25. Surkan MJ, Gibson W. Interventions to mobilize elderly patients and reduce length of hospital stay. *Can J Cardiol* 2018; **34**: 881–8
26. Bandholm T, Wainwright TW, Kehlet H. Rehabilitation strategies for optimisation of functional recovery after major joint replacement. *J Exp Orthop* 2018; **5**: 44
27. Hughes MJ, Hackney RJ, Lamb PJ, Wigmore SJ, Christopher Deans DA, Skipworth RJE. Prehabilitation before major abdominal surgery: a systematic review and meta-analysis. *World J Surg* 2019; **43**: 1661–8
28. Charette SL, Garcia MB, Reuben DB. Goal-oriented care. In: Bensadon BA, editor. *Psychology and geriatrics: integrated care for an aging population*. London: Academic Press; 2015. p. 1–19
29. Steele Gray C, Grudniewicz A, Armas A, Mold J, Im J, Boeckxstaens P. Goal-oriented care: a catalyst for person-centred system integration. *Int J Integr Care* 2020; **20**: 8

British Journal of Anaesthesia, 126 (3): 564–567 (2021)
doi: 10.1016/j.bja.2020.10.042
Advance Access Publication Date: 6 January 2021

What is the true worth of a *P*-value? Time for a change

George Hadjipavlou[1,*], Richard Siviter[2] and Birte Feix[2]

[1]Nuffield Department of Anaesthetics, Oxford University Hospitals NHS Foundation Trust, Oxford, UK and
[2]Neurosciences Intensive Care, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

*Corresponding author. E-mail: georgehadjipavlou@gmail.com

Keywords: Bayesian; Bradford-Hill; power; P value; statistics

Losing faith

The definition of a *P*-value is the probability that, for the sample data, no difference exists between the explored variables. It confers no meaning with respect to the cause–effect relationship, nor its size nor presence. Yet over time, the *P*-value seems to have acquired the unconscious assumption that if a study reports a significant *P*-value (in general, $P \leq 0.05$), then there must be a true difference between samples' representative populations. This is not an appropriate conclusion to make, and proving cause and effect remains a separate issue.

The origins of the *P*-value cut-off of $P<0.05$ for significance can be traced back to the mid-1920s and were proposed by Fisher in describing robust ways to identify significance in agricultural field tests.[1] Yet a value of $P<0.05$ meant for agriculture, seems to be applied throughout medical research without real justification. In medicine, where peoples' lives may change on the outcomes of these trials, an individual might expect a more stringent *P*-value cut-off when the cost of being wrong is more consequential.

As a group, we have been struck by the prevalence of misconceptions regarding interpretation of statistics. This may be attributable to insufficient teaching within clinical and research training, and poor reporting of statistical methods within articles. These factors conspire to generate a fear of statistics; and admitting this is difficult. This in turn has led to both a poor understanding of and an over-reliance on the *P*-value as some form of currency of how good a study's conclusions are. The fault is not entirely with us though, as access to a statistician, particularly one who also understands the medical field, is difficult.

Unsurprisingly then, there is growing dissatisfaction with the *P*-value as researchers spend huge resources to achieve a statistically significant result only to find it overturned on study replication or by a meta-analysis. An extreme example of this is the banning of *P*-values in some journals.[2] There is so much concern about *P*-value misuse that the American Statistical Association issued a statement on statistical significance and *P*-values in 2016[3] summarising it essentially as:

- An indicator of how incompatible the data are with the specific statistical model.
- Not a measure of the probability that the studied hypothesis is true, or the that the data were produced by random chance.
- Scientific conclusions and business or policy decisions should not depend solely on the *P*-value, with scientific inference requiring full reporting and transparency.
- Not a measure of effect size or importance.
- Nor is it by itself a measure of the evidence for a model or hypothesis.

