Appendix

How to Evaluate Scientific Literature

As new studies are published, the evidence base increases for understanding the risks and benefits of treatment options. A basic understanding of the types of studies and the meaning of the analyses helps healthcare professionals to evaluate the evidence and implications for clinical practice. Readers in search of more sophisticated discussions are urged to consult current clinical epidemiology texts.

Study types

Two major types of studies are *experimental* and *observational*.[1,2] There are also meta-analyses, which pool the results of clinical trials or other types of studies.

In experimental studies, interventions and conditions are strictly defined and controlled by the investigators. In observational studies, investigators observe outcomes in relation to variables of interest. They do not assign participants to an exposure of interest.

Experimental studies

Types of experimental studies are randomized controlled trials (RCTs), crossover trials, and quasi-experimental studies.

- *Randomized, controlled trials* are considered the strongest for therapeutic interventions. In RCTs, a group of participants with similar characteristics is identified (eg, low bone density). Each participant is then randomly assigned (similar to a flip of a coin; neither the participants nor the investigator chooses) to an *intervention group* (or groups) or to a *control group*. Participants typically have an equal and unbiased (random) chance of being assigned to each treatment under study. Randomized, controlled trials have the best chance of avoiding selection bias if the randomization is adequate and the study size is large. This is because known and unknown characteristics should be the same in each group. The baseline characteristics of the 2 groups should be presented and statistically compared in an RCT's final report. The Women's Health Initiative is an example of an RCT.

Randomized, controlled trials are best suited to situations in which exposure to treatment is modifiable, a legitimate uncertainty exists regarding benefit or harm of treatment, and outcomes are reasonably common. However, inclusion and exclusion criteria may limit the extrapolation of the results to other groups (limiting whether the results can be generalized).

The *power* of a trial is the likelihood that it will determine the effect of the intervention. The number of participants is determined when the trial is designed, based on the likelihood of the measured outcome events and the anticipated magnitude of the intervention's effect. If events do not occur as often as predicted, the trial may not have adequate power to determine the effect of the intervention. The *Methods* section of the published trial results will describe how the investigator calculated the number of participants required and will quantify the range of effect the study can detect (eg, the study was powered to detect more than a 20% reduction in heart attacks). *Publication bias* favors small trials with positive outcomes. An international registry of clinical trials requires submission of the planned trial (including design, interventions, and prespecified outcomes) so that a full accounting of the status of trials that are not completed or are completed, regardless of their outcomes, can be assessed.

Depending on the intervention, participants and investigators may be purposefully *blinded* or *masked* (ie, they do not know which treatment a participant is receiving). This helps reduce some forms of bias and the effects of the participant's or investigator's expectations of intervention benefit.

Classically, RCTs are used to assess for *efficacy* of the treatment in an ideal controlled setting. More often, an RCT will assess *effectiveness* (not efficacy) by studying the intervention under more usual circumstances. This is because a study for efficacy may not reflect its actual effectiveness in a real-world, clinical practice setting. Both types of RCTs often use a relatively narrowly

defined patient population. Even though an RCT is quite internally valid (ie, the study was done well), it may not be accurate to extrapolate (generalize) the results from one RCT to another patient population that was not studied in the trial.

Other important issues when evaluating the results of an RCT include checking to see whether all participants who started the trial were accounted for at trial conclusion and whether the groups were treated equally aside from the experimental intervention.

- *Crossover trials* allow participants to serve as their own controls. Participants are randomly assigned to one treatment arm and later switched to the other treatment arm. This crossover study methodology has often been used in trials to assess the efficacy of medications. The design is difficult to do well because of its potential for residual effects between interventions. Often there will be a *washout* period (time during which no treatment is given) between interventions.
- *Quasi-experimental studies* are of 2 general varieties. In one, 2 interventions are simultaneously compared in 2 groups of participants, but interventions are not randomized for any given participant (eg, 2 hospitals comparing 2 types of wound closure for the same type of surgery).

  Another common quasi-experimental study is when participants serve as their own controls, and the investigator controls the intervention. The intervention is neither randomized nor is there a control population to which the response can be compared. After baseline evaluation, the intervention is given, and the participants are reevaluated to observe any changes in characteristics because of the intervention.

Observational studies

Types of observational studies include purely *analytic* (or *descriptive*) studies. Analytic studies (including cohort studies, case-control studies, and cross-sectional studies) have a nonrandomized control group (eg, women who did not use hormone therapy [HT] for any number of reasons would be compared with women who elected to use HT). What is sampled first—risk factor, outcome, or risk factor and outcome simultaneously—determines which analytic study design is being used. Case reports and case series are not analytic because they do not have control comparisons. Observational studies can assess associations and correlations between exposures and outcomes, but they cannot confirm cause-and-effect relationships.

- *Cohort studies (or longitudinal studies)* begin with a defined group of participants (eg, persons of a certain age or those who work in a certain industry) called the *cohort*. These studies sample persons from the general population and determine whether they have exposures or risk factors of interest (eg, HT users vs nonusers). This cohort of persons is then followed over time to study a variety of outcomes. Data are collected in a similar manner on all participants from the beginning of the study (the baseline) and frequently at set intervals during follow-up. The Nurses' Health Study is one of the best-known prospective (occurring over time) cohort studies.

