Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care

Theodore J. Iwashyna[1,2,3], James F. Burke[4], Jeremy B. Sussman[1,3], Hallie C. Prescott[1], Rodney A. Hayward[1,3], and Derek C. Angus[5]

[1]Department of Internal Medicine and [4]Department of Neurology, University of Michigan, Ann Arbor, Michigan; [2]Australian and New Zealand Intensive Care Research Centre, Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Victoria, Australia; [3]Center for Clinical Management Research, Department of Veterans Affairs Ann Arbor Health System, Ann Arbor, Michigan; and [5]Clinical Research, Investigation, and Systems Modeling of Acute Illness Laboratory, Department of Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania

ORCID ID: 0000-0002-4226-9310 (T.J.I.).

Abstract

Randomized clinical trials (RCTs) are conducted to guide clinicians' selection of therapies for individual patients. Currently, RCTs in critical care often report an overall mean effect and selected individual subgroups. Yet work in other fields suggests that such reporting practices can be improved. Specifically, this Critical Care Perspective reviews recent work on so-called "heterogeneity of treatment effect" (HTE) by baseline risk and extends that work to examine its applicability to trials of acute respiratory failure and severe sepsis. Because patients in RCTs in critical care medicine—and patients in intensive care units—have wide variability in their risk of death, these patients will have wide variability in the absolute benefit that they can derive from a given therapy. If the side effects of the therapy are not perfectly collinear with the treatment benefits, this will result in HTE, where different patients experience quite different expected benefits of a therapy. We use simulations of RCTs to demonstrate that such HTE could result in apparent paradoxes, including: (*1*) positive trials of therapies that are beneficial overall but consistently harm or have little benefit to low-risk patients who met enrollment criteria, and (*2*) overall negative trials of therapies that still consistently benefit high-risk patients. We further show that these results persist even in the presence of causes of death unmodified by the treatment under study. These results have implications for reporting and analyzing RCT data, both to better understand how our therapies work and to improve the bedside applicability of RCTs. We suggest a plan for measurement in future RCTs in the critically ill.

Keywords: heterogeneity of treatment effect; acute respiratory failure; sepsis; randomized clinical trials; precision medicine

Guyatt and colleagues' classic *User's Guide to the Medical Literature II* (1) states, "if the patient would have been enrolled in the study had she been there—that is, she meets all the inclusion criteria, and doesn't violate any of the exclusion criteria—there is little question that the results are applicable." If this is true and there are no contraindications, then it is often argued that the patient should receive the treatment found to be superior by such a well-conducted randomized clinical trial (RCT). This logic motivates many guidelines and quality measures.

However, there are clear examples where such logic fails patients. Perhaps the best example is carotid endarterectomy for

(Received in original form November 28, 2014; accepted in final form July 14, 2015)

Supported by grants R21 AG044752 (T.J.I.), K08 NS082597 (J.F.B.), R01 MD008879 (J.F.B.), T32 HL007749 (H.C.P.), P50 GM076659 (D.C.A.), R01 GM101197 (D.C.A.), and the National Institute of Diabetes and Digestive and Kidney Diseases–funded Michigan Center for Diabetes Translational Research Methods Core grant P30DK020572 from the National Institutes of Health, and Department of Veterans Affairs Health Services Research and Development Service grants IIR 11-109 (T.J.I.) and IIR11-088 (R.A.H. and J.B.S.).

This work does not necessarily represent the views of the U.S. Government or the Department of Veterans Affairs.

Author Contributions: All authors were involved in the conceptualization of the analyses and critical revision of the manuscript for intellectual content; T.J.I. performed the analyses and drafted the original manuscript.

Correspondence and requests for reprints should be addressed to Theodore J. Iwashyna, M.D., Ph.D., VA Center for Clinical Management Research, 2800 Plymouth Road, NCRC Building 16, Room 326W, Ann Arbor, MI 48109. E-mail: tiwashyn@umich.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 192, Iss 9, pp 1045–1051, Nov 1, 2015

Originally Published in Press as DOI: 10.1164/rccm.201411-2125CP on July 15, 2015
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Although randomized clinical trials (RCTs) provide the best evidence of the effect of a treatment on a population, the average effect from an RCT may be a misleading guide to how any given patient will do.

