International perception of lung sounds: a comparison of classification across some European borders

Juan Carlos Aviles-Solis,[1] Sophie Vanbelle,[2] Peder A Halvorsen,[1] Nick Francis,[3] Jochen W L Cals,[4] Elena A Andreeva,[5] Alda Marques,[6] Päivi Piirilä,[7] Hans Pasterkamp,[8] Hasse Melbye[1]

ABSTRACT

Introduction Lung auscultation is helpful in the diagnosis of lung and heart diseases; however, the diagnostic value of lung sounds may be questioned due to interobserver variation. This situation may also impair clinical research in this area to generate evidence-based knowledge about the role that chest auscultation has in a modern clinical setting. The recording and visual display of lung sounds is a method that is both repeatable and feasible to use in large samples, and the aim of this study was to evaluate interobserver agreement using this method.

Methods With a microphone in a stethoscope tube, we collected digital recordings of lung sounds from six sites on the chest surface in 20 subjects aged 40 years or older with and without lung and heart diseases. A total of 120 recordings and their spectrograms were independently classified by 28 observers from seven different countries. We employed absolute agreement and kappa coefficients to explore interobserver agreement in classifying crackles and wheezes within and between subgroups of four observers.

Results When evaluating agreement on crackles (inspiratory or expiratory) in each subgroup, observers agreed on between 65% and 87% of the cases. Conger's kappa ranged from 0.20 to 0.58 and four out of seven groups reached a kappa of ≥0.49. In the classification of wheezes, we observed a probability of agreement between 69% and 99.6% and kappa values from 0.09 to 0.97. Four out of seven groups reached a kappa ≥0.62.

Conclusions The kappa values we observed in our study ranged widely but, when addressing its limitations, we find the method of recording and presenting lung sounds with spectrograms sufficient for both clinic and research. Standardisation of terminology across countries would improve international communication on lung auscultation findings.

Key messages

- We found variation in the level of agreement when clinicians classify lung sounds.
- Digital recordings with the use of spectrograms is a method suitable for research of lung sounds.
- Standardisation of the terminology of lung sounds would improve internatonal communication on the subject.

crackles and κ=0.51 for wheezes when clinicians classified lung sounds.[9] Similar results have been found in other studies.[10–14] Lower agreement levels have also been found.[7 15]

However, most of these agreement measures were based on clinicians sequentially listening to patients with a stethoscope. Clinicians working in the same hospital department have rated the sounds in these studies making the sample homogeneous and applicability of the results may be questioned.[1 11–14] In addition, the use of such methods would be difficult to implement in large epidemiological studies due to logistical challenges. New methods are needed for clinical research in this area to generate evidence-based knowledge about the role that lung sounds have in a modern clinical setting.

Studies of interobserver agreement using lung sound recordings, rather than traditional auscultation, may be a good alternative.[15–17] Recorded sounds may be presented with a visual display, and creating spectrograms of lung sounds is already an option in the software of electronic stethoscopes. Recording and visual display of lung sounds may be applied in large samples and classifications of the sounds may be repeated. However, we still do not know the reliability of such classifications.

The aim of the present study was to describe the interobserver agreement among an international sample of raters, including general

For numbered affiliations see end of article.

Correspondence to
Dr Juan Carlos Aviles-Solis; juan.c.solis@uit.no

INTRODUCTION

Lung auscultation is an old and well-known technique in clinical medicine. Adventitious lung sounds, such as wheezes and crackles, are helpful in the diagnosis of several lung and heart-related conditions.[1–5] However, the diagnostic value of chest auscultation may be questioned due to variability in recognising lung sounds.[6–8] In a scale from 0 to 1, a study by Spiteri *et al* found a kappa of κ=0.41 for

practitioners (GP), pulmonologists and medical students, when classifying lung sounds in adults aged 40 years or older using audio recordings with display of spectrograms.

