CHEST

Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis

Michael W. Sjoding, MD; Timothy P. Hofer, MD; Ivan Co, MD; Anthony Courey, MD; Colin R. Cooke, MD; and Theodore J. Iwashyna, MD, PhD

BACKGROUND: Failure to reliably diagnose ARDS may be a major driver of negative clinical trials and underrecognition and treatment in clinical practice. We sought to examine the interobserver reliability of the Berlin ARDS definition and examine strategies for improving the reliability of ARDS diagnosis.

METHODS: Two hundred five patients with hypoxic respiratory failure from four ICUs were reviewed independently by three clinicians, who evaluated whether patients had ARDS, the diagnostic confidence of the reviewers, whether patients met individual ARDS criteria, and the time when criteria were met.

RESULTS: Interobserver reliability of an ARDS diagnosis was “moderate” (kappa = 0.50; 95% CI, 0.40-0.59). Sixty-seven percent of diagnostic disagreements between clinicians reviewing the same patient was explained by differences in how chest imaging studies were interpreted, with other ARDS criteria contributing less (identification of ARDS risk factor, 15%; cardiac edema/volume overload exclusion, 7%). Combining the independent reviews of three clinicians can increase reliability to “substantial” (kappa = 0.75; 95% CI, 0.68-0.80). When a clinician diagnosed ARDS with “high confidence,” all other clinicians agreed with the diagnosis in 72% of reviews. There was close agreement between clinicians about the time when a patient met all ARDS criteria if ARDS developed within the first 48 hours of hospitalization (median difference, 5 hours).

CONCLUSIONS: The reliability of the Berlin ARDS definition is moderate, driven primarily by differences in chest imaging interpretation. Combining independent reviews by multiple clinicians or improving methods to identify bilateral infiltrates on chest imaging are important strategies for improving the reliability of ARDS diagnosis.

CHEST 2018; 153(2):361-367

KEY WORDS: acute lung injury; ARDS; clinical trials; diagnosis

ABBREVIATIONS: ICC = intraclass correlation coefficient

AFFILIATIONS: From the Department of Internal Medicine (Drs Sjoding, Hofer, Co, Courey, Cooke, and Iwashyna), the Institute for Healthcare Policy and Innovation (Drs Sjoding, Hofer, and Cooke), and the Department of Emergency Medicine (Dr Co), University of Michigan; the VA Center for Clinical Management Research (Drs Hofer and Iwashyna); and the Institute for Social Research (Dr Iwashyna), Ann Arbor, MI.

A portion of this work was presented in abstract form at the American Thoracic Society International Conference, May 19-24, 2017, Washington, DC.

FUNDING/SUPPORT: This work was supported by a grant to M. W. S. from the National Heart, Lung, and Blood Institute [K01HL136687], a grant to T. J. I. from the Department of Veterans Affairs Health Services Research and Development Services [IIR 13-079], and a grant to C. R. C. from the Agency for Healthcare Research and Quality [K08HS020672].

CORRESPONDENCE TO: Michael Sjoding, MD, University of Michigan, G027W Bldg 16 NCRC, 2800 Plymouth Rd, Ann Arbor, MI 48109; e-mail: msjoding@umich.edu

DOI: https://doi.org/10.1016/j.chest.2017.11.037

Reliable clinical diagnostic criteria are essential for any medical condition. Such criteria provide a framework for practicing clinicians so that they can consistently identify patients who have a similar response to medical treatment.[1] Reliable clinical diagnostic criteria are also necessary to advance medical research, helping researchers identify patients for enrollment into translational studies and clinical trials. Clinicians' failure to reliably identify ARDS may be a driver of negative ARDS clinical trials and slow progress in understanding ARDS pathobiology.[2-5] This failure may also contribute to the underrecognition and undertreatment of patients with ARDS in clinical practice.[6,7]

The 2012 revision to the ARDS definition sought to improve the validity and reliability of the previous American-European Consensus Conference definition.[8] However, the Berlin definition's success in improving the reliability of ARDS diagnosis in clinical practice is unknown. There has not been a rigorous evaluation of the interobserver reliability of the new Berlin ARDS definition or any of the specific nonradiographic ARDS clinical criteria.[9,10] Moreover, although early institution of lung-protective ventilation is the major tenant of ARDS treatment,[11-13] it is also unknown how closely clinicians agree on the time point when a patient meets all ARDS criteria.

