Acute Respiratory Distress Syndrome Measurement Error Potential Effect on Clinical Study Results

Michael W. Sjöding^{1,2}, Colin R. Cooke^{1,2}, Theodore J. Iwashyna^{1,3,4,5}, and Timothy P. Hofer^{1,2,3}

¹Department of Internal Medicine and ²Institute for Healthcare Policy & Innovation, University of Michigan, Ann Arbor, Michigan; ³VA Center for Clinical Management Research, Ann Arbor, Michigan; ⁴Institute for Social Research, Ann Arbor, Michigan; and ⁵Australian and New Zealand Intensive Care Research Centre, Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

Abstract

Rationale: Identifying patients with acute respiratory distress syndrome (ARDS) is a recognized challenge. Experts often have only moderate agreement when applying the clinical definition of ARDS to patients. However, no study has fully examined the implications of low reliability measurement of ARDS on clinical studies.

Objectives: To investigate how the degree of variability in ARDS measurement commonly reported in clinical studies affects study power, the accuracy of treatment effect estimates, and the measured strength of risk factor associations.

Methods: We examined the effect of ARDS measurement error in randomized clinical trials (RCTs) of ARDS-specific treatments and cohort studies using simulations. We varied the reliability of ARDS diagnosis, quantified as the interobserver reliability (κ -statistic) between two reviewers. In RCT simulations, patients identified as having ARDS were enrolled, and when measurement error was present, patients without ARDS could be enrolled. In cohort studies, risk factors as potential predictors were analyzed using reviewer-identified ARDS as the outcome variable.

Measurements and Main Results: Lower reliability measurement of ARDS during patient enrollment in RCTs seriously degraded study power. Holding effect size constant, the sample size necessary to attain adequate statistical power increased by more than 50% as reliability declined, although the result was sensitive to ARDS prevalence. In a 1,400-patient clinical trial, the sample size necessary to maintain similar statistical power increased to over 1,900 when reliability declined from perfect to substantial ($\kappa = 0.72$). Lower reliability measurement diminished the apparent effectiveness of an ARDS-specific treatment from a 15.2% (95% confidence interval, 9.4–20.9%) absolute risk reduction in mortality to 10.9% (95% confidence interval, 4.7–16.2%) when reliability declined to moderate ($\kappa = 0.51$). In cohort studies, the effect on risk factor associations was similar.

Conclusions: ARDS measurement error can seriously degrade statistical power and effect size estimates of clinical studies. The reliability of ARDS measurement warrants careful attention in future ARDS clinical studies.

Keywords: acute lung injury; diagnosis; clinical trial; observational study; bias

(Received in original form January 25, 2016; accepted in final form March 7, 2016)

This work was supported by National Heart, Lung, and Blood Institute grant T32HL007749 (M.W.S.), Department of Veterans Affairs Health Services Research and Development Services grant IIR 11-109 (T.J.I.), and Agency for Healthcare Research and Quality grant K08HS020672 (C.R.C.). The study funders had no role in the study design, analysis or interpretation of the data, or writing of the report for these analyses.

Author Contributions: M.W.S.: takes full responsibility for the content of the article, including the data integrity and accuracy of the analysis; M.W.S., C.R.C., T.J.I., and T.P.H.: were responsible for the simulation study design, analysis and interpretation of results, drafting of the manuscript, critical revision of the manuscript for important intellectual content, and approval of the final manuscript.

Correspondence and requests for reprints should be addressed to Michael W. Sjöding, 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 13, No 7, pp 1123–1128, Jul 2016

Copyright © 2016 by the American Thoracic Society

DOI: 10.1513/AnnalsATS.201601-072OC

Internet address: www.atsjournals.org

An estimated 200,000 patients develop the acute respiratory distress syndrome (ARDS) and 74,500 die as a result of it in the United States each year (1). Accurately identifying

patients with ARDS is a recognized challenge (2, 3). ARDS is currently identified using clinical criteria and cannot be measured with perfect precision or

reliability (4). When the reliability of ARDS diagnosis has been examined, studies have demonstrated that even ARDS experts often have only moderate agreement when

applying the clinical definition to patients (5–7). In clinical practice, where less attention may be paid to ARDS measurement accuracy, there is reason to believe reliability is even lower. For example, diagnoses made using other tools, such as echocardiography, have been shown to be less reliable at local centers compared with central laboratories (8).

