Accepted Manuscript

Inter-observer reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis

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

PII: S0012-3692(17)33219-1
DOI: 10.1016/j.chest.2017.11.037
Reference: CHEST 1470

To appear in: CHEST

Received Date: 19 July 2017
Revised Date: 27 September 2017
Accepted Date: 6 November 2017

Please cite this article as: Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ, Inter-observer reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis, CHEST (2018), doi: 10.1016/j.chest.2017.11.037.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Inter-observer reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis.

Michael W. Sjoding, MD, MSc\textsuperscript{1,2}, Timothy P. Hofer MD, MS\textsuperscript{1,2,3}, Ivan Co MD\textsuperscript{1,4}, Anthony Courey MD\textsuperscript{1}, Colin R. Cooke MD, MSc, MS\textsuperscript{1,2}, Theodore J. Iwashyna MD PhD\textsuperscript{1,3,5}

Affiliations:
\textsuperscript{1}Department of Internal Medicine, University of Michigan, Ann Arbor, MI
\textsuperscript{2}Institute for Healthcare Policy & Innovation, University of Michigan, Ann Arbor, MI
\textsuperscript{3}VA Center for Clinical Management Research, Ann Arbor, MI
\textsuperscript{4}Department of Emergency Medicine, University of Michigan, Ann Arbor, MI
\textsuperscript{5}Institute for Social Research, Ann Arbor, MI

Corresponding Author: Michael Sjoding, MD, MSc, University of Michigan, 3916 Taubman Center, 1500 E. Medical Center Dr., SPC 5360, Ann Arbor, MI 48109-5360. Email: msjoding@umich.edu

Author contributions: Dr. Sjoding 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.

Funding/Support: This work was supported by grants to Dr. Sjoding from the NHLBI K01HL136687, Dr. Iwashyna from the Department of Veterans Affairs Health Services Research & Development Services - IIR 13-079, and Dr. Cooke from the AHRQ K08HS020672.

All authors report no conflicts of interest to disclose

A portion of this work was presented in abstract form at the American Thoracic Society International Conference, May 2017, Washington, DC. USA
Abbreviation list

ARDS – Acute Respiratory Distress Syndrome

CI – Confidence Interval

ICC – intra-class correlation coefficient
Abstract

Background: Failure to reliably diagnose the Acute Respiratory Distress Syndrome (ARDS) may be a major driver of negative clinical trials and under-recognition and treatment in clinical practice. We sought to examine the inter-observer reliability of the Berlin ARDS definition and examine strategies for improving the reliability of ARDS diagnosis.

Methods: 205 patients with hypoxic respiratory failure from four intensive care units were reviewed independently by three clinicians, who evaluated whether patients developed ARDS, their diagnostic confidence, whether patients met individual ARDS criteria, and the time when criteria were met.

Results: Inter-observer reliability of ARDS diagnosis was “moderate,” kappa = 0.50 (95% CI, 0.40 – 0.59). 67% of diagnostic disagreement between clinicians reviewing the same patient was explained by differences in how chest imaging studies were interpreted, with other ARDS criteria contributing less (15% identification of ARDS risk factor, 7% cardiac edema/volume overload exclusion). Combining independent reviews of three clinicians can increase reliability to “substantial”, 0.75 (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 in the time when a patient met all ARDS criteria when ARDS developed within the first 48 hours of hospitalization (median difference 5 hours).
**Conclusion:** 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 reliability of ARDS diagnosis.
Introduction

Reliable clinical diagnostic criteria are essential for any medical condition. Such criteria provide a framework for practicing clinicians, so they can consistently identify patients who have a similar response to medical treatment. 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 the Acute Respiratory Distress Syndrome (ARDS) may be a driver of negative ARDS clinical trials and slow progress in understanding ARDS pathobiology. This failure may also contribute to the under-recognition and under-treatment of patients with ARDS in clinical practice.

The 2012 revision to the ARDS definition sought to improve the validity and reliability of the previous American-European Consensus Conference definition. However, the Berlin definitions' success in improving the reliability of ARDS diagnosis in clinical practice is unknown. There has not been a rigorous evaluation of the inter-observer reliability of the new Berlin ARDS definition or any of the specific non-radiographic ARDS clinical criteria. Moreover, although early institution of lung protective ventilation is the major tenant of ARDS treatment, 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 inter-observer reliability of each aspect of the Berlin ARDS definition. We hypothesized that ARDS diagnosis and individual ARDS criteria would have low reliability when applied to patients with hypoxic respiratory failure. We specifically examined patients with a PaO$_2$/FiO$_2$ ratio $\leq$ 300 while receiving invasive mechanical ventilation, the patient population where 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?

