

The *p*-value – a well-understood and properly used statistical concept?

doi:10.1111/j.1600-0536.2011.02002.x

The *p*-value as the result of a statistical test is a universally used concept in medical research. For example, going through the 66 original articles of *Contact Dermatitis* in 2010, we found 44 articles [66% (95% confidence interval (CI) 55–78%)] that explicitly report at least one *p*-value or refer to the result of a statistical test.

As it is used this frequently, one might assume that both authors and readers understand what the *p*-value actually means. There has been empirical research on how medical doctors understand basic biostatistical concepts. According to Windish et al. (1), 59% (95% CI 53–65%) of 277 medical residents correctly interpreted the meaning of a *p*-value in a four-item multiple-choice question (Table 1). In this study, the majority [88% (95% CI 84–92%)] of residents reported fair to complete confidence in their understanding of *p*-values. Using the same questionnaire, Darmon and Tabah (2) found that 68% (95% CI 61–74%) of 208 French intensivists gave the correct answers to the *p*-value question. However, and as pointed out by Goodman (3) 'in an exquisite irony', none of the four answers offered for the *p*-value question was correct.

There are several misconceptions regarding the correct meaning of the *p*-value, and Goodman (3) gave as many as 12 of them. In this editorial, we focus on what we feel is the most common and most serious of them: interpreting the *p*-value as the probability of the null hypothesis being true. Regarding the other 11

misconceptions, we highly recommend readers to refer to the original Goodman paper. Interestingly, the first paper pointing out this fallacy appeared as early as 1942 in the statistical literature (4), but obviously did not attract much attention.

To explain this misconception in more detail, we use the analogy between a diagnostic and a statistical test (5). Consider the example of the standard patch test for detection of sensitization to nickel. In Table 2, we give the four possible results for a single patient when this procedure is used. Each patient can be classified by the true disease status as measured by an external gold standard (True nickel sensitization? Yes or No) and by the result of the diagnostic test (Positive patch test result? Yes or No). Depending on the true disease status and the test result, a patient can be termed true-positive, false-positive, false-negative, or true-negative.

To judge the quality of the patch test, we have to study more than one patient, and we give the results of such a study (6) in Table 3. To assess the accuracy of the patch test, one can easily calculate the two common measures of sensitivity (the probability that a truly sensitized patient has a positive patch test result) and specificity (the probability that a truly non-sensitized patient has a negative patch test result); these values here amount to 62/122 [51% (95% CI 42–60%)] and 615/665 [92% (95% CI 90–94%)], respectively. We note for further elaboration that 1 – specificity equals the proportion of false-positives (here: 50/665).

It is often overlooked that, in routine diagnostic testing, sensitivity and specificity are of limited value because they use the fact that the true disease status is known. However, in routine diagnostic testing, the true disease status is not known before the gold standard diagnostic test has been applied. Actually, we apply the diagnostic test because we want to learn about the disease status.

Table 1. Question 6 from the survey of Windish et al. (1) on the *p*-value

In a placebo-controlled trial of the use of aspirin and dipyridamole to prevent arterial restenosis after coronary angioplasty, 38% of patients receiving the treatment had restenosis, and 39% of patients receiving placebo had restenosis. In reporting this finding, the authors stated that $p > 0.05$. This means

- the chances are greater than 1 in 20 that a difference would be found again if the study were repeated.
- the probability is less than 1 in 20 that a difference this large could occur by chance alone.
- the probability is greater than 1 in 20 that a difference this large could occur by chance alone.
- the chance is 95% that the study is correct.

The authors claimed answer (c) to be the correct one, but the correct answer would be rather 'the probability is greater than 1 in 20 that a difference this large or larger could occur by chance alone'.