Back to basic principles

There are also growing efforts to move away from *P*-values towards other measures to better portray the reliability of conclusions. In our opinion, this is incorrect as the *P*-value is the cornerstone of statistical testing. Many replacements are offered such as confidence intervals along with *P*-values, Bayesian likelihood of the null hypothesis *vs* the

alternative hypothesis, Akaike criterion, and false-positive risk.[4] In our opinion, the Bayesian approach makes the most sense, but ultimately replacing the *P*-value with something else does not change the fundamental problem of using statistics to replace basic scientific principles needed to establish cause and effect. The Bradford Hill criteria[5] are better at this and are ultimately what we should rely on; they are:

1. Strength of the relationship
2. Reproducibility/replicability
3. Specificity: the more specific, the greater the probability of cause–effect
4. Temporality: one variable precedes the other
5. Biological gradient, that is a dose–response effect
6. Plausibility: there is a credible theory that can explain the result
7. Agreement between laboratory and epidemiological results
8. Supporting experimental evidence of the effect
9. Similar circumstances reproduce similar results
10. Reversibility: the effect can be undone by removing the cause

A simplified approach

Instead of abandoning the *P*-value, which is essential to statistical testing, we can do much better by estimating the true worth of a study's conclusions (H0 *vs* H1) generated by using the *P*-value by interpreting it alongside the Bradford Hill criteria. To do this, we also need some estimate of the prior probability of our study effect existing. This at best is often a guess, but informed guesswork is better than nothing.

Table 1 shows an approach for estimating the worth of a study's conclusions, that is the *P*-value's worth, using epidemiological odds tables. We can create a positive and negative predictive value (PPV and NPV, respectively) of the desired 'study result', using the prior assumption of the likelihood of the cause–effect relationship existing in real life, such as disease incidence, along with the planned *P*-value and study power.

From the field of anaesthesia, we take the B-Aware trial as an example.[6] In this study ~2500 individuals were randomised to receive anaesthesia guided by bispectral index (BIS) monitoring (an electroencephalographic monitoring tool) or routine care. The study hypothesis was that BIS-guided anaesthesia led to fewer cases of awareness. They initially determined their sample size of 1090 using a power of 80% and a cut-off *P*-value of 5%. They observed 2/1225 awareness cases in the BIS group and 11/1238 in the routine care group and concluded with a $P=0.022$ that BIS-guided anaesthesia reduced intraoperative awareness by 82% with a confidence interval of 17–98%.

Let us assume that we have no idea beforehand whether BIS is going to reduce awareness. The odds are 50:50 – meaning it may be there or it may not. Using the *P*-value of 5% and power of 80%, Table 1 shows how the numbers play out. We find the PPV of a significant *P*-value of 5% is 83%, and the NPV of no effect is 94%. In this specific context, concluding BIS reduces awareness may be the correct conclusion ~80% of the time. This result allows us to be more objective about the study conclusion. The fact that another trial of similar sample size and incidence of awareness found no difference[7] shows how delicate these probabilities are. As the study is replicated, our prior belief can be adjusted by the evidence.

What if our prior belief is not 50:50? If we have a low or high prior belief of a cause–effect relationship existing with a power of 80% and *P*-value of 5%, we discover that the PPV and NPV can have very different meaning (Fig. 1a). With a low prior (10% chance of effect existing), we see that the NPV is 99% and so very useful, whereas the PPV of 35% is not. This means that a conclusion that an effect exists is quite probably incorrect when there is either good evidence it does not or that it is scientifically implausible (low satisfaction of the Bradford Hill criteria). With a high prior (90% chance of effect existing), we see that the PPV is 98% and NPV is 64%, and so the reverse is now true – that is a conclusion that the effect exists is helpful whereas a conclusion it does not is less so. If the power of the study falls for some unknown reason (Fig. 1b), it is the PPV that is most affected when the prior is low or unknown (50:50), whereas it is the NPV that is affected with a high prior.

