Outcomes and Statistical Power in Adult Critical Care Randomized Trials

Michael O. Harhay¹, Jason Wagner², Sarah J. Ratcliffe¹, Rachel S. Bronheim³, Anand Gopal¹, Sydney Green¹, Elizabeth Cooney¹, Mark E. Mikkelsen², Meeta Prasad Kerlin^{1,2}, Dylan S. Small⁴, and Scott D. Halpern^{1,2,3}

¹Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, ²Division of Pulmonary, Allergy, and Critical Care Medicine, ³Department of Medical Ethics and Health Policy at the Perelman School of Medicine, and ⁴Department of Statistics at the Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania

Abstract

Rationale: Intensive care unit (ICU)-based randomized clinical trials (RCTs) among adult critically ill patients commonly fail to detect treatment benefits.

Objectives: Appraise the rates of success, outcomes used, statistical power, and design characteristics of published trials.

Methods: One hundred forty-six ICU-based RCTs of diagnostic, therapeutic, or process/systems interventions published from January 2007 to May 2013 in 16 high-impact general or critical care journals were studied.

Measurement and Main Results: Of 146 RCTs, 54 (37%) were positive (i.e., the *a priori* hypothesis was found to be statistically significant). The most common primary outcomes were mortality ($n = 40$ trials), infection-related outcomes ($n = 33$), and ventilation-related outcomes ($n = 30$), with positive results found in 10, 58, and 43%, respectively. Statistical power was discussed in 135 RCTs

(92%); 92 cited a rationale for their power parameters. Twenty trials failed to achieve at least 95% of their reported target sample size, including 11 that were stopped early due to insufficient accrual/logistical issues. Of 34 superiority RCTs comparing mortality between treatment arms, 13 (38%) accrued a sample size large enough to find an absolute mortality reduction of 10% or less. In 22 of these trials the observed control-arm mortality rate differed from the predicted rate by at least 7.5%.

Conclusions: ICU-based RCTs are commonly negative and powered to identify what appear to be unrealistic treatment effects, particularly when using mortality as the primary outcome. Additional concerns include a lack of standardized methods for assessing common outcomes, unclear justifications for statistical power calculations, insufficient patient accrual, and incorrect predictions of baseline event rates.

Keywords: intensive care unit; critical care; intensive care; randomized clinical trial; randomized controlled trial

Randomized clinical trials (RCTs) are considered the “gold standard” for examining the efficacy and safety of medical interventions because randomization probabilistically balances treatment groups on measured and unmeasured baseline covariates, thereby mitigating selection biases (1).

Unfortunately, most published RCTs of critical care interventions that aim to reduce mortality have produced negative results (2–5), and even these reports may be overly optimistic because negative trials are less likely to be published and identified. Although several RCTs have revolutionized critical care practice (6–8),

the results of critical care trials on the whole have been so disappointing that some leaders in the field have suggested a renewed focus on nonexperimental study designs (9, 10).

However, truly negative trials are valuable because they prevent the use of interventions that are either costly but

(Received in original form January 10, 2014; accepted in final form April 23, 2014)

M.O.H. is supported by National Cancer Institute grant R01CA159932 awarded to S.D.H. Supported by Agency for Healthcare Research and Quality grant K08HS018406 (S.D.H.); National Heart, Lung, and Blood Institute grant K08HL116771 (M.P.K.); the Summer Undergraduate Research Minority Research program at the University of Pennsylvania (A.G. and S.G.); and an internship from the University of Pennsylvania Center for Bioethics (R.B.).

Author Contributions: Conception and design: M.O.H., J.W., S.J.R., E.C., M.E.M., M.P.K., D.S.S., and S.D.H. Acquisition of data: M.O.H., R.S.B., A.G., and S.G. Interpretation of data and drafting and revising manuscript: M.O.H., J.W., S.J.R., R.S.B., A.G., S.G., E.C., M.E.M., M.P.K., D.S.S., and S.D.H.

Correspondence and requests for reprints should be addressed to Michael O. Harhay, M.P.H., Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, 708 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. E-mail: mharhay@mail.med.upenn.edu

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 189, Iss 12, pp 1469–1478, Jun 15, 2014

Copyright © 2014 by the American Thoracic Society

Originally Published in Press as DOI: 10.1164/rccm.201401-0056CP on April 30, 2014

Internet address: www.atsjournals.org

nonbeneficial or even harmful (e.g., intensive insulin therapy [11] and hydroxyethyl starch [12, 13]). Furthermore, there are many reasons that trials may not demonstrate a treatment effect, including ineffective interventions, difficulty recruiting adequate sample sizes, postrandomization patient attrition, heterogeneous patient populations or treatment-effect heterogeneity, use of inappropriate outcomes, unreasonable assumptions (e.g., predicted effect sizes) used in power calculations, and/or smaller than appreciated attributable morbidity and mortality fractions (2–5, 14–18). Understanding an evidence base requires the ability to distinguish among these reasons so as to differentiate trials that are truly negative from those that may be falsely negative.