**Materials and Methods**

We performed a retrospective cohort study of 205 adult patients (Age 18 ≥ years) who received invasive mechanical ventilation in one of four intensive care units (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 October through November 2016. Patients were excluded if they did not have a documented PaO$_2$/FiO$_2$ ≤ 300 while receiving at least 12 hours of invasive mechanical ventilation or were transferred from an outside hospital.

**ARDS reviews**

Eight critical care trained clinicians (4 faculty and 4 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 3 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 graphical display of all PaO\textsubscript{2}/FiO\textsubscript{2} values and the time periods when patients received ≥ 5 mm H\textsubscript{2}O positive end expiratory pressure during invasive or non-invasive ventilation (example available in Appendix).

An electronic ARDS review questionnaire was developed for the study in Redcap (included in Appendix). 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 the patient developed ARDS within the 24 hours after onset of invasive mechanical ventilation, or at any point during the first 6 hospital days. If the clinician felt the patient 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.\textsuperscript{14} The tool and patient summary sheets were used by all clinicians on a training set of 4 patients not included in the main study. Clinicians were also provided the chest x-rays set associated with the published Berlin definition for additional pre-study training.\textsuperscript{15}

*Statistical Analysis*
To calculate inter-observer reliability of ARDS diagnosis, kappa for multiple, non-unique raters\textsuperscript{16} was used because of its common use in studies evaluating ARDS diagnostic reliability. To qualify agreement, kappa 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, and <0.2 as poor agreement.\textsuperscript{17} Confidence intervals 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 developed ARDS by at least 2 of 3 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 patient was fit, treating patient as a random effect, and calculated the intra-class correlation coefficient (ICC). The ICC represents the correlation in ARDS diagnosis among reviews on the same patient or the proportion of variance in ARDS diagnosis explained by the patient. The rating of each individual ARDS criteria 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.\textsuperscript{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.\textsuperscript{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 reliability of each individual ARDS criteria. In this setting, use of Cohen’s kappa to calculate inter-observer reliability is controversial and calculation of additional measures of agreement are recommended.\textsuperscript{20-22} Further details are provided in the Appendix.

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 where 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 upon meeting all ARDS criteria. Details of the simulation are provided in the Appendix.

Statistical analysis was performed using Stata 14 (StataCorp, College Station, TX), The institutional review board of the University of Michigan approved the study (HUM00104714).

\textbf{Results}

Among 205 patients with a PaO\textsubscript{2}/FiO\textsubscript{2} ≤ 300 while receiving invasive mechanical ventilation, 61 patients were felt to have developed ARDS by at least 2 of 3 clinicians.
Table 1 describes characteristics of the cohort stratified by whether a majority of clinicians felt they developed ARDS. Patients with ARDS had lower minimum PaO$_2$/FiO$_2$ and longer durations of mechanical ventilation.

There was “moderate” agreement (inter-observer reliability) among clinicians in the diagnosis of ARDS (Figure 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 hospital days had a kappa of 0.50 (95% CI, 0.40 – 0.59). Clinicians had higher agreement rates about patients who did not to develop ARDS (84%) compared to patients who did develop ARDS (66%). 67% of the disagreement in the diagnosis of ARDS was explained by differences in how clinicians interpreted chest imaging. Risk factor identification and cardiac edema exclusion explained 15% and 7% of the disagreement, while 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-table 2, e-table 3).

The median difference in time when two clinicians felt 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, while the difference was 13 hours for patients who met criteria after 48 hours (e-Figure 1). In 262 of 615 reviews, a clinician felt a patient met all individual ARDS criteria at some point (i.e. had at least one consistent chest x-ray and met other criteria), and in 74% of these reviews, the clinician felt all ARDS criteria were present simultaneously and the overall presentation was consistent with ARDS.
Combining reviews made independently by clinicians and averaging them substantially improved the reliability of ARDS diagnosis (Figure 2). When the diagnosis of ARDS during first 6 hospital days was made by a combination of three independent reviews instead of a single review, reliability improved from 0.50 (0.42 – 0.58) to 0.75 (0.68 – 0.80).