METHODS

In August to October 2014 we conducted a cross-sectional study to explore agreement in the classification of lung sounds. In order to obtain material to classify, we recruited a convenience sample of 20 subjects aged 40 years or older. We took contact with a rehabilitation programme in northern Norway for patients with heart and lung-related diseases (lung cancer, chronic obstructive pulmonary disease, heart failure, and so on). We got permission to hold a presentation about lung sounds and at the end of the presentation we invited the patients to be part of our research project as subjects. Fourteen patients attending the rehabilitation programme agreed to participate and we recorded the lung sounds that same evening. The patients were 67.43 years old on average (44–84) and nine were female. To hold a balanced sample (concerning the prevalence of wheezes, crackles and normal lung sounds), we obtained the rest of our recordings from six self-reported healthy employees at our university aged 51.83 years old on average (46–67) and five were female. We registered the following information about the subjects: age, gender and self-reported history of heart or lung disease. No personal information was registered that could link the sound recordings to the individual subjects.

Recording of lung sounds

To record the lung sounds, we used a microphone MKE 2-EW with a wireless system EW 112-P G3-G (Sennheiser electronic, Wedemark, Germany) placed in the tube of a Littmann Master Classic II stethoscope (3M, Maplewood, MN, USA) at a distance of 10 cm from the headpiece. The microphone was connected to a digital sound Handy recorder H4n (Zoom, Tokyo, Japan).

We placed the membrane of the stethoscope against the naked thorax of the subjects. We asked the subjects to breathe deeply while keeping their mouth open. We started the recording with an inspiration and continued for approximately 20 s trying to capture three full respiratory cycles with good quality sound. We performed this same procedure at six different locations (figure 1). The researcher collecting recordings used a headphone as an audio monitor to evaluate the quality of the recording. When too much noise or cough was heard during the recording, a second attempt was performed.

We obtained a total of 120 audio files. The audio files were in '.wav' format and recorded at a sample rate of 44 100 Hz and 16 bit depth in a single monophonic channel. We did not perform postprocessing of the sound files or implement filters.

Presentation of the sounds

One researcher (HM) selected the sections with less noise according to his acoustic perception. Breathing
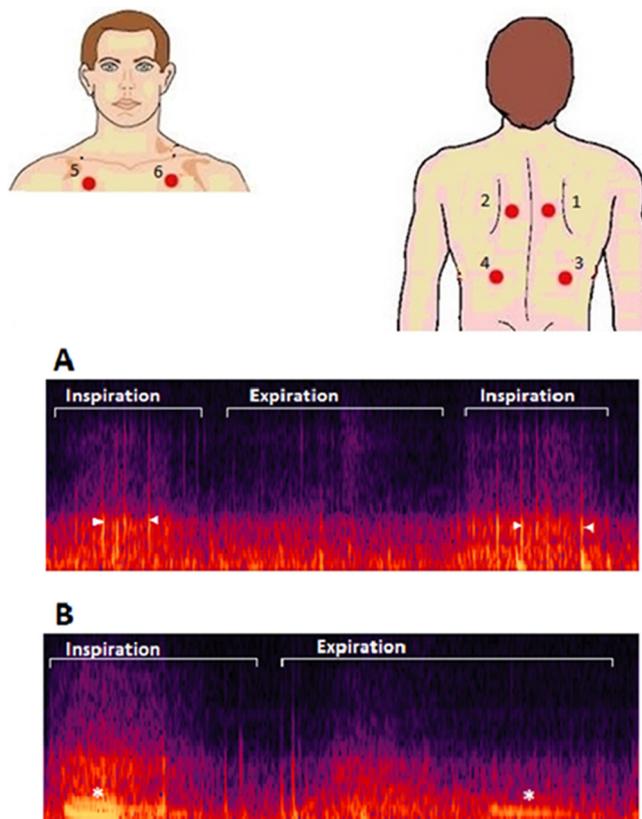

Figure 1 Upper: illustration showing the different places where lung sounds were recorded. (1_2) Between the spine and the medial border of the scapula at the level of T4–T5; (3_4) at the middle point between the spine and the mid-axillary line at the level of T9–T10; (5_6) at the intersection of the mid-clavicular line and second intercostal space. Lower: image showing two different spectrograms containing crackles (A) and wheezes (B). Crackles appear as vertical lines (arrowheads) and wheezes as horizontal lines (*).