As a first step in enhancing understanding of clinical trials in adult critical care, we created a contemporary database of the design, analysis, and reporting of intensive care unit (ICU)-based RCTs. Herein, we describe the development of this database, the characteristics of RCTs published in the past 6 years with a specific focus on the outcome measures used, the quality of these RCTs using selected quality metrics, and the extent to which several issues germane to statistical power may contribute to trials' outcomes. Part of this research was presented at the 2013 American Thoracic Society conference (19).

Methods

A detailed summary of the generation of the analytic database and eligibility criteria is available in the online supplement. Briefly, a group of physicians, epidemiologists, and statisticians, guided by the 2007 CONSORT (20) (Consolidated Standards of Reporting Trials) statement, Jadad scale (21, 22), and prior work and commentaries on the topic (2–5, 10, 14–18, 23), identified RCT elements to be abstracted. We began our search for published RCTs in 2007, as this approximated the end of prior review periods (2, 3), and continued our search through May 2013. We examined only RCTs of diagnostic, therapeutic, or process and systems interventions among adult patients conducted in an ICU published in 16 prominent general or critical care journals (see Tables E1 and E2 in the online

supplement). Intermediate and physiologic outcomes were excluded because our goal was to identify trials testing interventions that were sufficiently mature as to be applied clinically as opposed to those that were primarily hypothesis generating.

Data Abstraction

Using the Research Electronic Data Capture (REDCap) platform hosted at the University of Pennsylvania (24), two investigators independently abstracted the primary and secondary outcomes, as reported by the authors in each trial, and the result (positive or negative) for each RCT. A superiority study was considered positive if the *P* value for the analysis of the primary outcome was less than 0.05 or the adjusted significance level after interim analyses, based on the reporting in each RCT. An equivalence or noninferiority study was considered positive if the difference between study arms fell between the predetermined margins (confidence intervals [CIs]) and met the equivalence or noninferiority hypothesis at the *P* value declared by the study's authors. When a study had more than two arms, outcomes were recorded from the control arm and the arm using an intervention of maximal dose or degree. Data were also extracted on study funding, type of intervention tested, target patient population, enrollment and retention, and statistical power.

To assess statistical power, we abstracted three specific methodological elements: (1) discussion of the power calculation used for the trial, (2) rationales for the parameters used in the sample size or power estimation, and (3) participant accrual. Discussion of the power calculation was defined as reporting the inputs used to calculate power or sample size, such as the baseline (control group) event rate and the expected treatment effect size for binary endpoints. The rationales for sample size or power estimation inputs could include prior research results, pilot studies, or other objective data. Participant accrual was tracked by assessing CONSORT diagrams, when available, indicating the number of patients screened, randomized, and ultimately analyzed (20).

The data abstractors achieved greater than 90% agreement for individual data elements, including primary and secondary outcomes, funding, target sample size, and

reason(s) for study exclusion. The first author adjudicated the discrepancies that arose. STATA 13 (StataCorp, College Station, TX) was used for database management and analysis.

Analyses to Assess Statistical Power

For each RCT with a binary outcome, we abstracted the predicted and observed risk difference on the absolute scale. We used the absolute, rather than relative, risk difference because absolute differences are used to determine the clinical significance of effects (25). For example, to calculate the number needed to treat, the absolute risk reduction is required. For negative trials, we evaluated whether (nonsignificant) reductions in the primary outcome of 3% or greater were identified. Our choice of a 3% cutoff is somewhat arbitrary but was chosen *a priori* based on the view that any treatment-associated absolute effect of this size would clearly be important to patients and that effects less than 3%, albeit potentially important, could also more easily be attributable to noise or random error. These assessments were limited to trials that reported power calculations, so as to enable uniform determinations of whether or not these RCTs were powered to document these effect sizes as significant.

We also explored the related phenomenon of “delta inflation bias” (3, 26), whereby unrealistically large treatment effects are predicted in power calculations, resulting in target sample sizes that may fail to detect clinically important differences. Specifically, we calculated the power for a range of detectable differences for each trial using the equation outlined in the online supplement. Using the actual enrolled sample sizes in the control and treatment arms, and the observed baseline mortality rate, we calculated the power of each study to observe a clinically significant, treatment-associated mortality reduction from 3 to 15%. The number of trials with 80% power was tallied for each treatment-associated mortality reduction of 3 to 15%. We then performed each calculation again using the predicted baseline mortality from the published RCT. Comparison of power obtained using the observed and predicted rates therefore highlights how often mispredictions of baseline event rates influence power. Similar

analyses were undertaken using all RCTs with binary nonmortal primary outcomes.

