Evaluation of the Use of *P* Values in Neurosurgical Literature: from Statistical Significance to Clinical Irrelevance

Iris S.C. Verploegh[1], *Nicole A. Lazar*[2], *Ronald H.M.A. Bartels*[3], *Victor Volovici*[1]

The application and interpretation of *P* values have caused debate for several decades, and this debate has become particularly relevant in the past few years. The *P* value represents the probability of seeing results as extreme or more extreme than those observed in a data analysis, were the null hypothesis and other underlying assumptions to be true. While *P* values are useful in pointing out where an effect may be present, they have often been misused in an attempt to oversell "statistically significant" findings. As *P* values rely on the spread and number of measurements, a smaller *P* value does not necessarily imply a larger effect size, which is better assessed via an effect estimate and confidence interval interpreted in the context of the study. The clinical relevance of a computed *P* value is context dependent. We investigated the current use of *P* values in a small sample of recent neurosurgical literature. Only a minority of manuscripts that reported statistical significance described confounder adjustment, or effect sizes. A common, incorrect assumption often observed was that statistical significance equals clinical relevance. To enable correct interpretation of clinical significance, it is crucial that authors describe the clinical implications of their findings.

INTRODUCTION

Since the P value was first described in 1900 by Pearson,[1] it has come to be considered—rightly or wrongly—as the most important summary measure of statistical testing. However, use (and misuse) of the P value has been a cause of controversy particularly in recent years.[2,3] Authors might feel that they have to show "statistically significant" results for their manuscript to be accepted for publication or to be deemed relevant. Such incentives may lead to misuse of statistical testing and "data dredging." Moreover, most scientists naturally are not trained statisticians and do not know all of the nuances and assumptions made by each statistical test, leading to misinterpretation of results. In reaction to this state of affairs, 1 journal even banned P values completely from its papers.[4] Other attempts at improving the usage of P values have been education on alternative statistical methods such as Bayesian inference or stimulating more elaborate reporting by also providing confidence intervals interpreted in the context of the study data. To clarify the current controversy on P values, the American Statistical Association published a statement on the background and purpose of P values.[2] The statement can be used as a helpful tool for correct interpretation of statistical results.

Our aim is to elaborate on the correct usage and point out common pitfalls related to the use of P values. We also assess reporting of P values in a small sample of recent neurosurgical literature and report several recommendations for improvement.

METHODS

To provide context for the usage of P values in the neurosurgical literature, we assessed a target of 10 papers reporting P values, when available, from each of 2 issues published in 2021 in the following neurosurgical journals: *World Neurosurgery; Journal of Neurology, Neurosurgery, and Psychiatry; Journal of Neurosurgery: Neurosurgery;* and *Neurosurgical Focus* (a goal of 100 papers). From each article we extracted the following variables: type of study (cohort, case-control, randomized controlled, meta-analysis); prospective/retrospective; subspecialty; number of patients; whether confounder adjustment was performed; whether statistical tests were used; whether a confidence interval was reported;

Key words
- Confidence interval
- Effect size
- Hypothesis testing
- *P* value
- Statistical significance

From the [1]*Department of Neurosurgery, Erasmus Medical Center, Rotterdam, The Netherlands;* [2]*Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, USA; and* [3]*Radboud University Medical Center, Nijmegen, The Netherlands*

To whom correspondence should be addressed: Iris S. C. Verploegh, M.Sc. [E-mail: i.verploegh@erasmusmc.nl]

Supplementary digital content available online.

Citation: World Neurosurg. (2022) 161:280-283.
https://doi.org/10.1016/j.wneu.2022.02.018

*Journal homepage: **www.journals.elsevier.com/world-neurosurgery***

*Available online: **www.sciencedirect.com***

and primary outcome (if reported). We assessed whether the reported statistical significance was of true clinical relevance and whether the evaluation was in concordance with the evaluation of the authors. This analysis was also performed per subspecialty, of which at least 5 papers were present in our dataset. Clinical significance was inferred on the basis of consultations with a panel of experts in the specific subspecialty field. Descriptive statistics were calculated, and images were designed in R version 4.0.5.[5]

RESULTS

A total of 89 papers were included in the assessment (**Supplementary Table 1**). In some issues, fewer than 10 papers reporting statistical significance were published, which prevented us from reaching the goal of 10 manuscripts per issue. Out of the 89 papers, there were 71 cohort studies, 12 case-control studies, 5 randomized controlled trials, and 1 meta-analysis. Most studies (55 of 89) were retrospective.

