CHEST Online Supplement

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

CHEST 2018; 153(2):361-367

Online supplements are not copyedited prior to posting and the author(s) take full responsibility for the accuracy of all data.

DOI: 10.1016/j.chest.2017.11.037

e-Appendix 1.

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

ARDS criterion	ICC between clinicians	Residual ICC after criterion added to model	Proportion of variance explained by ARDS criterion
chest imaging	0.500	0.164	0.672
event timing	0.500	0.487	0.026
edema exclusion	0.500	0.467	0.066
risk factor	0.500	0.427	0.146

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.

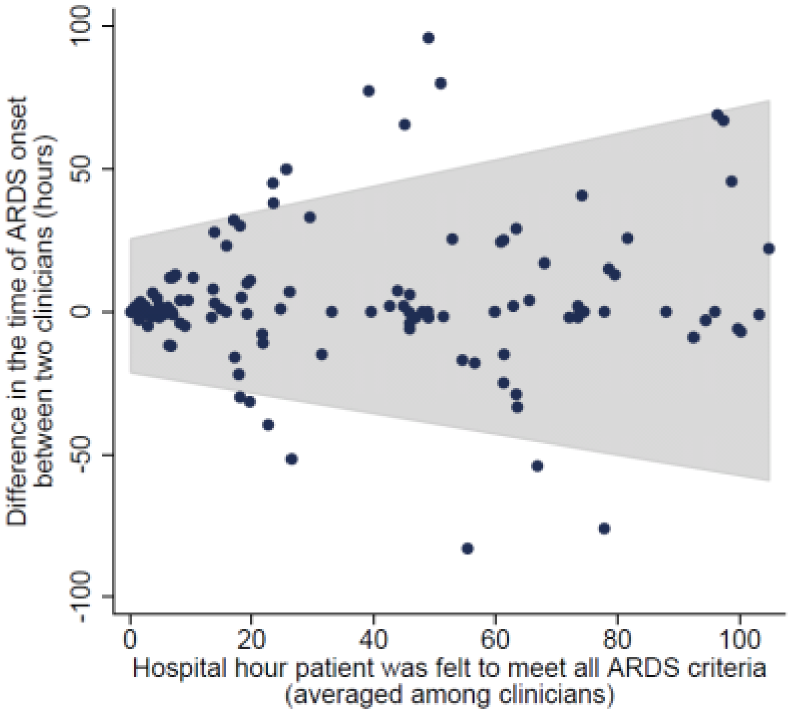
e-Table 2. Measures of agreement for each individual ARDS criteria

ARDS criteria	Prevalence	Raw agreemen t	Positive agreement	Negative agreement	Kappa	PABAK
Acute onset	0.95	0.91	0.95	0.05	0.00	0.82
ARDS risk factor	0.79	0.83	0.89	0.58	0.47	0.65
Bilateral opacities	0.41	0.73	0.67	0.78	0.45	0.47
Cardiac edema excluded	0.87	0.85	0.91	0.41	0.32	0.70

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

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



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

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

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" */

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

        *Simulate random mis-classification each reveiwers 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(.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

1. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in medicine.* 2012;31(29):3972-3981.
2. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of clinical epidemiology.* 1993;46(5):423-429.
3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical methods in medical research.* 1999;8(2):135-160.
4. 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.
5. 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 (London, England).* 2015;19:282.
6. Zompatori M, Ciccarese F, Fasano L. Overview of current lung imaging in acute respiratory distress syndrome. *European respiratory review : an official journal of the European Respiratory Society.* 2014;23(134):519-530.
7. Zaglam N, Jouvet P, Flechelles O, Emeriaud G, Cheriet F. Computer-aided diagnosis system for the Acute Respiratory Distress Syndrome from chest radiographs. *Computers in biology and medicine.* 2014;52:41-48.