Understanding results: P-values, confidence intervals, and number need to treat

Lawrence Flechner, Timothy Y. Tseng

Department of Urology, University of California, San Francisco, CA, 1Department of Urology, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

ABSTRACT

Objectives: With the increasing emphasis on evidence-based medicine, the urology literature has seen a rapid growth in the number of high-quality randomized controlled trials along with increased statistical rigor in the reporting of study results. P-values, CI, and number needed to treat (NNT) are becoming increasingly common in the literature. This paper seeks to familiarize the reader with statistical measures commonly used in the evidence-based literature.

Materials and Methods: The meaning and appropriate interpretation of these statistical measures is reviewed through the use of a clinical scenario.

Results: The reader will be better able to understand such statistical measures and apply them to the critical appraisal of the literature.

Conclusions: P-values, CI, and NNT each provide a slightly different estimate of statistical truth. Together, they provide a more complete picture of the true effect observed in a study. An understanding of these measures is essential to the critical appraisal of study results in evidence-based medicine.

Key words: Confidence intervals, evidence-based medicine, number needed to treat, statistical significance

INTRODUCTION

Evidence-based clinical practice requires that practitioners consider the best available evidence in support of potential clinical decisions. The highest category of evidence includes randomized controlled trials.1 With the increasing emphasis on evidence-based medicine, the urology literature has seen a rapid growth in the number of high-quality randomized controlled trials along with increased statistical rigor in the reporting of study results.2 Statistical measures including P-values, confidence intervals (CI), and number needed to treat (NNT) are becoming increasingly common in the urological literature. The meaning and appropriate interpretation of such measures are reviewed through the use of a clinical scenario.

CASE SCENARIO

A 60-year-old otherwise healthy man has just undergone a transrectal ultrasound-guided prostate biopsy for a persistently elevated prostate-specific antigen (PSA) of 6.2 ng/mL. His biopsy pathology was negative for prostate cancer. He now inquires whether there is any medication that he can take to decrease his risk of developing prostate cancer in the future. You are aware that 5-a-reductase inhibitors have been investigated for the chemoprevention of prostate cancer and decide to examine the literature for the current best evidence on this topic.

CLINICAL QUESTION

You define the clinical question using the “PICO” mnemonic. In men with negative prostate biopsies (population), how effective are 5-a-reductase inhibitors (intervention) when compared to no treatment (comparison) for the prevention of prostate cancer (outcome)?
FINDING THE BEST EVIDENCE

You decide to examine the literature on this clinical question by using PubMed. A search for all articles on “5-alpha-reductase inhibitors” yields 1596 articles. A separate search for “prostate cancer” yields 85,759 articles. When these two search terms are combined with the AND function, 350 articles are found. By applying the “Randomized Controlled Trial” article type limit, 21 titles are identified. One article in this list that addresses your clinical question is a randomized controlled trial entitled, “Effect of dutasteride on the risk of prostate cancer.”

In this study’s final analysis, 6729 men between the ages of 50 and 75 years of age with an elevated PSA and negative baseline prostate biopsies were randomized to treatment with dutasteride 0.5 mg or placebo daily. These patients then underwent transrectal ultrasound-guided prostate biopsies at 2 and 4 years, and the number of patients who developed prostate cancer on biopsy was recorded.

EVALUATING THE EVIDENCE

Before looking at the results of the study, you first examine its methodology to determine if the study results would be valid. Using the Consolidated Standards of Reporting Trials (CONSORT) statement as a guide, you find that the methodology of this randomized controlled trial was strong. You now turn your attention to the results of the study.

The study’s primary endpoint was prostate cancer detection on biopsy at 2 and 4 years. Over the entire 4-year study period, 659 of the 3305 men in the dutasteride group and 858 of the 3424 men in the placebo group were found to have prostate cancer. This study reports this result as a relative risk reduction for the development of prostate cancer of 22.8% (95% CI: 15.2-29.8; \( P < 0.001 \)) for patients taking dutasteride compared to patients taking placebo.

WHAT IS A \( P \)-VALUE?

