How to Read a Clinical Trial Paper: A Lesson in Basic Trial Statistics

Shail M. Govani, MD, MSc, and Peter D. R. Higgins, MD, PhD, MSc

Dr. Govani is a Fellow and Dr. Higgins is an Assistant Professor of Internal Medicine, both in the Division of Gastroenterology of the Department of Internal Medicine at the University of Michigan in Ann Arbor, Michigan.

Address correspondence to:
Dr. Peter D. R. Higgins
1150 W. Medical Center Drive
SPC 5682
Ann Arbor, MI 48109;
Tel: 734-647-2964;
Fax: 734-763-2535;
E-mail: phiggins@umich.edu

Abstract: While the number of clinical trials performed yearly is increasing, the application of these results to individual patients is quite difficult. This article reviews key portions of the process of applying research results to clinical practice. The first step involves defining the study population and determining whether these patients are similar to the patients seen in clinical practice in terms of demographics, disease type, and disease severity. The dropout rate should be compared between the different study arms. Design aspects, including randomization and blinding, should be checked for signs of bias. When comparing studies, clinicians should be aware that the outcomes being studied may vary greatly from one study to another, and some outcomes are much more reliable and valuable than others. The definition of clinical response should also be scrutinized, as it may be too lenient. Surrogate outcomes should be viewed cautiously, and their use should be well justified. Clinicians should also note that statistical significance, as defined by a *P*-value cutoff, may be the result of a large sample size rather than a clinically significant difference. The treatment effect can be estimated by calculating the number needed to treat, which will demonstrate whether changes in clinical practice are worthwhile. Finally, this article discusses some common issues that can arise with figures.

Keywords
Statistics, clinical trials, randomized controlled trials, outcomes, clinical research

In the last decade, the number of publications dedicated to research in gastroenterology has expanded dramatically.[1] In parallel, the number of clinical trials has been increasing, with more than 18,000 ongoing clinical trials in the United States alone.[2] With this rapid growth in research, clinicians find themselves trying to keep up with wave after wave of studies and trying to determine how these studies apply to their patients. As researchers learn more about disease processes, clinical trial design is also changing and becoming more complex. Medical school and subsequent clinical training provide limited education on how to evaluate these stud-

Table 1. Differences in Sample Characteristics Depend on Recruitment Method

	UK primary care practice (n=121)	US newspaper advertisement (n=72)	US gastroenterology clinics (n=52)	*P*-value
Average age (yrs)	43.3	49	42.3	<.05*†
Female (%)	78	74	81	NS
Smokers (%)	22	10	10	<.01**†
Psychiatric visit in the last 5 years (%)	20	40	23	<.01*
Antidepressant use (%)	18	29	42	<.01**
Irritable bowel syndrome symptoms in the last week				
Moderate (%)	87	97	77	<.05* <.001†
Severe (%)	13	3	23	<.05* <.001†

*Difference between UK primary care practice and US newspaper advertisement.
**Difference between UK primary care practice and US gastroenterology clinics.
†Difference between US newspaper advertisement and US gastroenterology clinics.
NS=not significant.
Adapted from Longstreth GF, et al.[4]

ies and incorporate their results into practice. Thus, this article will review some important concepts in the evaluation of therapeutic trials.

Are These Subjects Like My Patients?

The first step in the process of reviewing a trial should involve determining whether the subjects in the study are representative of the patients a clinician sees in his or her practice. Were the study subjects sicker, or did they have more complicated disease? Did they fail more prior medications, or did they have longer disease durations? Patients who enroll in clinical trials are usually different from the average patient with a particular disease. This trend has been well studied in oncology studies and is also a factor in gastroenterology clinical trials.[3] Compared to the average patient, study subjects often have a longer duration of disease, more complications, more active disease, and more previous medication failures.

Recruitment methods also influence the make-up of the study population and can lead to different response rates. This effect was demonstrated in a mock irritable bowel syndrome (IBS) study in which 3 patient samples were recruited and then examined.[4] One sample was recruited from a primary care practice in the United Kingdom, another was recruited from gastroenterologists' offices in the United States, and the third was recruited by newspaper advertisement in the United States. These 3 samples had significant clinical and demographic differences (Table 1).

