

University of Groningen

Statistics in publishing

Broekstra, Dieuwke C.; de Boer, Michiel R.; Stunt, Jonáh J.

Published in:
Journal of Hand Surgery: European Volume

DOI:
[10.1177/17531934221095377](https://doi.org/10.1177/17531934221095377)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Broekstra, D. C., de Boer, M. R., & Stunt, J. J. (2022). Statistics in publishing: the (mis)use of the p-value (part 1). *Journal of Hand Surgery: European Volume*, 47(6), 677-680.
<https://doi.org/10.1177/17531934221095377>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Statistics in publishing: the (mis)use of the p -value (part 1)

Journal of Hand Surgery
(European Volume)
2022, Vol. 47(6) 677–680
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17531934221095377
journals.sagepub.com/home/jhs



Introduction

In hand surgery research, most studies, whether observational studies or randomized controlled trials (RCT), are aimed at finding out whether there is an effect (association or difference) of a certain determinant on a specific outcome. This is usually determined using null-hypothesis significance testing (NHST), in which a p -value <0.05 is considered as evidence that the findings are significant. Although this method is widely used, it has been criticized since its inception. The critique has been mainly focused on the misuse of NHST, but also more conceptually on the method itself. In part 1 of this two-part article, we discuss some examples of how the p -value can be misused, using a simulated dataset partly based on real data from an RCT (Broekstra et al., 2022). In part 2, we will discuss the conceptual criticism and offer some guidance on alternatives.

In this example study, women with a distal radial fracture were randomized either to an intervention (cast + rehabilitation programme) or control (cast only) group in a 1:1 ratio. The intervention was aimed at restoring hand function, which was measured using the Patient-Rated Wrist Evaluation (PRWE), a validated patient-reported outcome measure for determining hand function in patients with wrist problems, with a score ranging between 0 (no problems) and 100 (severe problems).

Testing for baseline differences

One of the first steps that most researchers would take after gathering their data from an RCT is to look for any differences between the randomized groups that they want to compare, whether this be in terms of demographics, clinical characteristics or co-treatments received. The reasons given are (1) to make sure the randomization was successful, and (2) to examine whether potentially found differences are critical. Although this may seem a reasonable thing to do, it is not recommended to do this as comparing baseline characteristics of an RCT using NHST is not necessary and even problematic (Altman and Doré, 1990; Moher et al., 2010). Why may this be the case?

First, a statistically significant difference at baseline, for instance there is a difference between the two groups in terms of age, does not mean that the randomization process was incorrect. Because of chance, even correctly executed randomization with blinded allocation can yield differences at baseline resulting in differences in age, gender or injury patterns. This is especially the case in small samples as shown in our example (Table 1). Additionally, there is no cut-off point at which we can conclude that the randomization had failed, or that the statistically significant difference is a result of chance. It may be better to define how differences at baseline are treated during the design of the study, which brings us to the second reason often given by researchers to test for baseline differences.

In contrast to what these researchers believe, the decision to adjust for specific variables should not be made based on statistical significance (i.e. a p -value <0.05 when comparing baseline differences between groups), but rather on the conceptual framework determining whether this variable is prognostic for the outcome. In other words, one should determine whether variables might be of influence on the outcome before the study is even conducted and account for these variables in the analysis.

If we look at Table 1 again, using NHST we will conclude that there are no differences at baseline between the two groups in terms of age, fracture type or affected side, as indicated by p -values >0.05 . Therefore, some authors might conclude that it is better not to adjust the analysis for any co-variables. However, conceptually, one might expect that all these variables included in Table 1 should influence functional outcome. It may therefore be better to adjust the analyses for all these variables.

Based on the results of univariable linear regression analysis (which is in fact equal to an independent t -test in this case), we would conclude that the intervention is beneficial since there is a statistically significant difference between the groups (Table 2, unadjusted model; $p < 0.001$). However, if we would base our decision on the conceptual framework instead of the p -value and adjust for the other

Table 1. Descriptive statistics of the two study groups.

	Intervention (<i>n</i> = 33)	Control (<i>n</i> = 17)	<i>p</i> -value
Age at inclusion in years (mean, SD)	63.7 (4.3)	61.1 (5.4)	<i>p</i> = 0.10 ^a
Fracture type			<i>p</i> = 0.06 ^b
Extra-articular	17 (51%)	14 (82%)	
Intra-articular	16 (49%)	3 (18%)	
Affected side			<i>p</i> = 0.09 ^b
Dominant	27 (82%)	10 (59%)	
Non-dominant	6 (18%)	7 (41%)	

SD: standard deviation.

^aTested with independent t-test.

^bTested with Fisher's exact test.

Table 2. Regression coefficients for the unadjusted and adjusted models.

