Knee Surg Sports Traumatol Arthrosc
DOI 10.1007/s00167-017-4582-y

EDITORIAL

While modern medicine evolves continuously, evidence-based research methodology remains: how register studies should be interpreted and appreciated

Eleonor Svantesson[2] · Eric Hamrin Senorski[2] · Kurt P. Spindler[3] · Olufemi R. Ayeni[4] · Freddie H. Fu[5] · Jón Karlsson[1,2] · Kristian Samuelsson[1,2]

In just a few decades, the scientific stage has undergone some dramatic changes. Novel studies are produced at a "faster than ever" pace, and technological advances enable insights into areas that would previously have been referred to as science fiction. However, the purpose of research will always be the same—to serve as a firm foundation to practise evidence-based medicine and ultimately improve the treatment of our patients. Is the explosive evolvement of research publications and technological advances always beneficial when it comes to fulfilling this purpose? As we are served with a steady stream of new "significant" findings, it is more important than ever critically to evaluate the evidence that is presented and be aware of the limitations and pitfalls that we encounter every day as modern scientists and clinicians.

Look! A significant result!

One of the goals for researchers is to get their work published and acknowledged, preferably with multiple citations. A winning tactic to accomplish this is to present novel results and findings. Interestingly, it often happens that the most cited papers are those that contradict other reports or are proved to be fundamentally wrong [14]. So, it does not really matter how likely a result is to be true or clinically valuable—a spectacular result can entrench the findings of a study and influence clinical practice. It goes without saying that the most important factor of all in this quest is that a significant *P* value is presented. Today, it is generally accepted that significance, often defined as a *P* value of <0.05, means impact and evidence. However, this is an incorrect appreciation of the *P* value and could lead to an inappropriate approach to this statistical method. It has been shown that *P* values and hypothesis-testing methods are commonly misunderstood by researchers [6, 11, 17]

✉ Kristian Samuelsson
kristian@samuelsson.cc

1 Department of Orthopaedics, Sahlgrenska University Hospital, Mölndal, Sweden

2 Department of Orthopaedics, Institute of Clinical Sciences, The Sahlgrenska Academy, University of Gothenburg, 431 80 Gothenburg, Sweden

3 Cleveland Clinic Sports Health Center, Garfield Heights, OH, USA

4 Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

5 Department of Orthopedic Surgery, University of Pittsburgh, Pittsburgh, PA, USA

Published online: 13 June 2017

and instead tend to lead to a limited perspective in relation to a study result.

Sir Ronald Fisher is regarded as one of the founders of modern statistics and is probably most associated with the concept of the *P* value [10, 23]. Fisher suggested that the *P* value reflected the probability that the result being observed was compatible with the null hypothesis. In other words, if it were true that there was no (null) difference between the factors being investigated, the *P* value would give an estimation of the likelihood of observing a difference as extreme as or more extreme than your outcome showed. However, Fisher never propagated the $P < 0.05$ criterion that is currently almost glorified as our ultimate means of conclusion making. On the contrary, Fisher appeared not to give much consideration to the actual *P* value number [19]. The most important thing, according to Fisher, was to repeat the experiments until the investigator felt that he or she had a plausible certainty of declaring how the experiment should be performed and interpreted, something that is infrequently implemented nowadays. The *P* value was originally an indicative tool throughout this process, not something synonymous with evidence.

