EDITORIAL

The use of enrichment to reduce statistically indeterminate or negative trials in critical care

Randomised controlled trials (RCTs) are clinical experiments that assess whether the intervention tested improves the primary outcome. Random allocation of subjects and assessment of their outcomes in experiments to infer causality is credited to Ronald Fisher and Jerzy Neyman [1]. The underpinning theory of RCTs is that the difference in primary outcome between the treated and the controls provides an unbiased estimate of the causal effect of the intervention on the outcome. All RCTs start with a study population derived by inclusion and exclusion criteria. There is equipoise about the treatment effect. Randomisation ensures that measurable and, more importantly, unmeasurable characteristics are distributed in a way that reduces confounding to chance. The allocation of the next subject is not predictable. The subjects and investigators are unware of the allocation, which minimises bias. Compliance with the trial protocol is assessed and reported. The a-priori-designed analyses plans are adhered to when reporting. Thus, well-designed and conducted RCTs potentially provide the best graded evidence base for clinical care [2].

Most RCTs randomly allocate individuals and test whether the intervention is superior to standard of care, although we could also test for equivalence and non-inferiority of the new treatment [3]. Positive result from a well conducted RCT is accepted and incorporated into clinical practice, especially when replicated in two or more trials. When it is a statistically indeterminate or negative RCT and the intervention tested has biological plausibility, it makes us consider whether there are patient subsets who might have benefited and could we characterise them to inform future trial design? One reason for considering this argument is that critical care (ICU) trials are most often conducted in acute respiratory distress syndrome (ARDS [4]) and sepsis [5,6] and enroll heterogeneous populations. In this editorial, we highlight '*The only formula of physiological statistics*' [7] proposed by David Sackett as a way of reasoning statistically negative RCTs (Fig. 1). For understanding risk domains in RCT participants, the readers are directed to in depth review by Feudtner and colleagues [8]. The focus of our paper is to briefly discuss how signal enrichment and noise reduction could help reduce statistically negative RCTs. We use examples from ARDS and sepsis literature, as they have major overlaps in their biology [9,10] and the most common aetiology of ARDS is sepsis [11].

Heterogeneity

Heterogeneity is the norm in sepsis and ARDS patients. Heterogeneity is the inter-individual variation in susceptibility to illness or outcome from illness that arises from factors that do not clearly delineate a high risk sub-population. Individuals vary in their susceptibility to both the illness itself and the outcomes from the illness. Birth cohort studies have highlighted that the variation in risk of outcomes and susceptibility to illness is determined very early on in life [12]. Although genetic and environmental factors are important contributors, heterogeneity is often a random event [13]. The heterogeneity of sepsis and ARDS populations in their risk factors for critical illness, age, comorbidities, number and severity of organ failures is well established [14,15]. Although less well established, this is presumably true for the underlying mechanisms of these syndromes [9,10,16]. Susceptibility to any specific therapy is likely to be driven by the biology of an individual patient's

sepsis or ARDS as well as their underlying risk of death. There are a number of scenarios that may generate different treatment responses:

Grouping based on susceptibility to tested treatment

It seems intuitive that patients enrolled in a clinical trial should have a mechanism that can be responsive to that intervention. Trials of inotropes in septic shock should target patients with decreased cardiac output [17], trials of PEEP in ARDS should enrol patients who have recruitable lung [18], and trials targeted at endotoxin removal should be endotoxinaemic [19]. Despite this, critical care trials, either for reasons of pragmatic design, lack of a biomarker, or lack of agreement on the mechanism of the treatment are often not designed with this in mind. However, recent advances suggest several novel strategies.

Enrichment can be achieved by using stratification on a biomarker that is linked to the tested treatment. Meisel and colleagues proposed a sepsis subgroup with low monocyte HLA-DR expression as a marker of immunosuppression who had an enhanced response to granulocyte–macrophage colony-stimulating factor (GM-CSF) in an early phase trial [20]. In ICU patients, cytomegalovirus (CMV) reactivation is associated with adverse outcomes [21]. Therefore, CMV-seropositive status in ICU patients could stratify patients for administering prophylaxis to prevent CMV reactivation and improve outcomes [22].

