Measurement of respiratory rate by multiple raters in a clinical setting is unreliable

A cross-sectional simulation study

Brabrand, Mikkel; Hallas, Peter; Folkestad, Lars; Lautrup-Larsen, Cecilie Hovitz; Brodersen, Jacob Broder

Published in: Journal of Critical Care


Publication date: 2018

Document version: Accepted manuscript

Document license: CC BY-NC-ND

Citation for published version (APA):
Measurement of respiratory rate by multiple raters in a clinical setting is unreliable: A cross-sectional simulation study

Mikkel Brabrand, Peter Hallas, Lars Folkestad, Cecilie Hovitz Lautrup-Larsen, Jacob Broder Brodersen

PII: S0883-9441(17)31331-X
Reference: YJCRC 52820

To appear in:

Please cite this article as: Mikkel Brabrand, Peter Hallas, Lars Folkestad, Cecilie Hovitz Lautrup-Larsen, Jacob Broder Brodersen, Measurement of respiratory rate by multiple raters in a clinical setting is unreliable: A cross-sectional simulation study. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Yjcrc(2017), https://doi.org/10.1016/j.jcrc.2017.12.020

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.
Measurement of respiratory rate by multiple raters in a clinical setting is unreliable: a cross-sectional simulation study

Mikkel Brabrand MD PhD¹,², Peter Hallas MD³, Lars Folkestad MD⁴,⁵, Cecilie Hovitz Lautrup-Larsen BA², Jacob Broder Brodersen MD²

¹ Department of Emergency Medicine, Odense University Hospital, Sdr. Boulevard 29, DK-5000 Odense C, Denmark
² Department of Emergency Medicine, Hospital of South West Jutland, Finsensgade 35, DK-6700 Esbjerg, Denmark
³ Department of Anaesthesiology, Juliane Marie Centret, Rigshospitalet, Blegdamsvej 9, 2100 København Ø
⁴ Department of Endocrinology, Hospital of South West Jutland, Finsensgade 35, DK-6700 Esbjerg, Denmark
⁵ Institute of Clinical Research, University of Southern Denmark, Campusvej 55, DK-5000 Odense, Denmark

* Corresponding author: E-mail: mikkel.brabrand@rsyd.dk, phone: +45 79 18 20 00

Words 829.
Abstract

Objective

To evaluate the inter-observer reliability of nurses assessing respiratory rate.

Methods

We presented seven minimum 60-seconds long videos of thoraces of non-identifiable patients breathing to experienced nurses from several Danish emergency departments. Two videos were assessed by 50 raters while five were reviewed by eight. The videos were shown using an online system that also recorded the counted respiratory rate.

Results

A total of 140 nurses participated with a median of 15 years’ experience. The range of counted respiratory rate was minimum 10 on each video. For videos evaluated by eight nurses, average Inter Class Coefficient (ICC) was 0.662 (0.000-0.960) and individual ICC 0.197 (0.000-0.750). For the two case-videos analyzed by 50 nurses, average ICC was 0.0 (0.000-0.991) and individual ICC 0.0 (0.000-0.677).

Conclusions

We found a wide variation in measurements of RR with both few and many observers assessing exactly the same patients.
Introduction

Respiratory rate (RR) is part of most risk stratification tools, and abnormal values are correlated to poor outcome. However, the value of using RR in risk scores highly depends on the reliability of the estimation of RR. The existing studies assessing interrater reliability have been performed in laboratory settings or have used few raters per patient. We present data on interrater reliability assessing RR in a clinical setting where multiple nurses were asked, based on video recordings of patients, to triage the patients using different scoring systems. The aim of this sub-study was to evaluate nurses’ interrater reliability assessing RR.