Poor reliability during enrollment in ARDS clinical trials will lead to misclassification of some patients, potentially reducing statistical power and the ability to measure the true efficacy of an intervention. This problem may be magnified if reliability in real-world practice is substantially worse than in clinical trials. As a consequence of poor reliability in practice, some patients with ARDS will go unrecognized and miss out on lifesaving therapies (2), while others without ARDS could receive ARDS-specific treatments that may be of questionable benefit to patients without ARDS. Although the implications of ARDS measurement error have been conceptually described (9, 10), the true impact of such errors on clinical study results and patient care has not been thoroughly quantified.

In the present study, we examined the potential effect of ARDS measurement error on clinical study results and statistical power by simulating clinical studies in which ARDS was measured with varying reliability. We hypothesized that the degree of ARDS measurement or diagnostic error commonly reported in clinical studies has a substantial effect on study results, impacting a study's ability to accurately measure the relationship between a risk factor or a treatment and clinical outcomes. A portion of this work has been accepted as an abstract to be presented at the American Thoracic Society annual conference in San Francisco, CA, May 2016 (11).

Methods

We simulated two study types where ARDS measurement error may impact study results: (1) randomized clinical trials (RCTs) of targeted ARDS-specific treatments and (2) observational cohort studies investigating ARDS risk factors among patients receiving mechanical ventilation. An overview of each simulation is provided below; technical details and

simulation code for Stata 14.0 statistical software (StataCorp, College Station, TX) are provided in the online supplement.

We reviewed the literature to determine the range of ARDS measurement reliabilities to use in the simulations (*see* online supplement for details), finding published κ -values ranging between 0.47 and 0.91 (1, 6, 7, 12). We also reviewed all clinical trials published in top-tier medical journals within the last 10 years to determine how often reliability is assessed in the context of clinical trials. Of 43 ARDS clinical studies identified, we found no study in which researchers measured or reported the reliability of ARDS measurement.

Simulating ARDS Measurement Error

To model measurement error, we simulated hypothetical reviewers with set misclassification rates (sensitivity or specificity <100%). On the basis of these misclassification rates, some patients with ARDS were randomly misclassified as not having ARDS, while others without ARDS were randomly misclassified as having ARDS. Although the true ARDS status of each patient was always predetermined in each simulation, all study analyses was performed using the imperfect reviewer assignment to replicate the impact of measurement error in the real world.

During each simulated study, the classification procedure was performed twice on 100 randomly selected patients to estimate the reliability of ARDS measurement using the κ -statistic. When classification error rates were zero and interrater reliability was perfect ($\kappa = 1.0$), all patients were analyzed correctly. To qualify agreement, κ -values of 0.8–1 were defined as almost perfect agreement, 0.61–0.8 as substantial agreement, 0.41–0.6 as moderate agreement, and 0.21–0.4 as fair agreement, following the widely used classification of Landis and Koch (13).

Estimating ARDS Clinical Trial Power

To examine how measurement error can affect the statistical power of ARDS clinical trials, we performed power calculations for an ARDS RCT under varying degrees of ARDS measurement error. Power calculations were based on a newly planned ARDS RCT (14). The primary outcome was all-cause 90-day mortality, expected to be 35% in the control arm and 27% in the

treatment arm. Patients without ARDS enrolled due to misclassification were assumed to have a mortality rate similar to the 27% in the treatment arm. Power calculations were based on the clinical scenario where patients with ARDS were screened and enrolled from a patient population with hypoxic respiratory failure with an ARDS prevalence of between 25% and 75%. The rate of patients without ARDS incorrectly enrolled was estimated on the basis of these misclassification rates leading to low reliability measurement. The actual mortality rate of patients enrolled in each arm was calculated as the weighted average among the enrolled patients with and without ARDS. The sample size necessary to detect a mortality difference between groups with 90% power, as well as the power of a 1,500-patient trial, was calculated on the basis of a comparison of binomial proportions with an overall $\alpha = 0.05$.

RCT Treatment Effect Estimates

To simulate how ARDS measurement error could bias treatment effect estimates in RCTs, we created a hypothetical patient cohort that closely modeled hospital mortality risk in patients receiving mechanical ventilation in U.S. Department of Veterans Affairs hospitals. This dataset has a mortality prediction score with a c-statistic greater than 0.85 (15). Development of ARDS among patients in this cohort was randomly assigned, but it was weighted such that patients with a higher risk of hospital mortality were more likely to develop ARDS, consistent with multiple studies (1, 16, 17). The simulations were designed so that the underlying prevalence of ARDS in the cohort would be 25%.

