Critical Appraisal

Table of Contents

Authors:
Elizabeth O'Connor, Ph.D.
Jennifer Bellamy, Ph.D.
Bonnie Spring, Ph.D.

I. Overview of Critical Appraisal

Introduction

This module will focus on the critical appraisal of studies that attempt to determine whether an intervention works. These include:

- Controlled trials, both randomized (or experimental) and non-randomized
- Time series research designs, which include interrupted time series, as well as within-subject or single case designs

- Systematic reviews, including meta-analysis

Conducting and Reporting an Experimental Research Study

Each research project begins by asking a specific, testable question. The project is designed to answer that specific question. In critical appraisal of a research study, we ask ourselves, "How well did this project answer the question?"

Most experimental research studies involve some or all of the following steps:

- Sample Selection
- Group Assignment
- Assessment/Data Collection
- Conducting the Intervention
- Data Analysis

We will discuss the relevant steps for each research design that we cover. For each step, we will provide:

- A fictitious example from a sample methods or results section
- Questions that will help evaluate internal validity
- Questions that will help evaluate external validity (if the methods used for that step are likely to affect generalizability)

Important Terms

These terms and others can be found in the glossary for this module. They will also be clickable throughout if you'd like further explanation.

NOTE: Two additional types of validity are statistical and construct validity. Statistical validity is the extent to which statistical methods are used and interpreted appropriately. We will address this type of validity in the sections covering data analysis.

Construct validity is the extent to which the study tests underlying theoretical constructs as intended. This is particularly important in the development of the intervention and measurement instruments but will not be addressed in this learning module.

- Internal validity is the extent to which the results of a study are true. That is, it really is the intervention that caused the change in behavior, and not some other extraneous factor, such as differences in assessment procedures between intervention and control conditions.

  For example...

- External validity is the extent to which the results can be generalized to a population of interest. For the purposes of this module, the "population of interest" is up to you to determine. We will point out factors that you should take into account when deciding how well the results apply to your population.
- Bias is the systematic distortion of the real, true effect that results from the way a study is conducted. This can lead to invalid conclusions about whether an intervention works. Bias in research can make a treatment look better or worse than it really is. Bias can occur at almost any stage in the research process.

For example...

Suppose that a weight loss study used different follow-up procedures for experimental and control group participants. The researchers assess weight data after one year by telephoning control group participants, but they have the intervention participants come in to the clinic to be weighed. Then the weight differences between the groups could be due to differing assessment procedures, rather than to the intervention.

II. Controlled Trials

What is a Controlled Trial?

- A controlled trial is a study in which participants are assigned to a study group. If random assignment was used to create the study groups, then this is called a randomized controlled trial (RCT). Non-random assignment introduces a greater risk of bias, so random assignment is usually preferable, though it is not always possible.
- Study groups are also called study arms or treatment conditions.
- Procedures are controlled to ensure that all participants in all study groups are treated the same except for the factor that is unique to their group.
- The control group in this kind of study may receive no treatment at all or a weak treatment. Or the researcher may compare the effectiveness of two different active treatment approaches.

Sample Selection – Excerpt

Let's critically appraise a study to see whether sample selection affected the internal and external validity of the research.

The public health department in Arizona observed a high rate of infant mortality from motor vehicle accidents in rural areas. Researchers wanted to learn whether a hospital-based program that provides low-cost child safety seats could reduce infant injuries in motor vehicle accidents in rural counties in Arizona.

Sample Selection – Target Population

We'll use this study to discuss how to evaluate the internal and external validity of the sample selection.

- **Is the target population identified?**

  A good quality trial:

  - Specifies the target population clearly. (The researchers did this: they specified the target population as residents of rural counties in Arizona.)

- Pulls the intervention and control participants from the same population.

In theory, the researchers targeted the same population for both the intervention and control conditions in this study. But in reality, they recruited intervention and control participants from two different locations. This increases the risk that any difference in the study outcome (infant motor vehicle injuries) will reflect pre-existing differences in the kinds of people assigned to the study conditions. This is known as selection bias. For example, what if the control country has a higher proportion of Native American women who give birth at home? A consequence is that a smaller proportion of the new mothers in that particular county get exposed to any intervention delivered in a hospital. The study results might occur not because the intervention was effective, but rather, because a hospital-based intervention in one county reaches more of the population than it does in the other.

- **Are inclusion or exclusion criteria clearly explained?**

  A good quality trial tells the reader any characteristics that people needed to have (or not have) to be eligible for the study. Specification of enrollment criteria shows that the researchers had a clearly-identified target population and were thoughtful and systematic in their sample selection. It also helps readers to determine how well the sample reflects the population of interest to them.

  The Arizona researchers did not list inclusion criteria other than giving birth in one of the study hospitals. Conceivably, the researchers may really have attempted to counsel every woman (which would increase the generalizability of the research). We wonder, though, whether they implemented—but did not report—certain inclusion rules regarding the women's spoken language, vision, or access to transportation.

Selecting the Sample

How did the researchers select the sample from the population? In other words, how did they decide which potentially eligible people to invite to join the study?

- **Who do they invite?**

  A good quality trial specifies exactly how the researchers selected who to invite to participate in the research. Did they invite everyone in the population that

met their inclusion criteria? Just a sample of those eligible?

The Arizona researchers say that they recruited all women giving birth in hospitals located in the study counties. Confidence in the internal validity of a study lessens if the selection process is not described. That omission increases the likelihood that the selection was not systematic and therefore biased. For example, did the county hospitals all have nurse educators available seven days per week to recruit new mothers into the study? If not, then mothers discharged on days without a nurse educator may have been missed. We do not think recruitment omissions constituted a serious problem in the Arizona study, however, since nurse educators in the county hospitals contacted all but 2% of mothers of newborns (61/2959).

- **Were the same procedures followed to recruit people to the intervention and control conditions?**

  To minimize selection bias, a study can:

  - Select intervention and control participants from the same location (e.g., county, clinic, school, classroom), or recruit participants from more than one location for each study group

  If multiple locations are used, then a good quality trial:

  - Matches similar locations and randomly assigns the locations in each pair to their study arms
  - Assesses the characteristics of the locations and makes sure they are comparable

  The Arizona study recruited intervention and control participants from two different counties—one for each treatment group—which can be a concern. They did attempt to select two counties that were similar on some relevant characteristics, and the fact that they reported no differences in baseline characteristics of enrolled participants in the two counties was reassuring. However, since every county has unique features, we would have been more confident about the study's internal validity if they had randomized multiple counties to each study condition.

- **Were the recruitment staff blind to group assignment?**

  In a good quality trial, the person selecting study candidates is unable to figure out the treatment arm to which they will be assigned if found eligible. This is called "allocation concealment." If the recruiter is not blind to treatment allocation, this can bias the results of a study.

  The researchers did not say whether the person recruiting Arizona counties for the study knew which county would be assigned to the child safety seat

intervention. It would, however, constitute a serious problem of bias if the researcher knew the condition to which a county would be assigned. That would introduce the chance that the recruiter might bias sample selection intentionally or inadvertently, by recruiting counties that give the intervention a better chance of succeeding.

In the Arizona study, the nurse educators who were delivering the intervention also recruited participants into the study, so they were not blind to treatment assignment. This is a concern. However, it is not a major flaw for two reasons. First, the nurse educator did not determine whether the participant was eligible, but instead attempted to recruit everyone she saw. There is a much greater risk of selection bias if the recruiter is required to make a judgment about the participant's suitability for enrollment. Second, since the proportion who did not receive the intervention is low and similar between the two groups, the risk of bias in enrollment is minimal.

- **How was the eligibility assessment conducted?**

  A good quality trial conducts eligibility assessment in such a way as to maximize accuracy. For example, if family members were able to see or hear assessments, the participant may not have felt comfortable answering all questions honestly.

  This might have been a problem if the researchers targeted the intervention only to women who reported that their significant others were unsafe drivers. However, in this study, no eligibility assessment was conducted since all women in the study hospitals were considered eligible.

- **How many participants were selected?**

  A good quality trial provides:

  - A power calculation, showing the number of subjects needed for an expected effect size
  - A reasonable estimate of the intervention's expected effect size

  This study did not report a power calculation. However, since they attempted to enroll all eligible women (rather than selecting a subgroup), who are assumed to be nearly the entire eligible population, power calculations are not critical.

Sample Selection – Validity and

Representativeness

Here are some issues to consider when specifically evaluating the external validity of the sample selection.

- **Is the target population representative of the population to which we want to generalize?**

  A good quality trial states inclusion and exclusion rules to help us decide how well the study population mirrors our own.

- **Is the sample a random (unbiased) representation of the population of interest?**

  Even if the sample matches our population of interest perfectly in terms of demographics such as age and ethnicity, sometimes recruitment procedures lead the study sample to be a select, non-representative subgroup of the population.