In 40% (36/89) of the studied manuscripts adjusted effect sizes and their *P* values were reported. However, sometimes when the *P* value was >0.05 after confounder adjustment, this adjusted outcome was no longer mentioned in the discussion of the article and instead the unadjusted *P* value was further used. Of the 75 papers with "statistically significant" reported outcomes, 54 had clinically relevant conclusions and 21 did not. Papers with clinically relevant results performed adjustment for confounders: 43% (23/54) of the time compared with 33% (7/21) of those with statistically significant but clinically irrelevant results.

In our set of neurosurgical papers, 48% (43/89) reported confidence intervals related to the effect size of the primary outcome measure. However, when small effect sizes were present, the authors usually opted to report *P* values only. Nevertheless, 61% (33/54) of manuscripts with clinically relevant findings reported confidence intervals compared with 29% (6/21) of those with statistical significance but without clinical relevance.

We then assessed if clinical relevance of the statistically significant results and effect sizes were specifically discussed. Often, when the main result showed no significant difference, the clinical relevance of the outcomes was elaborately discussed (**Figure 1A**). However, in the case of results showing statistical significance, but lacking clinical relevance, the latter was only mentioned in the discussion of 19% (4/21) of manuscripts. By contrast, in 35% (19/54) of papers with both clinical and statistical significance, clinical significance was highlighted in the discussion. This trend was present within all analyzed subspecialties except policy studies, with imaging studies showing the strongest trend (see **Figure 1B**). In our dataset of recent neurosurgical studies, we regularly encountered well-executed use of the *P* value. Nevertheless, one third of papers screened reported results that were statistically significant but clinically irrelevant in the context of the research area.

DISCUSSION

Significance of a *P* Value

The *P* value is the probability for the used statistical model that, when the null hypothesis is true, the statistical summary would be equal to or more extreme than the actual observed measurement.[2] The null hypothesis most often states that there is "no difference" between 2 or more conditions or "no correlation" between the studied variable of interest and the outcome, though other values and other forms of null are possible. Note, however, that the statistical model also includes assumptions, as well as taking into account the number of observations, the types of measurement,

