Power Calculations to Select Instruments for Clinical Trial Secondary Endpoints

A Case Study of Instrument Selection for Post-Traumatic Stress Symptoms in Subjects with Acute Respiratory Distress Syndrome

Michael W. Sjoding^{1,2}, David A. Schoenfeld³, Samuel M. Brown⁴, Catherine L. Hough⁵, Donald M. Yealy⁶, Marc Moss⁷, Derek C. Angus⁸, and Theodore J. Iwashyna^{1,9}; NHLBI Prevention and Early Treatment of Acute Lung Injury (PETAL) Network

¹Department of Internal Medicine and ²Institute for Healthcare Policy & Innovation, University of Michigan, Ann Arbor, Michigan; ³Biostatistics Unit, Massachusetts General Hospital and Harvard University, Boston, Massachusetts; ⁴Department of Pulmonary and Critical Care Medicine, Intermountain Medical Center and University of Utah School of Medicine, Salt Lake City, Utah; ⁵Division of Pulmonary and Critical Care Medicine, Harborview Medical Center, University of Washington, Seattle, Washington; ⁶Department of Emergency Medicine and ⁸Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; ⁷Department of Medicine, University of Colorado, Denver, Colorado; and ⁹Veterans Affairs Center for Clinical Management Research, Ann Arbor, Michigan

Abstract

Rationale: After the sample size of a randomized clinical trial (RCT) is set by the power requirement of its primary endpoint, investigators select secondary endpoints while unable to further adjust sample size. How the sensitivity and specificity of an instrument used to measure these outcomes, together with their expected underlying event rates, affect an RCT's power to measure significant differences in these outcomes is poorly understood.

Objectives: Motivated by the design of an RCT of neuromuscular blockade in acute respiratory distress syndrome, we examined how power to detect a difference in secondary endpoints varies with the sensitivity and specificity of the instrument used to measure such outcomes.

Methods: We derived a general formula and Stata code for calculating an RCT's power to detect differences in binary outcomes when such outcomes are measured with imperfect sensitivity and specificity. The formula informed the choice of instrument for

measuring post-traumatic stress–like symptoms in the Reevaluation of Systemic Early Neuromuscular Blockade RCT (www.clinicaltrials.gov identifier NCT02509078).

Measurements and Main Results: On the basis of published sensitivities and specificities, the Impact of Events Scale-Revised was predicted to measure a 36% symptom rate, whereas the Post-Traumatic Stress Symptoms instrument was predicted to measure a 23% rate, if the true underlying rate of post-traumatic stress symptoms were 25%. Despite its lower sensitivity, the briefer Post-Traumatic Stress Symptoms instrument provided superior power to detect a difference in rates between trial arms, owing to its higher specificity.

Conclusions: Examining instruments' power to detect differences in outcomes may guide their selection when multiple instruments exist, each with different sensitivities and specificities.

Keywords: clinical trials; critical care outcomes; bias; sensitivity; specificity

(Received in original form August 3, 2016; accepted in final form October 9, 2016)

Supported by the National Institutes of Health under U.S. National Heart, Lung, and Blood Institute (NHLBI) U01 grants HL123009-01, HL123010-01, HL123004-01, HL123022-01, HL122989-01, HL123008-01, HL123027-01, HL123020-01, HL123018-01, HL123031-01, HL123033-01, HL122998-01, and HL123023-01, and by a grant to M.W.S. from the NHLBI (T32HL007749).

Author Contributions: All authors were involved in the conceptualization of the analyses, critical revision of the manuscript for intellectual content, and approval of the final manuscript. M.W.S. performed all statistical analysis and drafted the original manuscript.

