Review

Evidence-based medicine: Classifying the evidence from clinical trials – the need to consider other dimensions

Rinaldo Bellomo[1,2] and Sean M Bagshaw[1]

[1]Department of Intensive Care, Austin Hospital, Studley Rd, Heidelberg, Victoria 3084, Australia
[2]Faculty of Medicine, University of Melbourne, Royal Parade, Parkville, Victoria 3052, Australia

Corresponding author: Rinaldo Bellomo, rinaldo.bellomo@austin.org.au

Published: 4 October 2006
This article is online at http://ccforum.com/content/10/5/232

Critical Care 2006, **10:**232 (doi:10.1186/cc5045)

Abstract

The current approach to assessing the quality of evidence obtained from clinical trials focuses on three dimensions: the quality of the design (with double-blinded randomised controlled trials representing the highest level of such design); the statistical power (beta) and the level of significance (alpha). While these aspects are important, we argue that other significant aspects of trial quality impinge upon the truthfulness of the findings: biological plausibility, reproducibility and generalisability. We present several recent studies in critical care medicine where the design, beta and alpha components of the study are seemingly satisfactory but where the aspects of biological plausibility, reproducibility and generalisability show serious limitations. Accordingly, we argue for more reflection, definition and consensus on these aspects of the evaluation of evidence.

> "The extent to which beliefs are based on evidence is very much less than believers suppose."
>
> Bertrand Russell (1928)
> *Sceptical Essays*

Introduction

The evidence-based medicine (EBM) movement has brought about a paradigm shift not only in medical practice and education, but also in study design and in the appraisal and classification of published research in the field of critical care medicine, as well as medicine in general [1,2]. The principles created by pioneers in the field of EBM are now widely accepted as the standard not only for appraising the quality of evidence, but also for evaluating the strength of evidence produced by research [1,2]. These principles allow for evidence to be classified into different 'levels' according to specific characteristics. Accordingly, from these levels of evidence, recommendations are issued, each with its own 'grade' [3] (Table 1). These recommendations then typically influence clinical practice around the world through the promotion of consensus conferences, clinical practice guidelines, systematic reviews or editorials on specific aspects of patient care [4,5].

In this review, we will argue that the present system for how we classify the quality of evidence and formulate recommendations from such evidence would benefit from a refinement. We will argue that a refined system should ideally integrate several dimensions of evidence, in particular related to study design, conduct and applicability that were not explicitly discussed at the beginning of the EBM movement nor are presently considered or incorporated in widely accepted classification systems. In this context, we will further comment on the newly proposed hierarchical system, the Grades of Recommendation Assessment, Development and Evaluation (GRADE) system, for gauging the quality of evidence and strength of recommendations from research evidence. Our intent in this editorial is to generate dialogue and debate about how we currently evaluate evidence from research. We aim to create impetus for a broad consensus, which may both highlight limitations and promote important changes in how we currently classify evidence and, hopefully, lead to an improvement not only in the design and reporting of trials but also the quality of clinical practice in critical care medicine.

Reflections on predicting the future, the truth and evidence

In ideal circumstances, critical care physicians would be capable of predicting the biological future and clinical outcome of their patients with complete and unbiased accuracy and thus employ this knowledge to take care of them. For example, they would know that early administration of tissue plasminogen activator to a given patient with acute submassive pulmonary embolism would allow survival

ARDS = acute respiratory distress syndrome; EBM = evidence-based medicine; GRADE = Grades of Recommendation Assessment, Development and Evaluation; HFOV = high-frequency oscillatory ventilation.

Table 1

Overview of a simplified and traditional hierarchy for grading the quality of evidence and strength of recommendations

Levels of Evidence	
Level I	Well conducted, suitably powered RCT
Level II	Well conducted, but small and under-powered RCT
Level III	Non-randomized observational studies
Level IV	Non-randomized study with historical controls
Level V	Case series without controls
Grades of recommendations	
Grade A	Level I
Grade B	Level II
Grade C	Level III or lower

Levels of evidence are for an individual research investigation. Grading of recommendations is based on levels of evidence. Adapted from [1,2]. RCT, randomized controlled trial.

whereas other interventions would not [6]. Likewise, the clinician would know with certainty that this patient would not suffer any undue adverse consequences or harm as a result of treatment with tissue plasminogen activator.