Simulated trials were performed under two treatment scenarios. In the first, the treatment reduced the risk of death by 33% only among patients with true ARDS. However, the treatment provided no benefit to patients without ARDS and caused no adverse effects. In the second, the treatment's beneficial effect was the same but also had a 3% fatal adverse event rate for all patients (18). During each simulated clinical trial, patients from the hypothetical cohort were assessed for ARDS (with measurement error as described above). Those patients identified with ARDS were enrolled and randomized to the treatment or control arm of the trial. Whether patients died during the

study was then simulated, conditional on their baseline mortality risk, true ARDS status, and treatment received. (Details are described in the online supplement.) Finally, the measured absolute risk reduction in mortality was calculated for the treatment.

Simulating ARDS Risk Factor Studies

To simulate observational cohort studies investigating ARDS risk factors, we first built a hypothetical patient cohort that replicated characteristics of patients described in a published cohort of patients receiving mechanical ventilation (19). Patients were mechanically ventilated for more than 48 hours and did not have ARDS at mechanical ventilation onset. Patients were assigned a specific reason for respiratory failure (pneumonia, sepsis, trauma, or other diagnosis) so that rates of each diagnosis matched those in the published cohort. A mechanical ventilation tidal volume was assigned to each patient so that the distribution of tidal volumes matched those in the published study. We were unable to model the correlation structure of covariates, as this data was unavailable. We estimated each patient's probability of developing ARDS based upon their specific set of covariates and the cohort's baseline risk. We used odds ratios published in the study by Gajic and colleagues (19) to represent the true strength of risk factor associations with ARDS development in the simulations.

During each simulation, we first sampled patients from the hypothetical cohort. Whether patients in the sample actually developed ARDS was then simulated on the basis of their individual ARDS risk. An assessment of each patient's ARDS status was then performed

(with measurement error as described above). Finally, the strength of association between each risk factor and ARDS development was measured using multivariable logistic regression, setting the measured ARDS status (not a patient's true status) as the outcome variable. Risk factors with $P < 0.05$ were considered statistically significant.

Statistical Analysis

All simulations were performed 1,000 times to calculate median and 95% confidence interval (CI) estimates of the primary outcome (odds ratio or absolute mortality reduction). Code for simulations is provided in the online supplement. In this simulation work, we did not use individual patient data, and the study did not require institutional review board approval.

Results

RCT Simulations

We calculated the number of patients necessary to attain 90% power in RCTs of an ARDS-specific therapy when ARDS measurement was imperfect. Assuming event rates of 27% in the treatment arm and 35% in the control arm, with perfect ARDS measurement, the number of patients needed to detect a statistically significant benefit was 1,402. However, when the reliability of ARDS measurement declined but remained substantial ($\kappa = 0.72$), the sample size necessary to attain similar power increased to 1,968. When reliability was moderate ($\kappa = 0.51$), the sample size necessary further increased to 2,726 (Table 1). These sample size requirements were sensitive to the prevalence of ARDS among patients screened, and lower

reliability measurement had less effect on power when the underlying prevalence of ARDS in the population was higher (Figure 1).

In simulated RCTs of a targeted ARDS treatment, the absolute mortality benefit of an ARDS-specific treatment declined as the reliability of ARDS measurement decreased and the number of patients without ARDS enrolled increased (Table 2). Patients were enrolled from a population receiving mechanical ventilation with an ARDS prevalence of 25%, and patients with ARDS had a 10% higher average mortality risk than patients without ARDS (45% vs. 35%). When $\kappa = 1.0$ and the true treatment effect was a relative risk reduction of 33%, the absolute mortality reduction with treatment was 15.3% (95% CI, 9.4–20.9%). However, the absolute mortality reduction measured in the study declined to 10.9% (95% CI, 4.7–16.2%) when reliability was moderate ($\kappa = 0.51$ between reviews). In this scenario, on average, 290 of the 1,000 patients enrolled did not have ARDS. In addition, when the ARDS-specific treatment had a 3% adverse event rate, 145 patients without ARDS were recruited into the treatment arm, and 4 patients without ARDS had a fatal adverse event.