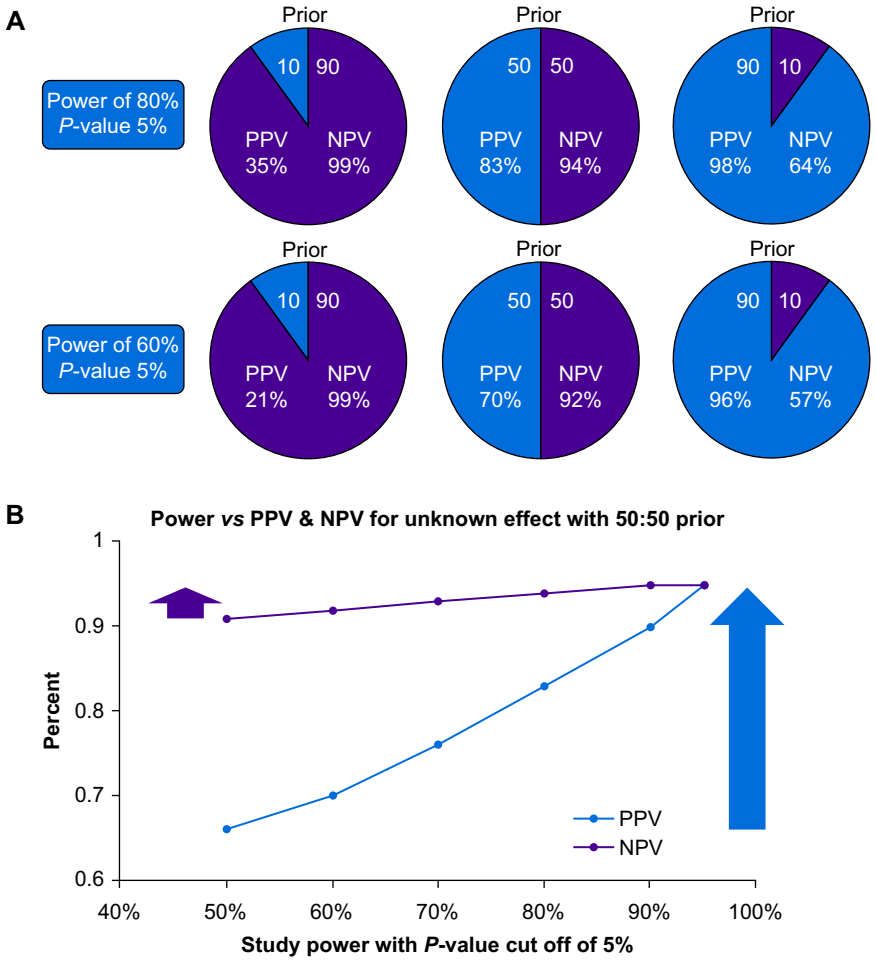
This relationship between prior, *P*-value, power, and their effects on PPV and NPV is complex, and so we recommend using Figure 1c to help plan any specific study. Overall, though, power appears to be more influential in the 'unknown' case scenario which is more commonly faced for new research.

Figure 1c shows the impact of chosen power and *P*-value cut-off on the PPV and NPV at a study's planning stage for the *a priori* chance of 50:50 of a difference existing. We recommend seeking a power > 90%, and a *P*-value < 0.05 as at this level and above, the PPVs and NPVs are both at 90% or higher.

A *posthoc* analysis of a study using this approach is also informative. A poorly powered study (60%) using a cut-off *P*-

Table 1 A Bayesian approach to the probability of an effect using epidemiological principles

	Effect exists	**Effect does not exist**	
The study finds the effect	A; (1–*P*-value)×N_1	B; (1–power)×N_2	PPV=A/(A+B)
The study does not find the effect	C; *P*-value×N_1	D; Power×N_2	NPV=D/(C+D)
	N_1=assumed probability of effect existing (0–1)	N_2=1–N_1	
The Bayesian approach of the existence of an effect assuming a study power of 80%, *P*-value of 5%, and a likelihood of effect existence of 50%, i.e. unknown			
	Effect exists	**Effect does not exist**	
The study finds the effect	0.475	0.1	PPV=83%=0.475/(0.475+0.1)
The study does not find the effect	0.025	0.4	NPV=94%=0.4/(0.025+0.4)
Prior probabilities	0.5	0.5	



C For unknown prior (50:50) of cause and effect

Negative predictive value

P-value \ Power	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0.01	0.98	0.98	0.98	0.99	0.99	0.99	0.99
0.025	0.94	0.95	0.96	0.97	0.97	0.97	0.97
0.05	0.89	0.91	0.92	0.93	0.94	0.95	0.95
0.10	0.80	0.83	0.86	0.88	0.89	0.90	0.90
0.25	0.62	0.67	0.71	0.74	0.76	0.78	0.79
0.50	0.44	0.50	0.55	0.58	0.62	0.64	0.66
0.80	0.33	0.38	0.43	0.47	0.50	0.53	0.54

