Journal of Physiotherapy 68 (2022) 1–4

AUSTRALIAN PHYSIOTHERAPY ASSOCIATION

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Editorial

Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors[1]

Mark R Elkins [a,b], Rafael Zambelli Pinto [a,c], Arianne Verhagen [a,b], Monika Grygorowicz [d], Anne Söderlund [e], Matthieu Guemann [f], Antonia Gómez-Conesa [g], Sarah Blanton [h], Jean-Michel Brismée [i], Clare Ardern [j], Shabnam Agarwal [k], Alan Jette [l], Sven Karstens [m], Michele Harms [n], Geert Verheyden [o], Umer Sheikh [p]

[a] International Society of Physiotherapy Journal Editors executive; [b] Journal of Physiotherapy; [c] Brazilian Journal of Physical Therapy/Revista Brasileira de Fisioterapia; [d] BMC Sports Science, Medicine and Rehabilitation; [e] European Journal of Physiotherapy; [f] European Rehabilitation Journal; [g] Fisioterapia; [h] Journal of Humanities in Rehabilitation; [i] Journal of Manual & Manipulative Therapy; [j] Journal of Orthopaedic & Sports Physical Therapy; [k] Journal of Society of Indian Physiotherapists; [l] Physical Therapy; [m] physioscience; [n] Physiotherapy; [o] Physiotherapy Research International; [p] The Journal of Physiotherapy & Sports Medicine

Null hypothesis statistical tests are often conducted in healthcare research,[1] including in the physiotherapy field.[2] Despite their widespread use, null hypothesis statistical tests have important limitations. This co-published editorial explains statistical inference using null hypothesis statistical tests and the problems inherent to this approach; examines an alternative approach for statistical inference (known as estimation); and encourages readers of physiotherapy research to become familiar with estimation methods and how the results are interpreted. It also advises researchers that some physiotherapy journals that are members of the International Society of Physiotherapy Journal Editors (ISPJE) will be expecting manuscripts to use estimation methods instead of null hypothesis statistical tests.

What is statistical inference?

Statistical inference is the process of making inferences about populations using data from samples.[1] Imagine, for example, that some researchers want to investigate something (perhaps the effect of an intervention, the prevalence of a comorbidity or the usefulness of a prognostic model) in people after stroke. It is unfeasible for the researchers to test all stroke survivors in the world; instead, the researchers can only recruit a sample of stroke survivors and conduct their study with that sample. Typically, such a sample makes up a miniscule fraction of the population, so the result from the sample is likely to differ from the result in the population.[3] Researchers must therefore use their statistical analysis of the data from the sample to infer what the result is likely to be in the population.

What are null hypothesis statistical tests?

Traditionally, statistical inference has relied on null hypothesis statistical tests. Such tests involve positing a null hypothesis (eg, that there is no effect of an intervention on an outcome, that there is no effect of exposure on risk or that there is no relationship between two variables). Such tests also involve calculating a *p*-value, which quantifies the probability (if the study were to be repeated many times) of observing an effect or relationship at least as large as the one that was observed in the study sample, if the null hypothesis is true. Note that the null hypothesis refers to the population, not the study sample.

Because the reasoning behind these tests is linked to imagined repetition of the study, they are said to be conducted within a 'frequentist' framework. In this framework, the focus is on how much a statistical result (eg, a mean difference, a proportion or a correlation) would vary among the repeats of the study. If the data obtained from the study sample indicate that the result is likely to be similar among the imagined repeats of the study, this is interpreted as an indication that the result is in some way more credible.

