Critical evaluation of three chest radiograph scores in cystic fibrosis

S M Sawyer, J B Carlin, M DeCampo, G Bowes

Abstract

Background – A number of chest radiographic scores have been developed to assess the severity of respiratory disease in cystic fibrosis but critical statistical evaluation has been limited. In particular, the chest radiograph component of the National Institutes of Health (NIH) clinical score has not previously been validated. Three different chest radiograph scores have been compared and the association between them and lung function tests investigated.

Methods – The interobserver and intra-observer variation of the Brasfield, NIH chest radiograph, and the Royal Children's Hospital (RCH) chest radiograph score was assessed by three observers – a paediatric radiologist, a junior and a senior respiratory physician – who independently scored, on separate occasions, 62 chest radiographs randomly selected from three age strata of patients ranging from 7 to 18 years. Lung function tests were available for 61 patients obtained within three months of the chest radiograph. Two way analysis of variance was used to estimate components of variation in scores.

Results – Results were similar for the Brasfield and NIH scores, both of which demonstrated greater precision than the RCH score, but the estimated repeatability of the Brasfield and NIH scores can be expected to differ by up to 20% of the maximum score. The reliabilities (intraclass correlation) are all reasonably high at 0.74, 0.73, and 0.61 for the Brasfield, NIH, and RCH scores, respectively. The estimated correlation between radiographic scores and lung function tests, adjusted for attenuation caused by measurement error, showed a similar correlation for all three scoring methods ranging from 0.55 to 0.78. Correlations were slightly greater with FEV₁% than FVC%. These correlations are substantial but not high, indicating that a large proportion of the variability in radiographic scores cannot be explained by lung function measurements.

Conclusions – The Brasfield and NIH chest radiograph scores have very similar statistical profiles and can be equally recommended if a chest radiograph score is to be used. The RCH radiographic score appears to be less reliable. The limitations of these scores need to be understood.

(Thorax 1994;49:863-866)

The severity of respiratory disease is the most important single factor affecting mortality in cystic fibrosis.¹ In addition to a clinical assessment, lung function tests, clinical scores and chest radiographic scores have been used to document the severity of respiratory disease, with lung function tests currently accepted as the most informative.

Chest radiograph scores are incorporated into the major clinical cystic fibrosis scores²⁻⁴ and a number of pure radiograph scores have been specifically developed.⁵⁻⁸ Evaluation of the repeatability and reliability of these scores is often incomplete or inadequate, and comparison of chest radiograph scores with lung function tests has been limited.

The respiratory complications of cystic fibrosis ensure that most patients will have many chest radiographs as part of their clinical assessment to determine the progression of bronchiectasis or of acute changes such as a pneumothorax. The broader role of routine chest radiograph scores and the value of including these scores in clinical scores for cystic fibrosis is less clear, as it is not known if radiograph scores provide extra information to lung function tests. The time and effort required to score routine chest radiographs, as well as the potentially harmful effects of multiple chest radiographs, can be justified in a life-limiting condition when there is proven benefit, but this is not the case so far in cystic fibrosis. Complete evaluation of chest radiograph scores is necessary so that the wider issue of the role of chest radiographs in disease evaluation can be fully considered.

We have studied the repeatability, reproducibility, and reliability of three chest radiograph scores and have correlated the three scores with lung function tests in an attempt to provide a complete evaluation of these scoring systems.

Methods

Sample selection was based on the stratification of patients with cystic fibrosis aged 7-18 years by sex and four year age groupings (7-10, 11-14, 15-18 years), with 10 radiographs to be randomly selected from each of the six strata. Ultimately, the study consisted of 62 frontal chest radiographs, as two additional radiographs were included from the older age grouping. Radiographs of stable patients were selected; those at times of acute respiratory illness were not used in the analysis. Each radiograph had identifying information such as patient name and age covered.

The three observers (SMS, GB, MDC) were

Department of
Thoracic Medicine
S M Sawyer
G Bowes

Department of
Clinical Epidemiology
and Biostatistics Unit
J B Carlin

Department of
Radiology
M DeCampo

Royal Children's
Hospital, Parkville,
Victoria 3052,
Australia

Reprint requests to:
Dr S M Sawyer, Respiratory
Biology Program 1-1302,
Harvard School of Public
Health, 665 Huntington
Avenue, Boston,
Massachusetts 02115, USA.

