Beautiful small: Misleading large randomized controlled trials? The example of colloids for volume resuscitation

Christian J Wiedermann, Wolfgang Wiedermann[1,2]

Department of Internal Medicine, Central Hospital of Bolzano, Teaching Hospital of the Medical University of Innsbruck, Bolzano, Italy, [1]Department of Psychology, Unit of Quantitative Methods, University of Vienna, Vienna, Austria, [2]Department of Educational, School and Counseling Psychology, College of Education, University of Missouri, Columbia, MO, USA

Abstract

In anesthesia and intensive care, treatment benefits that were claimed on the basis of small or modest-sized trials have repeatedly failed to be confirmed in large randomized controlled trials. A well-designed small trial in a homogeneous patient population with high event rates could yield conclusive results; however, patient populations in anesthesia and intensive care are typically heterogeneous because of comorbidities. The size of the anticipated effects of therapeutic interventions is generally low in relation to relevant endpoints. For regulatory purposes, trials are required to demonstrate efficacy in clinically important endpoints, and therefore must be large because clinically important study endpoints such as death, sepsis, or pneumonia are dichotomous and infrequently occur. The rarer endpoint events occur in the study population; that is, the lower the signal-to-noise ratio, the larger the trials must be to prevent random events from being overemphasized. In addition to trial design, sample size determination on the basis of event rates, clinically meaningful risk ratio reductions and actual patient numbers studied are among the most important characteristics when interpreting study results. Trial size is a critical determinant of generalizability of study results to larger or general patient populations. Typical characteristics of small single-center studies responsible for their known fragility include low variability of outcome measures for surrogate parameters and selective publication and reporting. For anesthesiology and intensive care medicine, findings in volume resuscitation research on intravenous infusion of colloids exemplify this, since both the safety of albumin infusion and the adverse effects of the artificial colloid hydroxyethyl starch have been confirmed only in large-sized trials.

Key words: Bias, consistency, design, meta-analysis, randomized controlled trial, sample size determination

Introduction

Evidence-based medicine is the integration of best research evidence with clinical expertise and patient values.[1] Practical medicine depends on physician experience and is guidance-oriented. Many clinicians, however, feel unqualified to critically appraise the medical literature.[2] Clinicians more or less follow recommendations knowing that guideline recommendations are subject to the influence of intense marketing strategies of pharmaceutical, diagnostics and device industries.[3]

Address for correspondence: Prof. Christian J. Wiedermann,
Department of Internal Medicine, Central Hospital of Bolzano,
Lorenz-Böhler Street 5, 39100 Bolzano, Italy.
E-mail: christian.wiedermann@asbz.it

Access this article online

Quick Response Code:

Website:
www.joacp.org

DOI:
10.4103/0970-9185.161680

Multiple study sites are required for large randomized controlled trials (RCTs) leading to heterogeneity due to differences in each site's standards of care. Pragmatic trial design with simplified inclusion criteria is a typical consequence and turns out as a potential limitation when it comes to the implementation of study results in daily clinical practice. In the field of volume resuscitation, the generalizability of results of large RCT results has, therefore, been questioned, and greater reliance on small RCT may appear attractive to some. This narrative review article is aimed at helping doctors better understand the meaning of the results of RCTs and the role of the design and size of studies that are appraised.

Randomized Controlled Trials and the Hierarchy of Studies

Guidelines are made for the general public, but must be tailored to the individual patient. For this to be successful, besides the practitioners' personal experience from routine clinical practice, medical knowledge gained from independent, competent evaluation of results from best research evidence is also critical. For this reason, the importance of self-directed learning through reading is emphasized.[4] Thus, for clinical medicine, and for all practicing physicians, it has become necessary to understand the methodology of clinical trials in order to critically appraise clinical study results and to be able to integrate best research evidence with clinical expertise and patient values.