In publications of clinical trials, the first table in the paper usually summarizes the characteristics of the study sample, and readers should examine this table to determine not only the age, race, and gender of the population but also the typical disease severity, disease duration, and medication history. Clinicians can only apply the results of a study to patient care if all of these factors appear to be similar to those of the patients seen in clinical practice.

What Happened to the Subjects? Did They Drop Out? Why?

After comparing real-world patients and the study sample, the next step is a review of the Consolidated Standards of Reporting Trials (CONSORT) diagram. This flow diagram is usually the first figure in the paper, and it shows the flow of subjects from recruitment through the end of the study (or early exit from the study). Special attention should be paid to the number of patients in each arm who dropped out of the study. Subjects drop out of studies for many reasons, but 2 of the most important reasons are lack of benefit (they were too sick to stay in the study) and side effects (which made them want to leave the study). Clinicians can compare the percentage of subjects in each arm of the study who drop out due to lack of benefit. If the subjects consider the treatment to be more effective than placebo, then the dropout rate due to lack of benefit should be lower in the treatment arm than in the (less effective) placebo arm. If this is the case, this finding not only validates the statistical signifi-

cance of the study results but also indicates that patients consider this difference to be clinically significant. In some studies, the reasons for leaving the study are not specified, and readers have to make do with comparing the exit rate for each arm. While the exit rate is a cruder measure, the more effective treatment arm should have a lower exit rate (if the treatment is well tolerated).

Is the Study Design Biased?

While readers may be tempted to skip the methods section of a paper, the study design should be reviewed carefully, with a particular focus on the comparator group, allocation of subjects to treatment arms, and blinding. In disease states with established therapies such as inflammatory bowel disease (IBD), the comparator should be the standard-of-care therapy. Functional disease states should incorporate a placebo group and/or a comparator that has proven to be effective (if available) to ensure that the placebo effect is measured. For example, in studies of rifaximin (Xifaxan, Salix Pharmaceuticals) versus placebo for treatment of IBS, 32% of the placebo-treated patients reported relief of symptoms, while 41% of subjects in the treatment arm reported improvement.[5] Without the placebo comparator, the effect of rifaximin on IBS symptoms would appear quite large, but readers need to discount results by the size of the placebo effect.

The allocation of subjects to each treatment arm should be well defined. The CONSORT diagram should be easy to follow and similarly detailed for each arm. To avoid bias, subject allocation must be concealed so that investigators cannot anticipate which arm the next subject will enter. If the investigators knew a very sick patient would be likely to receive placebo, they could avoid enrolling that patient, which could result in biased enrollment. Randomization may be at the population level or clustered at the site level; in the latter case, all the subjects at 1 site are in a single arm.

The blinding process is defined as single-blind if the subject is unaware of the allocation, while double-blind refers to a study in which the investigators are also unaware of the individual patient's allocation. Blinding of both participants and evaluators is equally important, as patients who are aware of their intervention may have lower compliance, and investigators who are aware of the intervention may overestimate a treatment effect. Not surprisingly, studies comparing outcomes with and without blinding have shown a significant overestimation of the treatment effect in studies conducted without blinding.[6,7]

Unfortunately, study investigators rarely report whether the blinding process was successful. In a 2007 analysis of 1,599 studies, only 2% reported the success of the blinding process; of those, almost half determined that blinding was unsuccessful (ie, patients could determine or guess whether they were receiving active drug or placebo).[8] The blinding process can be difficult in many studies, especially trials involving a procedure. Subjects often go to remarkable lengths to determine whether their study medication is active drug or placebo; especially given the information available on the Internet, extreme care must be taken to preserve the blind. If patients in a placebo-controlled trial can deduce that they are in the placebo arm, they could be less likely to report improvement, leading to underestimation of the placebo effect and overestimation of the relative treatment effect.

Does the Study Include an Intention-to-Treat Analysis?

All subjects who have been randomized should be counted in the analysis. Some subjects may not receive the intervention or may drop out of the study after a short period of time, but for the purposes of analysis, they should still be considered as having been assigned to their treatment arm. An analysis that includes all randomized subjects is called an intention-to-treat (ITT) analysis. As soon as subjects are randomized, the intention to treat them with their assigned intervention is established, even if they never receive this intervention.