  For example...

- **What proportion of the eligible invitees agreed to enroll and were randomized?**

  A good quality trial enrolls a high proportion of participants who are invited and known to be eligible for the study. If a large number of potentially eligible participants refused or were not randomized for some other reason, then the study sample may be a select, non-representative subgroup of the larger population.

  If we were public health officials in a rural U.S. county, we would probably conclude that sample selection procedures in the excerpted study had no detrimental effect on how well their study sample applies to our county. The degree to which a study in rural Arizona is applicable to our county is purely subjective.

For example...

When studies recruit participants through flyers and media announcements, these participants are likely to be different from the general population in the following

ways:

- They may be more public-service-minded.
- They may have stronger motivation to participate in the intervention.
- They may have more free time available for study participation.
- They may have a greater need for incentives provided by the study.

Group Assignment – Excerpt

We just reviewed how sample selection affects a study's validity. Now we consider the impact of how participants get assigned to a study condition. What follows is an excerpt from the methods and results sections of a multi-site trial comparing two family-based approaches to ADHD treatment. The researchers posed the question:

In school-aged children, will a group-based family intervention be more successful than a single family intervention at reducing the level of externalizing behavior, off-task behavior, and family conflict?

- **From the Methods section:**

  Once eligibility was confirmed and consent was received, families were randomized to either the group-based family intervention or single-family intervention arm. Randomization was computer-generated and took place off-site at the coordinating center, which notified both the participants and the project staff of the randomization results. Staff who conducted either intake interviews or outcomes assessment were blind to the treatment condition of the participants. Separate randomization tables were created for families where the index patient was a boy vs. a girl in order to maintain balanced groups.

- **From the Results section:**

  Baseline characteristics are presented in Table 1 and were compared between the two treatment arms. The average family size was larger in the group treatment arm (4.2 vs. 3.8), but no other baseline group differences were found. Therefore, family size was included as a covariate in all multivariate adjusted regression models. The average age of the index child was 11.2, and 69% of the index children were boys.

Group Assignment – Questions

We'll use this study to present some issues to consider when evaluating the internal and external validity of group assignment procedures.

- **Allocation Concealment**

  A good quality trial conceals treatment allocation when initial assessments are being made. As already noted, it is crucial that the person conducting initial assessments is blind to the participant's likely treatment assignment, especially if there are any judgments to be made about the participants' eligibility for the study. If allocation is not concealed, this is a serious concern. Assessment staff may unconsciously evaluate and treat participants differently depending on the condition to which they expect the person to be assigned.

  A strength of the ADHD study is that research staff conducting baseline assessments were blind to treatment allocation.

- **Assignment to Treatment Groups**

  A good quality trial provides enough information about the treatment assignment process to reassure the reader that intake assessors could not influence treatment assignment.

    - For example, a good quality trial explains how the group assignment was generated. Ideally the researcher assigns participants using a random number generator programmed into a computer or a random number table. Sometimes researchers use some other attributes like the person's name or social security number or birth date, or assign them to treatment groups based on the day of the week when the intake was conducted. The concern with these methods is the assignment may not be truly random. In that case, the person conducting the intake interview may be able to tell which group the prospective participant will be assigned to. That, in turn, can introduce bias.

  Group assignment in the ADHD study was excellent, inspiring confidence in the study's internal validity. Independent, off-site, research staff generated the random sequence via computer and kept on-site research staff and participants blind to treatment assignment.

- **Stratified Randomization**

  A good quality trial assesses and controls for differences in study locations and other important variables that may be related to the success of the intervention.

- If there are known differences in demographics between the recruitment locations, it is helpful to stratify participant randomization within levels of these factors. For example...

Since the researchers knew that there would be many more eligible boys than girls, they had separate randomization tables for boys and girls. This helped reduce the likelihood that sex would be a confounding factor. If one group were to wind up with a greater proportion of boys and had a less successful outcome, then it would be difficult to determine if the difference in results was due to the different interventions or the difference in gender distribution in the study arms.

The researchers did not explicitly state whether they had separate randomization tables for each site. However, one purpose of a coordinating center is to ensure high quality, consistent data between study sites. So although we aren't given specific information about whether separate tables were used at each study site, the presence of a coordinating center and off-site randomization is reassuring since it suggests attention to the quality of random assignment.

- **Baseline Characteristics**

A good quality trial tests for differences in the baseline characteristics of the treatment groups and, if found, accounts for these in the study analyses. Commonly examined baseline characteristics include age, gender, ethnicity, socio-economic status, level of severity and treatment history for the condition of interest.

The ADHD study reported baseline characteristics of the groups and found only one, probably minor, difference between them. We are reassured by the lack of differences found on the many characteristics the researchers examined.

For example...

If one location is 70% female and another is 40% female, it can improve internal validity if the researcher develops separate randomization tables for each study site, or separate tables for males and females.

Assessment and Data Collection – Excerpt

Now we will turn our attention to how the assessment and data collection procedures of a study might influence the study's internal and external validity. Mr. Flores oversees services for veterans with chronic mental illnesses in a large Veteran's Administration facility in the Northeastern United States. Due to concerns about lack of improvement among patients in their outpatient day treatment facility, he and his colleagues developed an intervention to improve social functioning in people diagnosed with schizophrenia. They wanted to see if their intervention would do better than usual care at improving social functioning among these patients in outpatient care. Here you see excerpts from the Methods and Results sections of the study.

- **From the Methods section:**

  Assessments were performed at baseline, and at 6, 12, and 18-months post-baseline. Data were collected in face-to-face interviews by staff who were not involved in the intervention and were unaware of the treatment group assignment of the participants. The primary outcome was the Social Functioning Scale, developed and validated on people diagnosed with schizophrenia. The seven subscales of the Social Functioning Scale measure withdrawal, relationships, social activities, recreation, independence competence, independence performance, and employment.

- **From the Results section:**

  There were no differences between the intervention and usual care groups in attrition at any of the followup assessments. Overall, 92% completed the 6-month assessment, 83% completed the 12-month assessment, and 82% completed the 18-month assessment.

Assessment and Data Collection – Questions

We will use the schizophrenia study to discuss how to evaluate the internal and external validity of a study's assessment and data collection.

- **Are the procedures the same for all treatment groups?**

  A good quality trial assesses all treatment groups with the same instruments, in the same settings, in the same way. Any differences in assessment procedures between groups could bias the results. That is, the differences in procedures could be the real reason why the groups differ at the end of the study.

  The schizophrenia study appears to have used the same assessment procedures in the intervention and control groups.

- **Are the assessment staff blind to treatment allocation?**

  A good quality trial ensures that assessment staff are unaware of the treatment group assignment of the participants they assess. Knowing which treatment group a participant is in could subtly bias an interviewer.

  The researchers who conducted the schizophrenia study explicitly state that assessment staff were unaware of patients' treatment group assignment.

- **Did they use valid and reliable assessment methods?**

  A good quality trial provides references or information describing what is known about the reliability and validity of its data collection instruments. It can be a concern if the researcher uses instruments that have not been developed and tested previously.

  Researchers should take special care to describe the validity and reliability of instruments they use to assess behaviors that are complex or difficult to measure, or that have strong cultural influences. The researchers in this study used an established instrument (the Social Functioning Scale) to assess their primary outcome. However, it appears that their assessment tool was a symptom rating scale that requires a trained evaluator. It would have been preferable if the researchers had described their assessment staff's training and interrater reliability with the measurement tool.

- **Is the amount of loss-to-followup reasonable and comparable between the groups?**

  A good quality trial has low overall loss-to-followup (also known as attrition) and similar loss-to-followup between groups. If attrition is high, then we can't adequately assess the intervention's impact on the sample as a whole. Any systematic differences between groups in attrition introduces bias and reduces confidence that group differences at the end of the study are caused by the intervention. For example...

  In the schizophrenia study, attrition was within reasonable bounds. Also the

researchers specifically tested for between group differences in attrition and found none.

For example...

People dropping out of control conditions often have different reasons for attrition than those dropping out of the intervention conditions. Those dropping out of the intervention group may have found the treatment unconvincing or burdensome, or may feel that they have let down the intervention staff by doing poorly. Consequently, those remaining in the treatment and control arms at the end of the study probably differ systematically and in different ways from all who were randomized. That is a very serious and probably unresolvable problem of bias.

Even if reasons for dropping out are similar between the groups, we have to make guesses about the final outcome for participants who dropped out. So, for example, if more participants are lost from the intervention group then we will have less reliable information about that group.