Cohort Study Simulations

In simulated observational cohort studies, the association between pneumonia and ARDS development declined as ARDS measurement error increased and reliability decreased (Table 3). Patients in this cohort had characteristics identical to patients in the International Mechanical Ventilation Study database (19, 20), with an average age of 59 years, average Simplified Acute Physiology Score II of 45 at intensive care unit admission, 16% with a primary

Table 1. Randomized clinical trial sample size and trial power estimation when acute respiratory distress syndrome enrollment is imperfect from a patient cohort with 25% acute respiratory distress syndrome prevalence

Interobserver Agreement	κ -Statistic	Power in 1,500-Patient Trial	Sample Size for 90% Power	Sample Inflation
Perfect	1.00	0.92	1,402	1.00
Almost perfect	0.85	0.87	1,664	1.19
Substantial	0.72	0.81	1,968	1.40
	0.61	0.74	2,320	1.65
Moderate	0.51	0.67	2,726	1.94
	0.42	0.60	3,198	2.28

For power calculations, the primary outcome was all-cause 90-day mortality, assumed to be 35% in control arm and 27% in the treatment arm (modeled after a newly planned acute respiratory distress syndrome [ARDS] clinical trial). Patients without ARDS enrolled due to misclassification were assumed to have a 27% mortality rate. Power calculations for a 1,500-patient trial were based on a comparison of binomial proportions with an $\alpha = 0.05$.

Figure 1. Sample size requirement for adequate statistical power in a randomized clinical trial as acute respiratory distress syndrome (ARDS) measurement reliability worsens. Sample inflation is the amount the original sample size must be multiplied by to maintain adequate power as reliability worsens. Power calculations were done for the primary outcome of all-cause 90-day mortality, assumed to be 35% in the control arm and 27% in the treatment arm. Patients without ARDS enrolled due to misclassification were assumed to have a 27% mortality rate. Power calculations were based on a comparison of binomial proportions with an $\alpha = 0.05$.

diagnosis of pneumonia, and 6.2% of the cohort who developed ARDS. When ARDS was measured perfectly ($\kappa = 1.0$ between reviews), the odds ratio for ARDS development among patients with pneumonia was 2.05 (95% CI, 1.40–2.98). However, the effect size declined to 1.60

(95% CI, 1.12–2.20) in simulations where ARDS was measured with moderate reliability ($\kappa = 0.52$). The percentage of studies concluding this true association was statistically significant also declined. In studies of 3,000 patients, 94% correctly identified the existence of an association

between pneumonia and ARDS that was statistically significant when $\kappa = 1.0$, but this declined to 77% when the κ -statistic was 0.52. For studies of 1,000, only 54% concluded the association was significant when $\kappa = 1.0$ and only 28% concluded the association was significant when the κ -statistic was 0.52.

Discussion

In the present study, we simulated the results of several prototypical ARDS clinical studies to quantify the impact of ARDS measurement error on their results. We found that measuring ARDS with lower reliability—here defined as “moderate” agreement between two reviewers by the Landis and Koch classification (13)—had a substantial effect on study results and statistical power. As studies were performed with increasing measurement error, their ability to accurately measure true relationships between risk factors or treatments and clinical outcomes rapidly diminished. Studies were equally affected whether ARDS was the primary outcome variable or when patients were enrolled on

Table 2. Measured effect of an acute respiratory distress syndrome–specific treatment on mortality in a 1,000-patient randomized clinical trial when acute respiratory distress syndrome is measured imperfectly

ARDS-Specific Treatment with 33% Relative Risk Reduction in Mortality			
Agreement	κ -Statistic	Measured Absolute Mortality Reduction	Patients without ARDS Enrolled
Perfect	1.00	15.3% (9.4–20.9%)	0
Almost perfect	0.85	14% (8–19.9%)	86
Substantial	0.72	12.7% (7.1–18.8%)	160
	0.61	11.8% (5.9–17.6%)	228
Moderate	0.51	10.9% (4.7–16.2%)	290
	0.42	10.1% (4–16.2%)	347

ARDS-Specific Treatment with 33% Relative Risk Reduction and 3% Fatal Adverse Event Rate				
Agreement	κ -Statistic	Measured Absolute Mortality Reduction	Patients without ARDS Given Treatment	Patients without ARDS with Fatal Adverse Events
Perfect	1.00	13.2% (7.3–19.6%)	0	0
Almost perfect	0.85	12% (6.3–17.8%)	55	1
Substantial	0.72	10.8% (4.7–16.8%)	97	2
	0.61	9.7% (3.5–15.4%)	133	3
Moderate	0.51	8.8% (3.3–14.4%)	167	4
	0.42	8.1% (1.5–14.2%)	197	5