What This Study Adds to the Field: We suggest that a patient's baseline risk of death, measured just before treatment, may be an important determinant of how much a treatment will help any given patient. This Critical Care Perspective shows why that might be true and how changes in the analysis and reporting of RCTs might improve the usefulness of RCTs at the bedside.

stroke reduction in symptomatic patients. In a major RCT, the average effect of surgery was greater than 10% absolute risk reduction in stroke or death at 3 years—a tremendous result (2). However, the benefits were largely accrued by patients at highest risk of imminent stroke. Even in the context of this very positive RCT, surgery worsened the risk of stroke for those trial participants with lower baseline risk (3). This is not merely a problem of the external validity. Instead, in this case, a well conducted RCT resulted in an average outcome that poorly reflected outcomes of many patients *within the RCT*. In direct contradiction of the received wisdom articulated by Guyatt and colleagues quoted above, there are times when one should not apply the results of an RCT at the bedside, even to patients who would clearly have been eligible for that RCT.

RCTs are the gold standard for testing the effect of a treatment on outcome, and this Critical Care Perspective fully agrees with that view. However, we should remember that the main result of an RCT is the average effect across the tested population (4–8), which does not directly address which particular patients will benefit, or how much. It might be that few patients can expect the average benefit, as the magnitude of benefit and risk of adverse effects can vary significantly across a tested population. This raises an important question—are there subgroups of patients in a trial who are particularly likely or unlikely to benefit? The carotid endarterectomy example is not unique. Similar examples exist in other areas of medicine (9, 10). It is increasingly urgent that we understand whether such so-called heterogeneity of treatment effect (HTE) might be clinically relevant in critical care medicine, and, if so, how changes in reporting of trials might improve the translation potential of RCTs.

RCTs in acute respiratory failure and sepsis often report subgroup analyses in recognition that there may be meaningful HTE between patients. In real intensive care units (ICUs), however, high-risk patients are high risk because of multiple risk factors, but conventionally reported one-at-a-time subgroup analyses fail to capture such variation in risk. The population of patients at higher risk from a given acute organ dysfunction will nonetheless include many younger patients lacking other organ dysfunctions, blunting the separation in risk that can be achieved in such subgroup analyses (11, 12). For example, there are 17 variables in the Simplified Acute Physiology Score (SAPS) II score (13); only the risk difference between a Glasgow Coma Scale of less than 6 and greater than 14 is as large as the interquartile range (IQR) of SAPS II scores seen in recent ARDSnet trials (*see* Frequently Asked Question [FAQ] 1 in Appendix 1 in the online supplement). We need to move beyond conventional subgroup analyses, because they simply do not capture enough variation in risk.

This Critical Care Perspective examines the implication of risk-stratified HTE analysis for the reporting and bedside use of RCTs. To do so, we conducted simulations of RCTs in acute respiratory failure, drawing from distributions of actual patient data (all results also hold for severe sepsis, as shown in Appendix 2). First, we demonstrate the extent to which HTE could plausibly produce overall trial results that substantially over- or underestimate the true benefit of subsets of patients identifiable through risk-stratified analysis. Second, we discuss the implications of this phenomenon for the reporting and application of RCTs. Finally, in Appendix 1, we present an FAQ where we directly address some of the concerns raised during our development of these ideas.

Could HTE Matter in Trials in Critical Care?

The core argument for examining HTE by baseline risk of death is as follows (14). If a treatment results in a consistent relative risk reduction (RRR) and if baseline risk varies, then the treatment must have substantial variation in absolute risk reduction between patients with different baseline risk. If a treatment reduces mortality due to respiratory failure by 25% in all patients, then 1 death from respiratory failure is prevented for every 8 patients treated when baseline risk is 50%, whereas only 1 death is prevented for every 200 patients treated when baseline risk is 2%. Thus, patients with low risk of dying from the treatment's target condition have such a small chance of benefitting that it can easily be overwhelmed by even a small treatment-related harm, such as increasing renal damage, susceptibility to infections, or later complications—especially if the risks of such adverse effects are independent of the risks of dying of the target condition.

Although the logic is straightforward, its relevance to critical care medicine is less clear. For example, if critical care RCTs enroll only very sick patients at high risk of dying from the condition of interest, then risk may simply not vary much between patients in our RCTs. But more generally, there is the possibility that RCTs in critical care are different than those in stroke or outpatient cardiovascular prevention, where much of this logic was first recognized. In the ICU, our primary outcome is typically mortality. This is not merely due to the high lethality of the conditions we treat, but also because both the conditions themselves and the therapies used to treat them can have diverse multisystem effects—many of which are challenging to monitor or underrecognized at the time of phase III RCTs. As such, in the ICU, we do not have the luxury of a clean separation between therapy-caused adverse events and the natural history of the conditions that we are seeking to treat. As such, this Critical Care Perspective also asks whether the logic of HTE by baseline risk of death is relevant in such a complex situation.

