilitates comparisons with multivariable models (column 3). The odds ratios in the second column for wave two, three, and four are comparisons to wave one. Note that while the odds ratios for wave two and three are less than one, the confidence intervals cross one and are not significant. The odds ratio for wave four is statistically different from one (the confidence intervals do not cross one), indicating that by wave four the men report that they are significantly less likely to have unprotected intercourse.

Part of why I chose this study to illustrate serial cross-sectional studies is that it demonstrates the strengths and weaknesses of this design. In terms of

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Two groups bivariate</th>
<th>Three or more groups bivariate</th>
<th>Multivariable analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dichotomous</td>
<td>Chi-squared, Fisher’s exact Odds ratio</td>
<td>Chi-squared</td>
<td>Multiple logistic regression</td>
</tr>
<tr>
<td>Interval, normal</td>
<td>$t$-test</td>
<td>ANOVA</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>distribution</td>
<td>Interval, non-normal</td>
<td>Kruskal–Wallis test</td>
<td>ANCOVA</td>
</tr>
<tr>
<td>distribution</td>
<td>Mann–Whitney test</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Mann–Whitney test</td>
<td>Kruskal–Wallis test</td>
<td>Proportional odds regression</td>
</tr>
</tbody>
</table>

*Attempt to transform the outcome variable so that it fits the assumptions of normality and equal variance.


Statistical analysis of intervention trials

Table 6.11. Change in frequency in unprotected anal intercourse with male partners over four waves of the study.

<table>
<thead>
<tr>
<th>Wave</th>
<th>Reported behavior, % (no.)</th>
<th>Unadjusted OR (95% CI)</th>
<th>Adjusted† OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 1</td>
<td>42.1 (119/283)</td>
<td>1.00 (reference)</td>
<td>1.00 (reference)</td>
</tr>
<tr>
<td>Wave 2</td>
<td>34.5 (99/287)</td>
<td>0.73 (0.52, 1.02)</td>
<td>0.73* (0.53, 1.00)</td>
</tr>
<tr>
<td>Wave 3</td>
<td>36.0 (109/303)</td>
<td>0.77 (0.56, 1.08)</td>
<td>0.77 (0.57, 1.06)</td>
</tr>
<tr>
<td>Wave 4</td>
<td>28.7 (77/268)</td>
<td>0.56** (0.39, 0.79)</td>
<td>0.57*** (0.41, 0.80)</td>
</tr>
</tbody>
</table>

†OR adjusted for city where survey took place, employment status, and whether the respondent had ever been to jail.

* P < 0.05; ** P < 0.01; *** P < 0.001

strengths, the data in the second column would be much less compelling if it was a pre- versus post-intervention cohort study because of the possibility that subjects were telling the investigators what they wanted to hear (social desirability responses, Section 3.7). Although there are objective markers of unsafe sex (e.g., HIV incidence, STD acquisition) these would have been difficult to incorporate in a study because they occur at low incidence.

In terms of weaknesses, because this is a serial cross-sectional design, subjects interviewed in wave 1 are not necessarily comparable to the subjects interviewed in wave 4. Indeed, the investigators found there were differences in sample characteristics between the waves. To adjust for these differences the investigators performed multivariable logistic regression (column 3). They included in their model those variables that were associated with the wave or with the outcome. The analysis shows that there is a statistically significant decrease of unprotected intercourse at wave 2 and at wave 4.

I will not illustrate all the other statistics listed in Table 6.10 for determining whether there are statistically significant differences between the pre-intervention and post-intervention assessments because the statistics themselves are the same bivariate and multivariable statistics used for any comparison of groups with independent observations. Because performing multivariable analysis with an interval variable with a non-normal distribution is difficult, the easiest recourse is to transform the outcome (e.g. logarithmic transformation) so that it fits assumptions of normality and equal variance. There will be other illustrations of these statistics in Sections 6.5 and 6.6.

6.4 Is the difference between the pre-intervention and the post-intervention assessment of the intervention group statistically greater (lesser) than that seen in the comparison group(s) in a longitudinal cohort study?

Tests for comparing differences in change over time between the intervention group and the comparison group(s) in longitudinal cohorts are shown
Evaluating Clinical and Public Health Interventions

### Table 6.12. Tests for comparing changes over time between the intervention group and the comparison group(s), longitudinal cohort study.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unadjusted comparison between groups (bivariate)</th>
<th>Adjusted comparison between groups (multivariable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean difference between two time points, two groups</td>
<td>Difference between the mean difference of two groups and 95% confidence intervals</td>
<td>Multiple linear regression ANOVA ANCOVA</td>
</tr>
<tr>
<td>Mean difference between two time points, three or more groups</td>
<td>t-test</td>
<td>Multiple linear regression ANOVA ANCOVA</td>
</tr>
<tr>
<td>Interval outcome, normal distribution, two or more time points, two or more groups</td>
<td>ANOVA</td>
<td>ANOVA ANCOVA</td>
</tr>
<tr>
<td>Proportion, two points, two groups</td>
<td>Repeated measures ANOVA</td>
<td>Repeated measures ANOVA ANCOVA Mixed-effects models</td>
</tr>
<tr>
<td>Proportion, three or more time points, two or more groups</td>
<td>Two-sample generalization of McNemar’s test</td>
<td>Conditional logistic regression Mixed-effects models</td>
</tr>
<tr>
<td>Difference in time to outcome</td>
<td>Conditional logistic regression</td>
<td>Conditional logistic regression Mixed-effects models</td>
</tr>
<tr>
<td></td>
<td>Kaplan–Meier</td>
<td>Proportional hazards regression</td>
</tr>
</tbody>
</table>

in Table 6.12. Because there are often baseline differences between the intervention group and the comparison group that require statistical adjustment, I have included multivariable methods as well as bivariate analysis in Table 6.12.

#### 6.4.A Mean difference between two time points, two groups, bivariate analysis

When you have calculated the mean between two time points, an easy way to look for differences between the groups is to calculate the difference between the mean change of the intervention group and the mean change of the comparison group.

Look back at Table 6.3. Using data from the Bright Bodies study, I illustrated the use of mean difference and the 95% confidence intervals to quantify the difference between pre-intervention and post-intervention for a group of overweight children. We saw that there was a significant weight loss among the children who received the intervention. Now, let’s see whether the weight loss over time was greater in the intervention group than in the control group.
**Table 6.13.** Comparison of mean weight loss among overweight children enrolled in Bright Bodies intervention versus the control group.

<table>
<thead>
<tr>
<th></th>
<th>Intervention group</th>
<th>Control group</th>
<th>Treatment – Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean weight change at 6 months</td>
<td>−2.6 kg</td>
<td>5.0 kg</td>
<td>7.6 kg</td>
</tr>
<tr>
<td>95% confidence intervals</td>
<td>−4.2 to −0.9 kg</td>
<td>2.9 to 7.2 kg</td>
<td>4.3 to 10.8 kg</td>
</tr>
</tbody>
</table>


Column 1 of Table 6.13 shows the loss of weight that occurred in the intervention group in the first 6 months (loss of 2.6 kg), just as in Table 6.3. Column 2 shows that the control group did not lose weight during the same period, but actually gained an average of 5.0 kg. In the last column of the table, you can see the difference between the mean weight loss of the two groups (7.6 kg) that occurred in the first 6 months. The 95% confidence intervals exclude zero, indicating that the difference that occurred between the two groups over time is not likely to be due to chance. The addition of the control group strengthens the inference that the intervention was effective.

Another method of comparing the change in the intervention group to the change in the control group is to use an unpaired *t* -test. For example, Walter and Vaughan evaluated the effectiveness of an AIDS-preventive curriculum on the knowledge, beliefs, self-efficacy, and behavioral risk factors of 9th and 11th grade students.16

Looking across the rows of Table 6.14 note that for each measure there is a baseline score, a follow-up score, and the difference between the baseline and follow-up score for both the intervention group and the comparison group. The statistical significance of the *t* -test is reflected in the final column, the *P* value, which is the probability that the observed difference between the change that occurred in the intervention group and the change that occurred in the comparison group could have occurred by chance alone.

The downside of the use of *P* values as opposed to the mean difference with the 95% confidence intervals (shown in Table 6.13) is that *P* values do not convey the effect size. If your sample is very large (e.g., 10,000 patients) even trivial differences may be statistically significant. Conversely, with a small sample size clinically important differences may not be statistically significant.

---

Table 6.14. Comparison of changes over time between the intervention and the comparison group.

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>Intervention</th>
<th>Comparison</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Follow-up</td>
<td>Difference</td>
<td>Baseline</td>
<td>Follow-up</td>
</tr>
<tr>
<td>Knowledge</td>
<td>75.6</td>
<td>85.5</td>
<td>9.9</td>
<td>78.8</td>
<td>81.2</td>
</tr>
<tr>
<td>Beliefs susceptibility</td>
<td>2.5</td>
<td>2.1</td>
<td>−0.4</td>
<td>2.5</td>
<td>2.3</td>
</tr>
<tr>
<td>Benefits</td>
<td>3.5</td>
<td>3.8</td>
<td>0.3</td>
<td>3.7</td>
<td>3.8</td>
</tr>
<tr>
<td>Barriers</td>
<td>4.4</td>
<td>4.5</td>
<td>0.1</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Values</td>
<td>5.4</td>
<td>5.5</td>
<td>0.1</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Norms</td>
<td>2.8</td>
<td>2.9</td>
<td>0.1</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>3.7</td>
<td>3.9</td>
<td>0.2</td>
<td>3.7</td>
<td>3.8</td>
</tr>
<tr>
<td>Behavior risk index</td>
<td>1.5</td>
<td>1.3</td>
<td>−0.2</td>
<td>1.0</td>
<td>1.3</td>
</tr>
</tbody>
</table>


6.4.B Mean difference between two time points, three groups, bivariate analysis

With three or more groups you cannot use a t-test to compare the changes among your groups. Instead, use ANOVA. For example, Tan evaluated the efficacy of phototherapy for reducing neonatal jaundice.\textsuperscript{17} As you can see in Table 6.15, the mean decrease in bilirubin in 24 hours differed in the three groups.

The significant $P$ value does not tell you where the difference lies. However, in situations like this, where the $P$ value of the ANOVA is significant, you may then perform pairwise comparisons. Because you will be essentially performing multiple comparisons (formula-fed versus breast-fed infants; formula-fed versus formula- and breast-fed infants; breast-fed versus formula- and breast-fed infants), the $P$ values should be adjusted to be more stringent. One popular method is the Bonferroni correction.\textsuperscript{18}

Although the study of infants did not report the Bonferroni correction it is easy to do. It is simply the significance level assuming a single test (usually $P = 0.05$) divided by the number of comparisons performed (in this case three). This then yields a new more stringent $P$ value of 0.017 (0.05/3). The pairwise comparison between breast-fed infants and formula- and breast-fed infants had a $P$ value of 0.007, below the more stringent $P$ value.


\textsuperscript{18} Although easy to compute and conceptually appealing the Bonferroni correction can be too conservative, especially with small sample sizes and a large number of pairwise comparisons. For a list and brief description of pairwise tests, see: Pairwise comparisons in SAS and SPSS at www.uky.edu/ComputingCenter/SSTARS/www/documentation/MultipleComparisons_3.htm; for a more thorough discussion see: Glantz, S. A. Primer of Biostatistics (5th edition). New York: McGraw-Hill, 2002: pp. 89–107.
**Statistical analysis of intervention trials**

You may see articles using ANOVA to compare the difference scores of only two groups. If the difference scores are being compared without adjustment for other factors, then the ANOVA is identical to an unpaired \( t \)-test.

6.4.C **Mean difference between two time points, two or more groups, multivariable analysis**

When you are comparing the mean difference among two or more different groups, you may need to adjust for baseline differences between the groups. The simplest method for adjusting for baseline differences is multiple linear regression. For example, in the study (above) evaluating the effectiveness of an AIDS-preventive intervention, the investigators used multiple linear regression to adjust their data for baseline differences.

In their linear regression model the outcome was the follow-up score, and the independent variables were the intervention group, the baseline score, age, gender, and race/ethnicity. The standardized regression coefficients shown in Table 6.16 represent the adjusted change in the outcome variable that is due to the intervention. All the changes were statistically significant except that for the variable “values.”

If you have more than two groups, you will need to create multiple dichotomous variables to represent the groups. The number of variables you will need will be 1 less than the number of groups you have. For example, if you have three groups A, B and C, you will create two variables: group A (yes/no) and group B (yes/no) and enter them into the model as independent variables.\(^{19}\)

You can also adjust for baseline differences using ANOVA or an ANCOVA. With ANOVA you can include baseline variables that are dichotomous or nominal such as sex and ethnicity. When you need to adjust for interval covariates instead of, or in addition to, dichotomous or nominal variables use analysis of covariance (ANCOVA).

---

Evaluating Clinical and Public Health Interventions

Table 6.16. Standardized linear regression coefficients for intervention effect on outcome variables.

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>Standardized coefficients*</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>0.25</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Beliefs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Susceptibility</td>
<td>-0.08</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Benefits</td>
<td>0.08</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Barriers</td>
<td>0.05</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Values</td>
<td>0.03</td>
<td>NS</td>
</tr>
<tr>
<td>Norms</td>
<td>0.09</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>0.07</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Behavior risk index</td>
<td>-0.10</td>
<td></td>
</tr>
</tbody>
</table>


6.4.D Interval outcome, two or more time points, two or more groups

When you are comparing two or more groups, use repeated measures ANOVA to compare changes across two or more time points.

Repeated measures ANOVA is an extension of ANOVA that can be used to compare changes over time at multiple points with two or more groups. For example, in the study by Belfort and colleagues evaluating the effectiveness of the medication pioglitazone (mentioned above) in treating non-alcoholic steatohepatitis, the researchers compared the effect of pioglitazone to placebo on several parameters (Table 6.17). The P values shown under the headings pioglitazone and placebo are based on a paired t-test. This is a within-group comparison (i.e., is there a statistically significant change between the before and after treatment assessment). It is exactly the same test as discussed in Section 6.2.A.

The P value in Table 6.17 for pioglitazone versus placebo (final column) is based on a repeated measures analysis of variance. It answers the statistical question: is the change over time different for the subjects who received pioglitazone than for those who received placebo? This is a between-group comparison, and the answer in this case is that the difference between the two assessments is greater in the pioglitazone group than in the placebo group.

In Section 6.2 we considered an intervention to have worked if there was a significant difference between the pre-intervention and the post-intervention assessments. However, with a control group we only conclude that an intervention works when the change in the intervention group is significantly larger than the change in the control group.

If you have any doubt that there must be a significantly larger change in the intervention group than in the control group to conclude that the intervention...
works, consider the following situation. An investigator evaluates the effectiveness of an intervention to decrease symptoms of depression; a control group is included. The investigator finds that over time both groups report decreased symptoms of depression, but the within-group improvement for the intervention arm is just under the $P < 0.05$ threshold, while the within-group improvement for the control group is just over the threshold. In a case such as this, the between-group comparison will not be statistically significant and it is likely that some factor other than the intervention resulted in improvement in both groups.

A study by Anderson and colleagues illustrates the use of repeated measures analysis of variance for an outcome measured at multiple points in time. The investigators randomized obese women to two different interventions: diet plus aerobic exercise or diet plus lifestyle activity. As you can see in Figure 6.1, both groups lost weight, but there was no significant difference between the two groups in terms of how much weight was lost.\(^{20}\)

### 6.4.E Interval outcome, two or more time points, two or more groups, multivariable analysis

The technique repeated measures ANOVA allows adjustment for baseline variables if the variable is dichotomous or nominal. If the baseline variable is interval use repeated measures ANCOVA. For example, Lautenschlager and colleagues evaluated the impact of a physical activity program on older adults with memory problems.\(^{21}\) Subjects were randomized to an activity group or to usual care. A variety of cognitive indices were measured at baseline, and


<table>
<thead>
<tr>
<th>Variable</th>
<th>Pioglitazone (n = 26)</th>
<th>Placebo (n = 21)</th>
<th>$P$ value (pioglitazone vs. placebo)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before treatment</td>
<td>After treatment</td>
<td>$P$ value</td>
</tr>
<tr>
<td>Body-mass index</td>
<td>33.5</td>
<td>34.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>93.7</td>
<td>96.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Body fat (%)</td>
<td>33.7</td>
<td>35.2</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Fasting plasma glucose (mg/dl)</td>
<td>119</td>
<td>99</td>
<td>0.004</td>
</tr>
<tr>
<td>Glycated hemoglobin (%)</td>
<td>6.2</td>
<td>5.5</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

---


Evaluating Clinical and Public Health Interventions

6 months, 12 months, and 18 months. The investigators used repeated measures ANCOVA so that they could assess changes over time adjusted for both interval variables (e.g., age, premorbid IQ) and categorical variables (e.g., sex, marital status). They found that exercise was associated with a modest improvement in cognition.

Repeated measures ANOVA and repeated measures ANCOVA have several weaknesses. For example, there must be an equal number of observations per subject and fixed periods of time between observations. Also, repeated measures ANOVA and repeated measures ANCOVA assume sphericity. In a longitudinal study sphericity means that the correlation between any two measurements at different time points is the same and that within subjects there is equal variance of measurements. When you have three or more measurements this assumption is often not met. For these reasons, investigators are increasingly using mixed-effects models or generalized estimating equations when comparing changes over time in an interval variable over time. Unfortunately, these procedures are beyond the scope of this book.²²

6.4.F Dichotomous outcome, two or more time points, two or more groups, bivariate and multivariable

You will remember from Section 6.2.C that McNemar’s test is used to assess the change occurring between two time points (e.g., pre-intervention and post-intervention) in one sample when you have a dichotomous variable. When you have two samples, the difference can be compared using a two-sample generalization of McNemar’s test. For example, Kegeles and colleagues developed and implemented a community-level HIV-prevention program for young gay men in a mid-sized community. Young men in the intervention community were surveyed before and after the intervention and young men in a comparison community were surveyed at two similar times. The intervention was successful at decreasing high-risk behavior in the intervention community compared to the comparison community based on the two-sample generalization of McNemar’s test (Table 6.18).

In situations where you have more than two time points or when you need to adjust for baseline differences, you will not be able to use the two-sample generalization of McNemar’s test. Instead, one option is to use conditional logistic regression analysis. Conditional logistic regression, like standard logistic regression, is used with dichotomous outcomes, but unlike standard logistic regression it can incorporate multiple observations of the same subject.

