Pediatr Radiol (1999) 29: 459–462

Frank H. Bloomfield
Rita L. Teele
Maurice Voss
David B. Knight
Jane E. Harding

Inter- and intra-observer variability in the assessment of atelectasis and consolidation in neonatal chest radiographs

Received: 20 July 1998
Accepted: 30 November 1998

F. H. Bloomfield · D. B. Knight · J. E. Harding
Department of Paediatrics,
National Women's Hospital, Claude Road,
Epsom, Auckland, New Zealand

F. H. Bloomfield (✉)
Department of Paediatrics,
Private Bag 92019, University of Auckland,
Auckland, New Zealand

R. L. Teele · M. Voss
Department of Imaging,
National Women's Hospital, Epsom,
Auckland, New Zealand

Abstract *Background.* Radiology is an essential part of neonatal intensive care. Interpretation of chest radiographs frequently contributes to respiratory management of neonates, but there has been little assessment of the consistency of this interpretation.
Objective. To assess the inter- and intra-observer variability for the reporting of atelectasis and/or consolidation in neonatal chest radiographs.
Materials and methods. A total of 585 chest radiographs from the 220 babies ventilated in our nursery over a 2-year period were coded by two radiologists for generalised, lobar and segmental atelectasis and/or consolidation. Two months later one of the radiologists re-coded a random sample of these films (n = 117, 20 %). Agreement was assessed by the kappa statistic and by proportions of agreement for normality and abnormality.
Results. The reported incidence of focal atelectasis was low (5–6 %). Focal changes of any nature were found in 21–26 % of films. Inter-observer agreement was fair to moderate (kappa = 0.25–0.44). Intra-observer agreement was mostly moderate to good (kappa = 0.38–0.66).
Conclusion. The poor inter-observer agreement for the diagnosis of pulmonary parenchymal abnormalities on chest radiographs of neonates receiving intensive care suggests that abnormalities should be described rather than diagnoses given or that a list of differential diagnoses be offered. When research involves radiographic interpretation, the potential lack of consistency in reporting abnormalities must be borne in mind.

Introduction

A neonate requiring intensive care is at risk of developing pulmonary atelectasis and consolidation. Atelectasis is a complication of extubation, occurring in 10–50 % of cases [1, 2]. The diagnosis of atelectasis or consolidation, which is made radiographically, may result in a change in management such as the institution of chest physiotherapy [3, 4]. A recent report of inter-observer variability in the assessment of chest radiographs in neonatal chronic lung disease found that there was considerable variation among radiologists [5]. However, to our knowledge, the inter- and intra-observer variability for the assessment of atelectasis and/or consolidation in neonatal chest radiographs has not been reported previously.

We have recently conducted a retrospective review of all post-extubation radiographs taken in our unit over a 2-year period to assess the efficacy of chest physiotherapy in preventing post-extubation atelectasis [6]. We found that there was no difference in the incidence of post-extubation atelectasis among babies who received physiotherapy and those who did not, or between pre- and post-extubation films. The films had been reviewed by two radiologists and the results were the same regardless of which radiologist's data were used.

Table 1 Incidences of abnormality for each radiologist. Values are percent ($n = 585$). Focal changes include lobar and segmental abnormalities

	Atelectasis		Consolidation		Volume loss	
	Any	Focal	Any	Focal	Any	Focal
Radiologist 1 (RLT)	18.1	5.8	37.9	18.6	43.2	21.5
Radiologist 2 (MV)	9.4	5.0	36.4	23.1	41.4	26.2

Only one data set was used in the previous publication [6]. However, we wished to look at the level of agreement between the two radiologists more closely. We therefore assessed the inter-observer variability of the two radiologists and the intra-observer variability of one of the radiologists who reviewed a subset of the films.

Materials and methods

Two radiologists – a registrar and a consultant – coded 585 AP chest radiographs from the 220 babies who were ventilated on our unit over a period of 2 years (1993–1995). We identified post-extubation films by hand, searching the radiology packets of every baby admitted to our intensive care unit. During this period, babies routinely had a chest radiograph taken 4 h after extubation. We also selected the film immediately preceding the post-extubation film, which was mostly taken in the preceding 24 h. If a baby was extubated on more than one occasion, each episode was eligible. During the first year, all babies received peri-extubation physiotherapy; during the second year, no baby did. During the study period, babies suspected of having hyaline membrane disease routinely received surfactant. Most babies (and all very low birth weight babies) received continuous positive airway pressure immediately after extubation.