phases were determined by listening to the recordings (which usually started with inspiration) and visual analysis of the spectrograms. A spectrogram for each of these recordings was created using Adobe Audition V.5.0 (Adobe Systems, San Jose, CA, USA) (figure 1). The spectrograms showed time on the x-axis, frequency on the y-axis and intensity by colour saturation. Videos of the selected spectrograms, where an indicator bar follows the sound, were made from the computer screen using Camtasia Studio V.8 software (TechSmith, Okemos, MI, USA). We compiled these 120 videos of lung sounds in a PowerPoint presentation (Microsoft, Redmond, WA, USA). Age, gender and recording location, but no clinical information, were presented about the subjects. The majority of the recordings started during inspiration and if that was not the case, this was specified.

Recruitment of the raters and classification of the files

We recruited seven groups of four raters to classify the 120 recordings: We wanted a heterogeneous sample,

Table 1 Prevalence, probability of agreement, Conger's kappa (SE) and 95% CI for the seven groups of observers when classifying 120 sound files for the presence of crackles and wheezes

	Prevalence	P (agree)	Kappa	SE (kappa)	95% CI
Crackles					
Experts	0.21	0.86	0.56	0.080	0.40 to 0.72
GP Norway	0.23	0.85	0.58	0.083	0.42 to 0.74
GP Russia	0.31	0.65	0.20	0.051	0.10 to 0.30
GP UK	0.17	0.87	0.53	0.089	0.36 to 0.70
GP Netherlands	0.17	0.86	0.49	0.105	0.28 to 0.70
Students	0.27	0.76	0.40	0.086	0.23 to 0.57
Pulmonologists	0.29	0.74	0.37	0.082	0.21 to 0.53
Wheezes					
Experts	0.079	0.96	0.75	0.125	0.51 to 1
GP Norway	0.083	0.94	0.62	0.163	0.30 to 0.94
GP Russia	0.22	0.69	0.09	0.076	−0.06 to 0.24
GP UK	0.065	0.99	0.97	0.024	0.92 to 1.00
GP Netherlands	0.050	0.94	0.39	0.087	0.22 to 0.56
Students	0.073	0.95	0.66	0.042	0.58 to 0.74
Pulmonologists	0.14	0.82	0.27	0.102	0.07 to 0.47

GP, general practitioner.

therefore we included GPs from the Netherlands, Wales, Russia, and Norway, pulmonologists working at the University Hospital of North Norway, an international group of experts (researchers) in the field of lung sounds (Pasterkamp H, Piirila P, Sovijärvi A, Marques A) and sixth year medical students at the Faculty of Health Sciences at UiT, The Arctic University of Norway. We chose to have four raters in each group for pairwise comparisons. The mean age of the groups of raters varied between 25 (the students) and 59 years (the lung sound researchers), and years of experience from 0 (the students) to 28.5 (the lung sound researchers).

All the 28 observers independently classified the 120 recordings. We first asked the observers to classify the lung sounds as normal or abnormal. If abnormal, they had to further classify them as containing crackles, wheezes or other abnormal sounds. It was possible to mark more than one option. The observers specified whether the abnormalities occurred in inspiration or expiration. In addition, they could mark if there was noise present in the recording. We offered two options for answering the survey: an electronic form in Microsoft Access (Microsoft), and a printed version of the questionnaire. We did not perform training of the raters. To make the raters familiar with sounds and spectrograms, the PowerPoint presentation with the 120 recordings started with a demonstration of the three examples, one with normal lung sounds, one with crackles and one with wheezes. The raters were free to play the videos (containing the sound recording and the spectrogram simultaneously) several times and to go back and forth through the cases ad libitum. We used English language

in the presentation of the videos and the survey forms. In Russia and the Netherlands, observers were offered translations of the terms included in the survey. These translations were taken from previous studies using lung sound terminology.[18 19]

Statistical analysis

We calculated the probability of agreement and multi-rater Conger's kappa using the delta method for the analysis of multilevel data.[20] Conger's kappa coefficient was chosen over Fleiss' kappa because the observers classifying the sounds were the same for all sounds. We analysed the intragroup agreement in each of the seven groups of observers when classifying the recordings for the presence of wheezes and crackles disregarding the breathing phase. We used the statistical software 'R' V.3.2.1 together with the package 'multiagree' for the statistical analysis of kappa statistics.[21]