Statistical Analyses

We conducted unadjusted comparisons of proportions using χ^2 tests to examine the differences in proportions of successful trials across trial characteristics. We used multivariable regression to identify study-level characteristics associated with a trial's being positive. For this purpose, given limited degrees of freedom, we limited our assessments to the following trial characteristics: (1) mortal versus nonmortal primary outcome, (2) funding source, (3) single versus 2 to 10 centers versus more than 10 centers, and (4) type of intervention. Odds ratios (ORs) from

a logistic regression and prevalence rate ratios (PRR) from a Poisson regression with a robust variance estimator are presented, because the ORs will overestimate relative risks with event (positive trial) rates greater than 10% (27).

Results

Our search identified 376 potential studies published between January 2007 and May 2013 (Figure 1). Of these, 146 met the prespecified inclusion criteria (Table E1). The most commonly tested types of interventions were protocols (49%) and drug therapies (40%) (Table 1). Most trials (92%) compared two intervention arms

(max = 5). Overall, 54 (37%) were positive; that is, these RCTs demonstrated a significant difference between study groups in the primary outcome as hypothesized (Table 1). In addition to the 19 (13%) RCTs stopped early for safety or futility, an additional 4 RCTs (3%) revealed statistically significant findings of inferiority (i.e., effects contrary to the primary hypothesis) (Table 2).

The most common primary outcomes were measures of mortality over a specified time period (27%), followed by outcomes related to healthcare-associated infections (23%), ventilation (21%) (e.g., time to extubation, ventilator-free days, or required mechanical ventilation), and quality (10%) (e.g., complications or adverse events). The incidence of positive trials varied

Figure 1. Analytic sample of published randomized clinical trials (RCTs) of critical care interventions. *See Figure 5 legend. ICU = intensive care unit.

Table 1. Characteristics of Adult Randomized Clinical Trials in Critical Care

Characteristic	No. (%)	No. (%) with a Positive Primary Outcome
Total	146 (100)	54 (37)
Funding		
No industry	80 (55)	26 (33)
Some industry	42 (29)	13 (31)
No funding/not reported	24 (16)	15 (63)
Single center	54 (37)	25 (46)
Multicenter	92 (63)	29 (32)
≤10 ICUs	40 (27)	18 (45)
11–25 ICUs	25 (17)	6 (24)
>25 ICUs	27 (18)	5 (19)
Type of intervention studied		
Protocol	71 (49)	30 (42)
Drug	59 (40)	18 (31)
Device/monitoring	5 (3)	1 (20)
Other	11 (8)	5 (45)
Primary target patient populations		
General ICU	52 (36)	30 (58)
Sepsis spectrum	22 (15)	0
Cardiac critical care	17 (12)	7 (41)
Acute lung injury/acute respiratory distress syndrome	16 (11)	2 (13)
Unit of randomization		
Patient, surrogate, or family	137 (94)	49 (36)
ICU, cluster randomization	9 (6)	5 (56)
Primary outcome (1 per trial, n = 146 RCTs)		
Mortality (e.g., hospital, ICU, 28 d)	40 (27)	4 (10)
Infection related	33 (23)	19 (58)
Ventilation related	30 (21)	13 (43)
Quality, complications/adverse outcomes	14 (10)	7 (50)
Organ failure	8 (5)	1 (13)
Composite outcome	7 (5)	2 (29)
Delirium	5 (3)	2 (40)
Hospital discharge disposition, functional status	3 (2)	1 (33)
Length of stay	3 (2)	2 (67)
Smoking cessation	2 (1)	2 (100)
Quality of sleep	1 (1)	1 (100)
Most frequent secondary outcomes (multiple possible per RCT)		
Mortality		
ICU	47 (32)	4 (9)
In-hospital	44 (30)	2 (5)
28 d	29 (20)	4 (14)
29–180 d	35 (24)	5 (14)
Ventilation		
Duration of MV	55 (38)	12 (22)
Ventilator-free days	22 (15)	6 (27)
Length of stay		
ICU	93 (64)	12 (13)
Hospital	71 (49)	5 (7)
Quality, complications/adverse outcomes	60 (41)	14 (23)
Infection related	36 (25)	8 (22)
Organ failure	17 (12)	2 (12)

Definition of abbreviations: ICU = intensive care unit; MV = mechanical ventilation; RCT = randomized clinical trial.

depending on the primary outcome. The success rates for trials using these four above-mentioned outcomes were 10, 58, 43, and 50%, respectively. Two of the four positive mortality trials were only significant after prespecified adjustment (28, 29).

