Paediatric Respiratory Reviews 25 (2018) 88–93

Contents lists available at ScienceDirect

Paediatric Respiratory Reviews

Review

Lies, damned lies and statistics: Clinical importance versus statistical significance in research

Craig Mellis

Emeritus Professor of Medicine, Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia

EDUCATIONAL AIMS

The reader will come to appreciate:

- How to better interpret commonly used statistics when reading the clinical research literature.
- How to avoid the risk of being misled by poorly performed biostatistics and badly misinterpreted statistics.
- The many ways in which statistics, particularly p values, can mislead.

ARTICLE INFO

Keywords:
P values
Statistical significance
Clinical relevance
95% Confidence Intervals (95% CI)
Effect size (Absolute vs Ratios)
Bayes theory

ABSTRACT

Correctly performed and interpreted statistics play a crucial role for both those who 'produce' clinical research, and for those who 'consume' this research. Unfortunately, however, there are many misunderstandings and misinterpretations of statistics by both groups. In particular, there is a widespread lack of appreciation for the severe limitations with p values. This is a particular problem with small sample sizes and low event rates - common features of many published clinical trials. These issues have resulted in increasing numbers of false positive clinical trials (false 'discoveries'), and the well-publicised inability to replicate many of the findings. While chance clearly plays a role in these errors, many more are due to either poorly performed or badly misinterpreted statistics. Consequently, it is essential that whenever p values appear, these need be accompanied by both 95% confidence limits and effect sizes. These will enable readers to immediately assess the plausible range of results, and whether or not the effect is clinically meaningful.

INTRODUCTION

For clinicians, statistics are simply the terms, or numbers, utilised by authors to summarise the results of clinical research studies. And hopefully, these summary statistics were calculated and interpreted with the assistance of an expert biostatistician. The focus of this paper will be on the correct interpretation of commonly used statistics – and particularly, what to be wary of. As clinicians, what we need are clear, easily interpreted statistics – numbers that convey a meaning that cannot be misunderstood. We want these summary statistics to give us information on a number of key issues in clinical trials, the majority of which will be the comparison between two interventions, preferably randomised.

Essentially, what we need is the following: Firstly, the 'statistical significance' of the results – so we can get a sense as to whether these results are likely to be due to chance – or not. Secondly, we need an appropriate measure of the effect size (magnitude of the difference), to enable us to decide whether this effect size is clinically relevant – or not. Thirdly, we need the 'margin of error', or confidence interval, around the best estimate (ie, the mean difference). Lastly, we want information on the underlying 'power' of the study to detect a meaningful difference [1–4].

WHAT DO CLINICIANS NEED TO KNOW ABOUT STATISTICS?

This is a frequently asked question, and like most things, it all depends. If you are a clinician researcher ('generator' of evidence)

E-mail address: craig.mellis@sydney.edu.au.

http://dx.doi.org/10.1016/j.prrv.2017.02.002
1526-0542/

the best advice is work closely with an expert biostatistician. And you need to be working together right from the beginning of your projects. *Rule number one*: Don't wait until you have completed your data collection, and then consult a statistician to analyse your data!

If you are a 'consumer' of evidence, you must be able to interpret commonly used statistics in clinical research publications. These include: p values, the many different measures of effect size (e.g. absolute risk difference; relative risk), 95% confidence intervals (CI 95%) and limitations of statistical testing.

While incorrectly analysed and misinterpreted statistics have resulted in high rates of false positive conclusions, the other major source is the risk of bias in the research methodology and/or execution of the study. Risk of bias in published papers is detected by critical appraisal, utilising validity checklists. This topic will not be covered here, but it is strongly recommend that the reader consult one of the many outstanding evidence based medicine texts [5,6].