In the simulation, trial screening and enrollment occurred among patients receiving mechanical ventilation with a mortality risk distribution modeled after patients receiving mechanical ventilation in Veterans Affairs hospitals. Simulations were designed so that the acute respiratory distress syndrome (ARDS) prevalence was 25% in this cohort, and patients with ARDS had a 10% higher mortality risk than patients without ARDS (45% vs. 35%). The ARDS-specific treatment reduced this mortality risk by 33% among patients with ARDS, but in the second simulation it also had a 3% fatal adverse event rate in all patients. When measurement error was present, as reflected in a lower κ -statistic, some patients without ARDS would incorrectly be enrolled.

Table 3. Strength of a risk factor association with acute respiratory distress syndrome development and rates of concluding the association is statistically significant when acute respiratory distress syndrome is measured imperfectly

Interobserver Agreement	κ -Statistic	Pneumonia (95% CI)	Statistical Significance Rate	
			Studies with 3,000 Subjects	Studies with 1,000 Subjects
Perfect	1.00	2.05 (1.40–2.98)	0.94	0.54
Almost perfect	0.80	1.85 (1.29–2.66)	0.91	0.49
Substantial	0.64	1.71 (1.19–2.40)	0.83	0.46
Moderate	0.52	1.60 (1.12–2.20)	0.77	0.36
	0.40	1.51 (1.06–2.11)	0.67	0.28

In the simulation, a cohort of patients receiving mechanical ventilation with characteristics identical to those described in the International Mechanical Ventilation Study database were used (19). Acute respiratory distress syndrome (ARDS) development was simulated on the basis of each patient's underlying probability of ARDS development, and the true association between pneumonia and ARDS development was an odds ratio of 2.05 (18). When measurement error was present, as reflected in a lower κ -statistic, reviewers misclassified some patients' ARDS status, but associations between risk factors and ARDS development were analyzed using reviewer-identified ARDS as the outcome variable.

the basis of meeting the clinical definition of ARDS. Yet, our review of recent ARDS clinical studies showed that none assessed the reliability of ARDS diagnosis.

Measurements are often described as having two basic properties: validity and reliability (21). Concerns with the validity of ARDS measurement have been raised in studies demonstrating that the clinical definition of ARDS is only loosely correlated with pathologic findings of diffuse alveolar damage on lung biopsy or autopsy (22, 23). Reliability concerns have been raised by studies demonstrating that two expert reviewers have only moderate agreement when evaluating a cohort for patients with ARDS (5). In the present study, we examined the effects of poor reliability of ARDS measurement and demonstrated its substantial effect on study results across study types.

There are a number of approaches to handling low-reliability measures. The most straightforward is to ensure reviewers undergo standardized training, use standardized abstraction tools, and have extensive quality control in place. With this approach, Rubenfeld and colleagues reported a $\kappa = 0.91$ when examining the incidence of ARDS in King County, Washington (1). When very high reliability is necessary, another approach is to average multiple measurements by different reviewers. It is also possible to adjust for measurement error during study analysis (24). However, while analytic procedures after the collection of data can correct for bias in the estimate of effect, they cannot compensate for the loss of precision (or power) that necessarily accompanies measurement error. For this reason, efforts to reduce measurement error by improving a study's design are preferable to

making corrections for measurement error in the analysis, but these not mutually exclusive options.

Performing an analysis on a more reliably identified subgroup, through either biomarkers or a modified clinical definition, may also be appropriate in some studies, at the risk of reducing the generalizability of the results (7). This would also limit the number of patients available for analysis, reducing a study's power to detect statistically significant effects if additional patients are not recruited into the study. Finally, consensus-based modifications to a clinical definition can improve reliability. Yet, this is a substantial undertaking that to our knowledge has occurred only twice in the case of ARDS (4, 25).

In practice, pilot work for major clinical trials of ARDS-specific treatments could estimate the reliability with which patients are classified as having ARDS and recruited into the trial. During the pilot study, if several independent determinations of whether patients have ARDS are performed, an estimate of the measures' reliability can be calculated. This information could allow investigators to better account for imperfect ARDS recruitment when making sample size calculations for the full trial. It might also motivate changes to the ARDS measurement procedure used during recruitment itself, so as to improve reliability of recruitment for the trial and in turn decrease the sample size requirement.