Correspondence and requests for reprints should be addressed to Michael Sjoding, M.D., Department of Internal Medicine, University of Michigan, 3916 Taubman Center, 1500 East Medical Center Drive, SPC 5360, Ann Arbor, MI 48109-5360. E-mail: msjoding@med.umich.edu

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

Ann Am Thorac Soc Vol 14, No 1, pp 110–117, Jan 2017

Copyright © 2017 by the American Thoracic Society

DOI: 10.1513/AnnalsATS.201608-585OC

Internet address: www.atsjournals.org

When choosing endpoints for randomized clinical trials (RCTs), trial designers must first determine the domains of scientific and clinical interest relevant to the study intervention. A primary study endpoint such as mortality is often chosen to assess the overall benefit of a new treatment. However, additional endpoints may also be relevant, particularly if a new treatment has important risks or the potential for adverse events. These safety endpoints are often relegated to the study's secondary outcomes and given less importance in a study's design. Yet, when deciding whether a new treatment is clinically useful, a clinician must weigh both the benefits and the risks of a particular treatment before deciding whether to use it in clinical practice. Therefore, it is important that a clinical trial be designed to detect important differences in secondary outcomes or adverse events if those outcomes have the potential to tip the balance of the risk-to-benefit ratio for a particular treatment.

The matter of choosing clinical trial endpoints for certain domains of interest is complicated by the fact that some outcomes, such as physical disability or psychiatric symptoms, are not quantified directly. These outcomes must be assessed using a test or measurement instrument (e.g., a survey administered to patients) that has an imperfect ability to measure such outcomes. Often more than one validated instrument is available to measure an outcome, each with known but distinct properties—that is, different combinations of sensitivities, specificities, and implementation burdens. In clinical practice, a clinician ought to choose a particular test by weighing the importance of the condition, burden of treatment, risks associated with a false-positive or false-negative result, and preferences of the patient. However, in a clinical trial, where the primary goal is to determine whether the risk-benefit burden of one treatment is significantly different from another, the selection of a test or instrument used to detect potential differences in a clinical endpoint should be guided on the basis of this specific goal.

As an example, the Reevaluation of Systemic Early Neuromuscular Blockade (ROSE) RCT (www.clinicaltrials.gov identifier NCT02509078) is a multicenter, prospective, two-arm, nonblinded RCT of the effect of early neuromuscular

blockade on 90-day in-hospital mortality of patients with moderate to severe acute respiratory distress syndrome (ARDS) (1). The ROSE sample size of 1,408 patients was determined on the basis of the power necessary to detect a difference in the primary endpoint of 90-day in-hospital mortality. An important secondary endpoint is the effect of randomization to neuromuscular blockade on long-term outcomes, including post-traumatic stress (PTS)-like symptoms. There are multiple instruments available to assess PTS-like symptoms, with no simple consensus as to which is best (2–4). Among potential instruments considered for ROSE, the Impact of Event Scale-Revised (IES-R) had a higher sensitivity in a published validation study, whereas the Post-Traumatic Stress Symptoms instrument (PTSS-14) had a reported higher specificity and lower implementation burden based on its fewer number of questions.

Because it was unclear how to balance these differences in measurement properties when choosing an instrument for the ROSE trial, the present study's aim was to determine whether either of the available instruments offers superior power to detect differences in PTS symptoms across study arms over a plausible range of symptom rates. Because the sample size of a clinical trial is determined on the basis of its power to detect important differences in the primary endpoint, we hypothesized that the selection of an instrument to detect statistically significant differences in an important endpoint could be guided on the basis of each instrument's power to detect potential differences in the endpoint.

Methods

Case Study: Reevaluation of Systemic Early Neuromuscular Blockade Randomized Clinical Trial

The motivation for this investigation was the planning of the ROSE RCT, which began enrolling patients in January 2016 (1). In this trial, patients with moderate to severe ARDS are randomized to either cisatracurium besylate bolus 15 mg, followed by a continuous infusion of 37.5 mg/h for 48 hours or no routine early neuromuscular blockade. While 90-day in-hospital mortality is the study's primary outcome, an important secondary outcome

is the development of significant PTS symptoms. There are at least four credible mechanisms for development of such symptoms after receiving neuromuscular blockade. Neuromuscular blockade may

1. Require high doses of sedation, which has been linked to worse PTS-like symptoms (5);
2. Alternatively, be accompanied by inadequate sedation, resulting in “awake paralysis” (6);
3. Cause neuromuscular weakness (7), prolonging intensive care unit (ICU) stay and increasing PTS-like symptoms either directly via response to weakness or indirectly via prolonged critical illness; or
4. Reduce the inflammatory burden in ARDS, reducing ICU stay and decreasing PTS symptoms by the converse of mechanisms listed by Twigg and colleagues (3).