Enrichment can also be achieved by using transient response to an intervention that is going to be tested in the trial. Goligher and colleagues re-analysed data from two ARDS trials to test the hypothesis that improvement in oxygenation in response to positive end expiratory pressure (PEEP) will identify a treatment-responsive ARDS subgroup. The authors highlight that a lower risk of death in the PEEP-responsive group will benefit from a RCT reassessing the value of higher versus lower PEEP in ARDS [23]. This approach is particularly appealing in trials of physiologic interventions with an immediately measurable physiologic response, and relies on the assumption, which needs to be validated, that physiologic response will translate in to outcome benefit. The design issue here is that physiologic response is used as an enrolment criterion not an outcome.

In many cases, the mechanistic endotypes of critical illness syndromes are less clear. In these situations, investigators have relied on empiric clustering algorithms to group patients based on clinical and biologic variables. There are numerous methods such as latent class analyses and clustering algorithms, depending on the data available to identify these groups [24, 25].

Sometimes, subgroups identified with these techniques suggest underlying biologic responses to illness. For example, Davenport et al. used whole blood leukocyte gene expression in sepsis due to community acquired pneumonia, and observed two patient clusters based on empiric algorithms (sepsis response signatures or SRS). A patient cluster identified as SRS1 had an immunosuppressed phenotype and greater mortality, which could identify a biological response group for immune reconstitution therapies in future RCTs [26]. Recently, Calfee et al. re-analysed data from three ARDS RCTs [27,28] by including additional biomarkers of inflammation, alveolar and endothelial injury and tested whether subgroups of patients identified with latent class analysis could be identified

within ARDS RCT populations. In all three RCTs, they consistently identified two subgroups of ARDS, with one subgroup showing greater inflammation amongst other biomarker characteristics and mortality. This ARDS subgroup also had a statistically significant interaction with the treatments tested in the RCTs – mechanical ventilation and fluid management. The implication of statistical interaction with the treatment is that it potentially identifies a subgroup of patients who on average are likely to have different treatment response to the intervention tested. The principal challenges from these two studies for future trial design includes the retrospective lookback analyses, measurement of selected biomarkers and use of latent class analyses that removes data granularity by using standard normal distribution. Thus, the questions remain as to whether we could prospectively identify similar patient groups for a trial inclusion and could we do this using a limited set of clinically feasible markers [27-29].

Wong et al. identified two septic shock subsets in paediatric population using gene expression data. The authors then tested whether these septic shock subgroups differed in the association to treatment effect of corticosteroids. They highlighted two main phenotypes differentiated by inflammation and a four-fold difference in mortality by steroid treatment category. They proposed that this gene expression profile could enrich septic shock population for future RCTs [30,31].

This concept of corticosteroid responsiveness has also been investigated in adult septic shock population by Bentzer et al. Corticosteroid responsiveness was defined as improved odds of survival. Importantly, they used plasma cytokine levels and had propensity score-matched septic shock patients who didn't receive corticosteroids as controls. Corticosteroid responsiveness was identified by detection of interleukin-3 (IL3), IL6, or chemokine ligand-4 above predefined threshold values [32]. The additional value of this study was that plasma cytokine measurement could be done as a point-of-care test.

Grouping based on difference in susceptibility to death

The differences in the effect of treatments across the range of severity of illness is well described [33]. This is true because the effect of treatment is relative to the baseline risk of death, and therefore the attributable benefit of treatment declines as baseline risk declines [34]. The difference in outcome over the changing baseline risk is called the heterogeneity in treatment effect. These findings are more problematic as the risks of therapy are often fixed and not related to underlying risk of death. Extending this argument, if a trial population has different representations of low and very high risk populations, the treatment effect is likely to vary significantly. Iwashyna et al. recently demonstrated this mathematical phenomenon using simulations of RCTs in sepsis and in ARDS. They showed that as the risk of death changes from low to high, the difference in mortality between the intervention and the control arm increases, thereby highlighting a sub-group that is likely to benefit from the intervention [35].

This model may itself be overly simplistic. Let us consider the relationship between sepsis and mortality. The term-attributable risk of sepsis refers to the difference in probability of death between patients with sepsis and those without sepsis. In the context of a RCT, the attributable benefit of an intervention refers to improvement in mortality in the treated compared with controls.

