MINERVA ANESTESIOL 2007;73:595-601

REVIEW ARTICLE

Feasibility, limits and problems of clinical studies in Intensive Care Unit

G. GRASSELLI [1], L. GATTINONI [2], B. KAVANAGH [3], R. LATINI [4], A. LAUPACIS [5], F. LEMAIRE [6], A. PESENTI [2], P. SUTER [7], A. SLUTSKY [8], G. TOGNONI [9]

[1]Department of PerioperativeMedicine and IntensiveCare, San Gerardo Hospital, University Milano Bicocca, Milan, Italy; [2]Department of Anesthesia and IntensiveCare, IRCCS Foundation, " Maggiore Policlinico, Mangiagalli, Regina Elena" Hospital, University of Milan, Milan, Italy; [3]Department of Anesthesiology and Critical Care Medicine, Hospital for Sick Children, University of Toronto, Toronto Canada; [4]Department of Cardiovascular Research, Mario Negri Institute, Milan, Italy, [5]Li Ka Shing Knowledge Institute, To ronto, Canada; [6]Assistance Publique-Hôpitaux de Paris, H. Mondor Hospital, University Paris XII, Créteil, France; [7]University of Geneva, Geneva, Switzerland; [8]Division of Respiratory Medicine, St. Michael's Hospital, University of Toronto, Toronto, Canada; [9]Consorzio Mario Negri Sud, Santa Maria Imbaro, Chieti, Italy

ABSTRACT

In critical care medicine there is still a paucity of evidence on how to manage most of the clinical problems commonly encountered in critically ill patients. Randomized controlled trials (RCTs) are the most powerful instruments to evaluate the efficacy of a therapeutic intervention and to generate evidence for clinical practice. Unfortunately, the design and conduct of RCTs in our field are particularly complicated, because of some intrinsic and structural problems (*e.g.* lack of reliable nosography, concomitant use of different therapies, problems in the definition of end-points besides mortality) that will be discussed in this review Further challenges are represented by the lack of tradition of large ICU networks, difficulties in linking or integrating physiologic and therapeutic objectives in designing clinical protocols, scarcity of independent or non-profit funds. A particularly stimulating opportunity of development is represented also by the relationship of critical care to EBM. Because of the above problems, metanalyses could be less informative than in other areas of medicine, as they are based on few trials which are often contradictory and of unsatisfactory quality. Few suggestions are formulated which could help looking forwards.

Key words: Randomized controlled trial - Evidence-based medicine - Data interpretation, statistical - Mortality.

The process of a clinical trial

Clinical trials are today recognized as the preferred, and somehow mandatory tool, to produce an information on drugs and/or strategies of intervention which could be considered sufficiently reliable to guide routine care. There is now a broad agreement on the criteria and conditions for the planning and execution of physiologic as well as of therapeutic trials, which can be summarize d in the following points, where the main strengths as well as weaknesses are underlined.

1. Motivation(s) and feasibility. The scenarios are many:

a. the need to define whether and how an "intervention" (not necessarily centered on a drug) produces effects which correspond to a well-predefined hypothesis leads to "physiologic" trials, which allow a more thorough understanding of the interplay of the variables and/or indicators/markers of a clinical condition and of its evolution;

b. a therapeutic trial aims to assess the existence, the direction, the size, the clinical and statistical significance of an intervention on the morbidity-mortality of a target population;

c. a limited number of well defined and carefully monitored patients is usually required for physiologic studies; much larger populations representatives of the real conditions of care are needed to test the efficacy (and the associated safety) of an intervention on clinical (and/or epidemiological) outcomes of relevant morbidities and mortality;

d. the combination of the two approaches would be obviously more informative, but it is rarely planned and implemented, for logistics, financial, but mainly motivational reasons of the promoters;

e. it is increasingly recognized that in many instances the motivations for a trial reflect more a pressure from the market (mainly for therapeutic trials).

2. Issue definition: this requirements encompasses both the need for an explicit and univocal definition of the condition to be treated, and of the expected outcome(s) or result(s). The "laboratory-like" approach requested for a physiologic study is more compatible with the above requirements. Therapeutic trials on complex conditions, such as those more specifically met in critical care, could be facing difficulties in the process of definition(s) (*e.g.* sepsis or acute respiratory distress syndrome, ARDS).

3. Protocol design: a correct description of the study design, of inclusion-exclusion criteria and of required sample size is of cornerstone importance. Ethical issues (informed consent) and identification of the type and number of participating units also deserve special attention.