  These studies provide a clearer temporal sequence of exposures and outcomes than other analytic studies, are well suited for common exposures, and can study multiple exposures and outcomes. However, they can be time-consuming and expensive, they have the potential for many forms of bias, and participants may be lost during follow-up. When too many patients are lost, the validity of the study is compromised.

  Cohort studies may follow relevant events as they occur over time (prospective), but they may also be performed in a historical or a concurrent (cross-sectional) manner. Evidence from prospective cohort studies is considered stronger than the other forms of analytic studies because data on exposures are collected before the outcomes occur.

  The term *retrospective* is sometimes used when referring to a historical cohort study, but it can be confusing. If the data are easily accessible, the researcher can retrospectively evaluate a cohort that was followed in time, but the time was in the past moving forward (historical cohort), not progressing from current time onward (concurrent cohort). An example of a retrospective cohort is an occupational cohort with known exposure to a carcinogen or toxin in the remote past and whose participants have already accrued outcomes. In the Nurses' Health Study and the Framingham Study, information was gleaned in a concurrent and prospective fashion (in contrast to a historical cohort). All participants in each of these circumstances were followed longitudinally forward in some time frame.
- *Case-control studies* begin with an outcome or disease of interest (eg, myocardial infarction [MI], breast cancer) and then compare the characteristics or exposures of participants with the outcome *cases* to *controls* who do not have the outcome or disease of interest. Case-control studies are prone to many more forms of bias. A frequent one is *recall bias* (ie, participants cannot remember exposures or risk factors accurately).

  Matching participants for specific characteristics and defining strict eligibility criteria lessens but cannot eliminate the possibility that the results are *confounded*. For example, women who use HT are known to smoke less and lead generally healthier lifestyles. Hormone therapy users have less cardiovascular disease primarily because of better lifestyle habits rather than from any beneficial effect of HT use. Smoking or other lifestyle patterns can confound the results when observational studies analyze

HT use and health outcomes. Matching cases and controls for smoking status (or adjusting for this variable) helps reduce this confounding.

Despite these limitations, case-control studies have many advantages. Because they begin with an outcome of interest, they can be performed efficiently and at less cost than cohort studies. They are important in situations in which it would be unethical to assign participants to an exposure (eg, chemotherapy) or when an outcome is rare (eg, X-chromosome abnormalities associated with primary ovarian insufficiency).

- *Cross-sectional studies* are snapshots in time. Here, cases and controls are evaluated at the same time for risk factors or characteristics and outcomes of interest. Cross-sectional studies are very useful for determining prevalence, for planning for healthcare needs, and for generating hypotheses.
- *Case reports* and *case series* describe the experience of a single patient or series of patients. Such reports are useful in bringing new diseases or phenomena to the attention of the clinical and scientific community and for generating new hypotheses. However, lacking a control group, case reports or series without further study are only suggestive.

Many of these basic designs can be modified or combined, and many hybrid studies exist. An example is a case-control study within a cohort; this is a very useful study design and can provide many advantages, including cost efficiency.

Meta-analyses

Meta-analysis describes an analytic technique used to pool the results of clinical trials or other types of studies (not only RCTs). Often, a meta-analysis is performed on a group of studies that are too small to have statistical significance by themselves but that may show significance when pooled. Specific criteria (eg, eligibility criteria of participants, follow-up rates, data quality) are established to determine which studies will be included in the analysis. Inasmuch as any biases present in the contributing studies will be present in the meta-analysis, the outcome of a meta-analysis is only as good as the studies included.

In general, meta-analyses are difficult to perform. They are best performed based on the original data obtained from each investigator from each individual study. International guidelines provide checklists to understand the quality of a meta-analysis of clinical trials (eg, Consolidated Standards of Reporting Trials [CONSORT] guidelines).[3,4]

Analyzing study results

The bottom-line question when evaluating a study is: "What are the results?" The results of cohort studies and clinical trials are most frequently presented as a *relative risk* (RR)—the likely level of greater or lower risk (eg, for HT users vs nonusers). The RR can be determined because these study designs follow participants longitudinally, and risk (which is time-dependent) can be determined in each comparison group.

Rate/Risk

The term *rate* is the number of events per the number of participants per the time interval (eg, 44/10,000/y). Knowing the exact number of events over time is very useful, because this determines the risk.

The Council for International Organizations of Medical Sciences Task Force has provided nomenclature to guide the interpretation of risks[5]:

- Rare = Less than or equal to 10 in 10,000 per year
- Very rare = Less than or equal to 1 in 10,000 per year

Rare outcomes would not be of such great concern to an individual woman making a decision about treatment. However, it is important to recognize that common exposures that produce rare outcomes can still have profound public health effects.

Relative risk

The RR is a ratio—the rate of disease or the outcome of interest in a group exposed to a potential risk factor or treatment, or having a characteristic of interest, divided by the rate of disease of interest in an unexposed group (ie, those without the risk factor, treatment, or characteristic of interest). The RR should be used only for prospective studies.