Mathematically, the attributable benefit of a therapy must vary with the baseline risk of death. Therefore, enrolling patients at higher risk of death or other outcomes will reduce the study sample size requirements, which is prognostic enrichment.

However, there may be more interesting phenomena occurring here. First, the baseline risk of death in a sepsis RCT population is related to illness severity, which is often measured using APACHE-2 and organ dysfunction scores. Our assumption in a RCT is that illness severity is caused by sepsis and therefore should get better with the treatment tested. One could argue that these severity scores may well be markers for unrelated biologic phenomenon. This has been reported in sepsis. If we account for pre-illness deterioration, very different sepsis-outcome relationships emerge [36]. Therefore, an argument would be, treatment-effect heterogeneity may be a simple way of clinical trial enrichment for poorly understood mechanistic phenomena. Second, for a given illness, the outcome of interest has more than one causal mechanism or many risk factors [37]. A patient with more attributable risk from sepsis is likely to have greater attributable benefit from treatment. An extreme example would be: two patients with similar illness severity score but one is a healthy 20 year-old with sepsis and the other is a 70 year-old with pneumonia and severe comorbidity.

There are two ways to take advantage of this phenomenon. First, in the design phase of the trial, enrolment can be restricted to the more severe end of the spectrum. A good example to consider in this context is the recombinant activated Protein C (drotrecogin alfa (activated) (DrotAA) trials in sepsis. The first phase-3 trial was stopped early for efficacy as crude mortality in the intention-to-treat population was reduced by 6.1% with a relative risk reduction of 19.4% [38]. However, sub-group analyses implied that this unexpected efficacy was due to differences in treatment response in a high risk of death sub-population. This lead to two further trials, one in low risk of death population [39] and a second trial in a high risk of death population enriched by persistent vasopressor requirement after resuscitation [40]. Both trials were statistically negative. The first [39] highlights the important concept that as the risk of death due to underlying illness and the benefit of therapy declines, the risk of therapy predominates, as highlighted by increased risk of bleeding. Thus, it is not always effective as an enrolment strategy to focus of high risk of outcome event population, especially when the outcome is determined by non-modifiable risk factors such as age or comorbidities.

Second, in the analysis phase of trials, we can delineate sub-groups based on treatment effect with mathematical approaches to incorporating baseline risk with limited covariates such as a logistic regression model or the more recently described Virtual Twins method [41].

In summary, trial enrichment is not always efficient, as it represents a trade-off between the underpinning research time it takes to identify how best to perform enrichment, the patients lost during the process of enrichment as they do not meet the criteria, and the magnitude of increase in signal that is derived. Our discussions thus far have deliberately avoided the terms ‘personalised’ and ‘precision’ medicine. When using enrichment approaches as described above, we are increasing the pre-test probability and likelihood of treatment response at a cohort level, not at an individual level. Our paper highlights the need to have a deeper biological and granular understanding of these two

critical illnesses to perform enrichment. In any RCT, there will be patients who survive irrespective of treatment allocation either due to: low-risk of death; placebo response; spontaneous improvement; or survival 'due' to the intervention (true responders). Similarly, in any RCT there will be patients who will die irrespective of the intervention. The likelihood of a positive trial could be improved by enriching true responders based on their susceptibility to either the tested treatment or to death or ideally a combination of the two approaches.

Acknowledgements

MS-H is supported by the National Institute for Health Research Clinician Scientist Award (NIHR-CS-2016-16-011). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. No other conflicts of interest declared.

M. Shankar-Hari
Consultant, Department of Intensive Care Medicine
Guy's and St Thomas' NHS Foundation Trust
Honorary Senior Lecturer,
Division of Asthma, Allergy and Lung Biology
Kings College London, UK
email: manu.shankar-hari@kcl.ac.uk

G D. Rubenfeld
Professor of Medicine
Interdepartmental Division of Critical Care Medicine, University of Toronto
Chief, Program in Trauma, Emergency and Critical Care, Sunnybrook Health Sciences Center
2075 Bayview Avenue, Room D108c
Toronto, ON M4N 3M5

