Leukemia (2016) **30**, 1965–1967

www.nature.com/leu

EDITORIAL

What is the (p-) value of the *P*-value?

Leukemia (2016) **30,** 1965–1967; doi:10.1038/leu.2016.193; published online 26 August 2016

One should try everything in life except incest, folk dancing and calculating a P-value.

After Sir Thomas Beecham, 2nd Baronet, CH

Inferential processes including notions of cause-and-effect are increasingly dominated by statistical analyses, a process in which the concept of the *P*-value has become central. Every scientific researcher thinks he/she knows what a *P*-value is. Some can calculate it themselves, others need help from a statistician or sometimes several statisticians to find an agreeable *P*-value. Why shop around (commonly termed P-hacking)? Because without a statistically significant *P*-value it is difficult or impossible to publish your research in a scientific journal with a high impact factor. (More on impact factors in a subsequent Editorial.) Usually statistical significance in this context is defined as a pre-set *P*-value < 0.05. Get a *P*-value of 0.055 and you are out of luck. *Journal of Irreproducible Results* here we come (impact factor − 2.3). Use of *P*-values in scientific reports is increasing. About 16% of articles in MEDLINE in 2014 including about 35% of meta-analyses, 40% of clinical trials and about 55% of randomized clinical trials mentioned a *P*-value.[1] (One wonders about the 45% of published randomized clinical trials with no *P*-value?) Most reported *P*-values reported cluster around $P = 0.05$. Ninety-six percent of articles reported in PubMed Central with one or more *P*-values claimed a statistically significant result with more than one *P*-value < 0.05.

Interestingly, and somewhat alarmingly for us, statisticians are less certain than are researchers what precisely a *P*-value is, means and implies about reproducibility and replicability of scientific data. This is entirely understandable given the complex history of the *P*-value. Credit for introducing the *P*-value into scientific research usually goes to the distinguished statistician Fisher.[2] Fisher defined the *P*-value as: 'the probability of the observed result, plus more extreme results, if the null hypothesis were true.' Importantly for what follows, the Fisher *P*-value was not meant to stand alone but to provide a measure of statistical inference in the context of the more complex process of scientific inference. Proper use of the Fisher definition of the *P*-value requires several assumptions including no relationship between the causal factor and outcome, no systematic error(s) such as misclassification or confounding and use of the appropriate statistical test. Given these assumptions it is easy to see why some researchers think the *P*-value indicates the probability the null hypothesis is true or that a high *P*-value indicates there is no association between a variable and an outcome. As we will see, both notions are incorrect. The next advance in the interaction between inferential thinking and statistical inference came with Neyman and Pearson who, in contrast to Fisher, postulated a formal, mutually exclusive alternative hypothesis to the null hypothesis and a preselected value level to reject the null hypothesis. The Neyman–Pearson approach was intended as a decision-making analysis whereas the Fisher approach was not. The inherent rigidity of the Neyman–Pearson approach comes with two unavoidable potential errors, rejecting the null hypothesis when it is really true (type-1 error) and accepting the null hypothesis when it is really false (type-2 error). It is important to realize that the Fisher and the Neyman–Pearson approaches are frequentist; both ignore a third approach, Bayesian inductive reasoning (see below).

"You can't keep revising the data to prove you would be the best Valentine date for Kate Upton"

Let's consider a recent example of use and misuse of a *P*-value. When physicists discovered the Higgs boson, they said the strength of the data supporting it is 5-sigma, meaning a *P*-value of $3 \times 10E-7$ or about 1 in 3.5 million. Contrary to what you may think, this is not the probability the Higgs boson exists. Rather it is the probability if the Higgs boson does not exist (the null hypothesis) the experimental data obtained would be at least as or more extreme as what was observed. Put otherwise, these results would occur in only 1 in 3.5 million replicate experiments. Bear in mind that the first experiment could be the one with that result, and replicating the experiment 3.5 million times more would produce the same less-extreme results. This type of statement annoys most people who are interested in knowing whether the Higgs boson probably exists. Sorry, this is not what the *P*-value is about. Also, it should be noted that, despite widespread use of $P < 0.05$ or even $P < 0.01$ as evidence of statistical significance in biomedical research, other scientific disciplines such as high-energy physics require lower *P*-values of $P < 0.003$ or even $P < 0.0000003$ depending on the novelty and importance of the claim.[3]