There is no single best strategy for measuring the development of PTS-like symptoms. The diagnosis of post-traumatic stress disorder (PTSD) in ICU patients is a subject of a nuanced literature; we have simplified it here to focus our discussion on the general issue of selecting secondary endpoints. In general, a formal diagnosis of PTSD is made by a trained clinician, often using the Clinically Administered PTSD Scale (CAPS), which can take up to 45 minutes to administer (8).

Our literature review identified two promising telephone questionnaires that had been evaluated in ICU patients. Both used reasonable references for assessing a serious burden of symptoms consistent with PTSD, which is the outcome the RCT is ultimately seeking to measure. (Because PTSD is a clinical diagnosis, we speak here of significant or burdensome PTS-like symptoms.) The IES-R is a 22-item instrument with a reported sensitivity of 100% and specificity of 85% in a critically ill population at the recommended diagnostic threshold (2). The PTSS-14 is a 14-item instrument with a reported sensitivity of 86% and specificity of 97% in a critically ill population (3). An instrument's sensitivity is its ability to correctly identify all patients who develop a condition, whereas its specificity is its ability to correctly identify all patients without the condition.

In a clinical trial, shorter instruments are advantageous because they are less costly to administer and require less respondent burden, holding all else equal. Yet, it was unclear how to trade off these measurement characteristics, given the ultimate goal of the RCT, which is to detect significant differences in the outcome between study arms, should such differences truly exist.

Power Calculations for Reevaluation of Systemic Early Neuromuscular Blockade Secondary Endpoint

The ROSE trial primary endpoint power calculation assumed a 32% 90-day mortality in the control arm. After accounting for early mortality and loss to follow-up, we estimated that approximately 1,000 patients out of the original 1,408 would be interviewed after discharge. We further simplified the present analysis by assuming that an equal number of patients would be interviewed at follow-up in each arm ($n = 500$). Best estimates of rates of PTS-like symptoms vary widely in the literature. A systematic review and meta-analysis reported prevalence rates of 4–52% across studies, and a pooled prevalence of 22% (9, 10). However, prevalence estimates do not typically adjust for the imperfect measurement characteristics of the various instruments used in these studies (11). Given the range of clinically significant PTS-like symptoms reported in the literature, we assumed a 25% rate of serious symptoms for our primary calculation, but we also examined a low estimate of 10% and a high estimate of 50% to test whether the results were sensitive to this unknown parameter.

After assuming the rate of clinical significant symptoms was 25% in the control arm, we examined scenarios of increasing symptom rates in the treatment arm and created power curves of each instrument's ability (at the recommended diagnostic threshold) to detect a difference in symptoms between trial arms. Because PTS-like symptoms must be measured using an instrument with imperfect measurement characteristics, the relationship between a true rate and an observed rate (as measured by the instrument) can be determined when the instrument's sensitivity and specificity are known:

$$p_{obs} = p_{true} \cdot (\text{sensitivity}) + (1 - p_{true}) \cdot (1 - \text{specificity}),$$

Figure 1. Comparison of the ability of the Post-Traumatic Stress Symptoms instrument (PTSS-14) and the Impact of Events Scale-Revised (IES-R) to detect a difference in post-traumatic stress-like symptoms between study arms. The PTSS-14 is a 14-item instrument with a reported sensitivity of 86% and specificity of 97% for detecting post-traumatic stress-like symptoms in patients recovering from critical illness. The IES-R is a 22-item instrument with a reported sensitivity of 100% and specificity of 85% for detecting post-traumatic stress-like symptoms in patients recovering from critical illness. A 25% baseline symptom rate was assumed in the control arm, and an increased level was assumed in the treatment arm. Power calculations assume a 500-patient sample size in each arm; the true test characteristics of each instrument were assumed to be the reported point estimates of the sensitivity and specificity of each instrument in its validation studies, and test performance would be similar over the course of the trial.