Testing the Potential Relevance of HTE

To begin exploring the potential impact of HTE, we constructed simulated clinical trial populations, taking great care to ensure the realism of the simulations. We chose simulation, rather than analysis of actual critical care trials, because simulation allows for the design of multiple scenarios, facilitating exploration and quantification of the impact of varying key factors affecting patient outcome, such as baseline risk of death, RRRs, adverse-effect rates, and so forth (12). Simulations have the additional virtue that the "true" effect is known as a consequence of the simulation specification, and the frequency with which the "truth" can be detected under realistic scenarios can be estimated.

We simulated RCTs using data from 7,255 nonsurgical mechanically ventilated patients from the hospitals of the U.S. Department of Veterans Affairs (15–19). The dataset has numerous strengths, including granular clinical data and detailed risk adjustment using a validated score (16, 17, 20), with an area under the receiver operating characteristics curve greater than 0.85 for 30-day mortality. The patients in these data are described in Table 1. We used the patients' risk of death calculated on their ICU admission day, modeling RCTs where enrollment and randomization happen early in the ICU stay. The mean risk of 30-day mortality was 39.8%, with a median of 34.8% (IQR = 21.5–54.6%). The smallest predicted risk of 30-day mortality was 3.6%, with 99% having a risk of death predicted to be 7.5% or greater. This is similar to the ranges of mortality risk seen in critical care RCTs; for example, across the ARDSnet trials ARMA (lower tidal volume), LaSRS (Late Steroid Rescue Study), High v Low PEEP, and FACTT (Fluid and Catheter Treatment Trial), the patients had a median SAPS II predicted risk of death at randomization of 37.0% (IQR = 19.6–59.8%; A. Walkey, personal communication; *see also* Ref. 11)

Table 1. Characteristics of Patients Receiving Nonsurgical Mechanical Ventilation

Characteristics	Value
Total, n (%)	7,295 (100)
Age, yr, mean (SD)	66.35 (11.65)
Sex, n (%)	
Male	7,081 (97.1)
Female	214 (2.9)
Race, n (%)	
White	5,120 (70.2)
African American/black	1,437 (19.7)
Unknown	636 (8.7)
Other	102 (1.4)
Total hospital LOS, d, mean (SD)	15.93 (15.71)
Admission source, n (%)	
VA emergency department	3,719 (51.0)
VA outpatient clinic	2,639 (36.2)
Other	937 (12.8)
Discharge status, n (%)	
Outpatient (home)	2,973 (40.8)
Nursing facility	973 (13.3)
Death	2,803 (38.4)
Other	546 (7.5)
Severe sepsis, n (%)	5,243 (71.9)
30-d mortality, n (%)	2,901 (39.8)

Definition of abbreviations: LOS = length of stay; VA = Department of Veterans Affairs.

In the simulations, we assumed that each individual's odds of 30-day mortality were influenced by four factors: (*1*) severity of acute respiratory failure; (*2*) comorbid conditions (such as chronic health problems or other features of the critical illness); (*3*) the treatment's reduction in mortality from the primary illness; and (*4*) the treatment's fatal adverse-effect rate. As a starting point for our modeling, we assumed that the treatment has a constant RRR for the primary illness (and therefore a larger absolute reduction in patients at higher risk), consistent with current RCT practice of reporting a single summary RRR for the entire population (*see* FAQs 4 and 5 for discussion of alternative assumptions). Similarly, as a starting point, we assumed any adverse effects would be independent of other health factors (e.g., the risk of death due to drug-related bleeding was not correlated with the risk of dying from the primary illness; *see* FAQ 5 for an examination of when adverse events increase with baseline risk of death; our main findings hold in that situation as well). We generated three clinical trial scenarios. For each scenario, we simulated 10,000 trials of 2,500 patients with 1:1 randomization to therapy or placebo. This sample size of 2,500 was chosen to represent current larger RCTs being planned in critical care, erring on the conservative side.

Appendix 3 presents a more detailed description of the logic behind the simulations. Appendix 4 provides all the Stata 13 code necessary to conduct the simulations.

Scenario 1: "Positive" Trials of a Therapy That Reduces Acute Respiratory Failure–related Mortality

For scenarios 1 and 2, we assumed that patients' baseline risk of 30-day mortality is completely due to acute respiratory failure. Adverse outcomes resulting from the therapy—both recognized adverse events and unknown effects—occur in 3 of 100 patients treated. These numbers were chosen to result in a well powered RCT with a statistically significant overall effect size in the range for which many current trials are powered and a plausible, nontrivial, serious adverse effect rate (21). In scenario 1, the therapy reduces death due to the primary illness by a constant RRR 20%, and "positive" trials were identified as those simulations with a statistically significant RRR for the trial as a whole, with P less than 0.05 as the conventional significance threshold.

