

Interpreting Negative Studies

STEPHEN D. SIMON

From the Office of Medical Research, Children's Mercy Hospital, Kansas City, Missouri.

How many research subjects does it take to change a light bulb? At least 300 if you want the bulb to have adequate power.

One of the trickiest problems for readers of *Journal of Andrology* and other research journals is in assessing negative findings. An example is when a new therapy does not achieve statistical significance relative to a standard therapy. Does the lack of statistical significance tell us that the 2 therapies are equivalent? Or is the result caused by an inadequate sample size?

Two things can help us decide whether the sample size was large enough in a negative study. First, we could look for a power calculation, ideally conducted prior to the collection of any data. Second, we could look at the width of the confidence intervals in the research paper. But don't try to assess a negative finding by looking only at the *P* value.

A large *P* value does not tell us that a therapy is ineffective because it is difficult to prove a negative (Gardner and Altman, 1986). A good analogy is the "not guilty" verdict in a criminal trial in the United States. Such a verdict does not "prove" innocence. What a not-guilty verdict tells us is that there is insufficient evidence to convict. Similarly, a large *P* value says there is insufficient evidence to conclude that a new therapy is better.

Are Small Sample Sizes Really a Problem?

Inadequate sample sizes do occur quite often. Freiman et al (1992) provides one of the best examples of this. These researchers identified 71 publications that had a negative finding and that used a binary outcome measure such as mortality (live/dead) or side effects (present/absent). They also studied whether the number of research subjects gave them a reasonable chance of detecting a moderate improvement in therapy (a 25% relative shift) or a large improvement in therapy (a 50% relative shift). As an example, suppose that the control group in a research study has a true mortality rate of

Andrology Lab Corner

20%. A moderate improvement would represent a therapy that has a true mortality rate of 15%. A large improvement would represent a therapy that has a true mortality rate of 10%.

The results in Freiman et al were very depressing. Fifty-seven out of 71 studies (80%) had a sample size so small that there was less than a 50-50 chance of correctly identifying a therapy that could produce moderate improvements. Thirty-four out of the 71 studies (49%) had less than a 50-50 chance of identifying a therapy that could produce large improvements.

Think about what this means. Any therapy that can produce a large improvement (eg, cutting mortality or side effects in half) is well worth identifying. But many studies are incapable of reliably detecting this size difference. To add to the depressing news, Freiman et al did a second analysis of more recently produced studies and found a similar proportion of studies with inadequate sample sizes.

Other studies bear the same grim news. In an analysis of 2000 research studies of schizophrenia over a 50-year period, Thornley and Adams (1998) found (using a different criteria) that only 60 (3%) studies had an adequate sample size. Furthermore, they noted that there was no trend toward larger studies in more recent times.

I am unaware of any examination of sample size adequacy for studies specific to andrology. Berndston et al (1997) point out, however, that the outcome measures for mating trials have large inherent variability, which makes the issue of sample size all the more important.

Do We Know the Range of Clinical Indifference?

To properly assess a negative study, we need to define a range of clinical indifference. This is a range of values, when comparing a new therapy to a standard therapy, that is so small that we as clinicians would not want to change our practice.

Another way to define this interval is to define the smallest value that is large enough to have a clinical impact. We might see this referred to as the clinically relevant difference or the minimum important difference. Either way, we need to define a boundary. Inside the boundary, we do not have enough of a reason to change. Outside the boundary, the new therapy is sufficiently helpful to convince our colleagues to adopt it.

Specifying the range of clinical indifference requires that one balance the relative merits of several factors. In a clinical setting, we need to examine the cost and side effect profile of the new and standard therapies and the

Correspondence to: Stephen Simon, PhD, Research Biostatistician, Office of Medical Research, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO 64108.

Received for publication September 13, 2000; accepted for publication September 13, 2000.

severity of the condition being treated. In a laboratory setting, we balance the time and expense of a new method, instrument, or protocol against the amount of benefit that the change would provide. In an epidemiology study, we weigh the prevalence of an exposure and the cost of remediation of that exposure against the amount and severity of morbidity and mortality that it might cause.

Determining the range of clinical indifference is a subjective judgment, and there is no external standard that we can rely on for all situations. This range may even vary from doctor to doctor and from patient to patient. For example, 24 surgeons were asked about 2 types of surgery for operable gastric cancer (Fayers et al, 2000). The first surgery, D1, was less extensive than the alternative, D2, so doctors would only prefer D2 if it demonstrated a marked improvement in survival. The doctors disagreed, however, on how much of an improvement they would need to see before they would adopt D2. Most doctors said that as little as a 5% or 10% absolute increase in survival would justify using D2; others said that as much as a 20% absolute increase would be needed.

