Observer Agreement, Chest Auscultation, and Crackles in Asbestos-Exposed Workers*

Peter Workum, M.D.; Elizabeth A. DelBono, M.P.H.; Stephen K. Holford, Sc.D.; and Raymond L. H. Murphy, Jr., M.D., F.C.C.P.

Investigators cite observer variability as a problem in using crackles to diagnose asbestosis. We measured agreement on the presence or absence of crackles noted during auscultation of 64 asbestos-exposed workers in order to clarify this question. There was 89 percent agreement between two observers who simultaneously examined subjects breathing from functional residual capacity (FRC). Kappa (κ), a statistic accounting for chance agreement, was 0.73. Unanimous agreement between four observers who listened to tape recordings of the breath sounds was 81 percent (κ = 0.69). When the subjects breathed from residual volume (RV) there was 78 percent (κ = 0.53) and 67 percent (κ = 0.60) agreement, respectively. Comparing direct to tape-playback auscultation, there was 90 percent (κ = 0.77) and 84 percent (κ = 0.58) intraobserver agreement when the subjects breathed from FRC and 90 percent (κ = 0.79) and 75 percent (κ = 0.39) when they breathed from RV. We conclude that observer variability is sufficiently low to allow trained observers to monitor asbestos-exposed workers for crackles directly and during tape-playback auscultation.

Crackles may be associated with a variety of pulmonary diseases1-4 and are included as one of the criteria for diagnosing asbestosis.5 They are also useful in screening industrial populations exposed to asbestos;6 however, some investigators have shown that physicians may not agree on common pulmonary signs,7-11 and Elmes12 specifically questioned the use of crackles to diagnose asbestosis because of the possibility of observer variability. We addressed this problem by measuring the agreement between trained observers on the presence or absence of crackles noted during direct and tape-playback auscultation of 64 asbestos-exposed workers. Additionally, we compared the results of tape-playback to direct auscultation to determine if recordings made by a technician during an industrial survey can be reliably used to monitor workers for crackles.

Materials and Methods

During an industrial survey a pulmonary physician using an acoustic stethoscope and a trained pulmonary technician using an electronic stethoscope performed chest auscultation on 64 asbestos-exposed workers (direct auscultation). The electronic stethoscope was a Kent Cambridge Leatham-type stethoscope microphone coupled to the chest wall by an air chamber, and the microphone's signal was preamplified and recorded using an audiotape recorder with a tape speed of 7½ in/sec. While straddling a chair, the subjects breathed through their mouths at tidal volumes followed by slow deep breaths from functional residual capacity (FRC) (tidal-slow deep breathing). They then breathed deeply and rapidly after expiring to residual volume (RV) (rapid deep breathing). They did not cough prior to or during the examination. The two on-site observers listened simultaneously while recording the breath sounds at ten basilar sites (right and left paravertebral, midcapular, posterior and anterior axillary, and midclavicular lines). They did not auscultate the same site at the same time. In an attempt to ensure the independence of each examiner's evaluation, they shielded the paper on which they took notes and did not discuss the results of their findings.

Tape recordings of the breath sounds were made at the time of direct auscultation. These were subsequently listened to and the presence or absence of crackles noted by the same two on-site observers approximately six weeks after the initial examination and by two physicians to whom copies of the tapes were mailed (tape-playback auscultation).

We considered that crackles were present if they were noted at two or more of ten sites. Interobserver agreement on the presence or absence of crackles was calculated between the physician and technician, first during direct auscultation and second during tape-playback auscultation. Unanimous (all four agree) and majority (three or more agree) interobserver agreement was calculated between the physician, technician, and the two physicians to whom copies of the tape recordings were mailed. Intraobserver agreement was measured both for the physician and technician, comparing the results of direct to tape-playback auscultation. Kappa (κ), a statistic accounting for chance agreement, was calculated by the following formula: \( \kappa = \frac{Po-Pc}{1-Pc} \), where Po is observed agreement and Pc is chance agreement.

Results

The subjects ranged in age from 20 to 71 years (mean, 43 years) and were exposed to asbestos from 0 to 41 years (mean, 12 years). Three of the subjects had asbestosis according to the criteria of Murphy et al.5 Crackles were noted by both on-site examiners in 15 subjects during the tidal-slow deep breathing maneuver and in 14 persons during rapid deep breathing. At direct auscultation, there was 89 percent agreement (κ = 0.73) between the physician and technician during


Manuscript received April 3; revision accepted June 24.

Reprint requests: Dr. Workum, Leominster Hospital, Leominster, MA 01523
the tidal-slow deep breathing maneuver and 81 percent agreement (κ = 0.57) during rapid deep breathing. When these same two observers listened to tape recordings, there was 89 percent (κ = 0.69) and 78 percent (κ = 0.53) agreement, respectively. When comparing the four observers who listened to the tape recordings, there was 92 percent majority agreement (three or more agree) (κ = 0.53) during the tidal-slow deep breathing maneuver and 84 percent agreement (κ = 0.48) during rapid deep breathing. There was 81 percent unanimous agreement (all four agree) (κ = 0.69) during tidal-slow deep breathing and 67 percent agreement (κ = 0.60) during rapid deep breathing. The intraobserver agreement for the technician was 90 percent (κ = 0.77 and 0.79) for both the tidal-slow deep breathing and rapid deep breathing maneuvers, while for the physician, it was 84 percent (κ = 0.58) and 75 percent (κ = 0.39), respectively.

