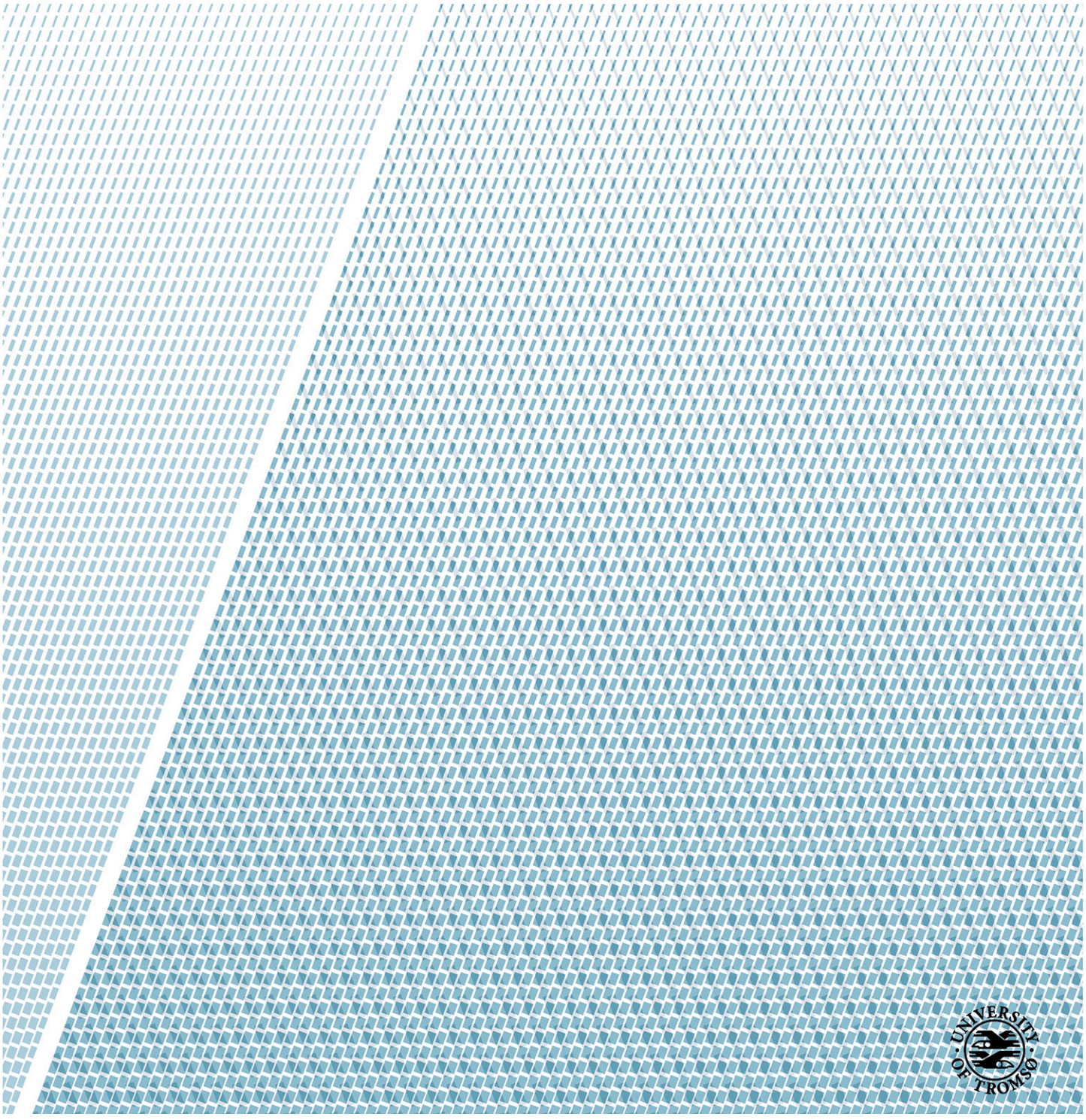
UiT
THE ARCTIC UNIVERSITY OF NORWAY

Intra- and interobserver variation in lung sound classification. Effect of training.

Med-3950 –Fifth year task at the University of Tromsø, the Arctic University of Norway.

Fredrik Ostad Grunnreis
Faculty of Health Science, Medical profession, class of 2011
Supervisor: Hasse Melbye

Tromsø Mai 2016.



Index

Preface

This study serves to inform about the inconsistency in reporting lung sounds among sixth year medical students at the University of Tromsø.

Training on lung sound classification is thought to make the students less inconsistent, therefore a course is included as part of the study given to one half of the attending students.

The project is funded by General Practice Research Unit at the University and was initiated by Hasse Melbye, supervisor and head of General Practice Research Unit.

Colleague in the project has been PhD Student Juan Carlos Aviles Solis. I would like to give Solis a big thank for the important help he has given me on the project. He has been indispensable. Also a big thank to Melbye who invited me to be part of his group of researchers, and have given me important help in the process of writing and interpretation of the results.

Tromsø 29.05.16

Summary

This study explores how last year medical students at the University of Tromsø, the Arctic University of Norway, interpret and describe different lung sounds. This is done by measuring intra- and interobserver variation in agreement among 16 students, when reporting abnormal lung sounds after listening to audio recordings. Agreement with a reference standard is included, and testing of effect on training on these agreements. To test the training effect the students were separated in two groups, one of them having an intervention, a 3 hour course.

The results serves to inform the medical society about the inconsistency in reporting lung sounds in this particular population, and hopefully also help finding measures to obtain better agreement.

Cohens kappa have been used to measure intraobserver agreement and agreement with the reference standard, Fleiss kappa to measure interobserver agreement. An “exact” Mann-Whitney U test for testing the effect of the course. The kappa level of agreement set to define acceptable agreement is “moderate”, with a lower limit of .41.

The results indicate highly acceptable intraobserver agreement, and the agreement tended to improve in both the intervention and the control group. The agreement with the reference standard was also highly acceptable for the category wheezes and acceptable for crackles and the abnormal category. A tendency to positive change in the intervention group when compared to the control group was found, but the difference was only statistically significant for the abnormal category in the agreement against reference standard. The interobserver agreement did not reach the limit of acceptable, except for wheezes. Summarized, a weak effect of the intervention was observed.

Introduction

History

Today the most prominent sign of a doctor appearing is perhaps the use of a stethoscope as a clinical instrument. Doctors have used it as part of the clinical examination since the 1820´s. The inventor Rene Theophile Hyacinthe Laënnec (1781-1826), a French physician, inspired by two children playing with a piece of wood, made the first prototype in 1816, a wooden tube. It all started with Laënnec visiting a patient suggested to have heart problems. The standard way of doing heart examination these days would have included the Hippocratic method of putting the ear to the chest of the patient. Laënnec did not want to be inappropriately intimate with his female patient, so this led him to roll a piece of paper and place it at the chest instead. He then discovered that this method made the acoustics more prominent, and he therefore started to work on a portable and better device to explore acoustic phenomenons.