First of all, it is necessary to understand the various categories of study types.[1] The two general categories, experimental and observational, are based on whether the investigator assigns the exposures or not. Experimental trials are subdivided into randomized and nonrandomized. Systems to stratify evidence by quality have been developed.[5] These are based on a hierarchy of clinical studies. Evidence obtained from at least one properly designed RCT is at the highest level in treatment and prevention research. RCTs are ubiquitous in clinical research and are scientific routine today. Although the concept of randomization can be traced back to early works of the statistician Fisher[6,7] (for a discussion of historical developments see, e.g., ref.[8]), the history of RCTs is relatively short. Only about 60 years ago, the first study of its kind was performed according to modern criteria.[9] Since then, this form of clinical research has undergone rapid development.[10]

Sample size and trial design

A well-designed trial in a homogeneous patient population with high event rates can yield conclusive results even if small.[11] In fact, as recently as 10 years ago, a trial with a few hundred patients was considered large in intensive care or anesthesia settings. Modern studies in these fields now-a-days need to include several thousands of patients to be considered large enough to be meaningful. Reasons for this are obvious: The size of the anticipated effects of therapeutic interventions, in particular of drugs in anesthesia or intensive care, is generally low in relation to relevant endpoints. Effect sizes measured in successful trials such as relative risk or risk ratio (RR) reductions are mostly in the range of 20-25%. This is because the trials usually investigate interventions in disease complications that are multifactorial. Furthermore, in contrast to the example, hereditary diseases or well-characterized disease entities, the patient population is particularly heterogeneous. Modern interventions usually are highly targeted to affect single pathophysiological steps in the development of disease complications which cannot be expected to correct a multifactorial pathophysiology more strongly than that usually resulting in an RR reduction of about 20-25%.

On the other hand, large studies are not always more reliable. Trial design is variable that needs also to be taken into account. A current controversy related to colloids for volume resuscitation concerns large "pragmatic" trials that do not enforce strict fluid protocols.[12,13] Specific design weaknesses of particular large trials need to be taken into account such as heterogeneous interventions, confounding concomitant treatments and baseline risk imbalances.[14] Another design feature of importance is the length of follow-up. Short follow-up was a major problem in some recent meta-analyses of colloids.[15,16]

Meta-Analysis of Randomized Controlled Trials

Logistical, financial and administrative reasons render "large" trials notoriously difficult to carry out. Evidence base for many interventions in intensive care and anesthesia, therefore, consists largely of "small" studies, requiring the use of statistics for help in clinical action. Meta-analysis aims to combine the results of individual studies in order to increase their analytical power usually given in systematic reviews. By statistically combining the outcomes of similar studies, estimates of treatment effects can be made more precise, and it can be assessed whether treatment effects are similar in similar situations. Variability arises from sources that are intrinsic to the patients leading to population differences in multi-center RCTs; sources of variability that are external to the patient include ways in which patients are recruited and managed. The decision about whether or not the results of individual studies are similar enough to be combined in a meta-analysis will impact on the validity of the result.[17]

The colloid controversy

In colloid resuscitation research, the meta-analysis by Van Der Linden *et al.*[15] has been criticized for combining studies that were considered too different.[18] The evolution from meta-analyses based on small studies to results of more meaningful, larger studies in anesthesiology and critical care medicine and their impact on updated meta-analysis can be best illustrated with the following example: In response to a Cochrane group meta-analysis of volume resuscitation and expansion studies in the critically ill, there were dramatic changes in guidelines on the use of albumin solutions. This and subsequent meta-analyses of preferentially small studies have resulted in contradictory recommendations.[19,20]

Originally in the treatment of hypovolemia and hypoalbuminemia, use of the natural colloid albumin had been recommended for critically ill intensive care unit (ICU) patients. In 1998, a Cochrane analysis for albumin administration in patients in intensive care was published in the British Medical Journal.[21] This systematic review and meta-analysis of 24 selected, small RCTs led to the conclusion of a significant increase in patient mortality when patients were treated with albumin rather than with other resuscitation fluids.[21] The number of patients included in the studies ranged between 14 and 219, making them consistently small studies. The sum effect showed a significantly increased mortality of almost 70% in patients receiving albumin. The results of this meta-analysis led to a dramatic reduction in the general use of albumin solutions for colloidal volume resuscitation. In Europe, albumin was replaced almost entirely with artificial colloids, in particular with hydroxyethyl starch (HES) solutions.[22]