Table 2. The four possible results in a diagnostic study as exemplified for the patch test for detection of nickel sensitization for a single patient

		True nickel sensitization?	
		Yes	No
Patch test positive?	Yes	True-positive	False-positive
	No	False-negative	True-negative

Table 3. Results from a study on the diagnostic accuracy of the patch test for detection of nickel sensitization as reported by Dickel et al. (6)

		True nickel sensitization?		Σ
		Yes	No	
Patch test positive?	Yes	62	50	112
	No	60	615	675
	Σ	112	675	787

What we really want to learn about is the probability that a patient with a positive patch test result is truly sensitized to nickel. This probability is known as the positive predictive value (the negative predictive value is analogously defined for non-sensitizations and negative patch test results), and can be calculated by the Bayesian rule, which translates a prior probability into a posterior probability by including the study data (7). To apply this rule, we also need information on the true prevalence of the disease, here the percentage of nickel-sensitized patients, to calculate the positive predictive value. In the study of Dickel et al. (6), we used the observed prevalence in the study population of 14% (112/787) to compute a positive predictive value of 55% (95% CI 46–64%). That is, a single patient with a positive patch test result from this study population has a probability of 55% of being truly sensitized to nickel. This calculation depends crucially on the prevalence of nickel sensitization, and is expected to vary by population characteristics; for example, the value would be smaller in a population with a lower prevalence of sensitization. This dependency of the positive predictive value on disease prevalence has been studied more deeply by Diepgen and Coenraads (8) in this journal, and we refer to their excellent article for additional insights.

We now turn to statistical tests. Here, we consider the example of the association between hand eczema and having an office job (9). In their study, Apfelbacher et al. found an unadjusted odds ratio of 0.50 (95% CI 0.27–0.91) with a p -value of 0.02 of the corresponding chi-square test. In perfect analogy to a diagnostic test for a single patient, the result of a statistical test in a single study can be given in a four-fold table (Table 4). The p -value is defined as the probability of observing the study results or even more extreme study results under the condition that the null hypothesis is true. If that probability is lower than the prespecified type I error (alpha error), the null hypothesis is rejected. Note that this is not what we actually want to know from a study. We would wish to have a probability statement about the null hypothesis, but what we get from the p -value is a probability statement about the result of a statistical procedure under the condition that the null

Table 4. The four possible results of a research study as exemplified by the association between having an office job and suffering from hand eczema

		True association between having an office job and suffering from hand eczema?	
		Yes	No
Statistical test rejects null hypothesis of no association?	Yes	True-positive	False-positive
	No	False-negative	True-negative

hypothesis is true. Again by analogy with diagnostic tests, the p -value can be considered as the proportion of false-positives or $1 - \text{specificity}$. And just as we actually need the predictive value in routine diagnostic testing, in statistical testing we would wish to have a statement about the probability of the null hypothesis being true given the result of the statistical test. A solution is, of course, again the application of the Bayesian rule to calculate the requested posterior probability of the null hypothesis being true. But just as the predictive value in diagnostic testing depends on the prevalence, this so-called posterior probability in statistical testing depends on the *a priori* probability of the null hypothesis being true. In our example, this prior probability is the probability that there is no association between having an office job and suffering from hand eczema, and it is probably difficult to determine this prior probability for this situation. Perhaps it is most convenient to assume that the null and alternative hypotheses are equally probable, that is $p_{\text{prior}} = 0.5$. With the use of Jeffreys' non-informative prior, which additionally introduces a small variance (0.125) around the prior probability of 0.5, the posterior probability (i.e. the probability that the null hypothesis is true) in the Apfelbacher data becomes $p_{\text{post}} = 0.18$. With two other approaches, of Goodman (minimum Bayes factor (10)) and Sellke et al. (11), which have been proposed as general rules for calculating posterior probabilities for null hypotheses, we achieve $p_{\text{post, Goodman}} = 0.07$ and $p_{\text{post, Sellke}} = 0.19$. Regardless of which approach we use, the posterior probability for the null hypothesis is overestimated if we misinterpret the p -value as a posterior probability, in the Sellke case by nearly 10-fold.