It is also worth considering the so-called rare events that occur when sufficient numbers of experiments are done. For example, if you flip a fair (even-sided) coin 100 times it is rather unlikely you will get 66 heads and 34 tails. However, if 1000 people each flip the same coin 100 times a few will get >60 heads and <40 tails. So, even if one observer gets this unlikely outcome he/she would be mistaken to conclude the coin is not even-sided. This notion is referred to as the improbability principle and is reviewed elsewhere.[4]

Another point to consider is that the *P*-value depends not only on the data but on how data were analyzed. For example, with the exact same data a reported *P*-value will differ depending on whether there were 1, 2 or more interim analyses. Also, the *P*-value approach must account for not only the observed data but more extreme data, otherwise *P*-values would be very small. If you toss a fair coin 6 times the probability of getting 4 heads and 2 tails is 0.23. But to test the null hypothesis (that the coin is not biased) you need to add the probability of 5 heads and 1 tail (0.09) and

the probability of 6 heads and no tails (0.01) to the probability of the actual data. Hence the *P*-value would be about 0.33 and the null hypothesis would not be rejected.

Recently, the American Statistical Association published a report on the *P*-value: 'what it is, what it means and how P-values should be interpreted.'[5] This is an extraordinary event, one that the Association, the world's largest professional organization of statisticians, has never done before indicating how important this issue has become. The Association solicited input from statisticians, many well-known to readers of *LEUKEMIA* including Donald Berry, Sander Greenland, John Ioaniddis and Kenneth Rothman. The expert panel prepared and published a report. It would be a misrepresentation to term this a consensus statement as there is often no consensus. We interpret this controversy favorably because consensus implies a degree of agreement not achieved by the panel. After all, to quote Michal Crichton: 'Historically consensus has been the first refuge of scoundrels. It is a way to avoid debate by claiming that the matter is settled.' Doubt it? Think of consensus that the world is flat or Earth is the center of the universe. Both proved false despite the so-called consensus. Although the threat of excommunication was a good incentive for consensus previously (think Galileo and Cardano), nowadays it is the fear of missing your Friday afternoon flight from Dulles airport after a National Institutes of Health consensus conference.

The issue the Association addressed is the context, process and purpose of the *P*-value. The issue was prompted by a query to the Association from Professor George Cobb who asked (rhetorically): '(1) why do so many colleges and graduate schools teach $P = 0.05$; and (2) why do so many people still use $P = 0.05$.' The answers seemed to be 'that's what the scientific community and journal editors want and because that's what students were taught in college or graduate school.' These explanations are not terribly reassuring. Nor are these concerns new. Worries over uses and abuses of $P < 0.05$ as the basis for rejecting the null hypothesis have been described as having more flaws than Facebook's privacy policy! [6] Misuse of *P*-values in hypothesis testing has led to a reproducibility crisis in modern science and especially in clinical medicine.[7] Readers will recall the study by Amgen scientists who were unable to reproduce results of 47 of 53 landmark cancer typescripts reported in the scientific literature.[8] Because of this controversy, some journals have gone so far as to ban use of *P*-values.[9]

The American Statistical Association panel report pointed out the notion of statistical significance that underpins many scientific conclusions, a concept in which the implication of statistical significance is derived or inferred from an index termed the *P*-value. Unfortunately, this index is commonly misused and/or -interpreted. To quote from their report: 'The validity of scientific conclusions including their reproducibility depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that **uncertainty surrounding them is represented properly'** (emphasis ours).

So what does the report say? First, it defines a *P*-value as the probability under a specified statistical model; a statistical summary of the data would be equal or more extreme than the observed value(s). We emphasize the notion under a specified statistical model. When we calculate a *P*-value we are not testing whether the difference between groups or cohorts occurred by chance but rather consistency of the data with a proposed statistical model. Often the statistical model being tested in clinical trials is the so-called null hypothesis. Consequently, when we consider the *P*-value we need to understand that it neither addresses whether the null hypothesis is true nor whether the statistical analysis of the results can be accounted for by chance. Namely, it does not provide an explanation of the results observed but rather how well they match the hypothesized statistical model. Another consideration is that the *P*-value does not reflect effect size. The report states 'scientific conclusions and business or policy decisions should not be based only whether a *P*-value passes a specific threshold.' We agree. Clinically unimportant effects can have very low *P*-values when the sample size is large, whereas clinically more important effects can have higher *P*-values because of a small sample size. Estimated clinically important effects with confidence intervals/bands and *P*-values should be transparently reported. For example, hazard ratios of time-to-event outcomes can be large and statistically significant even when the effect size is small. For this reason restricted mean survival times-based measures should also be reported. [10]

Some biomedical studies cherry-pick results with *P*-values < 0.05 based on multiple subgroup analyses disregarding the small sample sizes in these subgroups. Researchers should report all statistical analyses done and all hypotheses tested so the reader can consider false discovery rates which may need to be considered when multiple comparisons are done. The report does not deal with power (type-2 error), an important issue in evaluating a not statistically significant *P*-value.