Twenty-four of the 40 trials in which mortality was the primary outcome studied 28- or 30-day mortality (Table E3). Five additional RCTs included a mortality endpoint as part of a composite primary outcome with nonmortal measures, and one RCT was powered on mortality despite being listed as a secondary outcome; of these, one trial was positive (Table E4). The most common secondary outcomes across all RCTs were ICU (64%) and hospital (49%) length of stay (Table 1).

Of the 122 (84%) trials that disclosed the funding source, 34% reported receipt of industry funding, and 66% reported no industry funding. There was no relationship between industry funding and the probability that a trial would be positive (33 vs. 31%, $P = 0.9$). The remaining 24 trials did not disclose any sources of funding, and these were more likely to be positive (63%, $P = 0.005$ for comparison with all studies reporting funding sources). Single-center RCTs ($n = 54$) were less common than multicenter RCTs ($n = 92$). However, multicenter RCTs were less likely to be positive, and the rate decreased as the number of participating ICUs increased ($P = 0.03$) in univariate analyses. In the multivariable regressions, RCTs that did not report any funding source (OR = 3.3; 95% CI, 1.2–9.4) and RCTs that did not study a primary mortality outcome (OR = 6.8; 95% CI, 2.1–22.7) were significantly more likely to be successful (Figure 2).

Power or sample size were discussed in 135 RCTs (92%); however, only 68% of these studies cited prior research, a pilot study, or examination of other data (e.g., from the authors' center) to justify the inputs used in calculating the required sample size (Table 2). A CONSORT diagram portraying participant flow was reported in 119 RCTs (82%).

A total of 101 (69%) RCTs used a binary primary outcome. Of these, 40 examined a mortality outcome and 61 used other nonmortal outcomes (e.g., incidence of ventilator-associated pneumonia) (Figure 1; see Tables E3 and E4 for a full list of outcomes). Twenty-three of the 40 RCTs with mortality as a primary outcome explained the rationale for their predicted treatment-associated mortality reduction. Thirty-four of these 40 RCTs reported the values for their power calculation and specified that they were superiority trials (i.e., powered for a specific treatment-associated mortality

Table 2. Power and Sample Size Characteristics of Randomized Clinical Trials in Critical Care

Characteristic	No. (%) of RCTs	No. (%) Positive
Total	146	
Included a CONSORT diagram, patient flow	119 (82)	
Rationale for power parameters (e.g., baseline rate, predicted delta, expected time to event)	92 (63)	
Type of outcome		
Binary outcome	101 (69)	31 (31)
Duration (e.g., event-free days) or time-to-event outcome	35 (24)	16 (46)
Rate (e.g., per 1,000 patient-days)	7 (5)	6 (86)
Continuous	3 (2)	1 (33)
RCT stopped early	32 (22)	
Futility	10 (7)	
Safety	9 (6)	
Recruitment/logistical issues	11 (8)	
Power or sample size plan discussed, including cluster trials	135 (92)	
RCT reported a targeted <i>a priori</i> sample size	130/135 (96)	
Recruited <95% of target or stopped early due to recruitment/logistical issues	20/130 (15)	4/20 (20)
Recruited 95–110% of target sample size or stopped early for futility or efficacy	88/130 (68)	36/88 (41)
Recruited >110% of target sample size	13/130 (10)	4/13 (31)
Stopped early for safety reasons	9/130 (7)	

Definition of abbreviation: RCT = randomized clinical trial.

reduction). Of these 34 mortality endpoint superiority trials, three were positive (two only after prespecified adjustment), and 11 (33%) had nonsignificant absolute treatment effects in the hypothesized direction that were larger than 3% (Figure 3; see Table E3).

Of the 61 RCTs with a primary nonmortal binary outcome, 47 were two-arm superiority trials and reported the predicted treatment-associated reduction they used for their power calculation (Figure 1; see Table E4). Of these 47 RCTs, 20 were positive and 27 were statistically nonsignificant, of which 12 (44.4%) observed absolute treatment effects in the hypothesized direction that were larger than 3% (Figure 4; see Table E4).

Among the 33 superiority trials without adaptive control arms reporting expected control group mortality rates, the actual control group mortality differed

from the expected value by 7.5% or more in 22 RCTs (Figure E1). Despite these frequent differences between expected and observed control group mortality rates, this rarely accounted for a study's inability to detect a given effect size as significant. For example, 12 (out of 30) negative mortality trials that tested for superiority could have detected a 10% mortality reduction with the observed control group mortality rate, compared with 13 such trials if the expected control group mortality had been observed (Figure 5). Among the 46 (of 47) nonmortal superiority trials with a binary endpoint in which expected control group rates were reported in the manuscripts, the actual control group rate differed from the expected value by 7.5% or more in 21 RCTs. Similar to the aforementioned results for mortality trials, misspecification of control group

rates rarely accounted for a study's inability to detect a given effect size as significant.