However, as is true of the family of ANOVA procedures, conditional logistic regression can only be used if there are the same number of observations at each time point. Mixed-effects models and generalized estimating equations are more flexible in that they allow for multiple data points of the same persons and differences in the numbers of observations among the subjects (see footnote 22 above).

---

### Table 6.18. Impact of a community mobilization in reducing high-risk sex among young gay men.

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-intervention</th>
<th>Post-intervention</th>
<th>Pre-to post-intervention difference, %</th>
<th>P (within group)</th>
<th>P (intervention vs. control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprotected anal intercourse in prior 2 months</td>
<td>Intervention 41.0</td>
<td>Control 38.6</td>
<td>−11.0</td>
<td>&lt;0.05</td>
<td>P &lt;0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>30.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>39.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NS = non-significant.


---

6.4.G Comparison of time to outcome, two or more groups, bivariate and multivariable

We are often interested in whether an intervention affects time to outcome (e.g., time to death, time to recurrence of cancer). To determine time to outcome we use survival analyses, a group of procedures that incorporate time into the analysis. Survival curves are used to describe the proportion of persons experiencing an event over a period of time. The most common method of calculating a survival curve is the Kaplan–Meier method. To determine whether an intervention affects time to outcome we can compare the Kaplan–Meier curves of two or more groups using the log-rank statistic.

For example, Johnson and colleagues used Kaplan–Meier curves to assess the efficacy of topiramate for treating alcohol dependence. They found that those subjects randomized to topiramate had shorter times to continuous abstinence than those who received placebo (Figure 6.2). The difference between the two curves was statistically significant ($P$ value of log-rank test $<0.001$).

An important feature of survival curves, and of survival analyses generally, is that the technique enables us to incorporate subjects who have been lost to follow-up, withdrawn from the study without allowing themselves to be followed, or have an outcome that precludes them from being assessed for the outcome of interest (e.g., a subject enrolled in a study of time to cancer dies in a car accident). The feature is called censoring and it enables us to include subjects with different lengths of follow-up.

For example, in the study of topiramate versus placebo, 179 subjects were enrolled in the topiramate arm and 185 were enrolled in the placebo arm. However, only 112 subjects in the topiramate arm and 144 subjects in the placebo arm actually finished the trial. Other subjects experienced a limiting side effect, withdrew, were lost to follow-up or did not complete the trial for other reasons. With censoring, subjects contribute their experience until the time that they leave the study. In this way, vital observation time is not lost, as would occur if all subjects who did not complete a trial were excluded.

The numbers under the survival curve of Figure 6.2 indicate the number of people who are at risk for the outcome in the next period of time. Conversely, those people who had a limiting side effect, withdrew, were lost to follow-up, did not complete the study for some other reason, or had already experienced the outcome of interest (abstinence) are no longer counted. That’s the reason as time progresses fewer and fewer people are shown in the legend to be at risk for

---


the outcome. Investigators should always show the numbers of subjects at risk in the legend of their Kaplan–Meier curves otherwise the reader cannot assess the validity of the results; as the sample size shrinks the estimates of percentage outcome in the next period of time become less certain.

Often when comparing the time to outcome of different groups we will want to adjust for baseline differences. For example, in Section 3.4 I cited a study by Hannan and colleagues as an example of using a procedure database to evaluate an intervention.

In that study the investigators assessed how well drug-eluting stents (a new technology) worked compared to coronary artery bypass grafting (CABG), among persons with multivessel coronary artery disease. Since patients were identified through two registries and were not randomized into two groups, there were significant differences between patients in the two groups. Patients undergoing CABG were older, more likely to be male, to be non-Hispanic, to be white, to have lower ejection fractions, to have had previous myocardial infarctions, to have other coexisting conditions, and to have three-vessel disease. To adjust for these differences, the investigators used proportional hazards analysis. They found that patients who received a CABG had a lower risk of death and myocardial infarction than patients who received a drug-eluting stent (Table 6.19). Despite the lack of randomization the results of the study have been broadly accepted and CABG remains the treatment of choice for persons with multivessel disease.
### Table 6.19. Hazard ratios for death and for death or myocardial infarction after CABG and after treatment with a drug-eluting stent, according to number of diseased vessels.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Death</th>
<th>Death or myocardial infarction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted relative hazard* (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td>Three diseased vessels With or without proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.80 (0.65–0.97)</td>
<td>0.03</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>With proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.79 (0.61–1.02)</td>
<td>0.07</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Without proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.79 (0.58–1.09)</td>
<td>0.15</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Two diseased vessels With or without proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.71 (0.57–0.89)</td>
<td>0.003</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>With proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.71 (0.53–0.96)</td>
<td>0.02</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Without proximal LAD artery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.69 (0.48–0.98)</td>
<td>0.04</td>
</tr>
<tr>
<td>Stent</td>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

CABG denotes coronary-artery bypass grafting, and LAD left anterior descending.

* Relative hazards are adjusted for age, sex; ejection fraction; hemodynamic state; history or no history of myocardial infarction before procedure; presence or absence of cerebrovascular disease, peripheral arterial disease, congestive heart failure, chronic obstructive pulmonary disease, diabetes, and renal failure; and involvement of the proximal LAD artery.


### Table 6.20. Multivariable techniques for assessing whether the change from baseline to follow-up is greater (less) in the intervention group than in the comparison group.

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Multivariable analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dichotomous</td>
<td>Multiple logistic regression</td>
</tr>
<tr>
<td>Interval, normal distribution</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td></td>
<td>ANOVA</td>
</tr>
<tr>
<td></td>
<td>ANCOVA</td>
</tr>
<tr>
<td>Interval, non-normal distribution</td>
<td>–*</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Proportional odds regression</td>
</tr>
<tr>
<td>Time to outcome</td>
<td>Proportional hazards (Cox) analysis</td>
</tr>
</tbody>
</table>

* Attempt to transform the outcome variable so that it fits the assumptions of normality and equal variance.
6.5 Is the difference between the pre-intervention and the post-intervention assessment of the intervention group statistically greater (lesser) than that seen in the comparison group(s) in a serial cross-sectional study?

As with cohort studies, serial cross-sectional studies are also strengthened by having a comparison group. For serial cross-sectional studies with a control group multivariable analysis will almost always be necessary because there will be differences within groups and between groups. Therefore, I have shown only multivariable techniques for this type of study design (Table 6.20). As you can see the appropriate multivariable technique depends on the type of outcome measure you have.

Beyond the standard multivariable analyses listed in Table 6.20, generalized estimating equations and mixed-effects models can also be used to answer whether the change is greater in the intervention group than in the comparison group. These models are very flexible and can also incorporate linked observations.

To illustrate the statistical analysis of serial cross-sectional data with a control group, let’s look at a study by Snyder and Anderson. The investigators assessed whether participation in a quality improvement organization improved the care of Medicare beneficiaries. They abstracted the medical records of one group of patients at baseline and the medical records of a different set of patients at follow-up; they assessed improvement on 15 quality indicators that were the focus of the quality improvement organizations during the period between the baseline and follow-up surveys.

Results for two of the 15 indicators are shown in Table 6.21. The $P$ values for the (within-group) differences between the pre-intervention and post-intervention are based on chi-squared analyses, just as you would do if you were performing a serial cross-sectional study without a comparison group (Section 6.3).

Answering the question of whether the change is greater in the intervention group than in the comparison group will require multivariable analysis because we will need to adjust for differences in the samples. The outcome for this study is performance of the quality indicator (yes/no). Because the outcome is dichotomous the authors used multiple logistic regression. The independent variables were participation in the quality improvement organization (yes/no), period (baseline/follow-up), and an interaction between participation in the quality improvement program and period. The interaction tells you whether participating hospitals during the study period improved more or less than nonparticipating hospitals during the study period. In these models the

Table 6.21. Comparison of changes on two quality indicators among hospitals that did or did not participate in quality improvement organization.

<table>
<thead>
<tr>
<th>Intervention effect</th>
<th>Participated in quality improvement organization</th>
<th>Did not participate in quality improvement organization</th>
<th>Intervention effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>Follow-up</td>
<td>Change</td>
</tr>
<tr>
<td>Aspirin prescribed at discharge following heart attack</td>
<td>85.3</td>
<td>88.0</td>
<td>2.7</td>
</tr>
<tr>
<td>Patient screened for or given pneumococcal vaccine</td>
<td>15.5</td>
<td>40.9</td>
<td>25.4</td>
</tr>
</tbody>
</table>

* Odds ratios are adjusted for patient age, sex, and race and hospital bed size and profit status.

Data from Snyder, C., and Andersow, G. "Do quality improvement organizations improve the quality of hospital care for medicaid beneficiaries?" *JAMA* 293 (2005): 2900–7.
investigators adjusted for patient age, sex, and race, and hospital bed size and profit status.

The results of the multiple logistic regression analysis appear in the final column. There was no change in the likelihood of aspirin prescription, but patients in participating hospitals were more likely to be screened or given pneumococcal vaccine.

Because multiple patients were seen in the same hospital, the authors supplemented their main analysis, by performing generalized estimating equations to adjust for the effect of clustering within hospitals. Results were similar so the investigators only reported the results from the more conventional multiple logistic regression analysis.

6.6 Is the post-intervention assessment of the intervention group significantly different than the corresponding assessment of the comparison group?

As discussed in Section 1.2.C, there are times when no pre-intervention measure is available to evaluate an intervention. Instead, an intervention group is compared to a comparison group on a specific outcome. The bivariate and multivariable statistics for comparing a post-intervention assessment of an intervention group to the corresponding assessment of the control group are shown in Table 6.22. As was true of the other statistics listed in this chapter, the choice of statistic depends on the nature of the outcome variable.

You will note that these are the same statistics for comparing pre-intervention versus post-intervention measures with serial cross-sectional data (Section 6.3). This is because in both cases the observations are unlinked.

To demonstrate differences between the group that received the intervention and the control group you will generally need to use multivariable statistics because there will likely be baseline differences between the groups (unless there was a natural randomization that formed the groups – an unlikely circumstance).

For example, Tseng and colleagues evaluated whether Medicare drug benefit caps affected the ability of seniors to obtain their medication. After the drug caps were implemented, the investigators surveyed two groups of seniors: those who had a $750 or $1200 yearly cap and had exceeded it in the prior year (intervention group) and those who had a $2000 cap which they had not exceeded

---

Evaluating Clinical and Public Health Interventions

(control group) – since they had not exceeded it we can assume that they were essentially unaffected by the cap.

The investigators performed a stratified sample of potential controls to identify controls that were similar in age and monthly drug expenditures to those affected by the intervention. Nonetheless, there were significant differences between the intervention and the control group. Controls were more likely to be women, less likely to be married, had lower household income, higher rates of hypertension and emphysema, and a higher mean number of health conditions.

As you can see in Table 6.23, seniors in the intervention group were more likely to adopt strategies to decrease their medication usage, including taking less medication than prescribed, switching medications, using medication samples, calling pharmacies to find the best price, and receiving a discount for being over 65 years. The percentages and the P values were based on multiple logistic regression analysis adjusting for age, sex, education, ethnicity, household income, health status, functional limitations, number of health problems, type of survey, and whether income data were present.

Although the analyses were adjusted for differences between participants in the intervention group and those in the control group, there remains the possibility that underlying differences between these groups explain the differences in their strategies to decrease medication usage. The fact that the different caps were offered in different counties meant that individuals did not self-select to the different caps (which would have made the study design much more problematic) but the baseline characteristics tell us that the participants in the different counties were different.

This problem could have been eliminated if the caps could have been randomized, but randomizing types of insurance coverage for government

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Two groups bivariate</th>
<th>Three or more groups bivariate</th>
<th>Multivariable analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dichotomous</td>
<td>Chi-squared</td>
<td>Chi-squared</td>
<td>Multiple logistic regression</td>
</tr>
<tr>
<td>Interval, normal distribution</td>
<td>t-test</td>
<td>ANOVA</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>Interval, non-normal distribution</td>
<td>Mann–Whitney test</td>
<td>Kruskal–Wallis test</td>
<td>* –</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Mann–Whitney test</td>
<td>Kruskal–Wallis test</td>
<td>Proportional odds regression</td>
</tr>
</tbody>
</table>

*Attempt to transform the outcome variable so that it fits the assumptions of normality and equal variance.
Statistical analysis of intervention trials

Table 6.23. Adjusted percentages of elders who used strategies to lower medication costs.

<table>
<thead>
<tr>
<th>Medication strategies for decreasing costs that involve decreasing use</th>
<th>Intervention participants: exceeded $750 or $1200 cap (n = 665)</th>
<th>Control participants: did not exceed $2000 cap (n = 643)</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used less medication than was prescribed or less than desired (e.g., skipped dose)</td>
<td>18</td>
<td>10</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stopped a medication</td>
<td>8</td>
<td>8</td>
<td>0.86</td>
</tr>
<tr>
<td>Did not fill a prescription for a new medication (i.e., did not start a new medication)</td>
<td>6</td>
<td>5</td>
<td>0.39</td>
</tr>
<tr>
<td>Adopted at least 1 strategy above that decreased medication use</td>
<td>24</td>
<td>16</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Strategies for decreasing cost that may or may not involve decreasing use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switched medications</td>
<td>15</td>
<td>9</td>
<td>0.002</td>
</tr>
<tr>
<td>Received free medication samples</td>
<td>34</td>
<td>27</td>
<td>0.006</td>
</tr>
<tr>
<td>Took someone else’s medication</td>
<td>2</td>
<td>1</td>
<td>0.26</td>
</tr>
<tr>
<td>Called pharmacies to find the best price</td>
<td>46</td>
<td>29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Strategies for obtaining discounts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received discount for being age &gt;65 yr</td>
<td>12</td>
<td>7</td>
<td>0.003</td>
</tr>
<tr>
<td>Received discount for being in Medicare</td>
<td>10</td>
<td>7</td>
<td>0.13</td>
</tr>
<tr>
<td>Received discount from a pharmaceutical company</td>
<td>2</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>Used mail order</td>
<td>63</td>
<td>62</td>
<td>0.64</td>
</tr>
<tr>
<td>Bought medications outside the United States</td>
<td>3</td>
<td>3</td>
<td>0.92</td>
</tr>
</tbody>
</table>

* The adjusted difference is based on multivariate logistic regression including age, sex, education, ethnicity, annual household income, general health status, functional limitations in instrumental activities of daily living, number of health problems, written vs. telephone survey, and whether or not participants had income data as covariates.


Programs is not a very feasible study design. A pre-intervention measure of the same strategies would have made this study stronger – even stronger would have been evidence that seniors actually stopped taking medicines that they were taking prior to the caps. However, in the real world of health policy, this study makes a compelling case that cost caps change the behavior of seniors.

As discussed in Section 1.2.C, case-control studies can be used to assess whether an intervention has been successful in cases where there is no
Evaluating Clinical and Public Health Interventions

pre-intervention measurement. The statistics used to analyze case-control studies are the same as those listed in Table 6.21, as long as the cases and controls are not one-to-one matched.\textsuperscript{28} For example, Werler and colleagues used a case-control design to demonstrate that folic acid was effective in reducing neural tube defects. They identified 436 mothers with a baby born with a neural tube defect (cases) and 2615 mothers with a baby born with a malformation other than neural tube (controls). They asked the mothers about their consumption of folic acid-containing supplements in the periconceptual period. Eight percent of the cases versus 13% of the controls had taken daily folic acid in the preconceptual period, indicating that folic acid was associated with a significant reduction in the odds of having a neural tube defect (OR = 0.57, 95% CI = 0.4–0.8).\textsuperscript{29}

To minimize the effect of confounding, the investigators also performed multiple logistic regression adjusting for all those variables that altered the crude odds ratio. With adjustment, daily folic acid was still associated with decreased odds of a neural tube defect. As with any study that relies on subjects reporting prior events, there is the possibility that there was differential reporting of folic acid use. The investigators addressed differential reporting by limiting their analysis to women who had not heard of the hypothesis that folic acid reduced birth defects. In this subgroup, preconceptual use of folic acid was still associated with significantly reduced neural tube defects (multivariable OR = 0.4, 95% CI = 0.2–0.6).

Of course, the chance that there were other important differences between the cases and controls on recall or other issues remains. Nonetheless, when prospective studies were performed, folic acid was found to reduce neural tube defects.\textsuperscript{30} This case-control study helped to spark further research and ultimately contributed to the recommendation that all women who may become pregnant in the near future take folic acid.


\textsuperscript{29} If you look at the original article the authors report relative risk but state that they estimated it from the odds ratio. I couldn’t reproduce their crude odds ratio exactly so the one I report (calculated from the raw data with www.openepi.com) is slightly different from what is in the reported article. It is true that when the exposure (in this case taking folic acid) is uncommon (<15%), the odds ratio approximates the relative risk.

7 Methods for adjusting for baseline differences between treatment groups

7.1 How do I adjust for baseline differences between persons who receive the intervention and those who don’t?

In Section 5.3, I illustrated how baseline differences between treatment groups could confound the results of a study. The specific example was influenza vaccination among the elderly. Because physicians tended to immunize the sickest patients, it appeared that vaccination was associated with higher rates of hospitalization for respiratory conditions. However, once the investigators used multivariable analysis to adjust for confounders, elders who were vaccinated were less likely to have been hospitalized due to respiratory conditions than elders who were not vaccinated.

Although multivariable analysis is a powerful tool for adjusting for baseline differences, at times, especially with nonrandomized studies, it is not sufficient. In such situations, two advanced methods are increasingly being used for achieving comparable groups: propensity score analysis and instrumental variable analysis. Because these methods require multivariable analysis, I explain their advantages and disadvantages compared to performing multivariable analysis alone, as well as compared to each other (Table 7.1).

7.2 What are propensity scores? How are they calculated?

Propensity scores are a statistical method for adjusting for baseline differences between study groups.

The use of propensity scores is a two-part process: (1) the development of the score and (2) the use of the score for adjusting for baseline differences.

Propensity scores are calculations of the probability of a subject being in a particular group, conditional upon that subject’s values on variables that influence group membership. To calculate the propensity score first identify those variables that influence study group membership (e.g., demographic characteristics, disease severity).
Evaluating Clinical and Public Health Interventions

The variables that influence group membership are then entered into a logistic model (assuming two groups) estimating the likelihood of being in a particular group. A logistic model will yield a propensity score for each subject ranging from 0 to 1; the score is the estimated probability of being in one group versus the other, conditional on a weighted score of that subject’s values on the set of variables used to create the score.