In addition to research settings, ARDS measurement error is a well-described problem in clinical practice. When poor sensitivity drives such error, many patients with ARDS may go unrecognized and may not receive evidence-based treatments

(2, 26). Low reliability of ARDS diagnoses in clinical practice could reduce or eliminate the population benefit of disseminating interventions found to be efficacious in clinical trials. Thus, there is a pressing need to design more tools that help clinicians correctly identify and care for patients with ARDS and build these tools into electronic health record systems (27). Clinical trials could estimate a minimum diagnostic precision necessary for the clinical benefits found in the trial to be realized in practice.

Limitations

Our study should be interpreted in the context of several limitations. These results are simulations of hypothetical ARDS clinical studies. For most published studies, the reliability of ARDS diagnosis or the possible range over which reliability might vary is unknown. In the present study, these simulations required additional modeling assumptions that may not have fully replicate clinical populations; however, when possible, the simulations were modeled after actual published ARDS studies or real patient data. The hypothetical treatments were effective only in patients with true ARDS, not other patients with respiratory failure. In practice, some important ARDS treatments may convey some benefit to other high-risk patients, although the degree to which this is the case is unknown (28). Finally, we modeled reliability, not validity. Threats to validity would also introduce substantial bias if the current clinical definition of ARDS failed to accurately identify patients with the syndrome.

As shown in the present study, random (nondifferential) measurement error in a

single variable reduces its association with another variable, echoing a commonly held belief that measurement error always biases to the null. However, it is not uncommon for measurement error to be nonrandom or to be present in several correlated variables (e.g., pneumonia and ARDS). In these scenarios, it is nearly impossible to predict the bias of the resulting estimates, further emphasizing the importance of estimating and reducing measurement error of key variables in any analysis.

At the reliability level that ARDS is currently measured, ARDS clinical trials may be significantly underpowered and treatment effects may be underestimated. These results can be generalized to any other medical condition that has a clinical definition with low reliability. For example, a published report of the reliability of the sepsis definition ($\kappa = 0.66$ among experts) suggests these would be immediately applicable to patients with sepsis (29).

Conclusions

ARDS measurement error can seriously degrade statistical power and effect size estimates of clinical studies. These results would also apply to any other medical condition that has a clinical definition that could be measured with error. The reliability of ARDS measurement warrants careful attention in future ARDS clinical studies. ■