Only a few years later, the Cochrane meta-analysis suggesting increased mortality in albumin-treated critically ill patients was disproved by an updated meta-analysis, in which 55 trials involving 3504 randomly assigned patients had been included and 525 deaths occurred.[23] The median number of patients who underwent randomization per trial was 52 (range, 10-300), and significant small-trial bias became evident. An observed small-trial bias favored the control group, since RR was substantially lower in large trials than in small trials.[23] The Cochrane albumin meta-analysis proved irreproducible not only because of small-trial bias, which was a contributing factor, but also because of the assembly of a small biased subset of relevant randomized trials.[24] Later, based on new evidence from larger trials, results of this updated meta-analysis were confirmed.[25]

Whether the administration of albumin in the critically ill increases mortality or other relevant adverse events was finally clarified in "The Saline versus Albumin Fluid Evaluation" study, a large RCT with approximately 7000 patients.[25] Based on results of this sufficiently large RCT, the hypothesis from small RCT results was rejected and albumin administration for volume resuscitation or correction of hypoalbuminemia is now considered safe. This is an example where a meta-analysis put a question on the table and triggered a clinical trial to achieve "definitive proof."

More recently, additional evidence from large RCTs further confirmed the safety of albumin infusions.[26,27] In all three large RCT's, evaluating the use of albumin in adults with severe sepsis,[25-27] mortality was lower in the group allocated to albumin than crystalloid, and the pooled mortality reduction for all three trials was statistically significant (pooled RR: 0.92; 95% confidence interval [CI]: 0.84-1.0; $P = 0.046$).[28]

Recent large-scale RCTs have also helped settle long-debated questions about the safety of HES solutions. In the Scandinavian Starch for Severe Sepsis/Septic Shock (6S) trial of 798 patients, HES 130/0.42 increased mortality and the need for renal replacement therapy.[12] Utilization of renal replacement therapy was increased by HES 130/0.4 in the Crystalloid Versus Hydroxyethyl Starch Trial (CHEST) of 7000 ICU patients although no significant effect on mortality was observed.[13] In the Colloids Versus Crystalloids for the Resuscitation of the Critically Ill (CRISTAL) randomized trial of 2857 ICU patients, no effect of colloids, predominantly HES, on 28 days mortality or need for renal replacement therapy was observed.[14] In an exploratory analysis of the CRISTAL data, mortality was lower in the colloids group at 90 days.

These trials also exemplify methodological issues of potential importance beyond size and statistical power. One persistent criticism of studies purporting to demonstrate the safety of HES has been short follow-up. Short follow-up was a major problem in some recent meta-analyses of colloids.[15,16] Support for this criticism comes from the 6S trial, in which a significant effect on mortality could be shown at 90 days but not 28 days. Additionally, 6S, CHEST and CRISTAL were all pragmatic trials in which fluid management strategies were at the discretion of the attending clinicians rather than in accordance with a strict protocol. It remains possible that the implementation of particular protocols might modify outcomes although this would need to be demonstrated.

Lastly, even large trials must be conducted in such a manner to minimize biases. For instance, in CRISTAL the attending clinicians were not blinded to treatment assignment. Furthermore, in that trial there was striking evidence of flawed randomization, since patients receiving crystalloids in the 12 h before ICU admission were 33% more likely to be randomized to colloids ($P = 10^{-7}$), while those receiving prior colloids were 15% more likely to be assigned to crystalloids ($P = 0.001$). It is extremely unlikely such imbalances could have arisen by chance.