DEFINING THE UBIQUITOUS P VALUE

It is important to recall the technical definition of a p value: Namely, the probability of an observed result (or more extreme result), *given the assumption that the null hypothesis is true* [7,8]. The difficulty lies is translating this definition into something that can be easily understood [9,10]. Arbitrarily, the threshold for accepting or rejecting the null hypothesis is when there is a less than one in 20 ($p < 0.05$) chance that the extreme result observed (or more extreme results) – would occur, *under the assumption there is no difference*. Of crucial importance is the p value you calculate (or read in the literature), which refers only to *your* (or their) specific sample – and on its own, that p value is of strictly limited value [3]. The observed p value is not necessarily 'the truth', and does not necessarily reflect the true value for the defined 'population' of, for example, all children with asthma. Though, of course, we all live in hope that our own clinical trial observations are just that!

WHY ARE P VALUES UNDER ATTACK?

Even statisticians agree that making inferences about p values is 'risky business' [11]. An excellent summary of some of the problems with the interpretation of p values is the paper appropriately entitled: "A dirty dozen: Twelve p value misconceptions" [12]. You may be aware that sadly, a substantial proportion of the published clinical research cannot be replicated [1]. Indeed, it has been suggested that up to 50% of the published research is simply incorrect! [3,13]. Unsurprisingly, errors strongly favour false positives! For example, the ratio of False Positive to False Negative publications in epidemiologic studies, especially genetic epidemiology, is reported to be as high has 100:1 [14,15].

At least some of the blame for this inability to replicate results has been attributed to misunderstandings about the ever present 'p value'. This is not a new problem, and indeed the lay press is well aware of the issue, with newspaper headlines such as: "p-value misuse running rampant" [16,17]. Erroneous positive conclusions are particularly likely with the initial publication of a new intervention – and because of the novelty value ('newsworthiness'), it is very likely to be published in a high impact journal and receive considerable press coverage [18,5].

WHY THE HIGH FALSE POSITIVE ERROR RATE?

Obviously, there are many possible explanations for the high rate of non-replicable, erroneous research findings [19,20]. Common examples include: chance, biased methodology, biased study execution, biased reporting, small sample sizes, low event rates, mis-use and/or mis-interpretation of p-values, data-mining for the elusive $p < 0.05$ value, publication bias, selective reporting of p values, the ever present ethos of 'publish or perish', together with our human failing – the desire to find support (any support!) for our pet hypotheses, and, unfortunately, fraud.

IS THERE A NEW UNDERSTANDING OF TYPE I (FALSE POSITIVE) ERRORS?

Most clinicians have been taught that when p equals 0.05 (or <0.05), chance is unlikely to explain the extreme result (given the *assumption of 'no difference'*). Consequently, we reject the null hypothesis in favour of the alternative hypothesis (ie, the intervention is better than control). Further, we presume that our risk of making a false positive conclusion is 5% (ie, a 1 in 20 risk). That is, the so-called 'Type I error' – or risk of a false positive conclusion [7]. However, expert statisticians totally disagree with that interpretation, as outlined in Table 1.To quote Colquhoun [3]: "If you use p = 0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time [Not 5% !]. And if, as is often the case, experiments are underpowered, you will be wrong most of the time!" [3,13].

ARE THERE OTHER PROBLEMS WITH P VALUES?

It is important to be aware of how widely p values fluctuate, from study to study, despite what appears to be similar experimental conditions. Wide fluctuations are particularly likely when the sample size is small, with consequent low numbers of events (such as asthma exacerbations). This is unfortunately a common feature of most clinical trials [19].

Apart from lack of reproducibility, there are many other limitations with p values. They are an oversimplification, giving a black and white, 'yes or no' answer to the question posed by the clinical trial. It clearly illogical to make a different clinical conclusion between a trial with a result 'insignificant' because p = 0.051, from another we consider 'significant' because p = 0.049. Moreover, p values give no indication of the size of the difference between the two treatments. Thus, a tiny, clinically irrelevant difference could be statistically significantly different if the sample size is very large. Additionally, p values give no indication of the 'margin of error', nor any real information regarding the power of the study.

ARE P VALUES BEING USED MORE FREQUENTLY?