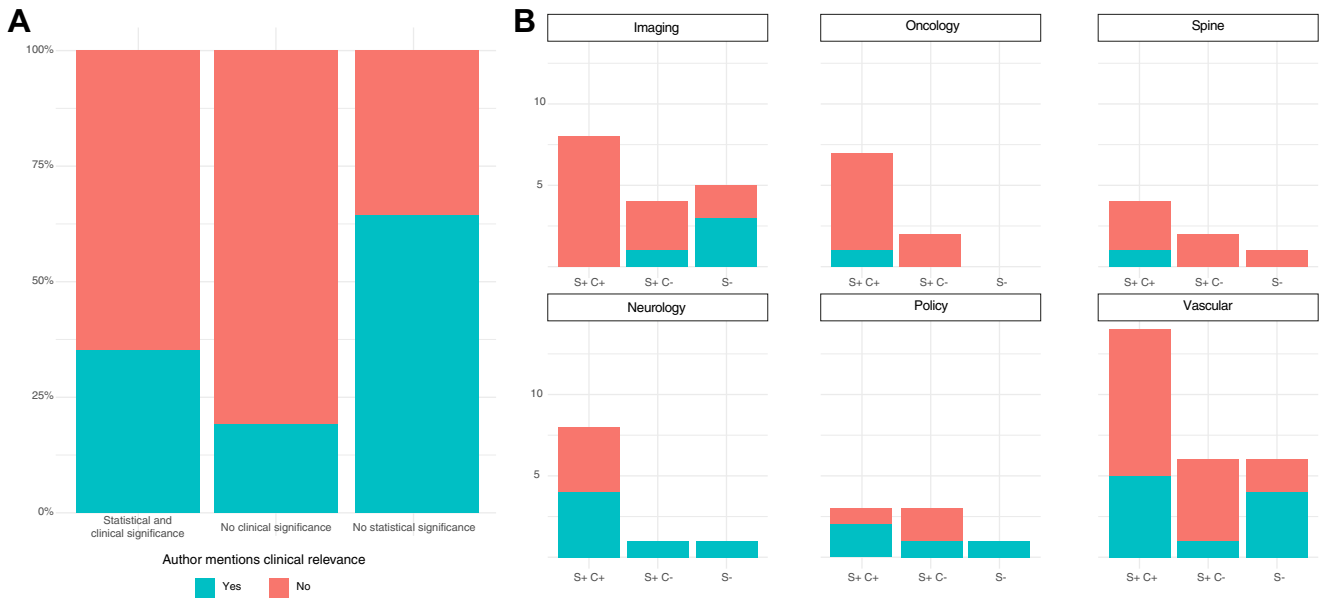

Figure 1. Relation of statistical and clinical significance. (**A**) Percentage of manuscript in which authors mention clinical relevance of their statistical outcome, stratified by clinical and statistical significance. (**B**) Number of studies describing statistical significance *(S)* and clinical relevance *(C)* per subspecialty.

the spread of the data, and the summary measure. A low P value can therefore imply lack of compatibility with the null hypothesis value, or violations of any of the other assumptions that went into the model (e.g., assumed distribution of the data, used to compute the test statistic; independence of the observations). In most clinical studies a P value of 0.05 is interpreted as being "statistically significant," though this value is arbitrarily set. Studies discussing large (e.g., genomic, transcriptomic) datasets often set a lower threshold of significance as many more tests are being performed.

Common Misconceptions About the *P* Value

Even though P values may provide helpful insight into interpreting data, they are often misused and misinterpreted. Here we list some common misconceptions[6,7]:

1. A P value tells us the chance that the alternative hypothesis is right. P values only describe a part of the evidence we have to support the null hypothesis. The compatibility with the alternative hypothesis is not tested. Many other factors can contribute to rejection of the null hypothesis that are not represented in the alternative hypothesis. Such factors include the study design, methods of data collection, and modeling assumptions. To reduce the influence of possible confounders on the acceptance or rejection of the null hypothesis, P value adjustment and sensitivity analysis can be performed.
2. A large P value implies that the alternative hypothesis is false. A commonly encountered misconception is that a large P value (no compelling evidence against the null hypothesis) implies noninferiority. A case-control study might compare 2 techniques (1 old and 1 new) and conclude, on the basis of a large P value, that the new technique is noninferior to the old one. This finding, in turn, would provide evidence for U.S. Food and Drug Administration approval, for example. Nevertheless, a noninferiority study design is particular, in that the null hypothesis means there is a difference between treatments. The study has to be powered to detect noninferiority using a so-called noninferiority margin. Simply showing a large P value does not automatically mean the noninferiority hypothesis is satisfied. A smaller sample size increases the chance of finding a large P value, leading to the conclusion of noninferiority. Therefore one should always assess the distribution of the raw datapoints to evaluate whether based on the used data, the rejection of the alternative hypothesis was correct and noninferiority should be powered as such.

P Values and Effect Size

A common pitfall in the interpretation of a P value is that the smaller the P value, the larger the effect or stronger the correlation. As stated before, when calculating the P value, the sample size and variability are also taken into account. This means that P values are sensitive to sample size and precision of measurement, which are not directly indicative of the effect size. If there is a statistical difference between the control group and intervention group, other measures are necessary, particularly effect size estimates and confidence intervals. Care also needs to be taken in interpreting these, for example, by not focusing exclusively on whether a confidence interval includes 0 when exploring the mean difference between 2 groups. Instead, the range of values in the interval, as well as their clinical importance, should be considered.

Statistical Significance and Clinical Irrelevance

The ultimate goal of clinical research is to improve patient care by better detection of disease, improved treatment, and better assessment of therapy response. To do so, researchers must be able to extract clinical significance from the published literature.[8]

Still, statistical significance is often interpreted by authors as clinical significance. As discussed earlier, we explored the clinical significance in papers in which the primary outcome had a reported "statistically significant" result. This assessment was based on all statistical and background information provided by the authors. No thresholds can be used for this, as every research question has its own implications and range of effect sizes that can be achieved. Authors of manuscripts should provide enough information to enable the reader to make a well-informed clinical decision, based primarily on the clinical relevance of findings. In one third of the papers we surveyed, statistically significant but clinically irrelevant results were highlighted. A solution for this could be to define a clinically relevant effect size and clinically relevant outcomes for the field at the start of the study. This strategy can aid investigators in focusing on clinical relevance and effect size instead of statistical parameters. One third of the most recent papers reporting clinically irrelevant but "statistically significant" results suggests the problem is still quite prevalent, despite more sophisticated methodology being employed nowadays.

