Intensive Care Med (2017) 43:575–577
DOI 10.1007/s00134-017-4696-1

EDITORIAL

Small observational studies and data sharing: fuel for debate and coins for the piggy bank of evidence

Daniele Poole*

In an article recently published in *Intensive Care Medicine*, Constant et al. [1] report the results of their observational study investigating the effectiveness of targeted temperature management (TTM) in improving outcome in patients who were successfully resuscitated after cardiac arrest during surgical procedures [2]. The authors retrospectively reviewed 101 cases that occurred between 2008 and 2013 in 11 centres, 30 treated with TTM. Using logistic regression TTM did not turn out to be an independent predictor of favourable neurological outcome. Consistently with data from the literature, shockable rhythms were strongly protective while emergency surgery worsened the prognosis [2, 3].

The use of logistic regression was advisable to compensate the unbalances between the study and the control group in terms of important prognostic variables, estimating their prognostic weight independently one of the other. The multivariable approach increases the reliability of the findings, although it accounts only for measured variables and not for unmeasured ones (including those that are unknown) as randomized controlled trials do.

Moreover, in an observational study investigating the efficacy of a treatment we have a further complication. The probability of receiving the study treatment is different between study arms and may be related to important prognostic factors. For example, physicians may not prescribe a treatment, especially when it is expensive or demanding in terms of workload, to patients whose condition is too severe. In such cases a beneficial effect would be untruly attributed to treatment.

*Correspondence: daniele.poole@alice.it
Operative Unit of Anesthesia and Intensive Care, S. Martino Hospital, Belluno, Italy

From the statistical perspective, the study by Constant et al. has two weak points.

First, its small sample size does not allow one to draw conclusions of causal relation between the variables included in the multivariable model and the outcome. Actually, when this is the purpose of the analysis all the most important predictors should be included in the model [4]. An underfitted model (i.e. including an insufficient number of variables) may not include important causal factors and wrongly estimate weights of the included variables. On the other hand, overfitting can occur when too many variables in relation to the number of outcomes are included in the model, with a high risk of generating biased estimates of variables weights [5]. It has been demonstrated that when dealing with binary outcomes (yes or no events, the outcome in logistic regression) having at least ten outcomes for each variable is a safe threshold [6, 7], although it has been suggested that the risk of bias could be acceptable when the outcome/variable ratio is between 5 and 10 [8]. What researchers often disregard is that this ratio should be computed using the initial number of variables entering the model and not those left after the selection process has been carried out.

Thus, Constant et al. were dealing with a small sample for explanatory purposes (i.e. seeking causal relations between TTM and long-term neurological outcome) but were caught between the risk of over- and underfitting. The researchers included in their logistic regression model only six variables (eventually, only two variables turned out to be statistically significant) and could not include more because of the risk of overfitting. Moreover, because of the limited sample size the model was probably underpowered to detect important predictors. For example, time from cardiac arrest to restoration of

spontaneous circulation, a plausible predictor of poor neurological outcome [2], was barely statistically non-significant ($p = 0.11$). Is this because of insufficient power or because of specific features of intraoperative cardiac arrest? Although the no-flow time was very short, and it is reasonable to think it had little or any influence on the outcome, technically the study cannot provide the answer.

The second weak point is that their logistic regression does not account for prognostically important variables linked to indications for TTM. This can be done by developing a propensity score, usually using logistic regression including baseline variables as predictors and treatment as the dependent variable [7]. Patients with the same propensity score will, hence, have the same chance to receive or not to receive the treatment. The propensity score is then included in the final logistic regression model, which thus measures the prognostic weight of model variables (including TTM) given the same probabilities of receiving the treatment [9].

Alternatively, the propensity score can be used to create matched pairs of treated and untreated patients, creating a study and a control arm with the same probabilities of receiving the treatment and resembling randomized controlled trials, with the (not negligible) limit of being based only on measured variables. Constant et al. have adopted this approach as a sensitivity analysis. Propensity score matching was probably a secondary analysis because it allows one to assess the effectiveness of TTM but does not measure the prognostic weight of other variables (such as shockable rhythm) as logistic regression does. However, because of the small sample size, the propensity score was limited to three variables, not accounting for the complexity of criteria that rule the decision whether or not to give a specific treatment, with a high risk of generating a biased model [10].