In this study, we examined the interobserver reliability of each aspect of the Berlin ARDS definition. We hypothesized that an ARDS diagnosis and individual ARDS criteria would have low reliability when applied to patients with hypoxic respiratory failure. We specifically examined patients with a $\text{Pa}\text{O}_2/\text{F}\text{IO}_2$ ratio ≤ 300 while they were receiving invasive mechanical ventilation, as this is the patient population in whom early identification of ARDS is most important for implementing current evidence-based treatments. We sought to answer the following questions: How reliable is the Berlin definition of ARDS in this population and what are the major factors that explain differences in diagnosis? As patients evolve over time, can physicians agree on the time when all criteria are met? Which of the potential targets for improvement would yield the highest overall increase in diagnostic reliability?

Methods

We performed a retrospective cohort study of 205 adult patients (aged $\geq$ 18 years) who received invasive mechanical ventilation in one of four ICUs (medical, surgical, cardiac, and trauma) at a single tertiary care hospital during two periods in 2016. Patients were identified consecutively from January through March and from October through November 2016. Patients were excluded if they did not have a documented $\text{Pa}\text{O}_2/\text{F}\text{IO}_2$ ratio $\leq$ 300 while receiving at least 12 hours of invasive mechanical ventilation or if they were transferred from an outside hospital.

ARDS Reviews

Eight critical care-trained clinicians (four faculty and four senior fellows) reviewed patients to determine whether ARDS developed during the first 6 days of a patient's hospitalization. Patients were assigned among clinicians so that each patient was independently reviewed by three clinicians. The number of patients reviewed by clinicians ranged from 25 to 139.

To increase the uniformity of reviews, clinicians were provided a detailed summary sheet of clinical data as they reviewed each patient's electronic records and chest images. Summary sheets included a graphic display of all $\text{Pa}\text{O}_2/\text{F}\text{IO}_2$ values and the periods when patients received $\geq$ 5 mm H_2O positive end-expiratory pressure during invasive or noninvasive ventilation (e-Appendix 1).

An electronic ARDS review questionnaire was developed for the study in REDCap (e-Appendix 1). The questionnaire asked whether patients met each Berlin ARDS criterion individually and prompted the clinician to personally review each chest radiograph individually. Explicit instruction on whether or not to review the radiologist's report while reviewing chest imaging was not provided. The questionnaire then asked whether ARDS developed within the 24 hours after onset of invasive mechanical ventilation or at any point during the first 6 days of hospitalization. If the clinician believed that the patient had developed ARDS, they were then prompted to provide the time when all ARDS criteria were first met. Questions about individual ARDS criteria or ARDS diagnosis had yes or no answers and were followed by questions assessing confidence in the answer ("equivocal, slightly confident, moderately confident, highly confident").

The ARDS review tool was developed iteratively to ensure clarity of questions and minimize ambiguity in responses.[14] The tool and patient summary sheets were used by all clinicians on a training set of four patients not included in the main study. Clinicians were also provided the chest radiographs associated with the published Berlin definition for additional prestudy training.[15]

Statistical Analysis

To calculate interobserver reliability of ARDS diagnosis, the kappa for multiple nonunique raters[16] was used because of its common use in studies evaluating ARDS diagnostic reliability. To qualify agreement, kappa values of 0.8 to 1 were defined as almost perfect agreement, 0.61 to 0.8 as substantial agreement, 0.41 to 0.6 as moderate agreement, and 0.21 to 0.4 as fair agreement, and $<$ 0.2 as poor agreement.[17] CIs of kappa scores were calculated by taking 95% interval estimates after bootstrap resampling patients with 10,000 replications. We also calculated raw agreement between clinicians, agreement among ARDS cases (positive agreement), and agreement among non-ARDS cases (negative agreement). For patients considered to have acquired ARDS by at least two of three reviewers, the difference in the time when ARDS criteria were met as reported by each clinician was examined.