One type of null hypothesis statistical test is *significance testing*, developed by Fisher.[4–6] In significance testing, if a result at least as large as the result observed in the study would be unlikely to occur in the imagined repeats of the study if the null hypothesis is true (as reflected by $p < 0.05$), then this is interpreted as evidence that the null hypothesis is false. Another type of null hypothesis statistical test is *hypothesis testing*, developed by Neyman and Pearson.[4–6] Here, two hypotheses are posited: the null hypothesis (ie, that there is no difference in the population) and the alternative hypothesis (ie, that there is a difference in the population). The *p*-value tells the researchers which hypothesis to accept: if $p \geq 0.05$, retain the null hypothesis; if $p < 0.05$, reject the null hypothesis and accept the alternative. Although these two approaches are mathematically similar, they differ substantially in how they should be interpreted and reported. Despite this, many researchers do not recognise the distinction and analyse their data using an unreasoned hybrid of the two methods.

Problems with null hypothesis statistical tests

Regardless of whether significance testing or hypothesis testing (or a hybrid) is considered, null hypothesis statistical tests have numerous problems.[4,5,7] Five crucial problems are explained in Box 1. Each of these problems is fundamental enough to make null hypothesis statistical tests unfit for use in research. This may surprise many readers, given how widely such tests are used in published research.[1,2]

It is also surprising that the widespread use of null hypothesis statistical tests has persisted for so long, given that the problems in Box 1 have been repeatedly raised in healthcare journals for decades,[8,9] including physiotherapy journals.[10,11] There has been some movement away from null hypothesis statistical tests, but the use of alternative methods of statistical inference has increased slowly over decades, as seen in analyses of healthcare research, including physiotherapy trials.[2,12] This is despite the availability of alternative

[1] This article is a joint publication by several member journals of the *International Society of Physiotherapy Journal Editors*.

https://doi.org/10.1016/j.jphys.2021.12.001

1836-9553/© 2021 Published by Elsevier B.V. on behalf of Australian Physiotherapy Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Box 1. Problems with null hypothesis statistical tests. Modified from Herbert (2019).[26]

Problem	Explanation
A *p*-value is not the probability that a hypothesis is (or is not) true	• Researchers need to know the probability that the null hypothesis is true given the data observed in their study. • A *p*-value instead is the probability of observing the observed data given that the null hypothesis is true. • These two probabilities may seem interchangeable but they are not. • Therefore, *p*-values do not equate to a probability that researchers need to know.
A *p*-value does not constitute evidence	• As explained above, a *p*-value is the probability of an observation given that a particular hypothesis is true. • Any probability of an observation given a particular hypothesis cannot provide evidence for or against that hypothesis. • It is only possible to quantify the strength of evidence for a hypothesis by comparing it with another hypothesis.
Statistically significant findings are not very replicable	• If a study is repeated with a new random sample from the same population, the result (and therefore the *p*-value) is likely to vary. • Imagine a study with a *p*-value between 0.005 and 0.05. • If this study was repeated with a new random sample from the same population, there would be a 33% chance that the *p*-value would be non-significant.[27]
In most clinical trials, the null hypothesis must be false	• The null hypothesis is that the effect of interest is exactly nil. • Almost all interventions would be expected to have some effect, even if that effect was trivially small. • Almost all trials (even those with the most robust methods) would be expected to have some bias, even if that bias was trivially small. • All trials should therefore identify an effect (because the null hypothesis is not true, ie, the effect of interest is not exactly nil). • This implies that every statistically non-significant result is actually a failure to detect an effect that does exist.
Researchers need information about the size of effects	• Researchers need to know more than just whether an effect does or does not exist. • Researchers need to know about the size of the effect. • A *p*-value gives no information about the size or direction of an effect.

methods of statistical inference and promotion of those methods in statistical, medical and physiotherapy journals.[10,13–16]

Estimation as an alternative approach for statistical inference

Although there are multiple alternative approaches to statistical inference,[13] the simplest is estimation.[17] Estimation is based on a frequentist framework but, unlike null hypothesis statistical tests, its aim is to estimate parameters of populations using data collected from the study sample. The uncertainty or imprecision of those estimates is communicated with confidence intervals.[10,14]

A confidence interval can be calculated from the observed study data, the size of the sample, the variability in the sample and the confidence level. The confidence level is chosen by the researcher, conventionally at 95%. This means that if hypothetically the study were to be repeated many times, 95% of the confidence intervals would contain the true population parameter. Roughly speaking, a 95% confidence interval is the range of values within which we can be 95% certain that the true parameter in the population actually lies.

