***P*-values should not be used for decision making in the practice of clinical medical physics**

Parminder Basran¹ · Giuseppe Palma² · Clive Baldock³

Accepted: 18 October 2021 / Published online: 26 October 2021

© Australasian College of Physical Scientists and Engineers in Medicine 2021, corrected publication 2021

Introduction and overview: Clive Baldock, moderator

With an increase in the quantification of large and complex datasets in clinical medical physics there is much scope for the appropriate application of statistical methods in the analysis of data. However, whilst this can result in much progress in developing new areas of scientific and clinical research, it potentially raises underlying concerns regarding the robustness of conclusions arrived at with ongoing questions regarding the replication and reproducibility of results.

Whilst the validity of scientific conclusions depends on the statistical methods used, appropriately chosen experimental techniques and data analysis along with the correct interpretation of statistical results ensure that conclusions arrived at are justifiable with uncertainties represented correctly. Many scientific conclusions are based on statistical significance assessed with the *p*-value. However, a number of authors have argued that *p*-values are commonly misused and misinterpreted [1].

In 2005 John Ioannidis stated that most research findings are false [2]. This led to concerns regarding the so-called replication crisis in which many research studies were considered difficult or impossible to replicate or reproduce with it affecting most disciplines in the natural, social and medical sciences [3, 4]. In subsequent years much has been written regarding the replication crisis which has been partially explained through poor experimental design, inappropriate applications of statistical tests and the use of *p*-values [5].

In this spirited debate Parminder Basran and Giuseppe Palma debate whether *p*-values should be used for decision making in the practice of clinical medical physics.¹

Arguing for the Proposition is Parminder Basran, PhD (University of Calgary, 2002), MSc, BSc (University of Alberta, 1994, 1997). Dr Basran is an Associate Research Professor in the College of Veterinary Medicine at Cornell University, New York. He is a medical physicist combining expertise in physics and medicine to help people and animals. He has published in a variety of fields, including safety and quality improvement in human oncology, machine learning methods from medical images and medical image processing, and is currently exploring the use of AI with medical images in the veterinary setting, including detecting diseases in cats and dairy cows, and preventing injuries in thoroughbred racehorses. He has a keen interest in medical physics outreach as Chair of the American Association of Physicists in Medicine (AAPM) Summer Undergraduate Fellowship and Outreach Program, and as a board member for Medical Physics for World Benefit, a not-for-profit organization which helps in delivering safe patient care for those in low to middle-income countries.

✉ Clive Baldock
c.baldock@westernsydney.edu.au

¹ Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA

² Institute of Biostructures and Bioimaging, Italian National Research Council (CNR), Naples, Italy

³ Graduate Research School, Western Sydney University, Penrith, NSW 2747, Australia

¹ Contributors to Topical Debates are selected for their knowledge and expertise. Their position for or against a proposition may or may not reflect their personal opinions.

Parminder Basran

Giuseppe Palma

Arguing against the Proposition is Giuseppe Palma, PhD. Dr. Palma is from Calimera, a small Greek-speaking town in the Heel of the Italian Boot. He received his BSc in physics (2003, high-energy cosmic rays) and his MSc in theoretical physics (2005, gamma-ray bursts) from the University of Pisa. He earned the Diploma in Sciences (2005, corrugational instabilities of hyper-relativistic shocks from hypernova models) and his PhD in physics (2010, high Lorentz-factor astrophysical flows) from the Scuola Normale Superiore of Pisa. He was visiting student at the École Normale Supérieure of Paris in 2004 and at the University of Colorado at Boulder in 2007. From 2008 he turned his attention to Magnetic Resonance Imaging (MRI) sequence design, and in 2011 he moved to Naples as researcher of the Italian National Research Council (CNR) at the Institute of Biostructures and Bioimaging. In 2014 he contributed to first unveil the text of two cyphered missives of the Renaissance duchess Lucrezia Borgia. In 2018 he co-founded a start-up company for e-care platforms that has been acknowledged as corporate spin-off of the Italian CNR. He is the primary inventor of an international patent on quantitative MRI, and co-authored one book and more than 70 papers on astrophysics, medical physics, cryptography and medieval historiography. His main research interests include radiation therapy outcome modelling, quantitative MRI, image processing and analysis.

Opening statement: Parminder Basran

P-values should not be used in clinical decision making by medical physicists or engineers because *p*-values are often misused and misunderstood, and there are few -if any- practical uses of it in clinical practice.