Assessment and Data Collection – External Validity and Representativeness

Consider the following questions when evaluating the external validity of the Assessment and Data Collection:

- Are the assessment procedures so onerous that only a select sample would be willing to participate or complete the study?
- Was there high loss-to-followup? It is problematic if there is excessive attrition. Findings from a sample of "completers" may be less representative of the target population than results based on the full sample. "Excessive attrition" can be a matter of judgment. Most would agree that if 45% or more of the sample lacks followup data, the results are highly suspect. Similarly, most agree that 95% followup or greater is satisfactory and would have few concerns about bias in the results. You can usually assume that those who drop out are different from those who stay in the study in some systematic way, although

the exact nature of the differences is usually impossible to fully characterize. If attrition is in that gray zone between 5 and 45 percent, it is useful to have as much information as possible about differences between those who remained in the study and those who dropped out, to give you an idea of those to whom the results really apply.

If we were considering adopting the new social functioning intervention for schizophrenic patients in our facility, we would conclude that the schizophrenia study's data collection procedures do not impede generalizing to our population. If the study's outpatient sample was similar to patients in our own clinic, we would be satisfied that the study provided a valid test of the intervention for patients like ours. The assessment procedures did not appear too onerous, and although loss to followup was fairly high, it was not so high as to cast significant doubt on the results.

Intervention – Excerpt

Now let's look at a description of a smoking prevention intervention conducted by a state health department. The intervention targeted adolescents. We will discuss factors to consider when evaluating how well an intervention is conducted and described.

Read over this excerpt from the Methods section of the study:

The media campaign ran from October, 2001 through February, 2002. The campaign included radio, billboard, television and internet-based commercials. The campaign contrasted media images that show happy, healthy teens smoking cigarettes with images illustrating the increased risk of death and other negative outcomes associated with cigarette use under the slogan "The Rest of the Story." A website was also developed and its URL was printed on all media images. The website presented information on deception in advertising and tobacco-related mortality in a highly engaging fashion. The themes covered in the website were developed with the aid of experts and focus groups.

Intervention – Excerpt

(continued)

A series of video vignettes was highlighted that showed teenagers describing how their own smoking had interfered with their lives and expressing their anger and sadness that they became smokers. These vignettes were also used in some radio and television advertisements. The website also included a "Lives Lost" calculator, which estimates the number of people dying of tobacco-related causes for a period of time (e.g., number of days or years) entered by the consumer, and the estimated numbers of children and grandchildren affected.

Intervention – Questions

We will discuss the adolescent smoking prevention study when evaluating the impact of intervention procedures on the internal and external validity of a study.

- **Was the intervention clearly described?**

  A good quality trial clearly describes all of the intervention conditions and control conditions. The description should include the overarching treatment model, specific topics covered and/or techniques used.

  - For individual-level interventions, the study should also describe the length and number of treatment sessions, and what kinds of materials were used (treatment manuals, patient hand-outs, video presentations, media used and content of media ads, etc.), who provided the intervention, and what kind of training they received on the study intervention. The setting in which the intervention took place should also be described.

  In the media intervention study to prevent adolescent smoking, the researchers clearly described all components of the intervention, including its dates. Dates can be important information for large-scale public health campaigns because news events or changes in laws and public policies may either enhance or interfere with the success of the intervention.

- **Was the intervention delivered as intended?**

  A good quality trial provides information on whether and how researchers assessed if the intervention was delivered as intended. This is sometimes

known as **treatment fidelity** or **treatment integrity**.

- In individual-level treatment studies, reports that interventionists were observed and rated or supervised provides some evidence that efforts were made to ensure treatment fidelity. If treatment fidelity was poor, then a study may not have provided an adequate test of the intervention.

Treatment fidelity is usually not problematic for media campaigns like that used in the smoking prevention study, although researchers should monitor to ensure that advertisements aired.

- **Was there adequate exposure or participation?**

A good quality trial will provide information on how well the intervention reached the target audience.

- For media campaigns like "The Rest of the Story," some measure of exposure to the campaign is analogous to participant participation or compliance. I.e., how many people saw one of the media announcements or visited a study website?
- In studies involving direct intervention with individuals, researchers should quantify exposure via the average number of sessions completed by participants, and completion of homework, if homework was an integral part of the intervention. If only 15% of participants participated in one important component of the treatment, then the study may not have provided an adequate test of the intervention.

The "Rest of the Story" study did not provide information on exposure, which is a flaw. If a treatment effect is not found, we will not know if it is because people didn't see the ads or if they saw them and did not find them compelling. The researchers could have described the number of billboards per capita, the frequency and duration of television and radio announcements, and the number of hits on the website. They could also have surveyed teenagers in the community to assess what proportion saw or heard about the videos.

Intervention – External Validity

Consider the following questions when evaluating the external validity of the intervention:

Is the intervention feasible in the "real world"?

- For community-wide public health interventions, could a typical community implement the intervention?
- For individual-level interventions, is this something that real-life clients are likely to be willing to participate in? Is it so intensive or logistically difficult that only those with copious free time could manage it?
- Was the intervention conducted in a setting that is similar to yours?

If the context is very different from yours, then the intervention's applicability to your community or clients may be limited.

If we were evaluating the "Rest of the Story" campaign to prevent adolescent smoking in community, we would conclude that it would be feasible to adopt the existing materials and websites. However, we would need more details about the level of exposure achieved in the study to see if that level of exposure would be possible in our state, given our budget.

Data Analysis/Results – Excerpt

Finally, let's use the following vignette to look at how to critically appraise the Data Analysis and Results section of a controlled trial. Ms. Biggs, a guidance counselor at a high school, is concerned about the number of youth she sees with symptoms of an eating disorder. With approval from her principal, she develops a learning module to incorporate into the required health class. She pilot-tested the program by delivering the intervention to two health classes every two weeks for a total of six 20-minute sessions. She used two classes that received no intervention as controls. Here is an excerpt from the Results section of Ms. Biggs' report:

Baseline characteristics were compared between the intervention and control groups using two-tailed t-tests (Table 2). No statistically significant differences were found between the groups at baseline on age, sex, baseline eating disorder symptoms, or body mass index. There were also no differences in followup rate between the two groups at either the 3-month or 6-month followup assessments. Eighty-six percent of the intervention group and 89% of the control group had complete followup.

Data Analysis/Results – Excerpt

(continued)

The two major outcomes, body dissatisfaction and bulimic symptoms, were examined using repeated measures ANOVAs. Analyses were first run only on those with complete data (n=84), and then repeated using the last observation carried forward (LOCF) method to substitute for missing data (n=95). The pattern of results was similar, though the effect sizes were larger in the analysis limited to those with complete data. We present the results from the LOCF analysis here.

For body dissatisfaction, we found an effect for time ($F(2, 92)=122.9$, $p<.001$) and time*treatment ($F(2,92)=13.1$, $p<.001$). Body dissatisfaction declined for all participants over time, but the decline was more pronounced in the intervention group (see Figure 1). There was also an effect of time on bulimic symptoms, but this effect did not differ by group. Both groups showed a gradual decline in bulimic symptoms over the course of the study, and the rate of decline did not differ between the two groups.

Data Analysis/Results – Questions

We will focus here on evaluating the internal and external validity of the data analysis for the body dissatisfaction study.

- **How do the analyses address missing data?**

  A good quality trial has minimal loss-to-followup and uses only conservative, appropriate methods for data imputation (substitution) when missing data are present. Data imputation methods allow you to estimate responses for those with missing data. They can be useful because they allow you to retain more people in your analysis, rather than dropping cases with missing data. If data imputation methods are used, the researcher should discuss how the results differ from those using the original data. A thorough discussion of data imputation methods is beyond the scope of this module, but if they are used, you should think carefully about how they might bias study results.

  Ms. Biggs' study had attrition of about 12%, which is not excessive. Because the followup assessment was a paper-and-pencil test administered in the classroom, missing data were likely due to school absences and not related to the students' experiences with the intervention. This study used the last

observation carried forward method (LOCF), which substitutes the last known observation for missing data. A more conservative approach is to substitute the student's baseline value for missing data, which assumes that the treatment had had no impact, rather than the last observation, which assumes that changes will be sustained. However, since in this case the LOCF analysis showed smaller effects than the "completers" analysis, we can conclude that the imputation methods did not exaggerate the effect size.

- **Is attrition similar between the treatment groups?**

  Differential attrition should be revisited when appraising the quality of the data analysis, especially if data imputation methods are used.

- **Were reasonable data analysis methods used?**

  For most study data there are multiple legitimate approaches to analysis. Some approaches have more limitations than others, and it is useful to have a good grounding in basic statistical approaches.

  In her study, Ms. Biggs conducted repeated measures ANOVAs. In most statistical packages, this analysis technique drops cases that have any missing observations, which is why Ms. Biggs used the LOCF method. Other more flexible statistical techniques can handle missing data better and might have slightly more power to detect group differences, but the ANOVA analysis was acceptable. Ms Biggs appropriately reported F- and p-values and related degrees of freedom.

  Click here to read about two common errors in data analysis.