4. Analysis of results: statistical methods must be rigorous and appropriate.

5. Publication: the site, or journal, is of utmost importance "Top journals" are obviously preferred for important positive or negative results, though sometimes reasons other than strict scientific quality determine the accessibility to "top journals", and therefore the potential impact of the results on medical practice.

Clinical trials in the intensive care unit: peculiarities and special issues

The process of planning and implementing trials according to the above steps is certainly more likely to be not easy in many conditions of critical care, because of their baseline complexity, variability, difficult definition(s), coexistence of many not easily standardized management strategies. The following considerations underline some of the main issues.

1. ICU are dealing more with syndromes than with well defined diseases. The diagnostic and prognostic definition of these syndromes is sometimes very difficult because it is based on nonspecific and nonselective criteria that are often a matter of debate. For example, diagnostic criteria for ARDS are extremely broad and nonspecific [1] and encompass several clinical situations that are clearly unrelated to each other.[2] The inclusion criteria define therefore inevitably populations which are expected to reflect important variabilities, which are likely not to be perfectly matched by the randomization process: higher "numbers" are desirable, but they imply very broad networks, where practices could however contribute further variability due to the sometimes different and not easily comparable management strategies.

2. Type of intervention: therapeutic interventions in the ICU are in general much more complicated than in other settings and are quite different from the simple administration of a new drug. In critical care we test interventions that are complex, require a relatively long learning time and sometimes are not well defined. In addition, the patients receive many concomitant therapies and this makes the evaluation of the effect of a single intervention quite difficult: in other words, it can be hard to distinguish the direct effect of the experimental treatment from the composite effect of all concurrent therapies. A corollary of this observation is that for many of the clinical conditions we have to face we do not have a "standard treatment" to be used as control. A clear example of this is the ARDSnet trial, that compared the effect of two different tidal volumes (12 mL/kg, predicted body weight as standard treatment *vs* 6 mL/kg, as the experimental arm) in patients with ARDS.[5] The trial unequivocally demonstrated that using

a tidal volume of 6 mL/kg is better than using the higher tidal volume in terms of overall survival. But can we affirm without doubt that a tidal volume of 12 mL/kg really represented standard treatment?[6, 7]

3. Outcome indicators: in ICU we treat conditions associated with very high mortality rates. This obviously makes mortality an important outcome (maybe the most important one) but potentially very hard to improve, so that large number of patients are required to demonstrate a significant survival advantage. To date, only a very limited number of the therapies used in critical care medicine have been proved to significantly reduce mortality. The lack of a clear impact on mortality has caused the rejection of several promising new therapies in the last decade. As pointed out by Petros *et al.* in 1995, the use of mortality as a primary endpoint in ICU-based trials is associated with some problems.[8] Besides the importance of large number of patients, two lines of methodological solutions could be possibly adopted.

The use of attributable mortality instead of all-cause mortality would be desirable, since it may allow a reduction of the sample size. To understand this concept, we can consider the patients with ARDS: overall mortality is 40% and mortality attributable to ventilator-induced lung injury (VILI) can be estimated around 10%. Let's imagine that we want to test a new ventilatory strategy that halves VILI-induced mortality: to demonstrate a drop in overall mortality from 40% to 35% we have to enrol 1 471 patients, whereas to show a reduction in VILI-attributable mortality from 10% to 5% the required sample size would be only of 435 subjects. Unfortunately, determining the attributable mortality can be difficult.

Though the use of different prognostic severity scores (APACHE, SAPS, ect.), and/or indicators of morbidity may be imprecise and misleading: an intelligent (and clearly pre-defined in terms of reliability and qualified relevance) adoption of surrogate and/or intermediate endpoints (instead of, or cumulative with mortality) could be explored and encouraged. For example, when testing a new ventilatory strategy, measures of its physiologic effect (such as gas exchanges or lung mechanics) or measures of morbidity (such as the rate of infections or duration of mechanical ventilation) may be more relevant endpoints than overall mortality.

The problem of evidence-based medicine and statistical considerations

Evidence-based medicine (EBM) can be defined as the application of the best evidence from research to clinical practice. The concept of EBM was first proposed in 1992;[9] since then, it has become increasingly popular and there are now many dedicated journals and more than 30 websites.