In a study recently published in *JAMA cardiology* [19], common misconceptions about *P* values were discussed. It was emphasised that, at best, the *P* value plays a minor role in defining the scientific or clinical importance of a study and that multiple elements, including effect size, precision of estimate of effect size and knowledge of prior relevant research, need to be integrated in the assessment [19]. This is strongly inconsistent with the concept of a *P* value of <0.05 as an indicator of a clinically or scientifically important difference. Moreover, the authors highlight the misconception that a small *P* value indicates reliable and replicable results by stating that what works in medicine is a process and not the product of a single experiment. No information about a given study regarding reproducibility can be made based on the *P* value, nor can the reliability be determined without considering other factors [19]. One frequently forgotten factor is how plausible the hypothesis was in the first place. It is easy to fall into the trap of thinking that a *P* value of <0.05 means that there is a 95% chance of true effect. However, as probability is always based on certain conditions, the most important question should be: what was the probability from the beginning? If the chance of a real effect from the beginning is small, a significant *P* value will only slightly increase the chances of a true effect. Or, as Regina Nuzzo put it in an article highlighting statistical errors in *Nature* [21]: "The more implausible the hypothesis—telepathy, aliens, homeopathy—the greater the chance that an exciting finding is a false alarm, no matter what the *P* value is" [21].

Moreover, the *P* value says nothing about the effect size. The *P* value is basically a calculation of two factors—the difference from null and the variance. In a study with a small standard deviation (high precision), even a very small difference from zero (treatment effect) can therefore result in a significant *P* value. How frequently do we ask ourselves: "From what numbers was this *P* value generated?" when reading a paper. It is not until we look at the effect size that it is really possible to determine whether the treatment of interest has an impact. Well then, what is the definition of impact? A term often used to describe the effectiveness of a treatment is the "minimum clinically important difference" (MCID). For a study to impact clinical decision-making, the measurement given must be greater than the MCID and, moreover, the absolute difference needs to be known. These factors determine the number needed to treat and thereby indicate the impact. However, current methods for determining MCID are subject of debate and it has been concluded that they are associated with shortcomings [9].

We should also remember that non-significant *P* values are sometimes used to conclude the interventions of interest as "equivalence" or "non-inferiority", which is extremely incorrect if the primary study design was not intended to investigate equivalence between two treatments [18]. Without primarily designing the study for this purpose, it is impossible to ascertain power for detecting the ideal clinically relevant difference that is needed for a declaration of equivalence. It can, in fact, have detrimental downstream effects on patient care if a true suboptimal treatment is declared as being non-inferior to a gold-standard treatment [12]. Instead, let us accept the fact that not all studies will show significant results, nor should they. There has been a bias against "negative trials", not showing significance, in the past and because of this we can only speculate about whether or not they could have impacted any of today's knowledge. If the acceptance of non-significant results increases, this could contribute to the elimination of publication bias.

The impact of study design

Regardless of study design, the optimal research study should give an estimate of the effectiveness of one treatment over another, with a minimised risk of systematic bias. The ability and validity of doing this for observational studies compared with randomised controlled trials (RCTs) has been the subject of an ongoing debate for decades. To determine the efficacy of a treatment or intervention (i.e. the extent to which a beneficial result is produced under ideal conditions), the RCTs remain the gold standard and are regarded as the most suitable tool for making the most precise estimates of treatment effect [22]. The only more highly valued study design is

the meta-analysis of large, well-conducted RCTs. Studies with an observational design are often conducted when determining the effectiveness of an intervention in "real-world" scenarios (i.e. the extent to which an intervention produces an outcome under normal day-to-day circumstances). A Cochrane Review published in 2014 [3] examined fourteen methodological reviews comparing quantitative effect size estimates measuring the efficacy or effectiveness of interventions tested in RCTs with those tested in observational studies. Eleven (79%) of the examined reviews showed no significant difference between observational studies and RCTs. Two reviews concluded that observational studies had smaller effects of interest, while one suggested the exact opposite. Moreover, the review underscored the importance of considering the heterogeneity of meta-analyses of RCTs or observational studies, in addition to focusing on the study design, as these factors influence the estimates reflective of true effectiveness [3].