- **Did the researchers provide enough detail for you to be able to evaluate the strength or size of the effect as well as its direction?**

  A good quality trial provides information about whether the effect could be considered clinically significant or meaningful, or offers benchmarks to help gauge the magnitude of the effect.

  Ms. Biggs did not provide information on the likely clinical significance of the group differences she reported. It would have been useful if she would have reported the average change in the scores for each group and some information about what that amount of change means. For example...

  For further information see, Statistical Significance Testing in Section III of the Randomized Clinical Trials Module.

- **Are there concerns about reporting bias?**

  A good quality trial reports clear aims and has a rationale for selecting the

specific, limited number of outcome measures they chose.

Does it seem as if the researcher examined a multitude of variables and only reported the ones with statistically significant results (also known as cherrypicking)? If many statistical tests were run on a wide variety of outcomes, it is likely that the researchers capitalized on chance associations. In that case, our confidence is reduced that the intervention was the reason for the change in the outcome.

If you are unsure whether the hypotheses being tested in the publication you are evaluating really were the ones the study was originally designed to test, one way to identify the original intention is to find the study at the website clinicaltrials.gov. When researchers register their controlled trials at this site, they must identify the aims and primary outcomes of the study.

Ms. Biggs' study reported on just two outcomes which were directly relevant to answering her research question. Reporting bias is unlikely in this study.

- **What is the external validity?**

  As has already been discussed, high attrition is the main concern related to external validity that comes up during data analysis.

Common errors

Here are some additional examples of two common statistical errors to be aware of:

- **Violating the assumption of independence of observations**
  - This error occurs when the researcher treats each person as if they are completely independent of all other participants when they actually are not, such as when members of the same family are in the same study. Correlated data such as these and other nested designs require special statistical methods, such as random effects modeling or general estimating equations. If participants were recruited from more than one study site (e.g., a clinic, school, or classroom), people at the same site maybe be more similar to each other than they are to the participants at other sites. At the very least, the researcher should check to see if study site is related to the outcome and include it as a control variable if it is. Often researchers will include the clinic or school as a covariate even if it has no statistically significant effect on the outcome, which is preferable to ignoring the nested nature of the data. Ms. Biggs' study made this

data analysis error. She sampled two classrooms in the intervention arm and two in the control arm. Students in a classroom probably responded more similarly to each other than to those in another classroom, but Ms. Biggs' analysis treated each student as completely independent of others. This raises a concern that the intervention effect might become nonsignificant if the analysis appropriately controlled for the clustering of responses within a classroom.

- **Failing to control for baseline differences between groups**
  - Researchers should test to see if the treatment groups differ on important demographic and clinical variables at baseline. If differences are found, the researcher should control for these variables in some way, such as by including them as covariates. The researchers in our example reported only a few baseline characteristics, but the ones they chose were appropriate and it is reassuring that the groups did not differ on them.

For example...

From the graph it appears that the body dissatisfaction scores of intervention participants lowered by about seven points, compared with a reduction of about four points in the control group. To convey what these changes mean clinically, Ms. Biggs could describe two students who began the trial at the mean body dissatisfaction score and whose symptoms decreased by 7 and by 4 points. Or, she could provide examples of how a student's responses on the items on the body dissatisfaction scale could yield a 7- and a 4-point change. Examples of this kind usually may be found in the paper's Discussion section.

III. Time Series Research Designs

What are Time Series Research Designs?

The defining feature of time series research designs is that each participant or sample is observed multiple times, and its performance is compared to its own prior performance. In other words, each participant or population serves as its own control. The outcome—depression or smoking rates, for example—is measured repeatedly for the same subject or population during one or more baseline and treatment conditions.

When the researcher studies only one or a few individuals, these are called Single Subject Research Designs (SSRD). They are particularly useful when:

- Few participants are available—problems with low incidence rates, for example
- Participants are relatively heterogeneous
- Participants demonstrate variability from day to day

When the researcher studies an entire population, such as a community, city, or health care delivery system, these are interrupted time series designs (ITSD). They are particularly useful when you want to evaluate the effects of a law, policy or public health campaign that has been implemented in a community.

We will examine both of these designs together because they share many similar considerations.

Basic Terms

In order to critically examine the quality of a time series study, it is helpful to understand some of the basic terms.

- **Baseline**

  Baseline refers to a period of time in which the target behavior or outcome (dependent variable) is observed and recorded as it occurs before introducing a new intervention. The baseline behavior provides the frame of reference against which future behavior is compared. In some designs, the term baseline can also refer to a period of time following a treatment in which conditions match what was present in the original baseline.

- **Treatment Condition**

  Treatment condition or treatment phase in these designs describes the period of time during which the experimental manipulation is introduced and the target behavior continues to be observed and recorded.

Design Types

A variety of study designs can be considered time series. The defining feature is that individuals or a population is measured many times before and after the intervention is introduced. Let's examine three types of time series designs, two that are used mostly with individual participants and one that can be used with either individuals or a population.

Design Types

Many variations of the basic time series design are possible.

NOTE: A variety of study designs can be considered time series. The defining feature is that an individual or a population is measured many times before and after the intervention is introduced. Here you see three types of time series designs, two that are used mostly with individual participants and one that can be used with either individuals or a population.

- **Interrupted Time Series Design**

  The simplest type of time series design is the interrupted time series. This design is typically used to evaluate the impact of a population-wide policy or intervention. It involves a single treatment group which is measured many times before and after the start of the intervention. It is called "interrupted" time series because the researcher graphs the data before and after the intervention, and looks for an interruption in the line or curve where the intervention was introduced. The interruption could be a change in:

  - Level or height of the curve (as in Figure TS-1a, which shows that

midway through 1993 the prevalence of smoking suddenly fell to a new, lower level)

- Trend or slope of the curve (as in Figure TS-1b), which shows that the prevalence of HIV in a very high-risk population rose by about 1% every 2 years in the 8-year period between 1985 and 1993 before the intervention was implemented. After the initiation of the intervention, the rate of increase in HIV prevalence slowed substantially: the entire next 7-year period showed less than a 1% increase in HIV prevalence

To be able to see the natural pattern of the behavior change, it is crucial to this type of design to have a sufficiently long period of baseline observation that includes usual seasonal or cyclical changes.

- **Reversal/ABAB Design**

  Another general family of time series designs are ABAB or reversal designs,. These are usually used with a single participant or a few individuals. The "A" refers to the baseline and "B" refers to the treatment. In this design the treatment is systematically provided and withdrawn. After a stable baseline is established (A), whereby the target behavior remains relatively constant, the treatment or intervention is provided (B). If the treatment is successful, we would expect to see a change in the target behavior in the predicted direction. After a period of treatment, the intervention is withdrawn. At that time the target behavior is expected to return to baseline levels (A). Treatment is once again initiated (B), and, if successful, should initiate a similar treatment response.

  This design provides an opportunity to repeatedly demonstrate, or replicate, the relationship between the treatment and outcomes of interest with a single participant. Figure SSRD-1 shows an example of this type of design. The panels show the effect of holding a favorite toy on the verbal utterances of an autistic boy. Without the toy, the boy does not speak. When he is given the toy, he speaks; when the toy is taken away he ceases talking, and so on.

- **Multiple Baseline Designs**

  Other time series studies can be characterized as multiple baseline designs. Reversal designs are not always feasible. Once treated, a behavior may not deteriorate back to baseline after treatment stops. Ethically, once a patient has responded to treatment, it may be considered too harmful to discontinue it. The multiple baseline design incorporates a baseline and an intervention condition (an A and a B) across multiple participants, behaviors, or contexts. The greater the number of replications, the more confident one can be that the treatment produced the observed changes.

Sample Selection – Excerpt

Now that we've covered some background on time series designs, let's discuss some issues unique to the critical appraisal of time series designs. Information covered under controlled trials also applies to many time series studies, so we won't revisit areas we've already covered.

Let's begin by talking about how sample selection can affect the internal and external validity of time series studies.

A team of clinicians who direct child and family services at an agency serving an immigrant Hispanic population in New York have observed that many of the mothers who bring their children in for treatment seem to struggle with depression. The team would like to begin providing services for these women. They are particularly concerned that the intervention be culturally relevant to their clients. The following is an excerpt describing the sample selection in a depression treatment time series study that one of the team members found and brought to the team's weekly supervision meeting.

The study participants were 3 Hispanic women between the ages of 22 and 39. The women were drawn from a population of clients who were receiving services from a community mental health center situated in a low-income neighborhood in a large urban metropolitan area in the Florida. To be included in the study, the women had to speak English, meet DSM-III-R criteria for a current episode of Major Depressive Disorder, and not currently meet criteria for alcohol or substance abuse /dependence disorder or other psychiatric disorders. When they were recruited into the study, each of the women was being treated by the same clinical social worker.

Sample Selection – Target Population

We'll use this study to discuss how to evaluate the internal and external validity of the sample selection.

