The Bayesian approach: may we learn a lesson from the ANDROMEDA-SHOCK trial?

Federico Franchi[1], Cesare Biuzzi[1], Emanuele Detti[1], Gabriele Cevenini[2], Sabino Scolletta[1]

[1]Department of Medicine, Surgery and Neurosciences, Anesthesia and Intensive Care Unit, [2]Department of Medical Biotechnologies, University Hospital of Siena, Siena, Italy

Correspondence to: Prof. Sabino Scolletta. Department of Medicine, Surgery and Neurosciences, Anesthesia and Intensive Care Unit, University Hospital of Siena, Viale Bracci 11, 53100 Siena, Italy. Email: sabino.scolletta@dbm.unisi.it.

Background: from physiology to the ANDROMEDA-SHOCK trial

Increasing serum lactate levels and signs of tissue hypoperfusion are frequently observed in patients with septic shock. Hyperlactatemia is strongly related to abnormal peripheral perfusion, organ failure and mortality (1). The Surviving Sepsis Campaign Guidelines suggest a resuscitation strategy aimed at obtaining serum lactate normalisation (2). However, lactate kinetics is complex, and an increase in lactate levels may not be directly correlated with tissue hypoperfusion, but could be due to decreased clearance or other mechanisms (3). Since lactate clearance may not serve as a timely response to treatment, serum lactate should be measured every 2 hours (4).

Other targets to guide resuscitation in septic patients have been proposed (5). Among these, one of the most commonly used is the capillary refill time (CRT) because it is a rapid, easy-to-use and resource-independent index (6). Since it was first described in 1947, it has been widely used in adults and children. Its physiological background is complex because CRT is influenced by many variables, such as blood driving pressure, arteriolar tone and constituents of blood. CRT measures the amount of time necessary for the skin to return to baseline colour after pressure (generally a fingertip) is applied to a soft tissue. One of its major limitations is inter-rater variability; thus, precise training is required to improve CRT reproducibility. CRT is a tool for assessing the severity of critical illness because of its association with hyperlactatemia and the higher sequential organ failure assessment score (SOFA). In addition, it is a predictive sign of increased mortality in septic shock patients (6,7).

Recently, in the ANDROMEDA-SHOCK trial, CRT and serum lactate levels normalisation were assessed as strategies of targeted resuscitation in septic shock patients (8). In this randomised controlled trial, targeting resuscitation interventions on normalisation of CRT instead of serum lactate levels did not reduce 28-day mortality. An absolute risk mortality difference between these two groups of 8.5% (34.9% *vs.* 43.4% for CRT and serum lactate levels normalisation, respectively) was found, but without statistical significance (P=0.06). In contrast, Zampieri *et al.*, in a Bayesian reanalysis of the ANDROMEDA-SHOCK trial (9), found peripheral perfusion-targeted resuscitation was related to lower mortality and faster resolution of organ dysfunction compared to the lactate-based strategy. It is not the first time that the Bayesian reanalysis approach has been adopted in the medical literature. For instance, a recent large randomised trial on the use of extracorporeal membrane oxygenation for acute distress respiratory syndrome was labelled 'negative' (P=0.09) (10), but a posterior Bayesian reanalysis overturned those results (11). One could wonder about the possibility of obtaining two opposite conclusions starting from the same data. In this editorial, we would like to discuss this issue and comment on the so-called 'significant results'.

Page 2 of 4

Franchi et al. Reanalyzing ANDROMEDA-SHOCK trial with Bayesian approach

Statistical inference: same numbers, different approaches

The American Statistical Association published a statement (12) exhorting researchers to discard the term 'statistical significance' and offering a 'not-to-do' list about P value. In summary, when commenting on the results of a study, we are encouraged not to draw conclusions based on an arbitrary 'statistical threshold' such as P<0.05. In the frequentist statistical approach, probability is seen as an objective value associated with the estimation of fixed and unknown parameters of a statistical model through inferential procedures applied to random samples of independent data. Its major drawback consists of interpreting the type-I probability (P value) as a direct measurement of the findings' validity, although it is only a measurement of model-associated data compatibility with the 'null hypothesis' of no differences among groups. To overcome this reductionist view of medicine and draw together statistics and medicine, a Bayesian approach could be used. Here, probability is seen as a subjective value and parameters as random variables, and the inferential procedure is based on the probability distribution of the parameters derived by observing data and having further information available (13). The starting point of these two approaches differs diametrically. With frequentist statistics, our trials are built on the probability of obtaining some data if the null hypothesis were true, starting from the end of the procedure in deductive logic. Instead, Bayesian inference allows us to start from the knowledge already acquired to measure the plausibility of our hypothesis in an inductive reasoning approach. Bayesian statistics quantitatively bring this external information into the probability calculation. A so-called prior, that is, a pre-test belief about the magnitude and distribution of the effect size of the treatment before having the data, is combined with a likelihood function that summarises the information about the parameters given the data set to produce new posterior probabilities. It is calculated using Bayes' theorem, which states that the posterior probability is directly proportional to the product of the likelihood and the prior. In other words, Bayesian inference updates the a priori probability through data evidence to reach a 'less uncertain' posterior probability. It represents a mathematical transposition of learning from experience, and it invites us to reconsider the strength of our previous ideas.