References

1. Rubin DB. Causal inference using potential outcomes. *Journal of the American Statistical Association* 2005; **100:** 322-31.
2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal* 2008; **336:** 924-6.
3. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG, Group C Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *Journal of the American Medical Association* 2012; **308:** 2594-604.
4. Force ADT, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin Definition. *Journal of the American Medical Association* 2012; **307:** 2526-33.
5. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Journal of the American Medical Association* 2016; **315:** 801-10.
6. Shankar-Hari M, Phillips GS, Levy ML, et al. Developing a new definition and assessing new clinical criteria for septic shock. *Journal of the American Medical Association* 2016; **315:** 775-87.
7. Sackett DL Why randomized controlled trials fail but needn't. *Canadian Medical Association Journal* 2001; **165:** 1226-37.
8. Feudtner C, Schreiner M, Lantos JD. Risks (and benefits) in comparative effectiveness research trials. *New England Journal of Medicine* 2013; **369:** 892-4.
9. Cohen J, Vincent JL, Adhikari NK, et al. Sepsis: a roadmap for future research. *Lancet Infectious Diseases* 2015; **15:** 581-614.
10. Sweeney RM, McAuley DF. Acute respiratory distress syndrome. *Lancet* 2016; **388:** 2416-30.
11. Bellani G, Laffey JG, Pham T, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *Journal of the American Medical Association* 2016; **315:** 788-800.
12. Pearson H. Epidemiology: study of a lifetime. *Nature* 2011; **471:** 20-4.
13. Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *International Journal of Epidemiology* 2011; **40:** 537-62.
14. Shankar-Hari M, Harrison DA, Rowan KM. Differences in impact of definitional elements on mortality precludes international comparisons of sepsis epidemiology-a cohort study illustrating the need for standardized reporting. *Critical Care Medicine* 2016; **44:** 2223-30.
15. Shaver CM, Bastarache JA. Clinical and biological heterogeneity in acute respiratory distress syndrome: direct versus indirect lung injury. *Clinics in Chest Medicine* 2014; **35:** 639-53.
16. Xiao W, Mindrinos MN, Seok J, et al. A genomic storm in critically injured humans. *Journal of Experimental Medicine* 2011; **208:** 2581-90.
17. Annane D, Vignon P, Renault A, et al. Norepinephrine plus dobutamine versus epinephrine alone for management of septic shock: a randomised trial. *Lancet* 2007; **370:** 676-84.
18. Goligher EC, Kavanagh BP, Rubenfeld GD, Ferguson ND. Pro: physiologic responsiveness should guide entry into randomized controlled trials. *American Journal of Respiratory and Critical Care Medicine* 2015; 192: 1416-9.
19. Klein DJ, Foster D, Schorr CA, Kazempour K, Walker PM, Dellinger RP. The EUPHRATES trial (evaluating the use of polymyxin B hemoperfusion in a randomized controlled trial of adults treated for endotoxemia and septic shock): study protocol for a randomized controlled trial. *Trials* 2014; **15:** 218.
20. Meisel C, Schefold JC, Pschowski R, et al. Granulocyte-macrophage colony-stimulating factor to reverse sepsis-associated immunosuppression: a double-blind, randomized, placebo-controlled multicenter trial. *American Journal of Respiratory and Critical Care Medicine* 2009; **180:** 640-8.
21. Limaye AP, Kirby KA, Rubenfeld GD, et al. Cytomegalovirus reactivation in critically ill immunocompetent patients. J*ournal of the American Medical Association* 2008; **300:** 413-22.
22. Cowley NJ, Owen A, Millar J, et al. Antiviral prophylaxis inhibits cytomegalovirus reactivation in critical illness. *Critical Care* 2015; **19:** P115.
23. Goligher EC, Kavanagh BP, Rubenfeld GD, et al. Oxygenation response to positive end-expiratory pressure predicts mortality in acute respiratory distress syndrome. A secondary analysis of the LOVS and ExPress trials. *American Journal of Respiratory and Critical Care Medicine* 2014; **190:** 70-6.
24. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5:** 21-7.