We obscured all identification on the films. The radiologists coded the films randomly and independently in terms of generalised, lobar or segmental atelectasis and/or consolidation. Atelectasis was defined as areas of volume loss, consolidation as opacity of the air spaces. During the course of this study, it became clear that in this neonatal population it was often difficult to differentiate atelectasis from consolidation. Therefore, during the analysis, atelectasis and consolidation were also combined as 'volume loss' to determine whether this led to an improvement in the inter-observer variability (see "Results").

Radiographers attempted to take all chest radiographs during the inspiratory phase and any loss of volume on the chest films was coded as generalised atelectasis rather than attributing lungs of small volume to an expiratory film: this was to avoid missing any cases of atelectasis. After a delay of 2 months, one of the radiologists (R.L.T.) coded a randomly selected subset of films a second time to assess intra-observer variability.

We have assessed inter- and intra-observer variability by the kappa statistic and by proportions of agreement for normality and abnormality [7]. Kappa < 0.2 is generally accepted to represent poor agreement, 0.21–0.4 fair agreement, 0.41–0.6 moderate agreement and 0.61–0.8 good agreement. Kappa > 0.8 represents excellent agreement [8]. We looked for observer bias with McNemar's test [7].

Results

A total of 585 films were reviewed from the 220 babies ventilated on our unit during 1994 and 1995. The median (range) birth weight was 1228 (510–4595) g and the median (range) gestational age was 29 (24–42) weeks. There were 297 post-extubation films and 288 pre-extubation films. All 585 films were treated as one group, since there were no differences in the incidence of atelectasis or consolidation between the pre-extubation and the post-extubation films, nor between babies who received peri-extubation physiotherapy and those who did not [6]. Focal atelectasis (lobar or seg-

Table 2 The P value for kappa refers to the probability that agreement is different from that expected by chance. *** $P < 0.001$, NS not significant, P_{abn} proportion of agreement for abnormality, P_{norm} proportion of agreement for normality

	Inter-observer variability ($n = 585$)				Intra-observer variability ($n = 117$)		
	kappa	Observer bias *(P)*	P_{abn} (95% CI)	P_{norm} (95% CI)	kappa	P_{abn} (95% CI)	P_{norm} (95% CI)
Any atelectasis	0.27***	< 0.001	0.22 (0.15–0.29)	0.81 (0.78–0.85)	0.38***	0.31 (0.14–0.48)	0.81 (0.74–0.89)
Any consolidation	0.42***	NS	0.46 (0.41–0.52)	0.64 (0.60–0.69)	0.66***	0.63 (0.49–0.76)	0.79 (0.71–0.88)
Any volume loss	0.44***	NS	0.49 (0.44–0.55)	0.63 (0.59–0.68)	0.65***	0.63 (0.49–0.76)	0.78 (0.69–0.87)
Focal atelectasis	0.35***	NS	0.24 (0.12–0.35)	0.93 (0.91–0.95)	0.43***	0.30 (0.02–0.58)	0.94 (0.89–0.98)
Focal consolidation	0.25***	< 0.05	0.25 (0.19–0.31)	0.73 (0.69–0.77)	0.54***	0.43 (0.23–0.64)	0.88 (0.82–0.94)
Focal volume loss	0.28***	< 0.05	0.29 (0.23–0.35)	0.71 (0.67–0.75)	0.54***	0.45 (0.27–0.63)	0.85 (0.78–0.92)

Table 3 Inter-observer variability for each individual lobe. (** $P < 0.01$, *** $P < 0.001$, P_{abn} and P_{norm} proportion of agreement for abnormality and normality, respectively)