To better understand why clinicians disagreed about the diagnosis of ARDS, we used linear mixed models to examine how differences in ARDS diagnosis were related to differences in a clinician's assessment of individual ARDS criteria. An empty model of ARDS reviews nested within patients was fit, treating the patient as a random effect, and the intraclass correlation coefficient (ICC) was calculated. The ICC represents the correlation in ARDS diagnosis among reviews of the same patient or the proportion of variance in ARDS diagnosis explained by the patient. The rating of each individual ARDS criterion was then added as a model covariate, the model was refit, and the residual ICC was calculated. The percent change in ICC between both models represents the proportion of variability in ARDS diagnosis explained by the individual ARDS criteria.[18]

To estimate the improvement in the reliability of ARDS diagnosis when independent reviews performed by three clinicians are combined, we calculated the ICC and used the Spearman-Brown prophecy formula to calculate the estimated reliability of ARDS diagnosis when three independent reviews are averaged.[19]

Because individual ARDS criteria have differing prevalence rates in the cohort, and the acute-onset criterion had extremely high prevalence, we calculated multiple measures of agreement to evaluate and compare the reliability of each individual ARDS criteria. In this setting, the use of Cohen’s kappa to calculate interobserver reliability is controversial, and calculation of additional measures of agreement are recommended.[20-22] Further details are provided in e-Appendix 1.

To estimate how improvements in the reliability of an individual ARDS criterion could impact the reliability of ARDS diagnosis, we performed statistical simulations. We simulated scenarios in which there was increasing agreement in an individual ARDS criterion and evaluated the effect on the reliability of ARDS diagnosis. For these simulations, ARDS diagnosis was based on meeting all ARDS criteria. Details of the simulation are provided in e-Appendix 1.

Statistical analysis was performed using Stata 14 (StataCorp LLC). The Institutional Review Board of the University of Michigan approved the study (HUM00104714).

Results

Among 205 patients with a $\text{Pao}_2/\text{Fio}_2$ ratio ≤ 300 while receiving invasive mechanical ventilation, 61 patients were thought to have acquired ARDS by at least two of three clinicians. Table 1 describes characteristics of the cohort stratified by whether a majority of clinicians believed that they had acquired ARDS. Patients with ARDS had a lower minimum $\text{Pao}_2/\text{Fio}_2$ ratio and longer durations of mechanical ventilation.

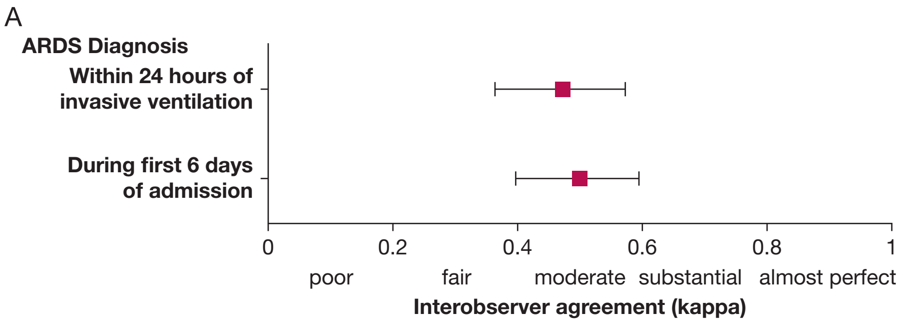
TABLE 1] Characteristics of Patients With and Those Without ARDS in the Cohort[a]

Characteristic	No ARDS (n = 144)	ARDS (n = 61)
Age, mean (SD)	60 (15)	54 (19)
Female sex	37	46
ICU type		
Medical	47	77
Surgical	26	13
Cardiac	14	5
Trauma/burn	13	5
Minimum $\text{Pao}_2/\text{Fio}_2$ ratio		
200-300	32	10
100-200	49	46
< 100	19	44
Duration of mechanical ventilation, median h (IQR)	48 (25-105)	108 (46-223)
Hospital length of stay, median d (IQR)	10 (5-18)	13 (6-23)
In-hospital mortality	22	39

Results are percentages unless otherwise stated. IQR = interquartile range.
[a]ARDS status determined based the simple average of three independent reviews.

There was “moderate” agreement (interobserver reliability) among clinicians in the diagnosis of ARDS (Fig 1). Diagnosis of ARDS within 24 hours after the onset of mechanical ventilation had a kappa of 0.47 (95% CI, 0.36-0.57) for agreement, and the diagnosis of ARDS at any point during the first 6 days of hospitalization had a kappa of 0.50 (95% CI, 0.40-0.59). Clinicians had higher agreement rates about patients who did not to acquire ARDS (84%) compared with patients who did acquire ARDS (66%). Sixty-seven percent of the disagreement in the diagnosis of ARDS was explained by differences in how clinicians interpreted chest images. Risk factor identification and cardiac edema exclusion explained 15% and 7% of the disagreement, respectively, whereas the acute-onset criterion explained 3% (e-Table 1). Among individual ARDS criteria, the criterion with the lowest agreement depended on the measure of agreement used (e-Tables 2, 3).