25. Dalton L, Ballarin V, Brun M. Clustering algorithms: on learning, validation, performance, and applications to genomics. *Current Genomics* 2009; **10:** 430-45.
26. Davenport EE, Burnham KL, Radhakrishnan J, et al. Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *Lancet Respiratory Medicine* 2016; **4:** 259-71.
27. Famous KR, Delucchi K, Ware LB, et al. ARDS Subphenotypes Respond Differently to Randomized Fluid Management Strategy. *American Journal of Respiratory and Critical Care Medicine* 2016.
28. Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respiratory Medicine* 2014; **2:** 611-20.
29. Shankar-Hari M, McAuley DF. Acute respiratory distress syndrome phenotypes and identifying treatable traits. The dawn of personalized medicine for ARDS. *American Journal of Respiratory and Critical Care Medicine* 2017; **195:** 280-1.
30. Wong HR, Atkinson SJ, Cvijanovich NZ, et al. Combining prognostic and predictive enrichment strategies to identify children with septic shock responsive to corticosteroids. *Critical Care Medicine* 2016; **44:** e1000-3.
31. Wong HR, Cvijanovich NZ, Anas N, et al. Developing a clinically feasible personalized medicine approach to pediatric septic shock. *American Journal of Respiratory and Critical Care Medicine* 2015; **191:** 309-15.
32. Bentzer P, Fjell C, Walley KR, Boyd J, Russell JA. Plasma cytokine levels predict response to corticosteroids in septic shock. *Intensive Care Medicine* 2016; **42:** 1970-9.
33. Willke RJ, Zheng Z, Subedi P, Althin R, Mullins CD. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Medical Research Methodology* 2012; **12:** 185.
34. Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *American Journal of Epidemiology* 1998; **148:** 1117-26.
35. Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of heterogeneity of treatment effect for reporting and analysis of randomized trials in critical care. *American Journal of Respiratory and Critical Care Medicine* 2015; **192:** 1045-51.
36. Iwashyna TJ, Netzer G, Langa KM, Cigolle C. Spurious inferences about long-term outcomes: the case of severe sepsis and geriatric conditions. *American Journal of Respiratory and Critical Care Medicine* 2012; **185:** 835-41.
37. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *American Journal of Public Health* 2005; **95 Suppl 1:** S144-50.
38. Bernard GR, Vincent JL, Laterre PF, et al. Efficacy and safety of recombinant human activated protein C for severe sepsis. *New England Journal of Medicine* 2001; **344:** 699-709.
39. Abraham E, Laterre P-F, Garg R, et al. Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death. *New England Journal of Medicine* 2005; **353:** 1332-41.
40. Ranieri VM, Thompson BT, Barie PS, et al. Drotrecogin alfa (activated) in adults with septic shock. *New England Journal of Medicine* 2012; **366:** 2055-64.
41. Foster JC, Taylor JM, Ruberg SJ Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011; **30:** 2867-80.

Figure 1 *'The only formula of physiological statistics*' [7]

$$\boldsymbol{Confidence\ in\ trial\ result} = \frac{\boldsymbol{Signal}}{\boldsymbol{Noise}}\ \boldsymbol{X}\ \sqrt{\boldsymbol{Samplesize}}$$

Let us consider a drug trial with mortality as the primary outcome for explaining this formula. The confidence in the trial result includes the mortality difference between the intervention and the control arm and the measure of precision around that difference, i.e. confidence interval (CI). The signal refers to the 'magnitude' of difference in mortality between the intervention and control arm, higher difference increases our confidence in the trial result, especially when it is replicated in many RCTs. The noise refers to all numeric and descriptive sources of variations (heterogeneity) in a trial and includes inter-individual variation in responses to treatment. Therefore, to get positive RCTs with confidence in their result, aside from increasing sample size, our options are to either increase the signal and/or reduce the noise. The easy explanation of a negative RCT is that intervention does not work. However, if the intervention tested has biological plausibility, then understanding how to optimise the signal is important. Superiority RCTs aim to demonstrate that the intervention improves outcome compared with standard care (i.e. intervention is better than standard care). Equivalence RCTs aim to demonstrate that the new intervention is as good as the old intervention, within a pre-specified range. Non-inferiority RCTs aim to demonstrate that although the new intervention is inferior to the old intervention in improving outcomes, the difference is not clinically significant. In all these RCTS, the conclusion about the effect of the intervention on primary outcome represents the average effect across the included population. It does not give any clue whether the patient in front of you will benefit or not. We also don't have any idea about the magnitude of benefit or harm for an individual patient that you are going to use this drug on.

RCT, randomised clinical trial