Some might argue that there is confusion over the interpretation and usage of *p*-values. More likely, its value is misplaced and subsequently misused. The *p*-value is the probability of an effect or association (hereafter collectively referred to as ‘effect’), and is most often computed through testing a (null) hypothesis like “the mean values from two samples have been obtained by random sampling from the same normal populations”. The *p*-value is not* the probability of this hypothesis. Full-stop. *P*-values cannot reveal the plausibility, presence, or importance of an effect: this is where scientists fall victim to Sir Francis Bacon’s failures in inductive reasoning—serious enough to warrant a new *idola statistica significationem* [6, 7]. Too common, evidence-based clinical decision making and justification for a clinical practice based on refutation of a null hypothesis are conflated. Yet these are not the same thing. Evidence-based

* In the original text, Professor Basran inadvertently excluded the word ‘not’ from the sentence ‘The *p*-value is not the probability of this hypothesis.’ This has been addressed by Professor Basran in his Letter to the editor (<https://doi.org/10.1007/s13246-021-01074-3>) which further clarifies the definition. Readers are asked to note that Professor Palma’s rebuttal in the debate is to the original text, not the amended text seen in this corrected version. Although this does slightly change aspects of the debate, it is important to correct the statement so that an incorrect definition does not propagate and lead to misinterpretation of the *P* Value.

decisions must include, among other factors, confirmation of an effect, which cannot be achieved solely via refutation of several null hypotheses [8]. It is, as described and suggested by inventor Ronald Fisher, one of several deductive tools in abductive reasoning. And while a statement like ‘ p -value = 0.01’ might exude confidence in the association between two datasets, that confidence diminishes when appreciating that the corresponding risk of no association, i.e., a “false alarm”, is 11% [9]. As per Wasserstein et al. *no p -value can reveal the plausibility, presence, truth, or importance of an association or effect* [1]. And even when it could have application, the underlying distribution may not be symmetric, normal, nor easily dichotomized.

Second, the appropriateness in the use of p -values for medical physics decision making is suspect. If p -values (a) do not confer if an effect is statistically significant; (b) do not establish existence or absence of an effect; (c) do not confer the probability that chance alone is related or unrelated to an effect; and (d) cannot be used to conclude anything of scientific or practical importance based on statistical significance, one has to ask: of what practical value is the p -value anyway [1]? The p -value may have value in dichotomizing data (i.e., an error exceeds some threshold), but more sophisticated tools, such as process control methods, are often more appropriate in medical physics decision making [10]. Confidence intervals along with a discussion on size-effects provide more insight on the relationship between datasets, whether an effect exists, and the extent of that effect.

In summary, a ‘one-size fits all’ solution for statistical inference in clinical medical physics practice rarely suffices. Agreeably, there could be instances where the p -value has relevance in clinical decision making; if doing so, it should be accompanied with an ATOM approach (Accept uncertainty, be Thoughtful, Open, and Modest [1]. But these instances herald misuse, and more sophisticated and appropriate methods should be employed.

Opening statement: Giuseppe Palma

It seems quite odd to state that “ p -values should not be used for decision making in the practice of clinical medical physics”. It sounds like saying that percentages or definite integrals should not be used in the above or another scientific practice. However, if no one would even imagine to rise against the calculus, it has recently become a somehow trendy whim to ostracize that particular, seemingly neutral, quantitative concept in statistical analysis [11, 12].

The history of scientific literature in the last half-century promptly provides a clue of the reasons that—to a certain extent—justify the skepticism toward p -values. They rest on

an impressive body of published misinformation that, in my view, can be labelled as *definition-* or *computation-related*.

The first category is most embarrassing, because it concerns the very meaning of the observed significance level. The involved claims typically rely on the confusion between frequency (likelihood) and hypothesis (posterior) probabilities or on the false equivalence between effect size and statistical significance [13, 14]. Such misconceptions are exacerbated by the dichotomization of the p -values according to arbitrary α -levels [1], whose repercussions are enjoyable only by a good sense of humour in a bunch of scientific memes.

A subtler bait is related to the computation of proper p -values. Anyone who has ever performed a “Hello world”-computation of some test knows that, when expressing *one* p -value in the analysis of a set of data, there is a huge variability in the resulting number, which may in turn abet p -hacking. It depends on the choice of a wide variety of issues (by too many researchers reduced to *technical issues*), usually related to the definition of the statistical models, which, in a nutshell, include but are not limited to the type of the test, the explanatory variables to account for, the correction for multiple comparisons, etc. But this ultimately involves the rigor and the mastery of the methods applied in the study design [15], on one hand, and the scientific integrity of the researchers [16], on the other. If one does not dislike litotes, one might remark that both aspects are not really fostered by Darwinism applied to current mechanisms of scientific careers.

Now, I am not sure I am saying something original, but the solution should not be to throw the baby out with the bath water. The concept of p -value is particularly effective in summarizing the compatibility of a dataset with a specified statistical model, and provides *one* way of quantifying evidence against the test hypothesis or other model assumptions [17]. The fact that a considerable number of researchers—including some physicists—are prone to misunderstand statistical concepts should not be a valid argument against their usage in decision making in clinical medical physics, as well as video games and rock music should not be banned simply because they might desensitize some people to violence. Rather, I believe we should invert the viewpoint, and keep non-skilled people out of decision making at all. Now, that would *significantly* improve my *confidence* in the decisions taken.