CONCLUSIONS

The P value is a powerful supportive tool for the interpretation of study results; however, P value adjustment, reporting of confidence intervals, and critical appraisal with respect to clinical relevance are necessary for optimal interpretation.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Iris S.C. Verploegh: Conceptualization, Data curation, Formal analysis, Writing — original draft. **Nicole A. Lazar:** Writing — review & editing. **Ronald H.M.A. Bartels:** Writing — review & editing. **Victor Volovici:** Conceptualization, Methodology, Validation, Supervision, Writing — review & editing.

REFERENCES

1. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag*. 1900;50:157-175.
2. Wasserstein R, Lazar N. The ASA statement on P values: context, process, and purpose. *Am Statistician*. 2016;70:129-133.
3. Greenland S. Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. *Am Statistician*. 2019;73(suppl 1):106-114.
4. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psychol*. 2015;37:1-2.

5. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/. Accessed February 28, 2022.

6. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-350.

7. Goodman S. A dirty dozen: twelve P-value misconceptions. *Semin Hematol*. 2008;45:135-140.

8. Betensky RA. The P-value requires context, not a threshold. *Am Statistician*. 2019;73(suppl 1):115-117.

Conflict of interest statement: The authors declare that the article content was composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received 30 November 2021; accepted 3 February 2022

Citation: World Neurosurg. (2022) 161:280-283.
https://doi.org/10.1016/j.wneu.2022.02.018

Journal homepage: ***www.journals.elsevier.com/world-neurosurgery***

Available online: ***www.sciencedirect.com***

SUPPLEMENTARY DATA

Supplementary Table 1. Summary Table of Included Studies

First Author	Journal	Volume	Type of Study	Timing	Subspecialty	Number	Adjustment	CI	Statistical Significance Equals Clinical Relevance	Author Mentions Clinical Relevance
Subramanian	*World Neurosurgery*	149	Cohort	Retrospective	Imaging	37	N	N	N	Y
Tsuji	*World Neurosurgery*	149	Cohort	Retrospective	Vascular	114	N	N	Y	N
Zhao	*World Neurosurgery*	149	Case control	Retrospective	Oncology	36	N	N	Y	N
Ruiter de	*World Neurosurgery*	149	Cohort	Prospective	Imaging	54	N	N	No statistical significance	N
Obo	*World Neurosurgery*	149	Case control	Retrospective	Imaging	100	N	N	N	N
Song	*World Neurosurgery*	149	Case control	Retrospective	Pain	63	Y	Y	N	N
Macki	*World Neurosurgery*	149	Cohort	Retrospective	Vascular	105	Y	Y	No statistical significance	N
Algahtany	*World Neurosurgery*	149	Cohort	Cross-sectional	Trauma	167357	N	Y	Y	N
Zhang	*World Neurosurgery*	149	Cohort	Retrospective	Vascular	265	N	Y	Y	N
Kimball	*World Neurosurgery*	149	Cohort	Prospective	Imaging	15	N	N	No statistical significance	Y
Budiono	*World Neurosurgery*	148	Cohort	Retrospective	Spine	68	Y	Y	Y	Y
Adamczak	*World Neurosurgery*	148	Cohort	Retrospective	Vascular	373	N	N	Y	Y
Otsuki	*World Neurosurgery*	148	Cohort	Retrospective	Policy	95	N	N	Y	N
Zimmerman	*World Neurosurgery*	148	Cohort	Cross-sectional	Pediatric	91	N	N	Y	Y
Almeida	*World Neurosurgery*	148	Cohort	Retrospective	Spine	61977	Y	Y	Y	N
Montgomery	*World Neurosurgery*	148	Cohort	Retrospective	Spine	156	Y	Y	No statistical significance	N
Kim	*World Neurosurgery*	148	Case control	Retrospective	Spine	100	N	N	N	N
Raman	*World Neurosurgery*	148	Cohort	Retrospective	Oncology	233	Y	Y	N	N
Wales	*World Neurosurgery*	148	Cohort	Prospective	Spine	40	N	N	Y	N
Wali	*World Neurosurgery*	148	Cohort	Retrospective	Policy	1429	N	N	No statistical significance	Y
Wiegertjes	*JNNP*	92-9	RCT	Prospective	Vascular	537	N	Y	Y	N
Nappini	*JNNP*	92-9	Cohort	Prospective	Vascular	464	Y	Y	N	N
Kojima	*JNNP*	92-9	Cohort	Retrospective	Neurology	53	Y	Y	Y	N
Ruiz	*JNNP*	92-9	Cohort	Retrospective	Neurology	48	N	N	Y	N
Colato	*JNNP*	92-9	Cohort	Retrospective	Imaging	988	Y	Y	Y	N
Pfeuffer	*JNNP*	92-9	Cohort	Prospective	Neurology	170	Y	Y	Y	N
Seiffge	*JNNP*	92-8	Cohort	Prospective	Vascular	2477	Y	Y	Y	N
Murphy	*JNNP*	92-8	Cohort	Retrospective	Neurology	100	Y	Y	Y	N
Haider	*JNNP*	92-8	Cohort	Prospective	Neurology	132	N	Y	Y	Y
Goh	*JNNP*	92-8	Meta-analysis	Meta-Analysis	Trauma	33	N	Y	Y	N
Sangha	*JNNP*	92-8	Cohort	Prospective	Neurology	181	Y	Y	Y	Y