In 8,706 of 10,000 simulations, the simulated RCT was positive (power = 87% in this scenario), showing a statistically significant median relative risk (RR) of 0.85 (95% confidence interval for trials in the scenario = 0.77–0.94; IQR = 0.82–0.87). This resulted in a mean reduction in absolute mortality from 39.8 to 33.6% in the 8,706 positive RCTs.

However, despite a constant treatment-related adverse-effect rate and RRR, patients in the 8,706 positive trials experience dramatically different likelihoods of benefiting from the treatment. Indeed, such "positive" trials routinely contain subgroups that are predictably more likely to be harmed than helped. Figure 1 is stratified by decile of prerandomization risk of death. Patients at high risk of dying of the condition have a dramatic reduction in that risk when treated—a mean absolute mortality reduction of 16.3% and a number needed to treat (NNT) of only 6. Indeed, the patients in the highest three deciles of risk all receive substantial benefits from the therapy, with NNT of approximately 10 or less. In contrast, treatment increases the chance of death in the lowest risk decile, albeit modestly, and an additional 20% of patients are in deciles of risk

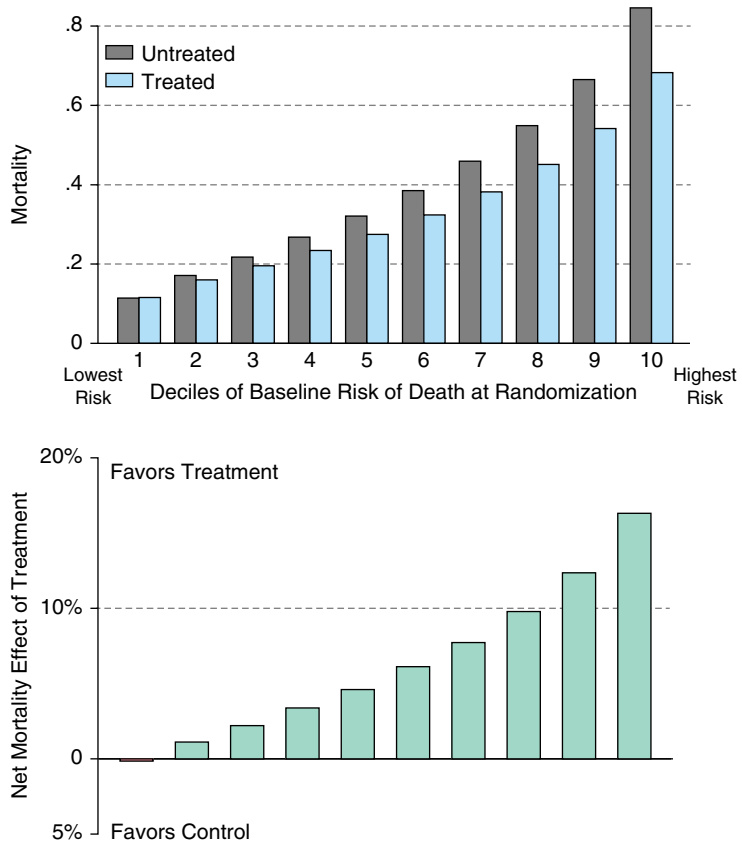

Figure 1. Scenario 1: heterogeneity of treatment effect in positive trials. This figure shows the results of the 8,706 simulated trials from scenario 1 (*see main text*) that showed a statistically significant benefit of the therapy in patients with acute respiratory failure requiring mechanical ventilation. Nonetheless, the lowest decile of risk showed very little effect. The *upper panel* shows the results in each arm, stratified by decile of risk; the *lower panel* shows the net effect: the rate of mortality in the control arm minus the rate in the treated arm.

receiving more marginal benefits, with NNTs of 45 and 90.

Interpretation and comment: proven therapies could predictably harm some patients. The most important of our results may be that the included population of RCTs of effective therapies may still contain subsets of patients who are more likely to be harmed than benefited by that therapy. This is not a facile restatement that even an effective treatment will have adverse effects so some patient will, *a posteriori*, be worse off. Instead, this is the finding that, when applying the results of acute respiratory failure trials to populations who would have met inclusion criteria for that trial, there may be *a priori*–identifiable groups of patients—for whom standard evidence-based medicine and guideline advice would recommend the treatment—who are nonetheless predictably made worse by that treatment. Such patients could be identified if trials were to examine risk-based HTE, and thereby spared such harm (or resource waste) when the trial results are translated into clinical practice. David Kent and others (8, 22, 23) have shown that such patients are best identified by stratifying patients based on composite risk scores, rather than traditional one-at-a-time subgroup testing.

