Intensive Care Med (2014) 40:1927–1929
DOI 10.1007/s00134-014-3510-6

Daniele Poole
Giovanni Nattino
Guido Bertolini

Overoptimism in the interpretation of statistics

The ethical role of statistical reviewers in medical journals

Received: 31 August 2014
Accepted: 26 September 2014
Published online: 7 October 2014

D. Poole (✉)
Operative Unit of Anesthesia and Intensive Care, San Martino Hospital, 32100 Belluno, Italy
e-mail: daniele.poole@alice.it

G. Nattino · G. Bertolini
GiViTI Coordinating Center, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Centro di Ricerche Cliniche per le Malattie Rare Aldo e Cele Daccò, 24020 Ranica, Bergamo, Italy

Introduction

A letter to the *BMJ* in 2000 unveiled macroscopic flaws regarding the calculations of means in a manuscript published by the journal [1]. Although such serious mistakes may be the exception, frequently statistical errors have been found in published articles, indicating reviewing system failures [2]. In our experience with intensive care medicine, however, we noticed that editors frequently involve statistical reviewers, taking full account of their revisions and requiring their final evaluation after authors have complied with reviewers' recommendations. This allows for the filtering of poor quality articles with evident mistakes such as applying exclusion criteria after randomisation, running multiple regression analyses on very small samples without accounting for the event-to-variable ratio, performing infinite bivariate comparisons to rule out basic differences between two study groups, and calculating sensitivities without having all patients submitted to the diagnostic test or reporting areas under the receiver operating characteristic curves values less than 0.5.

The high frequency of these errors and the insufficient reporting of statistics and study design make the reviewer assignment unduly difficult. At the same time, statistical revisions seem to be effective in improving the quality of published articles [3].

In some cases, however, although statistical analyses are correct, authors may overemphasize their interpretation, as reported in the examples below.

Diagnostic test interpretation

Aristotle gave a fundamental contribution to deductive reasoning, by introducing the concept of syllogism, his most famous one being "All men are mortal, Socrates is a man, therefore Socrates is mortal". Of course, syllogisms can be wrongly formulated. For example, "Italian presidential guards are tall, professional basketball players are tall, therefore Italian presidential guards are professional basketball players". This conclusion is obviously wrong. Still, in medical literature we often make similar errors. For example, given the high sensitivity (the rate of diseased patients with a positive test) of procalcitonin for infection diagnosis, it is not infrequent to find authors claiming its high predictive ability. In this case the syllogism would be "Infected patients have high procalcitonin, this patient has high procalcitonin, therefore this patient is infected", which is wrong since we are

not dealing with infected patients but with patients with a positive test that we want to correctly diagnose [4]. Suppose we have a 100 % sensitivity and 90 % specificity for procalcitonin, and of 200 admitted patients 20 % are infected yearly during the stay in a specific intensive care unit (ICU). The 40 infected patients will all have procalcitonin over the threshold while 16 of the 160 not infected will be false positives. Thus only 40 of 56 patients (71 %) with a positive test will also be infected, quite a low positive predictive value (PPV) in the face of very high sensitivity and specificity. Clearly, this is because PPV is dependent on disease prevalence (see Fig. 1), and clinicians should account for such dependence when applying the research results to their specific clinical contexts.

Meta-analysis interpretation

Sometimes conclusions of meta-analyses are similarly affected by excessive optimism. The main problem is that studies which are heterogeneous in design and case-mix are often combined to gain the power that single studies lack in order to demonstrate treatment effect. However, on a clinical basis it may be difficult to justify the extension of positive results to all the various categories of patients enrolled in the different negative studies. For example, a positive meta-analysis investigating the effect of prone position in ARDS included 11 trials, four of them involving patients receiving high-frequency oscillatory ventilation, comatose or trauma patients [5]. Could we reasonably extend the overall average result to each of these categories?

Another recent meta-analysis reported a barely non-significant result, but a subgroup analysis led the authors to conclude that pronation "improved survival among patients with ARDS who received protective lung ventilation" [6], maybe an excessively optimistic statement since subgroup analysis is known to generate spurious results, from which only hypotheses can be drawn [7]. This seems to be quite acknowledged for trials but somehow tolerated when dealing with meta-analysis.