Continues

Supplementary Table 1. Continued

First Author	Journal	Volume	Type of Study	Timing	Subspecialty	Number	Adjustment	CI	Statistical Significance Equals Clinical Relevance	Author Mentions Clinical Relevance
Singleton	*JNNP*	92-8	Cohort	Retrospective	Neurology	7	Y	N	No statistical significance	Y
Grimaldi	*JNNP*	92-8	Cohort	Retrospective	Neurology	85	Y	Y	Y	Y
Hilezian	*JNNP*	92-8	Cohort	Retrospective	Neurology	23	N	N	Y	Y
Bertalanffy	*JNS*	135-3	Cohort	Retrospective	Vascular	77	N	N	Y	Y
Cenzato	*JNS*	135-3	Cohort	Retrospective	Vascular	1274	N	N	N	N
Baek	*JNS*	135-3	Cohort	Retrospective	Imaging	235	Y	Y	Y	N
Zdunczyk	*JNS*	135-3	Cohort	Retrospective	Imaging	21	Y	N	Y	N
Khanna	*JNS*	135-3	Case control	Retrospective	Vascular	104	N	N	No statistical significance	Y
Kawashima	*JNS*	135-3	Cohort	Retrospective	Vascular	410	Y	Y	Y	N
Chen	*JNS*	135-3	Case control	Retrospective	Vascular	106	Y	Y	No statistical significance	Y
Rich	*JNS*	135-3	Cohort	Retrospective	Epilepsy	13	N	N	N	N
Eisenberg	*JNS*	135-3	Cohort	Prospective	Degeneration	20	N	Y	Y	Y
Horisawa	*JNS*	135-3	Cohort	Prospective	Degeneration	14	N	N	Y	Y
Mahajan	*JNS*	135-2	Cohort	Retrospective	Policy	4028	N	Y	N	N
Tang	*JNS*	135-2	Cohort	Retrospective	Policy	472938	Y	Y	Y	Y
Wright	*JNS*	135-2	Cohort	Prospective	Policy	48	N	N	N	N
Santos	*JNS*	135-2	Cohort	Cross-sectional	Vascular	67	Y	Y	Y	N
Tashiro	*JNS*	135-2	Cohort	Prospective	Vascular	66	Y	Y	Y	Y
Rumalla	*JNS*	135-2	Cohort	Retrospective	Vascular	2500	Y	Y	Y	Y
Ferreira de Melo Neto	*JNS*	135-2	Cohort	Retrospective	Vascular	201	N	N	Y	N
Shimizu	*JNS*	135-2	Cohort	Retrospective	Oncology	110	Y	Y	Y	N
Livermore	*JNS*	135-2	Case control	Retrospective	Imaging	73	N	Y	Y	N
Gogos	*JNS*	135-2	Cohort	Prospective	Oncology	657	N	N	Y	N
Gross	*NS*	89-3	Cohort	Retrospective	Epilepsy	101	N	N	Y	N
Scheffler	*NS*	89-3	Cohort	Retrospective	Hydrocephalus	97	N	Y	N	N
Bae	*NS*	89-3	Cohort	Cross-sectional	Imaging	34	N	N	Y	N
Alan	*NS*	89-3	Cohort	Retrospective	Trauma	111	Y	Y	Y	Y
Hale	*NS*	89-3	Case control	Cross-sectional	Neurology	3162	Y	N	N	Y
Lylyk	*NS*	89-3	Cohort	Retrospective	Vascular	835	N	Y	Y	N
Rocha	*NS*	89-3	Cohort	Prospective	Degeneration	13	N	Y	Y	Y
Krauss	*NS*	89-3	Cohort	Retrospective	Policy	1193	Y	Y	Y	Y
Khalafallah	*NS*	89-3	Case control	Retrospective	Policy	2749	Y	N	N	Y