Discussion

This contemporary study of 146 RCTs published in the leading medical and critical care journals yields several important findings. First, investigators choose a variety of primary outcomes for trials of ICU-based interventions (Tables 1, E3, and E4).

Some of this heterogeneity is appropriate, given different anticipated effects of various interventions. However, the variation of endpoints selected even among trials using some form of a mortality primary endpoint (Table E3) suggests little agreement on the optimal outcomes in critical care. These data complement a prior study showing variability in ventilation-associated outcomes in critical care RCTs (30). This lack of standardized definitions and methods for assessing common outcomes poses challenges for comparing and understanding differences between RCTs, replicating results, and conducting metaanalyses.

Second, a majority of RCTs are "negative" in the sense that they do not demonstrate a benefit from the tested intervention. This is particularly true when mortality is the primary outcome (10% positive rate, or 5% if only crude rates are considered), with higher proportions of positive trials when other outcomes are used (13–100% positive rate) (Table 1). Of note, a 5 to 10% positive rate is roughly the rate that would be expected assuming a conventional type I error rate of 0.05. A prior review of RCTs in both adults and children published in the journal *Intensive Care Medicine* from 2001 to 2010 found an overall success rate of 48.8% (of 217 RCTs) (26), somewhat higher than our observed rate of 37% (of 146 RCTs). Additionally, two reviews that focused on RCTs using mortality endpoints found success rates of 14% (10 of 72 RCTs published before August 2006) (2) and 18% (7 of 38 RCTs published from 1999–2009 in five major medical journals) (3), somewhat higher than our rate of 10%. Although it is possible that more trials are becoming negative over time, these differences may also be attributable to variability in the journals sampled and the eligibility criteria used to include RCTs. Because our study and all prior studies focused on published

Figure 2. Adjusted associations of selected randomized clinical trial (RCT) characteristics with positive primary outcomes. RCTs that included measures of morbidity or other clinical measures in the primary outcome were not categorized as mortality trials. ICU = intensive care unit.

RCTs, the true rates of successful trials are likely even lower.

The high rate of negative trials does not, itself, suggest a problem; a majority of trials may “appropriately” fail to detect significant reductions in mortality. Such “true negatives” could arise if more interventions being tested are truly ineffective, as may occur when a discipline matures. Alternatively, such findings may be attributable to the fact that mortality

in the ICU is heavily determined by physicians’ decisions to withhold or withdraw life support (31), crowding out any plausible effect of an intervention. Finally, 10 or 20% of trials should be negative by chance alone even when power is set to 90 or 80%, respectively.

Nonetheless, the present study suggests that in many cases, critical care RCTs, and especially those studying mortal endpoints, have not been designed to identify

realistic treatment effects. For example, we find that in a majority of negative RCTs, the results move in the predicted direction, often considerably so, yet fail to attain the predicted treatment effect on which the study was powered (Figures 3 and 4). This provides contemporary evidence in support of the notion that investigators commonly select implausibly large treatment effects on which to base sample size requirements (3). Although the problem of underpowered trials is certainly not unique to critical care, it does raise ethical concerns because such trials expose research participants to the risks and burdens of research without being (sufficiently) able to deliver on the purported benefits of expanding knowledge and improving future care (32, 33).

A third and related finding is that investigators commonly err in predicting the baseline event rate in their trials. With high-predicted background rates, large absolute risk reductions might seem plausible to investigators because they would reflect more modest relative risk reductions (25). However, we find that control group mortality rates are often considerably lower than predicted, which could make such large effects improbable. For instance, it may be unreasonable to assume that an intervention predicted to bring mortality down to 30%, assuming a base rate of 40%, would also reduce mortality to 10% if the base rate turned out to be 20%. Thus, as the baseline mortality rate declines, there will invariably be diminishing marginal returns for any intervention (i.e., a lower proportion of potentially savable patients).

Despite the possibility that overpredictions of control group event rates would contribute to critical care RCTs being negative, this appears to be only a minor piece of the problem. We found that even when large errors were made in predicted baseline mortality, this rarely changed whether a trial would or would not have detected a given difference as significant. This may be attributable to a counterbalancing phenomenon whereby as the baseline rate moves away from 50%, the sample size required to detect any given difference on an absolute scale decreases. Studies of secular declines in mortality rates for common pathologies, such as done with multicenter RCTs in sepsis (34) and acute lung injury (35), could better inform control group mortality rates and also guide selection of more reasonable treatment effects

Figure 3. Expected versus actual treatment effect on mortality in 34 superiority trials in which the primary outcome was mortality. See Table E3 in the online supplement for the numbers shown in this figure and for further information regarding all trials in which mortality was a primary outcome.

