Bone Marrow Transplantation (2016), 1–2

www.nature.com/bmt

SPECIAL REPORT

What is the *P*-value anyway?

RP Gale[1] and M-J Zhang[2]

Bone Marrow Transplantation advance online publication, 11 July 2016; doi:10.1038/bmt.2016.184

Without certainty science is nothing more than seemingly sophisticated guesswork.

Sir Francis Bacon

Statistical analyses and consequently statistical inferences are increasingly important components (but not all) of inferential processes. For example, consider a transplant study designed to determine whether a new drug decreases the likelihood of developing acute graft-*versus*-host disease (GvHD) compared with a placebo. (In statistics there is a distinction between likelihood and probability. The number that is the probability of some observed outcomes given a set of parameter values is regarded as the likelihood of the set of parameter values given the observed outcomes.) You do the study, collect the results and perform a statistical test, typically a test of a statistical model, often the *null hypothesis*. Colleagues and reviewers expect you to generate a *P*-value from these analyses. Usually *statistical significance* in this context is defined as a pre-set *P*-value < 0.05. A *P*-value of 0.055 is considered not *statistically significant*. Does a *P*-value 0.055 mean: (A) the new drug was ineffective? (B) The results can be accounted for by chance? (C) The *null hypothesis is true?* (D) All of the above? The correct answer is (E), none of the above. However, try to publish this study in *Bone Marrow Transplantation* and the Editors, Hillard Lazarus and Mohamad Mohty, are likely to send you a *rapid rejection* e-mail.

Although most scientific researchers think they know what the *P*-value is and what it implies about correct interpretation of their data, this is not so. When we informally sampled a cohort of scientific researchers including junior and senior scientists and clinicians most told us the *P*-value was a test of whether the *null hypothesis* was true or not and whether a factor was *significantly associated* with an outcome, notions that we will see are wrong.

Interestingly, statisticians are no more certain what a *P*-value is than are scientific researchers. To understand why we need to consider the complex history of the *P*-value. RA Fisher introduced the *P*-value into scientific research as a measure of statistical inference.[1] He defined it as *the probability of the observed result, plus more extreme results if the null hypothesis were true.* Fisher suggested the *P*-value could be used as a measure of statistical inference, a component, but not the only component, of the more complex process of causal inference. Several assumptions underlie correct use of Fisher's *P*-value. Unfortunately, many of these assumptions, such as no misclassification or confounding, are difficult to avoid in complex clinical trials in bone marrow transplantation. Consider the trial we mentioned of a new drug to prevent acute GvHD. Analyses of a survival end-point have to consider confounding between reducing the incidence of acute GvHD at the expense of increasing leukaemia relapse. Researchers should study, compare and report cumulative incidence rates of developing acute GvHD, relapse without acute GvHD and death without acute GvHD and relapse simultaneously because these outcomes are not mutually independent.

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."

Neyman and Pearson came next in *P*-value history. In contrast to Fisher they postulated a formal, mutually exclusive alternative hypothesis to the *null hypothesis* and a preselected *P*-value level to reject the *null hypothesis*. This subtle but important difference involves decision-making. The Neyman–Pearson approach results in a decision regarding casual inference whereas the Fisher approach does not. It is important to realize the Fisher and Neyman–Pearson approaches are frequentist, ignoring a third approach of Bayesian inductive reasoning (see below). (Frequentist inference is a type of statistical inference based on conclusions from sample data by emphasizing the frequency (or proportion) of the data.)

The reader may wonder why we are discussing such a seemingly simple-minded question of what the *P*-value really means at this late hour. However, we are not alone in our concern regarding widespread misunderstanding of what a *P*-value is and what it means. 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*.[2] To be clear, this is not a consensus statement; often there was considerable disagreement among expert statisticians on this question so readers need not feel perplexed if they are confused.