Occasionally, clinical trials will be greatly damaged by a high rate of early drop out, and investigators will be tempted to present a per-protocol analysis, in which only subjects who received their assigned intervention are considered. This analysis can be presented, but it should be presented only as a secondary endpoint, after the presentation of the ITT results. Despite the importance of performing an ITT analysis, fewer than 50% of papers in some well-regarded journals identify their analysis as an ITT analysis; even among these papers, verifying that the analysis was performed correctly is difficult.[9]

Is This a Test of Superiority? Equivalence? Noninferiority?

While most trials in the literature compare 2 therapies with the intent of proving that one is superior to the other, trials of equivalence and noninferiority are occasionally performed. In cases where a new therapy is more convenient or cheaper, an equivalence or noninferiority study can be appropriate. For example, a recent trial examined whether individualized duration of treatment for hepatitis C virus (HCV) infection was noninferior to standard treatment duration.[10]

Studies of noninferiority and equivalence can be particularly difficult to interpret due to the definitions of

noninferiority and equivalence. The acceptable amount of difference between the standard therapy and the alternative therapy must be defined in advance of the study, and the authors must provide a strong rationale for this definition. For example, in the HCV study mentioned above, the study authors assumed a sustained virologic response (SVR) rate of 48% in patients receiving standard therapy, and they allowed an acceptable margin (Δ) of 5% with individualized therapy. In other words, the authors were willing to declare individualized treatment noninferior to standard therapy if the SVR rate with individualized therapy was within 5% of the SVR rate achieved with standard therapy. Readers then have to decide whether they would accept using a treatment that is up-to-5% less effective than the standard-of-care therapy.

From the standpoint of study design, the sample size needed to determine noninferiority or equivalence is always much higher than the size needed to determine superiority. The methods section should define every aspect of the sample size calculation so that the reader can replicate this calculation if necessary. Typically, the design of an equivalence study requires that the 95% confidence interval of the difference between the treatments be less than the previously defined acceptable difference. This requirement often results in a required sample size about 4 times larger than the sample size needed for a superiority study. If a study is initially designed to test for superiority but this outcome is not found, switching to an equivalence study would be inappropriate for 2 reasons: (1) the acceptable difference would not have been defined in advance, and (2) the sample size would be too small.

Does the Measurement Matter? Is It Reproducible? Accurate?

One of the overlooked issues that can make a clinical trial difficult to generalize to the patients clinicians see in everyday practice is the study's measurement of success. Clinical trials require detailed definitions of success—for example, criteria for remission or clinical response in IBD trials—and these definitions usually involve several measurements, sometimes a clinical severity index, and some form of a scale. Several issues related to these measurements are discussed below.

First, is there good evidence that the measurement itself is reproducible and precise? Many measures are not very reproducible, including histologic assessment of dysplasia and endoscopy in patients with ulcerative colitis.[11,12] Endoscopic grading instruments are quite complex and suffer from a high level of disagreement in the middle of the scale, much like histologic grading of dysplasia. This disagreement creates noise in the data.

A second issue is whether the amount of change seen in the clinical index or score is clinically meaningful to patients. Does a change of 3 points on the index presented in a published paper translate to a meaningful improvement for real-world patients? This question is important to consider, as results can be statistically significant without being clinically meaningful.

A third issue is whether the measurement has been validated. Validity can have a number of meanings, but it generally includes whether the index is measuring all of the important aspects of a disease, whether it measures them accurately, whether the measurement is reproducible in subjects whose condition has not changed, and whether the instrument is responsive to small but clinically important changes. For example, the commonly used Crohn's Disease Activity Index (CDAI) can be influenced by postinflammatory IBS, which can lead to symptoms that will not respond to anti-inflammatory therapy. In the future, instruments measuring Crohn's disease activity will likely have separate subjective symptom scales and objective measures of inflammation to better define disease activity.