It did not take him long to complete the wooden stethoscope, but it took 3 years of research to produce and publish the work in a thesis. During the same time he also expanded the use to include lung auscultation. His diagnostic work using the stethoscope was supported by autopsies, and the thesis *De L'auscultation Mediate* (On Mediate Auscultation) was published in 1819.

The wooden stethoscope stayed in clinical practice until rubber tubes in latter half of the 19th century replaced it. Since then a lot of modifications have been made, ending up with the stethoscope, as we recognize it these days, with a combined bell and diaphragm chest-piece.

(1)

Background

Studies investigating intra- and interobserver evaluation of lung sounds, shows a marked difference in the use of nomenclature describing the same sound phenomenon and difference in the interpreting of a sound phenomenon (2-4). Pulmonary auscultation can therefore be considered to be a subjective method, based primarily on clinical experience rather than being evidence-based. The diagnostic value of the method is therefore

questionable, and clinical guidelines give low credit to chest findings, especially in the early diagnosis of COPD (5). The development in medical technology and laboratory medicine adds upon this, and are likely to contribute.
Despite this, two recent articles (2014, 2015) states that the method is important in diagnostics (6, 7). The message in the article by Sarkar et al (2015) is mediated with a structured and thoroughly passageway through physics and mechanisms of various types of breath sounds, also including categorization and sub classification of abnormalities. It is claimed that knowledge in the nature of lung sounds is necessary for being able to understand disease processes. It also notes the importance of the method being non-invasive, safe and inexpensive (7), supported by Bohadana et al (2014) (6).

The problem with inconsistency in the use of nomenclature may be explained in the lack of an international standardized agreement (3). The nomenclature in use today stems back from the time of Laënnec. He introduced terminology like “rattle” or “rale”, which he later replaced by “rhonchus” (8). Presumably translational error regarding the terminology of Laënnec and also between different languages is a significant source of the problem. The use of terminology have been unclear, dynamic and differed within borders for decades after the time of Laënnec (3, 8). The first clean up and organization of the nomenclature started in the mid 50´s. The terminology was simplified and categorized. In 1976 the International Lung Sound Association took this a step further. Classification in line with this approach was published by Bohadana et al in NEJM in February 2014, table 1 (6).

Clinical Characteristics and Correlations of Respiratory Sounds.

Respiratory Sound	Characteristics Clinical	Clinical Correlation
Normal tracheal sound	Hollow and nonmusical, clearly heard in both phases of respiratory cycle	Transports intrapulmonary sounds, indicating upper- airway patency; can be disturbed (e.g., become more noisy or even musical) if upper-airway patency is altered; used to monitor sleep apnea; serves as a good model of bronchial breathing
Normal lung sound	Soft, nonmusical, heard only on inspiration and on early expiration	Is diminished by factors affecting sound generation (e.g., hypoventilation, airway narrowing) or sound transmission (e.g., lung

		destruction, pleural effusion, pneumothorax); assessed as an aggregate score with normal breath sound; rules out clinically significant airway obstruction*
Bronchial breathing	Soft, nonmusical, heard on both phases of respiratory cycle (mimics tracheal sound)	Indicates patent airway surrounded by consolidated lung tissue (e.g., pneumonia) or fibrosis
Stridor	Musical, high-pitched, may be heard over the upper airways or at a distance without a stethoscope	Indicates upper-airway obstruction; associated with extrathoracic lesions (e.g., laryngomalacia, vocal- cord lesion, lesion after extubation) when heard on inspiration; associated with intrathoracic lesions (e.g., tracheomalacia, bronchomalacia, extrinsic compression) when heard on expiration; associated with fixed lesions (e.g., croup, paralysis of both vocal cords, laryngeal mass or web) when biphasic
Wheeze	Musical, high-pitched; heard on inspiration, expiration, or both	Suggests airway narrowing or blockage when localized (e.g., foreign body, tumor); associated with generalized airway narrowing and airflow limitation when widespread (e.g., in asthma, chronic obstructive lung disease); degree of airflow limitation proportional to number of airways generating wheezes; may be absent if airflow is too low (e.g., in severe asthma, destructive emphysema)
Rhonchus	Musical, low-pitched, similar to snoring; lower in pitch than wheeze; may be heard on inspiration, expiration, or both	Associated with rupture of fluid films and abnormal airway collapsibility; often clears with coughing, suggesting a role for secretions in larger airways; is nonspecific; is common with airway narrowing caused by mucosal thickening or edema or by bronchospasm (e.g., bronchitis and chronic obstructive pulmonary disease)
Fine crackle	Nonmusical, short, explosive; heard on mid-to-late inspiration and occasionally on expiration; unaffected by cough, gravity-dependent, not transmitted to	Unrelated to secretions; associated with various diseases (e.g., interstitial lung fibrosis, congestive heart failure, pneumonia); can be earliest sign of disease (e.g.,

	mouth	idiopathic pulmonary fibrosis, asbestosis); may be present before detection of changes on radiology
Coarse crackle	Nonmusical, short, explosive sounds; heard on early inspiration and throughout expiration; affected by cough; transmitted to mouth	Indicates intermittent airway opening, may be related to secretions (e.g., in chronic bronchitis)
Pleural friction rub	Nonmusical, explosive, usually biphasic sounds; typically heard over basal regions	Associated with pleural inflammation or pleural tumors
Squawk	Mixed sound with short musical component (short wheeze) accompanied or preceded by crackles	Associated with conditions affecting distal airways; may suggest hypersensitivity pneumonia or other types of interstitial lung disease in patients who are not acutely ill; may indicate pneumonia in patients who are acutely ill

Table 1

Computerized analysis of lung sounds have been available since the 80´s (8). The development in data technology has been formidable since that time, and new technology may help the standardization of terminology, since appropriately described recordings can be used in the education of doctors all over the world. The development also makes it theoretically possible to do bedside analysis, with for example a smartphone installed with an appropriate app. This could be the future of lung auscultation. Until this method by any chance becomes available, it is important to get insight in how medical practitioners interpret and describe lung sounds, and if possible, enhance their skills.

Despite clinical recommendations, research shows that abnormal findings on lung auscultation are associated with decisions regarding treatment: In general practice, positive findings in patients suffering from lower respiratory tract infections, increase the rate of antibiotic prescribing (3, 9, 10). This is an important finding, because it is probably one of many driving forces for the increasing development of antibiotic resistance, a problem that is predicted to be a disaster to modern medicine in few years, if not decelerated.
Some of the explanation may rest on the lack of knowledge in terms of how the sounds are generated. In that case, increased knowledge is potentially a factor that can help reduce the over prescribing of antibiotics.

Aim of study

In this study we wanted to explore how medical students, at their sixth and last year of medical school, at the University of Tromsø, the Arctic University of Norway, interpreted and described different lung sounds. This is done by measuring intra- and interobserver variation in reporting abnormal lung sounds after listening to audio recordings. We also wanted to test if there is an effect of training on these agreements.

The results will inform the medical society about the inconsistency in reporting lung sounds in a population of last year medical students today, and hopefully help finding measures to obtain a better agreement.