Scandinavian Starch for Severe Sepsis/Septic Shock and CHEST in particular has been decisive in prompting regulatory actions. Both the European Medicines Agency (www.ema.europa.eu/docs/en_GB/document_ library/Referrals_document/Solutions_for_infusion_containing_hydroxyethyl_starch/European_Commission_final_decision/WC500162361.pdf; accessed 8 July 2014) and the US Food and Drug Administration (www.fda.gov/BiologicsBloodVaccines/SafetyAvailability/ucm358271.htm; accessed 8 July 2014) have decided that HES solutions should no longer be used in critically ill patients, including those with sepsis. In Europe, use of HES to treat hypovolemia caused by acute blood loss will still be permitted; however, new risk minimization procedures are to be required, including monitoring of renal function for 90 days after HES infusion.

Event Rates Determine if a Study is Large Enough

Clinically important study endpoints such as death, sepsis, or pneumonia are dichotomous, infrequently occur and are, therefore, best studied in large RCTs. In anesthesia and in intensive care medicine, they typically occur in a setting where the patient has comorbidities and multiple organ dysfunction, and it is often difficult to determine their impact on clinically important outcomes of individual treatments. Depending on the inclusion criteria of RCTs, the frequency of such events is estimated to be within the range of 2-10%. Less important clinical endpoints such as significant changes in vital, hemodynamic or laboratory parameters, in contrast, occur more frequently although in many cases these surrogate parameters are not validated. The rarer endpoint events occur in the study population; that is, the lower the signal-to-noise ratio, the larger the RCTs must be to prevent random events from being overemphasized and to reach statistically meaningful study results. In the interpretation of a RCT's impact on evidence-based medicine, therefore, the event rate of the sample size must be sufficiently high to give the study the appropriate power to examine the hypothesis of interest. Sample size determination on the basis of event rates, clinically meaningful RR reductions and actual patient numbers studied are among the most important characteristics of RCTs when interpreting study results.[29]

The goal for most clinical research questions in anesthesia and intensive care is to prove that with the tested intervention, a RR reduction of 20-25% can be reached. For example, if a control group's mortality rate during a defined observation period is 8%, an RR reduction of 25% would be a reduction of mortality from eight to six deaths out of 100 participants, two deaths - A number so small that in a study size of 200 patients (100 patients per group) the influence of random deaths would be much too large. Statistical power refers to the probability of correctly rejecting the null hypothesis of a zero effect. Thus, the higher the statistical power, the higher is the probability of correctly rejecting the null hypothesis. The left panel of Figure 1 shows the total sample sizes needed in a study as a function of RR reduction and three commonly accepted power values (80%, 90%, and 95%) assuming a baseline rate of 8%. Sample sizes were calculated using the approach described in Agresti (p. 242).[30] Overall, the number of study subjects increases with the desired power and declines with the size of the RR reduction. For example, if the power of the RCT should be 80% (i.e., in 100 repetitions of the study, a significant result would be confirmed 80 times, given that the effect truly exists), 6450 patients would be required to prove a 20% RR reduction when the baseline event rate is 8%.

Any study that fails to perform a meaningful sample size calculation and to include the predicted number of patients runs the risk to be over- or under-powered. Over-powered trials refer to data situations in which very large samples lead to statistically significance results even in case of extremely small differences. Under-powered trials describe situations in which a study fails to detect a (truly existing) effect because sample sizes are too small. In both cases, valid conclusions cannot be drawn any more. If a RCT does not find a statistically significant effect, this may be due to the absence of a true effect or the study population may have been too small. The absence of a true effect can only be concluded if the trial is sufficiently powered.