To understand the strength of propensity scores, consider two subjects with identical propensity scores, one who received the intervention and one who did not. If we assume that the propensity score is based on all those factors that

<table>
<thead>
<tr>
<th>Multivariable analysis alone</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Backbone of all methods for adjusting for baseline differences. It is less frightening to non-statistically inclined readers than propensity-score and instrumental variable analyses.</td>
<td>Does not work well if there are many baseline differences relative to the number of outcomes in the study. Adjusting for a large number of variables may result in a misspecified model and may make it more difficult to perform diagnostic checks on the model. If the distribution of confounders does not sufficiently overlap between the groups, effect estimates are based on extrapolation and may be unreliable. Cannot adjust for unmeasured confounders.</td>
</tr>
</tbody>
</table>

| Propensity-score analysis | Can be used in situations where there are many baseline differences and relatively few outcomes, as long as the sample is relatively evenly split among the study groups. Can make unlimited attempts at balancing covariates without worrying about biasing your analysis. Makes no assumption about the relationship of the baseline characteristic and the outcome. | If group membership is not evenly split, there may not be a large enough sample size to develop valid propensity scores. There must be sufficient overlap between the groups for the technique to work. If you use propensity scores to match, your matched samples may no longer be representative of the underlying population. Cannot adjust for unmeasured confounders. |

| Instrumental variable analysis | Only method that can account for unmeasured confounders. Good for policy-level analyses. | Not always possible to identify an instrumental variable. Decreases precision of estimate of effect. Results are generalizable only to patients for whom the instrument determined whether they received the intervention (the “marginal” patient). Not appropriate for estimating the benefit of an intervention for an individual. |
Adjusting for baseline differences between groups

would affect the likelihood of receiving an intervention, then we can consider
the assignment of the cases to be essentially random.¹

Of course, even if your propensity score is based on all of the factors known
to influence group assignment, there remains the possibility that there are
unknown confounders. If you could include these unknown confounders, they
would change the propensity score. Therefore, two subjects with the same pro-
pensity score may not have the same likelihood of being in either group.²

Gum and colleagues used propensity scores to assess whether aspirin use
decreased mortality among patients suspected of coronary artery disease.³ Subjects were drawn from consecutive adult patients undergoing stress echocardiogram at a single center. Of 6174 patients meeting the eligibility criteria of the study, 2310 (37%) were taking aspirin and 3864 (63%) were not.

As you can see in Table 7.2 there were numerous baseline differences
between patients taking aspirin and those not taking aspirin. In particular,
those patients taking aspirin are older and sicker (more likely to be diabetic,
hypertensive, have prior coronary artery disease, etc.) and would therefore be
expected to have higher mortality. Therefore a comparison of the two groups
would not be valid unless the analysis was adjusted to make the groups more
comparable.

To create comparable groups, the investigators created propensity scores for
each subject using logistic regression. The outcome for the model (dependent
variable) was whether or not the patient received aspirin. The model included
34 baseline characteristics (independent variables), some of which are included
in Table 7.2. This resulted in a score for each patient, ranging from 0.03 to 0.98,
representing the probability that a patient would be using aspirin.

Note that the propensity score is calculated without regard to the outcome
(in the case of this study the outcome was mortality). This leads to an important
advantage of propensity scores. In comparison to the traditional approach of
adjusting for baseline differences by including them in a multivariable model,
the propensity score can be calculated without consideration of the outcome
variable. This means that it is possible to experiment with inclusion of different
combinations of variables in the propensity score without risking biasing your
model by choosing variables based on how they affect your estimate of outcome.

Once the propensity score is created, it can be used to adjust for baseline differences in four different ways: matching, stratification, as a covariate, and as a method for weighting observations. These different methods are discussed in Sections 7.5.A–E. However, first let’s answer two important questions with regard to how to calculate propensity scores: what variables to include in a propensity score and how to assess the adequacy of propensity scores.

### Table 7.2. Baseline characteristics of aspirin and non-aspirin users.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Aspirin (n = 2310)</th>
<th>No aspirin (n = 3864)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD), yr</td>
<td>62 (11)</td>
<td>56 (12)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Men, no. (%)</td>
<td>1779 (77)</td>
<td>2167 (56)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Clinical history</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes, no. (%)</td>
<td>388 (17)</td>
<td>432 (11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hypertension, no. (%)</td>
<td>1224 (53)</td>
<td>1569 (41)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Tobacco use, no. (%)</td>
<td>234 (10)</td>
<td>500 (13)</td>
<td>0.001</td>
</tr>
<tr>
<td>Prior coronary artery disease, no. (%)</td>
<td>1609 (70)</td>
<td>778 (20)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Prior coronary artery bypass graft, no. (%)</td>
<td>689 (30)</td>
<td>240 (6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Prior percutaneous coronary interv., no. (%)</td>
<td>667 (29)</td>
<td>148 (4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Prior Q-wave MI, no. (%)</td>
<td>369 (16)</td>
<td>285 (7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Atrial fibrillation, no. (%)</td>
<td>27 (1)</td>
<td>55 (1)</td>
<td>0.04</td>
</tr>
<tr>
<td>Congestive heart failure, no. (%)</td>
<td>127 (6)</td>
<td>178 (5)</td>
<td>0.12</td>
</tr>
<tr>
<td>Medication use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digoxin use, no. (%)</td>
<td>171 (7)</td>
<td>216 (6)</td>
<td>0.004</td>
</tr>
<tr>
<td>β-Blocker use, no. (%)</td>
<td>811 (35)</td>
<td>550 (14)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diltiazem/verapamil use, no. (%)</td>
<td>452 (20)</td>
<td>405 (10)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Nifedipine use, no. (%)</td>
<td>261 (11)</td>
<td>283 (7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lipid-lowering therapy, no. (%)</td>
<td>775 (34)</td>
<td>380 (10)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ACE inhibitor use, no. (%)</td>
<td>349 (15)</td>
<td>441 (11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cardiovascular assessment and exercise capacity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index, mean (SD), kg/m²</td>
<td>29 (5)</td>
<td>30 (7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ejection fraction, mean (SD), %</td>
<td>50 (9)</td>
<td>53 (7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Resting heart rate, mean (SD), beats/min</td>
<td>74 (13)</td>
<td>79 (14)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Systolic blood pressure, mean (SD), mmHg</td>
<td>141 (21)</td>
<td>138 (20)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diastolic blood pressure, mean (SD), mmHg</td>
<td>85 (11)</td>
<td>86 (11)</td>
<td>0.04</td>
</tr>
<tr>
<td>Purpose of test to evaluate chest pain, no. (%)</td>
<td>300 (13)</td>
<td>468 (12)</td>
<td>0.31</td>
</tr>
<tr>
<td>Mayo Risk Index &gt;1, no. (%)</td>
<td>2021 (87)</td>
<td>2517 (65)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Peak exercise capacity, mean (SD), METs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>8.6 (2.4)</td>
<td>9.1 (2.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Women</td>
<td>6.6 (2.0)</td>
<td>7.3 (2.1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart rate recovery, mean (SD), beats/min</td>
<td>28 (11)</td>
<td>30 (12)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ischemic ECG charges with stress, no. (%)</td>
<td>430 (24)</td>
<td>457 (14)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ejection fraction &lt; 40%, no. (%)</td>
<td>321 (14)</td>
<td>226 (6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stress-induced ischemia on echo, no. (%)</td>
<td>495 (21)</td>
<td>436 (11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fair or poor physical fitness, no. (%)</td>
<td>714 (31)</td>
<td>1248 (38)</td>
<td>0.26</td>
</tr>
</tbody>
</table>

7.3 Which variables should be included in a propensity score?

There is some controversy over which variables to include in creating a propensity score. Should you include: (1) all those variables that are associated with treatment assignment; (2) all those variables that are associated with the outcome; (3) only those variables that are associated with treatment and outcome (true confounders)?

Based on the discussion of confounding (Section 5.3) it might seem sufficient to include only those variables that are associated with both the treatment and the outcome (choice #3 above). After all, if a variable is not a confounder then its inclusion in a propensity score is not likely to improve the adjustment for differences between the two groups. Also by limiting the variables in the score to the confounders (1) the efficiency of the model creating the score will increase and (2) it will be easier to match by propensity score (Section 7.5.A).

But despite these advantages, I recommend that you include all those variables that are associated with treatment assignment (option #1). The reason is that limiting the variables to those associated with outcome (option #2) or to true confounders (option #3) requires assessing the association between the variables and the outcome. This undercuts one of the advantages of propensity scores: that they can be calculated without regard to the outcome.

As long as you do not consider how the variables that compose the propensity score (or the score itself) are related to outcome you can calculate the propensity score as many times as necessary without biasing your results. For example, a standard test for how well a propensity score works is to see whether cases and controls with similar propensity scores are similar on important covariates.

If you discover that there are differences between cases and controls with similar propensity scores on important covariates, you can recalculate the propensity score, perhaps including other variables, or a transformation (e.g., logarithmic transformation) of an existing variable, or an interaction between two variables to achieve better balance. Since these calculations are performed without consideration of the outcome, there is no risk of biasing your results by choosing the combination of variables for the propensity score that maximizes the difference between the groups on outcome.

Also, by including all those variables associated with treatment assignment, you do not risk missing a confounder. Missing a confounder is a much more

---


significant problem than including a variable that is associated with group membership but is not a confounder.  

7.4 How do you assess the adequacy of a propensity score?  

Once you have calculated a propensity score using those baseline variables that differ between the groups, there are three important checks on the adequacy of the score:

1. ability of propensity score to differentiate those who received the intervention from those who did not (Section 7.4.A);
2. whether there is sufficient overlap between the two groups on the propensity score (Section 7.4.B); and
3. whether the propensity score creates groups comparable on baseline characteristics (Section 7.4.C).

7.4.A Does the propensity score differentiate those who received the intervention from those who did not?  

To assess how well the propensity score differentiates those who receive the intervention from those who do not use the c index. A value of 0.5 indicates that the model does not distinguish any better than chance. The maximum value is one. The higher the c index the better the model is at distinguishing the two groups. The c index is equal to the area under the receiver operating characteristics curve (ROC). A ROC curve can be constructed by plotting the sensitivity of the propensity score (in predicting who received the intervention) on the y-axis and (1−specificity) on the x-axis.

7.4.B Is there sufficient overlap between the two groups on the propensity score?  

Although we want the propensity score to differentiate the two groups, for it to successfully adjust for baseline differences between the groups there also must be sufficient overlap of the propensity scores. One good way to assess this is to draw a box plot. For example, Landrum and Ayanian evaluated whether

---

6 In some circumstances Austin et al. (footnote 4 in this chapter) found that reduction in bias was marginally greater when only true confounders were included in the propensity score than when including all variables associated with treatment selection. However, in other circumstances there was no difference. Overall, I think the dangers of biasing your propensity score by choosing variables based on their relationship with the outcome outweigh the potential advantages of including only true confounders.

Adjusting for baseline differences between groups

specialty care decreased mortality among patients who suffered a myocardial infarction. They calculated a propensity score, representing the probability of receiving care by a cardiologist, based on 41 clinical and provider characteristics.

As you can see in Figure 7.1, the propensity scores of patients who only received care by generalists were lower than the scores of those who received care by cardiologists – just as you would expect. But the scores overlapped. Also of interest, there were a larger number of outliers – patients who had a low predicted likelihood of seeing a cardiologist who nonetheless received cardiac specialty care. The outliers are shown with the straight lines outside the range of values demarcated with the brackets. This raises the possibility that other factors not included in the propensity score affected whether or not patients received care from a cardiac specialist. In such situations, try to recalculate the propensity scores by including additional variables, transformations of existing variables, or interactions between variables to achieve fewer outliers.

Figure 7.1


TIP
Assess overlap of propensity scores by using a box plot.

TIP
If comparison of groups by propensity scores shows many outliers recalculate the scores by including additional variables, transformations of existing variables, or interactions between variables.

---

7.4.C Does the propensity score create groups that are comparable on baseline characteristics?

In addition to the treatment groups having sufficient overlap on the propensity scores, the groups should also be balanced on important covariates within narrow strata of propensity scores. This can be assessed by stratifying subjects into at least five strata and comparing the values of subjects on important covariates by intervention group.

For example, in Section 4.3 I cited an observational study by Connors and colleagues assessing the impact of right heart catheterization (RHC) on mortality. The investigators found in bivariate results that patients who received an RHC had shorter survival compared to patients who did not receive an RHC (30-day survival was 62% versus 69%, respectively). However, patients who received right heart catheterization were also sicker at baseline so an unadjusted comparison between the two groups was not valid.

Therefore the authors calculated a propensity score from 65 baseline characteristics. To see whether the propensity score adjusted for important baseline differences between subjects receiving and not receiving an RHC, they divided the two groups into five strata (quintiles). As you can see in Figure 7.2, patients who received an RHC were similar on APACHE III scores (a composite measure of illness severity), blood pressure, and partial pressure of arterial oxygen/fraction of inspired oxygen. The stratified comparison between the groups indicates the success of the propensity score in balancing baseline differences.

If the groups differ by propensity score on important covariates then you have two options. First, try to recalculate the propensity score by adding variables, transforming variables, or including interactions between variables, to achieve a propensity score on which the strata are more similar. Alternatively, include the covariates that are not sufficiently similar across treatment groups in the multivariable model in addition to the propensity score.

7.5 How are propensity scores used to adjust for baseline differences?

Once a propensity score is created, there are four common ways to use it to adjust for baseline differences in the groups (Table 7.3).

These different methods are discussed below in Sections 7.5.A to 7.5.E.

7.5.A Matching cases using propensity scores

In the case of the study evaluating the effect of aspirin use on mortality (discussed above), the investigators used the propensity score as a matching
Adjusting for baseline differences between groups

variable. Each aspirin user was matched to the non-aspirin user with the most similar propensity score. If no non-aspirin user could be identified whose propensity score was within 0.09 of the aspirin user’s score, the aspirin user was excluded. The preset maximum difference that can exist between a case and a control for a match to be made is referred to as a caliper.

DEFINITION
Caliper is the maximum distance that can exist on a score between a case and a control and still consider them a match.

Figure 7.2 Patients receiving a right heart catheterization (RHC) and those who did not were stratified into five quintiles by propensity scores on APACHE III scores, blood pressure, and partial pressure of arterial oxygenation/fraction of inspired oxygen. Reproduced with permission from Connors, A.F., et al. “The effectiveness of right heart catheterization in the initial care of critically ill patients.” *JAMA* **276** (1996): 889–97. Copyright © 1996 American Medical Association. All rights reserved.
Comparing Table 7.4 to Table 7.2 you can see the success of matching by propensity score in creating comparable groups. Note that although there still are some significant differences between the two groups (e.g., men in the aspirin group had higher peak exercise capacity than men in the non-aspirin group) the groups are much more comparable than they were.

Matching subjects by propensity score dramatically changed the results of the study. In bivariate analysis, survival for persons who used aspirin during 3.1 years of follow-up was identical to the survival of persons who did not use aspirin (4.5% vs. 4.5%). With propensity-based matching, aspirin use was associated with a lower risk of death (4% vs. 8%, \( P = 0.002 \)). This result is consistent with the results of randomized studies indicating that aspirin decreases cardiovascular morbidity. But, importantly, this study extends that finding to all-cause mortality and to patients seen in clinical practice (not just to those who would enter a clinical trial).

The method that Gum and colleagues used to match the cases (aspirin users) to the controls (non-aspirin users) – identifying for each case the control that had the closest propensity score – is the simplest way of matching cases and controls using propensity scores. However, in certain circumstances, it is critical to match the cases and controls not only on the propensity score but also on certain covariates. For example, it might be important to match cases and controls on sex.

For example, Stukel and colleagues assessed whether cardiac catheterization improved the survival of elderly patients following an acute myocardial
Adjusting for baseline differences between groups


The data were from a nonrandomized cohort of elderly persons. Following an analytic strategy similar to that of Gum and colleagues, they calculated a propensity score and then matched cases to controls whose
propensity scores differed by less than 0.10 (caliper of 0.10). The one difference is that Stukel and colleagues required that cases and controls also were within 5 years of age of one another.

The major advantage of matching on propensity scores is that compared to other propensity score methods it generally provides the least biased estimates of treatment effect.\(^{10}\)

In the statistical world, advantages usually come with disadvantages. A disadvantage of matching by propensity score is that the matched sample may be smaller and not representative of the study population, thereby decreasing generalizability of results. Comparison of Table 7.2 and Table 7.4 demonstrates that the match resulted in keeping only 58% of the aspirin users (1351/2310) and 35% of the non-aspirin users (1351/3864). The implication is that, based on this study, we can only say that aspirin benefits people similar to those aspirin users whose propensity scores could be matched.

### 7.5.B Stratifying analysis by propensity score

Since it is not always possible to find (near) exact matches on propensity score between cases and controls, you may instead stratify cases and controls by propensity scores. To use this method, create strata such that each stratum contains subjects with a similar propensity score.

Assuming that the subjects in each stratum have similar propensity scores and that all confounders were included in the calculation of the score then the subjects can be considered to have been randomly assigned to the study groups within each stratum. This is just a broadening of the concept that two subjects from different groups matched by propensity scores can be considered to have been randomly assigned to the group.

Once subjects are sorted into strata, the effectiveness of the intervention can be estimated for each stratum of patients. An overall measure of effectiveness can be estimated by averaging the within-stratum effects.

For example, Wang and colleagues compared mortality of elderly persons treated with conventional versus atypical antipsychotic medications.\(^{11}\) As is true of the other nonrandomized studies in this section, there were sizable differences between the two groups; patients starting on conventional drugs

---


Adjusting for baseline differences between groups

were more likely to be male and non-white, and have congestive heart failure, ischemic heart disease other than myocardial infarction, and cancer than patients who were started on atypical antipsychotic medications.

To make the groups more comparable, the investigators calculated a propensity score for each patient and then used the propensity score to stratify their sample into deciles. In a proportional hazards model stratified by the deciles of the propensity score, mortality was greater at 180 days among users of conventional antipsychotics (hazard ratio = 1.37; 95% confidence interval 1.27–1.49) than among users of atypical antipsychotics.

A weakness of stratification is that you may have residual confounding. In other words, within each stratum of propensity score there may be important differences on baseline characteristics. To assess residual confounding, you can compare cases and controls on important baseline characteristics by strata.