Confidence intervals are often discussed in relation to treatment effects in clinical trials,[18,19] but it is possible to put a confidence interval around any statistic, regardless of its use, including mean difference, risk, odds, relative risk, odds ratio, hazard ratio, correlation, proportion, absolute risk reduction, relative risk reduction, number needed to treat, sensitivity, specificity, likelihood ratios, diagnostic odds ratios, and difference in medians.

Interpretation of the results of the estimation approach

To use the estimation approach well, it is not sufficient simply to report confidence intervals. Researchers must also interpret the relevance of the information portrayed by the confidence intervals and consider the implications arising from that information. The path of migration of researchers from statistical significance and *p*-values to estimation methods is littered with examples of researchers calculating confidence intervals at the behest of editors, but then ignoring the confidence intervals and instead interpreting their study's result dichotomously as statistically significant or non-significant depending on the *p*-value.[20] Interpretation is crucial.

Some authors have proposed a ban on terms related to interpretation of null hypothesis statistical testing. One prominent example is an editorial published in *The American Statistician*,[13] which introduced a special issue on statistical inference. It states:

> The *American Statistical Association Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

This may seem radical and unworkable to researchers with a long history of null hypothesis statistical testing, but many concerns can be allayed. First, such a ban would not discard decades of existing research reported with null hypothesis statistical tests; the data generated in such studies maintain their validity and will often be reported in sufficient detail for confidence intervals to be calculated. Second, reframing the study's aim involves a simple shift in focus from whether the result is statistically significant to gauging how large and how precise the study's estimate of the population parameter is. (For example, instead of aiming to determine whether a treatment has an effect in stroke survivors, the aim is to estimate the size of the average effect. Instead of aiming to determine whether a prognostic model is predictive, the aim is to estimate how well the model predicts.) Third, the statistical imprecision of those estimates can be calculated readily. Existing statistical software packages

Box 2. Resources that provide additional information to respond to questions about the transition from null hypothesis statistical tests to estimation methods.

Question	Resources
Where can I find more detailed information about null hypothesis statistical testing and its problems?	This short paper details the problems inherent in significance testing and hypothesis testing.[25] https://doi.org/10.1016/j.jphys.2019.05.001
Is there widespread recognition of these problems and the need for an alternative?	This American Statistical Association's statement on *p*-values[28] shows that the problems are widely recognised by statisticians. Numerous fields of research have recognised the need to move beyond significance testing, such as medicine,[29] specific medical subdisciplines,[30,31] nursing,[32] psychology,[33] neuroscience,[34] pharmacy,[35] toxicology,[36] anthropology[37] and animal research.[38]
Is there a publication that explains confidence intervals from first principles?	These two editorials explain confidence intervals for continuous and dichotomous variables:[10,14] https://doi.org/10.1016/S0004-9514(14)60334-2 https://doi.org/10.1016/s0004-9514(14)60292-0
Are there published examples of how confidence intervals should be interpreted?	These two short papers explain confidence intervals and show examples of how they can be described in words:[18,19] https://doi.org/10.1016/j.bjpt.2019.01.003 https://www.jospt.org/doi/10.2519/jospt.2019.0706
How can I calculate confidence intervals from my raw data?	Existing statistical software packages already calculate confidence intervals, including free software such as R.[21,22]
How can I quickly calculate confidence intervals from the summary data in a published paper?	A free Excel-based confidence interval calculator is available to download from the PEDro website: https://pedro.org.au/english/resources/confidence-interval-calculator/

already calculate confidence intervals, including free software such as R.[21,22] Lastly, learning to interpret confidence intervals is relatively straightforward.