A clinician’s confidence in the development of ARDS was generally consistent with assessments of other clinicians reviewing the same patient (Figure 3). When a clinician had “high confidence” that ARDS 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 reliability of individual ARDS criteria on the overall diagnosis. Improving the reliability of chest imaging interpretation resulted in much larger improvement in the reliability of ARDS diagnosis, up to 0.29, compared to other ARDS criteria (Figure 4). For example, the reliability of ARDS diagnosis improved up to 0.07 by improving the reliability of cardiac edema exclusion. A 50% improvement in the reliability of chest radiograph interpretation, the amount expected if 3 clinicians independently reviewed chest radiographs, improved diagnostic reliability by 0.15.

Discussion

Clinicians had only moderate inter-observer 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 chest imaging was interpreted. Strategies such
as combining multiple independent reviews made by clinicians, or utilizing a clinician’s confidence in their review, can increase the uniformity of patients diagnosed with ARDS. When a simple majority of clinicians diagnosed a patient with ARDS, they agreed closely on the time when all ARDS criteria were if onset was during the first 48 hours of hospitalization.

The current study builds upon previous work examining inter-observer agreement of the ARDS radiographic criteria of bilateral infiltrates. In 1999, Rubenfeld et al presented chest radiographs to experts involved in ARDS clinical trials and found they had only moderate agreement when asked which images were consistent with the AECC 1994 ARDS definition, with a kappa of 0.55.9 Meade et al found similar reliability in chest radiograph interpretation in a study performed in 2000, but also found that reliability could improve after consensus training.10 The current study shows how low reliability in the current ARDS Berlin definition is primarily due to differences in chest x-ray interpretation, while 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. While the Meade et al study showed 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 use of computed tomography,24 lung ultrasound,25,26 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.\textsuperscript{28,29} The diagnostic certainty required to administer low tidal volume ventilation, a treatment with minimal harm, should be much lower than for prone-positioning, a treatment with potential harms.\textsuperscript{30,31} With the 2017 ARDS mechanical ventilation guidelines recommending prone-positioning for severe ARDS, the need for precise ARDS diagnosis exists.\textsuperscript{13} The current study suggests that clinicians should seek out colleagues to evaluate patients independently when higher certainty is required. In scenarios where 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 0.50 to 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 0.60 to 0.80 could lower the sample size necessary to detect a clinically important effect by as much as 30%.\textsuperscript{4} While independent triplicate review of patients might be technically difficult during prospective trial recruitment, one compromise is requiring chest imaging be reviewed in triplicate, which would still substantially improve ARDS reliability. Considering a clinician’s confidence in ARDS diagnosis has also been explored in ARDS clinical research. In work by Shah \textit{et al}, known ARDS risk factors were more strongly associated with ARDS development when patients categorized as “equivocal” ARDS were excluded from analysis.\textsuperscript{32}
The current study has some limitations. While the cohort of patients in this study was selected from four intensive care units, 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 inter-rater 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 similar number to many investigations of ARDS reliability, 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 may also access to additional information not recorded in a medical record. However, evaluation of chest imaging for bilateral infiltrates consistent with ARDS, the main driver of low reliability, may be expected to be similar.

Conclusion

We found the inter-observer reliability of ARDS diagnosis among clinicians to be only moderate, driven primarily by the low reliability of chest imaging interpretation. 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 imaging should be prioritized in future ARDS diagnostic research.
Acknowledgements

Author contributions: Dr. Sjoding 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.

D. Sjoding and Dr. Iwashyna contributed to the study design, analysis and interpretation of data, writing and revising the manuscript and approval of the final manuscript. Drs. Hofer, Co, Courey, and Cooke contributed to the analysis and interpretation of data, revising the manuscript for important intellectual content and approval of the final manuscript.

Funding/Support: This work was supported by grants to Dr. Sjoding from the NHLBI K01HL136687, Dr. Iwashyna from the Department of Veterans Affairs Health Services Research & Development Services - IIR 13-079, and Dr. Cooke from the AHRQ K08HS020672.

Role of sponsors: The funding organizations had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the preparation, review or approval of the manuscript

Disclaimer: This manuscript does not necessarily represent the view of the U.S. Government of the Department of Veterans Affairs.
References


Table 1. Characteristics of patients with and without ARDS in the cohort. ARDS status determined based on the simple average of three independent reviews.