HTE in general, by baseline risk of death in particular, is a problem both within a trial and in applying trials in clinical practice. Within a trial, the more uneven the distribution of patients (with a majority of patients at lower risk than the population *mean* risk, as so often occurs [24]), the greater the danger that a positive mean outcome obscures treatment-related harm in many lower baseline risk subjects.(7, 8) (*see* sepsis results in Appendix 2). When applying a trial at the bedside, the clinician is faced with the reality that most patients are not at the average baseline risk of death. Presenting results stratified by baseline risk better informs decisions on the appropriateness of the therapy for the particular patient—although the extent to which these stratified results should influence practice depends on the extent to which the HTE analyses were prespecified in the analyses plan and have been replicated.

Scenario 2: "Negative" Trials of a Therapy That Reduces Acute Respiratory Failure–related Mortality

In scenario 2, we simulate a "negative" clinical trial of a potentially efficacious therapy, one which reduced the RR by 15% (rather than 20% in scenario 1) so as to yield a large number of negative trials. All other factors, including the adverse event rate, were kept constant, resulting in a smaller population-averaged treatment effect (median RR = 0.90 [95% confidence interval = 0.81–0.99]) and lower statistical power (55%) than that in scenario 1. Examining the "negative" trials (45% of all simulated trials), we again see substantial HTE (Figure 2). Those patients in the lowest three deciles had a net increase in mortality from the tested therapy, but hidden in these "negative" trials were patients at the highest risk of dying who had a substantial risk reduction from the therapy (NNT = 9 and 14 in the two highest-risk deciles of patients). This means that, in these RCTs with no statistically significant difference in the average treatment effect—and many patients for whom the treatment offered little benefit or frank harm—there were still patients for whom the trial revealed that the therapy was highly effective.

Interpretation and comment: negative trials that benefited some patients. The same logic of HTE by baseline risk also applies to "negative" trials, those without an average difference between treated and controls. Such therapies may, of course, be truly ineffective for all patients. However, as shown in scenario 2, it may be that the trial population included many patients at low risk for the outcome of interest. In such cases, analysis on the basis of an aggregate risk score would have increased ability to detect the subpopulations in which this therapy could be beneficial. Of course, this would be only preliminary evidence, demanding further testing in a new RCT with more refined inclusion criteria—but it

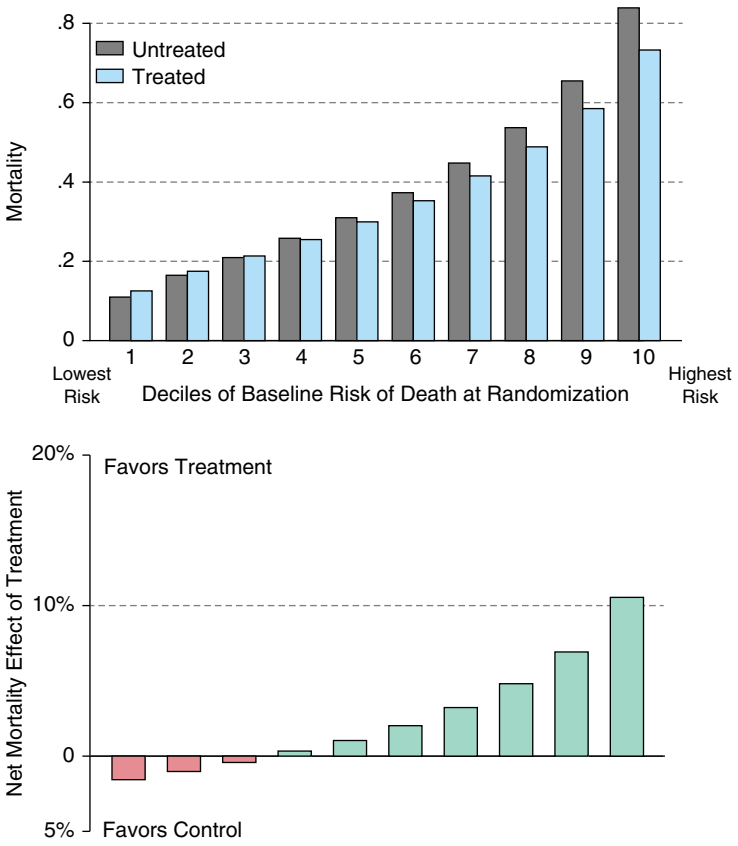

Figure 2. Scenario 2: heterogeneity of treatment effect in negative trials. This figure shows the results of the 4,504 simulated trials from scenario 2 (*see main text*) that failed to show a statistically significant benefit of the therapy in patients with acute respiratory failure requiring mechanical ventilation. Nonetheless, the highest deciles of risk showed consistent benefit (lower mortality in treated arm than in the control arm). The *upper panel* shows the results in each arm, stratified by decile of risk; the *lower panel* shows the net effect: the rate of mortality in the control arm minus the rate in the treated arm.

could prevent efficacious therapies from being erroneously discarded.