**Discussion**

In an excellent review on the reliability of clinical information, Koran showed that historical, physical, and laboratory data are all subject to error. Schilling et al showed that there was "fair" agreement between two observers in diagnosing byssinosis by history, 24 percent disagreement on the detection of "abnormal chest sounds," and some of the clinical data were too unreliable for use in the diagnosis of byssinosis. Fletcher reviewed the literature on observer variation in the initial diagnosis of chest diseases and pointed out that there may be marked variation in the history, examination, and roentgenographic interpretation of patients with pulmonary disease. Smylie et al showed that the agreement between nine physicians on 20 respiratory signs in 20 patients fell on the average between chance and 100 percent agreement. Hudson et al reported that for three observers who examined 100 consecutive patients with rales, there was only 47 percent agreement on the qualitative adjectives fine, medium or coarse, and moist or dry.

Our agreement on the presence or absence of crackles is better than some of the previously reported results of observer variability for a number of reasons. First, it has been shown that observers trained for a particular task will agree more often than those without training. The physicians were trained in pulmonary medicine, and the technician had worked in our laboratory for five years, listened to numerous recordings of breath sounds, and had the opportunity to compare the results of her auscultation with pulmonary physicians. Secondly, the more diagnostic categories considered during the examination, the poorer the agreement, and in this study the observers considered only one sign, the presence or absence of crackles. Thirdly, agreement on normality is usually greater than agreement on abnormality, and the majority of the subjects examined did not have crackles. Fourthly, agreement on the presence or absence of a sign will usually be higher than agreement on qualitative variabilities. Fifthly, observer error caused by different methods and timing of examination was avoided by having the two on-site examiners listen to the subjects simultaneously and having the four tape-playback examiners listen to a copy of or the original tape recording. Lastly, during direct auscultation the patients were examined carefully in a quiet setting, while during tape playback, the recordings were listened to at the observer's convenience in a controlled environment.

The agreement during rapid deep breathing was not as good as during tidal-slow deep breathing, probably because of the accentuation of the normal vesicular breath sounds and the movement of the stethoscope over the skin and hairs during rapid inspiration, making it more difficult to distinguish artifact from crackles. On subsequent surveys, we have had the subjects inspire slowly after expiring to RV, which appears to make the detection of crackles easier.

The intraobserver agreement is likely better for the technician than the physician because the technician listened with the electronic stethoscope during direct auscultation and heard the same signal during tape playback, while the physician listened with an acoustic stethoscope during direct auscultation and heard an electronic signal during the tape playback. Additionally, the technician has had more experience listening to tape playback than the physician.

We measured the overall observer agreement, which is the percentage of patients about whom the observers agreed on the presence or absence of crackles. Koran and Spitzer and Fleiss have pointed out that one of the drawbacks of the measurement of overall agreement is that it does not account for chance agreement and that the agreement between two equally skilled observers depends in part upon the percentage of cases each observer considers normal. The statistic, kappa, has been developed to take account for the chance agreement between two or more observers. It can be thought of as observed agreement not accounted for by chance divided by best possible agreement not accounted for by chance. It varies from negative values for agreement less than chance, to zero for chance agreement, and to +1.0 for perfect agreement. The advantage of kappa is that it improves the comparability of results between studies by accounting for chance agreement. It also may help to verbally describe the relative strength of agreement between observers which, in our study, was moderate or better in all cases but one. Because a technician tape-recording breath sounds during an industrial survey can collect a large amount of data which can be subsequently analyzed in a
laboratory or hospital by a specialist, we wanted to determine if information obtained in this way can be used reliably to follow asbestos-exposed subjects. The results of interobserver and intraobserver agreement show that the presence or absence of crackles noted during tape-playback auscultation is a reproducible sign, suggesting that tape recordings of breath sounds, when they are recorded carefully and listened to by trained observers, may be one way to monitor asbestos-exposed subjects for crackles; the observations made during tidal-slow deep breathing may be less variable than during rapid deep breathing.

The results of interobserver and intraobserver agreement noted in this study show that the presence or absence of crackles is as reproducible a physical sign as many of the other historic, physical, and laboratory data commonly used to diagnose medical diseases. We conclude, therefore, that it is acceptable for trained observers in a controlled setting to monitor asbestos-exposed workers for the presence or absence of crackles either during direct or tape-playback auscultation.

ACKNOWLEDGMENT: We thank Dr. Edward Gaensler, Thoracic Services, Boston University School of Medicine, for allowing us to participate in an industrial survey. Drs. Fumio Shirai and Kazuo Sada, Nara Medical University, Nara, Japan, who listened to the tape recordings; and Ms. Raffaela Murphy and Ms. Pat Gagne for secretarial help in preparing the manuscript.

REFERENCES
26. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975; 31:651-59
27. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74

CHEST / 89 / 1 / JANUARY, 1986 29