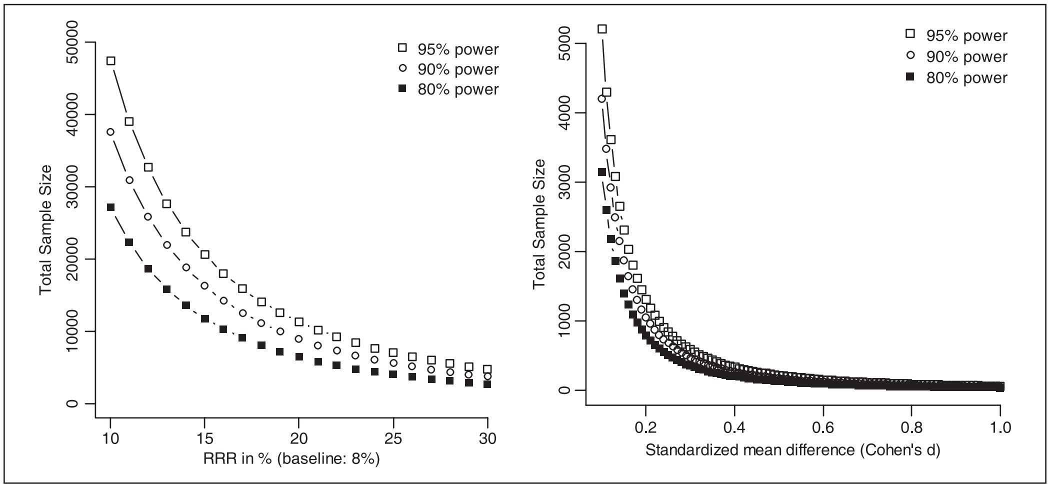

Figure 1: Number of study subjects needed as a function of effect size and power assuming a 5% significance level (left panel: Dichotomous variables assuming a baseline event rate of 8%; right panel: Continuous variables assuming homogeneous variances)

The Fragility of Small Studies in Anesthesia and Intensive Care Medicine

Small RCTs with statistically significant results often prove incorrect particularly when the study population is heterogeneous as in the ICU or the operating room. If repeated, results often do not reach statistical significance. One reason for this is that even the smallest changes in event frequencies, P-values may change substantially from significant ($P < 0.05$) to nonsignificant ($P > 0.05$). For example, in a 400 patients study (200 subjects in the intervention and 200 subjects in the placebo group), with 10 events in the control group and two events in the treatment group, we obtain a χ^2-value for the 2×2 contingency table of 4.21 with one degree of freedom (df). Thus, the difference would be statistically significant with a $P = 0.04$; if the event rate reduction were from 10 to 3 instead of 10 to 2, one obtains $\chi^2 = 2.86$ (df = 1) and a nonsignificant $P = 0.09$. The significant impact on results of statistical analyses of minimal changes in trial event rates is seen when the event rates are low and, thus, contribute to the fragility of small RCTs. In consequence, statistically significant $P < 0.05$ is not necessarily of clinical relevance. Formally, a $P < 0.05$ means that the probability of obtaining a difference in study groups at least as extreme as the one actually observed is smaller than 5%, while assuming that the true difference is zero. Obviously, this probability does not make any statements on the clinical relevance of the observed difference. In analogy, example, a RR of 45% with a 95% CI of 4-90% means that in 95 out of 100 repetitions of the study, the confidence range overlaps the true RR. Thus, the confidence level (e.g., 95%) determines the degree of (un) certainty. To derive clinically relevant conclusions from such study results, the CI is therefore of critical importance. If the CI is too wide, it is very likely that the RCT size was too small.[29]

Sample Size in Meta-analyses

The term "small study effect" describes a tendency for small trials to report greater treatment benefits than large trials in the same meta-analysis. Sample size varies greatly among trials even within a meta-analysis investigating the same question.[31] Using a single threshold to distinguish between small and large trials is not straightforward because required trial size also depends on the medical condition studied. If the condition characteristics enables identification of a less heterogeneous patient population, such as in hereditary conditions or certain liver diseases, the signal-to-noise ratio may be reliably high and studies less fragile even if small. The distinction between sufficient and insufficient trial size may be better reflected by the size of the CI.[29]

When assessing the influence of trial sample size on treatment effect estimates in a large collection of meta-analyses of various medical conditions and interventions, treatment effect estimates differed within meta-analyses solely based on trial sample size, and stronger effects were seen in the smaller studies.[31] Therefore, robustness of the conclusions of a meta-analysis should be assessed by checking whether the result for the overall meta-analysis agrees with the results for the quarter of the largest. Pooled results require cautious interpretation when this is not the case.