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

References

- 1 Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, Stern EJ, Hudson LD. Incidence and outcomes of acute lung injury. *N Engl J Med* 2005;353:1685–1693.
- 2 Herasevich V, Yilmaz M, Khan H, Hubmayr RD, Gajic O. Validation of an electronic surveillance system for acute lung injury. *Intensive Care Med* 2009;35:1018–1023.
- 3 Ferguson ND, Frutos-Vivar F, Esteban A, Fernández-Segoviano P, Aramburu JA, Nájera L, Stewart TE. Acute respiratory distress syndrome: underrecognition by clinicians and diagnostic accuracy of three clinical definitions. *Crit Care Med* 2005;33:2228–2234.
- 4 Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS; ARDS Definition Task Force. Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012;307:2526–2533.
- 5 Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver variability in applying a radiographic definition for ARDS. *Chest* 1999;116:1347–1353.
- 6 Meade MO, Cook RJ, Guyatt GH, Groll R, Kachura JR, Bedard M, Cook DJ, Slutsky AS, Stewart TE. Interobserver variation in interpreting chest radiographs for the diagnosis of acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2000;161:85–90.
- 7 Shah CV, Lanken PN, Localio AR, Gallop R, Bellamy S, Ma SF, Flores C, Kahn JM, Finkel B, Fuchs BD, et al. An alternative method of acute lung injury classification for use in observational studies. *Chest* 2010;138:1054–1061.
- 8 Hole T, Otterstad JE, St John Sutton M, Frøland G, Holme I, Skjærpe T. Differences between echocardiographic measurements of left ventricular dimensions and function by local investigators and a core laboratory in a 2-year follow-up study of patients with an acute myocardial infarction. *Eur J Echocardiogr* 2002;3:263–270.
- 9 Schuster DP. Identifying patients with ARDS: time for a different approach. *Intensive Care Med* 1997;23:1197–1203.
- 10 Wood KA, Huang D, Angus DC. Improving clinical trial design in acute lung injury. *Crit Care Med* 2003;31(4, Suppl):S305–S311.
- 11 Sjoding MW, Iwashyna TJ, Cooke CR, Hofer TP. Potential impact of acute respiratory distress syndrome measurement error on clinical study results [abstract]. *Am J Respir Crit Care Med* 193;2016:A1857.
- 12 Hendrickson CM, Dobbins S, Redick BJ, Greenberg MD, Calfee CS, Cohen MJ. Misclassification of acute respiratory distress syndrome after traumatic injury: the cost of less rigorous approaches. *J Trauma Acute Care Surg* 2015;79:417–424.
- 13 Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363–374.
- 14 Reevaluation of Systemic Early Neuromuscular Blockade (ROSE). ClinicalTrials.gov identifier NCT02509078 [updated 2016 Jan 19; accessed 2016 May 2]. Available from: <https://clinicaltrials.gov/ct2/show/NCT02509078>
- 15 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.
- 16 Shah CV, Localio AR, Lanken PN, Kahn JM, Bellamy S, Gallop R, Finkel B, Gracias VH, Fuchs BD, Christie JD. The impact of development of acute lung injury on hospital mortality in critically ill trauma patients. *Crit Care Med* 2008;36:2309–2315.
- 17 Mikkelsen ME, Shah CV, Meyer NJ, Gaieski DF, Lyon S, Miltiades AN, Goyal M, Fuchs BD, Bellamy SL, Christie JD. The epidemiology of acute respiratory distress syndrome in patients presenting to the emergency department with severe sepsis. *Shock* 2013;40:375–381.
- 18 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.
- 19 Gajic O, Frutos-Vivar F, Esteban A, Hubmayr RD, Anzueto A. Ventilator settings as a risk factor for acute respiratory distress syndrome in mechanically ventilated patients. *Intensive Care Med* 2005;31:922–926.
- 20 Esteban A, Anzueto A, Frutos F, Alía I, Brochard L, Stewart TE, Benito S, Epstein SK, Apezteguia C, Nightingale P, et al.; Mechanical Ventilation International Study Group. Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. *JAMA* 2002;287:345–355.
- 21 Carmines EG, Zeller RA. Reliability and validity assessment. Beverly Hills, CA: Sage; 1979.
- 22 Thille AW, Esteban A, Fernández-Segoviano P, Rodriguez JM, Aramburu JA, Peñuelas O, Cortés-Puch I, Cardinal-Fernández P, Lorente JA, Frutos-Vivar F. Comparison of the Berlin Definition for acute respiratory distress syndrome with autopsy. *Am J Respir Crit Care Med* 2013;187:761–767.
- 23 Kao KC, Hu HC, Chang CH, Hung CY, Chiu LC, Li SH, Lin SW, Chuang LP, Wang CW, Li LF, et al. Diffuse alveolar damage associated mortality in selected acute respiratory distress syndrome patients with open lung biopsy. *Crit Care* 2015;19:228.
- 24 Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2006.
- 25 Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, Lamy M, Legall JR, Morris A, Spragg R. The American-European Consensus Conference on ARDS: definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am J Respir Crit Care Med* 1994;149:818–824.
- 26 Fuller BM, Mohr NM, Miller CN, Deitchman AR, Levine BJ, Castagno N, Hassebroek EC, Dhedhi A, Scott-Wittenborn N, Grace E, et al. Mechanical ventilation and ARDS in the ED: a multicenter, observational, prospective, cross-sectional study. *Chest* 2015;148:365–374.
- 27 Azzam HC, Khalsa SS, Urbani R, Shah CV, Christie JD, Lanken PN, Fuchs BD. Validation study of an automated electronic acute lung injury screening tool. *J Am Med Inform Assoc* 2009;16:503–508.
- 28 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.
- 29 Zhao H, Heard SO, Mullen MT, Crawford S, Goldberg RJ, Frendl G, Lilly CM. An evaluation of the diagnostic accuracy of the 1991 American College of Chest Physicians/Society of Critical Care Medicine and the 2001 Society of Critical Care Medicine/European Society of Intensive Care Medicine/American College of Chest Physicians/American Thoracic Society/Surgical Infection Society sepsis definition. *Crit Care Med* 2012;40:1700–1706.