<table>
<thead>
<tr>
<th></th>
<th>No ARDS (N = 144)</th>
<th>ARDS (N = 61)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD)</td>
<td>60 (15)</td>
<td>54 (19)</td>
</tr>
<tr>
<td>Female</td>
<td>37</td>
<td>46</td>
</tr>
<tr>
<td>ICU type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>47</td>
<td>77</td>
</tr>
<tr>
<td>Surgical</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>Cardiac</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Trauma/Burn</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Minimum PaO2/FiO2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200-300</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>100-200</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>&lt;100</td>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>Duration of MV, median hours (IQR)</td>
<td>48 (25-105)</td>
<td>108 (46-223)</td>
</tr>
<tr>
<td>Hospital LOS, median days (IQR)</td>
<td>10 (5-18)</td>
<td>13 (6-23)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>22</td>
<td>39</td>
</tr>
</tbody>
</table>

Results are percentages unless otherwise stated. ARDS determined based on the average of three independent clinical reviews. IQR – interquartile range.
Figure Legend

Figure 1. Inter-observer reliability between clinicians applying the Berlin ARDS definition to a cohort of 205 patients with acute hypoxic respiratory failure. A) Inter-observer reliability for ARDS diagnosis within 24 hours after the onset of invasive mechanical ventilation and at any point during the hospital. B) Additional measures of agreement. All patients were reviewed in triplicate and reliability was calculated using Cohen’s kappa for multiple, non-unique 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 felt to have developed ARDS. Negative agreement is rate of agreement among patients felt not to have developed ARDS.

Figure 2. Inter-observer agreement between two individual clinicians applying the Berlin ARDS definition and the inter-observer 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. Inter-observer agreement is calculated using the intra-class correlation coefficient.

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

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 the appendix. 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.
### A

**ARDS Diagnosis**
- **Within 24 hours of invasive ventilation**
- **During first 6 days of admission**

![Graph](image)

**Inter-observer agreement (kappa)**
- poor
- fair
- moderate
- substantial
- almost perfect

### B

<table>
<thead>
<tr>
<th>ARDS Diagnosis</th>
<th>Prevalence</th>
<th>Raw agreement</th>
<th>Positive Agreement</th>
<th>Negative Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within 24 hours of invasive ventilation</td>
<td>0.26</td>
<td>0.80</td>
<td>0.61</td>
<td>0.86</td>
</tr>
<tr>
<td>During first 6 days of admission</td>
<td>0.32</td>
<td>0.78</td>
<td>0.67</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Inter-observer reliability of ARDS between two individual clinicians

Inter-observer reliability of ARDS between two groups of 3 clinicians

Inter-observer reliability

0 poor 0.2 fair 0.4 moderate 0.6 substantial 0.8 almost perfect 1 almost perfect
Assessments of other clinicians:
- No other clinician felt ARDS developed
- One clinician felt ARDS developed
- Two clinicians felt ARDS developed

Number of reviews

Individual clinician’s confidence in ARDS diagnosis:
**Supplemental methods**

**Sample size calculation**
To determine the number of reviews necessary to estimate ARDS diagnostic reliability with adequate confidence, we made the conservative assumption that reliability would be 0.6. Using the method proposed by Zou\(^1\), we determined that at least 120 patients would need to be reviewed by 3 reviewers (or 196 patients by 2 reviewers) to obtain confidence intervals no wider than 0.1 with 90% assurance probability.

**e-Table 1.** Proportion of disagreement in the diagnosis of ARDS explained by individual ARDS criteria

<table>
<thead>
<tr>
<th>ARDS criterion</th>
<th>ICC between clinicians</th>
<th>Residual ICC after criterion added to model</th>
<th>Proportion of variance explained by ARDS criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>chest imaging</td>
<td>0.500</td>
<td>0.164</td>
<td>0.672</td>
</tr>
<tr>
<td>event timing</td>
<td>0.500</td>
<td>0.487</td>
<td>0.026</td>
</tr>
<tr>
<td>edema exclusion</td>
<td>0.500</td>
<td>0.467</td>
<td>0.066</td>
</tr>
<tr>
<td>risk factor</td>
<td>0.500</td>
<td>0.427</td>
<td>0.146</td>
</tr>
</tbody>
</table>

An empty linear mixed model of ARDS reviews nested within patient was fit, treating patient as a random effect, and the intra-class correlation coefficient (ICC) was calculated. The rating of each individual ARDS criteria was then added to the linear mixed model as a 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.