Scenario 3: "Positive" Trials Even Though Much of the Risk of Death Is Unrelated to Acute Respiratory Failure, and Thus Unaffected by the Therapy

In scenario 3, we recognize that death may also be due to other health factors unaffected by the therapy (e.g., comorbidities or organ failures already established prior to treatment initiation). We considered three levels of other causes of death for scenario 3, where the proportion of death due to acute respiratory failure (or sepsis, in Appendix 2) was 0.25, 0.5, and 0.75 of the total mortality rate. We kept mortality due to adverse effects at a constant absolute 3% and adjusted the RRR iteratively until the simulations produced roughly the same overall average results as those in scenario 1.

This scenario demonstrates that, as the fraction of total risk that is treatment responsive declines (and other causes of death increase), the treatment-responsive RRR needs to increase (*see* Appendix 2, Table E1 in the online supplement) to maintain an overall positive treatment effect. As shown in Table 2, considerable HTE persisted, despite wide variation in the likelihood that death was due to acute respiratory failure versus other causes.

Interpretation and comment: other causes of death do not solve HTE. In complex syndromic illnesses, such as acute respiratory failure and severe sepsis, it is less likely that a single therapy treats all possible mechanisms of death. This may be true if for no other reason than those syndromes often cause secondary organ failures, which then themselves have a risk of death independent of their etiology. Critically, such other causes would not be affected by the treatment being tested in the RCT. One might hope that in the presence of such alternative mechanisms of death, HTE—and the possibility of predictable treatment-related harm in lower-risk subgroups—might be less prominent. We found that, in trials of a given overall effect size, even with many other causes of death, HTE can persist—with lower-risk groups having NNTs an order of magnitude worse than those in higher-risk groups.

Implications of Potential HTE

The preceding simulations used real distributions of risk of death and plausible treatment effect sizes and adverse event rates, as seen in recent acute respiratory failure trials. These results demonstrate that these trials could potentially have clinically meaningful HTE hidden within positive and negative trials in both acute respiratory failure and, as shown in Appendix 2, severe sepsis. Wide variation in baseline risk of death is present in current RCT study populations, both in ARDSnet and the PROWESS (Recombinant Human Activated Protein C Worldwide Evaluation in Severe Sepsis) trial. In "positive" trials with statistically significant reductions in mortality, there may be an identifiable subgroup of patients for whom treatment offers little expected benefit or even expected harm. Similarly, even in some "negative" trials, it was possible to identify subgroups of patients for whom treatment substantially decreased their risk of dying. We have shown that this risk-based HTE may plausibly be present even if the risk of death responsive to treatment accounts for only a modest percentage of the overall risk of death for enrolled subjects. Patients who are at relatively lower risk of death can easily receive little benefit or even net harm from a treatment that has substantial net benefit in higher-risk subjects.

Implications for Clinical Trial Conduct and Reporting

This work, and the body of literature to which it contributes, have several key implications for future respiratory failure and severe sepsis trials. These stem from the fact that HTE is a clinical and mathematical reality when treatment-related harms and treatment-related benefits are less than perfectly collinear. The reality is that recent large clinical trials, even after restricting

Table 2. Heterogeneity in Treatment Effect in Positive Trials Persists with Increasing Other Causes of Death

	Treatment-Responsive Risk			
	100%	**75%**	**50%**	**25%**
Risk of other causes of death	0%	25%	50%	75%
Deciles of baseline risk, NNTs				
1 (lowest risk)	(642)	212	95	50
2	90	57	41	26
3	45	38	27	20
4	30	27	22	16
5	22	21	18	15
6	16	16	15	14
7	13	13	13	13
8	10	10	11	13
9	8	8	10	13
10 (highest risk)	6	6	8	11

Definition of abbreviation: NNTs = numbers needed to treat.
NNTs correspond to the absolute change in mortality in simulations associated with positive trials. The first column with 100% treatment-responsive risk is composed of the same data as scenario 1, shown in Figure 1. Numbers in parentheses represent numbers needed to harm.

entry based on biological criteria, include quite wide-ranging risks of death (*see also* FAQ 1.) Furthermore, the frequent multisystem effects (known and not yet known) of both our conditions under study and of our therapies require consideration of broad, patient-centered endpoints (such as mortality) for which HTE by baseline risk of death may be particularly prominent.