Regrettably, we live in a less than ideal world where a patient's biological and clinical future cannot be anticipated with such certainty. Instead, the clinician can only be partly reassured by knowing 'the operative truth' for questions about this intervention. What would result if all such patients with submassive pulmonary embolism were randomly allocated to receive either tissue plasminogen activator or an alternative treatment? Would one intervention increase survival over the other? By what magnitude would survival increase? How would such an increase in survival weigh against the potential harms? Thus, the clinician would use 'the operative truth' about such interventions to guide in the routine care of patients.

Again, regrettably, such truth in absolute terms is unknown and unobtainable. Rather, clinicians have to rely on estimation, probability and operative surrogates of the truth for the prediction of the biological and clinical future of their patients. Such estimation is obtained through 'evidence'.

Evidence, of course, comes in many forms: from personal experience, teaching by mentors, anecdotes, case series, retrospective accounts, prospective observations, non-interventional controlled observations, before-and-after studies, single center randomized evaluations, randomized evaluation in multiple centers in one or more countries to double-blinded randomized multicenter multinational studies. Evidence in each of these forms has both merits and shortcomings. However, our intent is not to examine each in detail here.

As argued above, 'the truth' is an unknowable construct, and as such, the epistemology of how evidence evolves is much debated. The process of understanding how new evidence that is generated is translated into what clinicians need to know and integrated into patient care remains a great challenge [7]. This is further complicated by the sheer magnitude of the evidence produced for any given issue in critical care. Evidence is accumulating so rapidly that clinicians are often not able to assess and weigh the importance of the entire scope in detail. It is, therefore, not surprising that several hierarchical systems for classifying the quality of evidence and generating recommendations have been created in order to guide the busy clinician for decision making and ultimately caring for patients [8].

How a hierarchy of evidence is built

On the basis of reasonable thought, common sense, rational analysis, and statistical principles (but no randomized double-blinded empirical demonstration), the apex of the pyramid of evidence is generally the well-conducted and suitably powered multicenter multinational double-blind placebo-controlled randomized trial. Such a trial would be defined by the demonstration that intervention X administered to patients with condition A significantly improves their survival, a patient-centered and clinically relevant outcome, compared to placebo, given a genuine and plausible treatment effect of intervention X. This would be considered as level I evidence that intervention X works for condition A (Table 1). In the absence of such a trial, many would also regard a high quality systematic review and meta-analysis as level I evidence. However, systematic reviews require cautious interpretation and may not warrant placement on the apex of the hierarchy of evidence due to poor quality, reporting and inclusion of evidence from trials of poor quality [9]. In our opinion, they are best considered as a hypothesis generating activity rather than high quality evidence.

At this point, however, findings from such a trial would elicit a strong recommendation (for example, grade A), concluding that intervention X should be administered to a patient with condition A, assuming that no contraindications exist and that said patient fulfils the criteria used to enrol patients in the trial. Yet, there are instances when such a strong recommendation may not be issued for an intervention based on the evidence from such a trial. For instance, when an intervention fails to show improvement in a clinically relevant and patient-centered outcome, but rather uses a surrogate outcome. Moreover, when the apparent harms related to an intervention potentially outweigh the benefits, a lower grade of recommendation can be made (for example, grade B).

In general, this process would appear reasonable and not worthy of criticism or refinement. However, such hierarchical systems for assessing the quality of evidence and grading recommendations have generally only taken into account three dimensions for defining, classifying and ranking the quality of

Table 2

Overview of the GRADE system for grading the quality of evidence: criteria for assigning grade of evidence

Criteria for assigning level of evidence	
Type of evidence	
Randomized trial	High
Observational study	Low
Any other type of research evidence	Very low
Increase level if:	
Strong association	(+1)
Very strong association	(+2)
Evidence of a dose response gradient	(+1)
Plausible confounders reduced the observed effect	(+1)
Decrease level if:	
Serious or very serious limitations to study quality	(–1) or (–2)
Important inconsistency	(–1)
Some or major uncertainty about directness	(–1) or (–2)
Imprecise or sparse data[a]	(–1)
High probability of reporting bias	(–1)

[a]Few outcome events or observations or wide confident limits around an effect estimate. Adapted from [10].

Table 3

Overview of the GRADE system for grading the quality of evidence: definitions in grading the quality of evidence

Level of evidence	Definition
High	Further research is not likely to change our confidence in the effect estimate
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very Low	Any estimate of effect is uncertain

evidence obtained from clinical trials. Specifically, these include: study design; probability of an alpha or type-I error; and probability of beta or type-II error. A recent response to some of these concerns (the GRADE system) and some analytical comments dealing with the above fundamental aspects of trial classification will now be discussed.