Many researchers and readers initially come to understand how to interpret confidence intervals around estimates of the effect of a treatment. In a study comparing a treatment versus control with a continuous outcome measure, the study's best estimate of the effect of the treatment is usually the average between-group difference in outcome. To account for the fact that estimates based on a sample may differ by chance from the true value in the population, the confidence interval provides an indication of the range of values above and below that estimate where the true average effect in the relevant clinical population may lie. The estimate and its confidence interval should be compared against the 'smallest worthwhile effect' of the intervention on that outcome in that population.[23] The smallest worthwhile effect is the smallest benefit from an intervention that patients feel outweighs its costs, risk and other inconveniences.[23] If the estimate and the ends of its confidence interval are all more favourable than the smallest worthwhile effect, then the treatment effect can be interpreted as typically considered worthwhile by patients in that clinical population. If the effect and its confidence interval are less favourable than the smallest worthwhile effect, then the treatment effect can be interpreted as typically considered trivial by patients in that clinical population. Results with confidence intervals that span the smallest worthwhile effect indicate a benefit with uncertainty about whether it is worthwhile. Results with a narrow confidence interval that spans no effect indicate that the treatment's effects are negligible, whereas results with a wide confidence interval that spans no effect indicate that the treatment's effects are uncertain. For readers unfamiliar with this sort of interpretation, some clear and non-technical papers with clinical physiotherapy examples are available.[10,14,18,19]

Interpretation of estimates of treatment effects and their confidence intervals relies on knowing the smallest worthwhile effect (sometimes called the minimum clinically important difference).[23] For some research questions, such a threshold has not been established or has been established with inadequate methods. In such cases, researchers should consider conducting a study to establish the threshold or at least to nominate the threshold prospectively.

Readers who understand the interpretation of confidence intervals around treatment effect estimates will find interpretation of confidence intervals around many other types of estimates quite familiar. Roughly speaking, the confidence interval indicates the range of values around the study's main estimate where the true population result probably lies. To interpret a confidence interval, we simply describe the practical implications of all values inside the confidence interval.[24] For example, in a diagnostic test accuracy study, the positive likelihood ratio tells us how much more likely a positive test finding is in people who have the condition than it is in people who do not have the condition. A diagnostic test with a positive likelihood ratio greater than about 3 is typically useful and greater than about 10 is very useful.[25] Therefore, if a diagnostic test had a positive likelihood ratio of 4.8 with a 95% confidence interval of 4.1 to 5.6, we could anticipate that the true positive likelihood ratio in the population is both useful and similar to the study's main estimate. Conversely, if a study estimated the prevalence of depression in people after anterior cruciate ligament rupture at 40% with a confidence interval from 5% to 75%, we may conclude that the main estimate is suggestive of a high prevalence but too imprecise to conclude that confidently.

ISPJE member journals' policy regarding the estimation approach

The executive of the ISPJE strongly recommends that member journals seek to foster use of the estimation approach in the papers they publish. In line with that recommendation, the editors who have co-authored this editorial advise researchers that their journals will expect manuscripts to use estimation methods instead of null hypothesis statistical tests. We acknowledge that it will take time to make this transition, so editors will give authors the opportunity to revise manuscripts to incorporate estimation methods if the manuscript seems otherwise potentially viable for publication. Editors may assist authors with those revisions where required.

Readers who require more detailed information to address questions about the topics raised in this editorial are referred to the resources in Box 2, such as the Research Note on the problems of significance and hypothesis testing[25] and an excellent textbook that addresses confidence intervals and the application of estimation methods in various research study designs with clinical physiotherapy examples.[26] Both are readily accessible to researchers and clinicians without any prior understanding of the issues.

Quantitative research studies in physiotherapy that are analysed and interpreted using confidence intervals will provide more valid and relevant information than those analysed and interpreted using null hypothesis statistical tests. The estimation approach is therefore of great potential value to the researchers, clinicians and consumers who rely upon physiotherapy research, and that is why ISPJE is recommending that member journals foster the use of estimation in the articles they publish.