In Tromsø medical students get an introduction of lungs sounds and practical use of the stethoscope at their third year of study, as part of the courses in clinical examination and pulmonary medicine. The educational model in Tromsø is much oriented against clinical experience, and the students start meeting patients already at their first year. During fifth year they are six months exclusively in clinical practice, with four months in hospital and two months at a GPs office. Therefore the students are regarded to have good insight in how to handle a stethoscope correctly for medical purposes.

Methodology

Work process

The project was initiated by Hasse Melbye, professor of general practice UIT and head of General Practice Research Unit in Tromsø. It is part of a study that serves as a pilot study for a population based study on lung sounds, a part of the Tromsø Study (Tromsø 7). The other part of the pilot study is led by Juan Carlos Aviles Solis, a medical doctor and a PhD. Student from Mexico. Dr. Solis has been an important team player in my project as well, and have helped me out in the recruiting of students, to carry out the data collection, to sort the data, to understand SPSS and to calculate the interobserver agreement, since our way of calculating multirater agreement was not an option in SPSS.

The project plan was very general, but has worked out fine since the process with the project started as early as September 2014. It started with the recruiting of students in the second half of September. Then the reporting of the lung sounds was done in October. The

work with the data collected, took place through the most of 2015. The written part of the task started in October -15.

Most of the sounds used in the project were already sampled and ready for use. To get enough files we had to sample some more. This was done in October 2014. The details about the project is described below.

The magnitude of the work is great with respect to the duration of time allocated in the curriculum. The process has therefore been challenging taken into consideration it has been on top of studying in the fourth year and the practical training in the fifth year. But it is no doubt this work has given me a start as future medical research scientist. This was also the main goal, beside creating new information about a clinical method performed in the everyday medical practice.

The conduct of the study

16 medical students were recruited. The recruitment was online, through a commercial posted at the university online service, Fronter, and on Facebook. Since the students all were at the same level of education they were expected to have the same level of qualification in rating lung sounds.

From now on I will refer to the students as “observers”.

The observers were divided into two groups, one control group and one intervention group. This was a random selection, where each observer was given a number from 1-16, when they registered for the study. These numbers were plotted in an online randomization program, “Random.org”. The program then created new random numbers from 1-16 for each observer. Afterwards it was decided with a flip of a coin that odd numbers were having intervention. The observers were not informed about which group they belonged to until after the first of the two parts of the study.

The study included two similar surveys, one before and one after the intervention, called survey 1 (S1) and survey 2 (S2). It was four weeks between the two surveys. After two weeks the odd number group were given a 3-hour course in lung sound analysis.

Each survey contained two rounds and a presentation of 40 sound files in each round. The same files were presented in both rounds in a randomized order, using the same program as

described above, with a 15 min break in between. Each round had duration of about 60 min. The observers were not informed that the sound files were the same in the two rounds.

A total of 80 files were used in the study (40 in each survey). These were authentic recordings from a mixed population of real patients, recruited from the LHL lung rehabilitation facility in Skibotn, Troms, and healthy employees at the University aged 40 years or more. The sounds were recorded with a wireless Sennheiser microphone (Sennheiser MKE 2-EW with Sennheiser wireless system EW 112-P G3-G) placed in the tube of a Littmann Classic stethoscope. The auscultations were carried out manually by Melbye, Solis and the author on six different locations on the thorax, figure 1. The patients were asked to breathe deeply with their mouth open when examined.

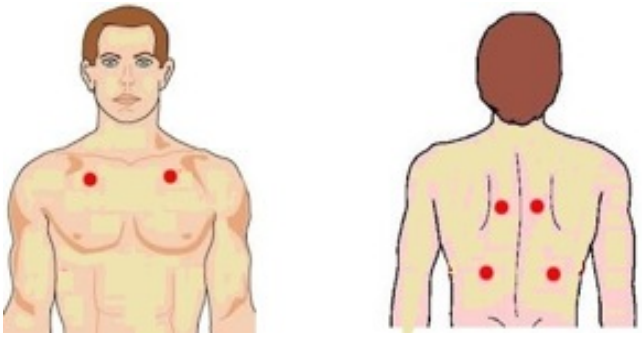


Figure 1: Locations of lung sound recordings

The sounds were presented through two loudspeakers in the front of a small classroom, and the observers were sitting in front of computers on two rows. Each sound file was between 5 and 10 seconds, and was played three times with some seconds in between.
An extra element was added to the presentation of the sounds; a spectrogram was generated by the software "Adobe Audition" for each sound for visualisation. A spectrogram is a visual presentation of the frequencies of a sound phenomenon, see figure 2. When the sounds were played the spectrogram was shown at a whiteboard in the front of the classroom. Lung sounds with spectrograms had been presented at one occasion in the third year of medical school, but no further information was given on how to interpret a spectrogram before survey 1.

Figure 2

Spectrograms were included because they are coming more into use in connection with lung sounds. Probably they represent the future in interpreting lung sounds as a clinician (11, 12), and it is applied by 3M in connection with their electronic stethoscopes as an option in their computer program StethAssist™ (13).

Before the surveys were initiated the observers were given information about how to rate the sounds, but not what kind of sound phenomenon they could expect. The form used for answering you can see in figure 2; first they had to decide if the sounds heard were normal or abnormal. If suggested to be abnormal they had to classify the abnormal sound and decide in what kind of respiratory phase the abnormality appeared. Wheezes and crackles were the default options, but they also had the opportunity to describe the sound with their own words.

At a few occasions some of the observers used own words when describing a wheeze or a crackle. The terms used were “translated” into the default categories. An example is rhonchus, which can be classified as a low-pitched wheeze. Recent literature suggest rhonchus as an own term (6), but the sound share similar clinical correlation of narrowed airways and have a continuous appearance, in contrast with crackles, and is therefore accepted here as a variant of the wheeze.