In order to permit the comparison of the agreement levels between and within groups, within and between-group agreement levels were summarised in a matrix, where the diagonal elements represent the mean agreement level between all possible pairs formed by two observers in the same group, and the off-diagonal elements represent the mean agreement level between all possible pairs with one observer in one group and the second observer of the pair in another group. This information was summarised in correlograms using the R package 'Corrplot'.[22]

This study has been reported according to the Guidelines for Reporting Reliability and Agreement Studies.[23]
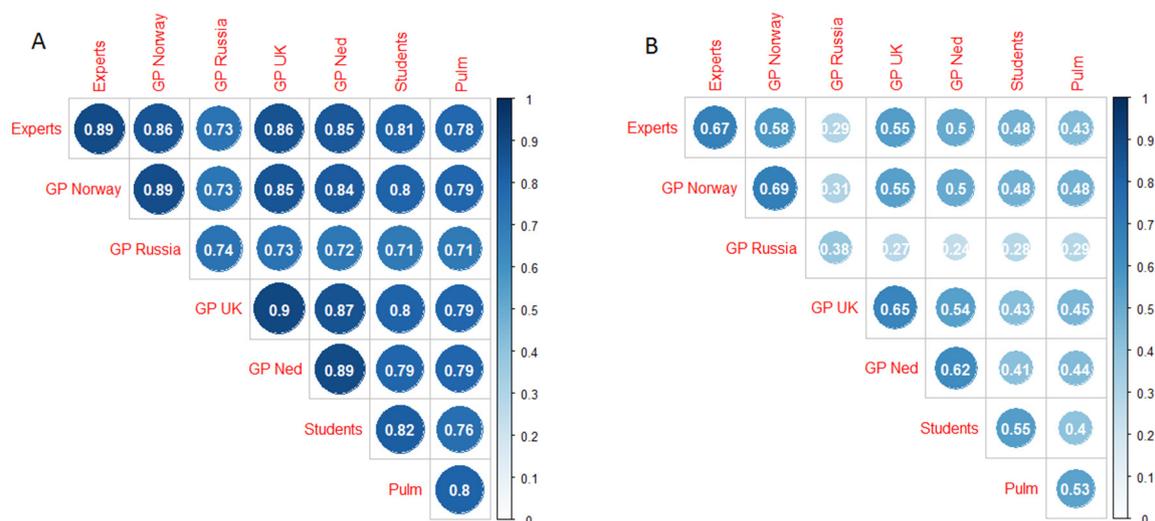
Figure 2 Average proportion of agreement (A) and kappa (B) between pairs of raters from the same (diagonal) and different (off-diagonal) groups when classifying the sounds for the presence of crackles. GP, general practitioner.

RESULTS

Prevalence of wheezes and crackles

All the 28 observers independently classified the 120 recordings. According to the experts' classification, crackles were present in 21% of the 120 recordings and wheezes in 7.9%. Per case (n=20), 15% of the individuals had wheezes, and 50% had crackles in one or more recordings. The prevalence of crackles and wheezes in the 120 recordings varied between groups with mean values among the four observers of 17.0%–29% for crackles and 5.0%–22% for wheezes (table 1). The group average noise reporting ranged from 1.46% to 17.70% (mean=7.5%) of the recordings. There was no significant correlation between the use of this variable and agreement or kappa coefficients. The groups with the highest level of agreement tended to use this variable more often.

Interobserver agreement within the same group

When evaluating interobserver agreement on crackles (inspiratory or expiratory) in each subgroup, observers agreed on between 65% and 87% of the cases. Conger's kappa ranged from 0.20 to 0.58 (table 1) and four out of seven groups reached a kappa of ≥0.49 (median). In the classification of wheezes, we observed a probability of agreement between 69% and 99.6% and kappa values from 0.09 to 0.97 (table 1). Four out of seven groups reached a kappa ≥0.62 (median).