The combined treatment effect may not be the best estimate of the true treatment effect. Hence, notion of the 'small study effect' raises the question whether all available evidence should be included in meta-analyses, because it could lead to seemingly more beneficial results. Not only should individual studies, in particular if they are small, be similar enough to be combined in a meta-analysis, the trials' hypotheses should also be based on similar pathophysiological rational and biological plausibility. Among the most important reasons for the "small study effect" is that smaller studies are more prone to publication bias due to the tendency for publication of reports of studies with significant rather than nonsignificant results.[32]

Empirically, meta-analyses usually agree with large RCTs.[33] Consistent treatment effects among small studies as summarized in a meta-analysis, can yield reliable results that are likely to be confirmed in large RCTs.[34] Indeed, consistency of treatment effects in large trials may allow conclusions to be drawn by meta-analysis that are not demonstrable in any of the individual large trials.[25-27] So consistency is another important variable beyond size *per se*.

Extreme Homogeneity — Additional Insights on Meta-analyses

Extreme between-study homogeneity provides useful insights on a meta-analysis and its constituent studies. Smaller trials are not only more likely to overestimate true treatment effects but are also more prone to reporting bias and fraud.[35-37] Recently in the field of anesthesiology and intensive care medicine, a total of 90 publications of small single-center clinical trials by Boldt *et al*. were retracted,[38] 88 of which in 2011 because of failure to involve ethics committees as well as fraud,[39] and additional two in 2014 because data fabrication was confirmed.[40,41] For the latter two, extreme homogeneity of treatment effects had already been observed in 2006 in a report suggesting that they might be fraudulent.[42] In a meta-analysis examining whether there is a difference in mortality with albumin or plasma protein fraction versus hydroxylethyl starch for fluid resuscitation,[43] there was overall extreme between-study homogeneity. However, 5 of the 10 studies with any events during follow-up (of a total of 20 studies in the meta-analysis) were performed apparently by Boldt *et al*. These five studies with almost identical designs accounted for 70% of the total number of events in the meta-analysis and they all came from the same center using the same stratification and same allocation schedule.[42] Biased and fraudulent study publications are often single-center with authors from only one department, but the patient population having characteristics that require the involvement of multiple departments in care. In addition, information on recruitment periods is frequently not provided.

Surrogate Endpoint Studies

Surrogate markers are used when the number of primary endpoints is very small, thus making it impractical to conduct a clinical trial to gather a statistically significant number of endpoints. Parameters are often biomarker or physiological with a continuous outcome measure intended to substitute for clinically meaningful endpoints. In order to be informative for clinical decision making, validation of surrogate endpoints is important requiring extensive research including RCTs with important clinical outcome.

With a continuous outcome measure, each person in a trial contributes information. Sample size calculations in such studies are based on the standardized mean differences such as Cohen's d we would wish reliably to detect. Cohen[44] suggested that d = 0.2, 0.5, and 0.8 refer to small, medium, and large effects, respectively. For a minimum difference of d = 0.5, example, a sample size of 2 × 64 study subjects would be sufficient using a nominal significance level of 5% and the desired power of 80%. In contrast, for a minimum difference of d = 0.2 (which is considered to be a small effect) 2 × 394 study subjects are necessary to achieve a power of 80% (again assuming a nominal significance level of 5%). The right panel of Figure 1 shows the total number of study subjects needed as a function of Cohen's d, the desired power, and a significance level of 5%. Again, total sample sizes increase with the power and decrease with the magnitude of effect size. Approximate calculations of this kind can help determine whether the study is "large enough" to support a firm conclusion. More detailed sample size calculation can be performed using software tools that can be downloaded free of charge, example, from http://dceg.cancer.gov/tools/design/power, http://www.gpower.hhu.de/, or http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize.