The Association panel report pointed out the *P*-value is commonly misused and/or misinterpreted. The report defines a

[1]Haematology Research Centre, Division of Experimental Medicine, Department of Medicine, Imperial College London, London, UK and [2]Division of Biostatistics and Center for International Bone Marrow Transplant Research, Medical College of Wisconsin, Milwaukee, WI, USA. Correspondence: Professor RP Gale, Haematology Research Centre, Division of Experimental Medicine, Department of Medicine, Imperial College London, London SW7 2AZ, UK.
E-mail: robertpetergale@alumni.ucla.edu
Received 25 May 2016; accepted 1 June 2016

P-value as the probability, *under a specified statistical model* a statistical summary of the data would be equal to or more extreme than the observed value(s). We emphasize *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 the consistency of the data with a proposed statistical model. The statistical model being tested in clinical trials is the *null hypothesis*. Consequently, when we consider the *P*-value we need to understand it does not address whether the *null hypothesis* is true nor whether the statistical analysis of the results can be accounted for by chance (common misconceptions in our survey).

In clinical trials it is also important to consider that the *P*-value does not reflect *effect size*. For example, a 5% decrease in the incidence of acute G*v*HD could be associated with a *P*-value < 0.05 when there is a very large sample size, whereas a true 50% decrease might be associated with a *P*-value > 0.05 when the sample size is small. Estimated clinically important effects with confidence intervals/bands and *P*-values should be transparently reported. 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 that the reader can consider false discovery rates, which should be considered when multiple comparisons are done.

The Association panel makes another important point for the readers of *Bone Marrow Transplantation,* namely, it is inadvisable to focus on an arbitrary point such as $P < 0.05$ to claim statistical inference. 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 underpinning. Second, other factors need consideration in deciding whether an outcome is *statistically significant* including study design, measurement accuracy, evidence external to the study, accuracy of measurements and validity of assumptions underlying the data analyses. For example, a survival end point will usually be more valid than a leukaemia-relapse end-point. To quote the report: [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*. This should be the take-home message from our typescript.

Up to this point we have discussed considerations in the realm of frequentist statistics. Although a discussion of using Bayesian inductive reasoning with a spectrum of probabilities (such as credibility limits) to express causal inference is beyond the scope of our discussion, this approach is increasingly considered, especially when there is uncertainty in the accuracy of measurements (such as who really has acute G*v*HD *versus* a virus infection or a drug-induced rash). A recent review by Kyriacou[3] discusses use and limitations of a Bayesian induction approach. Scientific inferences based on using frequentist and Bayesian methods are not mutually exclusive and often complementary, hence we urge readers to consider both.

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 evaluate the validity of the researchers' claims and conclusions and consider potential biases. This *P*-hacking is unfair, inappropriate and should be avoided. (This *data-dredging* or *fishing expedition* is not unlike multiple non-pre-specified subgroup analyses which should be considered hypothesis-generating, require confirmation and require statistical adjustment for multiple comparisons.) The bottom line is the *P*-value in isolation cannot be relied on to determine whether the *null hypothesis* is correct or not. There are several other important considerations regarding the *P*-value not covered in the Associations report and we refer interested readers to other reviews.[3,4]

The Editors tell us they plan no immediate change in the statistical review process for *Bone Marrow Transplantation*. However, it is important researchers submitting typescripts follow best statistical practices and acknowledge in their analyses and discussions limitations of the *P*-value in establishing causal inference. There will be a session on *P*-values and their correct interpretation sponsored by the Center for International Blood and Marrow Transplant Research at the next tandem meetings for transplant scientists and physicians seeking more detail.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

RPG acknowledges support from the National Institute of Health Research Biomedical Research Centre funding scheme. Parts of this typescript were published in *Leukemia*. Prof Hillard Lazarus kindly reviewed the typescript.

REFERENCES

1 Fisher RA. *Statistical Methods for Research Workers*. Oliver & Boyd: Edinburgh, UK, 1925.
2 Wasserstein RL, Lazar NA. The ASAs statement on p-values: context, process and purpose. *Am Stat* 2016; **70**: 129–133.
3 Kyriacou DN. The enduring evolution of the p value. *JAMA* 2016; **315**: 1113–1115.
4 Vickers A. *What is a p-Value Anyway?* Addison Wesley Longman: Boston, 2009, pp 212.