- **Is the target population identified?**

  A good quality time series study will clearly define the target population ahead of time. This increases the chances that participants were selected because they are typical of the target population, rather than being selected for

idiosyncratic reasons such as openness to trying new treatment approaches.

In the excerpted study, the target population was described as Hispanic women from a low-income neighborhood in Florida. The New York family service workers also work with Hispanic immigrants chiefly from Puerto Rico. However, they wonder whether the women in the Florida sample were also immigrants and parents. They also worry that the women from Florida might be of Cuban or Mexican origins, rather than Puerto Rican. These factors could influence whether the Florida intervention fits the needs of the New York population.

- **How many participants are included?**

  In time series designs, participants serve as their own control or comparison, so traditional power calculations are not applicable. In this study, confidence that any change in depression symptoms is due to the intervention increases when the effect replicates consistently across multiple participants. In population-based studies a community may be considered one participant.

  Some studies include only a single participant; our clinical example study evaluated three.

- **What are the inclusion and exclusion criteria?**

  Just as in experimental designs, inclusion and exclusion criteria help to describe how participants were selected into the study. A clear description of these criteria increases the likelihood that participants were selected because they are representative of the target population, rather than for idiosyncratic reasons which will not be replicable.

  The Florida example study clearly specified some inclusion and exclusion criteria regarding ethnicity, diagnosis, and geographic setting.

- **How were individual participants chosen?**

  Were participants chosen at random or through some other method (e.g., convenience)? Non-random methods have the potential to introduce bias, especially when study samples are very small.

  In our Florida study, it isn't clear how the three women were specifically chosen, except that all were patients of the same social worker. If they were not randomly selected out of a larger number of eligible participants who also met inclusion and exclusion criteria, we should be concerned about the potential for bias.

  Population studies sometimes use summary statistics for an entire community to inform policy or public health initiatives for that population. However, some population-based studies do select a sample from the population to help assess

the impact of an intervention. In these cases, random selection of a sample improves the likelihood that the sample is representative of the population. In addition, it is helpful if the researcher compares the characteristics of the study sample with those of the larger population to help determine how well the sample reflects the community.

Group (Factorial) Assignment in Single Subject Time Series Designs

Now that we've discussed how sample selection affects the quality of a study, we will briefly explore group or factorial assignment in time series designs.

- Classic single subject time series studies involve a single person, a single behavior, a single setting, and a single treatment that is intermittently present or absent
- However, parameters may be systematically varied to test whether a treatment effect replicates across a group of subjects, a group of settings, and several outcomes. For example...

For example...

Participants may be randomly assigned to one of several different presentation sequences of treatment and no treatment. Or the onset of the treatment for any participant may be chosen randomly to occur at one of several different dates. Figure SSRD-2 displays such a design. If the treatment effect replicates across all sequences or all onset dates, then that provides stronger evidence that the outcome is actually linked to the treatment rather than to an extraneous variable or event.

Group (Factorial) Assignment in Single Subject Time Series Designs

A good quality time-series study using group assignment:

- Uses a random process for assigning individuals or communities to conditions
- Clearly describes the process and level of randomization (person, classroom, county, etc.)

Assessment and Data Collection – Excerpt

A group of legislators are debating the renewal of a state seatbelt law. They commissioned a report to assess the effect of the initial seatbelt legislation. To begin our evaluation of the assessment and data collection, read over the following excerpt from their report on the effect of seat belt laws for preventing traffic injuries and deaths.

Monthly accident and fatality statistics are available from the state. Data on the number of accidents, deaths, and injuries broken down by level of severity were compiled for the period of January 1981 through January 1991. The date range was selected to provide an adequate number of observations before a mandatory seatbelt law was put into effect in 1986. Data before 1981 were not used because that is when the minimum drinking age was raised to 21, which has been demonstrated in other studies to have an impact on accidents and fatalities. For comparability, the pre- and post-intervention observation periods were both 5 years.

Click here to view the Fatalities Per Million Vehicle Miles Traveled chart.

Assessment and Data Collection – Questions

Most assessment and data collection issues involving time series designs are very similar to those already discussed for controlled trials that involve assessment of individuals. That discussion will not be repeated here. Some new issues can be especially important in time series designs.

- A good quality time series study uses an adequate number of data points in each phase (baseline and intervention) for each participant or community. The number of assessment points is adequate if the outcome measures have attained stability.

  In the sample study, the researchers provided a rationale for the years of data they collected, including an effort to establish a stable baseline with five years of data. In many cases, five observations might not be sufficient for data to stabilize. In this case, though, traffic fatalities are relatively stable, so five data points were sufficient.

- A good quality time series study provides sufficient detail about data collection procedures. Information about quality assurance methods is especially important when secondary data sources are used.

  Sometimes population-level studies are dependent on data sources that were created for other purposes, such as tracking traffic accidents. In such cases, researchers have little control over the collection of the data. Therefore the data are usually less trustworthy than data collected specifically for research purposes.

  An additional concern with using pre-existing data sources is that sometimes reporting systems change. When this happens it can appear as if the outcome is changing when it actually is not. For example...

  In the sample study, the authors provide little information about the quality of the data. It would be preferable if they described who was responsible for populating the data, whether clear instructions are provided, and what quality assurance checks are performed on the database.

For Example...

For example, a city may have implemented a new procedure whereby automobile

accidents can be reported online. If this makes it easier for drivers to report their accidents, then the data may show an increase in accidents when there was actually only an increase in the reporting of accidents.

Data Analysis/Results – Excerpt

Now that we've discussed topics related to assessment and data collection in time series designs, we turn to data analysis and results.

An advocacy group made up of health care practitioners and restaurant workers in a Midwestern state has been asked to testify in front of a legislative body that is debating a ban on smoking inside restaurants. They have collected studies examining the impact of similar bans in other states and cities.

Data Analysis/Results – Questions

Consider the following issues when evaluating the internal and external validity of the intervention.

- **Predictions**

  A good quality time series design specifies hypotheses about the expected intervention response in targeted outcomes. For example, the researchers may predict:

  - Direction of the change;
  - How soon response to the intervention will be seen; or
  - Which targeted behavior will show the greatest responsiveness.

  The more the results reflect a priori (pre-stated) predictions, the more confident we are that the results reflect the impact of the intervention and not

some other factor.

In this study the authors hypothesized that the ban would result in a reduction in the heart attack rate, which was confirmed by the results.

- **Visual Analysis**

  A good quality time series design:

  - Reports appropriate visual analysis that addresses level, trend and variability
  - Clearly labels x- and y-axes, indicates baseline and treatment phases with vertical lines, and uses consistent scales that accurately reflect the variability in the data

  Level refers to the frequency or intensity of the behavior in a particular phase of the study. In some studies the mean of each outcome is summarized for each baseline and treatment study phase.

  Trend refers to a pattern of change observed across data points. For example, if the outcome targeted by the intervention trends toward improvement during the baseline phase, before the intervention is introduced, this suggests that the outcome might have improved even in the absence of treatment.

  Variability refers to the consistency of the behavior or outcome. If it is highly variable during the baseline or treatment phases, stability may not have been attained, making it hard to determine the treatment response.

  In our sample study...

- **Statistical Analysis**

  The application of statistical analysis in time series studies is relatively new. Many existing studies do not use statistics to analyze results. However, statistical procedures may be particularly helpful in certain contexts. Examples of statistics used with time series studies include ARIMA models, celeration line approach, two-standard deviation band method, C-statistic, or other.

  Visual analyses are subjective as there is no standardized procedure for this approach to data analysis. Instances in which visual outcomes can be especially difficult to evaluate include studies where a stable baseline is not established, tests of new interventions where treatment effects are relatively unknown, or cases where treatment effects may be delayed or subtle.

  The use of both visual and statistical analyses enhances the quality of the smoking ban study. The advocacy group has some evidence that a smoking ban

may have an impact on heart attacks.

In our sample study...

In our sample study the authors depict heart attack rates during the baseline and treatment phases of the smoking ban. In addition they present variability and trend information using a reasonable scale and connect the monthly measured rates of heart attack admissions with a trend line. The figure could have been improved with the addition of a vertical line to indicate where the intervention was introduced.

Visual inspection of the data suggests some variability, but there is no apparent trend in the hypothesized direction before the smoking ban is implemented. If we had seen a trend toward a reduction in heart attacks before the ban was in place, we would worry that the reduction was due to a factor other than the ban.

IV. Systematic Reviews

What are Systematic Reviews?

Systematic reviews (also known as systematic evidence reviews) use rigorous and clearly documented methods to:

- Define a research question
- Search for relevant literature
- Select studies that help address the research question
- Combine the selected studies to answer the research question

Meta-analysis may be used to combine studies statistically. It is not always possible to combine studies statistically, however. Sometimes the studies are too different from each other to combine meaningfully. In such instances, the studies are combined in a narrative, qualitative fashion.