Received 6 December 1993
Returned to authors
5 April 1994
Revised version received
4 May 1994
Accepted for publication
9 May 1994

given a written description of each of the scoring systems and the details of each method were discussed and piloted on an additional 10 radiographs that were not part of the series. The study radiographs were scored independently by each observer using one score, with the process repeated for each scoring system. This process was repeated with a different order of presentation so that each radiograph was scored six times by each scorer. Observer 1 (MDC) is a senior paediatric radiologist, observer 2 (SMS) is a junior respiratory physician, and observer 3 (GB) is a senior respiratory physician.

The radiograph scoring methods used were (1) the Brasfield score, (2) the National Institute of Health (NIH) score, and (3) the Royal Children's Hospital (RCH) pulmonary score.

BRASFIELD SCORE

The Brasfield score⁶ has a total of 25 points that are reduced by increasing severity of radiological changes from each of five specific categories. It was specifically developed as a chest radiograph score for cystic fibrosis, is the best validated of the specific chest radiograph scores, and is commonly used instead of the chest radiograph score in the Shwachman clinical score.²

NATIONAL INSTITUTE OF HEALTH (NIH) SCORE

The NIH clinical score was specifically developed to give prognostic statements about individual patients.⁴ The chest radiograph component is a linear severity scale of 1–13, with a higher score for increasing severity. General guidelines divide the 13 point scale into four severity categories, with up to four additional points to allow for an acute infiltrate so that up to 17 points can be subtracted for chest radiographic changes out of a total of 100 for the NIH clinical score. Although the NIH clinical score is widely used, it remains essentially unvalidated and the chest radiograph component has not been independently assessed.

ROYAL CHILDREN'S HOSPITAL (RCH)

PULMONARY SCORE

The RCH pulmonary score is a simple linear severity scale.⁵ It is based on the four categories of cough, sputum, chest radiograph, and lung function tests (forced expiratory volume in one second as a percentage predicted (FEV₁)) which are rated on a scale from 0 to 4. The simplicity of this score is appealing, and it is used annually to rate the severity of respiratory

disease for patients at this hospital as well as those in the Australian cystic fibrosis database, but neither the chest radiograph score nor the overall score has been validated.

LUNG FUNCTION TESTS

Lung function tests were available for 61 patients from outpatient visits within three months of the chest radiograph during a period when the patients were stable. Forced vital capacity (FVC) and FEV₁ were recorded and expressed as the percentage predicted.⁹

STATISTICAL METHODS

Two way analysis of variance was used on the 372 scores obtained for each scoring method (three observers $\times$ 62 subjects $\times$ two replicates) in order to estimate components of score variation. As measurements were replicated for each subject it was possible to test for interaction between subjects and observers. Inter-observer and intraobserver variability were expressed as (i) repeatability ($2.83 \times$ standard deviation of measurement error) which gives the range within which 95% of differences between ratings performed on the same subject by the same observer should lie; (ii) reproducibility which gives the range within which 95% of differences between ratings performed by different observers on the same subject should lie (interobserver variability); and (iii) reliability or intraclass correlation which is an estimate of the correlation between two ratings on the same subject by different observers (also a measure of interobserver variability).

Correlations between chest radiograph scores and lung function tests were initially obtained using all raw score data ($n = 366$). To adjust for the fact that the measurement error in the radiographic scores causes these correlations to underestimate the true values that would be obtained in the absence of measurement error, two further correlations were obtained, the first correlating the mean radiograph score (six ratings) with lung function tests, and the second by dividing the raw correlation by a factor equal to one minus the ratio of the estimated measurement error variance to the total variance of the chest radiograph score.¹⁰

Results

The mean results of each observer for each score are summarised in table 1. Intra-observer mean differences were highly significant ($p < 0.001$) for all tests, although the