In terms of how many strata to create, five strata have been shown to eliminate about 90% of the bias in unadjusted analyses. An analysis comparing five strata to three or 10 equal-sized strata found similar results.

7.5.C Using propensity score as an independent variable in a multivariable model

Rather than using the propensity score to match cases or to create strata the score can be entered as an independent variable in a multivariable model. In other words, the propensity score would be an independent variable for which the value for each subject would be that subject’s propensity score.

For example, Levy and colleagues assessed whether management by critical care physicians (intensivists) improved the mortality of intensive care unit (ICU) patients. In unadjusted analysis, patients managed by intensivists had higher hospital mortality than patients not managed by intensivists (OR = 2.13; 95% CI 2.03–2.24; \( P < 0.001 \)).

Intensivists reading this result will immediately protest that patients managed by intensivists tend to be sicker than those patients not managed by intensivists. Indeed, in this study, patients managed by intensivists were more likely to require ventilation in the first 24 hours of their ICU stay and were more likely to require continuous sedation than those not managed by intensivists.


Also, ICUs that managed almost all their patients with intensivists were different (e.g., more likely to have an academic affiliation, sited in larger hospitals) than ICUs that managed few patients with intensivists. Often, academic hospitals receive the most challenging referrals; perhaps this accounts for why patients managed by intensivists have higher mortality. To adjust for these differences, the investigators calculated a propensity score based on 19 variables.

When the investigators entered the propensity score into a random-effects logistic model along with a scale measuring severity of illness, patients managed by intensivists still had higher hospital mortality (OR = 1.40; 95% CI 1.32–1.49; \( P < 0.001 \)). So either there was residual confounding due to measured or unmeasured factors or there may be something about how intensivists treat patients (e.g., more procedures) that results in patients being more likely to die.

You may wonder why Levy and colleagues entered an adjustment for severity of illness in addition to the propensity score given that a number of variables related to severity of illness were included in the propensity score. The answer is that there still may be residual confounding and or misspecification even with the propensity variable in the model.

### 7.5.D Using the propensity score to weight observations

Another way to adjust for baseline differences is to use the propensity score to weight the multivariable analysis. For example, McNiel and colleagues used propensity-weighted multivariable analysis to assess the effectiveness of a mental health court.\(^\text{15}\) As with the other nonrandomized studies discussed in this section, the challenge in assessing the effectiveness of the court was that those criminal defenders participating in the mental health court were different than those who received usual treatment. For example, those participating in the mental health court were more likely to be homeless and to have a severe mental disorder than those who received usual treatment.

To account for the differences between those participating in the court and those not participating, the investigators calculated a propensity score based on demographic characteristics, clinical variables, and criminal charges during the prior 12 months. The propensity score was then used to weight a proportional hazards model estimating the likelihood of recidivism; specifically (1) time to a new violent charge and (2) time to any new charge. The weight for a person receiving the intervention is the inverse of the propensity score; the weight for the control is the inverse of one minus the propensity score.

Adjusting for baseline differences between groups

The investigators found that participation in the mental health court resulted in longer times without any charges for violent crimes or any new criminal charges.

Some analysts have found that using propensity scores to weight observations is less subject to misspecification of the regression models used to predict average causal effect than using propensity-based stratification. A problem with weighting observations is that when the estimated propensity score is near zero or one, the weights may be extreme and unrealistic.

7.5.E Choosing among the different methods of using propensity scores

It is important to understand the advantages and disadvantages of the different approaches to using propensity scores. However, in practice, many people will use more than one method. Because matching produces the least biased estimates of treatment effect, there is a strong argument for using this method first. To overcome the limitation of decreased generalizability due to loss of subjects, researchers should also use the propensity scored to perform stratified analysis or include the propensity score as a covariate in a multivariable model.

In general, you should get similar results with different approaches. If you do, this will strengthen your findings. If the results differ, you will have learned something important about your data – perhaps that those cases for which matches are available are different from those cases that do not have appropriate matches.

7.6 What is instrumental variable analysis? How is it used to adjust for baseline differences between groups?

Instrumental variable analysis is a technique commonly used by economists to study the impact of social or economic policies (e.g., employment programs) on economic trends. Its utility is that it can potentially mimic randomization by adjusting for unknown or unmeasured confounders. In contrast, multivariable analysis and propensity-score analysis can only adjust for measured confounders.

The reason instrumental variable analysis has been used heavily in the economic arena is that economic policies are rarely randomized and economists are typically assessing the impact of these programs on the population. Recently, health researchers have noted the similarities between economic policy and health policy analysis, and shown increased interest in performing instrumental analysis to assess the impact of nonrandomized interventions. For example, instrumental variable analysis can be used to assess the impact of health insurance (an intervention that is rarely randomized) on health outcomes.

**DEFINITION**

To perform an instrumental variable analysis you must identify a variable that is associated with the receipt of an intervention, but does not have any effect on the outcome except through its effect on whether or not the intervention is received, as illustrated in Figure 7.3.

To illustrate how an instrumental variable can simulate randomization, let’s look at a study of the effectiveness of mammography for women over the age of 70. You might think that a question of this importance would have been addressed with a randomized study. Unfortunately the randomized controlled trials on this topic have generally excluded women over the age of 70 years.

---

19 Sections 7.6–7.10 on instrumental analysis come with a warning: it is a non-intuitive technique. I have done my best to keep the explanation simple. Although it is important for certain non-randomized designs most investigators have never used it, and perhaps you won’t have cause to use it either. Try to read all the sections through rather than getting stuck on the details. For two general references on instrumental analysis see: Hernan, M. A. and Robins, J. M. “Instruments for causal inference: An epidemiologist’s dream?” Epidemiology 17 (2006): 360–72; Greenland, S. “An introduction to instrumental variables for epidemiologists.” Int. J. Epidemiol. 29 (2000): 722–9.

Adjusting for baseline differences between groups

Table 7.5. Does mammography result in earlier diagnosis of breast cancer?

<table>
<thead>
<tr>
<th></th>
<th>No mammogram</th>
<th>Mammogram</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early cancer</td>
<td>59%</td>
<td>81%</td>
<td></td>
</tr>
<tr>
<td>Late cancer</td>
<td>41%</td>
<td>19%</td>
<td>( P = 0.001 )</td>
</tr>
</tbody>
</table>


Table 7.6. Baseline differences between women with cancer by whether or not they received a mammogram.

<table>
<thead>
<tr>
<th></th>
<th>No mammogram</th>
<th>Mammogram</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis</td>
<td>77</td>
<td>75</td>
<td>0.001</td>
</tr>
<tr>
<td>Charlson comorbidities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not hospitalized</td>
<td>27%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Hosp, no comorbidities</td>
<td>48%</td>
<td>54%</td>
<td></td>
</tr>
<tr>
<td>At least one comorbidity</td>
<td>25%</td>
<td>16%</td>
<td>0.001</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>6%</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Non-black</td>
<td>94.8%</td>
<td>97%</td>
<td>0.001</td>
</tr>
<tr>
<td>Income (median of zip code)</td>
<td>$42 030</td>
<td>$41 137</td>
<td>0.061</td>
</tr>
<tr>
<td>Primary care visits</td>
<td>5</td>
<td>11</td>
<td>0.001</td>
</tr>
</tbody>
</table>


Therefore the investigators attempted to answer this question using an observational design. They matched Medicare records to a tumor registry to identify women who were diagnosed with breast cancer. They found that women who had mammography were more likely to present at an early stage of illness (Table 7.5). Does this prove that mammography is effective at detecting cancer earlier among older women?

If these data came from a study in which women had been randomized to mammography or not, it would be strong evidence that mammography resulted in earlier diagnosis of breast cancer. But this is an observational study. As you can see, in Table 7.6, there were substantial baseline differences between women who did and did not receive a mammogram.

To address these baseline differences, the investigators conducted an instrumental variable analysis. The instrumental variable was the region where the women lived. Women from Seattle were much more likely to have a mammogram (64%) than women who lived in Connecticut (50%) or in Atlanta (45%). Although it is known that there are substantial geographic differences in use of screening tests
there is no reason to believe that women are more likely to present at an early stage of cancer in one area than another due to factors other than screening test use.

Using region as an instrumental variable, the investigators found that receipt of mammography was associated with increased likelihood of being diagnosed with an early stage cancer. The results of the instrumental variable analyses are shown in Table 7.7 in the right column and the results of a simpler multivariable model are shown in the left.

As the odds ratio for the impact of mammography on early diagnosis is similar in the simple multivariable model (2.97) and the instrumental model (3.01), you may wonder what value the instrumental variable adds. The answer is that by using region to simulate randomization, we are simulating adjustment for unmeasured characteristics. The standard multivariable model only adjusts for those measured characteristics you see in Table 7.7. The fact that the point estimates are similar for the two analyses tells us that the relationship between mammography and early diagnosis does not appear to be biased by unmeasured confounders.

### Table 7.7. Estimation of the impact of mammography use on early diagnosis of breast cancer.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Multivariable model</th>
<th>Instrumental variable analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>67–69 years</td>
<td>1.00 (reference)</td>
<td>1.00 (reference)</td>
</tr>
<tr>
<td>70–74 years</td>
<td>0.92 (0.75, 1.13)</td>
<td>0.92 (0.75, 1.13)</td>
</tr>
<tr>
<td>75–79 years</td>
<td>0.93 (0.75, 1.15)</td>
<td>0.93 (0.75, 1.16)</td>
</tr>
<tr>
<td>80–84 years</td>
<td>0.83 (0.66, 1.05)</td>
<td>0.82 (0.63, 1.07)</td>
</tr>
<tr>
<td>85+ years</td>
<td>1.02 (0.79, 1.32)</td>
<td>1.02 (0.70, 1.49)</td>
</tr>
<tr>
<td>Black</td>
<td>0.67 (0.50, 0.92)</td>
<td>0.70 (0.51, 0.98)</td>
</tr>
<tr>
<td>Comorbidities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not hospitalized</td>
<td>1.00 (reference)</td>
<td>1.00 (reference)</td>
</tr>
<tr>
<td>Hospital, no comorbidity</td>
<td>0.54 (0.45, 0.63)</td>
<td>0.55 (0.47, 0.65)</td>
</tr>
<tr>
<td>Hospital, comorbidity</td>
<td>0.48 (0.40, 0.59)</td>
<td>0.50 (0.40, 0.62)</td>
</tr>
<tr>
<td>High income</td>
<td>1.23 (1.08, 1.41)</td>
<td>1.22 (1.07, 1.41)</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connecticut</td>
<td>1.00 (reference)</td>
<td>–</td>
</tr>
<tr>
<td>Seattle</td>
<td>1.23 (1.06, 1.43)</td>
<td>–</td>
</tr>
<tr>
<td>Atlanta</td>
<td>1.17 (0.97, 1.41)</td>
<td>–</td>
</tr>
<tr>
<td>Primary care visits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1.00 (reference)</td>
<td>1.00 (reference)</td>
</tr>
<tr>
<td>1–3</td>
<td>0.77 (0.62, 0.94)</td>
<td>0.76 (0.55, 1.04)</td>
</tr>
<tr>
<td>4–12</td>
<td>0.97 (0.80, 1.17)</td>
<td>0.95 (0.58, 1.58)</td>
</tr>
<tr>
<td>13+</td>
<td>0.79 (0.64, 0.98)</td>
<td>0.80 (0.45, 1.45)</td>
</tr>
<tr>
<td>Mammogram performed</td>
<td>2.97 (2.56, 3.45)</td>
<td>3.01 (1.09, 8.34)</td>
</tr>
</tbody>
</table>

In the next section I will review the underlying assumptions of instrumental variable analysis, and then move on to how to conduct an instrumental variable analysis.

### 7.7 What are the underlying assumptions of instrumental variable analysis?

To conduct an instrumental variable analysis you need to identify an instrumental variable that fulfills four assumptions. Table 7.8 lists the four assumptions along with how to assess each of them. Greater detail is provided in the next four subsections.

#### 7.7.A Is the instrumental variable unassociated with outcome except through its relationship with the intervention?

It is impossible to prove that an instrumental variable is associated with an outcome variable only through its relationship with the intervention. Instead we choose instrumental variables that we believe, based on theoretical considerations, would be unlikely to be directly associated with the outcome.

To give you ideas for possible instrumental variables I have listed instrumental variables that have been used successfully in other studies along with the treatment assignments that they were used to predict and the outcome variable of the study (Table 7.9). The first one on the list is from the study of the effectiveness of mammography for older women discussed in Section 7.6. Note that

---

**Table 7.8. Assumptions underlying instrumental analysis.**

<table>
<thead>
<tr>
<th>Assumption</th>
<th>How to assess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrumental variable is not associated with outcome except through its relationship with the intervention.</td>
<td>Cannot be tested directly and must be assumed based on theoretical considerations.</td>
</tr>
<tr>
<td>Instrumental variable is strongly associated with group assignment.</td>
<td>Perform multivariable analysis with receipt of intervention as dependent variable and instrumental variable along with other important baseline characteristics as the independent variables; repeat the analysis removing instrumental variable; compare the models.</td>
</tr>
<tr>
<td>Instrumental variable is not associated with baseline characteristics that are associated with outcome.</td>
<td>Assess the association between the instrumental variable and baseline characteristics known to be associated with outcome.</td>
</tr>
<tr>
<td>Instrumental variable has a monotonic relationship with intervention status: i.e., everyone who would get the intervention with a low score predicting intervention would also get the intervention with a high score predicting intervention.</td>
<td>Cannot be tested directly. Must be assumed based on theoretical considerations.</td>
</tr>
</tbody>
</table>
what the instrumental variables listed in Table 7.9 have in common (region, frequency of procedure use, density of specialists, physician preference, or state benefits) is that you wouldn’t expect that these variables would influence health outcomes in ways other than affecting whether certain interventions occurred.

Note that several of these instrumental variables are sufficiently broad that they could be used for a range of outcomes. For example, distance to a specialty hospital, density of specialists in a particular area, or physician preferences are potential instrumental variables for a number of health service outcomes because a patient’s location shouldn’t affect health outcomes. Similarly, regional economic conditions may work as instrumental variables for health services research because the use of many procedures differs by region and not necessarily because of the characteristics of the patients within the region.

Table 7.9. Examples of instrumental variables used to estimate relationship between intervention and outcome.

<table>
<thead>
<tr>
<th>Instrumental variable</th>
<th>Intervention</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region where patient lives(^a)</td>
<td>Mammogram</td>
<td>Stage of diagnosis</td>
</tr>
<tr>
<td>Rate of procedure (catheterization) in region where patient lives(^b)</td>
<td>Cardiac catheterization</td>
<td>Mortality</td>
</tr>
<tr>
<td>Density of specialists (cardiologists) where patient lives(^c)</td>
<td>Ambulatory cardiology care</td>
<td>Mortality</td>
</tr>
<tr>
<td>Physician preference(^d)</td>
<td>Use of aprotinin during coronary artery bypass grafting</td>
<td>Mortality</td>
</tr>
<tr>
<td>Differential distance between specialty hospital</td>
<td>Cardiac catheterization</td>
<td>Mortality</td>
</tr>
<tr>
<td>(capable of performing catheterization) and nearest hospital in relation to patient’s home(^e)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>State unionization rate, state unemployment rate, and state tax rate (three instrumental variables) where patient lives(^f)</td>
<td>Insurance</td>
<td>Health index</td>
</tr>
<tr>
<td>Generosity of Medicaid and AIDS Drug Assistance Program benefits in the state where the patient lives (six variables)(^g)</td>
<td>Insurance</td>
<td>Mortality</td>
</tr>
</tbody>
</table>


Although it is not possible to prove that an instrumental variable is associated with an outcome only through its relationship with the intervention, it is possible to test for violations of this assumption. Specifically, if an instrumental variable is associated with a factor that may influence treatment choices and affect the outcome, then the assumption is violated. For example, if a potential instrumental variable, community of residence, is associated with health status (perhaps certain communities attract a higher percentage of healthy people because of being near to athletic facilities), and health status influences whether a person receives an intervention as well as their mortality, then the community a patient lives in cannot be used as an instrumental variable.  

(This issue is closely related to the assumption discussed in Section 7.7.C).

7.7.B Is the instrumental variable strongly associated with group assignment?

The stronger the association between an instrumental variable and receipt of the intervention, the more likely it is to work. The reason is that the stronger the instrumental variable, the larger is the group for whom the instrumental variable determines the treatment assignment and the more precise the treatment estimate will be.

The strength of the association between the instrumental variable and the group assignment is assessed by first performing a multivariable analysis with treatment assignment as the dependent variable (e.g., mammography would be the assignment variable in the study discussed in Section 7.6) and the instrumental variable (e.g., region), as well as potential confounders (e.g., age, comorbidities) as independent variables. Next, repeat the analysis removing the instrument from the model. Finally compare the two models (the one that includes the instrumental variable and the one that excludes the instrumental variable). If the treatment assignment is dichotomous (e.g., mammogram received or not), you might perform logistic regression. If the treatment assignment is interval (e.g., number of visits), you might perform multiple linear regression.

As a rule of thumb, the chi-squared value divided by the degrees of freedom for the comparison of the logistic models with and without the instrumental variable should be at least 20. Alternatively, the $F$ value for a linear regression model with and without the instrumental variable should be at least 10.  

In the analysis of the impact of mammography on stage of breast cancer diagnosis, the chi-squared statistic was 53.0 with two degrees of freedom. This

---


exceeds our threshold of 20 for the quotient of the chi-square divided by the degrees of freedom.

7.7.C *Is the instrumental variable associated with baseline characteristics that are associated with the outcome?*

For an instrumental variable analysis to be valid there should be no association between the instrumental variable and baseline characteristics associated with the outcome. To understand why, remember that our goal with instrumental variable analysis is to simulate a randomized study.

One of the major advantages of randomization is that it tends to produce groups that are comparable with respect to baseline characteristics (Section 4.2). The same should be true of an instrumental variable.

To test the association between the instrumental variable and the baseline characteristics associated with outcome, perform a bivariate analysis. If your instrumental variable is dichotomous or categorical (e.g., region in the mammography study), you can simply compare across categories. If it is continuous, you will need to divide the sample into categories to determine whether the instrumental variable is associated with prognostically important baseline characteristics.

For example, in Section 7.5.A, I described a study by Stukel and colleagues that used propensity scores to evaluate the efficacy of cardiac catheterization in patients who had previously suffered a myocardial infarction (heart attack). They also used instrumental variable analysis to answer the same question. Their instrumental variable was regional catheterization rate. The rate of cardiac catheterization following a myocardial infarction varied widely (29% to 82.3%) across the United States. For regional catheterization rate to function as an instrumental variable, prognostically important variables should be similar across regions.