For more detail about how systematic reviews are conducted see the EBBP module on

systematic reviews at http://www.ebbp.org/training.html.

Defining and Searching for Relevant Literature – Excerpt

Mr. Cohen is the police department liaison for the development commission in a medium-sized city in Washington. The commission is trying to develop a plan to address crime in residential areas of the city. Mr. Cohen and the police chief wonder if neighborhood watch schemes may be an effective way to reduce crime in several neighborhoods. To answer this question, they conducted a systematic review to assess the evidence about whether neighborhood watch programs are effective in reducing crime.

Here is an excerpt describing Mr. Cohen's process of searching for studies that meet the eligibility criteria for the review.

Eligible studies evaluated neighborhood watch schemes. An intervention was considered to be neighborhood watch if the following elements were present: (1) residents provide surveillance of the neighborhood; (2) residents report suspicious behavior to the police or a neighborhood coordinator; and (3) residents are expected to engage in cooperative problem-solving (rather than having a single person make all decisions and assign tasks). Comparative cohort and controlled trial designs were eligible for inclusion.

A librarian experienced in criminal justice designed the search strategy and searched the literature from 1985 through July 1, 2006 in the following electronic databases: International Bibliography of the Social Sciences (IBSS), Criminal Justice Abstracts, National Criminal Justice Reference Service Abstracts, Sociological Abstracts, Psychological Abstracts (PsycINFO), Social Science Abstracts, Government Publications, and Dissertation Abstracts (ASSIA), using the specific search terms. In addition to searching electronic databases, the authors manually reviewed reference lists from pertinent review articles and queried experts in the field.

Sample Selection – Questions

Read over the following issues to consider when evaluating the internal and external validity of the sample selection.

- **Is the relevant body of literature clearly defined?**

  A good quality systematic review clearly describes through detailed inclusion and exclusion rules what kinds of studies will be used to address the question. These rules specify study characteristics, such as population characteristics, interventions, comparison groups, outcomes, study design, publication date, and the required duration of follow-up. If those rules are not clearly defined (usually by a protocol) then two immediate problems can result:

  - Searching may not adequately locate all studies that are in the relevant body of literature
  - Articles whose characteristics fall into a gray area may be handled inconsistently or in a biased manner when it comes to deciding whether a study should be included in the review. For example...

  Mr. Cohen's review protocol specified required intervention elements and the research design types they would consider. They did not mention limits regarding language (can they translate non-English language publications?) or location of the study (are studies conducted outside of the U.S. within their scope of work?), or what types of comparison conditions are acceptable. It would have been preferable to have these factors specifically described.

- **Are search methods adequate?**

  A good quality systematic review:

  - Searches all relevant databases. Usually, more than one database is needed to capture all relevant studies. In addition, there should be evidence that the investigators utilized other methods such as hand-searching key journals and bibliographies of relevant articles and contacting researchers in the field to identify studies.
  - Provides information about the specific search terms used. Searching electronic databases can be a high-level task requiring specialized knowledge of databases and how indexed terms are used in each one. Input from a librarian can be extremely helpful, in many cases essential.

  Mr. Cohen's systematic review searched a number of relevant databases, used a specially-trained librarian to develop the search strategy (thus increasing the odds that the appropriate terms were selected and used correctly), and reported additional, non-electronic search strategies. All of these instill

confidence that the searching process was systematic and comprehensive.

- **How is the external validity affected?**

  As described under "Search Methods," the roster of databases searched should be as comprehensive as possible. Narrow searches can bias the results by excluding studies conducted by researchers from certain disciplines or geographic regions. Study inclusion and exclusion criteria also have a direct impact and should be carefully reviewed. If inclusion and exclusion rules are not clearly specified, then readers should carefully examine the characteristics of the individual trials for information about how the study samples reflect the population of interest to them.

For example...

If the criteria aren't clear (or aren't consistently applied), researchers may be more likely to include articles from top-tier journals or well known researchers. There are many good-quality studies published in lower-tier journals and by less-known researchers, and those studies should be given equal consideration.

Selecting Studies and Collecting Data – Excerpt

Now we will move to the next phase of a systematic review, which involves selecting the studies to be included and abstracting data from the included studies.

Read the following excerpt from a systematic review exploring the efficacy and effectiveness of Activation and Commitment Therapy for improving quality of life in people with chronic pain. It describes how studies were selected and how data was extracted from the studies:

Two raters reviewed 4,239 abstracts against inclusion and exclusion criteria, and pulled articles if either rater coded it as possibly meeting criteria. Two staff members then reviewed the resulting 327 articles to determine whether an article met inclusion/exclusion criteria. Disagreements that could not be resolved by discussion between the two reviewers were discussed at a team meeting where a final decision

was made. Next, articles were assessed for threats to validity and rated as "good," "fair," or "poor." Studies rated as poor were excluded from the review. Data were then abstracted using a standardized data abstraction form by one team member and checked for accuracy by another.

Selecting Studies and Collecting Data – Questions

Consider the following issues when evaluating study selection and data collection.

- **Are the procedures to evaluate abstracts and articles for inclusion in the review clear and free of apparent bias?**

  In a good quality systematic review, abstracts and articles are reviewed by at least two people to ensure that no important studies are excluded and that inclusion/exclusion criteria are applied accurately. Optimally, articles describe whether reviewers were working independently (i.e., blind to the other reviewers' results) and how disagreements were resolved when they arose.

  Some reviews use a blind assessment procedure whereby author and journal information are stripped from the copy of the article being assessed. This further reduces the risk of bias, but is not necessary in most cases, particularly if inclusion and exclusion criteria are well detailed.

  The Activation and Commitment Therapy systematic review provided adequate detail regarding the process of deciding which studies to include, including the numbers of abstracts and articles reviewed.

- **Were studies assessed for quality?**

  A good quality systematic review:

  - Critically appraises the quality of the included studies and reports the results
  - Describes how information on study quality has been used in the review. Some researchers exclude poor quality studies and others conduct analyses to see what (if any) effect study quality has on the treatment effect. Either method can be appropriate.
  - Describes the specific threats to validity they evaluated, or provides references for instruments they used to assess study quality. A

tremendous amount of information is available about assessing study quality. Some quality rating instruments have been developed and may be used. Other reviewers prefer to develop quality rating instruments specific to their review question.

The authors of the Activation and Commitment Therapy systematic review failed to provide information about what aspects of study quality they examined. They should have either provided an appendix listing the threats to validity that they assessed, or a reference if they used methods described elsewhere.

- **Are the procedures used to assess the quality of the articles clear and free of apparent bias?**

  Ideally, reviewers will be blind to other reviewers' assessment. This is particularly important for the assessment of quality, which involves some subjective judgment. For that reason, quality assessment is prone to bias or to being influenced by knowledge of other reviewers' conclusions. Ideally the researchers will also report procedures for resolving disagreements in quality ratings.

  The description of the quality assessment process in this review is weak. The reviewer does not say whether quality was assessed by more than one person for each study, much less whether the quality assessment was conducted with or without the knowledge of others' ratings.

- **Is data abstraction checked/verified?**

  A good quality systematic review clearly describes data quality assurance measures, such as having all or some data elements checked by a person other than the person who abstracted the data. This is especially important for primary outcomes and any data used in meta-analysis. It is very easy to make transcription errors, especially when there is a large volume of data.

  The authors appropriately described their data abstraction process as involving a standard form, and it is reassuring that data abstraction was checked by a rater other than the abstractor.

Data Analysis/Results

Systematic reviews synthesize the results of all the studies and provide a narrative

description of the synthesized results. If the studies are reasonably similar, the reviewer may combine the studies statistically in a meta-analysis, in which each study is a single observation in the analysis. A detailed description of meta-analysis is beyond the scope of this module; however, some basic principles that will apply to most meta-analyses will be discussed.

Data Analysis/Results – Excerpt

Now read over excerpts from the Methods and Results sections of a meta-analysis. This study examined the efficacy of physical activity interventions for preventing falls and injuries among adults living in assisted care settings.

- **From the Methods section:**

  The primary outcomes of interest were the proportion of participants who fell during the followup period and the proportion of participants who had a fall-related fracture. Because of the high degree of heterogeneity in patient populations, intervention approach, setting, and method of assessing the outcomes, we used random effects models for all meta-analyses rather than fixed effects. We analyzed the relative risks (RR) of falls and fractures. If the study did not provide the RR or the related standard error (SE), we calculated the missing statistics from raw data (i.e., numbers of events in each group and numbers randomized to each group). If they provided both unadjusted and adjusted RRs, the RR from the adjusted model was used. We calculated the I^2 statistic to assess statistical heterogeneity of the trials in each analysis.