Assessing the range of clinical indifference in an andrology study is likely to yield just as much disagreement. How much extra precision in a new computer-aided sperm assessment system would be needed to justify the added expense of buying the new equipment? Different laboratories will be likely to have different answers. How much of an improvement in pregnancy rates would be needed to justify the added expense and trouble of intracytoplasmic sperm injection over in vitro fertilization? Not every couple would (or should) require the same degree of superiority.

Ideally, the authors of the research publication should state their opinion about the range of clinical indifference. This is required, for example, by the CONSORT guidelines for randomized clinical trials (see Begg et al, 1996). When the authors specify the range of clinical indifference, we can then either accept that range or modify it to meet our specific needs. This is not an ideal world though, and the authors will rarely share their judgment about clinical indifference. So we will usually have to come up with a range on our own.

What Does a Power Calculation Tell Us?

Ideally, every study that involves testing a research hypothesis should begin with a power calculation (see Lang and Secic, 1997, page 12). Some researchers will calculate power after the study is complete, but this must be performed very carefully (Zumbo and Hubley, 1998).

Power for a research study is analogous to sensitivity for a diagnostic test. Sensitivity measures the probability that a diagnostic test will be positive if a patient has the disease being tested for. Power measures the probability that a research design will have a positive finding if the

new therapy is indeed effective. The word *effective* means that the difference between the 2 therapies is outside the range of clinical indifference.

Just as we would not rely on a diagnostic test that has low sensitivity, we should also not rely on a research study that has low power. Look for an adequate level of power, say 80% or 90%. Much lower levels of power tell us that the sample size is too small. Be sure that the power calculation specifies a difference that we would consider to be outside the range of clinical indifference, but not too far outside.

If a study has adequate power, and the findings are negative, we can take this as evidence that the new therapy does not have a large enough effect to convince us to change our practice. If the power is not adequate, we should probably not make any permanent decisions. We should probably wait for further publications that evaluate larger numbers of research subjects, or that use research methods that are more precise.

Many publications fail to mention power, which is a shame. For example, in the 2000 schizophrenia trials reviewed in Thornley et al (1998), only 20 (1%) made any mention of statistical power.

A good example of a power calculation appears in a paper by Biljan et al (1996), which examined whether the hypo-osmotic swelling (HOS) test was helpful in predicting fertilization, implantation, miscarriage, and live birth rates. The authors noted that they needed to study at least 320 patients to ensure that the study would have 90% power for detecting a 50% relative difference in the fertilization rate between semen samples with a normal HOS test and samples with an abnormal HOS test. The actual number of patients studied was 326. Thus, if we accept a 50% relative change as being just outside the range of clinical indifference, then we should give a high level of credibility to the mostly negative findings of this study.

What Does a Confidence Interval Tell Us?

A confidence interval places upper and lower limits around a statistical estimate to account for sampling error for that estimate. When comparing a new therapy to a standard therapy or control, the confidence interval provides a range of plausible values for the difference (or ratio) between the 2 therapies. We need to be careful about the interpretation of these intervals. A 95% confidence interval does not imply that there is a 95% probability that the therapy is effective or that 95% of the data lie inside that interval.

Like a *P* value, the confidence interval allows us to assess statistical significance. Does the confidence interval include the null value (the value that implies equivalence)? Typically, the null value is zero if we are com-

puting differences between the new and standard therapies, and 1 if we are computing ratios.

When the interval contains the null value (Figure 1), we have lack of statistical significance. The interval tells us that the null value is plausible. Therefore, it is reasonable to behave as if there were no difference between the new and standard therapies.

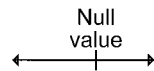


Figure 1. A confidence interval that indicates lack of statistical significance.

If the interval excludes the null value (Figure 2), then the results are statistically significant. The interval tells us that the null value is not plausible. Therefore, it is reasonable to behave as if there is a difference between the new and standard therapies.

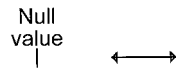


Figure 2. A confidence interval that indicates statistical significance.

Confidence intervals have an advantage over P values in that they also allow us to assess whether the results of an experiment have practical significance (Guyatt, 1995). When the confidence interval lies entirely inside the range of clinical indifference (Figure 3), we know that all plausible values have no clinical impact. With such a confidence interval, we know that the sample size was large enough to rule out a clinically meaningful difference. We can treat the study as a definitive negative finding.