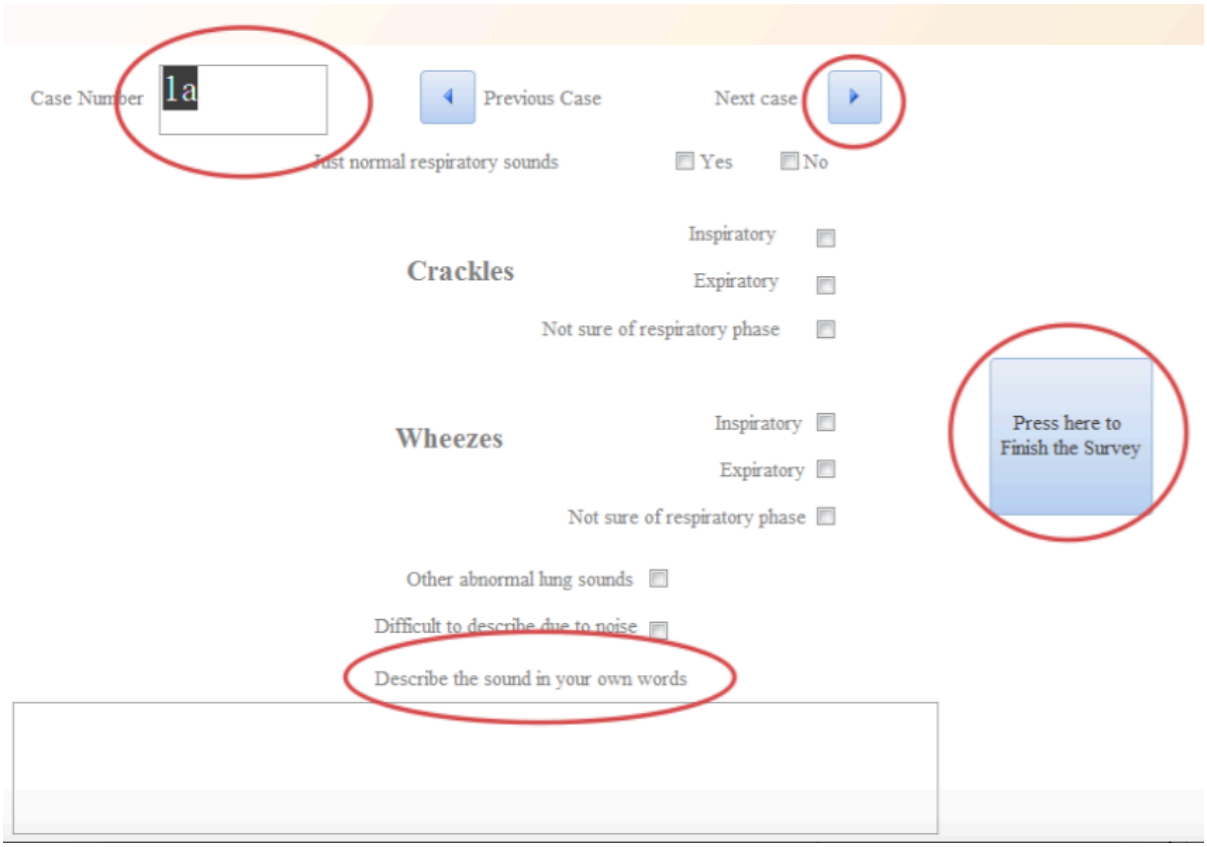


Figure 3: Scheme for evaluating the lung sounds

The sound material

Around 65% of the sound files comprised abnormalities, based on expert classification (described below). They vary between two different types of phenomenons, crackles and wheezes, represented in one or both phases of respiration. Both phenomenons could be represented in one sound file.

For the crackle sound files, both fine and coarse variants were represented. For the wheeze, high-pitched and low-pitched variants. The observers did not have to differentiate between these. The numerical prevalence of the lung sounds is illustrated in table 2 and 3 below.

In addition, a variable named “total” have been created. The totals reflect the summation of all the inspiratory and expiratory ratings for the crackle and the wheeze, and are the total registration of each phenomenon. This variable is also present in table 2 and 3.

The registration of abnormal sounds has been corrected; some of the observers forgot to mark for “abnormal” when registering their abnormality. This has been done by using a logic

command “or”. This means that abnormal is present if wheezes “or” crackles is present. In other words, abnormal could potentially have a lower value than the sum of “wheeze total” and “crackle total” because one case could potentially have both at the same time. This is also illustrated in the two tables below, where the prevalence of crackles total is close to the prevalence of abnormal.

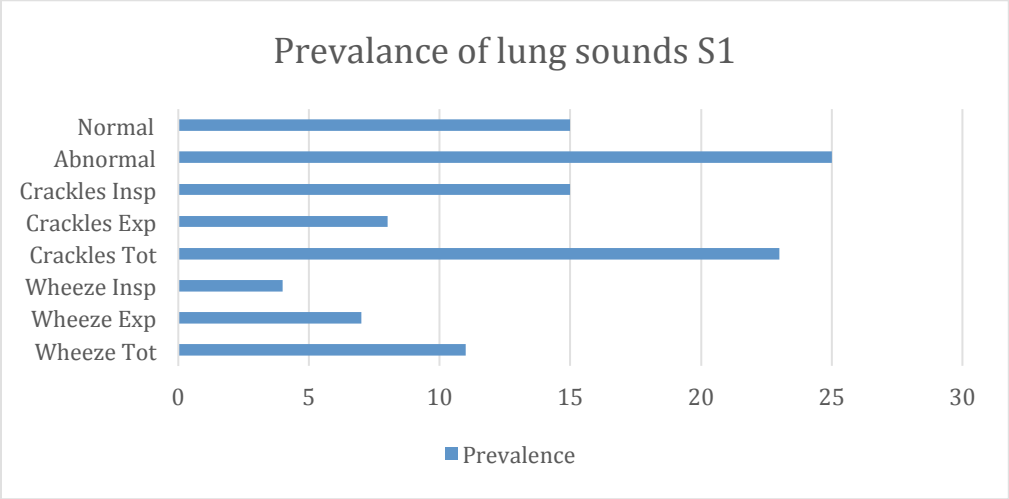


Table 2

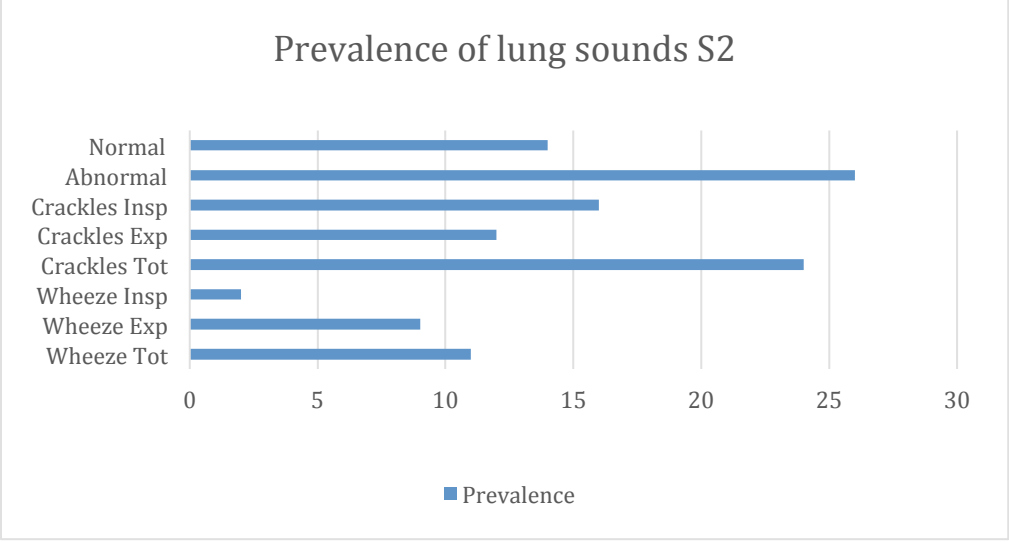


Table 3

The intervention

The course contained updated theory on the nature of the sounds, feedback about survey 1, nomenclature – what is recommended by the European Respiratory Society (ERS) task force for lung sounds, and audio examples connected to each example of nomenclature. The audio examples are published as interactive sound files and graphics at the website of NEJM,

together with the article by Bohadana et al (2014). The feedback was a review of 30 of the sound files from Survey 1 (all the abnormal were represented), were a standard reference was presented and handed out. The observers could themselves compare with their own answers and discuss with a supervisor and each other. Spectrograms were also shown and explained.