We must never take away the power and the validity of a well-conducted RCT. However, we need to underline the fact that evidence-based medicine is at risk if we focus myopically on the RCT study design and give it the false credibility of being able to answer all our questions. We must also acknowledge the weaknesses of RCTs and combine information obtained from this study design, while recognising the value of additional information from prospective longitudinal cohort studies. The Fragility Index (FI) is a method for determining the robustness of statistically significant findings in RCTs, and it was recently applied to 48 clinical trials related to sports medicine and arthroscopic surgery [16]. The FI represents the minimum number of patients in one arm of an RCT that is required to change the outcome, from a non-event to an event, in order to change a result from statistically significant to non-significant. So, the lower the number is, the more fragile the significant result. The systematic survey somewhat worryingly showed that the median FI of included studies was 2 [16]. Could it be that we are currently concluding evidence based on the outcome of two single patients in some orthopaedic sports medicine studies? The FI should be an indicative tool in future clinical studies which, in combination with other statistical summaries from a study, could identify results that should be interpreted cautiously or require further investigation [5].

Ultimately, the foundation of science is the ability to generalise the results of a study. The factors that affect the risk of an event or an outcome in a real-life situation are a result of the natural individual variation surrounding us. It is therefore somewhat paradoxical in RCTs to distribute risk factors evenly and eliminate all the factors that may interact with the intervention. We should remember that, when drawing conclusions from a RCT, this is based on many occasions on data obtained from highly specialised centres in one part of the world. The population is enrolled based on strict inclusion and exclusion criteria, which should always trigger the questions of "how many individuals failed to meet them?" and "could their participation have made any difference to the result?" Moreover, RCTs have also been criticised for not representing usual care, which may in fact be the case at a highly specialised centre for sports medicine [1].

High-quality observational studies—an asset in evidence-based medicine

In addition to the generalisability of the results, large observational studies originating from large registers offer advantages in terms of identifying incidences, understanding practices and determining the long-term effects of different types of exposure/intervention. In particular, adverse events can be identified and rare outcomes can be found [13]. Well-conducted, large cohort studies are regarded as the highest level of evidence among observational studies, as the temporality of events can be established. To put it another way, the cause of an event always precedes the effect [13]. The SPORT trial spine, MOON ACLR and MARS revision ACLR are examples of prospective longitudinal cohorts based on STROBE criteria [24] and multivariate modelling, where almost 100% of patients are enrolled. Further, they address the modelling of who will respond to an intervention. While an RCT determines an average of who the intervention will benefit, registers like these determine the individual patient to whom the treatment should be applied, as they model the multitude of risk factors a patient presents to clinicians.

On the other hand, observational studies are limited by indication bias and are subject to the potential effect of unmeasured confounders. The random variation, the confounding factors, must of course be reconciled with existing knowledge in observational studies. The more individual variation, the more the precision of what we are measuring is affected. There is variation in biological responses, in previous therapies, in activity levels and types of activity and variation in lifestyles, to mention just a few. However, we would like to underline the importance of seeing these factors as an opportunity in observational studies. An opportunity to acquire a greater knowledge of the relationship between factors of variance and the outcome, as well as possible underlying mechanisms. Using statistical approaches that adjust for confounders makes a good analysis possible [7, 20].

In order to improve the precision of results from the many registers being established around the world, we

must more clearly define and investigate the true confounders. With regard to anterior cruciate ligament (ACL) reconstruction, data from several large registers have enabled the valuable identification of predictors of outcome. However, there is as yet no existing predictive model for multivariate analysis where confounders are taken into account, which could potentially jeopardise the validity [2]. ACL reconstruction is one of the most researched areas in orthopaedic medicine, and this is therefore noteworthy because the lack of consensus in determining the factors that need to be included in a model may alter the results of studies investigating the same condition. Another key point for high-level register studies is that the representativeness of the cohort is ensured. When registers are being established, comprehensive data entry is needed and it is important that the investigators take responsibility for monitoring the enrolment and attrition of the cohort.