Table 1 Summary statistics of chest radiograph scores

Score	Score range	Observer 1		Observer 2		Observer 3	
		Mean score	Mean (SD) difference*	Mean score	Mean (SD) difference*	Mean score	Mean (SD) difference*
Brasfield	0–25	15.1	–0.1 (1.9)	15.2	0.8 (2.1)	13.1	0.1 (2.0)
NIH	0–17	8.8	–0.4 (1.2)	8.6	0.6 (1.3)	7.3	0.7 (1.2)
RCH	0–4	3.0	–0.2 (0.5)	2.7	–0.1 (0.4)	3.0	–0.05 (0.5)

* Mean (SD) of difference between replicate ratings.

differences did not appear to be related to the experience of the observer, nor to the score design (simple versus complex, linear severity versus multiple categories of response) and were therefore consistent with the notion of observers varying randomly. Similarly, significant interaction effects between observer and subject occurred for each score, but the variance associated with these effects was relatively small and not readily interpretable, so they were ignored in the final analysis.

Repeatability, reproducibility, and reliability were very similar for the Brasfield and NIH scores, both of which demonstrated greater precision than the RCH score (table 2). However, the estimated repeatability of both Brasfield and NIH scores shows that repeated assessments by the same observer can be expected to differ by up to 20% of the maximum score – that is, 4.8 units in the Brasfield and 3.4 units in the NIH score.

The reproducibility values shown in table 2 reflect interobserver variability in addition to the intraobserver variance included in the repeatability values, and the results show that observer variance does not add substantially to the inherent (intraobserver) measurement error. This finding must be tempered by the fact that our estimates of interobserver variance are relatively imprecise, based only on three observers.

Estimated correlations between chest radiograph scores and lung function tests (table 3) show that the three scoring methods correlate to a similar degree, with slightly greater correlations with FEV₁% than FVC% for all methods.

Discussion

The decline in respiratory function with time in cystic fibrosis has resulted in many attempts to develop precise methods for the classification of the severity of respiratory disease. Many chest radiographic scores have been developed to standardise the interpretation of chest radiographs in these patients, but none of the commonly used scores has had appropriate

repeatability studies performed. Increasingly, appropriate statistical techniques for assessing and comparing measurement methods are being applied in respiratory medicine,^{11,12} and this study has used such techniques to critically evaluate three chest radiographic scores.

We studied two radiographic scores that are incorporated in frequently used clinical scores for cystic fibrosis and showed that the greater detail required by using the Brasfield and NIH scores resulted in greater precision than the more straightforward Royal Children's Hospital score. The Brasfield score is currently recommended as the standard system for evaluation of patients with cystic fibrosis in the USA,¹³ but different countries favour different scoring methods. Our results show that the radiograph component of the NIH score is as precise as the better evaluated Brasfield score, and has the additional benefit of being easier to rate than the Brasfield score.

The reliabilities or intraclass correlations (table 2) are reasonably good, although not excellent; a value of 0.6 has been suggested¹⁴ as the minimum necessary for a useful scale. Again, these reflect the better performance of the Brasfield and NIH scores. In interpreting these coefficients it should be noted that they depend on the amount of intersubject variation in the population studied; the satisfactory values reflect, to a large extent, the wide range of subjects studied.¹⁵

Our detailed analysis has shown that considerable interobserver and intraobserver variation exists for all three scoring methods. Previous studies have relied solely on Pearson correlation coefficients for evaluating reliability, but high correlations between observers are not incompatible with substantial variability. For example, Pearson correlation coefficients between mean Brasfield scores by the three observers in this study ranged from 0.84 to 0.87, consistent with Brasfield,¹⁶ but results of analysis of variance showed that intraobserver differences may range by as much as 4.8 of the 25 point score. Moreover, previous studies^{6,16-19} have failed to discuss important methodological issues surrounding the calculation of correlations between chest radiograph scores and other measures such as lung function tests. In particular, the presence of measurement error leads to an attenuation that can be corrected for, either by correlating mean values or by an adjustment based on an estimate of measurement variance. Published results appear to be generally based on single observer scores, although Matthew *et al*¹⁷ and O'Laoide *et al*¹⁹ both used consensus estimates which would reduce error.