Despite their bad press, a recent survey of a large number of Medline abstracts and articles, found that p values are appearing

Table 1
Variable interpretation of Type I error.*

P value	A. What clinicians believe is risk of FALSE POSITIVE (FP) or Type I error rate	B. What statisticians calculate as FALSE POSITIVE (FP) or type I error rate	C. If small sample size (low event rate), FALSE POSITIVE (FP) or Type I error rate
=0.05	5%	23%	50%
=0.01	1%	7%	15%

* Derived from references [3,13].

with greater frequency over time (1990–2015). And of greater concern, almost all p values reported significant results – while few included confidence intervals or effect sizes [21,22]. Interestingly, at least one journal has addressed the problem by totally banning p values – as well as confidence intervals, and the words 'significant' and 'insignificant'! [23].

So, given that p values are so prone to abuse, misinterpretation and error, what are the take-home messages? Firstly, it is wise to maintain a healthy level of scepticism when you see the phrase "$p < 0.05$, statistically significant." Lack of reproducibility and high rates of false discovery have made the interpretation of hypothesis testing and p values a major concern [2,9,13,24]. The simple message appears to be:wait for replication! Clearly, using p values *alone* is unsatisfactory, and reporting effect size and Confidence Intervals (CI 95%) overcomes at least some of the drawbacks with p values [25,26].

EFFECT SIZE – WHICH ONE IS BEST?

While there are many ways of describing the outcome measures in a clinical trial, these can be broadly divided into 'patient important' outcome measures and 'surrogate' outcome measures [27]. As a general rule, surrogate outcome measures are numerical data, such as FEV_1, total IgE and serum bicarbonate. Numerical results will be expressed as a mean difference, or standardised mean difference (SMD). Surrogate outcomes provide research efficiency as they are easy to measure, objective, and frequent – compared to patient important outcomes like hospitalisation or death.

Clinicians need to take great care with surrogate outcome measures as these do not necessarily result in a patient important outcome. There are many examples in the literature where a surrogate outcome measure has gone in one direction, while the patient important outcome has gone in the opposite direction. A classic example is the use of an experimental agent in osteoporosis patients that improved measured bone mineral density (a surrogate outcome measure), but an increased risk of fractures (a patient important outcome measure). Presumably the agent resulted in more solid but more fragile bones [28]. The advice is clear: avoid changing your clinical practice on the basis of a surrogate outcome measure, such as FEV_1. It is best to wait until the intervention is proven to result in an improvement in an outcome that is valued by the patient.

Patient important outcome measures are usually "yes/no" answers. That is, binary data, usually expressed as the percentage that had the outcome of interest/event. For example, in an asthma study, the difference in the rate of asthma exacerbations, hospitalisations, or deaths would be compared between two interventions. This data will normally be presented as relative risk or risk ratio (RR), and Relative Risk Reduction (RRR).

SHOULD WE USE ABSOLUTE EFFECT SIZE OR RATIO – OR BOTH?

Clinicians and patients (parents) find it far easier to understand absolute values (eg Number needed to treat: NNT), while ratios can be confusing and misleading. Moreover, RRR invariably suggests an exaggerated effect size - the statistic most frequently used in pharmaceutical company advertisements [29,30] [Table 2]. So, why do we bother working out RR and RRR, and why are these statistics traditionally used in results? The reason is that these ratios are consistent across different risk groups. Thus, ratios can be applied to patients with different baseline risks of the outcome to those patients in the clinical trial. For example it may be quite obvious that your patient with asthma is more severe, and at a substantially higher risk (eg double the 10% control group risk, or 20%) of an exacerbation compared to the average patient in the randomised trial. If so, simply apply the relative risk and RRR from the published trial to your individual patient. Similarly, if it is clear that your patient is less severe, and at a lower risk (eg, half the 10% risk, or 5%) of an exacerbation than the patients in the clinical trial, you can use the RR and RRR, and apply it to your individual patient [31] [Table 3].

Table 2
Doing the simple maths for effect size.

Using a simple example, consider a randomised trial of 200 children with asthma, comparing inhaled corticosteroids (ICS) with placebo. Primary outcome measure; asthma exacerbations needing oral corticosteroids. Study duration of 12 months.