CI, confidence interval; N, no; Y, yes; JNNP, Journal of Neurology, Neurosurgery and Psychiatry; JNS, Journal of Neurosurgery; NS, Neurosurgery; NSFocus, Neurosurgery Focus; RCT, randomized controlled trial.

Continues

Supplementary Table 1. Continued

First Author	Journal	Volume	Type of Study	Timing	Subspecialty	Number	Adjustment	CI	Statistical Significance Equals Clinical Relevance	Author Mentions Clinical Relevance
Enriques-Marulanda	*NS*	89-3	Cohort	Retrospective	Vascular	402	Y	N	No statistical significance	Y
Rapaport	*NS*	89-2	Cohort	Prospective	Vascular	789	N	N	N	N
Bird	*NS*	89-2	Cohort	Retrospective	Vascular	122	Y	Y	No statistical significance	N
Rabinovich	*NS*	89-2	Cohort	Retrospective	Spine	184	Y	N	Y	N
Park	*NS*	89-2	Cohort	Retrospective	Imaging	86	Y	Y	Y	N
Chong	*NS*	89-2	Cohort	Retrospective	Trauma	370	Y	Y	Y	Y
Passeri	*NS*	89-2	Cohort	Retrospective	Oncology	32	N	Y	Y	N
Wang	*NS*	89-2	Cohort	Retrospective	Oncology	44	N	Y	Y	Y
Voormolen	*NS*	89-2	Cohort	Prospective	Oncology	262	Y	N	N	N
Lai	*NS*	89-2	Cohort	Retrospective	Vascular	95	Y	Y	N	N
Ishisaka	*NSFocus*	51-3	Case control	Retrospective	Vascular	86	N	N	No statistical significance	Y
Kanamori	*NSFocus*	51-3	Case control	Prospective	Vascular	20	N	N	N	N
Acker	*NSFocus*	51-3	Cohort	Prospective	Vascular	30	N	N	N	Y
Garcia	*NSFocus*	51-3	Cohort	Retrospective	Imaging	35	N	Y	Y	N
Wang	*NSFocus*	51-3	Cohort	Prospective	Vascular	106	Y	N	Y	N
Luzzi	*NSFocus*	51-2	Case control	Retrospective	Oncology	54	N	Y	Y	N
Van Gestel	*NSFocus*	51-2	RCT	Prospective	Hydrocephalus	128	N	N	Y	N
Knafo	*NSFocus*	51-2	Cohort	Prospective	Imaging	14	N	N	No statistical significance	Y
Yanni	*NSFocus*	51-2	Cohort	Prospective	Imaging	112	N	N	N	N
Steineke	*NSFocus*	51-2	Cohort	Retrospective	Vascular	21	N	N	Y	Y
Jean	*NSFocus*	51-2	Cohort	Retrospective	Oncology	15	N	N	Y	N
Roh	*NSFocus*	51-2	Cohort	Prospective	Imaging	31	N	N	N	N
Qi	*NSFocus*	51-2	RCT	Prospective	Spine	60	N	N	N	N
Greuter	*NSFocus*	51-2	RCT	Prospective	Imaging	80	N	N	No statistical significance	N
Davidovic	*NSFocus*	51-2	RCT	Prospective	Imaging	23	N	N	No statistical significance	Y
Qi	*NSFocus*	51-2	Cohort	Prospective	Imaging	37	N	N	Y	N

CI, confidence interval; N, no; Y, yes; JNNP, Journal of Neurology, Neurosurgery and Psychiatry; JNS, Journal of Neurosurgery; NS, Neurosurgery; NSFocus, Neurosurgery Focus; RCT, randomized controlled trial.