where p_{true} represents the true rate and p_{obs} represents the observed rate of a binary outcome (11). Additionally, the true difference in rates between two groups is

related to the observed difference by the following equation (12):

$$D_{true} = \frac{D_{obs}}{(\text{specificity} + \text{sensitivity} - 1)}$$

Figure 2. Comparison of the ability of the Post-Traumatic Stress Symptoms instrument (PTSS-14) and the Impact of Events Scale-Revised (IES-R) to detect a difference in post-traumatic stress-like symptoms between study arms when the IES-R's threshold is raised to improve its specificity. The PTSS-14 is a 14-item instrument with a reported sensitivity of 86% and specificity of 97% for detecting post-traumatic stress-like symptoms in patients recovering from critical illness. The IES-R is a 22-item instrument with a reported sensitivity of 75% and specificity of 92% in intensive care unit patients for detecting post-traumatic stress-like symptoms when using a higher diagnostic threshold. A 25% baseline symptom rate was assumed in the control arm, and an increased level was assumed in the treatment arm. Power calculations assume a 500-patient sample size in each arm; the true test characteristics of each instrument were assumed to be the reported point estimates of the sensitivity and specificity of each instrument in its validation studies, and test performance would be similar over the course of the trial.

Using $u = (\text{specificity} + \text{sensitivity} - 1)$ for simplicity, we can say that $D_{\text{obs}} = D_{\text{true}} \cdot u$ and $p_{\text{obs}} = (1 - \text{specificity}) + p_{\text{true}} \cdot u$. Using these formulas, one can calculate p_{obs} for each group and then calculate the power to detect a difference in proportions using standard statistical software. We calculated the power to detect a difference in rates between trial arms, given a sample size of 500 in each arm, the sensitivity and specificity of the instrument, and the underlying true event rates in each arm.

Derivation of Power Calculation and Visualization

Power calculations were performed using Stata software (version 14; StataCorp, College Station, TX); the code is provided as an appendix in the online supplement. We also derived a general formula using Mathematica software (version 10; Wolfram Research, Champaign, IL). To visualize the relationship between the sensitivity and specificity of an instrument and the power to detect a difference between two groups, we created three-dimensional surface plots in MATLAB (version 8.6; MathWorks, Inc., Natick, MA).

Human Subjects

We used no individual patient data for these analyses, so this study did not require human subjects or ethics review. The manuscript of this article was approved by the U.S. National Heart, Lung, and Blood Institute (NHLBI) Prevention and Early Treatment of Acute Lung Injury Network publications committee.

Results

Power Calculations for Secondary Outcomes in Reevaluation of Systemic Early Neuromuscular Blockade

If study patients had a 25% rate of PTS-like symptoms, the IES-R would be expected to measure a 36% symptom rate because of its lower specificity, leading to a high number of false positives. Conversely, the PTSS-14 would measure a 23% symptom rate because of its higher specificity and lower number of false positives.

Figure 1 displays the power of each instrument to measure a significant difference in symptom rates between treatment arms if a difference of a given

magnitude was truly present (i.e., the power of the trial for this secondary endpoint). The analysis assumes a 25% symptom rate in the control arm and an increasing level of symptoms in the treatment arm, with other assumptions described in the Figure 1 legend. As illustrated in the figure, the two instruments provide similar power when the baseline post-traumatic symptom rate is 25%.

However, if the underlying rate of symptoms in the control arm were 10%, the PTSS-14 would perform better, able to detect an 8% absolute difference in rate of significant symptoms between arms (10% in the control arm and 18% in the treatment arm) at 80% power. At the same

underlying rate of symptoms, the IES-R would detect only a 10% absolute difference in symptoms at the same power (10% in control and 20% in treatment). If the baseline rate of symptoms were 50%, the two instruments would perform equivalently, both able to detect an 11% absolute difference in the symptom rates at 80% power (50% in the control arm and 61% in the treatment arm). The analysis demonstrates that an instrument with high specificity, rather than high sensitivity, is better suited to detecting differences in rates between trial arms when event rates are low.

Recognizing the importance of an instrument's specificity for overall power,

A simplified version of the power calculations for comparing event rates in two arms of a clinical trial, with a two-side alpha of 0.05 is:

$$\text{Power} = \Phi \left(\frac{p_1 - p_2}{\sqrt{\sigma_{p_1 - p_2}^2}} - 1.96 \right) \quad \sigma_{p_1 - p_2}^2 = \frac{p_1(1 - p_1)}{n} + \frac{p_2(1 - p_2)}{n} \quad (1)$$

p_1 and p_2 are the event rates in each arm, Φ is the standard normal cumulative distribution function, $\sigma_{p_1 - p_2}^2$ is the variance of the sample difference, and n is the number of patients in each arm.