	Unweighted kappa	Weighted kappa	P_{abn} (95 % CI)	P_{norm} (95 % CI)
Atelectasis				
Right upper lobe	0.36***	0.40***	0.27 (0.09–0.46)	0.96 (0.96–0.98)
Right middle lobe	0.25***	0.24***	0.33 (– 0.20–0.87)	0.99 (0.98–1.00)
Right lower lobe	0.25***	0.25***	0.20 (– 0.15–0.55)	0.99 (0.98–1.00)
Left upper lobe	0.11**	0.21***	0.00 (0.00–0.00)	0.98 (0.97–0.99)
Lingula	0.00	0.00	0.00 (0.00–0.00)	0.99 (0.99–1.00)
Left lower lobe	0.19***	0.19***	0.20 (– 0.15–0.55)	0.99 (0.98–1.00)
Consolidation				
Right upper lobe	0.38***	0.42***	0.47 (0.33–0.61)	0.89 (0.87–0.92)
Right middle lobe	0.27***	0.28***	0.36 (0.16–0.56)	0.93 (0.91–0.95)
Right lower lobe	0.18***	0.18***	0.23 (0.12–0.34)	0.86 (0.83–0.89)
Left upper lobe	0.32***	0.32***	0.67 (0.49–0.84)	0.89 (0.86–0.92)
Lingula	0.11***	0.13***	0.13 (– 0.10–0.35)	0.96 (0.95–0.98)
Left lower lobe	0.15***	0.15***	0.39 (0.23–0.56)	0.84 (0.81–0.87)

mental abnormalities) was reported in approximately 5 %, consolidation in 37 % and volume loss in 42 % (Table 1). Agreement between the two observers for generalised or focal atelectasis and consolidation using the kappa statistic was poor (Table 2), a kappa > 0.4 being generally accepted to represent moderate agreement [8]. Closer examination of the raw data suggested that there were consistent differences between radiologists in the classification of parenchymal abnormalities. This suspicion was supported by the finding of significant observer bias and by the very poor proportions of agreement for abnormality, but good proportions of agreement for normality (Table 2). We therefore recalculated kappa after combining the findings of atelectasis and consolidation as volume loss. This led to an improvement in kappa to a moderate level of agreement (kappa = 0.44), with bias no longer significant. The proportion of agreement for abnormality also improved, although it was still less than 0.5 (Table 2).

We also analysed the findings in each lobe separately. As there were different grades of abnormality (normal, segmental abnormality and lobar abnormality) we calculated both unweighted and weighted kappa [9], with partial agreements between segmental and lobar abnormalities given a weighting of 0.75 and partial agreements between normal and segmental abnormalities given a weighting of 0.25. As expected, the weighted kappas were better than the unweighted scores; however, overall agreements were poor (Table 3). Interestingly, the best agreements were for the right upper lobe. The proportion of agreement for absence of abnormality was again excellent (0.84–0.99), but with very poor agreement on the classification of abnormality present (Table 3).

The results for intra-observer variability were better, although still only showing moderate agreement for the finding of atelectasis (Table 2).

Discussion

We have assessed the inter- and intra-observer variabilities for the finding of pulmonary parenchymal abnormalities (atelectasis and/or consolidation) in a large number of neonatal radiographs. The incidence of focal atelectasis (5–6 %) is lower than reported elsewhere [1, 2], although focal abnormality of some description was found in over 20 % of films.

The kappa statistic compares the observed and expected amounts of agreement. The expected amount of agreement represents that due to chance and is depen-

dent on the prevalence of the attribute being measured. The inter-observer kappa values reported here represent, at best, moderate agreement, but are not dissimilar to those described elsewhere for radiographic interpretation [10–13] and are substantially better than those described in a recent report of the interpretation of chest radiographs in neonatal chronic lung disease [5]. The highly significant *P* values for kappa that we report demonstrate that the agreement found is significantly different from that expected by chance. They do not indicate the level of agreement. In fact, kappa does not measure agreement, but association of assessments similar to the way that the correlation coefficient indicates associations for continuous variables.

A better way to measure the inter-observer variation of categorical variables is to calculate 95% limits of agreement, considering agreement for normal and abnormal assessments separately [12]. These are termed the proportions of agreement for normality and abnormality and are presented here along with the kappa statistic.