Results: Eight of the 100 children allocated to ICS had exacerbations (ie, risk of exacerbation = 8%), while 10 of the 100 allocated to placebo had exacerbations (ie, risk of exacerbation = 10%).

The simple maths for determining the absolute results and ratios are as follows:

Effect size as Absolutes:	
Control (placebo) Event Rate	=10%
Experimental (ICS) Event Rate	=8%
Absolute Risk Difference [ARD]	=10% minus 8% = 2%
Number Needed to Treat [NNT]	=1/ARD = 1/2% = 100/2 = 50
Effect size as Ratios:	
Relative Risk [Risk Ratio or RR]	=8%/10% = 0.8
Relative Risk Reduction [RRR]	=1.0 minus RR =1.0 − 0.8 = 0.2 = 20%

Table 3
Why we need RR & RRR.

From data in Table 2, RCT placebo vs ICS, we can estimate the risk of an acute exacerbation of asthma:

Placebo Group [CER]	=10%
ICS group [EER]	=8%
ARD	=2%; NNT = 50; RR = 0.80; RRR = 20%;

But it may be clear your individual patient is at HIGHER risk of an exacerbation than the controls in the RCT.

For example, if we assume your patient has double the 10% baseline risk of controls- your patient's baseline risk of an exacerbation is 20%

Therefore:
Since RR [0.8] and RRR [20%] are consistent across risk groups,
And since CER = 20%, so, now EER = 16% [ie, 20% × 0.8 = 16%]

However, absolute values will vary as follows:
Now, ARD = 4%; NNT = 25 [ie, NNT = 1/ARD = 1/4%, or 100/4 = 25]

The clear message demonstrated above is that as the baseline risk increases, ratios remain constant – but absolute results change. Specifically, the higher the baseline risk, the lower NNT (and vice versa).

WHEN IT IS APPROPRIATE TO USE A HAZARD RATIO?

A statistical term you will sometimes see in place of relative risk is a hazard ratio [HR]. A HR is derived from a survival analysis of a clinical trial, and the hazard ratio is simply the relative risk, averaged over the duration of the trial [32]. It is interpreted in exactly the same way as a relative risk, the further the HR he is away from 1.0, the greater the effect size. If the 95% CI around the hazard ratio crosses 1.0 (ie, the line of no effect) then the result is consistent with no difference (null hypothesis). For example, if the HR = 0.76 and CI 95 = 0.64 to 1.13, then the $p > 0.05$ is not significant. If a HR is quoted, expect to see survival curves, comparing the 'survival' (or time to event) in the experimental intervention to the control.

WHEN TO USE ODDS RATIOS?

While RR and HR are the appropriate effect size measures in randomised controlled trials and cohorts studies, in case-control

studies the true risk (or incidence) is unknown [33]. Therefore, we cannot calculate a true relative risk. Instead, we calculate the odds ratio, a rough approximation of the RR. The Odds Ratio is the odds of the relevant exposure (intervention) in the cases, compared to the odds of the exposure in the controls.

WHAT DO 95% CONFIDENCE INTERVALS (CI 95%) ADD?

Clinicians traditionally use a pragmatic (though not strictly true) interpretation of CI 95%: Namely, "we are 95% 'confident' the true result will be somewhere between the CI 95% limits" [34]. However, the individual study CI 95% relates specifically to that observation – and assumes the study was correctly performed – ie, adequately powered, numerous events, no confounding influences, no risk of bias and was appropriately analysed. Clearly, these are not common features of clinical trials! So, a qualified (but probably unrealistic) definition is more accurate: "in *correctly* performed studies we would *expect* the CI 95% to include the true value 95 percent of the time." [35,36,37].

Confidence Intervals address a number of the deficiencies with p-values. In particular, 95% confidence intervals give us a plausible or likely range of the effect size. In lay terms, CI 95% is the 'the margin of error', around our best estimate (ie, the mean or point estimate). Confidence intervals also supply valuable information regarding the power of the study; the narrower the width of the confidence interval, the better – ie, the more precise the results. While study power depends upon sample size, more important are the number of events observed.