When p_1 and p_2 are measured with imperfect sensitivity (sn) and specificity (sp), observed event rates and true events rate are related as follows:

$$p_{\text{obs}} = p_{\text{true}} \cdot sn + (1 - p_{\text{true}}) \cdot (1 - sp)$$

Which can be rearranged to $p_{\text{obs}} = (1 - sp) + (sp + sn - 1) \cdot p_{\text{true}}$. Letting $u = sp + sn - 1$ simplifies the equation to:

$$p_{\text{obs}} = (1 - sp) + u \cdot p_{\text{true}} \quad (2)$$

If two rates are measured, $p_{1\text{obs}}$ and $p_{2\text{obs}}$, their observed difference, D_{obs} , is related to the true difference D_{true} by:

$$\begin{aligned} D_{\text{obs}} &= p_{1\text{obs}} - p_{2\text{obs}} = (1 - sp + u \cdot p_{1\text{true}}) - (1 - sp + u \cdot p_{2\text{true}}) \\ &= u \cdot (p_{1\text{true}} - p_{2\text{true}}) \quad D_{\text{obs}} = u \cdot D_{\text{true}} \end{aligned} \quad (3)$$

To express power as a function of the true average event rate, p_{true} , and true difference in rates, D_{true} , substitute $p_{\text{true}} - D_{\text{true}}/2$ and $p_{\text{true}} + D_{\text{true}}/2$ for $p_{1\text{true}}$ and $p_{2\text{true}}$ respectively into (2), which yields:

$$p_{1\text{obs}} = (1 - sp) + u \cdot (p_{\text{true}} - D_{\text{true}}/2) \quad (4)$$

$$p_{2\text{obs}} = (1 - sp) + u \cdot (p_{\text{true}} + D_{\text{true}}/2) \quad (5)$$

Substituting (3), (4) and (5) into the original power formula (1) yields:

$$\begin{aligned} \text{Power} &= \Phi \left(\frac{D_{\text{true}}}{\sqrt{VF}} \sqrt{n} - 1.96 \right) \\ VF &= -\frac{D_{\text{true}}^2}{2} - 2 \cdot p_{\text{true}}^2 + \frac{2 \cdot sp}{u^2} - \frac{2 \cdot sp^2}{u^2} - \frac{2 \cdot p_{\text{true}}}{u} + \frac{4 \cdot p_{\text{true}} \cdot sp}{u} \end{aligned}$$

Where n is the sample size per group, VF is the variance factor, which substitutes for variance given the imperfect sensitivity (sn) and specificity (sp) of the measurement instrument, and $u = sp + sn - 1$.

we also investigated the performance of the IES-R when there is a higher diagnostic threshold for detecting significant PTS symptoms. As reported in the validation study, when the IES-R was set to a higher threshold, its specificity increased to 92%, whereas its sensitivity decreased to 75%. However, at this higher threshold, the analysis showed how the IES-R still did not perform better than the PTSS-14 at detecting differences between trial arms (Figure 2). We also considered the scenario where the treatment arm's mortality rate was lower than the control arm, leading to more patients being available for follow-up in the treatment arm. In this scenario, we found there was substantial difference in the two instruments' abilities to detect PTS symptoms compared with the scenario of equal mortality and follow-up (see Figure E1 in the online supplement).

Derivation of a General Formula for Estimating Power with an Imperfect Measurement Instrument

We show a mathematical derivation of a general formula for estimating power when outcome rates are measured with imperfect sensitivity and specificity in the text box. Table 1 further explains the abbreviations used in this derivation. The derived equation expresses power as a function of the average event rate (e.g., overall prevalence), the true difference between groups, sample size, and the measurement characteristics of the instrument (sensitivity and specificity).

Visualizing the Derived Power Formula

The general formula derived in the text box allows one to calculate the power to detect a true difference in outcome rates when such

rates are measured imperfectly, and to compare the power of two instruments to detect this difference, each with a different sensitivity and specificity. In Figure 3, we graphically illustrate the relationship between sensitivity, specificity, and trial power at varying underlying event rates. For this analysis, we assume that the true absolute difference in the event rates between control and experimental groups was 20%.