The median difference in time when two clinicians thought a patient met all ARDS criteria was 6 hours (interquartile range, 2-22 hours). Among patients who met ARDS criteria within the first 48 hours, the median difference was 5 hours, whereas the difference was 13 hours for patients who met criteria after 48 hours (e-Fig 1). In 262 of 615 reviews, a clinician believed that a patient met all individual ARDS criteria at some point (ie, there was at least one consistent chest radiograph and other criteria were met), and in 74% of these reviews, the clinician believed that all ARDS criteria were present simultaneously and that the overall presentation was consistent with ARDS.



B

ARDS Diagnosis	Prevalence	Raw agreement	Positive Agreement	Negative Agreement
Within 24 hours of invasive ventilation	0.26	0.80	0.61	0.86
During first 6 days of admission	0.32	0.78	0.67	0.84

Figure 1 – *Interobserver reliability between clinicians applying the Berlin ARDS definition to a cohort of 205 patients with acute hypoxic respiratory failure. A, Interobserver reliability for ARDS diagnosis within 24 hours after the onset of invasive mechanical ventilation and at any point during the hospitalization. B, Additional measures of agreement. All patients were reviewed in triplicate, and reliability was calculated using Cohen's kappa for multiple nonunique reviewers. Prevalence is the proportion of reviews in which ARDS was present. Raw agreement is the overall rate of agreement between clinicians. Positive agreement is the rate of agreement among patients believed to have acquired ARDS. Negative agreement is the rate of agreement among patients who were thought not to have acquired ARDS.*

Combining reviews made independently by clinicians and averaging them substantially improved the reliability of ARDS diagnosis (Fig 2). When the diagnosis of ARDS during the first 6 days of hospitalization was made by a combination of three independent reviews instead of a single review, reliability improved from a kappa of 0.50 (95% CI, 0.42-0.58) to a kappa of 0.75 (95% CI, 0.68-0.80).

A clinician's confidence that ARDS had developed was generally consistent with assessments of other clinicians reviewing the same patient (Fig 3). When a clinician had "high confidence" that ARDS had developed, both other clinicians agreed in 72% of reviews. Similarly, when a clinician had "high confidence" that ARDS did not develop, both other clinicians agreed that ARDS did not develop in 85% of reviews.

Simulations were performed to understand the potential effect of improving the reliability of individual ARDS criteria on the overall diagnosis. Improving the reliability of chest imaging interpretation resulted in a much larger improvement in the reliability of ARDS diagnosis, increasing kappa by up to 0.29, compared with other ARDS criteria (Fig 4). For example, improving the reliability of cardiac edema exclusion resulted in an improvement in the reliability of ARDS diagnosis by a kappa increase of up to 0.07. A 50% improvement in the reliability of chest radiograph interpretation, the amount expected if three clinicians independently reviewed chest radiographs, improved diagnostic reliability by a kappa of 0.15.

Discussion

Clinicians had only moderate interobserver agreement when diagnosing ARDS in patients with hypoxic respiratory failure under the Berlin criteria, and the major driver of this variability was differences in how

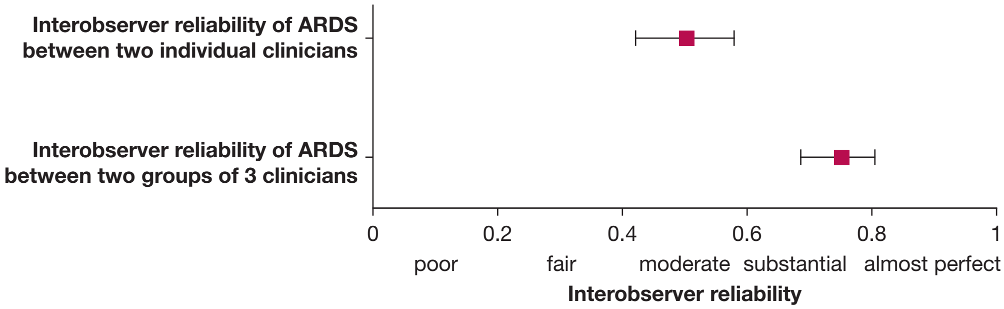


Figure 2 – *Interobserver agreement between two individual clinicians applying the Berlin ARDS definition and the interobserver agreement between two groups of three clinicians. In this approach, individuals perform ARDS reviews independently, and the group assessment is the combined average of three clinicians' individual assessments. Interobserver agreement is calculated using the intraclass correlation coefficient.*