The main supervisor at the intervention course was Melbye. Beside being a specialized general practitioner, and head of General Practice Research Unit he also is a researcher in pulmonary medicine, teaches medical students on different levels, is responsible for the video seminar in lung sounds at the third year of study, and is a member of the ERS task force on lung sounds, aiming at a standardization of nomenclature used when describing findings on lung auscultation (14). The other supervisor was Solis.

Data analysis

The data were analysed using kappa statistics to determine both intra- and interobserverer agreement. Cohen's kappa has been used for the intraobserver calculation, and Fleiss kappa for the multirater interobserver calculation. Fleiss kappa is an adaptation of Cohen's kappa, used to test for agreement when the number of observers exceeds two.
In kappa statistics it is taken into account that the possibility of agreement by chance may vary by the frequency of the observation. This is a strength compared to just using percentage agreement.

An "exact" Mann-Whitney U test was used to measure the effect on the course given to the intervention group. 2-tailed values is used. The groups were compared with respect to the change in intraobserver agreement (kappa value) between round 1 (R1) in the two surveys, and tested against each other. This test takes into consideration that the kappa values are not normally distributed, which is essential since the two groups are small (n=8). It is a non-parametric test based on ranks, and is equivalent with the Wilcoxon rank sum test (15).

The Kappa statistics is scaled from -1 to 1, where a value lower than 0 is considered to indicate less than chance agreement, and higher values than 0 indicates greater than chance agreement. 1 is perfect agreement. Negative values are rare, and are of lesser interest

because they are interpreted as worse than *expected*, kappa = 0. Large negative values can indicate problems with the group of observers or the instruments used in the study (16). In this study a few negative values have been detected, they are small, and can be expected by chance when the agreement is poor.

The interpretation of the kappa values is illustrated in table 4. This is a widely used way of organizing the values, and was first described by the inventor of the statistical method, Jacob Cohen. He suggested .41 as the lower limit of acceptable agreement, in other words, within the moderate level of agreement.

Value of Kappa	Agreement
.00	Chance/random
.01-.20	Slight
.21-.40	Fair
.41-.60	Moderate
.61-80	Substantial
.81-.99	Almost perfect

Table 4, modified from (17)

Study addition

An extra part has been added to the original plan of the study. This is a comparison with the answers from the observers and a reference standard, also using kappa statistics. The reference standard is based mainly on classification of the sounds done by an international independent expert panel, and partly by analysis performed by Melbye and Solis. 60 of the sound files (all in survey 1) were analysed by the expert panel, consisting of four lung sound researchers, two of them being medical doctors, and crackles and wheezes were regarded to be present when classified by three or more of the four experts. The experts had the same baseline for classification as the observers. The remaining 20 were classified by Melbye and Solis, based on consensus between the two. The reason why not all the sounds were analysed by the expert panel, is that some recordings had to be done after the analysing by the experts was performed.

The reference standard is interpreted as the “real truth” about the sounds, and makes the background for the prevalences of the lung sounds illustrated in table 2 and 3 above.

Results

When presenting the agreements, only calculations based on lumped categories (the totals) of inspiratory and expiratory sounds are included. For the observer agreement against reference standard and interobserver agreement, only round 1 in each survey is included. Only "mean values" of kappa is given attention. This is commented on in the "discussion" part.

S1=survey 1, S2=survey 2, R1=round 1, Diff.=difference

 # Observer 13 did not take part in survey 2.

Intraobserver agreement

Abnormal

Observer			
	S1	S2	Change
1	.41	.49	.08
2	.34	.70	.36
3	.50	.71	.21
4	.65	.62	-.03
5	.65	.74	.09
6	.22	.35	.13
7	.68	.50	-.18
8	.47	.61	.14
9	.75	.90	.15
10	.22	.30	.08
11	.33	.60	.27
12	.22	.56	.34
13	.52	#	#
14	.04	.29	.25
15	.37	.33	-.04
16	.58	.62	.04
Mean	**.43**	**.55**	**.12**

Table 5

All the observers as a total group is within moderate level of agreement in both surveys. There is a clear improvement between the surveys.

Crackles

Observer	Total		
	S1	S2	Change
1	.49	.30	-.19
2	.56	.87	.31

3	.65	.75	.10
4	.54	.53	-.01
5	.60	.56	-.04
6	.00	.32	.32
7	.43	.58	.15
8	.73	.87	.14
9	.83	.78	-.05
10	.39	.49	.10
11	.48	.60	.12
12	.52	.55	.03
13	.41	#	#
14	.53	.38	-.15
15	.53	.41	-.12
16	.65	.56	-.09
Mean	**.52**	**.57**	**.05**

Table 6

The total group is within moderate level of agreement in both surveys. There is a small improvement in their agreement between the surveys. In S2 the agreement is not far from the limit of substantial agreement.

Wheezes

Observer	Total		
	S1	S2	Change
1	.58	.66	.08
2	.65	.86	.21
3	.77	.93	.16
4	.81	.83	.02
5	.87	.92	.05
6	.47	.55	.08
7	.86	.92	.06
8	.61	.48	-.13
9	.80	.85	.05
10	.42	.63	.21
11	.73	.52	-.21
12	.56	.81	.25
13	.76	#	#
14	.55	.52	-.03
15	.65	.48	-.17
16	.81	.69	-.12
Mean	**.68**	**.71**	**.03**

Table 7

The total group is within substantial level of agreement in both surveys. There is a small improvement between the surveys.

Intraobserver agreement - intervention versus control

Abnormal

Intervention

Observer			
	S1	**S2**	**Change**
1	.41	.49	.08
3	.50	.71	.21
5	.65	.74	.09
7	.68	.50	-.18
9	.75	.90	.15
11	.33	.60	.27
13	.52	#	#
15	.37	.33	-.04
Mean	**.53**	**.61**	**.08**

Table 8

Control

Observer			
	S1	**S2**	**Change**
2	.34	.70	.36
4	.65	.62	-.03
6	.22	.35	.13
8	.47	.61	.14
10	.22	.30	.08
12	.22	.56	.34
14	.04	.29	.25
16	.58	.62	.04
Mean	**.34**	**.51**	**.17**

Table 9

The intervention group is within moderate level of agreement in S1 and just on limit of substantial level in S2. There is an improvement between the surveys. The control group is at fair level of agreement in S1, but increase their agreement to be clearly within moderate level in S2. There is no statistically significant difference in change between the groups. The mean rank is 7.1 for the intervention group, and 8.8 for the control group. The p-value is .50.