Figure 4. Expected versus actual treatment effect in 47 superiority trials with a binary nonmortal primary outcome. See Table E4 in the online supplement for the numbers shown in this figure and for further information regarding each trial and the primary outcomes.

when designing future RCTs. Furthermore, event-driven adaptive trial designs, such as used in the Prospective Recombinant Human Activated Protein C Worldwide Evaluation in Severe Sepsis and Septic Shock (PROWESS-SHOCK) trial (36), that adjust

(by increasing sample size) to lower than expected mortality in the control group offer an attractive solution to this issue.

Additional strategies for improving trial success might include use of prespecified covariate adjustment (37–39)

Figure 5. Simulation results of superiority trials where the primary outcome was mortality assuming 80% power to find a treatment-associated mortality reduction of 3 to 15%. $n = 33$ randomized clinical trials as noted in Figure 1 with the exception of the Prospective Recombinant Human Activated Protein C Worldwide Evaluation in Severe Sepsis and Septic Shock (PROWESS-SHOCK) trial due to the use of an adaptive control arm.

(e.g., see Jansen and colleagues [28]), larger target sample sizes, and more realistic and conservative treatment effect expectations (40) (Table 3). Additionally, innovative trial designs, such as Bayesian adaptive trials, may be particularly valuable for assessing drug therapies (35, 41). Regarding endpoints, some have questioned the conceptual propriety of using mortality as an endpoint for research or quality assessment on seriously or critically ill patients (42). Although many experts believe that mortality is the ultimate patient-centered outcome for critically ill patients, others have called for greater use of nonmortal clinical endpoints (35, 43). Unfortunately, nonmortal endpoints face several threats to validity, including, but not limited to, ascertainment bias (measurement error) and the limits of commonly used statistical methods for addressing the competing risks and informative dropout attributable to high ICU mortality rates. Indeed, our observation that RCTs of nonmortal endpoints were more likely to be positive may be an artifact of these measurement and analysis problems. Ongoing methodological work designed to offer new critical care outcome measures that incorporate mortality into the assessment of ICU length of stay or post-ICU quality of life may ultimately offer optimal approaches for quantifying the effects of interventions in the ICU.

This study has limitations. First, we only calculated power and detectable differences for trials using binary endpoints. We considered methods to assess effect sizes of trials using continuous or time-to-event outcome, such as ventilator-free days or time to extubation. However, potential effect size cutpoints (i.e., Cohen's d , Glass's Δ , or Hedges' g), are all based on assumptions of normally distributed data. Because we found these assumptions unrealistic for most critical care outcomes, and the inputs difficult, if not impossible, to back-calculate from the published findings, we limited our power assessments to trials using binary outcomes. Second, our review was limited to adult critical care RCTs published in 16 selected journals. Third, because we relied on published data (and online supplements when available), changes in journal requirements over time may have contributed to certain reporting omissions (e.g., funding information or CONSORT diagrams). Fourth, important design issues,

Table 3. Selected Recommendations for Critical Care Trial Design

Domain	Hypothesis	Recommendations to Potentially Improve Design
Study population	Treatment-effect heterogeneity might lead to a diluted effect estimate because although interventions work for certain patients, others are too sick and/or have too many competing risks for death for singular interventions to be of benefit.	Stratified randomization. Prespecified severity of illness adjustment when estimating treatment effects (39). Stratification of trial results based on severity of illness at baseline (46, 47). Adaptive trial designs (e.g., using biomarkers to stratify patients into more homogeneous subgroups [41], event-driven adaptive trials [36], or starting trials with several arms and then adjusting sample sizes [48] or narrowing arms based on observed interim safety and efficacy data [49]).
Participant accrual and retention	RCTs are sufficiently powered but patient attrition leads to appreciable postrandomization losses so that the intention-to-treat analyses are highly conservative.	Incorporation of patient attrition estimates when making sample size calculations. Improved models of informed consent (50) and potentially incentives for research participation (51).
Statistical power calculations	Even when the target sample size is achieved and retained, RCTs may be insufficiently powered to detect relatively small but important effects on appropriate outcomes.	Increased metastudies to better inform control arm event rates (e.g., [34]). Use of more realistic and conservative predicted treatment effects when estimating sample sizes. Use of continuous outcomes when possible. Reconstruction of binary endpoints into categorical endpoints to improve statistical efficiency (37, 52).
Outcome	Outcome measures are inappropriately specified or analyzed.	Consensus development among trial groups and intensivists about follow-up periods and definitions of outcomes for specific conditions to support comparisons across trials (e.g., metaanalysis) (30, 45). Novel methods for handling right-censoring due to deaths in analyses of quality of life and other nonmortal outcomes (53).

Definition of abbreviation: RCT = randomized clinical trial.
References 2–5 and 14–17 contributed ideas presented in this table.

such as allocation concealment, blinding or masking, and ascertainment bias, were not assessed. Finally, although we implemented an exhaustive search with oversight from a medical librarian, it is conceivable that our search strategy did not identify all eligible trials.