The panel emphasizes the inadvisability of focusing on an arbitrary point such as $P < 0.05$ to justify scientific claims. There are two issues here. First, considering $P = 0.05$ as a point for deciding on statistical significance is arbitrary and without a sensible mathematical rationale. The second is that other factors need consideration in evaluating whether an outcome is statistically significant, such as study-design, measurement accuracy, evidence external to the study, accuracy of measurements and validity of assumptions underlying the data analyses. For example, a survival endpoint will usually be more valid than a leukemia-relapse endpoint. The authors correctly state: [The] 'widespread use of statistical significance (generally accepted as $P < 0.05$) as a license for making a claim of scientific finding (or implied truth) leads to a considerable distortion of the scientific process.' These considerations remain in the realm of frequentist statisticians. Although a discussion of using Bayesian inductive reasoning with a spectrum of probabilities (such as credibility limits) to express casual inference is beyond the scope of our discussion, it should also be considered. A recent review by Kyriacou[11] discusses use and limitations of a Bayesian induction approach. Scientific inferences based on use of frequentist and Bayesian methods are not mutually exclusive and often complementary.

Another issue is that researchers often conduct multiple analyses of their data but may present only analyses with a statistically significant *P*-value. This does not allow the reader to fairly evaluate validity of the researchers' claims and conclusions and to consider potential biases. This p-hacking is unfair, inappropriate and should be avoided. The bottom line is that the *P*-value in isolation cannot be relied on to determine whether or not the null hypothesis is correct. There are several other important considerations regarding the *P*-value not covered in the Associations report, and we refer interested readers to other reviews.[11–13]

Although we plan no immediate change in the statistical review process for *LEUKEMIA*, it is important that researchers submitting typescripts follow best statistical practices and acknowledge in their analyses and discussions limitations of the *P*-value in establishing causal inference.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Prof Elihu Estey (University of Washington) and Dr Owen Hoffman (Oak Ridge Center for Risk Analysis) kindly reviewed the typescript. RPG acknowledges support

from the National Institute of Health Research (NIHR) Biomedical Research Centre funding scheme.

RP Gale[1], A Hochhaus[2] and M-J Zhang[3]
[1]*Haematology Research Centre, Division of Experimental Medicine, Department of Medicine, Imperial College London, London, UK;* [2]*Abteilung Hämatologie/Internistische Onkologie, Universitatsklinikum, Jena, Germany and* [3]*Division of Biostatistics and Center for International Bone Marrow Transplant Research, Medical College of Wisconsin, Milwaukee, WI, USA*
E-mail: robertpetergale@alumni.ucla.edu

REFERENCES

1 Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA* 2016; **315**: 1141–1148.
2 Fisher RA. *Statistical Methods for Research Workers*. Oliver & Boyd: Edinburgh, United Kingdom, 1925.
3 Lamb E 5 sigma what's that. Sci Am. 2012. Available at: http://blogs.scientificamerican.com/observations/five-sigmawhats-that/.
4 Hand DJ. *Why Coincidences, Miracles and Rare Events Happen Every Day*. Scientific American/Farrar, Straus and Giroux: New York, 2014.
5 Wasserstein RL, Lazar NA. The ASAs statement on p-values: context, process and purpose. *Am Stat* 2016; **70**: 129–133.
6 Siegfried T. Science News, 2014. Available at: https://www.sciencenews.org/blog/context/make-science-better-watch-out-statisticalflaws.
7 Peng R. The reproducibility crisis in science: a statistical counterattack. *Significance* 2015; **12**: 30–32.
8 Bagley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature* 2012; **483**: 531–533.
9 Trafimow D, Marks M. Editorial. *Basic Appl Soc Psychol* 2015; **37**: 1–2.
10 Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ration and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol* 2016; **34**: 1813–1819.
11 Kyriacou DN. The enduring evolution of the p value. *JAMA* 2016; **315**: 1113–1115.
12 Vickers A. *What is a p-value anyway?* Addison Wesley Longman: Boston, 2009; pp 212.
13 Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN *et al.* Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretation. *Eur J Epidemiol* 2016; **31**: 337–350.