Crackles

Intervention

Observer	Total		
	S1	S2	Change
1	.49	.30	-.19
3	.65	.75	.10
5	.60	.56	-.04
7	.43	.58	.15
9	.83	.78	-.05
11	.48	.60	.12
13	.41	#	#
15	.53	.41	-.12
Mean	**.55**	**.57**	**.02**

Table 10

Control

Observer	Total		
	S1	S2	Change
2	.56	.87	.31
4	.54	.53	-.01
6	.00	.32	.32
8	.73	.87	.14
10	.39	.49	.10
12	.52	.55	.03
14	.53	.38	-.15
16	.65	.56	-.09
Mean	**.49**	**.57**	**.08**

Table 11

Both groups are within moderate level of agreement in both surveys. There is an improvement between the surveys, small in the intervention group and a little more in the control group. Both groups are close to substantial level in S2.

There is no statistically significant difference in change between the groups. The mean rank is 6.9 for the intervention group, and 8.9 for the control group. The p-value is .40.

Wheezes

Intervention

Observer	Total		
	S1	S2	Change
1	.58	.66	.08
3	.77	.93	.16
5	.87	.92	.05
7	.86	.92	.06
9	.80	.85	.05
11	.73	.52	-.21
13	.76	#	#
15	.65	.48	-.17
Mean	**.75**	**.75**	**.00**

Table 12

Control

Observer	Total		
	S1	S2	Change
2	.65	.86	.21
4	.81	.83	.02
6	.47	.55	.08
8	.61	.48	-.13
10	.42	.63	.21
12	.56	.81	.25
14	.55	.52	-.03
16	.81	.69	-.12
Mean	**.61**	**.67**	**.06**

Table 13

Both groups are within substantial level of agreement in both surveys, clearly for the intervention group. There is no improvement between the surveys for the intervention group, and a small one for the control group.

There is no statistically significant difference in change between the groups. The mean rank is 7.1 for the intervention group, and 8.8 for the control group. The p-value is .50.

Observer agreement against reference standard

Abnormal

Observer			
	S1	S2	
	R1	R1	Change
1	.32	.34	.02
2	.49	.49	.00
3	.55	.52	-.03
4	.63	.49	-.14
5	.57	.63	.06
6	.50	.41	-.09
7	.44	.63	.19
8	.73	.52	-.21
9	.29	.49	.20
10	.57	.34	-.23
11	.15	.68	.53
12	.57	.49	-.08
13	.56	#	#
14	.00	.21	.21
15	.19	.47	.28
16	.53	.45	-.08
Mean	**.44**	**.48**	**.04**

Table 14

The total group is within moderate level of agreement in both surveys. There is a small improvement between the surveys.

Crackles

Observer	Total		
	S1	S2	
	R1	R1	Change
1	.14	.25	.11
2	.70	.47	-.23
3	.54	.48	-.06
4	.59	.52	-.07
5	.60	.50	-.10
6	.25	.45	.20
7	.51	.54	.03
8	.58	.37	-.21
9	.45	.33	-.12
10	.33	.38	.05
11	.52	.50	-.02
12	.54	.38	-.16

13	.45	#	#
14	.69	.32	-.37
15	.52	.42	-.10
16	.54	.36	-.18
Mean	**.50**	**.42**	**-.08**

Table 15

The total group is within moderate level of agreement in both surveys. There is a negative alteration between the surveys.

Wheezes

Observer	**Total**		
	S1	**S2**	
	R1	**R1**	**Change**
1	.64	.76	.12
2	.62	.71	.09
3	.88	.93	.05
4	.68	.76	.08
5	.73	.69	-.04
6	.42	.68	.26
7	.81	.84	.03
8	.53	.48	-.05
9	.47	.84	.37
10	.81	.63	-.18
11	.57	.81	.24
12	.60	.78	.18
13	.70	#	#
14	.57	.63	.06
15	.57	.55	-.02
16	.75	.76	.01
Mean	**.65**	**.72**	**.07**

Table 16

The total group is within substantial level of agreement in both surveys. There is an improvement between the surveys.

Observer agreement against reference standard - intervention versus control

Abnormal

Intervention

Observer			
	S1	S2	
	R1	R1	Change
1	.32	.34	.02
3	.55	.52	-.03
5	.57	.63	.06
7	.44	.63	.19
9	.29	.49	.20
11	.15	.68	.53
13	.56	#	#
15	.19	.47	.28
Mean	**.38**	**.54**	**.16**

Table 17

Control

Observer			
	S1	S2	
	R1	R1	Change
2	.49	.49	.00
4	.63	.49	-.14
6	.50	.41	-.09
8	.73	.52	-.21
10	.57	.34	-.23
12	.57	.49	-.08
14	.00	.21	.21
16	.53	.45	-.08
Mean	**.50**	**.43**	**-.07**

Table 18

The intervention group is in fair level of agreement in S1, but improve to be clearly within moderate level in S2. The control group is in moderate level of agreement in both surveys, but shows a little negative alteration between the surveys.

The change in the agreement between the groups is statistically significant, favouring the intervention group. The mean rank is 11.1 for the intervention group, and 5.3 for the control group. The p-value is .008.

Crackles

Intervention

Observer	Total		
	S1	S2	
	R1	R1	Change
1	.14	.25	.11
3	.54	.48	-.06
5	.60	.50	-.10
7	.51	.54	.03
9	.45	.33	-.12
11	.52	.50	-.02
13	.45	#	#
15	.52	.42	-.10
Mean	**.47**	**.43**	**-.04**

Table 19

Control

Observer	Total		
	S1	S2	
	R1	R1	Change
2	.70	.47	-.23
4	.59	.52	-.07
6	.25	.45	.20
8	.58	.37	-.21
10	.33	.38	.05
12	.54	.38	-.16
14	.69	.32	-.37
16	.54	.36	-.18
Mean	**.53**	**.41**	**-.12**

Table 20

Both groups are within moderate level of agreement in both surveys. Both groups have a negative alteration between the surveys, most prominent in the control group.

There is no statistically significant difference in change between the groups. The mean rank is 9.7 for the intervention group, and 6.5 for the control group. The p-value is .20.

Wheezes

Intervention

Observer	Total		
	S1	S2	
	R1	R1	Change
1	.64	.76	.12
3	.88	.93	.05
5	.73	.69	-.04
7	.81	.84	.03
9	.47	.84	.37
11	.57	.81	.24
13	.70	#	#
15	.57	.55	-.02
Mean	**.67**	**.77**	**.10**

Table 21

Control

Observer	Total		
	S1	S2	
	R1	R1	Change
2	.62	.71	.09
4	.68	.76	.08
6	.42	.68	.26
8	.53	.48	-.05
10	.81	.63	-.18
12	.60	.78	.18
14	.57	.63	.06
16	.75	.76	.01
Mean	**.62**	**.68**	**.06**

Table 22

Both groups are within substantial level of agreement in both surveys. The intervention group is in the upper level in S2. Both groups have an improvement between the surveys, with the intervention group a little better than the control group.

There is no statistically significant difference in change between the groups. The mean rank is 8.4 for the intervention group, and 7.6 for the control group. The p-value is .80.

Interobserver agreement

S1=Survey 1, R1=Round 1, CI=Confidence Interval, General=all observers together.