In summary, we believe greater dialogue is needed to determine the usefulness of nonmortal outcomes to patients, providers, and payers and to

identify elements of trial design and analysis that are associated with the significance of results (35, 44). Rather than abandoning RCTs, the results suggest opportunities for designing critical care trials more efficiently. Actionable first steps might include consensus building among the critical care community (including journal editors) regarding a minimum core outcome set (30, 45), methodological work to improve strategies for measuring these

outcomes, and closer scrutiny of submitted manuscripts to ensure an “honest” power calculation, which should in turn encourage more realistic trial design. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank our two anonymous reviewers for helpful comments on a prior version of this manuscript.

References

- Sevransky JE, Checkley W, Martin GS. Critical care trial design and interpretation: a primer. *Crit Care Med* 2010;38:1882–1889.
- Ospina-Tascon GA, Buchele GL, Vincent J-L. Multicenter, randomized, controlled trials evaluating mortality in intensive care: doomed to fail? *Crit Care Med* 2008;36:1311–1322.
- Aberegg SK, Richards DR, O'Brien JM. Delta inflation: a bias in the design of randomized controlled trials in critical care medicine. *Crit Care* 2010;14:R77.
- Angus DC, Mira JP, Vincent JL. Improving clinical trials in the critically ill. *Crit Care Med* 2010;38:527–532.
- Annane D. Improving clinical trials in the critically ill: unique challenge—sepsis. *Crit Care Med* 2009;37:S117–S128.
- 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.
- Girard TD, Kress JP, Fuchs BD, Thomason JW, Schweickert WD, Pun BT, Taichman DB, Dunn JG, Pohlman AS, Kinniry PA, et al. Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (Awakening and Breathing Controlled trial): a randomised controlled trial. *Lancet* 2008;371:126–134.
- Guerin C, Reignier J, Richard JC, Beuret P, Gacouin A, Boulain T, Mercier E, Badet M, Mercat A, Baudin O, et al. Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med* 2013;368:2159–2168.
- Dreyfuss D. Beyond randomized, controlled trials. *Curr Opin Crit Care* 2004;10:574–578.
- Vincent J-L. We should abandon randomized controlled trials in the intensive care unit. *Crit Care Med* 2010;38:S534–S538.
- 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.
- Perner A, Haase N, Guttormsen AB, Tenhunen J, Klemenzson G, Aneman A, Madsen KR, Moller MH, Elkjaer JM, Poulsen LM, et al. Hydroxyethyl starch 130/0.42 versus Ringer's acetate in severe sepsis. *N Engl J Med* 2012;367:124–134.
- Myburgh JA, Finfer S, Bellomo R, Billot L, Cass A, Gattas D, Glass P, Lipman J, Liu B, McArthur C, et al.; CHEST Investigators; Australian and New Zealand Intensive Care Society Clinical Trials Group. Hydroxyethyl starch or saline for fluid resuscitation in intensive care. *N Engl J Med* 2012;367:1901–1911.
- van Meurs M, Ligtgenberg JJ, Zijlstra JG. The randomized controlled trial needs critical care. *Crit Care Med* 2008;36:3118–3119, author reply 3119.
- Rubenfeld GD, Abraham E. When is a negative phase II trial truly negative? *Am J Respir Crit Care Med* 2008;178:554–555.
- Reade MC, Angus DC. The clinical research enterprise in critical care: what's right, what's wrong, and what's ahead? *Crit Care Med* 2009;37:S1–S9.
- McAuley DF, O'Kane C, Griffiths MJ. A stepwise approach to justify phase III randomized clinical trials and enhance the likelihood of a positive result. *Crit Care Med* 2010;38:S523–S527.
- Marini JJ. Limitations of clinical trials in acute lung injury and acute respiratory distress syndrome. *Curr Opin Crit Care* 2006;12:25–31.
- Harhay MO, Wagner J, Cooney E, Bronheim RS, Gopal A, Green S, Kerlin MP, Mikkelsen ME, Small D, Halpern SD. Epidemiology of published critical care randomized clinical trials, 2007–2012 [abstract]. *Am J Respir Crit Care Med* 2013;187:A1601.
- Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Schulz KF. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med* 2008;5:e20.
- Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1–12.
- Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323:42–46.
- Chiche JD, Angus DC. Testing protocols in the intensive care unit: complex trials of complex interventions for complex patients. *JAMA* 2008;299:693–695.
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–381.
- Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. New York: Churchill Livingstone; 1997.
- Latronico N, Metelli M, Turin M, Piva S, Rasulo FA, Minelli C. Quality of reporting of randomized controlled trials published in Intensive Care Medicine from 2001 to 2010. *Intensive Care Med* 2013;39:1386–1395.
- Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. *Occup Environ Med* 2008;65:501–506.
- Jansen TC, van Bommel J, Schoonderbeek FJ, Sleswijk Visser SJ, van der Klooster JM, Lima AP, Willemsen SP, Bakker J. Early lactate-guided therapy in intensive care unit patients: a multicenter, open-label, randomized controlled trial. *Am J Respir Crit Care Med* 2010;182:752–761.
- Papazian L, Forel JM, Gacouin A, Penot-Ragon C, Perrin G, Loundou A, Jaber S, Arnal JM, Perez D, Seghboyan JM, et al. Neuromuscular blockers in early acute respiratory distress syndrome. *N Engl J Med* 2010;363:1107–1116.
- Blackwood B, Clarke M, McAuley DF, McGuigan PJ, Marshall JC, Rose L. How outcomes are defined in clinical trials of mechanically ventilated adults and children. *Am J Respir Crit Care Med* 2014;189:886–893.
- Garland A, Connors AF. Physicians' influence over decisions to forego life support. *J Palliat Med* 2007;10:1298–1305.
- Luce JM, Cook DJ, Martin TR, Angus DC, Boushey HA, Curtis JR, Heffner JE, Lanken PN, Levy MM, Polite PY, et al.; American Thoracic Society. The ethical conduct of clinical research involving critically ill patients in the United States and Canada: principles and recommendations. *Am J Respir Crit Care Med* 2004;170:1375–1384.
- Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358–362.
- Stevenson EK, Rubenstein AR, Radin GT, Wiener RS, Walkey AJ. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis*. *Crit Care Med* 2014;42:625–631.
- Spragg RG, Bernard GR, Checkley W, Curtis JR, Gajic O, Guyatt G, Hall J, Israel E, Jain M, Needham DM, et al. Beyond mortality: future clinical research in acute lung injury. *Am J Respir Crit Care Med* 2010;181:1121–1127.
- Ranieri VM, Thompson BT, Barie PS, Dhainaut JF, Douglas IS, Finfer S, Gardlund B, Marshall JC, Rhodes A, Artigas A, et al. PROWESS-SHOCK Study Group. Drotrecogin alfa (activated) in adults with septic shock. *N Engl J Med* 2012;366:2055–2064.
- Roozenbeek B, Lingsma HF, Steyerberg EW, Maas AI, Group IS. Underpowered trials in critical care medicine: how to deal with them? *Crit Care* 2010;14:423.
- Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, McHugh GS, Weir J, Murray GD, Steyerberg EW, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med* 2009;37:2683–2690.
- Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004;57:454–460.
- Scales DC, Rubenfeld GD. Estimating sample size in critical care clinical trials. *J Crit Care* 2005;20:6–11.
- Angus DC, van der Poll T. Severe sepsis and septic shock. *N Engl J Med* 2013;369:840–851.
- Holloway RG, Quill TE. Mortality as a measure of quality: implications for palliative and end-of-life care. *JAMA* 2007;298:802–804.
- Ferguson ND, Scales DC, Pinto R, Wilcox ME, Cook DJ, Guyatt GH, Schünemann HJ, Marshall JC, Herridge MS, Meade MO; Canadian Critical Care Trials Group. Integrating mortality and morbidity outcomes: using quality-adjusted life years in critical care trials. *Am J Respir Crit Care Med* 2013;187:256–261.

44. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? *J Clin Epidemiol* 1994;47:787–795.
45. Young P, Hodgson C, Dulhunty J, Saxena M, Bailey M, Bellomo R, Davies A, Finfer S, Kruger P, Lipman J, *et al.* End points for phase II trials in intensive care: recommendations from the Australian and New Zealand Clinical Trials Group consensus panel meeting. *Crit Care Resusc* 2012;14:211–215.
46. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007;298:1209–1212.
47. Kent DM, Alsheikh-Ali A, Hayward RA. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* 2008;9:30.
48. Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biom J* 2006;48:537–555.
49. Lewis RJ, Viele K, Broglio K, Berry SM, Jones AE. An adaptive, phase II, dose-finding clinical trial design to evaluate L-carnitine in the treatment of septic shock based on efficacy and predictive probability of subsequent phase III success. *Crit Care Med* 2013;41:1674–1678.
50. Scales DC. Research to inform the consent-to-research process. *Intensive Care Med* 2013;39:1484–1486.
51. Halpern SD. Financial incentives for research participation: empirical questions, available answers and the burden of further proof. *Am J Med Sci* 2011;342:290–293.
52. McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, Weir J, Maas AI, Murray GD. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials* 2010;7:44–57.
53. Rosenbaum PR. Comment: the place of death in the quality of life. *Stat Sci* 2006;21:313–316.