Competing interests: Nil.

Source(s) of support: Nil.

Acknowledgements: We thank Professor Rob Herbert from Neuroscience Research Australia (NeuRA) for his presentation to the ISPJE on this topic and for comments on a draft of this editorial.

Provenance: Invited. Not peer reviewed.

Correspondence: Mark R Elkins, Centre for Education & Workforce Development, Sydney Local Health District, Sydney, Australia. Email: mark.elkins@sydney.edu.au

References

1. Nickerson RS. *Psychol Methods*. 2000;5:241–301.
2. Freire APCF, et al. *Braz J Phys Ther*. 2019;23:302–310.
3. Altman DG, Bland JM. *BMJ*. 2014;349.
4. Barnett V. *Comparative Statistical Inference*. London, New York: Wiley; 1973.
5. Royall RM. *Statistical Evidence: A Likelihood Paradigm*. 1st ed. London, New York: Chapman & Hall; 1997.
6. Gigerenzer G. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge, England: Cambridge University Press; 1989.
7. Goodman SN, Royall R. *Am J Public Health*. 1988;78:1568–1574.
8. Ziliak S, McCloskey D. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, USA: University of Michigan Press; 2008.
9. Hubbard R. *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. Thousand Oaks, USA: Sage; 2016.
10. Herbert RD. *Aust J Physiother*. 2000;46:229–235.
11. Maher CG, et al. *Phys Ther*. 2004;84:644–654.
12. Yi D, et al. *PLoS One*. 2015;10:e0140159.
13. Wasserstein RL, et al. *Am Stat*. 2019;73(Suppl. 1):1–19.
14. Herbert RD. *Aust J Physiother*. 2000;46:309–313.
15. Sim J, Reid N. *Phys Ther*. 1999;79:186–195.
16. Rothman KJ. *Eur J Epidemiol*. 2016;31:443–444.
17. Cumming G. Multivariate applications series. New York: Routledge; 2012.
18. Kamper SJ. *Braz J Phys Ther*. 2019;23:277.
19. Kamper SJ. *J Orthop Sports Phys Ther*. 2019;49:763–764.
20. Fidler F, et al. *Psychol Sci*. 2004;15:119–126.
21. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
22. RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, USA. http://www.rstudio.com/
23. Ferreira M. *J Physiother*. 2018;64:272–274.
24. Amrhein V, et al. *Nature*. 2019;567:305–307.
25. Herbert R. *J Physiother*. 2019;65:178–181.
26. Herbert RD, et al. *Practical Evidence-Based Physiotherapy*. 2nd ed. Oxford: Elsevier; 2011.
27. Boos DD, Stefanski LA. *Am Stat*. 2011;65:213–221.
28. Wasserstein R, Lazar N. *Am Stat*. 2016;70:129–133.
29. International Committee of Medical Journal Editors. ICMJE recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. 2013. http://www.icmje.org/icmje-recommendations.pdf
30. McGough JJ, Faraone SV. *Psychiatry*. 2009;6:21.
31. Hayat MJ, et al. *Otol Neurotol*. 2020;41:578–579.
32. Hayat MJ, et al. *Res Nurs Health*. 2019;42:244–245.
33. Cumming G, et al. *Aust J Psychol*. 2012;64:138–146.
34. Calin-Jageman RJ, Cumming G. *eNeuro*. 2019;6:eNeuro.0205-19.2019.
35. Schreiber JB. *Res Social Adm Pharm*. 2020;16:591–594.
36. Erickson RA, Rattner BA. *Environ Toxicol Chem*. 2020;39:1657–1669.
37. Smith RJ. *Am J Phys Anthropol*. 2020;172:521–527.
38. Du Sert NP, et al. *PLoS Biology*. 2020;18:e3000411.