The underlying event rate had a notable effect on the relationship between sensitivity, specificity, and power. When the prevalence is low, as in the case of Figure 3A, where the event rate is 10% in the control group and 30% in the treatment group, specificity is more important than sensitivity for overall power. If the specificity remains near 100%, power is maintained even as sensitivity drops to as low as 40%, whereas when sensitivity is maintained near 100%, power drops more quickly as specificity declines. This relationship is reversed at higher event rates, as shown in Figure 3C, where the event rate is 50% in the control group and 70% in the treatment group. In this case, if sensitivity remains near 100%, power is better maintained as specificity decreases.

Discussion

In this analysis, we show how clinical study power is a function of the measurement characteristics of the instrument (i.e., sensitivity and specificity) used to assess an outcome and the underlying outcome rate itself. We also show how the selection of an instrument in a clinical trial may be guided on the basis of each instrument's potential power to detect important differences in outcomes between

study arms. We used this approach to guide instrument selection for use in an RCT of early muscular blockade in ARDS, finding that the PTSS-14 may potentially provide higher power than the IES-R to detect significant differences in the rate of significant or burdensome PTS-like symptoms in this trial. As such, ROSE will use the shorter instrument, the PTSS-14, because the PTSS-14 will perform equivalently or better over the range of PTS-like symptom rates likely to be encountered during the study while minimizing respondent burden.

In clinical practice, when a clinician chooses a test to evaluate a patient for a specific disease, this choice is guided by balancing the risk associated with missing the disease or falsely diagnosing a disease and exposing the patient to an unnecessary treatment. When the condition cannot be missed or the treatment is not too burdensome, a test with high sensitivity is generally favored. When the condition is more benign or the treatment is more burdensome, a test with high specificity might be favored. For PTS symptoms, a test with high sensitivity might be preferred because of its clinical importance.

The goal of a clinical trial is to determine whether the risk-to-benefit ratio of one treatment is significantly different from that of another, which is different from the clinician's bedside task. Therefore, the selection of a test or instrument to use in a trial should be based on its ability to detect potential differences. When more than one instrument is available to measure an endpoint, our analysis shows how instrument selection can be guided by comparing each instrument's potential ability to detect these differences. When one instrument has both higher sensitivity and higher specificity, that instrument will clearly have higher power. When one instrument has higher sensitivity while the other has higher specificity, specificity tends to be more important than sensitivity for overall power when the prevalence of the event of interest is low. This may be somewhat contrary to clinical practice, where an instrument with high sensitivity may be favored because it would be less likely to miss patients who develop PTSD. Yet, at a certain point, even with very high specificity, an instrument with low sensitivity can still bring down clinical trial power if the sensitivity is so low that no true events are detected. Thus, the analytic

Table 1. Abbreviations used in mathematical formulas in the text

Abbreviation	Definition
Φ	Standard normal cumulative distribution function
D_{true}	True difference in event rates between groups
D_{obs}	Observed difference in event rates between groups
p_{true}	Overall average prevalence of the event
sn	Instrument sensitivity
sp	Instrument specificity
σ_{p1-p2}^2	Sample variance of the difference in event rates
u	Specificity + sensitivity - 1
VF	Variance factor, which is σ_{p1-p2}^2 while incorporating imperfect measurement
n	Sample size in each arm of the trial, so that total trial sample size is $2n$

Figure 3. Statistical power to measure a 20% absolute difference in the rate of an outcome across varying instrument sensitivity and specificity, and underlying outcome prevalence. (A) Outcome rate of 10% in control group and 30% in treatment group, (B) outcome rate of 25% in control group and 45% in treatment group, and (C) outcome rate of 50% in control group and 70% in treatment group.

Figure 3. (Continued).

approach we describe can be used to compare candidate instruments and examine these trade-offs when one has higher sensitivity while the other has higher specificity.

This approach may be of interest to others designing RCTs or seeking to maximize the power of cohort studies whose sample size is fixed. Good research practice seeks to estimate the plausible effect size for the primary endpoint on the basis of information in the literature and minimal clinically significant differences (whether in continuous scales or above a diagnostic threshold). This estimate guides calculation of the sample size necessary to detect such an effect size. Explicitly considering the impact of imperfect outcome detection can also maximize the power of the trial and lessen the risk of missing an important difference in an endpoint that could also tip the balance of the risk-to-benefit ratio for a particular treatment. The logic developed here for maximizing the power of secondary outcomes could also inform primary outcome selection when sample size is fixed for other pragmatic reasons; improving the reliability of all measurements in an RCT may substantially improve overall power (13).