It is easy to observe how the two statistical approaches converge when the sample size tends to the population size. An increase of the sample size in the frequentist approach involves a reduction of the confidence interval of the parameters, determined by the data, whereas in the Bayesian approach, we obtain a reduction of the credibility interval of the parameters associated with the probability of finding the parameter starting from the initial opinion (prior probability) refined by the data.

A new way of interpreting the ANDROMEDA-SHOCK trial

The ANDROMEDA-SHOCK trial aimed to demonstrate a 15% reduction in 28 all-cause mortalities targeting resuscitation by CRT instead of a strategy based on serum lactate levels in patients with septic shock. Secondary outcomes included a variety of measures of recovering from organ dysfunction. At 28 days, the mortality rates were 43.4% in the lactate-guided group and 34.9% in the CRT-guided group, with a hazard ratio of 0.75 (95% CI, 0.55 to 1.02) in favour of the CRT-guided group (P=0.06). There was less organ dysfunction at 72 hours in the CRT-guided group, but none of the other secondary outcomes showed a significant difference between the two groups. These results raised several concerns: the study could be interpreted as a failed test of the superiority of a CRT-based resuscitation, probably because it was underpowered for the main outcome. However, one should not ignore, at least, the non-inferiority of this strategy, even if the study was not designed to test for equivalence.

Using a Bayesian reanalysis approach, Zampieri *et al.* examined data from the ANDROMEDA-SHOCK trial and defined four priors, which were the mathematical representations of different opinions about the effect of the intervention (optimistic, neutral, null, and pessimistic). The enthusiastic prior was the best estimate and corresponded to the effect size used for the sample size calculation of the original study. Even in the absence of previous information or external evidence, this approach has been proposed in literature when analysing a frequentist trial (13). Testing the trial data with the predefined priors allowed to check if they were sensible to different beliefs. A reliable statement in favour of the intervention (CRT-guided resuscitation) would be made if the results changed minimally among differing priors. When the study data are sufficiently strong, differing priors have minimal influence on the calculated posteriors. In contrast, if the study data are relatively weak, the posteriors will not agree. Nonetheless, this lack of consensus is likely appropriate given the absence of

sufficiently compelling data.

The probability of whether CRT-guided therapy would reduce 28-day mortality (OR <1) was independent of the selected prior and maintained above 90%. When considering an OR <0.8, the posterior probability was equal or greater than 80% for all the priors, except for the pessimistic one (the latter represented a very pessimistic scenario in which, however, a 20% possibility that the intervention is beneficial remained). This trend favouring the CRT group persisted also at 90 days post-inclusion. Considering the absolute mortality between the two groups, there is an estimated reduction at 28 days ranging from 7% to 13%. The authors also designed an analysis for the first secondary endpoint of the original trial, the SOFA score at 72 hours. With the CRT strategy, patients had a higher probability of being in the lower quartile of the SOFA score (between 0 and 7).

Bayesian reanalysis weak points

Critics often question the confirmation bias behind this kind of unplanned Bayesian reanalysis because the latter could be altered by the known data that resulted from the study and thus be prone to be interpreted with subjectivity. One could argue that this analysis would have never been performed if the ANDROMEDA-SHOCK trial results were positive because a Bayesian approach served as a booster to overcome statistical significance. Actually, the Bayesian inference is much closer to the clinicians' inductive way of thinking, combining a previous belief with the results of a non-fully conclusive trial to obtain a posterior probability of the tested intervention. Priors do not work as a booster: in contrast, they bring back the observed effects to a clinical real context, moderating unrealistic results and otherwise supporting plausible ones. The subjectivity of Bayesian analysis seems to be a strong point because mathematically defining qualitative beliefs makes a posterior judgement more explicit and informative than just a fixed threshold value.

Zampieri *et al.* also provided the results of a reanalysis based on a frequentist approach. Instead of the Cox model used originally, with a logistic regression model, the OR for 28-day mortality was 0.61 (95% CI, 0.38–0.92), with a P=0.022. In this way, changing the statistical test led to overturning the interpretation of the trial. The primary endpoint of 28-day mortality reduction reached 'statistical significance'. If the authors of the ANDROMEDA-SHOCK

trial had used this analysis from the beginning, their study would have been considered conclusively in favour of a CRT strategy. This contrast should be considered proof of the risks of misinterpretation of the P value, rather than a reason to abandon frequentist statistics.