Methods

This was a sub-study of Triage of Acute medical Patients – Interobserver Reliability study (TAPIR), a study evaluating interobserver reliability in measuring vital signs and clinical scoring systems. This sub-study used an online questionnaire distributed via e-mail to members of the Danish Emergency Nurses Association. The questionnaire included anonymized videos of real patient cases and each participant only saw one case-video. To evaluate RR, the videos showed a close-up of at least 60 seconds of the patient’s chest movements. We used seven case-videos. The first two case videos were distributed to 60 participants each, and the latter five case videos were distributed to 10 participants each. We calculated Inter Class Coefficient (ICC) using the one-way random-effects model as a measure of agreement as well as absolute measurements. Data will be presented as median (range) or ICC (95% confidence interval). Stata 14.1 was used for analyses. The need for approval was waived by the regional ethics committee; all patients had consented to be filmed and the video to be used in the study.

Results

Of 170 invited participants, 140 (82%) replied. Participants had median 15 years clinical experience (2-40). When evaluating the five case-videos, each analyzed by eight (of 10 invited) nurses, we found an average ICC of 0.662 (0.000-0.960) and an individual ICC of 0.197 (0.000-0.750). Range of measured RR was 22-36, 24-40, 14-32, 12-30 and 22-32, respectively (figure 1). For the two case-videos analyzed by 50 nurses each (of 60 invited), we found an average ICC of 0.0 (0.000-0.991) and an individual ICC of 0.0 (0.000-0.677). Range of measured RR was 20-60 and 19-40, respectively (figure 1).
Discussion

Using our novel approach with multiple raters evaluating few patients in a completely standardized clinical environment, we found that few raters were able to reach suboptimal level of agreement while many raters for only two patients showed poor agreement. This was likewise illustrated (as seen in figure 1) by the fact that each case had a wide range of measurements, more than 10 breaths per minute.

Our videos all showed each patient for at least 60 seconds. This enabled the participants to measure RR according to the standard protocol. However, we cannot know how the participants performed the measurement; inconsistency in methodology of counting a RR (e.g. counting for 15 seconds and multiplying by 4) might explain some of the large variations we found. Lim et al. found a similar phenomenon with variation of 6 breaths per minute when few raters assessed multiple patients.

The use of videos limits our study, considering that the rater is limited to what is shown on the screen, the quality of the video feed, and, due to the videos’ focus only on chest movement, lacking a more global evaluation of the study subjects. However, the study-setup allowed multiple raters to see the same patient and evaluate the same time span in the patient’s clinical course, something that would be unrealistic to set up without a video. Our group has previously used video recordings in a laboratory-based setup. Here, a healthy subject breathed at a standardized rate without variation over 60 seconds. In this hypothetic environment, we obtained excellent agreement. This suggests that the rater’s approach/method change when not given the specific task of measuring RR, but merely measuring RR as part of triage – similar to a clinical environment, where several assessments are done simultaneously.

RR has been proven to be one of the best predictors of adverse events. However, the variation found could have a large impact on calculated risk scores, e.g. early warning scores. While a RR of 20 would result in a score of 0, the effect of a RR of 25 would be a score of 3 using the National Early Warning Score. Our data thus implies that bedside, visual measurement of RR is not reliable for the purpose of risk scoring because of a substantial interobserver variation. There is a risk that the large interobserver variation will be passed onto the risk score, rendering it a less reliable tool. A simple way to overcome this might be to use automated RR measurements, thereby limiting interobserver variation.
Conclusion
In conclusion, we found a wide variation in measurements of RR with both few and many observers assessing exactly the same patients. This is cause for concern as RR is an intricate part of most risk stratification tools in modern medicine.

Compliance with Ethical Standards

Funding: This study received no external funding.

Conflicts of interest: The authors declare that they have no conflicts of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References
Figure legends

Figure 1 – A bar graph showing the range in calculated respiratory rates for the seven cases. As seen, we found a wide variation in the count of breaths per minute.
Highlights

- Respiratory rate is part of most risk stratification tools, and abnormal values are correlated to poor outcome.
- Using video recordings of different patients breathing we found wide variation in measurements of respiratory rate by observers assessing exactly the same patients.
Figure 1
Figure 3