To test whether this were true of their data, the investigators divided their sample into five groups of approximately equal numbers of patients (Table 7.10). The proportion of patients in each group receiving catheterization (ranging from 42.8% in the first quintile to 65.0% in the fifth quintile) closely correlated with the regional catheterization rate, indicating that regional catheterization rate was strongly associated with whether an individual patient received catheterization. If you compare the patients in the five groups you can see that their baseline characteristics are similar (although not identical).

If the instrumental variable is associated with variables that are prognostically related to the outcome, you may still proceed with the analysis. However, you will need to adjust for these variables in the model. This is determining the influence of mammography on stage of diagnosis.” Health Serv. Outcomes Res. Methodol. 2 (2001): 279–90.
Adjusting for baseline differences between groups

The same approach that would be necessary if you performed a randomized controlled trial, and by chance there were important baseline differences.

The problem with this approach is that if you have strong differences in prognostic characteristics it raises the question of whether the instrumental variable is effectively simulating randomization. Continuing with the analogy to randomization, although you can by bad luck obtain unequal groups, if there were strong differences between the groups you would wonder whether there were a problem with your randomization.

You may find that an instrumental variable is associated with baseline differences but that these differences are not associated with the outcome. For example, if the instrumental variable was density of specialists and the outcome were mortality, the density of specialists may be associated with urbanicity of

<table>
<thead>
<tr>
<th>Quintile (range) of regional cardiac catheterization rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (29.2–48.1)</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td>No. of patients</td>
</tr>
<tr>
<td>Cardiac catheterization rate</td>
</tr>
<tr>
<td>Acute myocardial infarction severity</td>
</tr>
</tbody>
</table>

Demographics

<table>
<thead>
<tr>
<th>Age range, yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>65–74</td>
</tr>
<tr>
<td>75–84</td>
</tr>
</tbody>
</table>

Men

| Men | 53.7 | 54.2 | 55.0 | 55.6 | 56.4 |

Black

| Black | 4.1 | 8.1 | 6.3 | 5.5 | 5.4 |

Social Security income > $2600

| Social Security income > $2600 | 30.4 | 28.2 | 33.4 | 27.9 | 29.1 |

Comorbidities

<table>
<thead>
<tr>
<th>Comorbidities</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of angina</td>
<td>50.1</td>
<td>48.3</td>
<td>47.8</td>
<td>47.6</td>
<td>44.0</td>
</tr>
<tr>
<td>Previous myocardial infarction</td>
<td>30.1</td>
<td>29.8</td>
<td>29.2</td>
<td>28.7</td>
<td>26.9</td>
</tr>
<tr>
<td>Previous revascularization</td>
<td>16.5</td>
<td>18.6</td>
<td>20.8</td>
<td>20.2</td>
<td>22.1</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>18.4</td>
<td>18.0</td>
<td>17.3</td>
<td>16.9</td>
<td>15.1</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>32.9</td>
<td>32.5</td>
<td>32.3</td>
<td>31.3</td>
<td>30.0</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>10.5</td>
<td>10.9</td>
<td>11.0</td>
<td>10.4</td>
<td>10.0</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>21.1</td>
<td>20.2</td>
<td>20.3</td>
<td>20.3</td>
<td>20.7</td>
</tr>
<tr>
<td>Smoker</td>
<td>16.7</td>
<td>16.7</td>
<td>17.0</td>
<td>18.0</td>
<td>17.9</td>
</tr>
</tbody>
</table>

AMI clinical presentation characteristics

<table>
<thead>
<tr>
<th>AMI clinical presentation characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-ST-segment elevation AMI</td>
<td>40.4</td>
<td>41.2</td>
<td>40.5</td>
<td>39.3</td>
<td>39.0</td>
</tr>
<tr>
<td>Shock</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Hypotension</td>
<td>2.8</td>
<td>2.9</td>
<td>2.6</td>
<td>2.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Received CPR</td>
<td>1.6</td>
<td>1.7</td>
<td>1.7</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Peak creatinine kinase &gt;1000 U/L</td>
<td>30.3</td>
<td>30.5</td>
<td>30.4</td>
<td>31.7</td>
<td>32.6</td>
</tr>
</tbody>
</table>

the area. Presumably, urbanicity is not associated with mortality. Therefore, this should not pose a problem for an instrumental variable analysis.

7.7.D Instrumental variable has a monotonic relationship with intervention status

This assumption sounds much more complicated than it is. What it means is that persons who would get the intervention with a given score predicting intervention would also get the intervention with a higher score predicting intervention.\(^{23}\) In other words, having a higher score on the prediction would not work against the person obtaining the intervention. This assumption cannot be tested but has to be assumed on theoretical grounds to be true.

7.8 How is instrumental variable analysis performed?

The first job in conducting an instrumental variable analysis is to prove that the instrumental variable fits the two assumptions that can be empirically tested for: strength of association between the instrumental variable and the group assignment (Section 7.7.B) and lack of association between the instrumental variable and baseline characteristics associated with outcome (Section 7.7.C). For the two that cannot be proven, consider whether the instrument variable fits the assumption on theoretical grounds.

Having convinced yourself that the instrument fulfills the four assumptions, you can perform the analysis. As is true of propensity models, instrumental variable analysis involves two stages. In stage one the instrumental variable, along with any other predictors of the intervention, is used to predict the likelihood of receiving the intervention for each subject. In stage two, these predicted values replace the actual intervention assignment for each subject and are used to estimate the impact of the intervention on the outcome.

Look back at Table 7.7. Note that the confidence intervals for the estimates of the impact of mammogram use on breast cancer diagnosis were much wider for the model using the instrumental variable than for the multivariable model. Because instrumental variable analysis replaces the actual treatment assignment of each subject with the prediction of treatment assignment based on the first-stage model, the estimates tend to be less precise (have larger standard errors).\(^{24}\)

Note also that in the instrumental variable model the instrumental variable is not entered into the model. The reason is that the instruments are assumed


Adjusting for baseline differences between groups

to be unrelated to outcome except through their effect on outcome; therefore, it is not necessary to adjust the model for them.

7.9 What are the limitations of instrumental variable analysis?

Certainly, the biggest limitation to instrumental variable analysis is identifying instrumental variables that fit the assumptions listed in Table 7.8.

Beyond this, the results of an instrumental variable only apply to those patients whose treatment status is determined by their value on the instrumental variable. This is called the marginal population.

In the case of the study of mammography, the benefit is demonstrated only for the population of women whose receipt of mammography is influenced by living in an area where mammography is commonly provided. The results do not pertain to women who would never receive mammography or who would always receive mammography no matter where they lived. For these reasons, instrumental variable analysis cannot be used to predict the impact of interventions on specific individuals.

In the case of a study where you are predicting survival differences, the survival difference for the marginal population is referred to as the local average treatment effect (LATE). It is the average benefit gained by having the intervention among persons for whom the instrument determined that they received the intervention. For example, the study of the impact of cardiac catheterization on mortality (Section 7.5.A) found that cardiac catheterization was associated with a 16% decrease in mortality (adjusted RR 0.84; 95% CI 0.79–0.90). This is the average effect of cardiac catheterization for those persons for whom regional catheterization rate determined their receipt of catheterization.

Given this similarity, how well do instrumental variables simulate randomization? The answer depends on how sure you are that the instrumental variable does not have any effect on the outcome except through its effect on the intervention. When you randomize, assuming you have done so without bias, you can have great confidence that the randomization process is not associated with the outcome. In fact, the outcome is not known at the time of the randomization process. This is not the case with instrumental variable analysis, which is performed using retrospective data and the outcome is known.

Moreover, since an instrumental variable is not randomly assigned, you cannot take it on faith that there is no relationship between the instrumental

---


variable and the outcome other than through group assignment. Worse yet, you cannot assess whether the instrumental variable is unrelated to outcome except through its relationship to the outcome. Nonetheless, because instrumental variable analysis is the only method available for adjusting observational data for unmeasured characteristics, it is an important tool.

### 7.10 How do the results of multivariable adjustment alone, propensity scores, and instrumental variable analysis compare?

Several studies have compared the results of analyses using multivariable analysis alone, multivariable analysis with propensity scores, and instrumental variable analysis. For example, the study of the impact of mammography on stage of cancer diagnosis discussed in Section 7.6 used both propensity analysis and instrumental variable analysis; the two techniques produced similar results.

In contrast, the study of the impact of cardiac catheterization on mortality described in Section 7.7.C found that standard multivariable analysis and propensity-based matched analysis indicated a much stronger effect of cardiac catheterization than an instrumental variable analysis. The investigators interpreted this difference as being due to instrumental variable analysis being less biased. However, not all experts agree. 27

My own view is that there is too little experience to conclude that instrumental variable analysis is less biased than these other methods, although I think it is a promising technique for addressing problems that cannot be addressed with these other methods.

My general advice is when possible analyze your data using more than one method. When different methods produce similar answers you can have greater faith in the answer. When they produce different answers, focus on why they produced different answers rather than which method is better.

Also, remember that instrumental variable analysis, and to a lesser extent propensity analysis, are conceptually difficult procedures. Many readers will have difficulty understanding them. For this reason if your intervention can be evaluated adequately with multivariable analysis alone there may be no reason to perform a more complicated analysis.

### 7.11 What is sensitivity analysis?

Sensitivity analysis refers to a broad group of analyses in which the investigator varies certain inputs to see how such changes affect the results.

---

Adjusting for baseline differences between groups

In the context of this chapter, we are interested in a particular type of sensitivity analysis: how strong and how prevalent would an unmeasured confounder have to be to change the conclusion of the study? If, for example, we assume that there is an unmeasured confounder present among 20% of your subjects that is correlated with the intervention you are studying at 0.30, how strong would its relationship to outcome have to be to change your results? If the answer is not very strong, than your results may be biased by an unmeasured confounder because it is easy to miss weak confounders. If the answer is that the unmeasured confounder has to be very strong, then it is much less likely that such a confounder exists and yet investigators in the field are unaware of it.

For example, colleagues of mine and I at the San Francisco Department of Health conducted an observational study to evaluate the impact of housing on mortality of homeless persons with AIDS by matching a supportive housing registry to an AIDS registry. We identified 70 homeless persons who received supportive housing after an AIDS diagnosis and 606 homeless persons who were not housed by the program. To determine whether receipt of housing was associated with decreased mortality, we performed a proportional hazards model adjusting for age, ethnicity, CD4 count at time of diagnosis, and HIV risk group. We found that mortality was reduced by 80% for persons who received supportive housing (RH = 0.20; 95% confidence intervals 0.05 to 0.81).

Although we adjusted for a number of potential confounders, the question remained whether there could be some unmeasured confounder that explained why persons who received supportive housing were less likely to die. We had no good candidates for an instrumental variable analysis. Instead, we performed a sensitivity analysis. We found that an unmeasured confounder would need to exist at a prevalence level of 70%, be correlated at 0.5 with not receiving supportive housing, and predict at least a nine-fold increase in risk to eliminate the beneficial impact of supportive housing on mortality. Although not impossible, it is unlikely that such a strong and prevalent confounder exists and yet we (and other investigators in our field) are not aware of it. Other prevalence levels and correlations with receipt of supportive housing were even less plausible.


29 Our manuscript provides a formula in the appendix for calculating how different combinations of prevalence of risk factor and correlation with the risk factor would result in different strengths of a confounder necessary to eliminate the effect of an intervention on an outcome. Based on these scenarios, you can determine how plausible it would be for a confounder to explain your results.
8.1 How do I use time series analysis to analyze how an intervention affects a series of observations over time?

For some interventions, the best way of determining whether they are effective is to observe whether they cause a change in the pattern of a series of observations over time. For example, you may want to know whether an intervention affects the number of births, the number of deaths, the number of infections, or the number of accidents during a three-year period in a particular city. Methods for examining trends in a series of observations are called time series analyses. The specific method for determining whether an intervention affects a series of observations is called interrupted time series analysis.¹

Although the name “interrupted time series analysis” sounds complicated this method is just a fancy pre-intervention versus post-intervention design (Section 1.2.A) with a lot of pre-intervention and post-intervention points; the “interruption” is the intervention. The idea is that with a series of measurements over time you can detect changes in pattern. When a change coincides with the start of an intervention, it suggests that the intervention has had an effect.

Interrupted time series analyses are used most commonly to evaluate non-randomized interventions. Often the outcome data come from an administrative dataset (e.g., registries of birth, death, accidents). However, the method can be used for randomized designs as well.

To perform time series analyses the data must be in the form of events per time and the intervals must be equally spaced (e.g., births per day, accidents per week, infections per month). The data can be a series of observations of a population (e.g., births per month in a particular country) or a single person (e.g., number of times per day that a person experiences an episode of pain).

Time series analysis

Besides count data other kinds of data that can be used with time series are mean value (e.g., mean number of prescriptions dispensed per subject) or proportion (e.g., proportion of subjects who received HIV risk-reduction counseling), so long as the outcomes can be measured on a time scale (e.g., mean number of prescriptions dispensed per patient per month; proportion of subjects receiving counseling per year).

To increase the likelihood that the changes that appear to be temporally related to the intervention are actually related to the intervention, it is helpful to have a series of observations over the same period of time in a setting where the intervention did not occur. With a comparison group, an interrupted time series analysis is a fancy pre-intervention versus post-intervention with concurrent controls (Section 1.2.B).

With or without controls, it is desirable to test whether the changes that occur following an intervention are greater than would be expected by chance. The analytic tool for determining whether the change is greater than would be expected by chance is called segmental regression analysis.

To better understand the concepts of time series, interrupted time series, segmental regression analysis, and the use of control groups, let’s examine an evaluation of a change in HIV testing policy in San Francisco.\(^2\) Prior to May 2006 the San Francisco Public Health Department required written patient consent for HIV testing in its medical settings; beginning in May 2006 only verbal agreement was required. The new policy was instituted because we believed that written consent (as opposed to verbal agreement) was a barrier to persons receiving testing. The change in policy applied only to public health facilities in San Francisco and not to other providers in the county.

Figure 8.1 shows a graphic presentation of time series data. Each dot represents the number of HIV tests per patient visits at public health settings. The vertical line represents the interruption of the series in May 2006 when the new policy went into effect.

Visually inspecting the time series it would appear that in the period July 2004 to May 2006 the number of tests performed was gradually increasing, but that at the interruption – May 2006 when the new policy began – there was an increase in the number of tests performed, and an acceleration in the rate of increase.

This point is made in Figure 8.2, which is the same data as Figure 8.1, except I have added two segmental regression lines. The lines help you to see the sharp

\(^2\) Zetola, N. M., Grijalva, C. G., Gertler, S., et al. “Simplifying consent for HIV testing is associated with an increase in HIV testing and case detection in highest risk groups, San Francisco January 2003–2007.” *PloS One* 3 (2008): e2591. Some of the analyses reported here were not in the published article, but were supplied by the lead author.
increase in the level of testing following the change of policy in May 2006. The steeper slope of the top line compared to the bottom line illustrates how testing increased more rapidly following the policy change than it had prior to the change.
Note that each of the two segments seen on Figure 8.2 can be defined using two parameters: the level (where the segment starts) and the slope. To conduct an interrupted time series analysis we must also specify a transition period or lag time when no significant change is anticipated. The analysis of the new HIV testing policy assumed that the policy change took one month to become standard practice and therefore May of 2006 was considered the transition month and data from that month were excluded.

To statistically determine the impact of an intervention we use segmental regression analysis to see whether the level and the slope changed after the intervention, compared to the segment prior to the intervention. In the case of the data shown in Figure 8.2, we found that both the level and the slope increased significantly after the intervention.

In addition to determining whether the level and slope have significantly changed after the intervention, segmental regression analysis can also be used to adjust for other factors that may have changed during the study period. For example, segmental regression analyses can adjust for seasonal changes (as long as you have data from more than one year!) or changes in the population (e.g., more homeless persons seeking care).

In the case of the analysis of the HIV testing policy change, we adjusted for age, race, language, gender, homelessness status, insurance, and health care setting. With these terms in the model, we found that the average monthly rate of HIV tests increased by 4.38 tests over what would have been predicted by the baseline trend ($P < 0.001$).

An important aspect of time series data is that the error terms associated with each observation are not independent. Indeed, error terms of sequential observations are likely to be correlated. Special methods are needed to incorporate this dependence, otherwise the standard errors will be underestimated and the statistical significance will be overestimated. The most commonly used model for analyzing time series data with correlated errors is autoregressive integrative moving average (ARIMA), which uses prior values to estimate later values. By correctly modeling the errors (or noise) we can determine the effectiveness of the intervention.$^4$

---


Despite the ability of segmental regression analysis to adjust for changes in the population over time and auto-correlation, the results of the HIV testing analysis still leave an important question. Could there be some other factor that occurred at the same time as the new HIV testing policy that truly accounted for the increase in testing rates other than the variables the investigators adjusted for? For example, could there be some other factor that was increasing HIV testing at all hospitals in San Francisco or could there be some other factor that was increasing all kinds of tests in the San Francisco public health system. To explore these possibilities the investigators used two controls: HIV testing at a university-based medical center that did not change its HIV testing policy (Figure 8.3) and hematocrit testing within the public health system (Figure 8.4). Segmental regression analysis showed no significant change following May 2006 in the rates of testing for these controls.

Although segmental regression analysis of interrupted time series data is a powerful tool, if you are using it to analyze data from a nonrandomized study, all the same limitations exist as with any analysis of observational data: you cannot exclude confounding due to some unmeasured factor. In the case of this study, we cannot completely exclude the possibility that the increase in HIV testing rates seen in the San Francisco public facilities but not in the university hospital was due to some other factor.
Figure 8.4  Hematocrit tests per month at San Francisco public health facilities.

Nonetheless, this study had a major impact. The data were cited as supporting evidence for a legislative bill to change HIV testing policy in California. The bill passed and was signed by the Governor. Beginning in 2008 in California HIV testing could be done as an opt-out test similar to other medical tests. In other words, a patient would be told by the physician of the intention to order HIV testing, and the patient could tell the doctor if he or she did not want the test.

8.2 How many observations are needed for performing an interrupted time series analysis?

As with many statistical questions, the general answer is easy and the specific answer is hard. The general answer is the more data points you have the better. Classically, experts argued that you should have at least 100 data points.\(^5\) However, many valid analyses have been published with fewer.