Conclusions

In the last decade, clinical research in anesthesia and intensive care medicine has seen a change from mostly small studies to an increasing number of large RCTs. Experience from studies on volume resuscitation illustrates problems arising from over-interpretation of treatment effects seen in underpowered studies including meta-analyses. Small trials are misleading because they are often underpowered for important endpoints, fragile, and prone to bias. Meta-analyses of small trials may be informative when consistent with results of large trials. Large trials are able to evaluate small but clinically important treatment effects, and give better estimations of true treatment effects [Table 1]. Compared to small trials, they are less fragile. However, they cost more, require longer data-acquisition periods and as they need to be carried out at multiple centers, are logistically more complex to manage.

Table 1: Benefits and limitations of large sample size in RCTs

Benefit	Limitation
Ability to evaluate small, but clinically important, treatment effects	Patient heterogeneity
Ability to establishes effective widely practicable therapy for common diseases	Greater cost
Better estimation of treatment effect	May require multiple centers and longer data-acquisition periods
Result less fragile	Impossible for rare diseases and rare outcomes
Highest level of evidence	

RCTs = Randomized controlled trials

References

1. Grimes DA, Schulz KF. An overview of clinical research: The lay of the land. Lancet 2002;359:57-61.
2. Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA 1995;274:700-5.
3. Wiedermann CJ. Bioethics, the surviving sepsis campaign, and the industry. Wien Klin Wochenschr 2005;117:442-4.
4. Murad MH, Coto-Yglesias F, Varkey P, Prokop LJ, Murad AL. The effectiveness of self-directed learning in health professions education: A systematic review. Med Educ 2010;44:1057-68.
5. US Preventive Services Task Force. Guide to Clinical Preventive Services: Report of the U.S. Preventive Services Task Force. DIANE Publishing; 1989. p. xxiv. Available from: http://www.books.google.com/books?id=eQGJHgI_dR8C&pg=PR24. [Last accessed on 2015 Jan 01].
6. Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd; 1935.
7. Fisher RA. Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd; 1925.
8. Hall NS. R. A. Fisher and his advocacy of randomization. J Hist Biol 2007;40:295-325.
9. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948;2:769-82.
10. Sessler DI, Devereaux PJ. Emerging trends in clinical trial design. Anesth Analg 2013;116:258-61.
11. Sort P, Navasa M, Arroyo V, Aldeguer X, Planas R, Ruiz-del-Arbol L, *et al.* Effect of intravenous albumin on renal impairment and

mortality in patients with cirrhosis and spontaneous bacterial peritonitis. N Engl J Med 1999;341:403-9.