Our analysis also serves as a reminder of the importance of accounting for imperfect measurement when interpreting observations more generally. In our case study, the two instruments considered were predicted to measure outcome rates of 23% and 36%—a more than 50% relative difference—when the true underlying rate was 25%. This may help to explain the wide range of prevalence estimates of long-term outcomes across different cohort studies when different instruments are used (9). Adjustment for the imperfect measurement characteristics of an instrument could be done to improve the accuracy and comparability of studies, particularly systematic reviews. However, such an adjustment depends on whether the estimated sensitivity and specificity of tests are correct. The present analysis also depends on this assumption. Validation studies of tests cannot estimate a test's characteristics perfectly and necessarily report confidence intervals surrounding these estimates.

In the present analysis, we made the simplifying assumption that the point estimate of each instrument's sensitivity and specificity as reported in the validation

studies represented the true test characteristics. We did not attempt to account for the confidence intervals surrounding the sensitivity and specificity estimates. In future work, researchers could consider ways of incorporating confidence intervals from different validation samples to ensure that the most relevant measurement characteristics are incorporated.

In our primary analysis, we assumed that there would be no difference in mortality or availability for follow-up between trial arms. In our supplemental analysis, we evaluated the effect of unequal sample size between groups due to differences in mortality, and we assumed that mortality does not provide information about (and is uncorrelated with) rates of secondary outcome. Although a difference in a clinical trial's primary outcome may make secondary outcomes appear less important, there still may be the real possibility that a treatment provides benefit to some patients but also causes harm through a different mechanism (a form of treatment effect heterogeneity) (14). Thus, making measurement of differences in secondary outcomes as accurate as possible

is always of critical importance when quantifying the particular risks and benefits of a specific therapy for individual patients.

This work has additional limitations that may make the analysis not applicable to all scenarios. If the sensitivity and specificity of an instrument are not known in a population under study, the analysis is not feasible. Recent efforts to measure the psychometrics of commonly used instruments in critical care populations are therefore particularly important (2, 15–17). If the secondary outcome does not require a measurement instrument (e.g., reintubation), this type of analysis would also be unnecessary. If instruments give a continuous score that should not be dichotomized, the analytic approach to determining the best instrument would also be different. We chose to dichotomize the PTS syndrome endpoint, which both simplified the analysis and made it more interpretable. An alternative approach would be to keep the PTS symptom measures on a continuous scale and

compare mean scores between groups. This approach may improve the power to detect small differences in symptom levels between trial arms. However, by dichotomizing this outcome, we focus here on detecting clinically significant differences across a burden threshold, a common approach to diagnosis.

Optimizing the measurement of any specific secondary endpoint does not diminish the risks of making a false discovery when performing multiple comparisons across multiple secondary endpoints. We suggest using a structured approach to manage the potential for false discoveries rather than haphazardly diluting the power of a secondary endpoint that has substantial clinical importance (18). Finally, the derivation of the general formula for power in this specific context relies extensively on well-established principles of statistics; no claim of substantial statistical originality is made. Our contribution in this work is to examine implications for planning of RCTs.

Conclusions

The power of an RCT to detect differences in secondary outcomes depends not only on the true effect of the treatment under study but also on the underlying outcome rates and measurement characteristics of the measurement instrument. Selecting instruments for use during an RCT can be guided by maximizing this power. Doing so ensures that the unique opportunity afforded by an RCT will yield the greatest increase in scientific knowledge. ■

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

Acknowledgment: The authors thank all members of the U.S. National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury Network Long-Term Outcomes Committee and the ROSE Protocol Committee for their insightful comments that motivated us to write this article. This article does not necessarily represent the views of the U.S. government or the Department of Veterans Affairs.