Rebuttal: Parminder Basran

My esteemed colleague provides convincing arguments that first, the p -value is one of many tools and thus one should *not throw out the baby with the bathwater*, and second, to keep non-skilled people out of such decision making in the

first place. We concede on the second point, but with respect to the first—consider alternative and more powerful statistical tools [18]. If the need is establishing whether values fall within statistical margins, confidence intervals (e.g., 95%) transmit thresholds for ‘significance’ along with their precision, with no assumptions of the underlying distribution. If the need is hypothesis testing, a valid ‘null hypothesis’ in isolation rarely contributes to a clinical decision. At the risk of spiraling down the Fisherian versus Bayesian rabbit-hole, the magnitude of the alternative hypothesis, by way of the log Bayes Factor, provides a more robust interpretation of whether one outcome is more (or less) likely than another [19]. But perhaps of most clinical utility are magnitude-based inference methods, such as those in statistical process control methods, which provides both confidence limits and upper/lower bounds. The null hypothesis might be falsified for 100 s of VMAT QA measurements (i.e., mean value is not zero but 0.3%), but in practical terms, this means little in the presence of priors, such as knowing the daily output QA should be within $\pm 3.0\%$. On the flipside, if the p -value is not significant from 100 s of QA measurements, yet the mean is $+1.7\%$, one is not likely to ignore this potential systematic error. There are simply few, if any, reasons to continue using p -values in clinical decision making given the evolution of statistical methods. It is time to say farewell to the p -value and embrace more modern statistical approaches.

Rebuttal: Giuseppe Palma

My esteemed colleague’s opening statement—with its wide gamut of assertions including false p -value definitions, unexpected claims of hard-coded links between p -values and false-alarm probabilities (given the fair odds for genuine coin toss, I would happily rid banks of all the coins generating outcome sequences with $p \leq 0.01$ and reimburse for false-alarms in gold up to twice the value of 11% of the detected coins), and some reasonable advices—gives a glimpse of the disorientation in the field.

To clarify, the p -value is the probability of getting a result as or more extreme than the observed result under an assumed statistical model.

Accordingly, I agree with my esteemed colleague that p -value, regrettably, cannot per se evaluate hypothesis probability without introducing subjective estimates of the prior knowledge and marginal likelihoods of competing hypotheses [20, 21] (e.g., 11% is the minimum false-alarm probability after a $p=0.01$ experiment for a 1-to-1 prior odds hypothesis, assuming the $-eplogp$ minimum Bayes factor calibration). And yet, the very absence of further assumptions proves there is room in our studies for a clean, well-defined concept as p -value is.

We also differ in our attitudes towards his positivistic belief in the existence of “more sophisticated tools” that could likely substitute the fideistic p -value fetish (*idolum statisticae significantiae*) with another (*idolum evidentiae*?). In cargo-cult statistics the p -value concept is only under a century ahead of these or other Bayesian alternatives, which will rapidly close the gap if researchers are taught that the key for socially accepted literature is keeping p -values out of report templates. Since “no statistical method is immune to misinterpretation and misuse” [13], we should warn scientists against the risks of thinking in a burst of naïve inductivism [22] that a decision algorithm can replace the critical reasoning, rather than witch-hunt a probability.

References

1. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond “ $p < 0.05$.” *Am Stat* 73:1–19
2. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:e124
3. Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454
4. Serra-Garcia M, Gneezy U (2021) Nonreplicable publications are cited more than replicable ones. *Sci Adv* 7:eabd1705
5. Ritchie SJ (2020) Science fictions: exposing fraud, bias, negligence and hype in science. The Bodley Head, London
6. Ziliak ST (2019) How large are your G -values? Try Gosset’s Guinnessometrics when a little “ p ” is not enough. *Am Stat* 73:281–290
7. Bacon F (1878) *Novum organum*. Clarendon Press, Oxford
8. Significant (2021) xkcd. <https://xkcd.com/882/>. Accessed 18 May 2021
9. Nuzzo R (2014) Scientific method: statistical errors. *Nat News* 506:150
10. Pawlicki T, Yoo S, Court LE et al (2008) Process control analysis of IMRT QA: implications for clinical trials. *Phys Med Biol* 53:5193–5205
11. Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. *Nature* 567:305–307
12. Trafimow D, Marks M (2015) Editorial on null hypothesis testing. *Basic Appl Soc Psych* 37:1–2
13. Greenland S, Senn SJ, Rothman KJ et al (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 31:337–350
14. Altman DG, Bland JM (1995) Absence of evidence is not evidence of absence. *BMJ* 311:485
15. Stark PB, Saltelli A (2018) Cargo-cult statistics and scientific crisis. *Significance* 15:40–43
16. Feynman RP (1974) Cargo cult science. *Eng Sci* 37:10–13
17. Wasserstein RL, Lazar NA (2016) The ASA statement on p -values: context, process, and purpose. *Am Stat* 70:129–133
18. Cumming G (2014) The new statistics: why and how. *Psychol Sci* 25:7–29
19. Goodman SN (1999) Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med* 130:995–1004
20. Nuzzo R (2014) Scientific method: statistical errors. *Nature* 506:150–152
21. Held L, Ott M (2018) On p -values and bayes factors. *Annu Rev Stat Appl* 5:393–419
22. Chalmers A (2013) What is this thing called science? University of Queensland Press, Brisbane