For a valid ARIMA model, there should be at least 20 data points prior to the interruption.\(^6\) The reason for wording the guidance in terms of the number


of points prior to the intervention is that you need a solid baseline to evaluate changes due to an intervention.

A large number of points following the intervention is useful for establishing the stability of the change – but there may be times when an intervention really does result in a substantial change in level, but the change is not sustained over time because the impact of the intervention dissipates.

In the case of the study on the impact of a change in HIV testing consent on number of tests ordered (Section 8.1), we had 22 data points before the intervention and 13 data points after the intervention.
9.1 What methods are available for evaluating interventions that do not occur to all subjects at the same time?

It is common in observational cohorts to have different subjects start an intervention at different times, with some subjects never receiving it. This is particularly common when new medications are introduced into practice. In situations such as these, you may have no subjects on the medication (prior to approval) and then a gradual uptake in medication use. Let’s assume you want to evaluate the impact of a new medication designed to prevent osteoporosis. Your outcome is occurrence of a pathologic fracture (yes/no). How would you do it?

At first blush you might think that you could compare the number of fractures in the 280 persons who took the drug to the number of fractures among the 720 persons who did not take the drug.

But this would be wrong from several points of view. First, for those people who dropped out we would only know what happened to them prior to the drop-out (e.g., whether they started taking the medicine, that they didn’t develop a fracture). By year 5, some of them may have begun the medicine and some may have had a fracture. We could exclude dropouts from our analysis but then we would lose 15% of our total sample. And what about the deaths that occurred during the study prior to a fracture? Had these people lived some of them may have started taking the medicine and some of them may have had a fracture. Do we exclude the deaths as well and lose another 21% of our sample?

To minimize the impact of losses to follow-up we could use survival analysis. This would allow us to censor subjects when they dropped out; in other words, we would maintain the benefit of the time they were under observation (Section 6.4.G). The problem with this solution is that it creates survival bias. Specifically, persons who survive longer have a higher chance of receiving the new medication (because, as you can see in Table 9.1, rates of medication use are increasing over time in the cohort). Therefore, people who received a new
Evaluating Clinical and Public Health Interventions

Table 9.1. Uptake of a hypothetical medicine to prevent osteoporosis in a hypothetical cohort of 1000 patients.

<table>
<thead>
<tr>
<th></th>
<th>Year one</th>
<th>Year two</th>
<th>Year three</th>
<th>Year four</th>
<th>Year five</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative subjects</td>
<td>0</td>
<td>90</td>
<td>170</td>
<td>225</td>
<td>280</td>
</tr>
<tr>
<td>taking medicine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative pathologic</td>
<td>10</td>
<td>15</td>
<td>25</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>fractures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative drop outs</td>
<td>40</td>
<td>60</td>
<td>85</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td>(before a fracture)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative deaths (before</td>
<td>50</td>
<td>100</td>
<td>140</td>
<td>190</td>
<td>210</td>
</tr>
<tr>
<td>a fracture)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fracture-free sample size</td>
<td>900</td>
<td>825</td>
<td>750</td>
<td>675</td>
<td>600</td>
</tr>
<tr>
<td>at end of year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

medication within the cohort will tend to do better even if the medication does not work, simply because they lived long enough to receive the medication.

To avoid survival bias in cohort studies, use time-dependent covariates to represent the different times at which subjects begin the intervention. With a time-dependent covariate, the variable changes value at the time that the intervention occurs. In the case of a new medication, every one in the cohort illustrated in Table 9.1 would have a value of zero at the start of the study and then the variable would switch to a value of 1 at the time the subject begins the medication. If the subject never begins the medication during the study the value remains zero for that subject throughout the study. It is also possible to construct time-dependent variables that change value back and forth (if, for example, a person starts and stops taking a medication).

For example, De Martino and colleagues used time-dependent covariates to evaluate the effectiveness of antiretroviral therapy for children with perinatal HIV-1 infection.\(^1\) The investigators followed children born between 1980 and 1997; the outcome of the study was survival.

As you can see in Figure 9.1, the use of antiretroviral therapy increased over the study period. In the first panel you can see that none of the children was taking antiviral therapy in 1987 and over 80% were taking it by 1998. Moreover, during this period, therapy intensified. At first monotherapy grew in numbers, then double combination therapy, and in the later years triple combination therapy.

Had the investigators treated antiretroviral therapy as a yes/no variable, receipt of therapy would have certainly been associated with longer survival.

---

because children alive in the later years were more likely to receive treatment, especially triple combination therapy, which wasn’t used prior to 1997 and is markedly more effective than other types of therapy.

Therefore, the investigators created three time-dependent variables: one for monotherapy, one for double combination therapy, and one for triple combination therapy; the value for each of the three variables was set to zero for each subject at the start of the study and became one at the time (if ever) the therapy was begun. The results are shown in Table 9.2.

Time-dependent covariates do not have to be dichotomous. Time-dependent variables can be interval measures such as systolic blood pressure or CD4 lymphocyte counts. At each point at which the variable is measured it takes on a different value.
9.2 What if I want to prove that an intervention is not inferior or is equivalent to another intervention?

Up until this part of the book, we have been discussing how to determine whether persons in the intervention group do better (or worse) than subjects in the non-intervention group(s). This is called a superiority trial. However, sometimes the goal is to prove that one treatment is not inferior to another (a non-inferiority trial) or that two or more treatments are equivalent (an equivalence trial).²

The terms non-inferiority trial and equivalence trial are sometimes confused; although they have similarities, they are not the same. Because non-inferiority trials are much more common, I will review them first.

Non-inferiority trials seek to demonstrate that a treatment is not substantially worse than another treatment. Non-inferiority trials are particularly important if one intervention is substantially less expensive, has fewer side effects, or has a higher rate of compliance than the standard intervention.

To perform a non-inferiority trial you must first define the size of the difference between the treatments that would lead you to conclude that a treatment is worse than the standard treatment. The amount should be the smallest difference between two treatments that would be clinically significant. In other words, if a new treatment were this much worse, you would stick with the standard treatment. The amount can be defined in percentages or points on a scale.

Special topics

Power calculations and statistical testing are based on the 95% confidence interval for the difference between the treatments being less than the preset threshold for what would be considered clinically worse.

For example, it is well known that continuous exercise benefits persons with chronic obstructive pulmonary disease. The problem is that continuous exercise can be difficult for persons with COPD, resulting in low compliance. Puhan and colleagues assessed whether interval exercise would have similar beneficial effects to continuous exercise.³

The investigators defined 45 meters in a 6-minute walking test as a clinically meaningful standard. In other words, if participants in the continuous intervention arm could walk 45 meters more than the participants in the interval intervention arm, then the interval intervention is inferior; if those in the continuous intervention arm cannot walk 45 meters more than those in the interval intervention arm, then interval exercise is not inferior.

Following the two interventions the adjusted difference between the two groups on the 6-minute walking test (distance walked by interval group – difference walked by continuous exercise group) was 1.1 meters. The 95% confidence interval was (−25.4 to 27.6). Because the confidence intervals did not include −45 meters the study demonstrated non-inferiority of interval exercise. Persons in the interval exercise group were more likely to adhere to the protocol (48% vs. 24%; \( P = .0014 \)).

Two additional points about this study: I chose it because most people think of non-inferiority studies in the context of medication comparisons – I wanted to illustrate that the design can be used for non-medication trials. Second, the comparison of the two groups reported above was based on a per-protocol analysis, not an intention-to-treat analysis. From Section 4.8 you will remember my explaining that intention-to-treat analysis is preferred for randomized trials because it maintains the benefits of randomization. However, this is not necessarily the case for inferiority trials. The reason is that intention-to-treat analysis tends to bias towards the null finding, which would make it easier to demonstrate non-inferiority. That’s why non-inferiority trials should use per-protocol analysis; ideally, per-protocol and intention-to-treat analyses will produce similar answers. (Although the main analysis reported by the investigators in this trial was the per-protocol analysis, they also reported the intention-to-treat analysis, which showed similar results.)

While non-inferiority studies seek to prove that the difference between the effect of an intervention is no worse than a preset standard, equivalence studies

seek to prove that the difference between two (or more) interventions is no larger than a preset standard. In the language of statistical hypothesis testing, with non-inferiority studies we are only interested in one tail; with equivalence studies we are interested in two tails.

For example, Parienti et al. were interested in establishing that hand-rubbing with an aqueous alcohol solution and surgical hand-scrubbing were equivalent in terms of rates of surgical infections. The investigators chose 2% as the difference in surgical infections that would be considered clinically relevant. Therefore, the power calculation for the study was based on assuming that the difference between the infection rates of the two treatments was 0 and the confidence intervals for that difference would be −2.0% to +2.0%. In other words, if the difference between the infection rates for the two groups was more than 2%, the treatments would not be equivalent.

Surgical infections occurred in 2.44% of the patients in the hand-rubbing group and 2.48% of the patients in the hand-washing group for a difference of 0.04% (95% CI = −0.88% to 0.96%). Because the 95% confidence intervals are well within the −2.0% to +2.0% boundary we can conclude that the two treatments are equal.

9.3 How do I adjust for multiple comparisons?

As discussed in Section 3.10, it is often advantageous to test whether an intervention affects more than one outcome. However, when performing multiple comparisons you may need to adjust for the number of comparisons you make.

The reason to consider adjusting for multiple comparisons is the concern that if we perform 20 comparisons between an intervention and a comparison, we would expect that one of the comparisons would be statistically significant at the $P = 0.05$ level by chance alone. Therefore, if one of 20 comparisons is significant, do we conclude that the intervention is effective for that outcome or do we conclude that we have a statistical artifact (this is the same issue with pairwise comparisons (Sections 6.2.A, 6.2.B and 6.4.B), subgroup analyses (Section 9.5), and interim analyses (Section 9.4))?

If you are performing only two to four comparisons, the simplest adjustment method is to use the Bonferroni correction just as you would do when making pairwise comparisons for an ANOVA design (Section 6.4.B). So if you were assessing the impact of the intervention on four outcomes you would divide

---

the conventional $P$ value of 0.05 by 4 and require that any comparison have a $P$ value of less than 0.0125 (0.05/4) before concluding that the difference is statistically significant.

As with pairwise testing, the problem with the Bonferroni correction is that it is very conservative, especially when you have a large number of outcomes. For example, Morey and colleagues performed an intervention to improve functional outcomes among older, overweight cancer survivors.\(^5\) The primary outcome of their study was a change in physical function. In addition they assessed 17 secondary outcomes. If they used the Bonferroni correction for their secondary outcomes the only comparisons that would be considered statistically significant would be those with $P$ values less than 0.003 (0.05/17). Instead, as you can see in Table 9.3, they used an adaptation of the Bonferroni procedure, called the Holm procedure.

For the Holm procedure the $P$ values for the different outcomes are ranked in sequential order from the lowest to the highest. The lowest $P$ value has to be less than the Bonferroni threshold (in this case 0.003) to be statistically significant and to continue testing the remaining outcomes. As you continue assessing the outcomes in order, the denominator on the Bonferroni correction is decreased by one (0.003/16), (0.003/15), etc. This becomes the threshold against which you determine whether the comparison is statistically significant.

If you compare the final two columns you can see that in some cases a difference between the intervention group and the placebo group was below the 0.05 threshold, but did not reach statistical significance using the Holm test. For example, the difference between the groups on the advanced lower extremity function was not statistically significant using the more rigorous threshold.

For a variety of reasons, not everyone believes that adjustment for multiple comparisons is necessary.\(^6\) A major reason is the ambiguity of what comparisons to count. For example, let’s say you read the study by Morey and you were interested in whether the intervention improved sexual function of the elders (there is good evidence that exercise improves sexual function). You contact the investigators; they tell you that they did in fact collect data on sexual function and you are welcome to explore that question with their data (with appropriate attribution, of course). When you perform that comparison would you adjust your analysis for just the number of comparisons you were making (e.g., number of times a week for sex, satisfaction with sex, partner’s satisfaction


\(^6\) Rothman, K. J. “No adjustments are needed for multiple comparisons.” *Epidemiology* 1 (1990): 43–6.
Table 9.3. Differences between the intervention group and the control group on a primary outcome and 17 secondary outcomes.

<table>
<thead>
<tr>
<th></th>
<th>Intervention (n = 319) mean change at 12 mo</th>
<th>Control (n = 322) mean change at 12 mo</th>
<th>Mean group difference (95% CI)</th>
<th>Adjusted P value*</th>
<th>Holm procedure α level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary outcome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36 physical function (range, 0–100)</td>
<td>−2.15 (0.9)</td>
<td>−4.84 (0.9)</td>
<td>2.69 (0.17 to 5.21)</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td><strong>Secondary outcomes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLF basic lower extremity function (range, 45.6–100)</td>
<td>0.34 (0.6)</td>
<td>−1.89 (0.6)</td>
<td>2.24 (0.56 to 3.91)</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>LLF advanced lower extremity function (range, 0–100)</td>
<td>−0.37 (0.5)</td>
<td>−2.30 (0.6)</td>
<td>1.92 (0.45 to 3.39)</td>
<td>0.02</td>
<td>0.006</td>
</tr>
<tr>
<td><strong>Behavioral targets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of strength training exercise, min/wk (range, 0–600)</td>
<td>18.7 (2.4)</td>
<td>0.5 (2.7)</td>
<td>18.21 (11.21 to 25.21)</td>
<td>&lt;0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Duration of endurance exercise, min/wk (range, 0–149)</td>
<td>36.3 (4.9)</td>
<td>23.4 (3.6)</td>
<td>12.89 (1.89 to 27.58)</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Strength training exercise frequency, sessions/wk (range, 0–7)</td>
<td>1.4 (0.2)</td>
<td>0.2 (0.1)</td>
<td>1.12 (0.70 to 1.54)</td>
<td>&lt;0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Endurance exercise frequency, sessions/wk (range, 0–15)</td>
<td>1.6 (0.2)</td>
<td>0.5 (0.2)</td>
<td>1.05 (0.39 to 1.72)</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Daily servings of fruits and vegetables (range, 0–15.80)</td>
<td>1.24 (0.14)</td>
<td>0.13 (0.11)</td>
<td>1.11 (0.76 to 1.47)</td>
<td>&lt;0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Sat. fat intake, g/d (range, 2–57 g)</td>
<td>−3.06 (0.51)</td>
<td>−1.07 (0.49)</td>
<td>−1.99 (−0.58 to −3.40)</td>
<td>0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>Weight, kg (range, 59.1–125.5 kg)</td>
<td>−2.06 (0.19)</td>
<td>−0.92 (0.2)</td>
<td>−1.14 (−0.59 to −1.69)</td>
<td>&lt;0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>Body mass index (range, 25–47)</td>
<td>−0.69 (0.07)</td>
<td>−0.31 (0.08)</td>
<td>−0.38 (−0.19 to 0.57)</td>
<td>&lt;0.001</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>Health-related quality of life on SF-36</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General health (range, 15–100)</td>
<td>0.77 (0.72)</td>
<td>−1.94 (0.80)</td>
<td>2.71 (0.58 to 4.84)</td>
<td>0.03</td>
<td>0.007</td>
</tr>
<tr>
<td>Pain (range, 10–100)</td>
<td>−0.78 (1.07)</td>
<td>−3.19 (1.22)</td>
<td>2.40 (−0.79 to 5.59)</td>
<td>0.16</td>
<td>0.02</td>
</tr>
<tr>
<td>Vitality (range, 0–100)</td>
<td>−0.47 (0.89)</td>
<td>−2.42 (0.98)</td>
<td>1.95 (−0.64 to 4.55)</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>Social functioning (range, 12.5–100)</td>
<td>−1.29 (1.05)</td>
<td>−5.05 (1.22)</td>
<td>3.75 (0.58 to 6.92)</td>
<td>0.03</td>
<td>0.008</td>
</tr>
<tr>
<td>Mental health (range, 32–100)</td>
<td>0.50 (0.53)</td>
<td>−2.04 (0.74)</td>
<td>2.54 (0.75 to 4.33)</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>Physical role (range, 0–100)</td>
<td>−2.43 (2.02)</td>
<td>−4.68 (2.14)</td>
<td>2.25 (−3.54 to 8.05)</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>Emotional role (range, 0–100)</td>
<td>−0.73 (1.32)</td>
<td>−0.62 (1.38)</td>
<td>−0.11 (−3.86 to 3.64)</td>
<td>0.93</td>
<td>0.05</td>
</tr>
</tbody>
</table>

*All models were adjusted for baseline value of the outcome, age, race, number of comorbidities, number of symptoms, education, cancer type, body mass index, and SF-36 physical function subscale score.

with sex) or for the number of comparisons Morey and colleagues made in their paper plus the number you were making?

What if you identified 20 datasets that randomized persons to an exercise regimen and collected data on sexual function? Would you adjust for the number of comparisons you made using the 20 datasets? If you are thinking that it is far-fetched to adjust for comparisons in different datasets, consider this: if you were to compare the benefit of exercise in 20 different trials, you would expect that one of the trials would show benefit for exercise at the $P = 0.05$ threshold by chance alone. On the other hand, you wouldn’t expect the 20 different investigators to adjust their data for the other 19 studies. In fact, they wouldn’t even necessarily know the other studies were going on. It’s a conundrum.

My general advice on this complicated topic is to:

1. Pre-specify the comparisons you intend to perform in your analysis plan (Section 3.11), including being clear what your primary outcome is (it is not necessary to adjust the primary analysis for the number of secondary comparisons).
2. Tell the reader how many comparisons have been made; that way even if you do not adjust for multiple comparisons, the reader can take the number of comparisons performed into account.
3. Don’t be a slave to $P$ values. Whether a comparison is just over or just under a threshold, whether that threshold is adjusted or not, is unlikely to be the major determinant of whether an intervention works. Other factors, including biological plausibility, consistency with other empiric findings, clinical effect size, elimination of confounding, success of blinding, drop out rates, intensity of exposure, contamination, missing data, dose–response relationship, are more likely determinants of whether the intervention really works.