The Grades of Recommendation Assessment, Development and Evaluation system

An updated system for grading the quality of evidence and strength of recommendations have been proposed and published by the GRADE Working Group [8,10-13]. The primary aim of this informal collaboration was to generate consensus for a concise, simplified and explicit classification system that addressed many of the shortcomings of prior hierarchical systems. In addition, such a revised system might generate greater standardization and transparency when developing clinical practice guidelines.

The GRADE system defines the 'quality of evidence' as the amount of confidence that a clinician may have that an estimate of effect from research evidence is in fact correct for both beneficial and potentially harmful outcomes [11]. A global judgment on quality requires interrogation of the validity of individual studies through assessment of four key aspects: basic study design (for example, randomized trial, observational study); quality (for example, allocation concealment, blinding, attrition rate); consistency (for example, similarity in results across studies); and directness (for example, generalizability of evidence). Based on each of these elements and a few other modifying factors, evidence is then graded as high, moderate, low or very low [11] (Tables 2 and 3).

The 'strength of a recommendation' is then defined as the extent to which a clinician can be confident that adherence to the recommendation will result in greater benefit than harm for a patient [11]. Furthermore, additional factors affect the grading of the strength of a recommendation, such as target patient population, baseline risk, individual patients' values and costs.

The GRADE system represents a considerable improvement from the traditional hierarchies of grading the quality of evidence and strength of recommendations and has now been endorsed by the American College of Chest Physicians Task Force [14]. However, there are elements of evidence from research that have not been explicitly addressed in the GRADE system, which we believe require more detailed discussion.

Traditional measures of the quality of evidence from research

Study design

The design of a clinical trial is an important determinant for its outcome, just as is the 'true' effectiveness of the intervention. As an interesting example, let's consider the ARDS Network trial of low tidal volume ventilation [15]. This study was essentially designed to generate a large difference between the control and the protocol tidal volume interventions for treatment of acute respiratory distress syndrome (ARDS). Thus, this design maximized the likelihood of revealing a difference in treatment effect. However, whether the tidal volume prescribed in the control arm represented a realistic view of current clinical practice remains a matter of controversy [16].

However, the principles of EBM would typically focus on several simple key components of study design, such as measures aimed at reducing the probability of bias (that is, randomization, allocation concealment, blinding). Therefore, for a trial to be classified as level I or high level evidence, it essentially requires incorporation of these elements into the design. This approach, while meritorious, often fails to account for additional dimensions of study design that deserve consideration.

First, as outlined above in the ARDS Network trial, was the control group given a current or near-current accepted therapy or standard of practice in the study centers? Second, how are we to classify, categorize and compare trials of surgical interventions or devices (that is, extracorporeal membrane oxygenation (ECMO) or high-frequency oscillatory ventilation (HFOV)) where true blinding is impossible? Third, how can we classify trials that assess the implementation of protocols or assessment of changes in process of care, which, similarly, cannot be blinded? Finally, do the study investigators from all centers have genuine clinical equipoise with regards to whether a treatment effect exists across the intervention and control groups? If not, bias could certainly be introduced.

As an example, if a randomized multicenter multinational study of HFOV in severe ARDS found a significant relative decrease in mortality of 40% ($p < 0.0001$) compared to low tidal volume ventilation, would this be less 'true' than a randomized double-blind placebo controlled trial showing that recombinant human activated protein C decreases mortality in severe sepsis compared to placebo? If this is less 'true', what empirical proof do we have of that? If we have no empirical proof, why would this finding not be considered as level I or high level evidence, given that blinding of HFOV is not possible?

These questions suggest there is a need to consider refinement of how we currently classify the quality of evidence according to study design. At a minimum, this should include principles on how to classify device and protocol trials and how to incorporate a provision that demonstrates the control arm received 'standard therapy' (which of itself would require pre-trial evaluation of current practice in the trial centers).

Alpha error

An alpha or type I error describes the probability that a trial would, by chance, find a positive result for an intervention that is effective when, in fact, it is not (false-positive). In general, the alpha value for any given trial is traditionally and somewhat arbitrarily set at <0.05. While recent trends have brought greater recognition for hypothesis testing by use of confidence intervals, the use of an alpha value remains frequent for statistical purposes and sample size estimation in trial design.

The possibility of an alpha error is generally inversely related to study sample size. Thus, a study with a small sample size or relatively small imbalances between intervention groups (for example, age, co-morbidities, physiological status, and so on) or numerous interim analyses might be sufficient, alone or together, to lead to detectable differences in outcome not attributable to the intervention. Likewise, a trial with few observed outcome events, often resulting in wide confidence limits around an effect estimate, will be potentially prone to such an error.