The authors honestly acknowledge the limits and the exploratory nature of their study, calling for confirmatory research, without overemphasizing their results [11].

Why then should readers be interested in a study providing very low evidence in support or against TTM in intraoperative cardiac arrest?

In my opinion this study deserves credit for raising the important issue of how evidence should be applied in clinical practice. We have evidence from two randomized controlled trials that TTM is effective in improving prognosis after cardiac arrest [12, 13]. However, the authors stress the importance of not transferring automatically this evidence to intraoperative cardiac arrest, which bears specific features. I agree with this interpretation for several reasons. First, when general anaesthesia is performed a neuroprotective effect may be present. Second, in monitored patients detection of cardiac arrest is immediate, and consequently treatment is timely. Third, in patients under anaesthesia it is difficult to assess the presence of coma. Fourth, patients may frequently die because of causes related to surgical procedures and not because of anoxia following cardiac arrest. The combination of these factors can affect the severity and prognostic relevance of anoxia, the correct diagnosis of coma, and consequently the effectiveness of TTM.

Besides fuelling a debate on a clinically meaningful question, the study provides detailed descriptive data on the subject. Although the sample is limited, we have a clear picture of the characteristics of patients undergoing intraoperative cardiac arrest. Larger samples may be provided by permanent registries that, however, not being focused on the specific subject, usually lack important information.

Thus, small observational studies, based on clinically relevant hypotheses and carried out in unexplored fields where it is difficult to collect detailed data, are a valuable resource for medical science for the hypotheses they generate and for the descriptive data they provide, rather than for inferential analyses that are inherently weak. Their value would, however, be even greater if researchers made their data publicly available [14], stimulating the replication of their studies and favouring the progressive expansion of a common dataset. This would allow analyses carried out at patient level with greater statistical power and greater external validity, providing a contribution to evidence that each single small observational study could never provide.

Under the paradigm of data sharing, small observational studies could thus be saved in a common database as coins in the piggy bank of evidence.

Compliance with ethical standards

Conflicts of interest

The author declares no conflict of interest.

Received: 20 January 2017 Accepted: 25 January 2017
Published online: 20 February 2017

References

1. Constant AL, Mongardon N, Morelot Q, Pichon N, Grimaldi D, Bordenave L, Soummer A, Sauneuf B, Merceron S, Ricome S, Misset B, Bruel C, Schnell D, Boisramé-Helms J, Dubuisson E, Brunet J, Lasocki S, Cronier P, Bouhemad B, Carreira S, Begot E, Vandenbunder B, Dhonneur G, Jullien P, Resche-rigon M, Bedos JP, Montlahuc C, Legriel S (2017) Targeted temperature management after intraoperative cardiac arrest: a multicenter retrospective study. Intensive Care Med. doi:10.1007/s00134-017-4709-0
2. Dumas F, Grimaldi D, Zuber B et al (2011) Is hypothermia after cardiac arrest effective in both shockable and nonshockable patients? Insights from a large registry. Circulation 123(8):877–886
3. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270(24):2957–2963

4. Katz MH (2003) Multivariable analysis: a primer for readers of medical research. Ann Intern Med 138(8):644–650
5. Babyak MA (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 66(3):411–421
6. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15(4):361–387
7. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
8. Vittinghoff E, McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 165(6):710–718
9. Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res 46(3):399–424
10. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I (2013) Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. BMJ 347:f6409
11. Poole D, Nattino G, Bertolini G (2014) Overoptimism in the interpretation of statistics: the ethical role of statistical reviewers in medical journals. Intensive Care Med 40(12):1927–1929
12. Bernard SA, Gray TW, Buist MD et al (2002) Treatment of comatose survivors of out-of-hospital cardiac arrest with induced hypothermia. N Engl J Med 346(8):557–563
13. Hypothermia after Cardiac Arrest Study Group (2002) Mild therapeutic hypothermia to improve the neurologic outcome after cardiac arrest. N Engl J Med 346(8):549–556
14. Warren E (2016) Strengthening research through data sharing. N Engl J Med 375(5):401–403