	Unadjusted model		Adjusted model	
	Coefficient (95% CI)	<i>p</i> -value	Coefficient (95% CI)	<i>p</i> -value
Intercept	52.35 (50.32 to 54.37)	<0.001	33.44 (25.22 to 41.65)	<0.001
Change in PRWE score with intervention	4.44 (1.95 to 6.93)	<0.001	1.21 (-0.12 to 2.54)	0.070
Age	—		0.32 (0.19 to 0.45)	<0.001
Fracture type (intra-articular)	—		5.17 (3.87 to 6.46)	<0.001
Dominant hand affected (yes)	—		-3.56 (-4.94 to -2.17)	<0.001

PRWE: Patient-Rated Wrist Evaluation; CI: confidence interval.

variables that may influence the outcome, we would conclude that the beneficial effect is absent (Table 2, adjusted model; *p* = 0.070).

Significance versus clinical relevance

Another area where we need to be wary is the clinical implication of a *p*-value. Even when the *p*-value is smaller than 0.05, we need to interpret it carefully within the context of the research question and outcome measure chosen. For example, our unadjusted model indicates that there was a statistically significant difference between the two groups of 4.4 points in terms of the PRWE score. However, a statistically significant difference does not necessarily indicate that it is clinically relevant. What does this mean? In this case, the minimal change in PRWE score that is meaningful to the average patient, also known as the minimal clinically important change (MCIC), is 20 points (McCreary et al., 2020). This is much larger than the difference that we observed between the two groups. In other words, the statistically significant difference observed between the two groups is a difference that was too small for patients to notice, namely it is not clinically relevant.

In this aspect, it would be more appropriate to conclude that there was no beneficial effect of the intervention despite the *p*-value <0.05. It should be noted that this conclusion only provides information on group-level, that is for the average patient. Of course, it is possible that individual patients did have an increase in PRWE score larger than the MCIC of 20 points. However, if we are interested in such individual effects of the intervention, there are other better methods for analysis, such as responder analysis in which we would define the individual level outcome here as attaining a change of 20 points or more.

Figure 1 shows possible scenarios of an effect with respect to statistical significance and clinical relevance. The dots represent the point estimate, which is for instance the observed difference in means between groups when using t-tests, or the beta-coefficient in regression analysis. The whiskers represent the 95% confidence intervals (CI), which provide an indication of the precision with which this difference was estimated (this will be further elaborated in Part 2). Note though that CIs are also used to test for statistical significance, just like *p*-values. In fact, whenever the 95% CI does not

include the neutral value ('0' in case of testing a difference of means), the effect is statistically significant, and the corresponding p -value will be smaller than 0.05.

In Figure 1, there are two scenarios where statistical significance and clinical relevance are in line with each other: (1) the case in which there is a statistically significant difference and clinically relevant difference (red) as indicated by the 95% CI not overlapping 0, and a point estimate being larger than the MCIC; and (2) the case in which there is no statistically significant and no clinically relevant difference (blue) as indicated by the 95% CI containing 0 and the point estimate being smaller than the MCIC. In the first scenario, there is evidence that the new intervention is effective, and that it may be implemented in clinical practice. Clearly, other considerations, such as costs, side-effects or burden, should also be taken into account in the decision to implement it. Additionally, the decision to implement it should always be seen in the light of plausibility and ideally not be based on a single study. The second scenario shows absence of both statistical significance and clinical relevance, which provides evidence that the intervention should not (yet) be implemented in clinical practice. The other scenarios reflect situations where the two are not in line with each other or where one cannot tell based on the results. The third scenario (green) shows statistical significance, but there is absence of clinical relevance, which reflects the situation of our example

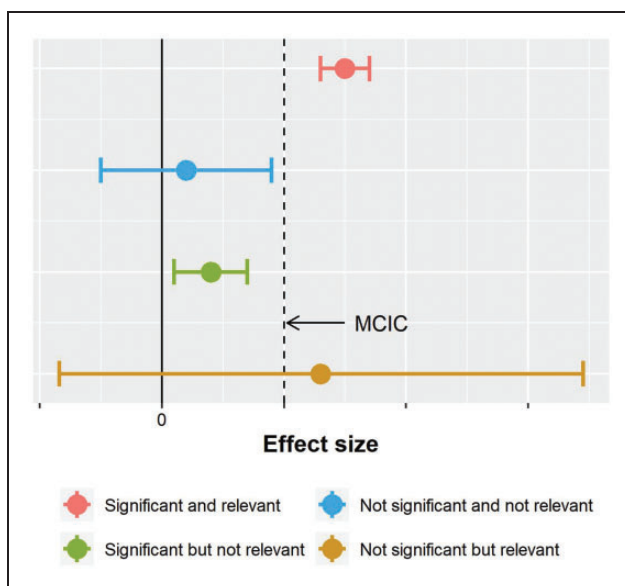


Figure 1. The four possible scenarios of statistically (non)significance and (no) clinical relevance. MCIC: minimal clinically important change.

(Table 2, mean difference 4.44 [95% CI: 1.95 to 6.93]). In the last scenario (yellow), the 95% CI reflects a very imprecise result that is clearly not statistically significant, and for which clinical relevance remains inconclusive since the 95% CI overlaps the MCIC. In this situation, further research is warranted.