The basic principle of EBM is that clinical practice should be based on the results of published trials.[10] In the attempt to define "the best evidence from research", a hierarchy of the possible sources has been established. RCTs are judged to provide the best possible evidence, followed by meta-analyses, case-control and case-series studies.[11] Appropriately sized and high quality RCTs are the most powerful tool, but this does not mean that the results of a single RCT should be passively accepted and become a new standard of care.[12] Many questions are still open and RCTs on the same subject often give conflicting results. A number of issues must be addressed: is the sample size adequate, are enrolment criteria and statistical methods stringent enough to allow the extrapolation of results to the general population, and are the results generalizable.[13]

We are dealing with inferential statistics, which is the art of using samples to reach conclusions (of specific reliability) about populations. With respect to this problem, some points should be highlighted. By convention, a difference between groups is usually reported as statistically significant if the P value is 0.05 or less: this means that the probability that the findings are due to chance is less than 5% (1/20). There is considerable debate among statisticians and epidemiologists on the importance of the P value:[12, 14, 15] this figure is in fact heavily dependent on sample size and may not give any information about the actual clinical relevance of the results. In an editorial published in 1968 in the *New England Journal of Medicine* it was stated: "*Significant, being one of the words that mean everything or nothing, is too convenient for the* *medical writer (...). In its vulgar sense, unrelated to statistical manipulations, it comfortably serves ambiguity (...). It permits an author to describe his findings as significant (...) without requiring him to use an acceptable measure of that importance".*[15]

According to Altman and Bland, the interpretation of non significant findings (P > 0.05) may be particularly misleading: lack of significance does not mean that there is no difference between groups, but only that there is no evidence of a difference.[16] In other words, the P value is just an acceptable level of the risk of being wrong.

When we observe a difference between treatment and control arms that may be clinically relevant but does not achieve statistical significance, there is the risk of rejecting a potentially useful therapy. In this case we have two options: the first is to decrease data dispersion by increasing sample size; the second (lazier) option, is to pool the data with other similar trials (*i.e.* meta-analysis). In highly selected situations, it may be possible to use a one-tail test instead of the usual two-tail test. The use of two-tail test is not a dogma. A one-tail test can be considered whenever an effect can only occur in one direction: an accepted indication is the preclinical screening of active molecules, when we are interested only in picking up effective compounds; other examples can be found in rare diseases, that for some aspects can be compared to ARDS. In the clinical setting the conditions for the use of one-tail test are rarely met.

As already stated, the level of significance (P value) is arbitrary and does not tell us anything about the clinical relevance of an observed difference or about cause-effect relationships. These considerations are particularly relevant in the case of very large RCTs, as shown in a recent paper by Villar *et al.*[12] It is well known that the most relevant data to describe the clinical impact of a treatment are the absolute risk reduction and the 95% confidence interval (CI) for this difference. Another important parameter is the number needed to treat (NNT, calculated as 1/absolute risk reduction), which describes the number of patients that need to be treated to prevent one adverse event. Unfortunately, in most clinical trials the NNT is not reported and the relative risk reduction is used instead of the absolute risk reduction: this may lead the clinician to overestimate the efficacy of a new treatment.[12]

Other problems arise from the widespread use of subgroup analysis.[17-19] In an attempt to improve the prediction of the individual effect, investigators often analyze data from trials by subgroups within which responses are expected to be similar but between which responses are suspected to differ. Unfortunately, the interpretation of results in subgroups is fraught with inferential problems. If for example we consider a trial comparing a worthless treatment with a control and we divide patients into 10 mutually exclusive subgroups, we will have a 20% chance (P < 0.05) to conclude that treatment is significantly better than control in at least one subgroup, but also a 20% chance to obtain the opposite results. To avoid these pitfalls, the statistical analysis plan must be clearly predefined; subgroup findings should be exploratory and only exceptionally used to affect a trial's conclusions.[19]

Similarly, the results of *interim* analyses leading to premature interruption of a trial can be particularly misleading and should be interpreted with great caution. One of the roles of Data Safety and Monitoring Boards (DSMB) is to carefully assess *interim* to ensure the safety of participating subjects.[20] Over the past decade the use of DSMBs has increased substantially, mainly due to the growing number of trials with mortality as the primary end-point and to the greater awareness of potential biases that can result with early stopping a study. The critical challenge is to determine when *interim* data can be considered reasonably conclusive, thus justifying an early termination of a trial because of evidence that one treatment has greater efficacy or causes greater harm than another. When interim data are analyzed multiple times, the use of a P value of 0.05 or less as the stopping criterion can lead to a false conclusion, due to the effect of multiple tests of significance.[21] Moreover, the results of clinical trials that are stopped early are likely to exaggerate the magnitude of a treatment effect.[22] In a recent paper, Montori *et al.* examing a large number of trials prematurely stopped showed that the percentage of trials that were stopped early increased from 0.5% in the years 1990-1994 to 1.2% in the interval 2000-2004, with an average accrual of 63% of planned sample size.[23]