Interobserver agreement between different groups

Lower range probability agreement (<0.8 for crackles and <0.9 for wheezes) within a group was associated with a lower range probability agreement with members of
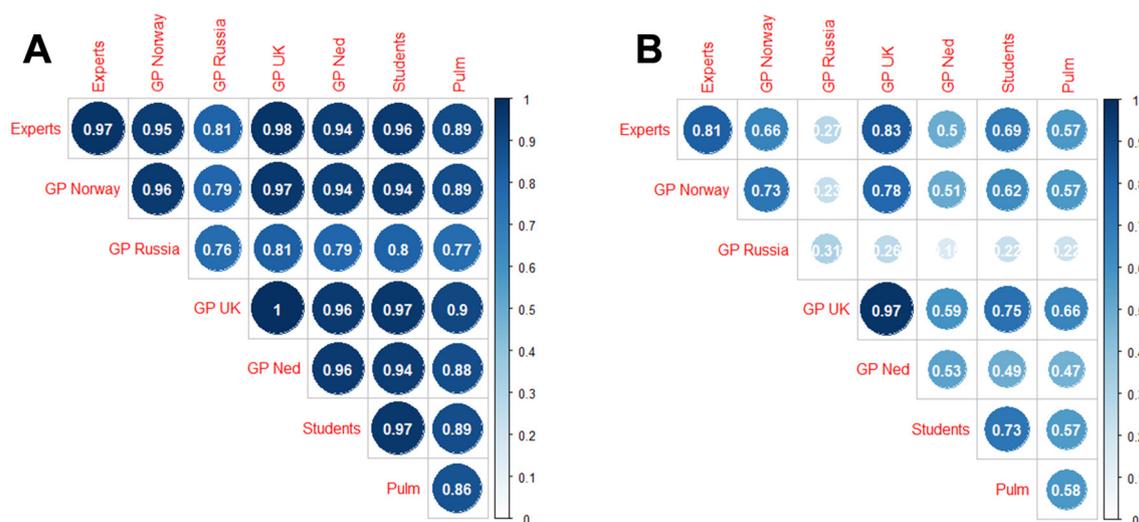

Figure 3 Average proportion of agreement (A) and kappa (B) between pairs of raters from the same (diagonal) and different (off-diagonal) groups when classifying the sounds for the presence of wheezes. GP, general practitioner.

other groups. Correspondingly, high agreement within a group was associated with high agreement with members of other groups (figures 2 and 3).

In particular, the probability of agreement between GPs and the experts was very similar to the probability of agreement within the group of experts (0.86 for crackles and 0.96 for wheezes), except for the group of Russian GPs. Students agreed slightly less with the experts (0.81 for crackles and 0.96 for wheeze) while pulmonologists showed even lower agreement levels with the experts (0.78 for crackles and 0.89 for wheeze). Similar conclusions can be drawn according to Cohen's kappa coefficient values (figures 2 and 3).

DISCUSSION

This study showed a median kappa agreement of 0.49 for crackles and 0.62 for wheezes in the observer groups. Even though kappa coefficients are not directly comparable, our results are similar to those found in other studies analysing interobserver agreement when classifying for wheezes,[10–12] crackles[16] or for both.[6 7 9 13–15 24] The kappa agreements we found were not inferior to those found for other widely accepted clinical examinations.[2 25–29]

In our study, when the agreement levels between clinicians from the same country were in a higher range, we also found a higher level of agreement with members of other groups and vice versa. This finding argues for a general understanding across groups about how to classify crackles and wheezes with some groups encountering greater difficulty in uniform classification.

We found the highest levels of agreement within the experts and some groups of GPs. GPs might be more familiar with the use of lung auscultation, since information from chest imaging, advanced lung function testing or blood gas analysis is not available. Also, GPs are more used to listening to normal lung sounds and sounds with discrete abnormalities. This may have been reflected in the similar levels of agreement between GPs from UK and Norway and the experts in this study.

Strengths and limitations

It was a strength of our study that we included a group of experienced lung sound researchers. They represent recommended use of terminology, and comparison with their classifications may be enlightening, although they were not used as a reference standard.