A direct estimate of the p value can be derived from the confidence intervals. For example, if in a randomised trial (Treatment A vs Treatment B), the effect measure is Relative Risk (RR) of hospitalisation, and our observed 95% confidence interval includes a Relative Risk on 1.0, for example: RR = 0.97, CI 95% = 0.78 to 1.22, we will interpret this result as not statistically significant ($p > 0.05$, NS) – so, we accept the null hypothesis. However, this result leaves clinicians with uncertainty, because at one extreme of the CI 95%, treatment A results in a *reduction* in hospitalisation (lower boundary of the CI 95% = 0.78; or a Relative Risk Reduction of 22%); but at the other extreme of the CI 95%, treatment A results in an *increase* in the risk of hospitalisation (upper boundary of the CI 95% = 1.22; or a Relative Risk Increase of 22%). Consequentially, when the CI 95% interval offers a different clinical decision at the extremes of the CI 95%, we are left with uncertainty re clinical decision making – in short, the study is underpowered!

If the primary outcome measure is numerical, for example the difference in FEV_1 between two randomised treatments, and if the 95% confidence interval includes zero, then again, we conclude the result is not statistically significant (eg, Mean FEV_1 difference = +7%, CI 95% minus 3% to plus 17%; $p > 0.05$. NS). Again, at the extremes of the CI 95% a totally different decision is offered. Despite the valuable additional information with CI 95%, interpretation of CI 95% still relies on the troublesome, close relationship with p values, and the dichotomous decisions re 'significant' vs 'non-significant' [38].

STATISTICAL ISSUES WITH DIFFERENT STUDY TYPES

Randomised controlled trials

The major statistical issues are the problems with interpretation of the p value, as outlined above. In particular, one should be especially wary of the initial study of a new intervention [39]. Erroneous conclusions are particularly prominent with the initial publication of randomised trials of a novel intervention. You'll recognise these articles: Always strongly positive, often with implausibly large effect sizes, published in a high impact medical journal, and given extensive media coverage. Unfortunately, these results are also likely to be subsequently proven to be either incorrect, or at best, to have a highly inflated effect size. This is what has been termed "regression to the truth"! [40]. Because of the well-known high risk of error in the initial study, it is wise to wait for replication – in a separate setting, and ideally in a study that has a total of at least 300 (patient important) events [41].

Systematic reviews/meta-analyses

The two key statistical issues are related to heterogeneity (inconsistency) and publication bias.

Heterogeneity

Statistical heterogeneity is currently best expressed as the I^2 statistic. This statistic gives a numerical score of the extent of variation in results between studies, expressed as a percentage from 0 to 100% [42]. The lower the percentage the better, ideally 0% – indicates no heterogeneity (other than by chance alone). An $I^2 > 75\%$ indicates a large amount of heterogeneity, and your level of confidence in the results will be necessarily lowered. If I^2 approaches 100%, the heterogeneity is so great that statistical pooling is probably not warranted. Readers can get a quick qualitative estimate of this variation by simply eyeballing the forest plot – specifically looking to see whether or not the 95% confidence intervals for the multiple trials are overlapping. If all the CI 95%s overlap, the variation between studies should not be of concern, and this will be reflected in a low I^2.

Publication bias

Non-publication of negative studies is a major, predictable risk in every systematic review, resulting in an inflated pooled effect size [43]. This bias is best detected by carefully reading the methodology to ensure the authors have done a comprehensive search, ideally including contacting authors in the field, hand searching, and checking out the 'grey literature'. Visual inspection of the degree of symmetry of the 'funnel plot' will give an indication of whether there are obvious 'missing' negative trials. However, unless there are a reasonable number of trials, it is not possible to generate a meaningful funnel plot [44]. Various statistical tests have also been developed to evaluate the risk of publication bias, but these will not be covered here [45].

The other statistical issue surrounds the method of combining data from the individual studies (ie, the 'meta-analysis'). Two methods are available: the fixed effects model, and the random effects model [46]. There is considerable controversy about which method is best, and in which situations each should be performed. Currently, the research world remains divided and some authors will report both analyses to give readers their choice.