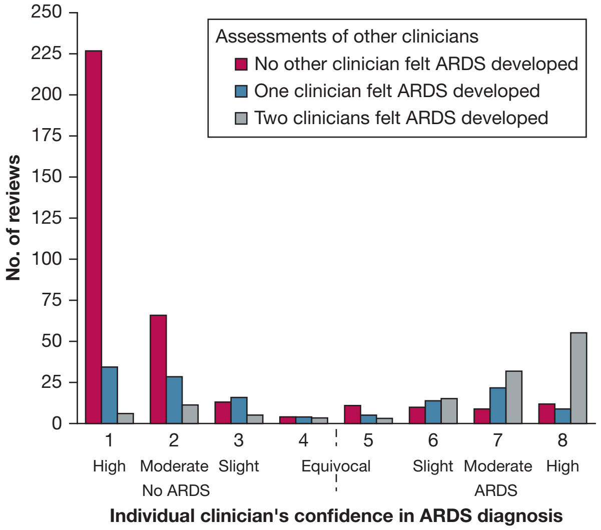


Figure 3 – *Relationship between an individual clinician's confidence in the diagnosis of ARDS and the assessment of other clinicians.*

chest images were interpreted. Strategies such as combining multiple independent reviews made by clinicians or using a clinician's confidence in their review can increase the uniformity of the diagnosis of ARDS. When a simple majority of clinicians diagnosed a patient with ARDS, they agreed closely on the time when all ARDS criteria were present if onset was during the first 48 hours of hospitalization.

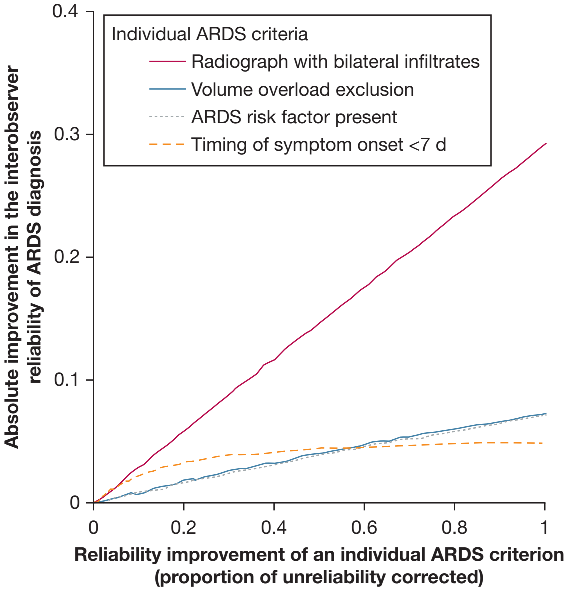


Figure 4 – *Potential for improvement in the reliability of ARDS diagnosis after improvements in individual ARDS criteria. Improvement in the reliability of individual ARDS criteria on the effect on ARDS diagnosis was simulated with assumption details described in e-Appendix 1. Absolute improvement in the reliability of ARDS diagnosis is calculated as the difference in the reliability of ARDS diagnosis before and after the reliability of the individual ARDS criteria was improved.*

The current study builds on previous work examining interobserver agreement of the ARDS radiographic criteria of bilateral infiltrates. In 1999, Rubenfeld et al[9] presented chest radiographs to experts involved in ARDS clinical trials and found that they had only moderate agreement when asked which images were consistent with the American European Consensus Conference 1994 ARDS definition, with a kappa of 0.55. Meade et al[10] found similar reliability in chest radiograph interpretation in a study performed in 2000, but they also found that reliability could improve after consensus training. The current study shows how low reliability in the current ARDS Berlin definition is primarily due to differences in chest radiograph interpretation, whereas other ARDS criteria make smaller contributions.

