[Original Research]

≋ 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 2017; ■(■):■-■

KEY WORDS: ARDS; clinical trials; diagnosis

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 PaO_2/FIO_2 ratio $\leq$ 300 while they were receiving invasive mechanical ventilation; 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 PaO_2/FIO_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 PaO_2/FIO_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 acquired 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