- **From the Results section:**

  Six trials assessed fall-related fractures and found that physical activity interventions did not reduce the probability of a fracture. The pooled relative risk was 0.83 (95% CI 0.61, 1.14) with low heterogeneity (I^2 = 16.3% percent). However, the thirteen trials reporting the proportion of participants with a fall found that physical activity interventions did reduce the probability of falling (RR=0.76, 95% CI 0.63, 0.93). Heterogeneity was low for this analysis as well (I2=5.2%).

Data Analysis/Results – Questions

Now we will walk through a number of questions to consider when evaluating a meta-analysis.

- **If studies were statistically combined in a meta-analysis, were they sufficiently alike to warrant doing so?**

  Three types of heterogeneity should be evaluated: clinical (variability in factors such populations, interventions, outcomes, settings), design (variability in the design features such as presence of random assignment and time to followup), and statistical. The first two are evaluated subjectively by the reviewer, and the third is evaluated using statistical tests.

  A good quality review:

  - Only combines studies statistically that are reasonably similar (What does "reasonably similar" mean?)
  - Uses statistical models that account for the degree of heterogeneity that is present (More on heterogeneity)
  - Provides a measure of statistical heterogeneity, such as I^2 (In this study...)
  - Provides enough information to the reader to evaluate how similar the studies are

  If the studies in a review are very different from each other, then an average may be meaningless. For example...

- **How many studies were included in the analysis and what was the quality of the studies?**

  Higher quality studies yield more trustworthy results. A large number of studies is also usually preferable to a smaller number, though there are some exceptions. Two to three very good quality, large trials would usually provide more trustworthy results than 10 small, fair-quality trials.

- **Do they describe which outcome(s) they chose?**

  A good quality systematic review:

  - Describes which specific measures were accepted as outcomes and which statistics were entered into the meta-analysis (e.g., relative risks, odds ratios, mean differences)

- Reports formulae, if they calculated outcome measures that were not reported in the individual studies
- Reports which outcomes were chosen if a study reported more than one relevant outcome, such as self-report vs. medical chart data, any fracture vs. only non-vertebral fractures, and falls in 6-month vs. 12-month window

- **Do they describe how they handled studies with three or more treatment arms?**

  Designs with multiple intervention and/or control groups are very common. Reviewers should not simply include multiple comparisons in the same meta-analysis as if they were from different studies. Two commonly used strategies are:

  - Select the intervention arm with the highest intensity
  - Take an average of the data across treatment arms

- **Do they provide a visual display of the data?**

  If they have multiple studies with similar outcomes, systematic reviews commonly include forest plots (see Figure SR-1 for an example), which show the average effect size for each study along with "whiskers" that show the 95% confidence interval. The size of the symbol showing the effect size is usually proportional to the size of the study. If the studies are sufficiently similar to warrant combining them statistically, then an average effect size is usually presented as well, either for the entire group of studies or for logical subgroups of studies. The plots help the reader understand how consistent the results are and whether one or two studies are exerting a strong influence on the summary statistics.

  The authors of the falls prevention review did not provide forest plot, despite having enough data to run a meta-analysis.

- **Are there concerns about reporting bias?**

  A good quality systematic review describes a priori a limited number of outcomes they will report.

  Sometimes an outcome is reported in only a few of the studies included in a systematic review. When that occurs, the possibility is raised that this outcome is reported selectively: i.e., it is more likely to be reported when the results are statistically significant. This review limited itself to two primary outcomes: falls and fall-related fractures, which reduced the risk of reporting bias.

- **Are there concerns about publication bias?**

  Publication bias occurs when studies with statistically significant results are more likely to be published than those without statistically significant results. A

good quality systematic review reports what they did to examine the potential for publication bias and whether there is any reason for concern. Certain ways of plotting the data (such as a funnel plot or L'abbe plot) are useful for assessing the likelihood that positive studies in an area were more likely to have been published. A funnel plot should have been provided or described for the falls prevention review. The plot could have helped the reader gauge the likelihood that unidentified, unpublished negative trials exist but were not captured by the review.

- **Are the reviewer's conclusions warranted by the data?**

  Sometimes a reviewer over-generalizes the findings of the review or uses statistical methods that exaggerate the size of the effect. In this example, the reviews report only relative risk statistics. So, we know that the risk of falling is 17% lower for people engaged in falls prevention interventions, but we need to know the absolute difference in risk in order to put the results in perspective. Reporting on relative risks can be misleading.

- **What is the impact on external validity?**

  The methods used for analyzing data have a limited impact on external validity. Readers may want to look for analyses of subgroups of the trials whose samples best match their population of interest. For example, a systematic review may include studies conducted in both residential and outpatient settings. If you work in a residential setting, then a meta-analysis limited to the subset of trials performed in residential settings may provide the more relevant information for your context.

For example...

For example, a fall prevention intervention involving only vitamin D supplements would likely have a very different effect than one involving comprehensive assessment and redesign of the home environment. The average of the two studies would be unlikely to describe the effect in either study accurately.

Reasonably similar

"Reasonably similar" is a judgement call. The reviewer should ask him or herself, "Would the average effect of these studies be meaningful or useful?"

Heterogeneity

Random effects models are most appropriate when clinical, design, or statistical heterogeneity is substantial. These models assume there is heterogeneity in the studies to be combined. To their credit, the reviewers did use random effects models for their analysis.

In this study...

The levels of statistical heterogeneity were low. The combination of heterogeneity in study features coupled with homogeneity in study effects is unusual and raises a concern. Might publication bias—the tendency for only positive results to be published—explain the discrepancy?

V. Appraising Guidelines and Recommendations

Critical Appraisal of Guidelines and Recommendations

Many professional groups conduct their own assessment of the literature in a particular area in order to develop treatment guidelines or recommendations. It is tempting to accept these guidelines at face value and assume that the experts have done an exemplary job of weighing the evidence. However, the process of developing

guidelines can vary considerably in rigor and objectivity. Guidelines, just like individual studies, must also be critically appraised.

Critical Appraisal of Guidelines and Recommendations

Guideline developers:

- Weigh the relative likelihood and magnitude of the benefits and harms of treatment
- Assess the overall quality of the body of evidence in order to provide an assessment of the strength of their recommendation

There are very few cases in which the evidence is clear and straightforward. Most bodies of literature show mixed evidence and the relevant studies demonstrate a range of limitations. In a good-quality guideline development process, the quality of the evidence is absolutely integral to the recommendation.

Grading of Recommendations Assessment, Development, and Evaluation (GRADE)

The GRADE approach to presenting and developing recommendations is widely used and provides a framework for critical appraisal of recommendations. The GRADE working group website (http://www.gradeworkinggroup.org/index.htm) provides links to publications describing the GRADE system, including a series of articles published in the British Medical Journal in 2008.

VI. Conclusions

Conclusions

The way in which a study is designed, conducted, and analyzed can have a huge effect on the outcome. Poor quality studies may overestimate or underestimate the size of an effect. We rely on researchers to write up their methods accurately, but the truth is that research is much messier than what can be captured in a journal article. The small idiosyncrasies are generally not reported, and even under the best of circumstances, we humans are messy creatures and we make mistakes.

On the front lines of research:

- Interviewers forget to probe for important details
- Participants don't accurately recall details of their lives
- Data entry personnel make keystroke errors
- Investigators make inconsistent decisions about how to handle anomalies

Conclusions

The methods and results of published articles tell us the broad strokes, and, assuming things really went as reported most of the time, they give us a sense for how strongly we should trust the findings of the study. Sometimes we find clues buried in publications that hint at greater difficulties, and as consumers we always need to be vigilant for those times when the data do not seem to reflect the real truth of the matter. We rely on replication and consistency between studies to demonstrate that an intervention really is effective, but those data are only meaningful if the quality of the studies is acceptable.

As busy practitioners, we are sometimes tempted to jump straight to the discussion section of an article for the bottom line, but it is imperative that we read studies with a critical eye if we are going to make policy or treatment decisions based on the results of research.

Resources

Resources

Further Reading:

- Greenhalgh, T. How to Read a Paper: The Basics of Evidence-based Medicine. Third Edition. Wiley-Blackwell. 2006
- Guyatt G, Rennie D, Meade MO, Cook DJ. Users' guide to the medical literature: A manual for evidence-based clinical practice, Second Edition. McGraw Hill Medical. 2008
- Rubin A. Practitioner's guide to using research for evidence-based practice. Wiley. 2008
- McMillan, J. H. (2004). Educational Research: Fundamentals for the Consumer, 4th Edition. Allyn and Bacon: Boston.
- Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1)

Critical Appraisal Checklists

- NICE Manual Guide 2009 - Appendix C: Methodology checklist: Systematic reviews and meta-analyses
- NICE Manual Guide 2009 - Appendix D: Methodology checklist: Randomized controlled trials
- Additional checklists can be found at the EBBP critical appraisal webpage

Glossary

Glossary

- **Allocation concealment**

When neither the participant nor the person assigning the participant to a study arm knows which study arm the participant will be randomized into until the moment of randomization. Any baseline assessments should take place either before randomization, or by a staff member who does not know the treatment group to which the participant has been assigned to. Lack of allocation concealment can introduce selection bias.