These results highlight a need for better approaches to identifying patients with bilateral airspace disease. Whether additional training improves reliability of chest radiograph interpretation is uncertain. Although the Meade et al[10] study showed that some reliability improvement is possible, another recent study evaluating the effect of additional training on chest radiograph interpretation among intensivists failed to show significant improvement.[23] Alternative approaches might include increasing the use of CT,[24] lung ultrasonography,[25,26] or automated processing of digital images,[27] or greater engagement with radiologists as independent reviewers.

Decisions about ARDS diagnosis should be made with specific treatments in mind, and the need for diagnostic certainty should be directly related to the potential harms of a particular treatment.[28,29] The diagnostic certainty required to administer low tidal volume ventilation, a treatment with minimal harm, should be much lower than that for prone positioning, a treatment with potential harms.[30,31] With the 2017 ARDS mechanical ventilation guidelines recommending prone positioning for severe ARDS, the need for precise ARDS diagnosis exists.[13] The current study suggests that clinicians should seek out colleagues to evaluate patients independently when higher certainty is required. In scenarios in which other clinicians are unavailable, diagnostic confidence is also a meaningful measure. In the current study, when a clinician diagnosed ARDS with "high confidence," other clinicians agreed with the diagnosis in most cases.

When independent reviews by three clinicians were combined, ARDS diagnostic reliability improved from a kappa of 0.50 to a kappa of 0.75. Such an improvement would have a major impact on ARDS clinical trials. Previous work suggests that improving the reliability of ARDS diagnosis from a kappa of 0.60 to a kappa of 0.80 could lower the sample size necessary to detect a clinically important effect by as much as 30%.[4] Although independent triplicate review of patients might be technically difficult during prospective trial recruitment, one compromise is requiring that chest images be reviewed in triplicate, which would still substantially improve ARDS diagnosis reliability. Considering a clinician's confidence in the ARDS diagnosis has also been explored in ARDS clinical research. In work by Shah et al,[32] known ARDS risk factors were more strongly associated with the development of ARDS when patients categorized with an "equivocal" ARDS diagnosis were excluded from analysis.

The current study has some limitations. Although the cohort of patients in this study was selected from four ICUs, including medical, surgical, cardiac, and trauma, reviewing patients from other populations or centers may produce different results. The study was also limited to patients with hypoxic respiratory failure. As measures of interrater reliability are dependent on the populations in which they are examined, results in populations with different patient mixes may vary. Reviews were performed by a group of eight investigators, including four faculty and four senior fellows, a number that is similar to many investigations of ARDS reliability,[10,32,33] but reliability may differ among other clinicians. Finally, reviews were retrospective, and it is unknown whether the reliability of ARDS diagnosis is similar when patients are evaluated prospectively, as performed in clinical practice. In this situation, clinicians cannot evaluate a patient's entire course of illness when assessing ARDS, but they may also have access to additional information not recorded in a medical record. However, evaluation of chest images for bilateral infiltrates consistent with ARDS, the main driver of low reliability, may be expected to be similar.

Conclusions

We found the interobserver reliability of ARDS diagnosis among clinicians to be only moderate, driven primarily by the low reliability of the interpretation of chest images. Combining independent reviews of patients increased reliability substantially and should be performed whenever possible when diagnosing ARDS. Efforts to improve detection of bilateral lung infiltrates on chest images should be prioritized in future ARDS diagnostic research.

Acknowledgments

Author contributions: M. W. S. had full access to all the data in the study and takes full responsibility for the integrity of the data and accuracy of the data analysis. M. W. S. and T. J. I. contributed to the study design, analysis, and interpretation of data; writing; revising the manuscript; and approval of the final manuscript. T. P. H., I. C., A. C., and C. R. C. contributed to the analysis and interpretation of data, revising the manuscript for important intellectual content, and approval of the final manuscript.

Financial/nonfinancial disclosures: None declared.

Role of sponsors: The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

Other contributions: This manuscript does not necessarily represent the view of the US Government or the Department of Veterans Affairs.

Additional information: The e-Appendix, e-Tables, and e-Figure can be found in the Supplemental Materials section of the online article.