Level I or high level evidence demands that trials should have a low probability of committing an alpha error. Naturally, this is highly desirable. However, how do we clinically or statistically measure a given trial's probability of alpha error? Is there a magic number of randomized patients or observed events in each arm that makes the probability of committing an alpha error sufficiently unlikely (no matter the condition or population) to justify classifying a study as level I or high level evidence? If so, how can such a magic number apply across many different situations as can be generated by diseases, trial design and treatment variability? How should the probability of a trial's given alpha error be adjusted to account for statistical significance? Should the burden of proof be adjusted according to the risk and cost of the intervention?

There are suggested remedies for recognizing the potential for bias due to an alpha error in a given trial by assessment of key aspects of the trial design and findings. These include whether the trial employed a patient-centered or surrogate measure as the primary outcome, evaluation of the strength of association between the intervention and primary outcome (for example, relative risk or odds ratio), assessment of the precision around the effect estimate (for example, confidence limits), and determination of the baseline or control group observed event rate. In the end, however, other than use of a patient-centered primary outcome, how should such an error be prevented? These unresolved questions suggest a need for both debate and consensus on the concept of alpha error and its practical application.

Beta error

The term beta or type II error describes a statistical error where a trial would find that an intervention is negative (that is, not effective) when, in fact, it is not (false-negative). A larger study sample size, and thus number of observed outcome events, reduces the probability of a trial committing a beta error on the assumption that a genuine difference in effect exists across intervention groups. In order to minimize the chance of a beta error, trials have to be suitably 'powered'. In general, the probability of beta error is traditionally and, again, arbitrarily set at 0.10 to 0.20 (for example, power 0.80 to 0.90) and used in the statistical design and justification of trial sample size. Inadequately powered trials risk missing small but potentially important

clinical differences in the hypothesized intervention [17,18]. Thus, of course, the ideal trial is one in which the power is high.

The risk of a beta error can be reduced by making rational assumptions, based on available evidence, on the likelihood of a given outcome being observed in the control arm of the trial and the size of treatment effect of the intervention (for example, absolute and relative risk reduction). However, such assumptions are often wide of the mark [19]. While maximizing the power of a given trial may seem logical, such an increase has both ethical and cost considerations [20]. Thus, power is expensive. For example, for a large multicenter multinational trial to decrease the probability of a beta error (for example, increase the power) from 0.20 to 0.10, the result would be greater recruitment, an increase in the number of patients exposed to placebo interventions, and possibly result in a multi-million dollar increase in cost. Is this money wisely spent? Should suitable power (and its cost) be a matter of statistical considerations only? If so, where should it be set for all future large trials? Or should power be subject to other considerations, such as the cost of the intervention being tested, the size of the population likely to benefit, the relevance of the clinical outcome being assessed, the future cost of the medication and other matters of public health? In addition, these issues need consideration in the context of trials of equivalency or non-superiority and for trials that are stopped at interim analyses for early benefit [21-23]. Finally, future trials need to address whether estimates of risk reduction used for sample size calculations for a given intervention are biologically plausible, supported by evidence and feasible in the context of the above mentioned considerations [24]. These issues deserve both debate and consensus on the concept of beta error and its practical application.

Additional dimensions to the quality of evidence from research

In the above paragraphs, we have discussed several controversial aspects of the three major dimensions used in generating and assessing the quality of evidence. In the next few paragraphs, we would like to introduce additional dimensions of evidence, which we believe should be formally considered or addressed in future revised consensus systems, such as the GRADE system, for grading the quality of evidence from research.

Biological plausibility

The evidence from trials does not and cannot stand on its own, independent of previous information or studies. While this might seem obvious, more subtle views of biological plausibility may not. For example, most, perhaps all, clinicians and researchers would reject the results of a randomized controlled study of retroactive intercessory prayer showing that such intervention leads to a statistically significant decrease in the duration of hospital stay in patients with positive blood cultures [25]. Such a study completely lacks biological plausibility [26]. Fewer clinicians, however, would have rejected the findings of the first interim analysis of the AML UK MRC study of 5 courses of chemotherapy compared to 4, when they showed a 53% decrease in the odds of death (odds ratio 0.47, 95% confidence interval 0.29 to 0.77, $p = 0.003$) [23]. Yet the data safety and monitoring committee continued the trial because these initial findings were considered too large to be clinically possible and lacked biological plausibility. The committee recommended the trial be continued and the final results (no difference between the two therapies) vindicated this apparent chance finding at interim analysis [23].