### 9.4 Should a trial have pre-specified interim analyses with early stopping rules?

When designing a trial you may want to plan for an interim data analysis. The purpose of interim analyses is to stop a trial early if there is sufficient difference between the groups to conclude that the intervention is effective or harmful. In the case of efficacy, early stopping may allow subjects in the placebo arm, and those not in the study at all, to receive the beneficial treatment sooner. In the case of side effects, early stopping may prevent unnecessary harm. Early stopping may also save money and facilitate earlier reporting and translation of the results to clinical practice.
The major problem with interim analysis is that by repeatedly “checking” to see whether there is a difference between the groups you are increasing the chance of making a type 1 error (falsely concluding that there is a difference between the groups when there is no difference). To appreciate why, imagine that you check whether there is a difference between your groups at 20 different times. We would expect, by chance alone, that the difference between the groups would be statistically significant at the $P = 0.05$ level at least once.

Therefore interim analyses set a more stringent $P$ value to avoid prematurely concluding the study. These interim analyses are performed at pre-specified times – time is usually based on a percentage of study accrual (e.g., when 60% of the subjects have been enrolled in the study). At each pre-specified time the investigators will specify how large a difference between the two groups will constitute sufficient evidence to terminate the study.

The combination of pre-specified times for the early analysis and the threshold of the $P$ value for ending the study early are referred to as early stopping rules. If there is more than one planned early analysis (e.g., when 50% of the subjects have been enrolled and when 70% of the subjects have been enrolled) the $P$ value for ending at the first look will be more stringent than the $P$ value for ending at a later time. The interim analysis should be performed by a data and safety monitoring committee (Section 3.14).

For example, Lallermant and colleagues planned interim analyses with early stopping rules for their trial of single-dose nevirapine plus zidovudine for the prevention of perinatal transmission of HIV. The investigators enrolled 1844 women to one of three groups: (1) single-dose nevirapine for mother and infant (nevirapine–nevirapine regimen); (2) single-dose nevirapine for mother and placebo for infant (nevirapine–placebo regimen); (3) placebo for mother and infant (placebo–placebo). Patients and investigators were masked to treatment assignment (double-blinded).

The investigators planned to conduct the first interim analysis when 40% of the women were enrolled and a second interim analysis when 70% of the women were enrolled. An increase in the rate of HIV transmission associated with any of the three treatment regimens would be considered significant if any nominal $P$ value was less than 0.0004 at the first interim analysis.

---

Indeed, after reviewing the data at the first interim analysis, the independent data and safety monitoring committee recommended stopping enrollment in the placebo–placebo group. The transmission rate in women in the nevirapine–nevirapine group (1.1%, 95% confidence interval 0.3 to 2.2) was significantly lower than that in the placebo–placebo group (6.3%, 95% confidence interval 3.8 to 8.9%). No serious adverse reactions were reported with the single dose.

I chose this example because the stakes of this study were so high (preventing HIV infection) that you can really appreciate the benefits of early stopping rules. We wouldn’t want to continue giving women placebo if we have enough information to say that infection rates are lower with treatment and the treatment is safe.

However, early stopping rules may not always be a good idea. Montori and colleagues conducted a systemic review of randomized controlled trials that were stopped early because of greater benefit in one group. They found that many studies did not report the planned sample size for interim analysis, or whether a stopping rule informed the decision, or whether the interim analysis was adjusted for interim monitoring. Although these are problems that can be fixed, they also found that many trials that were stopped with few events yielded implausibly large treatment effects, suggesting that they may have caught a random high.

The chance of catching a random high (or low) can be made smaller by using more stringent cut-offs for early termination; on the other hand, more stringent cut-offs may make it impossible for a study to be terminated earlier, in which case there is no purpose to an interim analysis.

There are two other issues of interest concerning interim analyses. The data and safety subcommittee members need not unblind the study to perform an interim analysis. They can simply compare the groups on the pre-specified criteria. If there is a sufficiently large difference between the groups that the study will be terminated early, the subcommittee members can then unblind the data and determine which group was doing better.

When you have a well-publicized study that is known to have interim analyses, researchers and clinicians in the field will draw conclusions when the interim analysis is conducted and no study termination occurs. They will assume that the treatment difference is not large. However, just as there can be random highs, there can be random lows, and a drug may turn out to be effective when a sufficient number of events have occurred.

9.5 How do I analyze subgroup differences in the response to an intervention?

It is often of interest to know whether an intervention works better (or worse) for different groups. The subgroups are typically defined by baseline characteristics such as demographics (e.g., does aspirin have a greater benefit among men than women in preventing heart disease?), presence of comorbid conditions (e.g., do ACE inhibitors have greater benefit among diabetics than non-diabetics?) and severity of disease (e.g., do HIV antiretroviral drugs have greater benefit in patients with severe immunodeficiency than patients with normal immune function?).

Given the obvious importance of subgroup analysis it may seem surprising that it has such a bad reputation. Why? The reason is that some investigators have mistakenly claimed to find an effect in subgroup analyses when all they really found was statistical artifact. Here’s how.

Let’s imagine that you compare an intervention to placebo in 200 people. You find that the intervention has no effect compared to placebo. You then compare the intervention to placebo in men and women separately. No difference. You then compare the intervention to placebo across ethnicities. No difference in any ethnic group. You then compare the intervention to placebo among the young and the old. You find that the intervention is significantly better for the old but not the young. You are excited to have finally identified a significant value and draft a manuscript on how the intervention works for older people. Wrong! Why?

If you compare an intervention to placebo in enough different subgroups you would expect that the intervention would differ from placebo in at least one of the groups by chance alone. Note this is the same reason that when performing multiple interim analyses (Section 9.4) you need to pre-specify when you will be performing the analysis and you must use a more stringent $P$ value to determine statistical significance.

Another reason subgroup analysis has a bad reputation is that investigators make a common statistical mistake in analyzing this type of data: concluding that there is a difference between groups based on separate analyses of each group (e.g., determining the odds ratio for the intervention versus placebo among men and separately determining the odds ratio for the intervention versus placebo among women). To demonstrate that there is a statistically significant difference across subgroups you must demonstrate that there is a statistically significant interaction term.

To construct an interaction term multiply the intervention term by the subgroup. If the intervention variable is (1 = yes; 0 = no) and the subgroup is male sex (1 = yes; 0 = no), then the interaction term will take on two values (0 or 1); the group with both characteristics will be the only group with a value of one as shown below:

\[
\begin{align*}
\text{Intervention group and male sex} & = 1 	imes 1 = 1 \\
\text{Intervention group and female sex} & = 1 	imes 0 = 0 \\
\text{Non-intervention group and male sex} & = 0 	imes 1 = 0 \\
\text{Non-intervention group and female sex} & = 0 	imes 0 = 0
\end{align*}
\]

Although the only group with a value of one is the first group, you can see from the columns of Table 9.4 that the three variables (shown in the rows) – the intervention term, the sex term, and the interaction term – define four unique groups.

Let’s assume that the outcome of this hypothetical study is rupture of an aortic aneurysm. If the intervention is effective in reducing ruptures then the variable intervention (row 1) will be statistically significant. If males have more ruptures than females the variable male sex will be statistically significant (row 2). If the intervention is statistically more or less effective among men than women, the interaction term (row 3) will be statistically significant.

Armed with the knowledge of how to test whether an interaction is statistically significant, there are two correct ways of performing subgroup analyses: with a pre-specified analysis plan (Section 3.11) or with tremendous humility.

At the outset of a study there may be reason to suspect that an intervention will work better in one group than another. Go ahead and describe in the analysis plan your intention to study whether the intervention works equally well in both groups. Power your analysis sufficiently for this comparison. If you are performing a randomized trial, you may wish to use a stratified allocation of subjects (Section 4.4) to be sure you have a sufficient number of subjects in each group.

Even if you don’t identify subgroup analysis as part of a pre-specified plan, you may discover in conducting your study, or in analyzing your data, that
there are differences across subgroups. It is fine to go ahead and explore those differences, but this is post-hoc analysis (Section 3.11). Show tremendous humility in reporting the results.

9.6 How should I deal with missing data in my intervention study?

Handling missing data is a challenge in any kind of study. With intervention studies, it is common to have missing data on baseline characteristics (covariates), especially if the subjects were drawn from medical records, registries, or administrative databases.

If you have a very small number of cases with missing data on covariates, it is easiest to simply delete those cases from your analysis. If the number of deleted cases is a small proportion of your total sample (e.g., < 3%), and the cases with missing data are not different from cases without missing data on other prognostic factors, deleting them is unlikely to bias your analysis.

At the other extreme, if you have variables with a large number of missing cases (e.g., > 15%), the best solution may be to delete these variables (rather than the cases) from the analysis. No matter how important the variables may be to your theory, if you have a lot of missing cases on these variables, including them in the analysis will lead to loss of cases (if you exclude cases with missing data) or potential measurement error (if you must estimate the missing values using multiple imputation: see below.)

When the number of cases you would lose by excluding those with missing values is not trivial and when your covariates have some (but not a substantial proportion) of missing values the best way to incorporate these cases is to use multiple imputation to estimate the missing values.\(^\text{10}\) Multiple imputation is a sophisticated method of estimating the value of missing values, based on the values of those subjects with complete data on the missing variable and its important correlates. In addition a random component with repetition is used to compute standard errors that take into account the uncertainty induced by imputing the value.

With a longitudinal study, where the outcome is measured repeatedly, there may also be missing outcome data (e.g., a subject attends the first five study appointments but misses the sixth one). In situations such as these, many analytic methods (e.g., ANOVA, conditional logistic regression) cannot be used because they require the same number of observations per subject.

In cases where there is a very small percentage of subjects with a missing observation, and these subjects are no different from those subjects with all their values, it is possible to exclude these subjects from your analysis. Unfortunately this results in loss of sample size.

In the past it was common to deal with this problem by carrying the last observation forward. With this method the missing value is replaced with the most recent valid non-missing value. Unfortunately, this method can result in underestimating or overestimating the treatment effect and should not be used.\(^{11}\)

A better method of dealing with missing observations is to use an analytic method that can incorporate missing observations such as generalized estimating equations and mixed-effects models. Both can deal with data that are missing randomly and may be able to deal with nonrandomly missing data when the outcome is interval (but not dichotomous).\(^{12}\)

**9.7 What special considerations are there for publishing the results of intervention trials?**

A major improvement in the field of intervention research has been the establishment of guidelines for publishing studies. The guidelines ensure that investigators include all relevant information in their manuscripts. By standardizing the presentation of results, guidelines also help readers to rapidly find the information they are looking for and make it easier to compare the results of trials, informally and through meta-analysis. There is a guideline for publishing randomized intervention trials\(^ {13}\) and an extension of this guideline for non-inferiority and equivalence randomized trials.\(^ {14}\) For publishing nonrandomized studies, two guidelines are available.\(^ {15}\)

---


The guidelines themselves are very detailed, and I won’t attempt to review them in full. But I do want to stress the importance of including a diagram showing the flow of subjects through a trial. A generic version of such a figure is shown in Figure 9.2.16

Figure 9.3 is from a study of physical activity on cognitive function in patients at risk for Alzheimer disease.17 Starting at the top of the figure and working your way down, note how the figure enables you to quickly assess who is excluded and why.

Continuing down from the branch point it is easy to see how many subjects were randomized into each group, and within each group, and how many subjects completed the 6-month, 12-month, and 18-month assessments, along

---


311 Individuals screened for eligibility by telephone

141 Excluded
52 Declined participation
48 Physical health exclusion
10 Mental health exclusion
19 Could not be recontacted
4 Participating in another trial
4 No memory complaints
2 Lived outside Perth
2 Poor English communication

170 Participants randomized

85 Randomized to exercise group
71 Completed 6-mo follow-up
8 Lost to follow-up
4 Refused participation
3 Medical illness
1 Died
6 Discontinued intervention
4 Medical illness
2 Gave no reason

85 Randomized to control group
80 Completed 6-mo follow-up
5 Lost to follow-up
4 Refused participation
1 Family illness

69 Completed 12-mo follow-up
2 Lost to follow-up
1 Refused participation
1 Missed visit and could not be contacted

68 Completed 12-mo follow-up
10 Lost to follow-up
5 Refused participation
3 Medical illness
1 Moved
1 Family illness
2 Missed visit
1 Family illness
1 Participant illness

69 Completed 18-mo visit
1 Returned to study
1 Lost to follow-up

69 Completed 18-mo visit
2 Returned to study
1 Lost to follow-up due to family illness

85 Included in primary analysis

85 Included in primary analysis

Evaluating Clinical and Public Health Interventions

with the number of subjects who were lost to follow-up or discontinued the intervention at each of these time points.

The very bottom of the flow diagram indicates that the primary analysis was performed on the same number of subjects as were initially randomized. This indicates that the investigators conducted an intention-to-treat analysis (Sections 3.9, 5.8).

9.8 What special considerations are there for publishing negative studies?

Studies where the intervention does not work are less likely to get published (Section 3.9) or at least less likely to get published in the most prestigious journals. There are several reasons for this.

One reason for not publishing a negative study is that it isn’t sufficiently powerful to exclude the possibility that the intervention works (type 2 error). For example, colleagues of mine were interested in answering an important question: how to get more persons at high risk of HIV to enroll in available HIV prevention programs. Our earlier work had shown that less than 6% of high-risk patients receiving HIV-negative test results at the San Francisco sexually transmitted disease (STD) clinic had received an HIV-prevention referral; how many of the 6% actually followed through on the referral was not known. 18

To increase the proportion of high-risk HIV-negative persons receiving and following through on HIV-prevention service referrals, we provided HIV counselors at the STD clinic with training on how to more effectively provide referrals to patients; counselors were also given detailed lists of available referrals. 19

Since counselors at the STD clinic believed that high-risk HIV-negative persons did not perceive themselves to be at risk for HIV and had little motivation to follow through on referrals, an intervention was developed to increase follow-through on referrals by having counselors give clients additional referral information, ongoing encouragement and assistance, reminders, and reimbursements for childcare and transportation.

Clients were randomized to the intervention or to usual care. A power calculation showed that to detect a 10% improvement in accessing referrals with the intervention (assuming the intervention group accessed referrals with 20% frequency and the non-intervention group with 10% frequency), a sample size


of 200 per group would be needed (assumes power of 80% and two-sided statistical testing at alpha = 0.05).

Given the high volume of high-risk patients coming to the STD clinic it should have been possible to recruit a sample of this size, but for a variety of reasons it wasn't. Over a 14-month period, only 113 of the 121 persons who met study criteria agreed to participate, and only 79 were available for baseline interviews. Of these 79 clients, 39 were randomized to the intervention and 40 to the non-intervention group. However, the effective sample size was even smaller than this because the intervention could only be done for patients who did not access referrals prior to the intervention and who were available for a follow-up interview. In this small group, there was no difference in follow-up of referrals between the two groups: only 2 of 16 intervention clients (12.5%) versus 2 of 22 non-intervention clients (9.1%) ($P = 1.0$).

The 95% confidence interval for the odds ratio of 1.4 was 0.18 to 11.38. In other words, the investigators could not rule out that the intervention could substantially increase or decrease follow-through on prevention referrals. For this reason, the data really could not inform next steps and were not published.

However, even if you power your study such that you can rule out that the intervention works, you may still find it hard to publish? Why? Because there remains a bias against negative studies among journal reviewers and editors. This is unfortunate. A well done study that is adequately powered and evaluates an intervention that is either in use or may be potentially useful deserves the same attention whether the intervention works or not. Favoring studies with positive results encourages investigators to slant their results to demonstrate effects when the results are equivocal. If you've done a good study, don't be any less proud of it just because the intervention didn't work.
10.1 How do you translate research into practice?

Many highly skilled researchers – able to develop, implement, and evaluate cutting-edge interventions – often fall short on the last, and most important, stage of the process: translating their research into practice.

Much of the problem is that the people who develop and evaluate interventions are usually academics, and academia is primarily concerned with generating new knowledge. Translating interventions into the messy world of government bureaucracies, public health departments, and social service agencies, as well as dealing with political and funding challenges, has never been the focus of academic work.

With medication and medical devices there exists a machinery – drug companies and medical device suppliers – to translate research findings into practice, including convincing doctors to prescribe them, patients to take them, and insurance companies to pay for them. No comparable machinery exists for behavioral or structural interventions (e.g., changes to the physical environment, changes to the law). Political leaders, public health and social service practitioners and advocates do not typically read research journals, and when they do they are often unsure how to translate projects that were conducted in a research setting into practice.

This gap between knowledge generation and program adoption must be bridged if we are to take full advantage of newly developed interventions. There are a number of steps that researchers can take to make translation of their work easier:

- Design interventions that are translatable.
- Articulate the benefits of the intervention in terms people will understand and be motivated by.
- Use the media to disseminate the results.
- Gain the support of interest groups.
- Send the results to policy makers.
- Develop a tool kit to make it easier to adopt the intervention.
In Section 2.2.C, I discussed the importance of developing interventions that are likely to be implemented. Interventions that have been shown to be effective, especially in real-world settings, are inexpensive, do not require highly specialized equipment or staff, and are of interest to the target population, are the ones most likely to be implemented.

For people to be interested in your intervention you will need to clearly articulate its benefits. Will people feel better? Will they save money? Will they live longer? If the intervention is aimed at physicians, will the intervention help them to practice better? Will it save them time? Will their patients feel better?

If you are aiming to convince politicians of the need to implement a law, they will want to know what advantage the law will provide for their constituents. Will it improve their services? Will it save them money?

Often local leaders can be persuaded to adopt a policy if it is considered to be a “best practice” in other parts of the world. For example, when I was convincing local politicians to have San Francisco become the first city in the United States to ban the sales of tobacco in pharmacies, I emphasized that a similar ban had been passed 10 years earlier in Canadian provinces. This made it seem like a less radical idea.

Many researchers are frightened by the press, thinking that reporters will try to sensationalize their results. Although it is true that the job of journalists is to sell newspapers (or magazines, TV advertisement time, web ads, etc.), journalists do not like to get the facts wrong. When stories are wrong it is often because no one took the time to explain to the reporter the findings in an understandable way.

To promote your findings, prepare a press release with a focus on what is new and exciting about your findings (detailing how your study confirms the results of prior studies is not likely to get much play). Send it out to newspapers and media outlets that will be interested. Give them lists of people to talk to, with phone numbers and emails, especially people not directly involved in the study who can attest to the broader implications of the work; follow up with phone conversations to the reporters.

Individuals do not change the world. Groups of individuals change the world. The best way to get the attention of policymakers is to line up several advocacy groups who support the intervention. For example, the American Heart Association and the American Lung Association were critical in pushing through smoking bans in the United States.

Who are the natural supporters of your intervention? Get them involved in advocating for it.