We provide pragmatic recommendations for prespecified analyses for HTE in Table 3, and some suggestions on the interpretation and follow-up of significant HTE analyses in FAQ 7. Such *a priori* specification is particularly important to allow credible examinations within negative trials of subsets of patients who would truly benefit from the therapy under study. Most examinations of HTE by baseline risk in an individual trial may be underpowered, but still worth reporting for transparency and to inform bedside decision making, by the same logic that RCTs now report subgroup analyses. Patient-level meta-analyses may be a particularly ripe opportunity to test for the magnitude of HTE by baseline risk, taking advantage of their larger sample size. It is worth noting that reporting HTE may be relevant even without clear treatment-related harm. Many new therapies—inside the ICU and out—are extremely expensive. Decisions to adopt those therapies may be made on their cost-effectiveness ratios, which are dramatically altered by an order of magnitude change in the NNT.

Table 3. Pragmatic Recommendations to Examine Heterogeneity of Treatment Effect by Baseline Risk of Death in a Trial with Mortality as Outcome

- Collect prerandomization data for an existing, well-validated severity of illness score (e.g., APACHE, SAPS) for the RCT's primary outcome. If existing scores are not deemed adequate to the population under study, collect relevant data points to develop an internally validated risk score (22). This is referred to as the baseline risk of death score.
- As primary subgroup analysis, present absolute rates of the primary outcome stratified by quintile* of baseline risk of death score.
- Test the statistical significance of the HTE as by testing the statistical significance of the interaction on the absolute scale between the baseline risk of death as a continuous* variable and the treatment (recognizing that this is likely underpowered in a single trial, as most subgroups are).

Definition of abbreviations: APACHE = Acute Physiology and Chronic Health Evaluation; HTE = heterogeneity of treatment effect; RCT = randomized clinical trial; SAPS = Simplified Acute Physiology Score.
Recommendations on interpreting such results are expanded upon in Frequently Asked Question 7 in the online supplement.
*Note that these are pragmatic recommendations at present; there is ongoing research into the optimal number of quantiles over which the stratification should be presented and the optimal parameterization of the significance testing for HTE.

Limitations

Our work has a number of limitations that should be kept in mind. Foremost, we have used simulations to demonstrate the possibility of these results. That is, these results constitute an "existence proof"—we have shown that, under plausible conditions relevant to critical care, HTE by baseline risk of death could be of a magnitude to be clinically relevant in the decision whether or not to apply a treatment to a given patient. However, we have not yet demonstrated that any specific treatment shows a consistently reduced benefit or harm in a specific group of patients. Given existing RCT reporting standards, which rarely call for stratification by aggregate risk scores, quantifying the magnitude of HTE in past trials may be challenging—but need not be so in future trials.

Beyond the limitations of this specific contribution, the literature on HTE by baseline risk of death is an active area of research where several important questions are not fully resolved. For example, better methodologies are needed to quantify and assess the clinical importance of HTE and its replicablity. A consensus statistical best practice for quantification of HTE has not yet emerged, but is needed. Equally important, additional theoretical work is necessary to understand the dynamics of HTE in the context of less-simplified situations we consider here. For example, although existing work routinely simplifies severity of illness to a single score, it may (or may not) be useful to distinguish those aspects of severity of illness driven by an acute illness, and those aspects driven by pre-existing comorbidity. Furthermore, there is work to be done understanding the interplay between HTE by baseline risk of death and HTE by specific biological mechanisms of a given treatment (*see* also FAQs 2 and 9). Nonetheless, we believe this HTE literature is sufficiently advanced that certain specific changes in reporting and analysis of RCTs are warranted at this stage, as proposed in Table 3. The implications of finding evidence of HTE are

discussed in FAQ 7, and potential misinterpretations of such a finding are discussed in FAQ 8.