Another problem with meta-analysis is the measurement and interpretation of heterogeneity that often leads to unsupported strong statements [8]. Statistical heterogeneity indicates variations across the results of studies and is commonly evaluated with the I^2 statistics that can range between 0 and 100 % [9]. A major drawback of this test is that it is underpowered given the number of studies commonly included in meta-analyses, so that a non-positive test in most cases does not rule out the possibility that heterogeneity does exist [8]. Moreover, the translation of statistical heterogeneity into clinical terms is quite a tricky exercise. For example, in a recent meta-analysis on prone positioning in ARDS, I^2 was reduced from 64 to 11 and 25 % after dividing the studies into two groups on the basis of tidal volumes (high vs

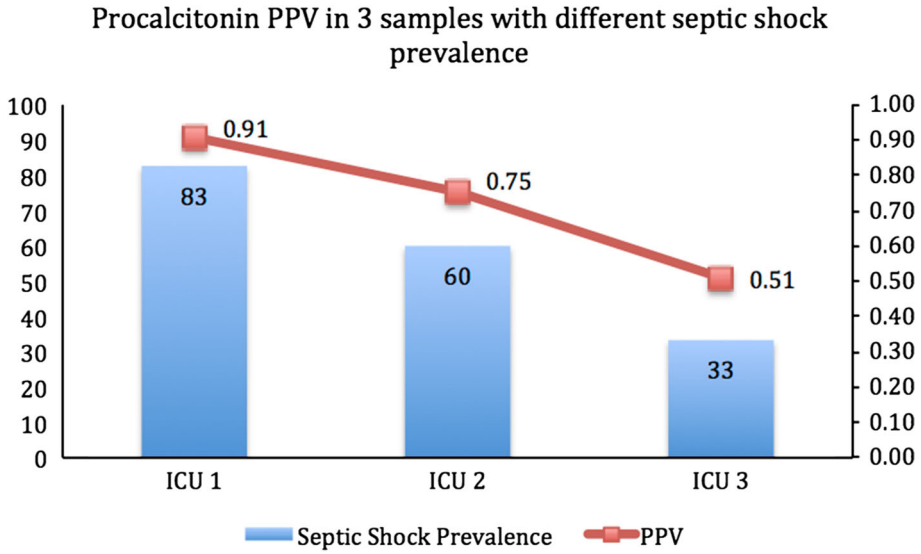

Fig. 1 The *left column* reports data from a study investigating the ability of procalcitonin to discriminate between septic shock and other forms of shock [15]. Out of 75 patients admitted to a single intensive care unit (ICU), 62 (83 %) had septic shock. Using a threshold of 1 ng/ml procalcitonin sensitivity was 0.95 and specificity 0.54. Using the same sensitivity and specificity, we tested the same threshold in two hypothetical ICUs (*2* and *3*, in the chart) having a higher rate of haemorrhagic shock due to trauma, and a lower prevalence of septic shock. The positive predictive value (PPV) of procalcitonin decreased proportionally. This demonstrates how the change of case-mix may impact on the PPV of a diagnostic test and highlights the issue of the generalizability of study results, especially when research is conducted in single centres

low) [10]. The authors concluded, probably with excessive confidence, that heterogeneity across studies was explained by this different ventilatory approach. Well, we meta-analysed the studies included in a 2010 meta-analysis [11] along with the last published trial [12], and then stratified them according to the initials of the last author (from A to I, and K to Z) and found that the I^2 decreased from 45 % to 0 and 12 %. Thus, modification of statistical heterogeneity by stratification can have multiple explanations and the plausibility of the hypothesis does not modify the fact that such analyses should be recognized as exploratory in nature and needing confirmation [8]. Moreover, the uncertainty of the I^2 measure is usually so wide as to provide largely overlapping confidence intervals (in our example ranging between 0 and more than 70 % in all cases), not supporting the existence of true estimate differences.