Surrogate Outcomes

Ideally, the outcome being studied should be the one in which clinicians are most interested. Primary outcomes—such as remission, cure, and nonrecurrence—are stronger and more meaningful than surrogate outcomes such as biomarker levels. While measuring surrogate biomarkers may appear to be an easy, inexpensive, and less invasive method of determining a clinical outcome, the validity of the surrogate markers should be scrutinized closely. A glaring example of the misuse of surrogate markers was the use of class 1c antiarrhythmic medications to prevent sudden cardiac death based on the evidence that these medications suppress the number of premature ventricular contractions (PVCs). Previous research had shown a correlation between PVCs and poor clinical outcomes. However, a prospective randomized trial comparing these drugs to placebo showed an increased risk of cardiac death in the treatment arms despite successful suppression of PVCs.[13] Such faulty reasoning likely led to many deaths and is a lesson of the importance of surrogate outcome choice.

Bucher and colleagues suggested a 3-step approach to assessing the validity of a surrogate outcome.[14] First, a strong independent association should exist between the surrogate and the desired outcome. Second, evidence from randomized controlled trials in other drug classes should show that improvement in the surrogate leads to an improvement in the clinical outcome. Lastly, randomized controlled trials within the same drug class should demonstrate similar improvements in the sur-

rogate, which lead to improvements in the desired clinical outcome. For example, mucosal healing (typically defined as endoscopic healing) is often used as a surrogate marker for clinical improvement in IBD. In applying these rules, readers will notice that steroids have not led to improvements in mucosal healing despite leading to clinical improvement.[15] However, other immunomodulators, including azathioprine, methotrexate, and infliximab (Remicade, Janssen Biotech), have shown improvement in mucosal healing and clinical improvement.[16,17] Any surrogate marker that has an inconsistent correlation with important clinical outcomes has to be viewed skeptically and should not be considered an ideal primary endpoint for clinical trials.

Dichotomous Outcomes, Continuous Outcomes, Correlations, and Time-to-Event Endpoints

Some studies focus on a dichotomous outcome endpoint like clinical remission or response. The definitions of these endpoints are critical and should be scrutinized carefully to ensure that they are not too lenient. For example, early studies of Crohn's disease therapies defined clinical response as a reduction in CDAI score of 70 points or more, and remission was defined as an overall CDAI score less than 150 points. More recently, the US Food and Drug Administration has encouraged changing the definition of clinical response to a decrease of 100 points or more. In addition, Sandborn and colleagues have suggested that decreases in CDAI scores of 70 or 100 points (Δ70, Δ100) should only be used as secondary outcome measures and should be coupled with a CDAI score less than 150 points.[18]

Other studies may present a continuous outcome measure, which provides more power to detect a difference between groups. For example, if a study evaluating 2 immunomodulator drugs for the treatment of Crohn's disease compared the differences in CDAI scores before and after drug use, the study could conclude that there is a difference between the 2 drugs even if that difference is only 5 points. On the other hand, when the outcome is dichotomized (hopefully into a clinically meaningful difference) and a statistically significant difference is found, these findings are more meaningful for clinical care.

Other outcomes, including correlations and time-to-event endpoints (ie, time to remission), should also be treated carefully. In testing for correlations, the null hypothesis is that there is no correlation. The *P*-value in such tests is very sensitive to weak correlations, as even small correlations will be considered significant. In general, it is probably best to ignore the *P*-value for correlations and instead look at the correlation itself (and its 95% confidence interval). A correlation of 90% with a confidence interval of 85–95% is impressive, while a correlation of 23% with a confidence interval of 1–57% is not very impressive, although the latter would have a significant *P*-value ($P<.05$). A study that reports a correlation and gives a *P*-value but not a 95% confidence interval should be viewed with caution.

Time-to-event data are also weaker endpoints than outcomes measured at a particular time point, such as remission. Remission rates at a particular time point (ie, 12 weeks) provide 1 data point for each subject, and the rates between study arms can be compared, usually with a *t*-test. The *t*-test for a dichotomous outcome (ie, remission vs nonremission) sets a high bar for success. In contrast, statistical tests for time-to-event data compare the survival rate, which is essentially an estimate of the slope across the entire follow-up period. This approach effectively counts each subject multiple times by looking at each subject at multiple time points, thus adding to the statistical power of the test.