In this vein, how does intensive insulin therapy provide large benefits for surgical but not medical patients [27,28]? Yet, few physicians would now reject the findings of a mortality benefit of an intensive insulin therapy trial in critically ill patients [28]. However, the point estimate of the relative reduction in hospital mortality in this trial was 32% (95% confidence interval 2% to 55%, $p < 0.04$), thus making the lowering of blood glucose by 3.9 mmol/l for a few days more biologically powerful than trials on the effect of thrombolytics in acute myocardial infarction (26%) or ACE inhibitors in congestive heart failure (27%) [29-31]. Is this biologically plausible? No one to date has sought to incorporate biological plausibility into the grading of the quality of evidence or strength of recommendations from such studies. We believe that future assessment of evidence should consider this dimension and develop a systematic consensus approach to how biological plausibility should influence the classification of evidence.

Reproducibility

Reproducibility in evidence refers to finding consistency in an effect of an intervention in subsequent trials and in diverse populations, settings, and across time. Such consistency essentially considers the ability of a given intervention applied in a trial to be easily reproduced elsewhere. For example, the PROWESS trial tested the efficacy of rhAPC in severe sepsis; however, it was limited in scope by the study inclusion criteria (that is, adults, weight <135 kg, age >18 years, and so on) [32]. Yet, evidence of effect in additional populations and settings is less certain [33-36]. In addition, this intervention carries such an extraordinary cost that it makes its applicability outside of wealthy countries near impossible and unfeasible [37,38].

Likewise, interventions that involve complex devices, therapies, protocols or processes (that is, HFOV, continuous renal replacement therapy, intensive insulin therapy or medical emergency teams) as applied in a given trial imply an entire infrastructure of medical, surgical and nursing availability, knowledge, expertise and logistics that are often not universally available [19,28,39,40]. The translation of a particular intervention in isolation to a setting outside of its

initial development may have negative and cost consequences in a different setting.

Due thought needs to be given to how the results of a trial can be translated into interventions that reliably work, are reproducible and can be applied elsewhere. These concerns should not be taken to encourage 'evidence-based relativism' or 'ignorance-based nihilism' such that no evidence is worth considering unless 'it was obtained here'. Rather, their aim is to generate a search for better trial designs and better evaluation of evidence. The GRADE system incorporates a subjective assessment of consistency as criteria for grading the quality of evidence and, in the setting of unexplained heterogeneity across trials, suggests a decrease in grade [11].

Generalizability

The generalizability of findings from a clinical trial represents a fundamental dimension of evidence, that of external validity. Narrow controls designed to optimize the internal validity of a trial (that is, inclusion/exclusion criteria, intervention protocol) can compete with and compromise overall generalizability [41]. Furthermore, an individual trial's generalizability can also be the result of additional factors. More subtly, the results of a trial might come from the application of a given therapy in a multicenter setting that included only large academic centers. Alternatively, use of a particular agent might significantly impact upon the results of an intervention (for example, etomidate use in the recent French study of the treatment of relative adrenal insufficiency [42]), whereas such an agent is simply not available elsewhere (as in Australia, where etomidate is not approved for patient use) [43]. Further, the power of the investigator-protagonist needs to be taken into account. Such investigators, when involved in single center studies, especially unblinded ones, have the power to profoundly influence outcome and behavior through their commitment to the cause, expertise, dedication and enthusiasm. Examples of such studies include use of early-goal directed therapy, higher volume continuous veno-venous hemofiltration, tight glycemic control or implementation of medical emergency teams [19,28,39,44]. These studies have several details in common. All these trials are single center, using complex interventions/protocols with a local protagonist.

How generalizable are the findings of a single center study, however well designed? Can or should level I or high level evidence ever come from single center trials? They currently do. How should we classify an intervention that works in a single center trial? For example, would early goal directed resuscitation really improve the outcome of all patients with septic shock presenting to emergency departments around the world or do the results of this trial simply reflect improvements in patient care in a single institution where there existed a very high pre-intervention mortality [44]? Similarly, would intensive insulin therapy really reduce mortality in all surgical intensive care unit patients worldwide or do these results merely reflect the consequences of increased patient care in a single institution where the mortality of the control cardiac surgery patients was particularly high [28]? Finally, would higher volume hemofiltration really reduce the mortality of all acute renal failure patients or are the results of this study a reflection of increased patient attention by a specific high-experience team in a center with a unique acute renal failure population and a very low incidence of sepsis [39]? These are more than idle questions because all of the above studies have profoundly influenced and are still shaping the practice of critical care around the world [5]. Yet two recent assessments of interventions that, in single center studies, looked extraordinarily promising (steroids for the fibro-proliferative phase of ARDS and introduction of a medical emergency team) failed to show a benefit when taken to a multicenter setting [19,45]. A similar fate might well await other single center studies that are currently being incorporated into guidelines.