Cohort studies

Although observational studies are clearly less reliable for clinical decision-making, in the absence of high quality studies (randomised controlled trials) these may represent the best available evidence. The key statistical issue is the inevitable difference between those exposed and non-exposed to the exposure of interest [47].

For example, we could consider a cohort study to test whether exposure to inhaled corticosteroids [ICS] could increase the risk of Community Acquired Pneumonia [CAP]. In this setting, there will always be additional risk factors, for example age, gender, socio-economic status, exposure to environmental tobacco smoke, co-morbidities, and disease severity, which will *not* be evenly balanced between exposed and non-exposed groups. These unevenly distributed risk factors ('confounders') are associated

with the outcome of interest (e.g. CAP), and will interfere with our ability to determine the true association between the exposure (ICS) and the outcome of interest (CAP). Consequently, these additional potential risk factors must be identified, accurately measured, and statistically "controlled for" (ie, adjusted). Obviously, this statistical adjustment is not as effective as randomisation, and unknown/un-measured risk factors (eg, genetic risk factors) cannot be controlled.

This means when you read the results of a cohort study, the focus must be on the adjusted Relative Risk [aRR], and not the 'raw' (unadjusted) RR. Because of the risks of bias, your level of confidence in the results from a cohort study will necessarily be less than that of a well-conducted RCT.

Case-control studies

These are inherently biased studies, and not generally useful for clinical decision-making [33]. However, they are important for hypothesis generation, before more definitive and expensive studies are justified. Moreover, for rare diseases, or harmful exposures, case control studies usually represent the only available evidence.

The major problem with these studies is in the methodology rather than with the statistical analysis. Obtaining accurate exposure status in both the cases and the controls is prone to error and is usually biased. In addition, there are major issues with accurate measurement of additional risk factors (confounders), preventing effective statistical control (adjustment) for the many predictable differences between cases and controls. Consequently, these studies have a history of being misleading.

DIRECTIONS FOR FUTURE RESEARCH

It is clear that there are some major statistical problems with hypothesis testing, and in particular, p values. While 95% confidence intervals [CI 95%] around effects size [eg, Relative Risk] supplies important addition information, CI 95% and p values are closely linked, and suffer from similar problems relating to the interpretation of whether they are 'significant or not significant'.

An approach, which is not new, but is now more feasible with the increasing sophistication of computing, is "Bayesian statistics" [17,48]. The beauty of Bayes' theory is that it makes use of all pre-existing knowledge/data concerning the hypothesis being tested. The downside however is that Bayes theory assumes there will be pre-existing data to enable investigators to factor in an estimate of the probability of the hypothesis being correct vs. incorrect.

In the meantime, the researcher should address measures to improve the accuracy of clinical research, as summarised in Table 4. There is no doubt that p values will continue to be utilised to estimate the role of chance. There is nothing inherently wrong with p values – it is their misuse and mis-interpretation that has given them their notoriety [49].

Table 4
Possible measures to reduce erroneous research.*

Research culture/research methods:
Registration of study protocols
Large-scale collaborative research
Improve quality of study design/methodology
Standardize outcome measures
Improve peer review, reporting, and dissemination of research
Educate scientific workforce in methods and statistical literacy
Statistical issues:
Ensure expert statistical support
More appropriate statistical methods
Increase thresholds for claiming discoveries or "successes"
Conflict of Interest ("Spin")
Containment of conflicted sponsors and authors

* Modified from Ionnadis et al. [20].

EDUCATIONAL ARTICLE

You can receive 1 CME credit by successfully answering these questions online.

(A) Visit the journal CME site at http://www.prrjournal.com.
(B) Complete the answers online, and receive your final score upon completion of the test.
(C) Should you successfully complete the test, you may download your accreditation certificate (subject to an administrative charge), accredited by the European Board for Accreditation in Pneumology.