The problem with a time-to-event analysis is that it can produce a significant *P*-value, but this value may not reflect a clinically meaningful difference. Examples of these types of studies include the placebo-controlled studies of budesonide as maintenance therapy for Crohn's disease.[19,20] These studies revealed that budesonide was associated with a longer time to relapse; however, the outcomes in the placebo and budesonide arms were the same at 18 months. In addition, the time-to-relapse curves were not presented with confidence intervals, making it difficult to interpret whether meaningful differences were achieved. While time-to-event analysis is considered an acceptable approach, especially in oncology, a statistically significant result with this design does not carry as much clinical weight as a statistically significant result from a *t*-test of a dichotomous, clinically meaningful endpoint evaluated at a single time point, given that the time-to-event analysis sets a much lower bar for success.

Multicenter Trials

Multicenter trials offer a chance to expand the sample size in studies of rare diseases and to ensure some population diversity. However, a potential weakness of a multicenter trial is the lack of standardization of both the intervention and the outcome. The protocol for the intervention has to be especially rigid and detailed in a multicenter trial. Similarly, the chosen outcome should be validated and reproducible across multiple centers. For endoscopic endpoints, reproducibility has proven problematic, leading many studies to move to grading of endoscopic appearance by central reviewers in order to improve consistency. Analyses of multicenter trials should always consider site as a covariate, as a few outlier sites can skew the outcome.

Shifting Target Outcomes

Since the first gastrointestinal clinical trial, Truelove and Witt's 1954 study of cortisone for the treatment of ulcerative colitis, target outcomes have changed and become more stringent.[21] While clinicians keep raising the bar as their body of knowledge grows, the incremental gain in patient outcomes does not appear to be as great. Over time, clinicians will need to determine whether there is an improved, long-term, benefit-to-risk ratio that results from achieving biologic or endoscopic remission in patients with IBD. In an example from another field of medicine, very tight control of blood pressure showed no survival benefit over usual control; to avoid similar missteps, clinicians need prospective evidence that tight control in IBD will improve outcomes.[22]

How Are Missing Data Addressed?

One of the challenges of all trials is how to deal with missing data. Subjects often drop out of a study before the primary endpoint date is reached, fail to show up for appointments, or change their mind and refuse to undergo the scheduled endoscopy at the final evaluation. There are 2 common methods for handling missing data. The most rigorous method is to consider any subject who has missing data as one who would have failed to meet the endpoint. This method is often called nonresponder imputation. It lowers success rates (and placebo rates), thus making results look less impressive.

Another approach is to use the last observation at which the subject was measured in place of the missing data; this approach is especially common in maintenance studies with repeated measurements. Called the last-observation-carried-forward approach, this method is reasonable, but it may inflate the success rates of both the primary intervention and the control arm, as a common reason patients do not return for visits is that they feel the treatment does not provide a clinical benefit.

Finally, a third approach for addressing missing data is called imputation, in which other data are used to estimate what the missing data would have been. This approach is generally considered appealing but questionable, as it is impossible to check the accuracy of these estimates.

Do the Design and Methods Conform to the Prestudy Guidelines?

A number of clinical journals and funding sources expect clinical trials to be registered on the ClinicalTrials.gov website. This website contains a large amount of information about the trial, including endpoints, date of first enrollment, inclusion and exclusion criteria, and a description of the intervention. An interested reader can view the registration page for the trial to ensure that the endpoints did not change during the study and that the trial was truly prospective. A review of the endpoints listed on this website may reveal that the trial's secondary endpoints were exploratory (ie, not defined prior to initiation of the trial). Readers may also discover that some of the outcomes mentioned on the initial registration page are not reported in the published paper. In these instances, the reader could presume that the outcome results were disappointing.

How Good Are the Results?

Many readers rely on *P*-values to determine if a clinically important effect is present. The *P*-value is based on the assumption that the null hypothesis is true (that the treatment arms are equally effective), and the *P*-value gives the probability that the difference observed between the 2 treatment arms is due to chance alone. For example, if testing the difference in blood pressures between 2 populations that truly had the same blood pressure resulted in a *P*-value of .04 (from a student's *t*-test), then there would only be a 4/100 chance of seeing a difference of that magnitude by chance alone.