Conclusions

In conclusion, this work suggests that a single average treatment effect from an RCT might be a misleading guide to physicians for their use of a given treatment in an individual patient, as we may not know what characteristics make the patient "average" in the relevant sense. Even among patients who meet enrollment criteria, some patients will generally have a much greater absolute risk reduction from treatment than others—and the treatment may even increase the risk of death in some. In short, the mean result should not be the only message that we hear from the rich data collected in RCTs. A hallmark of the modern ICU is the exceptionally wide range of short-term risk of death seen in conventional ICU patients (25)—a variation in risk that may exceed the variation found in most other clinical contexts. Because such variation in baseline risk is a major driver of clinically significant HTE, the ICU may be a pre-eminent site within which improved analyses and reporting of clinical trials can lead to more effective and efficient decisions, and smarter patient care. ■

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

Acknowledgment: The authors thank Kyle Kepreos for his expert assistance with data analysis, and Alisa M. Higgins and three anonymous reviewers for very helpful critiques.

References

1. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1994; 271:59–63.
2. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). *Lancet* 1998;351:1379–1387.
3. Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. *Lancet* 1999; 353:2105–2110.
4. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 2005;365:82–93.
5. Rothwell PM. Treating individuals 2: subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176–186.
6. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 2005;365:256–265.
7. Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers, and consumers. *Health Aff (Millwood)* 2005;24:1571–1581.
8. 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.
9. Hayward RA, Krumholz HM, Zulman DM, Timbie JW, Vijan S. Optimizing statin treatment for primary prevention of coronary artery disease. *Ann Intern Med* 2010;152:69–77.
10. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, Silvestri GA, Chaturvedi AK, Katki HA. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N Engl J Med* 2013;369:245–254.
11. Ely EW, Laterre PF, Angus DC, Helterbrand JD, Levy H, Dhainaut JF, Vincent JL, Macias WL, Bernard GR; PROWESS Investigators. Drotrecogin alfa (activated) administration across clinically important subgroups of patients with severe sepsis. *Crit Care Med* 2003;31: 12–19.
12. Clermont G, Bartels J, Kumar R, Constantine G, Vodovotz Y, Chow C. *In silico* design of clinical trials: a method coming of age. *Crit Care Med* 2004;32:2061–2070.
13. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270:2957–2963.
14. Kerr EA, Hayward RA. Patient-centered performance management: enhancing value for patients and health care systems. *JAMA* 2013; 310:137–138.
15. Render ML, Welsh DE, Kollef M, Lott JH III, Hui S, Weinberger M, Tsevat J, Hayward RA, Hofer TP. Automated computerized intensive care unit severity of illness measure in the Department of Veterans Affairs: preliminary results. SISVistA Investigators. Scrutiny of ICU Severity Veterans Health Sysyems Technology Architecture. *Crit Care Med* 2000;28:3540–3546.
16. Render ML, Kim HM, Welsh DE, Timmons S, Johnston J, Hui S, Connors AF Jr, Wagner D, Daley J, Hofer TP; VA ICU Project (VIP) Investigators. Automated intensive care unit risk adjustment: results from a National Veterans Affairs study. *Crit Care Med* 2003;31: 1638–1646.
17. Render ML, Deddens J, Freyberg R, Almenoff P, Connors AF Jr, Wagner D, Hofer TP. Veterans Affairs intensive care unit risk adjustment model: validation, updating, recalibration. *Crit Care Med* 2008;36:1031–1042.
18. Render ML, Freyberg R, Hasselback R, Hofer TP, Sales AE, Deddens J, Almenoff PL. Infrastructure for quality improvement: measurement and reporting in Veterans Administration Intensive Care Units. *Qual Saf Health Care* 2011;20:498–507.
19. Cooke CR, Kennedy EH, Wiitala WL, Almenoff PL, Sales AE, Iwashyna TJ. Despite variation in volume, Veterans Affairs hospitals show consistent outcomes among patients with non-postoperative mechanical ventilation. *Crit Care Med* 2012;40:2569–2575.
20. Render ML, Kim HM, Deddens J, Sivaganesin S, Welsh DE, Bickel K, Freyberg R, Timmons S, Johnston J, Connors AF Jr, *et al*. Variation in outcomes in Veterans Affairs intensive care units with a computerized severity measure. *Crit Care Med* 2005;33:930–939.
21. Harhay MO, Wagner J, Ratcliffe SJ, Bronheim RS, Gopal A, Green S, Cooney E, Mikkelsen ME, Kerlin MP, Small DS, *et al*. Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med* 2014;189:1469–1478.
22. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circ Cardiovasc Qual Outcomes* 2014;7:163–169.
23. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol* 2006;6:18.
24. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. *J Clin Epidemiol* 1997;50:1089–1098.
25. Wunsch H, Angus DC, Harrison DA, Linde-Zwirble WT, Rowan KM. Comparison of medical admissions to intensive care units in the United States and United Kingdom. *Am J Respir Crit Care Med* 2011; 183:1666–1673.