References

[1] Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Medicine* 2005;**2**. http://dx.doi.org/10.1371/journal.pmed.0020124. 696–701 [e124].
[2] Nuzzo R. *P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature* 2014;**506**:150–2.
[3] Colquhoun D. An investigation of the false discovery rate, and the misinterpretation of p-values. *R Soc Open Sci* 2014. http://dx.doi.org/10.1098/rsos.140216
[4] Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 2016;**70**(2):129–33. http://dx.doi.org/10.1080/00031305.2016.1154108
[5] Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: Manual for Evidence-Based Clinical Practice, Third Edition, New York: McGraw-Hill; 2015.
[6] Strauss S, Glasziou P, Richardson WS, Haynes RB. Evidence Based Medicine: How to practice and teach it, Fourth Edition, Edinburgh: Churchill Livingston; 2011.
[7] Armitage P, Berry G. Edition 2. In: Statistical Methods in Medical Research. Oxford: Blackwell Scientific Publications; 1987. Confidence Intervals 95%, p. 101–104.
[8] Altman D. Practical Statistics for Medical Research. London: Chapman & Hall; 1991 , Interpretation of p values, p. 167–171 Relation between confidence intervals and statistical significance, p. 175.
[9] 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 misinterpretations. *Eur J Epidemiol* 2016;**31**:337–50.
[10] Goodman SN. Toward evidence-based medical statistics. Part 1: The P value fallacy. *Ann Intern Med* 1999;**130**:995–1004.
[11] Aschwanden C. Not Even Scientists Can Easily Explain P-values; 2015 website: http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/.
[12] Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. *Semin Hematol* 2008;**45**:135–40.
[13] Sellke T, Bayarri MJ, Berger JO. Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician* 2001;**55**:62–71.
[14] Ioannidis JPA, Tarone R, McLaughlin JK. The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology* 2011;**22**:450–6.
[15] Ioannidis JPA, Ntzani EE, Trikalinos TA, Despina G, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nature Genetics* 2001;**29**:306–9.
[16] Ross J. p-value misuse running rampant. The Australian Newspaper, Mar 30, 2016.
[17] Ross J. p-value misuse running rampant. The Australian Newspaper, Mar 30, 2016.
[18] Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;**294**:218–28.
[19] Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews. *Neuroscience* 2013;**14**:365–75.
[20] Ioannidis JPA. How to Make More Published Research True. *PLOS Medicine* 2014;**11**:e1001747.
[21] Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting P Values in the Biomedical Literature, 1990–2015. *JAMA* 2016;**315**:1141–8.
[22] Kyriacou DN. Editorial: The Enduring Evolution of the P Value. *JAMA* 2016;**315**:1113–5.
[23] Trafimow D, Marks M. Editorial: Announcing ban on p values, & null hypothesis significance testing. *Basic and Applied Social Psychology* 2015;**37**:1–2.
[24] Ioannidis JPA. Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. *JAMA* 1998;**279**:281–6.
[25] Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education* 2012;**4**:279–82.