- **Autoregressive Integrated Moving Average (ARIMA)**

A data analysis technique appropriate for time series data, where autocorrelation (successive or seasonal observations that are serially dependent) is likely. This approach is particularly appropriate to identify significant shifts in data associated with policy or other population level interventions, independent of the observed regularities in the history of the dependent variable. The ARIMA model describes the stochastic autocorrelation structure of the data series and, in effect, filters out any variance in a dependent variable that is predictable on the basis of the past history of that variable.

- **Autoregressive Integrated Moving Average (ARIMA)**

A data analysis technique appropriate for Time series data, where autocorrelation (successive or seasonal observations that are serially dependent) is likely. This approach is particularly appropriate to identify significant shifts in data associated with policy or other population level interventions, independent of the observed regularities in the history of the dependent variable. The ARIMA model describes the stochastic autocorrelation structure of the data series and, in effect, filters out any variance in a dependent variable that is predictable on the basis of the past history of that variable.

- **Baseline (in single subject research designs)**

The period of time in which the target behavior (dependent variable) is observed and recorded as it occurs without a special or new intervention. Can also refer to a period of time following a treatment in which conditions match what was present in the original baseline.

- **Baseline (in controlled trials)**

This initial assessment that takes place before the intervention has been implemented.

- **Bias**

Influences on a study that can lead to invalid conclusions about a treatment or intervention. Bias in research can make a treatment look better or worse than it really is. Bias can even make it look as if the treatment works when it actually doesn't. Bias can occur by chance or as a result of systematic errors in the design and execution of a study. Bias can occur at different stages in the research process, e.g. in the collection, analysis, interpretation, publication or review of research data. Some commonly referred to types of bias are:

- **Non-response bias.** When participants who do not participate in a study or complete follow-up assessments are systematically different from those who do.
- **Performance bias.** Systematic differences in care provided apart from the intervention being evaluated. For example, if study participants know they are in the control group they may be more likely to use other forms of care; people who know they are in the experimental group may experience placebo effects, and care providers may treat patients differently according to what group they are in. Masking (blinding) of both the recipients and providers of care is used to protect against performance bias.
- **Publication bias.** Studies with statistically significant results are more likely to get published than those with non-significant results. Meta-analyses that are exclusively based on published literature may therefore produce biased results. This type of bias can be assessed by a funnel plot or L'abbe plot.
- **Recall bias.** When the study groups are systematically different in their ability to recall events that are key to assessing the effect of the intervention.
- **Reporting bias.** When studies are more likely to report, selectively, outcomes that are statistically significant.
- **Selection bias.** Selection bias has occurred if:
    - the characteristics of the sample differ from those of the wider population from which the sample has been drawn, OR
    - there are systematic differences between comparison groups of patients in a study in terms of prognosis or responsiveness to treatment.

- **Blinding (a.k.a., Masking)**

The practice of keeping the research staff or participants of a study ignorant of the group to which a participant has been assigned. For example, a clinical trial in which the participating patients or their doctors are unaware of whether they (the patients) are taking the experimental drug or a placebo (dummy treatment). The purpose of 'blinding' or 'masking' is to protect against bias. Unless a placebo medication is involved, it is usually impossible to blind the interventionist and participants to the group assignment of the participants, but it is critically important that staff conducting the assessments be blind to the group assignment.

- **Celeration line approach**

Also called the split-middle method of trend estimation. The procedure is designed to identify the trend of the data. A trend line is computed using data in the baseline phase. This line is then extended to the treatment phase to evaluate the effect of intervention on the subject's performance. The proportion of data points above and below the trend line is compared from the baseline to the treatment phase. If the treatment indicates no effect, the proportion of data points below and above the line should be equivalent in the baseline and the treatment phases.

- **Confounding**

Something that influences a study and can contribute to misleading findings if it is not

understood or appropriately dealt with. For example, if a group of people exercising regularly and a group of people who do not exercise have an important age difference then any difference found in outcomes about heart disease could well be due to one group being older than the other rather than due to the exercising. Age is the confounding factor here and the effect of exercising on heart disease cannot be assessed without adjusting for age differences in some way.

- **Construct validity**

The degree to which the observed pattern of how our treatment and assessments work corresponds to our theory of how they should work. The researcher has a theory of how the measures relate to one another and how the treatment works and relates to the outcome. When how things work in reality - as captured by the study's assessments - matches up with how we theorized they should work, that provides evidence of construct validity.

- **C-statistic**

Can be applied to a data series with as few as eight observations. The C-statistic produces a z value, which is interpreted using the normal probability table for z scores. The C-statistic is first calculated for the baseline data. If the baseline data do not contain a significant trend, the baseline and intervention data are combined and the C-statistic is again computed to determine whether a statistically significant change has occurred.

- **Effect size**

The magnitude of a treatment effect, independent of sample size. The effect size can be measured as either: a) the standardized difference between the treatment and control group means, or b) the correlation between the treatment group assignment (independent variable) and the outcome (dependent variable).

- **Exclusion criteria**

Specified characteristics that would prevent a potential participant from being included in the study.

- **External validity**

The degree to which the results of a study hold true in non-study situations, e.g. in routine clinical practice. May also be referred to as the generalizability of study results to non-study patients or populations.

- **Fidelity (of intervention)**

The degree to which the intervention was delivered as planned and was differentiated from the control condition as planned, at both conceptual and pragmatic levels. Conceptually, the issue is whether the intervention, but not the control condition, captured the theoretical constructs that the researcher believes produce its positive effect. Pragmatically, the issues are whether the interventionists followed the treatment plan by delivering the intended intervention elements and by not delivering any elements that were proscribed.

- **Forest plot**

A graphical display of results from individual studies on a common scale, allowing visual comparison of results and examination of the degree of heterogeneity between studies.

- **Funnel plot**

Funnel plots are simple scatter plots on a graph. They show the treatment effects estimated from separate studies on the horizontal axis against a measure of sample size on the vertical axis. Publication bias may lead to asymmetry in funnel plots.

- **Inclusion criteria**

Specified characteristics that would enable a potential participant to be included in the study.

- **Internal validity**

Refers to the integrity of the study design. Experimental studies with a high degree of internal validity give a high degree of confidence that differences seen between the groups are due to the intervention.

- **Interrater reliability**

The degree to which two or more different raters give the same results to the same rating opportunity. This measures consistency between raters.

- **Intrarater reliability**

The degree to which a single rater is consistent in their assessment results.

- **Masking**

See "Blinding"

- **Power analysis**

An analysis that allows the researcher to estimate how many participants would be needed to achieve statistical significance for a given expected effect size. Power is the ability of a study to demonstrate an association or causal relationship between two variables, given that an association exists. For example, 80% power in a clinical trial means that the study has an 80% chance of ending up with a p value of less than 5% in a statistical test (i.e. a statistically significant treatment effect) if there really was an important difference (e.g. 10% versus 5% mortality) between treatments. If the statistical power of a study is low, the study results will be questionable (the study might have been too small to detect any differences). By convention, 80% is an acceptable level of power.

- **Reliability**

Reliability refers to a method of measurement that consistently gives the same results. For example someone who has a high score on one occasion tends to have a high score if measured on another occasion very soon afterwards. If different clinicians make independent assessments in quick succession - and if their assessments tend to agree then the method of assessment is said to be reliable. (see also Interrater reliability and Intrarater reliability).

- **Replication**

When a study is repeated in a different sample. Replications sometimes involve minor changes that add new information in addition to confirming results of the previous study.

- **Sequence generation**

Creating a master list which assigns participants to a study arm. A random number table or computer-generated random numbers provide the basis for adequate sequence generation.

- **Statistical validity (or statistical conclusion validity)**

The degree to which a statistical result is due to real differences rather than to chance or random error. This has to due with the proper use and interpretation of statistical tests.

- **Two-Standard Deviation Band Method**

Also called the Shewart chart method. First the standard deviation is computed for the baseline data. Once the standard deviation is computed for the baseline data, bands are drawn on the graph that contain scores within 2 standard deviations from the mean. The treatment effect is considered significant if at least two consecutive data points lie outside of the bands.

- **Validity**

There are a number of types of validity, all of which describe the degree to which we can "believe" the results of a study, either for the specific sample of participants, or for how the results apply to other samples. Commonly referenced types of validity are internal, external, statistical, and construct.

- **I^2 statistic**

A measure of the degree of inconsistency in study results indicating the percentage of total variation across studies that is due to genuine differences in the studies rather than chance; calculated as 100%x(Q-df)/Q where Q=Cochran's Q and df=degrees of freedom.