As someone who works within a government bureaucracy, I chuckle when I read scholarly articles that include recommendations to policy makers. Few policy makers I know read scholarly articles. To engage policy makers, make an appointment with them and bring them copies of the research, but also “one-pagers” explaining the intervention and its value in easy-to-understand language. Tell them who supports your intervention and why. Also tell them who opposes it (policy makers hate being surprised by a group of angry constituents) and how their opposition can be neutralized.

Interventions take time to implement. To increase implementation of an intervention create a tool kit. A tool kit should include background on the problem that the intervention is meant to address, descriptions of how to conduct the intervention, and the materials necessary to accomplish the goal. For example, when my county created a universal health coverage program, we received many inquiries from other counties interested in pursuing similar initiatives. To make it as easy as possible, we asked a health foundation to create a tool kit for our program that included descriptions of the program, authorizing legislation, financing plan, quarterly charges, and point-of-service costs.

10.2 How is the translation of interventions assessed?

A framework has been developed for assessing the translation of public health interventions, called RE-AIM. Although the success of translating an intervention can usually be judged in less formal ways, the framework is helpful in assessing the different stages of implementation.

RE-AIM stands for reach, efficacy or effectiveness, adoption, implementation, and maintenance. Reach refers to the proportion of the target population who participated in the program and the characteristics of those persons. Participant characteristics are important to ensure that the program is reaching those who most need it (e.g., implementing an exercise program that attracts only persons who are already fit will not change obesity levels in the community).

Efficacy/effectiveness refers to the success rate when implemented as recommended. Adoption refers to the proportion of settings (e.g., clinics, community centers) that adopt the intervention. Implementation is the extent to which the program is implemented as recommended. Maintenance is the extent to which the program is sustained over time.


Research to action

For example, Li and colleagues used the RE-AIM framework to assess the translation of a tai chi intervention to six community senior centers in five cities in Oregon, USA. In terms of reach, of the 555 persons who usually attend activities at the six community centers from which participants were enrolled, 287 signed up for the class and 249 were eligible to participate (reach = 249/555 or 45%). Participants were similar to the general population of senior center users.

In terms of effectiveness, participants showed significant pre-intervention to post-intervention improvement (Section 6.2) on a variety of physical measures. Adoption of the intervention by the six centers was 100% and all centers successfully implemented the program (implementation = 100%). Maintenance of the program was high (five centers continued offering the implementation beyond the study period and one was waiting for instructor availability).

Conclusion

Studying interventions is hard work, especially when the interventions do not lend themselves to randomized controlled designs! This is one reason that risk-factor studies greatly outnumber intervention studies in many critical areas, such as disparities in health outcomes, HIV transmission, substance abuse, and violence.

But the pay-off of developing, studying, and translating interventions into practice – changing the world – is tremendous. Early in my career, I published many descriptive and risk-factor studies. They helped me hone my statistical skills and some may have been of help to others in designing interventions. But I am proudest of the work that I have done developing and implementing interventions, so much so that I no longer perform risk-factor studies (although I still help people with theirs, while trying to recruit them to do intervention studies). I hope that this book will recruit you to the world of intervention studies.

If you need help (or you want to tell me about your success developing, evaluating, or implementing interventions) email me at mhkatz59@yahoo.com.
adherence to treatment, 56–60
administrative databases, 30, 31
adolescent high-risk behaviours, 15–17
ADVANCE trial, type 2 diabetes, 18
AIDS:
  housing impact on mortality in homeless persons, 127
  preventive curriculum, 85, 87
alcohol dependence, 92–3
alcohol withdrawal, 39
allocation methods, 50–2
analysis of covariance (ANCOVA), 84, 87
  repeated measures ANCOVA, 84, 89–90
analysis of variance (ANOVA), 84, 86–7
  repeated measures ANOVA, 76–7, 84, 88, 89–90
analysis plan, 40–2
  post-hoc analyses, 42
ANOVA see analysis of variance (ANOVA)
Anderson, G., 95
Anderson, R. E., 89
ANOVA see analysis of variance (ANOVA)
antipsychotics, mortality and, 63–4, 112–13
antiretroviral therapy in children, 136–8
as-treated analysis, 57–60
aspirin, 103, 108–10, 112
autoregressive integrative moving average (ARIMA), 131, 133
Ayanian, J. Z., 106–7
Back, L., 77–8
bariatric surgery, 68–70
baseline differences, 47–9, 101
  adjusting for, 101, 108–27
  nonrandomization and, 65–8
  see also instrumental variable analysis
  propensity score analysis
Belfort, R., 77, 88
best practice, 155
between-group comparison, 88
bias:
  in case-control studies, 70
  publication bias, 45–6
  randomization and, 57
  survival bias, 135–6
caliper, 109
cancer survivors, functional outcomes, 141–3
cardiovascular disease, estrogen and, 18
Carleton, R. A., 62
carpal tunnel syndrome, 75–6
case-control studies, 12–14, 42, 70, 99–100
  limitation, 14
  censoring, 92–93
  chi-squared test, 82, 84, 95
chlamydia screening in adolescent girls, 54
chronic obstructive pulmonary disease (COPD), 139
cigarette smoking see smoking
clustered (group) randomization, 10, 52–4
  clusters, 71–2
  for nonrandomized designs, 71–2
  Cochran Q test, 79–80
  colorectal cancer, 31
  community advisory committee, 23
  community mobilization study, 62–3
  comparison group:
    change between pre- and post-intervention assessment, 7–11, 83–100
    outcome evaluation, 11–14
    time series analysis, 129
  see also control group selection
  computerized number generator, 47
  birth registries, 30, 32
  blinding, 55–6
  blocked randomization, 51
  blood pressure, dark chocolate impact on, 76–7
  Bonferroni correction, 86, 140–1
  box plot, 106–7
  breast cancer, mammography effectiveness, 116–18, 121, 124, 125
  Bright Bodies weight management program for overweight children, 75, 76, 84–5
c index, 106
chronic obstructive pulmonary disease (COPD), 139
clustered (group) randomization, 10, 52–4
  clusters, 71–2
  for nonrandomized designs, 71–2
  Cochran Q test, 79–80
  colorectal cancer, 31
  community advisory committee, 23
  community mobilization study, 62–3
  comparison group:
    change between pre- and post-intervention assessment, 7–11, 83–100
    outcome evaluation, 11–14
    time series analysis, 129
  see also control group selection
  computerized number generator, 47
Index

Concato, J., 29
concurrent controls, 26–7
conditional logistic regression, 84, 91
certainty intervals, 75, 85
country, 47
sensitivity analysis, 127
see also propensity score analysis, instrumental
variable analysis
country, 65–8, 100
by contraindication, 67
by indication, 65–7
country, 35
country randomization, 10, 52–4
country randomization, 10, 52–4
dark chocolate impact on blood pressure, 76–7
data collection, 23–4
data points, 6–7, 24
De Martino, M., 136
death registries, 32
depression, 89
diabetes:
  ADVANCE trial, 18
  lifestyle intervention, 27, 80–1
  pancreas transplant, 26
  steatohepatitis in, 77, 88
  thiazolidinediione (TZD) use, 32
dichotomous outcome, 74, 79–80, 91
DiClemente, R. J., 48–9
dose-response, 34
double-masked (blinded) study, 56
drug-eluting stents, 93
Dubinet, test, 78
early stopping rules, 143–5
effectiveness evaluation, 3–14
change related to comparison group, 7–11,
  83–100
  outcome related to comparison group, 11–14
pre-intervention versus post-intervention
  assessment, 4–7
effectiveness trials, 19, 20, 64
efficacy trials, 19–20, 64
electronic patient records, 29, 30
equal allocation, 51
  equipoise, 50
equivalence trials, 138, 139–40
  estrogen, cardiovascular disease and, 18
evaluation, 3
control group selection, 24–33
data collection, 23–4
exposure to intervention, 34–5
intervention effectiveness, 3–14
interventions without an evaluation plan,
  28–33
longitudinal cohort versus serial
cross-sectional design, 35–7
measurement instruments, 37–8
outcomes, 33
sample size, 42–3
stakeholder engagement, 23
exercise:
  chronic obstructive pulmonary disease and,
    139
in school-age children, 52–3
exposure to intervention, 34–5
treatment adherence, 56–60
Fagard, C., 79–80
Farley, T. A., 2
folic acid, neural tube defects and, 100
Friedman test, 77–8
Garfinkel, M. S., 75–6
generalizability, 17
increasing, 64
generalized estimating equations, 54, 90
Glasgow, R. E., 20
Golden, J. A., 28
Google search, 15
Grodstein, F., 31
group randomization, 10, 52–4
Gum, P. A., 103, 110
hand-rubbing with aqueous alcohol solution, 140
Hawkins, J. D., 15
Hawthorne effect, 50
healthy eating and exercise in school-age
  children, 52–3
heart disease, estrogen and, 18
hip protectors, 55
historic controls, 27–8
HIV:
  AIDS-preventive curriculum, 85, 87
  antiretroviral therapy in children, 136–8
  perinatal transmission, 144–5
  prevention, 48–9, 82–3, 91, 144–5, 152–3
  testing, 129–33, 134
  treatment interruptions, 79–80
Holm procedure, 77, 141
Holm, S., 77
homicide prevention in workplaces, 13–14
Hu, T., 2
human subjects committees see institutional
  review board
identification of comparable clusters (groups),
  71–2
Index

implementation issues, 154–7
  assessment, 156–7
independent data and safety monitoring
  committee, 45
individual matching, 68–70
influenza vaccination, 65–7, 72, 101
informed consent, 43–5
institutional review board (IRB), 43–5
instrumental variable analysis, 102, 115–26
definition, 116
limitations of, 125–6
performance of, 124
underlying assumptions, 119–24
versus multivariable analysis, 126
instrumental variables, 116–18
baseline characteristics and, 122–3
group assignment association, 121
monotonic relationship with intervention
status, 124
outcome association, 119–21
see also instrumental variable analysis
insurance program impact on access to care, 79
insurance records, 31–2
intensive care units (ICUs), 113–14
intention-to-treat analysis, 41, 57, 59, 139
interaction term, 146–7
interest group support, 155–6
interim data analysis, 143–5
International Committee of Medical Journal
Editors, 46
interrupted time series analysis, 128
see also time series analysis
interval outcome:
  non-normally distributed, 74, 77–8
  normal distribution, 74, 75–7
two or more time points, two or more groups,
  88–9
  two or more time points, two or more groups,
multivariable analysis, 89–90
interventions:
  design of, 1–2, 15–22
evaluation of see evaluation exposure to, 34–5
grounding in theory, 15–18
hypothesis, 38–40, 41
outcomes, 33
practicality issues, 19–22
reasons for study, 1–3
translation of research into practice, 20–2,
  154–7
treatment adherence, 56–60
types of, 15, 16
  without an evaluation plan, 28–33
Johnson, B. A., 92
Jones, K. T., 82–3
Kaplan–Meier curves, 84, 92–3
Kegeles, S. M., 91
Kiel, D. P., 55
Kruskal-Wallis test, 98
Lallermant, M., 144
Landrum, M. B., 106–7
last observation carried forward, 149
LATE see local average treatment effect (LATE)
Lautenschlager, N. T., 89–90
Levy, M. M., 113, 114
Lü, R., 157
Lipscombe, L. L., 32
local average treatment effect (LATE), 125
longitudinal cohort study, 4, 35
  advantages and disadvantages, 35–6
  statistical significance, 74–81, 83–94
Loomis, D., 13
lumbar disk herniation treatment, 57–60
Lyme disease prevention, 39–40
McNemar’s test, 75, 79, 91
McNeil, D. E., 114
mammography effectiveness, 116–18, 121, 124,
  125
Mann–Whitney test, 77, 98
marginal population, 125
masking, 55–6
matching, 68–71
  using propensity scores, 108–12
mean difference:
  two time points, three groups, bivariate
    analysis, 86–7
  two time points, two groups, bivariate
    analysis, 83–5
  two time points, two or more groups,
multivariate analysis, 87–8
measurement instruments, 37–8
Medicaid records, 31, 64
Medicare:
  databases, 31
drug benefit caps, 97–9
memory, physical activity and, 89–90
mental health:
  adolescents with divorced parents, 21–2
casino impact on native American children,
  2–3
mental health court, 114–15
Minnesota Heart Health program, 36–7
missing data, 148–9
mixed-effects models, 54, 84, 90
Montori, V. M., 145
Morey, M. C., 141–3
Moskowitz, J. M., 11–12
multiple comparisons, 40, 76–7, 140–3
multiple imputation, 148
multiple linear regression, 80–1, 84, 87, 98
multiple logistic regression, 80–1, 95–7, 98, 100
multivariable analysis, 42, 67–8, 80–1, 87, 89–97,
  98, 101
  versus instrumental variable analysis, 126

- 161 -
municipal administrative databases, 31
myocardial infarction:
cardiac catheterization, 110–12, 122–3
specialty care, 107
Nadelman, R. B., 39
National Health and Nutrition Examination
Survey (NHANES), 31
natural experiment, 2, 16
negative studies, 152–3
neonatal jaundice, phototherapy and, 86
nested case-controls, 31
neural tube defects, folic acid and, 100
new medication introduction, 135–6
95% confidence intervals, 75, 85
non-inferiority trials, 138–40
non-normal distribution, 74, 75, 77–8
nonrandomized studies, 61
disadvantages of, 65–8
identification of comparable clusters (groups),
70–2
matching, 68–70
reasons for, 61–5
normal distribution, 74, 75–7
null hypothesis, 4, 38, 73
Nurses’ Health Study, 31
obesity, 1–2, 89
bariatric surgery, 68–70
odds ratio, 82, 100
O’Donnell, L. N., 54–5
Office for Human Research Protections (OHRP),
44–5
O’Loughlin, J. L., 34
one-group pre-intervention versus
post-intervention design, 4–7
one-sided hypothesis, 39
ordinal outcome, 74, 77–8
organ transplantation, 26
outcome variables, 73–4
outcomes, 33
multiple outcomes, 40, 140–3
time to outcome, 74, 92–4
P values, 42, 85, 86, 88, 141
paired t-test, 75–6, 88
pairwise comparisons, 76
pancreas transplant, 26
Parienti, J. J., 140
passively observed controls, 27
Pell, J. P., 63
per-protocol analysis, 57, 139
periodic surveys, 31
phototherapy, neonatal jaundice and, 86
pilot studies, 19
pioglitazone, 77, 88
Pletcher, M. J., 39
Pneumocystis carinii pneumonia (PCP)
prevention, 28
policy makers, 156
post-hoc analyses, 42
posttraumatic stress disorder (PTSD), 25
press release, 155
Pronovost, P., 4–5, 43–4
propensity score analysis, 101–115
adequacy assessment, 106–8
adjusting for baseline differences, 108–15
as independent variables in a multivariable
models, 113–14
calculation of, 101–4
choice of approach, 115
matching cases, 108–12
overlap, 106–7
stratifying analysis and, 112–13
variables to be included, 105–6
versus instrumental variable analysis, 126
weighting observations, 114–15
proportional hazards analysis, 84
proportional odds regression, 98
prostate cancer screening, 29
proximal marker, 33
psychopathology see mental health
publication of results, 149–51
guidelines, 149–50
negative studies, 152–3
publication bias, 45–6
PubMed, 15
Puhan, M. A., 139
quality improvement organization, 95–7
random number generator, 47
randomization, 24–6, 47
adherence to treatment and, 56–60
advantages of, 47–9
allocation methods, 50–2
clustered (group) randomization, 10, 52–4
difficulties, 10–11
disadvantages of, 49–50
masking (blinding), 55–6
what can be randomized, 54–5
randomized controlled trial, 7–8, 50, 64
Raschke, R. A., 61
Ray, W.A., 63
RE-AIM framework, 156–7
receiver operating characteristics curve (ROC),
106
registration of trials, 45–6
registries:
birth, 32
death, 32
repeated measures analysis of covariance
(ANCOVA), 84, 89–90
repeated measures analysis of variance
(ANOVA), 75, 76–7, 84, 88, 89–90
research translation into practice, 19–22, 154–6
retrospective cohort, 11
assessment of, 156–7
Index

sample size, 42–3
Savoye, M., 75
screening:
  chlamydia in adolescent girls, 54
  prostate cancer, 29
segmental regression analysis, 129–33
sensitivity analysis, 126–7
serial cross-sectional study, 4, 35–7, 81–3, 95–7
  advantages and disadvantages, 36–7
  statistical significance, 81–3, 95–7
Shafer, M. B., 54
Sjostrom, L., 68
Sliﬁkin, R. T., 79
smoking:
  bans in public places, 11–12, 63
  taxation impact, 2
Snyder, C., 95
socially desirable responses, 36
stakeholder engagement, 23
statistical signiﬁcance:
  longitudinal study outcomes, 74–81, 83–94
  serial cross-sectional study outcomes, 81–3, 95–7
steatohepatitis in diabetes, 77, 88
Stein, B. D., 25
stratiﬁcation, 68, 108
  propensity score and, 112–13
stratiﬁed randomization, 51
Stukel, T. A., 110–12, 122
subgroup analysis, 146–8
  bad reputation of, 146
  interaction term, 146–7
subgroups, 146
superiority trials, 138
survival bias, 135–6
survival curves, 92–3
$t$-test:
  paired, 75–6, 88
  unpaired, 84, 85
tai chi, 157
Tan, K. L., 86
Taubert, D., 76–7
time series analysis: 128–34
  comparison group, 129
  error terms, 131
  number of observations, 133–4
  use of, 128–33
time to outcome, 74, 92–4
time-dependent covariates, 136–7
tonsillectomy, 77–8
tool kit, 156
topiramate, 92–3
translation of research into practice, 20–2, 154–6
  assessment of, 156–7
treatment adherence, 56–60
trial registration, 45–6
Tseng, C.-W., 97–9
Two Feathers, J., 27, 80–1
Two-sided hypothesis, 39
  two-tailed tests, 39–40
type 2 diabetes, ADVANCE trial, 18
unequal randomization, 52
unpaired $t$-test, 84, 85
variables, outcome, 73–4
  see also instrumental variables multivariable analysis
Vaughan, R. D., 85
Venstrom, J. M., 26
video education in STD clinic, 54–5
waiting list controls, 24–6
Walter, H. J., 85
Wang, P. S., 112–13
warfarin, 64–5
weight management program for overweight children, 75, 76, 84–5
weighting observations, 114–15
Wilcoxon signed-rank test, 75, 77
Wingood, G. M., 48–9
within-group comparison, 88
Wolchik, S. A., 21