\( P \)-values were introduced in the first-half of the 20th century as the probability of rejecting the null hypothesis that a treatment has no effect when, in actuality, the null hypothesis is true. Stated another way, a \( P \)-value is the probability that an observed difference is due to random chance when the null hypothesis is true. Statisticians Jerzy Neyman and Egon Pearson later described this as a Type I error in contrast to a Type II error where the null hypothesis is accepted when, in actuality, the null hypothesis is false. As statistician Ronald A. Fisher suggested, a \( P \)-value is an index for the strength of the evidence for the tested hypothesis against the null hypothesis. It is not, however, the probability that the null hypothesis is true. Although a \( P \)-value is appropriately considered a statistic interpretable across a range of values, in contemporary experimental studies, “statistical significance” is now conventionally set at a \( P \)-value of <0.05. This means that the null hypothesis is appropriately rejected if the probability of a Type I error is <5%.

In our study example, the hypothesis being tested is whether dutasteride decreases the risk of prostate cancer as detected by biopsy among men at increased risk for prostate cancer. The null hypothesis is therefore that dutasteride does not decrease the risk of prostate cancer. In this particular study, a \( P \)-value of 0.01 or less was used as the predetermined criterion for statistical significance. This means that the null hypothesis in which dutasteride does not decrease the risk of prostate cancer is appropriately rejected if the probability that the null hypothesis is rejected when it is in fact true is <1%. In this study, a relative risk reduction for the development of prostate cancer of 22.8% was found for the patients taking dutasteride compared to placebo. The \( P \)-value was <0.001. Therefore, the probability of rejecting the null hypothesis when it was in fact true was <0.1%, and the result is statistically significant by the study’s predetermined criterion. We can therefore have confidence that dutasteride does reduce the risk of prostate cancer in this patient population by 22.8%.

WHAT IS A CONFIDENCE INTERVAL?

Although a \( P \)-value is helpful in determining the reliability with which the null hypothesis can be rejected and therefore the strength of the observed result, it does not provide information regarding the precision of the result. To address this issue, CI can be calculated around the point estimate of the result to provide a range of values within which the true value is certain to exist with a given level of confidence. A wide CI suggests an imprecise result and indicates that the results should be interpreted with caution regardless of statistical significance. In accordance with the conventional acceptance of statistical significance at a \( P \)-value of 0.05 or 5%, CI are frequently calculated at a confidence level of 95%.

In general, if an observed result is statistically significant at a \( P \)-value of 0.05, then the null hypothesis should not fall within the 95% CI.

In our study example, the result is expressed as a relative risk reduction percentage. The null hypothesis would therefore be a relative risk reduction of 0%. In this study, the relative risk reduction over four years in men treated with dutasteride was 22.8% and the 95% CI was 15.2-29.8%. The null hypothesis of no risk reduction with dutasteride does not fall within the 95% CI, and the result is therefore statistically significant at a \( P \)-value of <0.05. Furthermore, the 95% CI is relatively narrow and suggests that the true relative risk reduction with dutasteride is between 15.2% and 29.8%. The fact that there is a reasonably large relative
risk reduction even at the low end of the confidence interval suggests that the result is not only statistically meaningful, but also clinically meaningful.

WHAT IS NUMBER NEEDED TO TREAT?

Although measures such as relative and absolute risk provide discrete statistical estimates of treatment effect, the magnitude of a treatment’s clinical effect is often difficult to determine through these statistics. The NNT, however, is a straightforward measure that conveys an estimate of a treatment’s clinical effect. For a study of treatment effect, the NNT is defined as the number of patients that must be treated in order to prevent one additional adverse outcome. It is easily calculated as the inverse of the absolute risk difference between treatment groups. In general, a large treatment effect is associated with a small NNT while a small treatment effect is associated with a large NNT.

In our study example, after 4 years of treatment, 659 of the 3305 men in the dutasteride group and 858 of the 3424 men in the placebo group were found to have prostate cancer. In this study, the absolute risk of developing prostate cancer in the treatment group was 659/3305 = 19.9%. The absolute risk of prostate cancer in the placebo group was 858/3424 = 25.1%. The absolute risk difference was therefore 25.1-19.9 = 5.2% or 0.052. The NNT is then the inverse of the absolute risk difference, which was 1/0.052 = 19.2. Therefore, approximately 19 men need to be treated with dutasteride to prevent one additional case of prostate cancer at 4 years.