A strength of our study was also the heterogeneity of the observers in terms of clinical background, experience and country of residency. We believe this gives us a better external validity than if we had included a homogenous sample. However, this factor also presented some challenges concerning language and terminology, which was a weakness of the study.

Different use of lung sound terminology may influence the interobserver agreement.[24 30] The group of Russian GPs had a lower intragroup and intergroup agreement.

We think this situation might be partly explained by confusion around the terminology. Anecdotally, we note that the Russian GPs were familiar with a terminology for lung sounds similar to the classic terminology of Laennec, which offers more options than the simple distinction between wheezes and crackles.[31] A higher agreement within the group and with the experts would probably be found if the study had been based on their own terminology. A similar problem was present in the Dutch sample, where the observers found it difficult to classify what they call 'rhonchi' as wheezes or crackles and used the variable 'other abnormal sounds' more frequently than the other groups. In contrast, a terminology restricted to wheezes and crackles is used in UK and Norway, and this has probably made it easier to obtain higher agreements in these countries.

We did not present audiological definitions of crackles and wheezes.[32] As indicated by the Russian and Dutch classifications, the example sounds and the translations to own language did not quite remove the terminology problems. However, clinicians are not familiar with audiological definitions, and we do not think such definitions would have been helpful.

Implications for research

For future research, it is important to be aware that it might be difficult to reach high kappa values when the prevalence of the trait of study is very low or very high, even though absolute agreement may be high.[33 34] This has probably had little impact on the kappa coefficients we observed, since the prevalence of crackles and wheezes was 21% and 7.9%, respectively. However, much lower prevalence of adventitious lung sounds could be found in real epidemiological data. Accordingly, specific measures should be implemented when using this method in epidemiological studies in order to improve its reliability such as training of raters, consensus agreement, multiple independent observations and standardisation of the terminology.[35]

Conclusion

The strength of agreement and correspondingly kappa values were wide ranging, and some groups found it more challenging to produce uniformity in breath sound classification than others. Although the technology was through our experience found to be quite suitable for research, standardisation of terminology across countries with supportive training could improve international communication on lung auscultation findings.

Author affiliations
[1]General Practice Research Unit, Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway
[2]Department of Methodology and Statistics, University of Maastricht, Maastricht, The Netherlands
[3]Department of Primary Care and Public Health, Cardiff University, Cardiff, UK
[4]Department of Family Medicine, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

⁵Department of Family Medicine, Northern State Medical University (NSMU), Arkhangelsk, Russia
⁶Lab 3R–Respiratory Research and Rehabilitation Laboratory, School of Health Sciences (ESSUA) and Institute for Research in Biomedicine (iBiMED), University of Aveiro, Aveiro, Portugal
⁷Unit of Clinical Physiology, HUS Medical Imaging Center, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
⁸Department of Pediatrics and Child Health, University of Manitoba College of Medicine, Winnipeg, Manitoba, Canada

