The problem with P values: defining clinical vs. statistical significance

By Lindsay Kobayashi
Posted: June 24, 2015

The problem with P values

Today we warmly welcome guest writer Sean Sinden to PLOS Public Health Perspectives. His biography is at the end of the post.

The practice of null hypothesis testing has traditionally been used to interpret the results of studies in a wide variety of scientific fields. Briefly, significance testing involves the calculation of an outcome statistic, known as the *P* value. The *P* value represents the probability of finding a difference, by chance, between two sets of values larger than that which was observed, assuming no difference between the two sets of values. Conventionally, if that probability is less than 0.05 the outcome is deemed "statistically significant". If this sounds confusing, it's because it is!

P values are commonly misinterpreted and misused to answer research questions, but in actuality they fail to provide much information to the reader (1). This method of statistical analysis has been met with criticism throughout its history and increasingly so in the past few decades. The shortcomings of significance testing have been identified in both the academic and non-academic literature. I encourage anyone interested in the mechanistic limitations of significance testing to seek out such resources.

> *The P value doesn't begin to explain the importance of a study's outcome or the amount of the effect observed, though many researchers mistakenly believe it to.*

Null hypothesis significance testing does not explain how much better – or worse – a group of individuals did compared with another group, just that there was a difference between the two groups (2). In short, significance testing only gives us statistical significance and says nothing about a study's practical significance or clinical applicability. There are numerous examples of a statistically significance result having no practical significance and vice-versa, two of which I have included below:

A primary HIV prevention medication known as Truvada – a combination of tenofovir and emtricitabine – was approved by the US Food and Drug Administration (FDA) in 2012. The two major side effects of this drug, taken as a daily oral dose, are a mild, non-progressive decrease in kidney function and a small decrease in bone mineral density (3). The effect on kidney function was found to be statistically significant ($P = 0.02$), but was considered to be "sub-clinical" by the authors (4). In a separate study, the same drug caused a small, but statistically significant decrease in bone mineral density from baseline, the clinical significance of which was unknown but was not associated with an increased occurrence of bone fracture during the study (5). Hence, Truvada does not appear to have any significant clinical side effects (see the Table below). However, in a cohort of 1603 at-risk individuals in the US, 185 people – roughly 12% of respondents – cited concern about side effects as their reason for not taking the medication (6). This is an alarming misconception for a disease that infects approximately 50,000 Americans and 2 million people globally per year (12,13), highlighting the need for better knowledge translation to improve public understanding of research outcomes.

Table. Statistical and clinical significance of Truvada side effects

Side effect	Statistically significant?	Clinically significant?
Reduced kidney function (4)	Yes: $P = 0.02$	No: the 'sub-clinical' effects on kidney function were reversible and manageable
Decreased bone mineral density (5)	Yes: $P = 0.02$ for a 3% decrease in bone mineral density	Unclear: no increase in fractures associated with loss of bone mineral density in the study

On the other hand, research results that are not found to be statistically significant may have clinical applications. In the field of exercise physiology, a 1% change in performance might be considered clinically applicable.

> *To put this in perspective, a difference of 1% in the 100m sprint is the difference between Donovan Bailey's 1996 Olympic record of 9.84 seconds, and Asafa Powell's 2007 world record of 9.74 seconds.*

Such a difference might not be deemed statistically significant in a typical research study, due to the difficulty of obtaining precision around such a small difference as a result of limited sample size or large variability in measurement (7). Another good example is from a study where researchers found a 2.9-minute improvement in a cycling time trial lasting 160 minutes with a supplement intervention. This difference was not statistically significant (8). However, if the researchers were to look at the results from an application standpoint, their

conclusion may have been different: the 2.9 minute time improvement translated into a 1.8% increase in performance, suggesting that competitive athletes would probably benefit from the supplement intervention and that further investigation is warranted (7).

In recent years there has been a move away from using null hypothesis significance testing alone, or at all, in many scientific fields. Making inferences using magnitude-based measures such as confidence intervals, which is becoming increasingly popular, allows researchers to estimate the **size** of an effect in relation to clinical and practical importance (1,9,10). *P*-values may still have a place in our statistical assessment of study outcomes, but not should be the defining value for accepting or rejecting an outcome (11). Researchers, scientific writers, and the public need to look at the outcomes of research from a practical standpoint, and not only use the outdated and dichotomous view of statistical significance. Do not make conclusions about research outcomes solely based on significance testing without considering the practical significance and applicability of the observed effect.

Sean Sinden is an MSc student in the School of Kinesiology at the University of British Columbia, specialising in exercise physiology. Sean Sinden's research is focusing on the physiology of doping and the impact of environmental conditions on asthmatic athletes. He is also interested in knowledge translation, science writing, and public understanding of scientific findings.

Twitter: @seanmsinden

References

1. Batterham AM, Hopkins WG. Making Meaningful Inferences About Magnitudes. Int J Sports Physiol Perform. 2006;1(1):50–7.
2. Wilhelmus KR. Beyond the P. I: Problems with probability. J Cataract Refract Surg. 2004;30(9):2005–6.
3. Krakower DS, Jain S, Mayer KH. Antiretrovirals for Primary HIV Prevention: the Current Status of Pre- and Post-exposure Prophylaxis. Curr HIV/AIDS Rep. 2015;12(1):127–38.
4. Solomon MM, Lama JR, Glidden DV, Mulligan K, McMahan V, Liu AY, et al. Changes in renal function associated with oral emtricitabine/tenofovir disoproxil fumarate use for HIV pre-exposure prophylaxis. AIDS. 2014;28(6):851–9.
5. Liu AY, Vittinghoff E, Sellmeyer DE, Irvin R, Mulligan K, Mayer K, et al. Bone Mineral Density in HIV-Negative Men Participating in a Tenofovir Pre-Exposure Prophylaxis Randomized Clinical Trial in San Francisco. PLoS ONE. 2011;6(8):e23688–11.
6. Grant RM, Anderson PL, McMahan V, Liu A, Amico KR, Mehrotra M, et al. Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: a cohort study. Lancet Infect Dis. 2014;14(9):820–9.
7. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc. 1999;31(3):472–85.
8. Madsen K, MacLean DA, Kiens B, Christensen D. Effects of glucose, glucose plus branched-chain amino acids, or placebo on bike performance over 100 km. J Appl Physiol. 1996;81(6):2644–50.
9. Wilkinson M, Winter EM. Clinical and practical importance vs statistical significance: Limitations of conventional statistical inference. Int J Ther Rehabil. 2014;21(10):488-95.
10. Cumming G. The new statistics: why and how. Psychol Sci. 2014;25(1):7–29.

11. Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology. 2013;24(1):62–8.
12. Joint United Nations Programme on HIV/AIDS. *UNAIDS Fact Sheet: Global Statistics.* http://www.unaids.org/en/resources/campaigns/2014/2014gapreport/factsheet/ (accessed June 11, 2015)
13. Centre for Disease Control and Prevention. *HIV in the United States: at a glance.* http://www.cdc.gov/hiv/statistics/basics/ataglance.html (accessed June 11, 2015)

About Lindsay Kobayashi

View all posts by Lindsay Kobayashi →

This entry was posted in Epidemiology, Guest Posts, Science Outreach and tagged communication, epidemiology, knowledge translation, public health. Bookmark the permalink.

2 Responses to *The problem with P values: defining clinical vs. statistical significance*

Pingback: *The problem with P values: defining clinical vs. statistical significance [Reblog] | Mulford Library Blog*

Pingback: *The problem with P values: defining clinical vs. statistical significance | WilliamSiebold.com*