Furthermore, we need to highlight and better understand the limitations of data from single center trials. We need to consider the meaning of multicenter and how it relates to grading the quality of evidence. We need to relate the control population studied in any single or multicenter trial to other large populations with respect to the same condition, so that we can consider the 'generalizability level' of a given study. We also need to give weight to the meaning of 'multinational' in terms of quality of evidence.

In addition, we may need to think more about the association between evidence and 'the unknowable' truth in the context of the limitations of randomized controlled trials. For example, a multicenter prospective epidemiological study of 10,000 patients showing a significant association between intervention X and patient outcome Y with narrow confidence limits and a $p < 0.0001$ after controlling for more than 50 major variables might also need to be taken into account. While this obviously overlaps with issues of study design, such an observational study might provide a better real world estimate of the effect of an intervention than a double-blind randomized controlled trial in a single center. Randomized trials, especially if associated with complex and strict protocols and many exclusion criteria, often give us the ability to know much but only about a world that does not exist. Large observational studies, on the other hand, carry much uncertainty about causality but do describe the 'real' world. Likewise, observational studies have the distinct advantage of examining the long-term effects or prognosis of an intervention and assessing for adverse or rare outcome events.

If we think that large observational studies approximate 'the truth' as much as small single center studies, we need to recognize this in our classification systems. The GRADE system has taken a positive step forward for recognizing the potential importance of high quality observational studies that clearly reveal a strong association between exposure and outcome (Tables 2 and 3).

Table 4

Summary of components to consider when evaluating the quality of evidence from research

Study design	Randomized
	Allocation concealment
	Blinding (if possible)[a]
	Clinically important and objective primary outcome
	Beta-error[b]
	Multi-center
Study conduct	Intention-to-treat analysis
	Follow-up or attrition rate
	Completion to planned numbers
Study findings	Biological plausibility
	Strength of estimate of effect
	Precision of estimate of effect
	Observed event rate
Study applicability	Consistency across similar studies
	Reproducibility
	Generalizability

[a]Blinding may not be possible in device or protocol/process trials. [b]Adequately powered, appropriate estimate of control event rate and relative or absolute reduction in clinically important primary outcome.

The need for further refinement and consensus

An argument can be made that proposed classification systems, especially the new GRADE system, are best left alone. They are reasonably simple, explicit, have been validated and now are increasingly endorsed. Furthermore, the dimensions of evidence discussed in this editorial (study design, biological plausibility, reproducibility and generalizability) are difficult to simply measure and their impact on how the findings of an individual trial approximate the 'truth' is hard to quantify (Table 4). However, we believe our arguments are valid and warrant discussion.

A classification system that is simple is indeed desirable but becomes a problem when, for the sake of simplicity, it fails to take into account important aspects of the growing complexity of the nature of the evidence available. We also accept that a classification system should seek to quantify its components and that some of the additional dimensions of evidence that we propose may be difficult to quantify. Some of them, however, are numerical (one center versus ten centers versus twenty centers or one nation versus two nations versus three nations) and could be quantified. For some of the issues we raise there will likely not be scientifically valid answers. In their absence, there is need for broad consensus.

We acknowledge the view that the issues we raised could simply be left to clinician judgement. However, while it is true that clinician judgement will always play a role, it is misleading to believe that busy clinicians can and do regularly read the published reports of trials in detail and integrate them within a fully informed assessment of the previous literature. The evidence to the contrary is clear.

Accordingly, summary classifications of the quality of evidence and strength of recommendations, such as the GRADE system, will continue to have an important and expanding role in medicine. We believe that as the GRADE system becomes more widely endorsed, additional refinements to the system will result in appropriate recognition of higher quality evidence and contribute to greater confidence in recommendations for clinical practice. We also believe that this field is very much 'work in progress' and needs to evolve more explicit recognition and classification of the dimensions of trial design discussed in this manuscript

Conclusion

In this review, we have argued in favor of the concept that assessment of the quality of evidence from trials in critical care medicine requires ongoing refinement. Such refinement should, in particular, reflect those dimensions of evidence that are currently not explicitly addressed. The GRADE Working Group has made considerable contributions to improving how the quality of research evidence and recommendations are graded. We believe that additional refinement is needed to explicitly address and quantify dimensions of evidence such as biological plausibility, reproducibility and generalizability. We believe such refinement should occur through consensus and we hope that this article will add further impetus for this process to continue and advance, especially in the field of critical care medicine. We also believe that such refinement would have lasting beneficial effects on clinical practice and on the future design and reporting of clinical trials and research.