Estimated correlations between the Brasfield (or closely related scores such as the Chrispin-Norman score⁷) and lung function tests in previous studies^{6,17-19} range between 0.50 and 0.68 for FEV₁% and 0.35 and 0.68 for FVC%. These correlations have been described as "good" and "very good",¹⁸ though our raw score correlations are not high. This is despite attempting to maximise the possible correlation between pulmonary function tests and chest radiograph scores by only including chest

Table 2 Coefficients of intraobserver and interobserver variability

Score	Repeatability	% of maximum	Reproducibility	% of maximum	Reliability
Brasfield	4.8	19.2	5.8	23.2	0.74
NIH	3.4	20.0	4.1	24.1	0.73
RCH	1.3	32.5	1.4	35.0	0.61

Table 3 Correlation of chest radiograph scores (raw, mean, and adjusted) with lung function

Score	FVC%			FEV ₁		
	Raw*	Mean†	Adjusted‡	Raw*	Mean†	Adjusted‡
Brasfield	0.52	0.59	0.64	0.58	0.65	0.70
NIH	0.44	0.50	0.55	0.52	0.59	0.64
RCH	-0.43	-0.52	-0.64	-0.52	-0.64	-0.78

* Correlation based on 366 individual chest radiograph scores.

† Correlation based on mean of chest radiograph scores for three observers and two occasions.

‡ Raw correlation adjusted for measurement error in chest radiograph score, as described in the text.

radiographs when patients were stable, as the resolution of chest radiographic changes is usually delayed when compared with improvement in lung function. We elected to use a frontal chest radiograph without a lateral film for scoring as a lateral radiograph does not provide major additional information in the assessment of diffuse lung disease such as cystic fibrosis,²⁰ and it is not the practice at the Royal Children's Hospital to routinely take lateral chest radiographs.

The wider issue is whether the radiographic score is a sufficiently independent assessment of the severity of respiratory disease that it justifies inclusion in a score of respiratory function, or whether it fails to contribute anything extra to lung function tests that are cheap and available in all cystic fibrosis centres. A patient with severe respiratory disease may have a chest radiograph characterised by widespread cystic changes, but lung function tests are equally likely to be poor. A perfect correlation ($r=1$) would indicate that chest radiograph scores are redundant, while the moderate values reported here and elsewhere, whether adjusted or not, indicate that a large proportion (40–80%) of the variability of chest radiograph scores cannot be explained by lung function test results. Whether this unexplained variability represents clinically meaningful information cannot be resolved until long term follow up studies are performed of the association between routine measurements of lung function and chest radiograph scores with survival or other clinical end points.

The situation for children who are too young to perform lung function tests is a separate issue. The recent development of the Wisconsin cystic fibrosis chest radiograph score is an example of the recognition of the limitations of currently used radiograph scores in cystic fibrosis.²¹ It was specifically developed to be more discriminating in patients with mild lung disease, unlike the existing scores which are more discriminating for those with moderate or severe disease. It is better validated than the Brasfield score, but is more complex than any of the existing scores as it comprises six major criteria requiring 24 component scores with the final result generated by computer software. As lung function tests currently have limited application in the quantitative assessment of disease severity in young children who generally have mild disease, this score is an exciting development, but longitudinal studies will be required to determine its role. It is noteworthy, however, that despite the increasing specificity of the score, the reproducibility is no better than for the scores we studied.

Although one might expect that creating a radiograph score from a number of distinct components would provide a more accurate measure of severity of a chest radiograph than using a single overall assessment, our results show that the complexity of the Brasfield radiograph score (which requires assessment in five separate categories) provides no more precision of measurement than the NIH radio-

graph score (which is a single linear severity score). Similarly, including a greater number of components in a score of overall respiratory severity will not necessarily provide more precision. Taussig argued that, as the links between particular organ systems in cystic fibrosis are poorly understood, scores for different organ systems should be reported separately.²² Until there is a better understanding of the independent contribution of chest radiograph scores and lung function tests to prediction of medium to long term survival, we recommend that results of single tests of respiratory severity, such as radiograph scores or lung function tests, are better reported separately without inclusion in a more general score of respiratory severity. This recommendation is reinforced by the substantial measurement variability in widely used radiograph scoring systems that we have documented here.