It is important to note that NNT does not incorporate the time to the observed effect. In many cases, a longer duration of treatment will increase the reduction in risk. In our sample study, the absolute risk difference at 2 years was only 3.8%. At 2 years, the NNT would therefore have been 26.3 patients. Because timing may change the result, it is important to report the timing of the observed effect along with the NNT. Furthermore, just as CI can be calculated to provide a more complete picture of point estimates of effect such as relative risk reduction, so too can confidence intervals be calculated for NNT. If the observed effect is statistically significant, this can be accomplished by taking the inverse of the upper and lower limits of the CI of the absolute risk difference. The calculation of CI for a nonsignificant result has been described previously and is beyond the scope of this paper.

APPLYING THE RESULTS TO THE CARE OF YOUR PATIENT

According to our sample study, the use of dutasteride in men with an elevated PSA between 2.5 and 10 ng/mL and a negative baseline prostate biopsy confers a significant relative risk reduction for the development of prostate cancer of 22.8% (95% CI: 15.2-29.8; P<0.001) at 4 years compared to patients who received placebo. The number of men that need to be treated with dutasteride to prevent one additional diagnosis of prostate cancer in 4 years is 19. This result appears to be not only statistically significant, but also robust with a narrow confidence interval that is relatively far from the null hypothesis.

A distinction should be made, however, between statistical significance and clinical significance. In general, a larger sample size will reduce the probability of a Type I error and therefore decrease the P-value. This raises the possibility that a clinically insignificant difference may be found to be statistically significant. The actual magnitude of the treatment effect should always be considered separately from considerations of statistical significance. The NNT is one tool that addresses the magnitude of the clinical effect. In particular, NNT places study results in a context that allows discussion of broader issues such as the cost of treatment to society and the exposure of patients to potential adverse reactions. In this study, the NNT is relatively low at 19 and provided additional benefits with regard to outcomes related to benign prostatic hyperplasia with a low side effect profile.

Based upon the results of this study, treatment of your patient with dutasteride may be appropriate as a means to decrease his risk of developing prostate cancer in the future.

CONCLUSION

As the quality of the evidence in the urological literature improves, urologists are increasingly faced with statistical measures such as P-values, CI, and NNT. Each provides a slightly different nuanced estimate of statistical truth. Together, these statistical measures allow a more complete picture of the true effect observed in any given study. An understanding of these measures is therefore essential to the critical appraisal of the study literature in evidence-based medicine.

REFERENCES

1. Scales CD Jr, Preminger GM, Keitz SA, Dahn P. Evidence based clinical practice: A primer for urologists. J Urol 2007;178:775-82.
2. Scales CD Jr, Norris RD, Keitz SA, Peterson BL, Preminger GM, Vieweg J, et al. A critical assessment of the quality of reporting of randomized, controlled trials in the urology literature. J Urol 2007;177:1090-4.
3. Krupski TL, Dahn P, Fesperman SF, Schardt CM. How to perform a literature search. J Urol 2008;179:1264-70.
4. Krupski TL, Schardt CM, Fesperman SF, Dahn P; Evidence Based Urology Working Group. Evidence-based urology in practice: How to use PubMed effectively. BJU Int 2009;103:1156-9.
5. Andreole GL, Bostwick DG, Brawley OW, Gomella LG, Marberger M, Montorsi F, et al. Effect of dutasteride on the risk of prostate cancer. N Engl J Med 2010;362:1192-202.
6. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials. Ann Intern Med 2010;152:726-32.
7. Sterne JA, Davey Smith G. Sifting the evidence-what’s wrong with significance tests? BMJ 2001;322:226-31.
8. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med J (Clin Res Ed) 1983;286:1489-93.
9. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1998;318:1728-33.
10. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317:1309-12.

How to cite this article: Flechner L, Tseng TY. Understanding results: P-values, confidence intervals, and number needed to treat. Indian J Urol 2011;27:532-5.

Source of Support: Nil, Conflict of Interest: None declared.