**e-Table 2.** Measures of agreement for each individual ARDS criteria

<table>
<thead>
<tr>
<th>ARDS criteria</th>
<th>Prevalence</th>
<th>Raw agreement</th>
<th>Positive agreement</th>
<th>Negative agreement</th>
<th>Kappa</th>
<th>PABAK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute onset</td>
<td>0.95</td>
<td>0.91</td>
<td>0.95</td>
<td>0.05</td>
<td>0.00</td>
<td>0.82</td>
</tr>
<tr>
<td>ARDS risk factor</td>
<td>0.79</td>
<td>0.83</td>
<td>0.89</td>
<td>0.58</td>
<td>0.47</td>
<td>0.65</td>
</tr>
<tr>
<td>Bilateral opacities</td>
<td>0.41</td>
<td>0.73</td>
<td>0.67</td>
<td>0.78</td>
<td>0.45</td>
<td>0.47</td>
</tr>
<tr>
<td>Cardiac edema excluded</td>
<td>0.87</td>
<td>0.85</td>
<td>0.91</td>
<td>0.41</td>
<td>0.32</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Prevalence is the proportion of reviews in which the specific criterion was felt to be met. Raw agreement is the overall rate of agreement between clinicians when evaluating the criterion for each patient. Positive agreement is the rate of agreement among patients felt to meet the criterion. Negative agreement is rate of agreement among patients felt to not have the criterion. Kappa is Cohen’s kappa for multiple non-unique raters. PABAK is the prevalence-adjusted bias-adjusted kappa.\(^2\)
e-Table 3. Measures of agreement for identifying specific ARDS risk factors

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Frequency</th>
<th>Raw agreement</th>
<th>Kappa</th>
<th>PABAK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pneumonia</td>
<td>0.88</td>
<td>0.82</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>Non-pulmonary sepsis</td>
<td>0.57</td>
<td>0.86</td>
<td>0.56</td>
<td>0.73</td>
</tr>
<tr>
<td>Non-cardiogenic shock</td>
<td>0.82</td>
<td>0.76</td>
<td>0.40</td>
<td>0.52</td>
</tr>
<tr>
<td>Multiple transfusions</td>
<td>0.84</td>
<td>0.90</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>Major trauma</td>
<td>0.18</td>
<td>0.98</td>
<td>0.85</td>
<td>0.97</td>
</tr>
<tr>
<td>High risk surgery</td>
<td>0.31</td>
<td>0.93</td>
<td>0.60</td>
<td>0.85</td>
</tr>
<tr>
<td>Aspiration</td>
<td>0.42</td>
<td>0.88</td>
<td>0.49</td>
<td>0.75</td>
</tr>
<tr>
<td>Pancreatitis</td>
<td>0.02</td>
<td>1.00</td>
<td>0.80</td>
<td>0.99</td>
</tr>
<tr>
<td>Severe burns</td>
<td>0.07</td>
<td>0.98</td>
<td>0.66</td>
<td>0.97</td>
</tr>
<tr>
<td>Inhalation injury</td>
<td>0.04</td>
<td>0.98</td>
<td>0.37</td>
<td>0.97</td>
</tr>
<tr>
<td>Pulmonary vasculitis</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Pulmonary contusion</td>
<td>0.05</td>
<td>0.98</td>
<td>0.29</td>
<td>0.95</td>
</tr>
<tr>
<td>Drowning</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

e-Figure 1. Differences in agreement on the time of ARDS onset among clinicians for the 61 patients who developed ARDS in the cohort. Shadow represents the 95% limits of agreement. 95% intervals of agreement using a regression approach described by Bland et al. because the standard deviation of measurement differences did not appear constant over time.³
e-Table 4. Possible approaches to improve imaging evaluation of bilateral infiltrates consistent with ARDS

<table>
<thead>
<tr>
<th>Method</th>
<th>Explanation/Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require multiple clinicians to review chest x-rays</td>
<td>As shown in the manuscript, averaging independent reviews by multiple clinicians increases reliability of the assessment</td>
</tr>
<tr>
<td>Engage radiologists as additional reviewer</td>
<td>Increased engagement with radiology might be useful, particularly in situations where other clinicians are unavailable</td>
</tr>
<tr>
<td>Lung ultrasonography</td>
<td>Lung ultrasound can be used to help differentiate ARDS from other causes of acute hypoxic respiratory failure(^4,5)</td>
</tr>
<tr>
<td>Computed tomography (CT)</td>
<td>CT imaging may help identify bilateral infiltrates consistent with ARDS, the underlying cause of ARDS and its complications(^6)</td>
</tr>
<tr>
<td>Imaging processing technology to automate detection</td>
<td>Digital image processing technology to identify ARDS may be possible,(^7) although further development is needed</td>
</tr>
</tbody>
</table>