Statistically, a *P*-value serves as a combination of 2 measurements: effect size (the difference between the 2 groups) and precision (the variation of that difference represented by the confidence interval). These 2 values should be displayed in a paper and should be examined separately. A very large study will have more precision, which will generally lead to a smaller *P*-value without a meaningful change in effect size. A very small study will inherently have less precision, in which case a clinically meaningful effect size may be missed due to a higher *P*-value. Thus, a small effect size—for example, a 5% response rate in 1 arm and 10% response rate in the other arm—will be statistically significant in a very large study (>500 patients per arm), but it will be unlikely to produce a significant *P*-value if the study has only 50 patients per arm.

To illustrate how a large sample size can make a small effect significant, consider the 2 examples in Figure 1. A large sample size, as shown in Figure 1A, can make a blood pressure difference of 2.5 mmHg between sample populations statistically significant. When the difference between the 2 groups is larger (12 mmHg), a sample size of 50 patients per group is only barely sufficient to yield a statistically significant difference between the 2 groups (Figure 1B). While these 2 hypothetical studies have similar *P*-values, the larger, more expensive study has, in effect, "bought" a better *P*-value with larger sample sizes. The smaller study may have a larger and more clinically relevant effect, but this effect is difficult to detect with a small sample size. If the smaller study had failed to reach

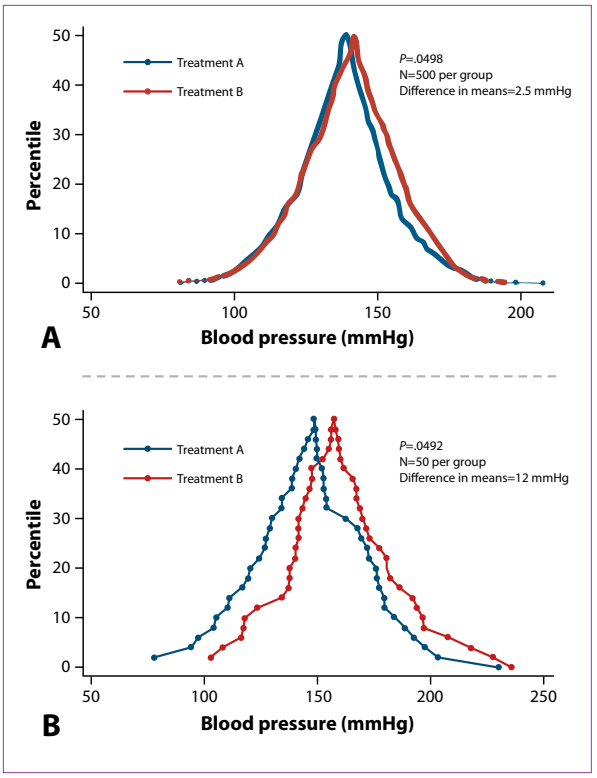

Figure 1. Figure **1A** shows overlaid mountain plots of the systolic blood pressures of two 500-subject arms of a hypothetical study. The difference between the 2 means is 2.5 mmHg. Because of the large sample size, this small difference achieves a statistically significant *P*-value. Figure **1B** shows overlaid mountain plots of the systolic blood pressures of two 50-subject arms of a much smaller hypothetical study, which has a difference between the means of 12 mmHg. Due to the smaller sample size, this study barely achieves the same *P*-value, despite having a larger, clinically significant effect.

its recruiting goal and had stopped at 40 patients per arm, the *P*-value would not have been significant.

As mentioned earlier, the dropout rate due to lack of benefit can be compared between the different arms of a study, and this rate should be lowest in the arm with the most effective treatment. This is called the "voting with their feet" endpoint. A meta-analysis measuring this outcome compared the clinical response in ulcerative colitis patients who were treated with 5-aminosalicylic acid drugs versus placebo, and it confirmed that the dropout rate was a powerful and intuitive way to confirm clinical response from the point of view of the patient, no matter how clinical response was measured in the study.[23] This endpoint should be generalizable to other clinical trials and is a quick and easy way to gauge the response rate from the point of view of the patient.