In the literature, there are some clear examples of large trials that at the *interim* analysis were close to being stopped because of a very strong trend favoring one treatment arm but that at the completion failed to show any significant difference between the 2 groups.[20, 24, 25] To overcome these problems statisticians have proposed a number of rules for *interim* analysis, a complete list of which is beyond the scope of this paper. We believe that the most reasonable approach is to use very stringent stopping criteria, so that trials are terminated early only when there is clear evidence that a treatment is better than another: for example, Peto *et al.* proposed that trials should be stopped only if the P value is less than 0.001 at any *interim* analysis.[26]

Despite these considerations, there is no doubt that RCTs are the gold standard in evaluation of the efficacy of clinical interventions. When large RCTs are not available, clinicians usually take into account meta-analyses as the preferred source of evidence to guide their clinical behaviour. The intrinsic strengths and weaknesses of meta-analyses have been the subject of several reviews [27, 28] and will not be discussed here. Of note, the ability of meta-analyses to predict the results of RCTs on the same subject is quite limited.[29] For example, Le Lorier *et al.* compared the results of 12 RCTs and 19 meta-analyses addressing the same questions and showed that the outcomes of the RCTs were not predicted 35% of the time by the meta-analyses previously published on the same subject.[30]

Summarizing this section, we believe that the use of EBM in the intensive care setting has been a useful development, but unfortunately many therapies have not been rigorously evaluated. We need large and unbiased RCTs to translate evidence coming from research into current clinical practice. Results from clinical trials must be interpreted critically and combined with evidence from other forms of research. EBM is not a "cookbook medicine" applicable in all situations, but rather a clinical decision making process that relies on the judicious use of results from clinical studies combined with clinical experience, clinical reasoning, understanding of pathophysiology and of patients' preferences. It is particularly important to stress the concept that EBM must not be viewed as an attempt to denigrate clinical experience and clinical intuition nor to de-emphasize the importance of understanding the physiopathologic mechanisms of diseases.[31] As stated by Brochard *et al.* "physiological and clinical understanding are needed as foundation stones before pillars of evidence can be erected".[10]

Importance of physiological studies

From the previous discussion, the importance of physiological studies should become evident. Studies exploring physiological issues are needed to increase our understanding of etiology and pathophysiology of diseases. This should also help to obtain more precise definitions of the syndromes associated with critical illness.

A strong physiological rationale is necessary for the design of meaningful clinical trials. In this sense the role of physiological studies can be compared to that of phase I-II studies in the process of drug development: these studies provide information on safety, dosage and spectrum of activity of the experimental drug that are essential for the design of phase III (randomized) trials.

For example, if the hypothesis behind the use of higher positive end expiratory pressure (PEEP) levels is that it improves recruitment, a randomized trial exploring two different PEEP levels in ARDS would have a greater chance of success if it was possible to predict the potential for recruitment in individual patients, an issue that can be addressed only exploring the physiological mechanisms of lung recruitment. By their nature, physiological studies should be performed in specialized and selected units, whereas participation in RCTs addressing clinical questions can be extended to a larger number of ICUs.

Unfortunately, physiological studies are becoming more difficult to perform even in highly specialized units, mainly due to economic and ethical issues. As suggested above, to overcome also these difficulties a promising strategy should be their nesting in, and/or articulation with, early (phase II) or therapeutic (phase III and IV) trials, so that their results could be more easily linked and finalized to the understanding and/or the better qualification of clinical-therapeutic results.

Conclusions

Critical care is a relativdy young branch of medicine in continuous and rapid evolution. Many of the therapeutic interventions in this field are still waiting for a rigorous evaluation. The only way to generate and increase our knowledge is research. The most powerful research tools to evaluate the efficacy of a therapeutic intervention are RCTs. In critical care, we need a cultural revolution to consider RCTs as the preferredway of conducting clinical trials: they should become the routine rather than an exception.

As discussed above, there are some intrinsic and structural problems (lack of reliable nosography, difficulties in identifying adequate control groups, concomitant use of different therapies, use of mortality as primary endpoint) that make the design and conduct of RCTs in our discipline particularly difficult. Enrolment of large numbers of patients is also hindered by the limited number of ICU beds. These considerations partially explain why so many clinical trials in our field give negative results, *i.e.* they fail to demonstrate a statistically significant difference between treated patients and controls.