References

- 1 ClincialTrials.gov. Reevaluation of Systemic Early Neuromuscular Blockade (ROSE) [accessed 2015 Dec 1]. Available from: <https://clinicaltrials.gov/ct2/show/NCT02509078>
- 2 Bienvenu OJ, Williams JB, Yang A, Hopkins RO, Needham DM. Posttraumatic stress disorder in survivors of acute lung injury: evaluating the Impact of Event Scale-Revised. *Chest* 2013;144:24–31.
- 3 Twigg E, Humphris G, Jones C, Bramwell R, Griffiths RD. Use of a screening questionnaire for post-traumatic stress disorder (PTSD) on a sample of UK ICU patients. *Acta Anaesthesiol Scand* 2008;52:202–208.
- 4 Stoll C, Kapfhammer HP, Rothenhäusler HB, Haller M, Briegel J, Schmidt M, Krauseneck T, Durst K, Schelling G. Sensitivity and specificity of a screening test to document traumatic experiences and to diagnose post-traumatic stress disorder in ARDS patients after intensive care treatment. *Intensive Care Med* 1999;25:697–704.
- 5 Jones C, Bäckman C, Capuzzo M, Flaatten H, Rylander C, Griffiths RD. Precipitants of post-traumatic stress disorder following intensive care: a hypothesis generating study of diversity in care. *Intensive Care Med* 2007;33:978–985.
- 6 Amot-Smith J, Smith AF. Patient safety incidents involving neuromuscular blockade: analysis of the UK National Reporting and Learning System data from 2006 to 2008. *Anaesthesia* 2010;65:1106–1113.
- 7 Leatherman JW, Fluegel WL, David WS, Davies SF, Iber C. Muscle weakness in mechanically ventilated patients with severe asthma. *Am J Respir Crit Care Med* 1996;153:1686–1690.
- 8 Blake DD, Weathers FW, Nagy LM, Kaloupek DG, Gusman FD, Charney DS, Keane TM. The development of a Clinician-Administered PTSD Scale. *J Trauma Stress* 1995;8:75–90.
- 9 Parker AM, Sricharoenchai T, Raparla S, Schneek KW, Bienvenu OJ, Needham DM. Posttraumatic stress disorder in critical illness survivors: a metaanalysis. *Crit Care Med* 2015;43:1121–1129.
- 10 Jackson JC, Pandharipande PP, Girard TD, Brummel NE, Thompson JL, Hughes CG, Pun BT, Vasilevskis EE, Morandi A, Shintani AK, et al.; Bringing to light the Risk Factors And Incidence of Neuropsychological dysfunction in ICU survivors (BRAIN-ICU) study investigators. Depression, post-traumatic stress disorder, and functional disability in survivors of critical illness in the BRAIN-ICU study: a longitudinal cohort study. *Lancet Respir Med* 2014;2:369–379.
- 11 Shoukri MM. Measures of interobserver agreement. Boca Raton, FL: Chapman & Hall/CRC; 2004.
- 12 Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Hoboken, NJ: John Wiley and Sons; 2003.
- 13 Sjoding MW, Cooke CR, Iwashyna TJ, Hofer TP. Acute respiratory distress syndrome measurement error: potential effect on clinical study results. *Ann Am Thorac Soc* 2016;13:1123–1128.
- 14 Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of heterogeneity of treatment effect for reporting and analysis of randomized trials in critical care. *Am J Respir Crit Care Med* 2015;192:1045–1051.
- 15 Jutte JE, Needham DM, Pfoh ER, Bienvenu OJ. Psychometric evaluation of the Hospital Anxiety and Depression Scale 3 months after acute lung injury. *J Crit Care* 2015;30:793–798.
- 16 Pfoh ER, Chan KS, Dinglas VD, Girard TD, Jackson JC, Morris PE, Hough CL, Mendez-Tellez PA, Ely EW, Huang M, et al.; NIH NHLBI ARDS Network. Cognitive screening among acute respiratory failure survivors: a cross-sectional evaluation of the Mini-Mental State Examination. *Crit Care* 2015;19:220.
- 17 Chan KS, Pfoh ER, Denehy L, Elliott D, Holland AE, Dinglas VD, Needham DM. Construct validity and minimal important difference of 6-minute walk distance in survivors of acute respiratory failure. *Chest* 2015;147:1316–1326.
- 18 Cao J, Zhang S. Multiple comparison procedures. *JAMA* 2014;312:543–544.