Table 2. Crude Exit Rate by Arm: Data From the SONIC Trial

	Primary endpoint	Exit rate endpoint
Treatment arm	**Percent achieving CDAI <150 points**	**Percent who dropped out**
Azathioprine	30.0	26.1
Infliximab	44.4	18.4
Azathioprine plus infliximab	56.8	10.1

CDAI=Crohn's Disease Activity Index.
Data from Colombel JF, et al.[24]

For example, applying this endpoint to the SONIC trial, in which patients with Crohn's disease were randomized to receive azathioprine, infliximab, or combination therapy, the overall exit rate was found to be lowest in the combination therapy group. The CONSORT diagram does not define what percentage of patients dropped out due to lack of effect, so the crude exit rate was used; the exit rate was defined as patients who were lost to follow-up, withdrew consent, or dropped out "for other reasons" (Table 2).[24] Using the chi-square test, the distribution between the dropout percentages in the 3 arms is statistically significant ($P<.001$), supporting the validity of the findings from the patient's point of view.

How Big of an Effect Is It?

To help judge the impact of a study's findings, it is worthwhile to calculate the number needed to treat (NNT). This number represents the reciprocal of the absolute risk reduction and gives the reader an understanding of how many patients a clinician would have to treat to see the desired outcome. For example, in a 2007 study assessing the benefit of high-dose proton pump inhibitor (PPI) therapy prior to endoscopy, high-dose therapy was shown to reduce the need for endoscopic therapy in patients with gastrointestinal bleeding.[25] In this study, 19.1% of patients in the omeprazole arm needed endoscopic therapy, while 28.4% of patients in the placebo arm needed endoscopic therapy. The absolute risk difference was .093 (.284–.191), leading to an NNT of 11 (1/.093). Therefore, a clinician would have to treat 11 patients presenting with upper gastrointestinal bleeding with high-dose PPI therapy prior to endoscopy in order to reduce the need for endoscopic therapy by 1 patient.

Checking the Figures

Finally, readers should consider a few points when reviewing figures. When illustrating a study's results, the authors can make interesting graphical choices, sometimes for

aesthetic reasons and at other times to make their results appear more compelling. The most basic (and easily overlooked) errors involve axis manipulation: the lack of a zero point on the y-axis or a change in scale in the middle of the axis. While the authors may be trying to make their graph more understandable by "cropping" the axis, the results leave the reader without perspective. Having an axis that covers the full range of possible results for a measurement is generally best. This may be 0–100% on the y-axis, rather than 10–40%, which would make differences appear larger; for a scale like the Mayo Clinic Score, which is used to evaluate patients with ulcerative colitis, the full range of possible scores that should be graphed is 0–12.

Another consideration with graphs is that, while plots of means without an indication of variability may be more aesthetically pleasing, these figures do not give the reader an understanding of the variation in the data. Ideally, this variability should be presented by plotting all the data points or, if the data are summarized, by showing means with 95% confidence interval bars instead of standard error or standard deviation bars.

Conclusion

While researchers have learned a great deal from randomized controlled trials, some hypotheses cannot be tested with prospective randomized controlled trials due to cost, duration, or ethical constraints. For example, a recent cross-sectional study of the relationship between fiber intake and diverticular disease is a reminder that other study designs serve an important function in medicine.[26] These studies can be important for hypothesis generation or for identifying risk associations. With this point in mind, we emphasize the importance of considering study designs beyond clinical trials, as other types of studies can address important questions that are not suited to a prospective randomized controlled trial design.