General

Lung sound	S1R1	95% CI		S2R1	95% CI		Change
Abnormal	**.37**	.26	.48	**.34**	.24	.44	**-.03**
CrTot	**.39**	.27	.51	**.34**	.22	.46	**-.05**
WhTot	**.61**	.47	.76	**.57**	.39	.76	**-.04**

Table 23

The agreement among the observers in the total group regarding abnormal is in the fair level, the same is observed for the total crackle. The total wheeze is just at limit of the substantial level in S1, and in the upper moderate level in S2. All the categories are showing a little negative alteration.

Intervention

Lung Sound	S1R1	95% CI		S2R1	95% CI		Change
Abnormal	**.37**	.24	.51	**.33**	.21	.45	**-.04**
CrTot	**.38**	.22	.54	**.28**	.16	.40	**-.10**
WhTot	**.67**	.50	.84	**.57**	.39	.74	**-.10**

Table 24

The agreement among the observers regarding abnormal is in the fair level. The same is observed for the total crackle. The total wheeze draws the same picture as the for the general group, except a little higher agreement in S1. The alteration is negative in all categories, minor for abnormal and a little more prominent for the two other categories.

Control

Lung Sound	S1R1	95% CI		S2R1	95% CI		Change
Abnormal	**.38**	.25	.52	**.34**	.21	.47	**-.04**
CrTot	**.39**	.26	.53	**.43**	.26	.59	**.04**
WhTot	**.53**	.35	.71	**.59**	.37	.81	**.06**

Table 25

The agreement among the observers regarding abnormal is in the fair level. The same is observed for the total crackle, except S2 which is in the moderate level. The total wheeze is in the moderate level. The alteration is negative for abnormal, and a little positive for the other two categories.

Discussion

Methodology

After the surveys the observers reported things that could have influenced the study in different ways. These are considered to have had minor impact, but have to be mentioned: Since the sounds were played three times each, and some observers used their own words to describe, tapping on the keyboard became a source of disturbance for some observers listening for the second and third time. Some reported that since the situation was new; it took a little while to feel comfortable in R1 of S1. The computer playing the sounds had to be restarted in the middle of S2, this lead to a 2 min unexpected break. Observer 13 did not show up for S2, and this left the intervention group with 7 instead of 8 observers.

The visualisation of the lung sounds using a spectrogram is questionable, since the effect of this is not known. Did it work as a remedy or was it just a source of confusion, especially in S1 when we did not explain it to the observers. Some observers indeed reported that they recognized the pattern at the spectrogram, and that way potentially took advantage of this also in S1.

When playing the sounds we used loudspeakers instead of headset, this makes the listening more different than necessary from the regular way of doing auscultation in a clinical situation. On the other hand, using taped sounds, secured that all the observers heard the same phenomenon. In a live setting with real patients it would not be possible for all the observers to listen at the same time, and this will create a source of bias. Time will alter a sound phenomenon since the patient is a dynamic source, a possible change may be disappearing of a wheeze when the patient has coughed.

A factor regarding the sound material is that the sounds could be more or less challenging in S2 compared to S1. This is difficult to measure, and has therefore not been given attention. The prevalence of sound phenomenons is anyway very similar in the two surveys, and this source of variation is of little importance when focusing on differences between groups.

All statistical methods have their weaknesses. In kappa statistics the interpretation of the values is a discussed problem. What level of agreement can be accepted? This debate has

led to the suggestion that kappa statistics should not be used alone in medical research measuring agreement (16). In this study .41 has been used as the lower limit of acceptable agreement. Recent literature suggest a more strict way of interpreting the values, cf. the article by L.McHugh (2012) (16). According to this source, higher level of agreement is important when a study has a potential for changing healthcare practice. That is not the goal in this study, and therefore it has been considered acceptable to use .41. Other recent medical studies, for example a radiology study exploring agreement among radiologists, by Timmers, Doorne-Nagtegaal and Verbeek et al 2011 also use .41 when interpreting the values (18).

Fleiss kappa was selected for testing interobserver agreement. It is mentionable that the study first included Intraclass Correlation Coefficient (ICC) calculations. The method was changed because it is a strength to keep within one statistical method throughout the study. The results indicated a difference of about 10%, with lower values using Fleiss Kappa.

It is questionable that some of the lung sounds were analysed by Melbye and Solis, since they were involved in the recording of the sounds and the planning of the study. It must be emphasized that their classifications were done before the data from the observers was looked at. This secured that the answers given by the observers did not affect them. Only 20 of the 80 sounds were analysed by Melbye and Solis, so the influence on the results, if any, has probably been minor.

Results

The study created a massive amount of results. Working with the results revealed a part of the task that became very complex and challenging regarding the ability getting a satisfying overview. Therefore it was decided to use the lumped (totals) categories of wheezes and crackles and to compare only R1 in each survey, even though interesting information is not to being published.

Only the mean values of the agreements are commented on. Large variations between observers could have been given attention, but is leaved out to limit the extent of the task.

General

The prevalence of crackles in the sound material is significantly higher than of the wheezes. This, together with the wheezes being a more pronounced phenomenon, could explain the significantly higher level of agreement observed for the wheezes throughout the study. The wheezes, with its musical appearance, is a characteristic sound that is probably more easy to be aware of. This is also shown in other similar studies, where the wheezes are recognised more accurate than for example the crackles (19, 20).

An other important factor is the presence of background noise in the sound material. Examples of this is chest hair rubbing against the diaphragm, heart sounds and bowel sounds. Sound phenomenons like this are parts of ordinary chest auscultations, but could distract the observers in a setting where they don't handle the stethoscope by themselves.

Intraobserver agreement

Crackles

Taken into consideration the crackle being a more challenging phenomenon to recognise, the agreement for the observers as a total group is relatively strong compared to other similar studies (19, 20). The agreement in both surveys was above .50, and in addition, improve with .05 from the relatively strong baseline.

Wheezes

Strong baseline, with .68. The improvement to .71 in S2, is minimal, but improvement from a strong baseline is a positive sign.

Intraobserver agreement - Intervention versus control

Abnormal

The improvement is better in the control group, but they had a poorer baseline compared to the intervention group, this makes them potentially more receptive for natural improvement, regression to the mean (15). The intervention group has nevertheless a better agreement than the control group in both surveys, clearly within acceptable level.

Crackles

A little poorer baseline for the control group, showing a little better improvement. The agreement is clearly within acceptable level in both groups in both surveys.

Wheezes

Strong agreement in both groups, especially in the intervention group, not far from being defined as “almost perfect”. Again, a poorer baseline for the control group which can explain their improvement compared to no improvement for the intervention group.

Observer agreement against reference standard

Abnormal

Acceptable agreement in both surveys.

Crackles

Acceptable agreement in both surveys, but the development is negative with -.05.
This can be explained by variation by chance.

Wheezes

Relatively strong agreement, especially in S2, with >.70. Improvement of .07 from a strong baseline is again a positive sign, but can also be explained by variation by chance.

Observer agreement against reference standard - intervention versus control

Abnormal

The intervention group is below acceptable level of agreement in S1, but improvement of .16 lead to clearly acceptable agreement in S2. The control group is above acceptable level of agreement in both surveys, but shows a negative development between the surveys of -.07. The improvement in the intervention group is significantly better than for the control group with a p-value <0.05. The baseline for the intervention groups is clearly poorer than for the control group, with .12, but the agreement in S2 is better than for the control group in S1.