Positive predictive value

P-value \ Power	0.40	0.50	0.60	0.70	0.80	0.90	0.95
0.01	0.62	0.66	0.71	0.77	0.83	0.91	0.95
0.025	0.62	0.66	0.71	0.76	0.83	0.91	0.95
0.05	0.61	0.66	0.70	0.76	0.83	0.90	0.95
0.10	0.60	0.64	0.69	0.75	0.82	0.90	0.95
0.25	0.56	0.60	0.65	0.71	0.79	0.88	0.94
0.50	0.45	0.50	0.56	0.63	0.71	0.83	0.91
0.80	0.25	0.29	0.33	0.40	0.50	0.67	0.80

Fig 1. Positive predictive values (PPV) and negative predictive values (NPV) for different priors of the cause–effect existing, and the effect of power on PPV and NPV.

value of 5% has a PPV of 70% and an NPV of 92%. Although researchers do not intentionally under-power their studies, poor power comes as a consequence of unaccounted measurement error and effect size over-estimation. A recent review of meta-analyses found that approximately 50% of the studies surveyed had a statistical power less than 20%.[8] This is a significant problem if we extrapolate this result across the medical field and logically explains why we observe so little success in the thousands of studies out there.

Where do we go from here?

We suggest that experiments reshape their planning to focus on how meaningful their conclusions are going to be scientifically and statistically.

For single research trials, the following should be planned for and reported either from the study data or past published reports:

1. Strength of the relationship
2. Biological gradient, that is a dose-response relationship
3. Specificity
4. Temporality
5. Similar circumstances reproduce similar results
6. Reversibility
7. A power and *P*-value to meet an assigned study PPV and NPV

For data syntheses of multiple studies, the following are also important:

8. Reproducibility
9. Agreement between laboratory and epidemiological results
10. A credible mechanism exists that can explain the result
11. Experimental evidence supporting the cause–effect

Study design should focus around meeting as many of the Bradford Hill criteria possible, and the power and *P*-value chosen to achieve the highest possible study PPV and NPV using calculations as demonstrated here. A conclusion should be grounded in the same principles. Researchers in the planning phase can use the approaches described in this article to help them achieve this. We must report our statistics better, focusing on how power and sample size analysis is done, including whether the sample and power is achieved in real life. Published studies should clearly state the level of statistical expertise involved in their design and analysis.

In conclusion, we should stop seeking more from the *P*-value than it can provide and move towards a more scientifically robust reporting structure for validating our conclusions.

Authors' contributions

Study concept, theory development: GH
Drafting of the article: GH, RS, BF
Interpretation of statistical approach: RS, BF

Declarations of interest

The authors declare that they have no conflicts of interest.

References

1. Fisher R. 048: the arrangement of field experiments. *J Min Agric Gr Br* 1926; **33**: 503–13
2. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psychol* 2015; **37**: 1–2
3. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *Am Stat* 2016; **70**: 129–33
4. Halsey LG. The reign of the *p*-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol Lett* 2019; **15**: 20190174
5. Hill S. Pharmacokinetics of drug infusions. *Contin Educ Anaesth Crit Care Pain* 2004; **4**: 76–80
6. Myles PS, Leslie K, McNeil J, et al. Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet* 2004; **363**: 1757–63
7. Avidan MS, Zhang L, Burnside BA, et al. Anesthesia awareness and the bispectral index. *N Engl J Med* 2008; **358**: 1097–108
8. Dumas-Mallet E, Button KS, Boraud T, et al. Low statistical power in biomedical science: a review of three human research domains. *R Soc Open Sci* 2017; **4**: 160254

British Journal of Anaesthesia, 126 (3): 567–571 (2021)
doi: 10.1016/j.bja.2020.10.023
Advance Access Publication Date: 17 December 2020

Preoperative considerations of new long-acting glucagon-like peptide-1 receptor agonists in diabetes mellitus

Abraham H. Hulst[1,*], Jorinde A. W. Polderman[1], Sarah E. Siegelaar[2], Daniel H. van Raalte[3], J. Hans DeVries[2,4], B. Preckel[1] and Jeroen Hermanides[1]

[1]Department of Anaesthesiology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands, [2]Department of Internal Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands, [3]Department of Internal Medicine, Amsterdam UMC, Amsterdam, the Netherlands and [4]Profil Institute for Metabolic Research, Neuss, Germany

*Corresponding author. E-mail: a.h.hulst@amsterdamumc.nl