REFERENCES

1. Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? *JAMA* 1997;277:1712–9.
2. Wang CS, FitzGerald JM, Schulzer M, *et al*. Does this dyspneic patient in the emergency department have congestive heart failure? *JAMA* 2005;294:1944–56.
3. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. *JAMA* 1997;278:1440–5.
4. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA* 1995;273:313–9.
5. Ponikowski P, Voors AA, Anker SD, *et al*. 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur J Heart Fail* 2016;18:891–975.
6. Mulrow CD, Dolmatch BL, Delong ER, *et al*. Observer variability in the pulmonary examination. *J Gen Intern Med* 1986;1:364–7.
7. Elphick HE, Lancaster GA, Solis A, *et al*. Validity and reliability of acoustic analysis of respiratory sounds in infants. *Arch Dis Child* 2004;89:1059–63.
8. Jauhar S. The demise of the physical exam. *N Engl J Med* 2006;354:548–51.
9. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet* 1988;1:873–5.
10. Badgett RG, Tanaka DJ, Hunt DK, *et al*. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med* 1993;94:188–96.
11. Badgett RG, Tanaka DJ, Hunt DK, *et al*. The clinical evaluation for diagnosing obstructive airways disease in high-risk patients. *Chest* 1994;106:1427–31.
12. Margolis PA, Ferkol TW, Marsocci S, *et al*. Accuracy of the clinical examination in detecting hypoxemia in infants with respiratory illness. *J Pediatr* 1994;124:552–60.
13. Wipf JE, Lipsky BA, Hirschmann JV, *et al*. Diagnosing pneumonia by physical examination: relevant or relic? *Arch Intern Med* 1999;159:1082–7.
14. Holleman DR, Simel DL, Goldberg JS. Diagnosis of obstructive airways disease from the clinical examination. *J Gen Intern Med* 1993;8:63–8.
15. Brooks D, Thomas J. Interrater reliability of auscultation of breath sounds among physical therapists. *Phys Ther* 1995;75:1082–8.
16. Workum P, DelBono EA, Holford SK, *et al*. Observer agreement, chest auscultation, and crackles in asbestos-exposed workers. *Chest* 1986;89:27–9.
17. McCollum ED, Park DE, Watson NL, *et al*. Listening panel agreement and characteristics of lung sounds digitally recorded from children aged 1-59 months enrolled in the Pneumonia Etiology Research for Child Health (PERCH) case-control study. *BMJ Open Respir Res* 2017;4:e000193.
18. Andreeva E, Melbye H. Usefulness of C-reactive protein testing in acute cough/respiratory tract infection: an open cluster-randomized clinical trial with C-reactive protein testing in the intervention group. *BMC Fam Pract* 2014;15:80.
19. Francis NA, Melbye H, Kelly MJ, *et al*. Variation in family physicians' recording of auscultation abnormalities in patients with acute cough is not explained by case mix. A study from 12 European networks. *Eur J Gen Pract* 2013;19:77–84.
20. Vanbelle S. Comparing dependent kappa coefficients obtained on multilevel data. *Biom J* 2017;59:1016–34.
21. Vanbelle S. *R package "multiagree": comparison of dependent kappa coefficients*, 2017.
22. Wei T, Simko V. *R package "corrplot": visualization of a correlation matrix*, 2017.
23. Kottner J, Audigé L, Brorson S, *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
24. Melbye H, Garcia-Marcos L, Brand P, *et al*. Wheezes, crackles and rhonchi: simplifying description of lung sounds increases the agreement on their classification: a study of 12 physicians' classification of lung sounds from video recordings. *BMJ Open Respir Res* 2016;3:e000136.
25. Veloso SG, Lima MF, Salles PG, *et al*. Interobserver agreement of gleason score and modified gleason score in needle biopsy and in surgical specimen of prostate cancer. *Int Braz J Urol* 2007;33:639–51.
26. Timmers JM, van Doorne-Nagtegaal HJ, Verbeek AL, *et al*. A dedicated BI-RADS training programme: effect on the inter-observer variation among screening radiologists. *Eur J Radiol* 2012;81:2184–8.
27. Brosnan M, La Gerche A, Kumar S, *et al*. Modest agreement in ECG interpretation limits the application of ECG screening in young athletes. *Heart Rhythm* 2015;12:130–6.
28. Halford JJ, Shiau D, Desrochers JA, *et al*. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clin Neurophysiol* 2015;126:1661–9.
29. Zhou S, Zha Y, Wang C, *et al*. [The clinical value of bedside lung ultrasound in the diagnosis of chronic obstructive pulmonary disease and cardiac pulmonary edema]. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue* 2014;26:558–62.
30. Pasterkamp H, Brand PL, Everard M, *et al*. Towards the standardisation of lung sound nomenclature. *Eur Respir J* 2016;47:724–32.
31. Robertson AJ, COOPE R. Rales, rhonchi, and laennec. *Lancet* 1957;273:417–23.
32. MLP SARA, Charbonneau G, Vanderschoot J, *et al*. Characteristics of breath sounds and adventitious respiratory sounds. *Eur Respir Rev* 2000;10:591–6.
33. Guggenmoos-Holzmann I. The meaning of kappa: probabilistic concepts of reliability and validity revisited. *J Clin Epidemiol* 1996;49:775–82.
34. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002;21:2109–29.
35. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. *J Clin Epidemiol* 1988;41:959–68.