12. Perner A, Haase N, Guttormsen AB, Tenhunen J, Klemenzson G, Åneman A, *et al.* Hydroxyethyl starch 130/0.42 versus Ringer's acetate in severe sepsis. N Engl J Med 2012;367:124-34.
13. Myburgh JA, Finfer S, Bellomo R, Billot L, Cass A, Gattas D, *et al.* Hydroxyethyl starch or saline for fluid resuscitation in intensive care. N Engl J Med 2012;367:1901-11.
14. Annane D, Siami S, Jaber S, Martin C, Elatrous S, Declère AD, *et al.* Effects of fluid resuscitation with colloids vs crystalloids on mortality in critically ill patients presenting with hypovolemic shock: The CRISTAL randomized trial. JAMA 2013;310:1809-17.
15. Van Der Linden P, James M, Mythen M, Weiskopf RB. Safety of modern starches used during surgery. Anesth Analg 2013;116: 35-48.
16. Martin C, Jacob M, Vicaut E, Guidet B, Van Aken H, Kurz A. Effect of waxy maize-derived hydroxyethyl starch 130/0.4 on renal function in surgical patients. Anesthesiology 2013;118:387-94.
17. The Cochrane Collaboration. Combining studies — What is meta-analysis. Available from: http://www.cochrane-net.org/openlearning/html/mod12-2.htm. [Last accessed on 2014 Feb 09].
18. Takala J, Hartog C, Reinhart K. Safety of modern starches used during surgery: Misleading conclusions. Anesth Analg 2013;117:527-8.
19. Bunn F, Lefebvre C, Li Wan Po A, Li L, Roberts I, Schierhout G. Human albumin solution for resuscitation and volume expansion in critically ill patients. The Albumin Reviewers. Cochrane Database Syst Rev 2000;CD001208.
20. Roberts I, Blackhall K, Alderson P, Bunn F, Schierhout G. Human albumin solution for resuscitation and volume expansion in critically ill patients. Cochrane Database Syst Rev 2011;CD001208.
21. Cochrane Injuries Group Albumin Reviewers. Human albumin administration in critically ill patients: Systematic review of randomised controlled trials. BMJ 1998;317:235-40.
22. Finfer S, Liu B, Taylor C, Bellomo R, Billot L, Cook D, *et al.* Resuscitation fluid use in critically ill adults: An international cross-sectional study in 391 intensive care units. Crit Care 2010;14:R185.
23. Wilkes MM, Navickis RJ. Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials. Ann Intern Med 2001;135:149-64.
24. Wilkes MM, Navickis RJ. Does albumin infusion affect survival? Review of meta-analytic findings. In: Vincent JL, editor. Yearbook of Intensive Care and Emergency Medicine. Berlin: Springer-Verlag; 2002. p. 454-64.
25. Finfer S, Bellomo R, Boyce N, French J, Myburgh J, Norton R, *et al*. SAFE study investigators. A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 2004;350:2247-56.
26. Caironi P, Tognoni G, Masson S, Fumagalli R, Pesenti A, Romero M, *et al.* Albumin replacement in patients with severe sepsis or septic shock. N Engl J Med 2014;370:1412-21.
27. Charpentier J, Mira JP. Early albumin resuscitation during septic shock. Intensive Care Med 2011;37 Suppl 1:S115.
28. Wiedermann CJ, Joannidis M. Albumin replacement in severe sepsis or septic shock. N Engl J Med 2014;371:83.
29. Glasziou P, Doll H. Was the study big enough? Two café rules. Evid Based Med 2006;11:69-70.
30. Agresti A. Categorical Data Analysis. 2nd ed. Hoboken, NJ: Wiley and Sons; 2002.
31. Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: Meta-epidemiological study. BMJ 2013;346:f2304.
32. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: Comparative study of articles indexed in PubMed. BMJ 2010;340:c723.
33. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. JAMA 1998;279:1089-93.
34. LeLorier J, Grégoire 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.
35. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA 2004;291:2457-65.
36. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, *et al.* The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010; 340:c365.
37. Wiedermann CJ. Reporting bias in trials of volume resuscitation with hydroxyethyl starch. Wien Klin Wochenschr 2014;126:189-94.
38. Wise J. Boldt: the great pretender. BMJ 2013;346:f1738.
39. Elia N, Wager E, Tramèr MR. Fate of articles that warranted retraction due to ethical concerns: A descriptive cross-sectional study. PLoS One 2014;9:e85846.
40. Retraction note: Volume therapy in the critically ill: Is there a difference? Intensive Care Med 2014;40:145.
41. Notice of formal retraction of an article by Dr Joachim Boldt. Br J Anaesth 2014;112:397.
42. Ioannidis JP, Trikalinos TA, Zintzaras E. Extreme between-study homogeneity in meta-analyses could offer useful insights. J Clin Epidemiol 2006;59:1023-32.
43. Bunn F, Alderson P, Hawkins V. Colloid solutions for fluid resuscitation. Cochrane Database Syst Rev 2003;CD001319.
44. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale NJ: Lawrence Erlbaum Associates; 1988.

How to cite this article: Wiedermann CJ, Wiedermann W. Beautiful small: Misleading large randomized controlled trials? The example of colloids for volume resuscitation. J Anaesthesiol Clin Pharmacol 2015;31:394-400.
Source of Support: Nil, **Conflicts of Interest:** CJW has received honoraria from Baxter, CSL Behring and the Plasma Protein Therapeutics Association, an organization representing the private sector manufacturers of plasma-derived and recombinant analogue therapies. WW has no conflicts of interest to declare.