Rate is used in both the numerator and the denominator. These are the numbers of events, per numbers of participants, per time interval (eg, 50/100,000/y). For example, if the annual rate of MI in women who smoke is 220 per 100,000, and the annual rate in women who do not smoke is 110 per 100,000, the RR associated with smoking is:

$$RR = \frac{220}{100{,}000/y} \div \frac{110}{100{,}000/y} = 2.00$$

This means that compared with nonsmoking women, the risk of MI for a woman who smokes is twice that of a woman who does not smoke in the study.

An RR less than 1.0 suggests that the factor decreases risk. For example, an RR of 0.50 means that there is a 50% less chance (or risk) of the outcome studied in those with the risk factor versus those without the risk factor of interest. An RR of 0.3 means a 70% lower relative risk.

An RR greater than 1.0 suggests that the factor increases risk. For example, an RR of 1.2 means there is a 20% increase in risk in the group with the risk factor versus the group without the risk factor. An RR of 2.0 means double the risk.

Odds ratio

The odds ratio (OR) is an estimate used in many analytic studies. It best approximates the RR when the outcome is rare.

Appendix

Confidence interval

The confidence interval (CI), usually cited with the RR or the OR, indicates with a certain degree of assurance the range within which lies the true magnitude of the measured effect. The CI has 2 components—the degree of certainty and the range (eg, 95% CI, 1.09-1.32). The point estimate (the RR or OR number) is the best mathematical estimate from the data. Understanding the upper and lower limits of the range is often clinically useful. If the CI is *wide*, the reader's confidence in the validity of the RR would be less than if the CI is *narrow* (ie, closer to the value of the RR).

Often, a 95% CI is used. A 95% CI gives the range of values that have a 95% probability of containing the true RR or OR. When a 95% CI does not contain the number 1.0 (eg, 0.40-0.80 or 1.12-1.37), the measured RR or OR is statistically significant by at least $P<.05$. The CI is more clinically useful than the *P* value because the CI helps the reader to understand the best estimate of the effect, and it provides the mathematical estimated limits, which are useful in determining the best-case and worse-case scenarios.

P value

This term is the probability of obtaining the observed RR or OR (or a more extreme value) by chance (random sampling) alone. A *P* value of .01 means that there is a 1% mathematical probability that the observed difference between 2 groups occurred by chance. By convention, *P* is generally deemed statistically significant if it is below 0.05. This means that if 20 outcomes are evaluated in a single study, 1 of these outcomes is likely to show a positive result just because of chance alone ($P=.05$, or 1/20). By the time $P=.001$, the likelihood is only 1 in 1,000 that the results occurred by chance—in other words, the finding is more likely to be real.

It is important to remember that a study can be statistically significant and not clinically significant. However, if it is not statistically significant, it cannot reach clinical significance, and the result could be clinically nonsignificant or inconclusive. An example is when the study is underpowered.

Absolute risk/Attributable risk

The effect of RR on a population and on an individual basis depends on *incidence* (ie, the number of new cases in a given period). This can be quantified by the absolute risk or attributable risk (AR), which is the difference between the incidence rates in the exposed and unexposed groups—in other words, the *risk difference*. The AR quantifies the effect of exposure, providing a measure of its public health effect. For example, for the calculation about the risk of MI in women who smoke, the AR is

$$\text{AR} = \frac{220}{100{,}000/\text{y}} - \frac{110}{100{,}000/\text{y}} = \frac{110}{100{,}000/\text{y}}$$

This means that for every 100,000 women who smoke, there would be 110 additional cases of MI per year.

Often, AR is more clinically useful than RR in explaining risk to patients. The US Food and Drug Administration requires that the absolute risk reduction be included on the product information sheet.

Number needed to treat

To communicate this risk difference to patients, the number needed to treat (NNT) can be useful. The NNT is merely the reciprocal of the AR (ie, 1 divided by the AR). For example, in a 1-year study, if the rate of an outcome was 20 per 1,000 in an untreated group and 10 per 1,000 in a treated group, the NNT is

$$\text{NNT} = \frac{1}{(20/1{,}000) - (10/1{,}000)} = \frac{1}{(10/1{,}000)} = \frac{100}{0.01}$$

This means that for every 100 people treated, there would be 1 fewer negative outcome over the year.

References

1. Koepsell TD, Weiss NS. *Epidemiologic Methods: Studying the Occurrence of Illness.* New York: Oxford University Press; 2003.
2. Porta M. ed. *A Dictionary of Epidemiology.* 5th ed. New York: Oxford University Press; 2008.
3. Moher D, Schulz KF, Altman D; CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA.* 2001;285(15):1987-1991.
4. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) Group. *JAMA.* 2000;283(15):2008-2012.
5. World Health Organization. *Guidelines for Preparing Core Clinical-Safety Information on Drugs.* 2nd ed. Report of CIOMS Working Groups III and IV. Geneva: World Health Organization; 1999.