Competing interests

The author declares that they have no competing interests.

References

1. Cook DJ, Guyatt GH, Laupacis A, Sackett DL: **Rules of evidence and clinical recommendations on the use of antithrombotic agents.** *Chest* 1992, **102:**305S-311S.
2. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ: **Clinical recommendations using levels of evidence for antithrombotic agents.** *Chest* 1995, **108:**227S-230S.
3. Guyatt GH, Cook DJ, Sackett DL, Eckman M, Pauker S: **Grades of recommendation for antithrombotic agents.** *Chest* 1998, **114:**441S-444S.
4. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P: **Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group.** *Crit Care* 2004, **8:**R204-212.
5. Dellinger RP, Carlet JM, Masur H, Gerlach H, Calandra T, Cohen J, Gea-Banacloche J, Keh D, Marshall JC, Parker MM, *et al.*: **Surviving Sepsis Campaign guidelines for management of severe sepsis and septic shock.** *Crit Care Med* 2004, **32:**858-873.

6. Konstantinides S, Geibel A, Heusel G, Heinrich F, Kasper W: **Heparin plus alteplase compared with heparin alone in patients with submassive pulmonary embolism.** *N Engl J Med* 2002, **347:**1143-1150.
7. Upshur RE: **The ethics of alpha: reflections on statistics, evidence and values in medicine.** *Theor Med Bioeth* 2001, **22:** 565-576.
8. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B, *et al.*: **Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group.** *BMC Health Serv Res* 2004, **4:**38.
9. Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, Doig CJ: **A systematic evaluation of the quality of meta-analyses in the critical care literature.** *Crit Care* 2005, **9:**R575-582.
10. **GRADE Working Group** [http://www.GradeWorkingGroup.org]
11. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, *et al.*: **Grading quality of evidence and strength of recommendations.** *BMJ* 2004, **328:**1490.
12. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, Hill S, Jaeschke R, Liberati A, Magrini N, *et al.*: **Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system.** *BMC Health Serv Res* 2005, **5:**25.
13. Schunemann HJ, Best D, Vist G, Oxman AD: **Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations.** *Can Med Assoc J* 2003, **169:**677-680.
14. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schunemann H: **Grading strength of recommendations and quality of evidence in clinical guidelines: report from an american college of chest physicians task force.** *Chest* 2006, **129:**174-181.
15. The Acute Respiratory Distress Syndrome Network: **Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome.** *N Engl J Med* 2000, **342:**1301-1308.
16. Oba Y, Salzman GA: **Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury.** *N Engl J Med* 2000, **343:**813; author reply 813-814.
17. Moher D, Dulberg CS, Wells GA: **Statistical power, sample size, and their reporting in randomized controlled trials.** *JAMA* 1994, **272:**122-124.
18. Halpern SD, Karlawish JH, Berlin JA: **The continuing unethical conduct of underpowered clinical trials.** *JAMA* 2002, **288:**358-362.
19. Hillman K, Chen J, Cretikos M, Bellomo R, Brown D, Doig G, Finfer S, Flabouris A: **Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial.** *Lancet* 2005, **365:**2091-2097.
20. Whitley E, Ball J: **Statistics review 4: sample size calculations.** *Crit Care* 2002, **6:**335-341.
21. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, *et al.*: **Randomized trials stopped early for benefit: a systematic review.** *JAMA* 2005, **294:**2203-2209.
22. Le Henanff A, Giraudeau B, Baron G, Ravaud P: **Quality of reporting of noninferiority and equivalence randomized trials.** *JAMA* 2006, **295:**1147-1151.
23. Wheatley K, Clayton D: **Be skeptical about unexpected large apparent treatment effects: the case of an MRC AML12 randomization.** *Control Clin Trials* 2003, **24:**66-70.
24. Finfer S, Bellomo R, Boyce N, French J, Myburgh J, Norton R: **A comparison of albumin and saline for fluid resuscitation in the intensive care unit.** *N Engl J Med* 2004, **350:**2247-2256.