[26] Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;**292**:746–50.
[27] Fleming TR, DeMets DL. Surrogate End Points in Clinical Trials: Are We Being Misled? *Ann Intern Med* 1996;**125**:605–13.
[28] Riggs BL, Hodgson SF, O'Fallon WM, Chao EYS, Wahner HW, Muhs JM, et al. Effect of fluoride on the fracture rate in postmenopausal women with osteoporosis. *NEJM* 1990;**322**:802–9.
[29] Horton R. Offline: What is medicine's 5 sigma? *Lancet* 2015;**385**:1380.
[30] Mayor S. Researchers claim clinical trials are reported with misleading statistics. *BMJ* 2002;**324**:1353.
[31] Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: Manual for Evidence-Based Clinical Practice, Third Edition, New York: McGraw-Hill; 2015, Applying results to individual patients. p. 235–247.
[32] Barratt A, Wyer PC, Hatala R, McGinn T, Dans AL, Keitz S, et al. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ* 2004;**171**:353–8.
[33] Prasad K, Jaeschke R, Wyer P, Keitz S, Guyatt G. Tips for Teachers of Evidence-Based Medicine: Understanding Odds Ratios and Their Relationship to Risk Ratios. *J Gen Intern Med* 2007;**23**:635–40.
[34] Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Moyer V, et al. Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *CMAJ* 2004;**171**:611–5.
[35] Armitage P, Berry G. Edition 2. In: Statistical Methods in Medical Research. Oxford: Blackwell Scientific Publications; 1987. Confidence Intervals 95%, p. 101–104.
[36] Altman D. Practical Statistics for Medical Research. London: Chapman & Hall; 1991, Relation between confidence intervals and statistical significance, p. 175.
[37] Johnson VE. Revised standards for statistical evidence. *PNAS* 2013;**110**:19313–7.
[38] Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. *Psychon Bull Rev* 2016;**23**:103–23.
[39] Ioannidis JPA. Why Most Discovered True Associations Are Inflated. *Epidemiology* 2008;**19**:640–8.
[40] Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? The contribution of regression to the truth. *J of Cardiac Failure* 2003;**9**:364–7. http://dx.doi.org/10.1054/S1071-9164(03)00018-6
[41] Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: Manual for Evidence-Based Clinical Practice, Third Edition, New York: McGraw-Hill; 2015, Misleading presentations of clinical trial results; p. 259–270.
[42] Hatala R, Keitz S, Wyer P, Guyatt G. Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ* 2005;**172**:661–5.
[43] Moher D, Liberati A, Tetziaff J, Altman D. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 2009;**6**(7):e1000097. http://dx.doi.org/10.1371/journal.pmed.1000097
[44] Sedgwick P. Meta-analyses: how to read a funnel plot. *BMJ* 2013;**346**:f1342. http://dx.doi.org/10.1136/bmj.f1342
[45] Duval S, Tweedie R. Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *BIOMETRIC* 2000;**56**:455–63.
[46] Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-Effects Meta-analysis of Inconsistent Effects: A Time for Change. *Ann Intern Med* 2014;**160**:267–70. http://dx.doi.org/10.7326/M13-2886
[47] Lawson JA, Janssen I, Bruner MW, Hossain A, Pickett W. Asthma incidence and risk factors in a national longitudinal sample of adolescent Canadians: a prospective cohort study. *BMC Pulmonary Medicine* 2014;**14**:51–60. http://dx.doi.org/10.1186/1471-2466-14-51
[48] Goodman SN. Toward evidence based medical statistics 2. The Bayes Factor. *Ann Intern Med* 1999;**130**:1005–13.
[49] Gelman A, Loken E. The Statistical Crisis in Science. *American Scientist* 2014;**102**:460–9, http://www.americanscientist.org/issues/num2/2014/6/the-statistical-crisis-in-science/1.

CME QUESTIONS

Q1:

Which of the following 5 statements is true or false?

The p value is . . .

A The probability that the null hypothesis is true?
B The probability that the alternative hypothesis is true?
C The probability that an initial finding will be replicated?
D Tells us the result is clinically clinically relevant?
E Tells us whether the result can be generalized to other populations?

Q2:

The following are results of five published clinical trials which compared the rate of community acquired pneumonia with inhaled corticosteroids vs control.

Which one of these trial results is not consistent with chance? [CI 95% = 95% Confidence Interval]

A Risk Ratio = 0.91 [CI95%: 0.89, 1.01]
B Odds Ratio = 1.04 [CI95%: 0.97, 1.11]
C Hazard Ratio = 0.83 [I95%: 0.65, 1.19]
D Relative Risk = 1.16 [CI95%: 1.01, 1.23]
E Relative Risk Reduction = 13% [CI95%: minus 6%, 21%]

Q3:

In a large randomised controlled trial, a new pertussis vaccine was compared to the current vaccine. After 12 months of follow-up, the following results were obtained: rate of proven pertussis with the new vaccine was 5%, compared to 8% with the current vaccine.

Which one of the following is the approximate Number Needed to Treat (NNT) with the new vaccine, compared to the current vaccine, to prevent one patient from developing pertussis?

(A) 150
(B) 100
(C) 50
(D) 33
(E) 25