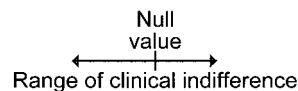


Figure 3. A confidence interval that indicates a definitive negative finding.

When the confidence interval lies partly inside and partly outside the range of clinical indifference (Figure 4), we don't know what to think. It is plausible to behave as if there is no important difference, but it is also plausible to behave as if there *is* an important difference. With this type of confidence interval, we know that the sample size is too small. The research did not have enough precision to distinguish between clinically trivial and clinically important findings.

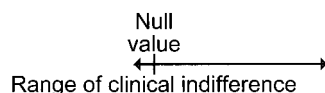


Figure 4. A confidence interval that indicates an inadequate sample size.

The classic example of small sample size is when a confidence interval for a ratio (such as an odds ratio or a relative risk) includes both the values of 1 and 2. What

this tells us is that the sample size is so small that the research design cannot distinguish between no change in risk and a doubling of risk. Equivalently, when we are looking for a decrease in risk, an interval that contains both 0.5 and 1.0 is usually evidence that the sample size is too small.

Some good examples of confidence interval appear in the Biljan et al (1996) paper. The authors present the confidence interval for the odds ratio relating an abnormal HOS test to the fertilization rate as 0.97 to 1.14. Notice first that the interval for this ratio includes the null value of 1.0, implying a lack of statistical significance. If we further accept a range of clinical indifference from 0.5 to 2.0, then this interval rules out the possibility that an abnormal HOS test has a clinically significant effect on the rate of fertilization.

Different readers may not agree about range of clinical indifference proposed above. But notice that even with a much narrower range of clinical indifference (eg, 0.8 to 1.25), the interval still rules out the possibility of a clinically significant finding.

Contrast this, however, with the finding in the same paper about miscarriage rates. The authors found that the miscarriage rate was much worse in the samples with abnormal HOS test results (odds ratio = 0.37, 95% confidence interval = 0.09 to 1.49). This confidence interval is far wider than the previous one because miscarriage was a rare event (only 21 total among all patients). As above, this interval includes the null ratio of 1.0, implying lack of statistical significance. Unlike the fertilization rate, however, the lower limit of the interval for miscarriage extends well outside the range of clinical indifference. The authors correctly argue that the increase in the rate of miscarriage in samples with abnormal HOS results may be clinically important. They also suggest that a new study examining the miscarriage rate with a larger sample size may be warranted.

Conclusion

Negative studies are difficult to interpret. We need to rule out the possibility that the negative finding was caused by an inadequate sample size. The P value by itself does not help us. Look for a power calculation performed prior to data collection. If the study has 80% to 90% power for detecting a change outside the range of clinical indifference, then we know that the sample size was large enough. Also examine the confidence interval for the primary outcome measure in the study. If it lies entirely inside the range of clinical indifference, we know that the sample size was large enough. If the study fails to mention any power calculation and it fails to present confidence intervals, then we do not have enough information to draw any meaningful conclusions.

References

- Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA*. 1996;276:637-639.
- Berndtson WE, Judd JE, Castro AC. Inherent variability among measures of fertility of rats and its implications in the design of mating trials. *J Androl*. 1997;186:717-724.
- Biljan MM, Buckett WM, Taylor CT, Luckas M, Aird I, Kingsland CR, Lewis-Jones DI. Effect of abnormal hypo-osmotic swelling test on fertilization rate and pregnancy outcome in in vitro fertilization cycles. *Fertil Steril*. 1996;66:412-416.
- Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinka B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. *Br J Cancer*. 2000;82:213-219.
- Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial. In: Bailar JC II, Mosteller F, eds. *Medical Uses of Statistics*. 2nd ed. Boston, Mass: NEJM Books; 1992: 357-373.
- Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J Clin Res Ed*. 1986; 292:746-750.
- Guyatt G. Interpreting study results: confidence intervals. *Can Med Assoc J*. 1995;152:169-173.
- Lang TA, Secic M. *How to Report Statistics in Medicine*. Philadelphia, Penn: American College of Physicians; 1997.
- Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50 years. *Br Med J*. 1998;317:1181-1184.
- Zumbo BC, Hubley AM. A note on misconceptions concerning prospective and retrospective power. *The Statistician*. 1998;47:385-388.

Andrology Lab Corner welcomes the submission of unsolicited manuscripts, requested reviews, and articles in a debate format. Manuscripts will be reviewed and edited by the Section Editor. All submissions should be sent to the **Journal of Andrology** Editorial Office. Letters to the editor in response to articles as well as suggested topics for future issues are encouraged.