Crackles

Both groups are within acceptable level of agreement in both surveys, and both show negative development between the surveys, most prominent in the control group, with -.12 compared to -.04. The baseline is a little stronger in the control group with .06.

Wheezes

The agreement in the intervention group is strong, especially in S2, being close to “almost perfect”. Both surveys shows stronger agreement than the control group. It must be emphasized that also the control group show relatively strong agreement. In addition of having a stronger baseline the intervention group also improve more than the control group

between the surveys, with .10 compared to .06. This is a minor difference, but is interesting since the baseline is stronger.

Interobserver agreement

The total group have a multirater agreement that is below acceptable level for abnormal and crackles, and clearly within acceptable level for wheezes. The development is also negative, which counts for all three categories. The intervention group draw a similar picture as the total group, but having a little stronger baseline for wheezes and a poorer agreement in S2 for crackles, with only .28. The development is also more negative. With -.10 for crackles and wheezes. The control group draw an opposite picture having an improvement between the surveys, except for abnormal that is similar with the intervention group. It must be emphasized that the intervention group had a significant higher baseline with .13 for wheezes, compared to the control group.

The control group showing improvement compared to negative development in the intervention group, is an interesting observation since the control group did not have any training. The picture is difficult to explain but may indicate that improvement is unevenly distributed among the group members.

Conclusion

Regarding the intraobserver agreement the observers as a total group shows an overall agreement clearly within acceptable level. The category with the highest agreement is wheezes, which is expected. It is an exclusively positive improvement between the surveys for all categories.

When looking at the intraobserver agreement with the observers in separate groups, the same pattern is observed. The highest agreement is observed for the intervention group in the wheeze category. The poorest agreement, which also is below acceptable level (kappa value of .34) is observed for the control group in the abnormal category. Also here the improvement is exclusively positive between the surveys in both groups. The best improvement is observed for the control group, but they had poorer baseline from S1 than the intervention group. No statistically significant difference was discovered. The intervention group overall clearly shows the highest agreement.

Regarding the agreement with the reference standard the total group shows an overall

agreement that also is clearly within acceptable level. Only crackles category shows negative development between the surveys, the other two is positive. Also here wheezes show the highest agreement, with the crackles and the abnormal overall being relatively similar.
The agreement in the separate groups is generally within acceptable level of agreement. The only one being below is abnormal category in S1 in the intervention group (kappa value of .38). For this category the improvement in the intervention group between the surveys compared to the control group is statistically significant. The wheezes show the highest agreement and the crackles show negative development in both groups, most prominent in the control group.

For the interobserver agreement the total group show agreement below acceptable level, except for the wheezes category. The intervention group has negative development between the surveys, while the control group has positive development, except the abnormal category. The negative development in the intervention group is assumed to be explained in normal variation, but it is also possible that the course made them agree less with each other.

The results indicate that sixth year medical students at the University of Tromsø, the Arctic University of Norway have highly acceptable intraobserver agreement, and the agreement tended to improve in both the intervention and the control group. The agreement with the reference standard was also highly acceptable for the category wheezes and acceptable for crackles and the abnormal category. A tendency to positive change in the intervention group when compared to the control group was found, but the difference was only statistically significant for the abnormal category in the agreement against reference standard. The interobserver agreement did not reach the limit of acceptable, except for wheezes. Summarized, a weak effect of the intervention was observed

Further investigation should be tried without the use of spectrograms. A method that probably work out better is the use headset to avoid disturbing background noise. The training should probably be different, for example divided in several and shorter sessions with spiral teaching, and include repeated rating.

1. Roguin A. Rene Theophile Hyacinthe Laënnec (1781-1826): The Man Behind the Stethoscope. Clinical Medicine & Research. 2006;4(3):230-5.
2. Robert L. Wilkins JRD, Raymond L. H. Murphy et al. Lung Sound Nomenclature Survey. Chest. 1990;98(4):886-9.
3. Nick A. Francis HM, Mark J. Kelly 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. European Journal of General Practice. 2013;19(2):77-84.
4. Cynthia D.Mulrow BLD, Elizabeth R.Delong et al. Observer Variability in the Pulmonary Examination. Journal of General Internal Medicine. 1986;1(Nov/Dec):364-7.
5. Claude Lenfant NKRPea, editor Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease: NHLBI/WHO Workshop Report1998: NATIONAL INSTITUTES OF HEALTH. National Heart, Lung and Blood Institute.; 2001.
6. Abraham Bohadana GI, Steve S. Kraman. Fundamentals of Lung Auscultation. The New England Journal of Medicine. 2014;370(8):744-51.
7. Malay Sarkar IM, Narasimhalu Niranjan et al. . Auscultation of the respiratory system. Annals of Thoracic Medicine. 2015;10(3):158-68.
8. Melbye H. Auscultation of the lungs - still a useful examination? Tidsskiftet for Den norske legeforening. 2001;121(4):451-4.
9. J. Macfarlane SAL, R. Macfarlane et al. . Contemporary use of antibiotics in 1089 adults presenting with acute lower respiratory tract illness in general practice in the U.K.: implications for developing management guidelines. Respiratory Medicine 1997;91(7):427-34.
10. RM Hopstaken CB, JW Muris et al. . Do clinical findings in lower respiratory tract infection help general practitioners prescribe antibiotics appropriately? An observational cohort study in general practice. Family Practice 2006;23(2):180-7.
11. Sahgal N. Monitoring and analysis of lung sounds remotely. International Journal of Chronic Obstructive Pulmonary Disease 2011;6:407-12.
12. Zhang Kexin WX, Han Fangfang et al. The detection of crackles based on mathematical morphology in spectrogram analysis. Technology and Health Care 2015;23:489-94.
13. 3M. 3M Littmann StethAssist Heart and Lung Sound Visualization Software User Manual - Version 3.
14. Hasse Melbye LG-M, Mark Everard et al. . Wheezes, crackles, rhonci: Agreement among members of the ERS task force on lung sounds. European Respiratory Journal. 2014;44(58).
15. Sterne BRKaJAC. Essential Medical Statistics. 2nd ed: Blackwell Science Ltd; 2003.
16. L.McHugh M. Interrater reliability: the kappa statistics. Biochem Med (Zagreb). 2012;22(3):276-82.
17. Garrett AJVaJM. Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine. 2005;37(5):360-3.
18. J.M.H Timmers HJvD-N, A.L.M. Verbeek et al. . A dedicated BI-RADS training programme: Effect on the inter-observer variation among screening radiologists. European Journal of Radiology. 2011;81(2012):2184-8.
19. Thomas DBaJ. Interrater Reliability of Auscultation of Breath Sounds Among Physical Therapists. Journal of The American Physical Therapy Association. 1995;75(12):1082-8.
20. al. SATWSJe. Accuracy and reliability of physiotherapists in the interpretation of tape-recorded lung sounds. Australian Physiotherapy 1995;41(3):179-84.