**Description of simulation assumptions and Stata code**

We performed simulations to answer the following question: how would improvement in the reliability of an individual ARDS criterion improve the reliability of the ARDS diagnosis overall? During each simulation, ratings on one of the individual ARDS criterion were varied in such a way that the inter-rater reliability of the criterion increased among reviewers. Then, whether a patient had ARDS was determined based upon meeting all ARDS criteria, and the simulated ratings of the ARDS criterion under question was used in this determination. Finally, the reliability of ARDS diagnosis was re-calculated to determine how much improvement in the reliability of the diagnosis of ARDS would be seen by improving the reliability of the individual criterion.

To simulate improvement in the reliability of an individual ARDS criterion, first, whether each patient met the criterion was determined based on the average assessment of three reviewers. Next, each reviewer’s rating on the ARDS criterion was compared against the group rating to determine each reviewer’s rate of miss-classifying patients. Finally, the reviewer’s initial ratings on the criterion were dropped and then simulated, based on these miss-classification rates. Over the course of multiple simulations, each reviewer’s miss-classification rate was reduced, resulting in increasing agreement among reviewers. As the miss-classification rate for each reviewer approached zero, the inter-rater reliability approached 1.0.
Stata code for the simulation

version 14
set seed 97302
drop _all
set more off
postutil clear

cap program drop calc_kappa
program calc_kappa, rclass
    syntax, var(varname)
    *Calculates the kappa for var between reviewers when data are in "long" form
    tempvar pos neg tag
    bysort patient_num: gen `tag' = _n==1
    bysort patient_num: egen `pos' = total(`var')
gen `neg' = 3 - `pos'
qui: kappa `pos' `neg' if `tag'==1
return scalar calc_kappa = r(kappa)
end

cap program drop calc
program calc, rclass
    use ards-reviews.dta, clear
    set more off
    syntax , num(real) var(varname) ARDSCriteria(string)
    /*
    Variables:
    num = tuning parameter adjusts amount of agreement for an ARDS criterion
        between reviewers, when num = 1, inter-rater reliability = 1
    var = ARDS criterion examined
    ARDSCriteria = the group of criteria used determine whether patient had
        ARDS, e.g. "`var'_1 == 1 & not_cardiac==1 & risk==1 & event_timing==1"
    */
    return scalar num = `num'
    *determine a patient's true status on the ARDS criterion's based on the group
    *assessment among reviewers
    bysort patient_num: egen val = mean(`var')
gen true = val > 0.5
    *Calculate each individual reviewer's rate of correctly classifying a patient
    bysort reviewer true: egen pos = total(`var')
bysort reviewer true: gen rate = pos/_N /*reviewer "sensitivity" */
replace rate = 1-rate if true==0 /* reviewer "specificity" */

*Generates new classification rate
gen rate1 = (1-rate)*`num' + rate

*Simulate random mis-classification each reviewers correct classification rate
gen `var'_1 = cond(runiform() < rate1, 1, 0) if true==1
replace `var'_1 = cond(runiform() < rate1, 0, 1) if true==0

*Calculate the reliability of the simulated ratings
calc_kappa, var(`var'_1)
return scalar var_kappa = r(calc_kappa)

*determine the patient's ARDS status based on meeting all criteria,
*now incorporated the simulated variable `var'_1
gen ards_1 = 1 if `ARDSCriteria'
replace ards_1 = 0 if ards_1==.

*Calculate the reliability of the patients newly determined ARDS status
calc_kappa, var(ards_1)
return scalar ards_kappa = r(calc_kappa)

drop ards_1 `var'_1 rate rate1 pos true val
end

*Now perform the simulation to examine how improvement in the reliability of the chest
*imaging criterion could impact the reliability of ARDS diagnosis

postfile sim num var_kappa ards_kappa using sim_cxr, replace

local var any_cxr
local ARDSCriteria = "`var'_1 == 1 & not_cardiac==1 & risk==1 & event_timing==1"

forval i = 0(0.02)1 {
    simulate num = r(num) ///
    var_kappa = r(var_kappa) ///
    ards_kappa = r(ards_kappa), ///
        reps(1000): calc, num(`i') var(`var') ARDSCriterian(`equation')
        mean num var_kappa ards_kappa
    post sim (_b[num]) (_b[var_kappa]) (_b[ards_kappa])
}
postclose sim
References