Superiority testing versus equivalence or non-inferiority testing

Using the example on distal radial fractures again, our hypothesis was that the intervention was beneficial for functional outcome. In other words, the example trial is a so-called superiority trial. However, we should note that often new interventions emerge that are not expected to be better than existing ones, but that are expected to have equivalent (similar) or at least not inferior (not worse) effectiveness with regard to the outcome. The latter is especially common in clinical trials, for instance, if we expect the (new) treatment to have less side effects or if costs are expected to be lower than the usual care. It is important to note that this is a completely different starting point. In such cases, regular (superiority) testing using NHST will not answer the research question as to whether the treatment outcomes are similar, because it is only geared to test for superiority. Testing for non-inferiority starts with the definition of a margin of non-inferiority. This margin indicates that the new treatment might result in a (slightly) worse outcome, without clinical implications. The margin is the cut-off after which we deem differences to indicate worse outcomes for the new treatment as compared with standard care. After results have been obtained, again confidence intervals can be used to make a statement about the results, in this case, whether non-inferiority of the new treatment can be assumed. When the lower limit of the confidence interval lies before the margin of non-inferiority (Figure 2, scenarios BD), the effects can be considered as non-inferior. In cases where the interval includes the margin (Figure 2, scenarios E-G), the result is inconclusive. When both the lower and upper limit of the confidence interval exceed the margin (Figure 2, scenario H), the new treatment can be deemed inferior. Note that the conclusions for many of these scenarios are very different from those from a superiority analysis. For example, in a superiority analysis, scenario B would imply that we cannot conclude that the new treatment is better than usual care, whereas in a non-inferiority analysis we conclude it is non-inferior.

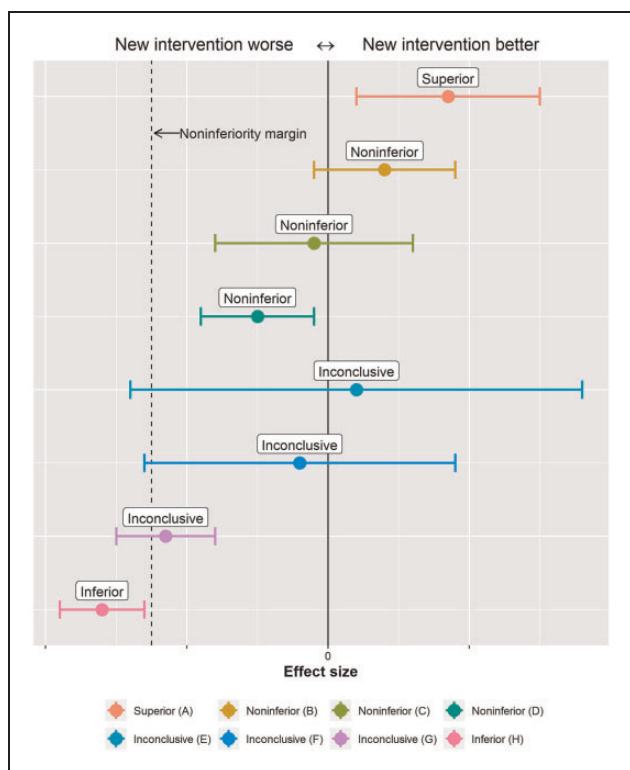



Figure 2. Different scenarios showing the difference between superiority and (non)inferiority.

Conclusion

In conclusion, we have shown some common forms of p -value and particularly NHST misuse, their possible consequences and offered practical solutions. The issues we raised are well known among methodologists, but less so among applied researchers and clinicians. Furthermore, statistical behaviour perpetuates with the behaviour of others as presented to us in the articles we read. We hope we

have provided an easy-to-understand explanation applied to a hand surgery example, and a few simple guidelines that can be followed in situations where common mistakes are made. It is important to note there is no 'one-stop shop' solution that will overcome all drawbacks related to both NHST (p -values) and the use of CI instead, and constant discussions with a statistician or methodologist is helpful. In part 2, we will discuss the conceptual problems of NHST in further detail and also look at more alternatives.

ORCID iD Dieuwke C. Broekstra  <https://orcid.org/0000-0002-7134-7007>

References

- Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990; 225: 149–53.
- Broekstra DC, Mouton LJ, van der Sluis CK, IJpma FFA, Stenekes MW. Hand function in patients with distal radius fractures after home-based kinaesthetic motor imagery training. *J Hand Surg Eur*. Online publication 01 February 2022. DOI: 10.1177/17531934221075945.
- McCreary DL, Sandberg BC, Bohn DC, Parikh HR, Cunningham BP. Interpreting patient-reported outcome results: is one minimum clinically important difference really enough? *Hand*. 2020; 15: 360–4.
- Moher D, Hopewell S, Schultz KF et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. 2010; 63: e1–37.

Dieuwke C. Broekstra^{1,*} , **Michiel R. de Boer²** and **Jonáh J. Stunt³**

¹Department of Plastic Surgery, University of Groningen, Groningen, The Netherlands

²Department of General Practice and Elderly Care Medicine, University of Groningen, Groningen, The Netherlands

³Department of Health Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

*Corresponding author: d.c.broekstra@umcg.nl

Twitter: @DCBroekstra