The crux of the matter

What can we learn from ANDROMEDA-SHOCK trial? Either serum lactate levels or CRT could be fundamental tools in the evaluation of tissue hypoperfusion. We observed with interest the results of the study, and we wondered why CRT worked better than lactate levels. Was it due to different interventions in the two groups (amount of fluids, vasopressors)? Was it due to different measurement intervals of the parameters (every 30 min CRT, every 2 h lactate)? The study was not designed to answer these questions.

Several articles in the literature are sceptical about the actual importance of 'numbers power'. A recent editorial stated that '*the problem is the whole concept of statistical significance*' (14): studies are categorised as positive (P<0.05) or negative (P>0.05) in a binary way. It is clear how this could be wrong because two studies with the same effect size but with a P value of 0.05 for the first and of 0.06 for the second would be reported as positive for the first and negative for the second. This is illogical. We think we should stay away from a dichotomous consideration of the reality, particularly when discussing results from large clinical trials. The use of a different statistical method, the Bayesian analysis, may fundamentally change this point of view. The Bayesian analysis, based on a different interpretation of statistical inference, allows clinicians to obtain a non-dichotomous view of the results, expressing them as probability and permitting us to interpret them with a 'clinical view' rather than as only pure numbers.

In conclusion, researchers should be encouraged to use additional statistical methods that could help readers get oriented with the results of large clinical trials. We endorse the attempt by Zampieri *et al.*, to improve the interpretation of the ANDROMEDA-SHOCK trial. We end this editorial by quoting a sentence from a recent article published by *Nature*: '*Inferences should be scientific, and that goes far beyond the merely statistical*' (15).

Acknowledgments

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Glenn Hernández and Guo-wei Tu) for the series "Hemodynamic Monitoring in Critically Ill Patients" published in *Annals of Translational Medicine*. The article was reviewed by the Guest Editor Dr. Guo-wei Tu.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi. org/10.21037/atm.2020.02.17). The series "Hemodynamic Monitoring in Critically Ill Patients" was commissioned by the editorial office without any funding or sponsorship. Prof. FF reports personal fees from VYGON, outside the submitted work. Prof. SS reports personal fees from EDWARDS LIFESCIENCES, personal fees from VYGON, outside the submitted work. The other authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Vincent JL, Quintairos A, Couto L Jr, et al. The value of blood lactate kinetics in critically ill patients: a systematic review. Crit Care 2016;20:257-70.
2. Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: international guidelines for management of sepsis and septic shock: 2016. Crit Care Med 2017;45:486-552.
3. Hernandez G, Bellomo R, Bakker J. The ten pitfalls of lactate clearance in sepsis. Intensive Care Med 2019;45:82-5.
4. Levy MM, Evans LE, Rhodes A. The Surviving Sepsis Campaign bundle: 2018 update. Crit Care Med 2018;46:997-1000.
5. Hariri G, Joffre J, Leblanc G, et al. Narrative review: clinical assessment of peripheral tissue perfusion in septic shock. Ann Intensive Care 2019;9:37-45.
6. Ait-Oufella H, Bige N, Boelle PY, et al. Capillary refill time exploration during septic shock. Intensive Care Med 2014;40:958-64.
7. Lima A, Jansen TC, van Bommel J, et al. The prognostic value of the subjective assessment of peripheral perfusion in critically ill patients. Crit Care Med 2009;37:934-8.
8. Hernández G, Ospina-Tascón GA, Damiani LP, et al. Effect of a Resuscitation Strategy Targeting Peripheral Perfusion Status vs Serum Lactate Levels on 28-Day Mortality Among Patients With Septic Shock: The ANDROMEDA-SHOCK Randomized Clinical Trial. JAMA 2019;321:654-64.
9. Zampieri FG, Damiani LP, Bakker J, et al. Effect of a Resuscitation Strategy Targeting Peripheral Perfusion Status vs Serum Lactate Levels on 28-Day Mortality Among Patients with Septic Shock: A Bayesian Reanalysis of the ANDROMEDA-SHOCK Trial. Am J Respir Crit Care Med 2019. doi: 10.1164/rccm.201905-0968OC.
10. Combes A, Hajage D, Capellier G. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. N Engl J Med 2018;378:1965-75.
11. Goligher EC, Tomlinson G, Hajage D. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. JAMA 2018;320:2251-9.
12. Wasserstein R, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat 2016;70:129-33.
13. Wijeysundera DN, Austin PC, Hux JE, et al. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2009;62:13-21.
14. Davidson A. Embracing uncertainty: The days of statistical significance are numbered. Paediatr Anaesth 2019;29:978-80.
15. Amrhein A, Greenland S, McShane B. Retire statistical significance. Nature 2019;567:305-7.