The relatively poor inter-observer kappa values we found may be related to differing levels of experience of the paediatric radiologist versus the registrar [10, 11] or to different levels of context bias [14]. However, the good proportions of agreement for normality tell us that the radiologists agreed on what is normal. The poor proportions of agreement for abnormality, together with statistically highly significant persistent observer bias, suggest that the radiologists were consistently coding abnormalities differently. This is supported by the improvement in kappa, and particularly in the proportion of agreement for abnormality, when the classifications of atelectasis and consolidation are combined as volume loss.

When comparing observations with more than two categories, as we have done for the changes in each lobe (Table 3), we would expect kappa to be lower than when simply classifying abnormalities as present or absent. This is because the opportunities for error and disagreement increase as the numbers of categories increase. The weighted kappa is used to adjust for the seriousness of different levels of disagreement. The weighting applied is decided arbitrarily. We have given the partial agreement between segmental and lobar changes a weighting of 0.75 and the partial disagreement between no abnormality and segmental abnormality a weighting of 0.25. This is because a disagreement over the degree of an abnormality is less serious than a disagreement over whether the abnormality exists or not. The very small improvements in kappa that resulted when these weightings were applied to the partial agreement and disagreement cells suggest that the differences between the two radiologists were not in the degree of abnormality.

Numerous chest radiographs are taken to assess neonates who have respiratory distress. It is well recognised that a degree of variability is a typical feature of radiological interpretation in the clinical setting. This study suggests that it is difficult to differentiate atelectasis from consolidation in the radiographs of babies who are receiving neonatal intensive care. Treatment decisions, such as institution of chest physiotherapy or antibiotic therapy, taken on the basis of radiographic diagnoses made by only one radiologist or neonatalogist are likely to lead to marked variations in the indication for treatment.

The lack of agreement reported here and elsewhere [5] is a reminder that firm diagnoses should not rely on relatively subtle parenchymal changes but that the changes themselves should be described. Furthermore, such findings emphasise the importance of blinded assessments by more than one observer when radiographic interpretation is involved in research.

References

1. Finer NN, Moriartey RR, Boyd J, et al (1979) Postextubation atelectasis: a retrospective review and a prospective controlled study. J Pediatr 94: 110–113
2. Odita JC, Kayyali M, Ammari A (1993) Postextubation atelectasis in ventilated newborn infants. Pediatr Radiol 23: 183–185
3. Flenady V, Bagley C, Tudehope D, et al (1997) Active chest physiotherapy practices in neonatal intensive care: a survey of units in Australia and New Zealand (Abstract). Proceedings of the First Annual Congress of the Perinatal Society of Australia and New Zealand, 45
4. Lewis JA, Lacey JL, Henderson-Smart DJ (1992) A review of chest physiotherapy in neonatal intensive care units in Australia. J Paediatr Child Health 28: 297–300
5. Fitzgerald DA, Van Asperen PP, Lam AH, et al (1996) Chest radiograph abnormalities in very low birthweight survivors of chronic neonatal lung disease. J Paediatr Child Health 32: 491–494
6. Bloomfield FH, Teele RL, Voss M, et al (1998) The role of neonatal chest physiotherapy in preventing post-extubation atelectasis. J Pediatr 133: 269–271
7. Grant JM (1991) The fetal heart rate trace is normal, isn't it? Observer agreement of categorical assessments. Lancet 337: 215–218
8. Brennan P, Silman A (1992) Statistical methods for assessing observer variability in clinical measures. BMJ 304: 1491–1494
9. Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70: 213–220
10. Melbye H, Dale K (1992) Interobserver variability in the radiographic diagnosis of adult outpatient pneumonia. Acta Radiol 33: 79–81
11. Collins CD, Wells AU, Hansell DM, et al (1994) Observer variation in pattern type and extent of disease in fibrosing alveolitis on thin section computed tomography and chest radiography. Clin Radiol 49: 236–240
12. Grenier P, Mourey-Gerosa I, Benali K, et al (1996) Abnormalities of the airways and lung parenchyma in asthmatics: CT observations in 50 patients and inter- and intraobserver variability. Eur Radiol 6: 199–206
13. Rosendahl K, Aslaksen A, Lie RT, et al (1995) Reliability of ultrasound in the early diagnosis of developmental dysplasia of the hip. Pediatr Radiol 25: 219–224
14. Egglin TK, Feinstein AR (1996) Context bias. A problem in diagnostic radiology. JAMA 276: 1752–1755