Every effort should be done to overcome these problems. In particular: 1) we need better definitions of the common critical care syndromes (*e.g.* sepsis, ARDS ...); 2) we also need to increase the number of physiologic studies, to increase our understanding of pathophysiology and etiology of diseases; and 3) the creation of a large network(s) of research ICUs (ideally with comparable level of patient care) to carry out randomized clinical trials is necessary to increase the enrolment of adequate numbers of patients, such as occurs in the oncology and cardiovascular setting.

References

1. Ware LB, Matthay MA. The acute respiratory distress syndrome. N Engl J Med 2000;342:1334-49.
2. Villar J, Perez-Mendez L, Kackmarek RM. Current definitions of acute lung injury and the acute respiratorydistress syndrome do not reflect their true severity and outcome. Intensive Care Med 1999;25:930-5.
3. Parshuram CS, Kavanagh BP. Positiveclinical trials – understand the control group before implementing the result. Am J Respir Crit Care Med 2004;170:223-6.
4. Silverman HJ, Miller FG. Control group selection in critical care randomized controlled trials evaluating interventional strategies: an ethical assessment. Crit Care Med 2004;32:852-7.
5. 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;342:1301-8.
6. Ferguson ND, Kackmarek RM, Chiche JD, Singh JM, Hallett DC, Mehta S *et al.* Screening of ARDS patients using standardized ventilator settings: influence on enrolment in a clinical trial. Intensive Care Med 2004;30:1111-6.
7. Ware LB. Prognostic determinants of acute respiratorydistress syndrome in adults: impact on clinical trial design. Crit Care Med 2005;33 Suppl:S217-22.
8. Petros AJ, Marshall JC, Van Saene HKF. Should morbidity replace mortality as an endpoint for clinical trials in intensive care? Lancet 1995;345:369-71.
9. Evidence-Based Medicine Working Group. Evidence-Based Medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420-5.
10. BrochardL, Mancebo J, Tobin M. Searching for evidence: don't forget the foundations. Intensive Care Med 2003;29:2109-11.
11. Playfor S, Jenkins I, Boyles C, Choonara I, Davies G, Haywood T *et al.* United Kingdom Paediatric IntensiveCare Society Sedation, Analgesia and Neuromuscular Blockade Working Group. Consensus guidelines on sedation and analgesia in critically ill children. Intensive Care Med 2006;32:1125-36.
12. Villar J, Perez-Mendez L, Aguirre-Jaime A, Kackmarek RM. Why are physicians so skeptical about positive randomized controlled clinical trials in critical care medicine? Intensive Care Med 2005;31:196-204.
13. Holmberg L, Baum M. Can results from clinical trials be generalized? Nat Med 1995;1:734-6.
14. Evans SJW, Mills P, Dawson J. The end of the p value? Br Heart J 1988;60:177-80.
15. Anonymous. Significance of significant. N Engl J Med 1968;278:1232-3.
16. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
17. Lord SJ, Gebski VJ, Keech AC. Multiple analyses in clinical trials: sound science or data dredging? Med J Aust 2004;181:452-4.
18. Cook DI, Gebski VJ, Keech AC. Subgroup analyses in clinical trials. Med J Aust 2004;180:289-91.
19. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064-9.
20. Slutsky AS, Lavery JV. Data Safety and Monitoring Boards. N Engl J Med 2004;350:1143-7.
21. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.
22. Pocock Sj, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Control Clin Trials 1989;10 Suppl:S209-21.
23. Montori VM, Deveraux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M *et al.* Randomized trials stopped early for benefit: a systematica review. JAMA 2005;294:2003-9.
24. Wells RJ, Gartside PS, McHenry CL. Ethical issues arising when interim data in clinical trials is restricted to independent data monitoring committees. IRB 2000;22:7-11.
25. 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.
26. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I – Introduction and design. Br J Cancer 1976;34:585-612.
27. Horwitz RI. “Large-scale randomized evidence: large, simple trials and overview of trials”: discussion: a clinician’s perspective on meta-analyses. J Clin Epidemiol 1995;48:41-4.
28. Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. J Clin Epidemiol 1995;48:71-9.
29. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomized controlled trials. Lancet 1995;345:772-6.
30. Le Lorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;337:536-42.
31. Laupacis A. The future of evidence-based medicine. Can J Clin Pharmacol 2001;8 Suppl A:6A-9A.

Received on June 20, 2007 - Accepted for publication on June 26, 2007.

Address reprint requests to: L. Gattinoni, Dipartimento di Anestesia, Rianimazione e Terapia del Dolore, Fondazione IRCCS, Ospedale Maggiore Policlinico, Mangiagalli, Regina Elena, Via Francesco Sforza 35, 20122 Milan, Italy. E-mail: gattinon@policlinico.mi.it