What do statisticians and epidemiologists wish the medical researcher to pick up from this situation? First, of course, nobody should interpret the p -value as a probability of the null hypothesis being true, as this would be a clear misinterpretation. Second, we would like to stress the dependency of this posterior probability on the *a priori* probability. Just as every clinician would

interpret the result of a diagnostic procedure in the context of the history of the patient (e.g. by asking whether it is probable that a patient with a positive patch test result is truly sensitized to nickel when the patient has never had any skin problems before), we would like the clinician to interpret the result of a statistical test against the background of the prior probability of the null hypothesis being true. Unexpected research findings (i.e. those with a high probability of the null hypothesis being true) need much more supporting evidence before they can be accepted by the scientific community. The permanent ignorance of these ideas led Ioannidis to the provocative assertion that 'most published research findings are false' (12).

In the future, there might be some deeper changes to the statistical paradigms in medical research. Some would make the case for completely switching to Bayesian statistics, whereby we would be able to use prior information, give prior probabilities for all of our model parameters, and then achieve posterior probabilities for all of our hypotheses. In this world, there would be no more *p*-values. The first steps in this direction have been taken; for example, the US Food and Drug Administration issued an explicit

guideline how to use Bayesian statistics in medical device trials (13), and Ioannidis et al. (14), as the new editors of the *European Journal of Clinical Investigation*, stated that 'statistical significance will not be a criterion for selecting papers for publication and we encourage authors to adopt Bayesian approaches, whenever possible and deemed appropriate and interpret judiciously their results in the context of prior evidence'.

In the meantime, we encourage researchers to use CIs to report the statistical variability of findings. A CI conveys information about both the magnitude of the effect and precision, and therefore is highly preferable to a single *p*-value (15). Only recently, we proposed CIs for measures that were not previously considered to be essential in patch testing (16, 17).

Oliver Kuss¹ and Andreas Stang²

¹Institut für Medizinische Epidemiologie, Biometrie und Informatik, Universitätsklinikum Halle und Medizinische Fakultät, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany and ²Institut für Klinische Epidemiologie, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany

References

- Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA* 2007; **298**: 1010–1022.
- Darmon M, Tabah A. Basic concepts of biostatistics are poorly understood by intensivists. *Intensive Care Med* 2009; **35**: 24.
- Goodman S. A dirty dozen: twelve *p*-value misconceptions. *Semin Hematol* 2008; **45**: 135–140.
- Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc* 1942; **37**: 325–335.
- Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987; **257**: 2459–2463.
- Dickel H, Kreft B, Kuss O et al. Increased sensitivity of patch testing by standardized tape stripping beforehand: a multicentre diagnostic accuracy study. *Contact Dermatitis* 2010; **62**: 294–302.
- Uter W, Schnuch A, Gefeller O. Guidelines for the descriptive presentation and statistical analysis of contact allergy data. *Contact Dermatitis* 2004; **51**: 47–56.
- Diepgen T L, Coenraads P J. Sensitivity, specificity and positive predictive value of patch testing: the more you test, the more you get? ESCD Working Party on Epidemiology. *Contact Dermatitis* 2000; **42**: 315–317.
- Apfelbacher C J, Funke U, Radulescu M, Diepgen T L. Determinants of current hand eczema: results from case-control studies nested in the PACO follow-up study (PACO II). *Contact Dermatitis* 2010; **62**: 363–370.
- Goodman S N. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999; **130**: 1005–1013.
- Sellke T, Bayarri M J, Berger J O. Calibration of *p* values for testing precise null hypotheses. *Am Statist* 2001; **55**: 62–71.
- Ioannidis J P. Why most published research findings are false. *PLoS Med* 2005; **2**: e124.
- FDA USFDA. Guidance for the use of bayesian statistics in medical device clinical trials, 2010.
- Ioannidis J P, Tatsioni A, Karassa F B. A vision for the European Journal of Clinical Investigation: note from the new editors. *Eur J Clin Invest* 2010; **40**: 1–3.
- Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol* 2010; **25**: 225–230.
- Kuss O, Dickel H. A confidence interval for the reaction index. *Contact Dermatitis* 2010; **62**: 252–253.
- Kuss O, Dickel H. A confidence interval for the difference between two reaction indices. *Contact Dermatitis* 2011; **64**: 164–165.