A similar problem of power is met with most meta-analyses when dealing with publication bias evaluation with formal assessment of asymmetry, generating equivocal interpretations of negative results (is publication bias absent or is the power insufficient?) [13].

The use of fixed or random effect models is another source of confusion. Fixed effect models are applicable under the assumption that all studies share a common effect size, an assumption that rarely holds in real life. However, the use of fixed effects generally reduces the width of the confidence interval range, thus increasing the likelihood of having a statistically significant result.

Conclusion

Interpretation of statistical analysis is a slippery ground for authors, who, in total good faith, may tend to overemphasize the results. The consequence may be the slavish translation into clinical practice of treatments for which benefits and risks have not been fully verified. Under this perspective statistical reviewers have an ethical role: they should not focus only on analyses but also closely evaluate results interpretation of submitted scientific manuscripts [14].

Conflicts of interest The authors declare they have no conflict of interest. DP has been part of the statistical board of *Intensive Care Medicine* since 2009. GB was a member of the editorial board of *Intensive Care Medicine* for statistical review from 2009 to 2012.

References

1. Bland M (2000) Fatigue and psychological distress. Statistics are improbable. BMJ 320:515–516
2. Altman DG (2002) Poor-quality medical research: what can journals do? J Am Med Assoc 287:2765–2767
3. Altman DG (1998) Statistical reviewing for medical journals. Stat Med 17:2661–2674
4. Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. BMJ 309:102
5. Lee JM, Bae W, Lee YJ, Cho YJ (2014) The efficacy and safety of prone positional ventilation in acute respiratory distress syndrome: updated study-level meta-analysis of 11 randomized controlled trials. Crit Care Med 42:1252–1262
6. Sud S, Friedrich JO, Adhikari NK, Taccone P, Mancebo J, Polli F, Latini R, Pesenti A, Curley MA, Fernandez R, Chan MC, Beuret P, Voggenreiter G, Sud M, Tognoni G, Gattinoni L, Guerin C (2014) Effect of prone positioning during mechanical ventilation on mortality among patients with acute respiratory distress syndrome: a systematic review and meta-analysis. CMAJ 186:E381–390
7. Assmann SF, Pocock SJ, Enos LE, Kasten LE (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355:1064–1069
8. Ioannidis JP (2008) Interpretation of tests of heterogeneity and bias in meta-analysis. J Eval Clin Pract 14:951–957
9. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560
10. Beitler JR, Shaefi S, Montesi SB, Devlin A, Loring SH, Talmor D, Malhotra A (2014) Prone positioning reduces mortality from acute respiratory distress syndrome in the low tidal volume era: a meta-analysis. Intensive Care Med 40:332–341
11. Sud S, Friedrich JO, Taccone P, Polli F, Adhikari NK, Latini R, Pesenti A, Guerin C, Mancebo J, Curley MA, Fernandez R, Chan MC, Beuret P, Voggenreiter G, Sud M, Tognoni G, Gattinoni L (2010) Prone ventilation reduces mortality in patients with acute respiratory failure and severe hypoxemia: systematic review and meta-analysis. Intensive Care Med 36:585–599
12. Guerin C, Reignier J, Richard JC, Beuret P, Gacouin A, Boulain T, Mercier E, Badet M, Mercat A, Baudin O, Clavel M, Chatellier D, Jaber S, Rosselli S, Mancebo J, Sirodot M, Hilbert G, Bengler C, Richecoeur J, Gainnier M, Bayle F, Bourdin G, Leray V, Girard R, Baboi L, Ayzac L, PROSEVA Study Group (2013) Prone positioning in severe acute respiratory distress syndrome. N Engl J Med 368:2159–2168
13. Ioannidis JP, Trikalinos TA (2007) The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ 176:1091–1096
14. Altman DG (1980) Statistics and ethics in medical research. Misuse of statistics is unethical. Br Med J 281:1182–1184
15. Clec'h C, Ferriere F, Karoubi P, Fosse JP, Cupa M, Hoang P, Cohen Y (2004) Diagnostic and prognostic value of procalcitonin in patients with septic shock. Crit Care Med 32:1166–1169