References

1. Chou LF. Medline-based bibliometric analysis of gastroenterology journals between 2001 and 2007. *World J Gastroenterol*. 2009;15:2933-2939.
2. ClinicalTrials.gov. Advanced search. http://www.clinicaltrials.gov/ct2/search/advanced. Accessed June 26, 2011.
3. Elting LS, Cooksley C, Bekele BN, et al. Generalizability of cancer clinical trial results: prognostic differences between participants and nonparticipants. *Cancer*. 2006;106:2452-2458.
4. Longstreth GF, Hawkey CJ, Mayer EA, et al. Characteristics of patients with irritable bowel syndrome recruited from three sources: implications for clinical trials. *Aliment Pharmacol Ther*. 2001;15:959-964.
5. Pimentel M, Lembo A, Chey WD, et al; TARGET Study Group. Rifaximin therapy for patients with irritable bowel syndrome without constipation. *N Engl J Med*. 2011;364:22-32.
6. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336:601-605.
7. Poolman RW, Struijs PA, Krips R, et al. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? *J Bone Joint Surg Am*. 2007;89:550-558.
8. Hrobjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *Int J Epidemiol*. 2007;36:654-663.
9. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*. 1999;319:670-674.
10. Sarrazin C, Schwendy S, Moller B, et al. Improved responses to pegylated interferon alfa-2b and ribavirin by individualizing treatment for 24-72 weeks. *Gastroenterology*. 2011;141:1656-1664.
11. Eaden J, Abrams K, McKay H, Denley H, Mayberry J. Inter-observer variation between general and specialist gastrointestinal pathologists when grading dysplasia in ulcerative colitis. *J Pathol*. 2001;194:152-157.
12. Cooney RM, Warren BF, Altman DG, Abreu MT, Travis SP. Outcome measurement in clinical trials for ulcerative colitis: towards standardisation. *Trials*. 2007;8:17.
13. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med*. 1989;321:406-412.
14. Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. *JAMA*. 1999;282:771-778.
15. Olaison G, Sjodahl R, Tagesson C. Glucocorticoid treatment in ileal Crohn's disease: relief of symptoms but not of endoscopically viewed inflammation. *Gut*. 1990;31:325-328.
16. D'Haens G, Van Deventer S, Van Hogezand R, et al. Endoscopic and histological healing with infliximab anti-tumor necrosis factor antibodies in Crohn's disease: a European multicenter trial. *Gastroenterology*. 1999;116:1029-1034.
17. D'Haens G, Geboes K, Rutgeerts P. Endoscopic and histologic healing of Crohn's (ileo-) colitis with azathioprine. *Gastrointest Endosc*. 1999;50:667-671.
18. Sandborn WJ, Feagan BG, Hanauer SB, et al. A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. *Gastroenterology*. 2002;122:512-530.
19. Lofberg R, Rutgeerts P, Malchow H, et al. Budesonide prolongs time to relapse in ileal and ileocaecal Crohn's disease. A placebo controlled one year study. *Gut*. 1996;39:82-86.
20. Sandborn WJ, Lofberg R, Feagan BG, Hanauer SB, Campieri M, Greenberg GR. Budesonide for maintenance of remission in patients with Crohn's disease in medically induced remission: a predetermined pooled analysis of four randomized, double-blind, placebo-controlled trials. *Am J Gastroenterol*. 2005;100:1780-1787.
21. Truelove SC, Witts LJ. Cortisone in ulcerative colitis; preliminary report on a therapeutic trial. *Br Med J*. 1954;2:375-378.
22. Cooper-DeHoff RM, Gong Y, Handberg EM, et al. Tight blood pressure control and cardiovascular outcomes among hypertensive patients with diabetes and coronary artery disease. *JAMA*. 2010;304:61-68.
23. Rangwalla SC, Waljee AK, Higgins PD. Voting with their feet (VWF) endpoint: a meta-analysis of an alternative endpoint in clinical trials, using 5-ASA induction studies in ulcerative colitis. *Inflamm Bowel Dis*. 2009;15:422-428.
24. Colombel JF, Sandborn WJ, Reinisch W, et al; SONIC Study Group. Infliximab, azathioprine, or combination therapy for Crohn's disease. *N Engl J Med*. 2010;362:1383-1395.
25. Lau JY, Leung WK, Wu JC, et al. Omeprazole before endoscopy in patients with gastrointestinal bleeding. *N Engl J Med*. 2007;356:1631-1640.
26. Peery AF, Barrett PR, Park D, et al. A high-fiber diet does not protect against asymptomatic diverticulosis. *Gastroenterology*. 2012;142:266.e1-272.e1.