References

1. Coggon D, Martyn C, Palmer KT, Evanoff B. Assessing case definitions in the absence of a diagnostic gold standard. *Int J Epidemiol*. 2005;34(4):949-952.
2. Rubenfeld GD. Confronting the frustrations of negative clinical trials in acute respiratory distress syndrome. *Ann Thorac Surg*. 2015;12(suppl 1): S58-S63.
3. Frohlich S, Murphy N, Boylan JF. ARDS: progress unlikely with non-biological definition. *Br J Anaesth*. 2013;111(5): 696-699.
4. Sjoding MW, Cooke CR, Iwashyna TJ, Hofer TP. Acute respiratory distress syndrome measurement error. Potential effect on clinical study results. *Ann Thorac Surg*. 2016;13(7):1123-1128.
5. Pham T, Rubenfeld GD. Fifty years of research in ARDS: the epidemiology of acute respiratory distress syndrome. A 50th birthday review. *Am J Respir Crit Care Med*. 2017;195(7):860-870.
6. Bellani G, Laffey JG, Pham T, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA*. 2016;315(8):788-800.
7. Weiss CH, Baker DW, Weiner S, et al. Low tidal volume ventilation use in acute respiratory distress syndrome. *Crit Care Med*. 2016;44(8):1515-1522.
8. Ranieri VM, Rubenfeld GD, Thompson BT, et al. Acute respiratory distress syndrome: the Berlin Definition. *JAMA*. 2012;307(23):2526-2533.
9. Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver variability in applying a radiographic definition for ARDS. *Chest*. 1999;116(5): 1347-1353.
10. Meade MO, Cook RJ, Guyatt GH, et al. Interobserver variation in interpreting chest radiographs for the diagnosis of acute respiratory distress syndrome. *Am J Respir Crit Care Med*. 2000;161(1):85-90.
11. Amato MB, Barbas CS, Medeiros DM, et al. Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome. *N Engl J Med*. 1998;338(6):347-354.
12. Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, Wheeler A. 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(18):1301-1308.

13. Fan E, Del Sorbo L, Goligher EC, et al. An Official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine Clinical Practice Guideline: mechanical ventilation in adult patients with acute respiratory distress syndrome. *Am J Respir Crit Care Med*. 2017;195(9):1253-1263.

14. Sudman S, Bradburn NM, Schwarz N. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass Inc.; 1996.

15. Ferguson ND, Fan E, Camporota L, et al. The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material. *Intensive Care Med*. 2012;38(10):1573-1582.

16. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2003.

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.

18. Snijders TAB, Cosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage; 2012.

19. Spearman CE. Correlation calculated from faulty data. *Br J Psychol*. 1910;3:271-295.

20. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-429.

21. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543-549.

22. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. *J Clin Epidemiol*. 2005;58(7):655-661.

23. Peng JM, Qian CY, Yu XY, et al. Does training improve diagnostic accuracy and inter-rater agreement in applying the Berlin radiographic definition of acute respiratory distress syndrome? A multicenter prospective study. *Crit Care*. 2017;21(1):12.

24. Pesenti A, Tagliabue P, Patroniti N, Fumagalli R. Computerised tomography scan imaging in acute respiratory distress syndrome. *Intensive Care Med*. 2001;27(4):631-639.

25. Bass CM, Sajed DR, Adedipe AA, West TE. Pulmonary ultrasound and pulse oximetry versus chest radiography and arterial blood gas analysis for the diagnosis of acute respiratory distress syndrome: a pilot study. *Critical Care*. 2015;19:282.

26. Sekiguchi H, Schenck LA, Horie R, et al. Critical care ultrasonography differentiates ARDS, pulmonary edema, and other causes in the early course of acute hypoxemic respiratory failure. *Chest*. 2015;148(4):912-918.

27. Zaglam N, Jouvet P, Flechelles O, Emeriaud G, Cheriet F. Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs. *Comput Biol Med*. 2014;52:41-48.

28. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302(20):1109-1117.

29. Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. *N Engl J Med*. 1989;320(22):1489-1491.

30. Guerin C, Gaillard S, Lemasson S, et al. Effects of systematic prone positioning in hypoxemic acute respiratory failure: a randomized controlled trial. *JAMA*. 2004;292(19):2379-2387.

31. Taccone P, Pesenti A, Latini R, et al. Prone positioning in patients with moderate and severe acute respiratory distress syndrome: a randomized controlled trial. *JAMA*. 2009;302(18):1977-1984.

32. Shah CV, Lanken PN, Localio AR, et al. An alternative method of acute lung injury classification for use in observational studies. *Chest*. 2010;138(5):1054-1061.

33. 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(3):417-424.