25. Leibovici L: **Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial.** *BMJ* 2001, **323:**1450-1451.
26. Hettiaratchy S, Hemsley C: **Effect of retroactive intercessory prayer. Paper proves power of statistics, not prayer.** *BMJ* 2002, **324:**1037; author reply 1038-1039.
27. Van den Berghe G, Wilmer A, Hermans G, Meersseman W, Wouters PJ, Milants I, Van Wijngaerden E, Bobbaers H, Bouillon R: **Intensive insulin therapy in the medical ICU.** *N Engl J Med* 2006, **354:**449-461.
28. van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R: **Intensive insulin therapy in the critically ill patients.** *N Engl J Med* 2001, **345:**1359-1367.
29. Wilcox RG, von der Lippe G, Olsson CG, Jensen G, Skene AM, Hampton JR: **Trial of tissue plasminogen activator for mortality reduction in acute myocardial infarction. Anglo-Scandinavian Study of Early Thrombolysis (ASSET).** *Lancet* 1988, **2:**525-530.
30. The CONSENSUS Trial Study Group: **Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS).** *N Engl J Med* 1987, **316:**1429-1435.
31. Swedberg K, Held P, Kjekshus J, Rasmussen K, Ryden L, Wedel H: **Effects of the early administration of enalapril on mortality in patients with acute myocardial infarction. Results of the Cooperative New Scandinavian Enalapril Survival Study II (CONSENSUS II).** *N Engl J Med* 1992, **327:**678-684.
32. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW, Fisher CJ Jr: **Efficacy and safety of recombinant human activated protein C for severe sepsis.** *N Engl J Med* 2001, **344:** 699-709.
33. Frassica JJ, Vinagre YM, Maas B: **Recombinant human-activated protein C (rhAPC) in childhood sepsis.** *J Intensive Care Med* 2004, **19:**56-57.
34. Kylat R, Ohlsson A: **Recombinant human activated protein C for severe sepsis in neonates.** *Cochrane Database Syst Rev* 2006, **3(34):**CD 005385.
35. Fry DE, Beilman G, Johnson S, Williams MD, Rodman G, Booth FV, Bates BM, McCollam JS, Lowry SF: **Safety of drotrecogin alfa (activated) in surgical patients with severe sepsis.** *Surg Infect (Larchmt)* 2004, **5:**253-259.
36. Abraham E, Laterre PF, Garg R, Levy H, Talwar D, Trzaskoma BL, Francois B, Guy JS, Bruckmann M, Rea-Neto A, *et al.*: **Drotrecogin alfa (activated) for adults with severe sepsis and a low risk of death.** *N Engl J Med* 2005, **353:**1332-1341.
37. Chalfin DB, Teres D, Rapoport J: **A price for cost-effectiveness: implications for recombinant human activated protein C (rhAPC).** *Crit Care Med* 2003, **31:**306-308.
38. Manns BJ, Lee H, Doig CJ, Johnson D, Donaldson C: **An economic evaluation of activated protein C treatment for severe sepsis.** *N Engl J Med* 2002, **347:**993-1000.
39. Ronco C, Bellomo R, Homel P, Brendolan A, Dan M, Piccinni P, La Greca G: **Effects of different doses in continuous venovenous haemofiltration on outcomes of acute renal failure: a prospective randomised trial.** *Lancet* 2000, **356:**26-30.
40. Derdak S, Mehta S, Stewart TE, Smith T, Rogers M, Buchman TG, Carlin B, Lowson S, Granton J: **High-frequency oscillatory ventilation for acute respiratory distress syndrome in adults: a randomized, controlled trial.** *Am J Respir Crit Care Med* 2002, **166:**801-808.
41. Green LW, Glasgow RE: **Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology.** *Eval Health Prof* 2006, **29:**126-153.
42. Malerba G, Romano-Girard F, Cravoisy A, Dousset B, Nace L, Levy B, Bollaert PE: **Risk factors of relative adrenocortical deficiency in intensive care patients needing mechanical ventilation.** *Intensive Care Med* 2005, **31:**388-392.
43. Annane D: **ICU physicians should abandon the use of etomidate!** *Intensive Care Med* 2005, **31:**325-326.
44. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M: **Early goal-directed therapy in the treatment of severe sepsis and septic shock.** *N Engl J Med* 2001, **345:**1368-1377.
45. Steinberg KP, Hudson LD, Goodman RB, Hough CL, Lanken PN, Hyzy R, Thompson BT, Ancukiewicz M: **Efficacy and safety of corticosteroids for persistent acute respiratory distress syndrome.** *N Engl J Med* 2006, **354:**1671-1684.