As researchers and clinicians, we sometimes need to take a step back and evaluate how we can continue to implement evidence-based medicine. We must understand that many factors contribute to what we choose to call evidence and that there is no single way to find it. Moreover, scientists before us recognised that only by repeated experiments is it possible to establish the reproducibility of the investigation and thereby get closer to the truth about the efficacy of our treatments. Instead of naively believing that significant results ($P < 0.05$) from any study are synonymous with evidence, we can take advantage of the strengths of different study designs. We should remember that many studies have found that the results of observational and RCT studies correlate well [4, 8, 15]. We encourage the performance of the best research whenever possible. Sometimes this is a well-conducted RCT or highly controlled prospective longitudinal cohorts, and at other times it is the establishment of large patient registers. With regard to the observational study design, future comprehensive prospective cohort studies can provide us with important knowledge and be significant contributors to evidence-based medicine. Nevertheless, the *P* value is a tool that can be helpful, but it must be applied thoughtfully and while appreciating its limitations and assumptions. Evidence-based medicine as defined by the original practitioners involves making a clinical decision by combining clinical experience and the best available evidence from RCTs and registers and by incorporating the patient's values and preferences.

References

1. Albert RK (2013) "Lies, damned lies…" and observational studies in comparative effectiveness research. Am J Respir Crit Care Med 187(11):1173–1177
2. An VV, Scholes C, Mhaskar VA, Hadden W, Parker D (2016) Limitations in predicting outcome following primary ACL reconstruction with single-bundle hamstring autograft—A systematic review. Knee 24(2):170–178
3. Anglemyer A, Horvath HT, Bero L (2014) Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 4:Mr000034
4. Benson K, Hartz AJ (2000) A comparison of observational studies and randomized, controlled trials. N Engl J Med 342(25):1878–1886
5. Carter RE, McKie PM, Storlie CB (2017) The Fragility Index: a P-value in sheep's clothing? Eur Heart J 38(5):346–348
6. Cohen HW (2011) P values: use and misuse in medical literature. Am J Hypertens 24(1):18–23
7. Concato J (2012) Is it time for medicine-based evidence? JAMA 307(15):1641–1643
8. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342(25):1887–1892
9. Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 7(5):541–546
10. Fisher R (1973) Statistical methods and scientific inference, 3rd edn. Hafner Publishing Company, New York
11. Goodman S (2008) A dirty dozen: twelve p-value misconceptions. Semin Hematol 45(3):135–140
12. Greene WL, Concato J, Feinstein AR (2000) Claims of equivalence in medical research: are they supported by the evidence? Ann Intern Med 132(9):715–722
13. Inacio MC, Paxton EW, Dillon MT (2016) Understanding orthopaedic registry studies: a comparison with clinical studies. J Bone Joint Surg Am 98(1):e3
14. Ioannidis JA (2005) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294(2):218–228
15. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J (2001) Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286(7):821–830
16. Khan M, Evaniew N, Gichuru M, Habib A, Ayeni OR, Bedi A, Walsh M, Devereaux PJ, Bhandari M (2016) The fragility of statistically significant findings from randomized trials in sports surgery. Am J Sports Med. doi:10.1177/0363546516674469
17. Kyriacou DN (2016) The enduring evolution of the P value. JAMA 315(11):1113–1115
18. Lowe WR (2016) Editorial Commentary: "There, It Fits!"—Justifying Nonsignificant P Values. Arthroscopy 32(11):2318–2321
19. Mark DB, Lee KL, Harrell FE Jr (2016) Understanding the role of P values and hypothesis tests in clinical research. JAMA Cardiol 1(9):1048–1054
20. Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI) (2012) Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. JAMA 307(15):1636–1640
21. Nuzzo R (2014) Statistical errors—P values, the 'golden standard' of statistical validity, are not as reliable as many scientists assume. Nature 508:150–152
22. Rosenberg W, Donald A (1995) Evidence based medicine: an approach to clinical problem-solving. BMJ 310(6987):1122–1126
23. Salsburg D (2002) The lady tasting tea, 31728th edn. Holt Paperbacks, New York
24. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370(9596):1453–1457