Dr Sawyer was supported by an Australian National Health and Medical Research Council postgraduate scholarship.

- 1 Boat TF, Welsh MJ, Beaudet AL. Cystic fibrosis. In: Scriver CR, Beaudet AL, Sly WS, Valle D, eds. *The metabolic basis of inherited disease*. New York: McGraw-Hill, 1989: 2649–80.
- 2 Schwachman H, Kulczycki LL. Long-term study of one hundred five patients with cystic fibrosis. *Am J Dis Child* 1958;96:6–15.
- 3 Doershuk CF, Matthews LW, Tucker AS, Nudelman H, Eddy G, Wise M, et al. A 5 year clinical evaluation of a therapeutic program for patients with cystic fibrosis. *J Pediatr* 1964;65:677–93.
- 4 Taussig LM, Kattwinkel J, Friedewald WT, di Sant'Agnese PA. A new prognostic score and clinical evaluation system for cystic fibrosis. *J Pediatr* 1973;82:380–90.
- 5 Holzer FJ, Olinsky A, Phelan PD. Variability of airways hyper-reactivity and allergy in cystic fibrosis. *Arch Dis Child* 1981;56:455–9.
- 6 Brasfield D, Hicks G, Soong SJ, Tiller R. The chest roentgenogram in cystic fibrosis: a new scoring system. *Pediatrics* 1979;63:24–9.
- 7 Chrispin AR, Norman AP. The systematic evaluation of the chest radiograph in cystic fibrosis. *Pediatr Radiol* 1974;2: 101–6.
- 8 Cooperman EM, Park M, McKee J, Assad JP. A simplified cystic fibrosis scoring system (a preliminary report). *Can Med Assoc J* 1971;105:580–2.
- 9 Zapletal A, Samenek TP. *Lung function in children and adolescents: methods and reference values*. Basel-München: Karger, 1987.
- 10 Armitage P, Berry G. *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications, 1987:271.
- 11 Chinn S, Britton JR, Burney PGJ, Tattersfield AE, Papacosta AO. Estimation and repeatability of the response to inhaled histamine in a community survey. *Thorax* 1987;42:45–52.
- 12 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
- 13 Wood RE, Farrell PM, eds. *Cystic fibrosis patient evaluations and scoring systems*. Cystic Fibrosis Foundation Conference Report 2 (2), 1978.
- 14 Chinn S. Repeatability and method comparison. *Thorax* 1991;46:454–6.
- 15 Dunn G. Design and analysis of reliability studies. *Statistical Methods in Medical Research* 1992;1:123–57.
- 16 Brasfield D, Hicks G, Soong SJ, Peters J, Tiller R. Evaluation of scoring system of the chest radiograph in cystic fibrosis: a collaborative study. *AJR* 1980;134:1195–8.
- 17 Matthew DJ, Warner JO, Chrispin AR, Norman AP. The relationship between chest radiograph scores and respiratory function tests in children with cystic fibrosis. *Pediatr Radiol* 1977;5:198–200.
- 18 Rosenberg SM, Howatt WF, Grum CM. Spirometry and chest roentgenographic appearance in adults with cystic fibrosis. *Chest* 1992;101:961–4.
- 19 O'Laoidé RM, Fahy J, Coffery M, Ward K, Malone D, Fitzgerald MX, Masterson J. A chest radiograph scoring system in adult cystic fibrosis: correlation with pulmonary function. *Clin Radiol* 1991;43:308–10.
- 20 Kennedy J, Dawson KP, Abbott GD. Should a lateral chest radiograph be routine in suspected pneumonia? *Aust Paediatr J* 1986;22:299–300.
- 21 Weatherly MR, Palmer CGS, Peters ME, Green CG, Fryback D, Langhough R, et al. Wisconsin cystic fibrosis chest radiograph scoring system. *Pediatrics* 1993;91:488–95.
- 22 Taussig LM. *Cystic fibrosis*. New York: Thieme-Stratton, 1984.