Validity and reliability of acoustic analysis of respiratory sounds in infants

H E Elphick, G A Lancaster, A Solis, A Majumdar, R Gupta, R L Smyth

Objective: To investigate the validity and reliability of computerised acoustic analysis in the detection of abnormal respiratory noises in infants.

Methods: Blinded, prospective comparison of acoustic analysis with stethoscope examination. Validity and reliability of acoustic analysis were assessed by calculating the degree of observer agreement using the $\kappa$ statistic with 95% confidence intervals (CI).

Results: 102 infants under 18 months were recruited. Convergent validity for agreement between stethoscope examination and acoustic analysis was poor for wheeze ($\kappa = 0.07$ (95% CI, $-0.13$ to 0.26)) and rattle ($\kappa = 0.11$ ($-0.05$ to 0.27)) and fair for crackles ($\kappa = 0.36$ (0.18 to 0.54)). Both the stethoscope and acoustic analysis distinguished well between sounds (discriminant validity). Agreement between observers for the presence of wheeze was poor for both stethoscope examination and acoustic analysis. Agreement for rattle was moderate for the stethoscope but poor for acoustic analysis. Agreement for crackles was moderate using both techniques. Within-observer reliability for all sounds using acoustic analysis was moderate to good.

Conclusions: The stethoscope is unreliable for assessing respiratory sounds in infants. This has important implications for its use as a diagnostic tool for lung disorders in infants, and confirms that it cannot be used as a gold standard. Because of the unreliability of the stethoscope, the validity of acoustic analysis could not be demonstrated, although it could discriminate between sounds well and showed good within-observer reliability. For acoustic analysis, targeted training and the development of computerised pattern recognition systems may improve reliability so that it can be used in clinical practice.

The evaluation of findings using the clinical history and examination is an important component of clinical diagnosis. Despite this, few studies have assessed their precision and accuracy. Studies in adult subjects suggest that the level of agreement among independent observers for the presence of respiratory signs using a stethoscope is poor.1–3 Inability to identify respiratory signs accurately has been shown to lead to an incorrect diagnosis by physicians on 28% of occasions.4 In spite of its importance, there is no information in current reports on agreement between observers for examination of children and infants with a stethoscope.

Computerised acoustic analysis is a technique used to evaluate the acoustic properties of respiratory sounds. Guidelines on methods of recording and analysis of breath sounds have been published recently in an attempt to standardise terminology.1 Although there were no references to infants in this document, this is a useful technique to apply to infants because it does not require any sedation and is non-invasive. The equipment is portable and measurements can therefore be made at the bedside.

No study has previously compared acoustic analysis with stethoscope examination in infants. This is a particularly important group in which to use terminology with accuracy, as the underlying respiratory diagnosis in this age group can be difficult to establish.5 The present study was designed to compare the findings of acoustic analysis with those of the clinical examination. The aim was to investigate the validity and reliability of acoustic analysis in the detection of abnormal respiratory noises in infants.

METHODS

The study design was a blinded, prospective comparison of computerised acoustic analysis with stethoscope examination. Because of the poor agreement reported between observers for stethoscope examination in previous studies, this method could not be regarded as sufficiently reliable to be used as a gold (reference) standard against which to compare the acoustic analysis technique. Instead, we chose to measure agreement between the two methods.

The study received approval from the local paediatric research ethics committee, and informed consent was obtained from the parents of all participants.

Participants

We recruited 102 subjects from inpatients in the general paediatric wards at Alder Hey Children’s Hospital, Liverpool, with noisy breathing arising from lower respiratory tract illness. Infants of either sex under the age of 18 months were included. There were no exclusion criteria. Clinical information was not shared with any of the investigators before the study measurements took place. On the days of recruitment a consecutive sample was taken, regardless of the clinical condition of the patient, and all participants included were subjected to both measurements, except for two who were not recorded by acoustic analysis—one because of breakdown of recording equipment and one because of acute deterioration in clinical status. Each child was examined using a stethoscope by two experienced clinicians, and the findings on each infant were recorded independently of each other and independently from the acoustic analysis findings. Blinding was ensured by asking each clinician to complete evaluation forms, which were then placed into a coded envelope which was not opened until the acoustic analysis had been completed. No treatments were given between tests, which were done either concurrently or immediately sequentially.
Stethoscope examination

The clinicians (AM and RG) were specialist registrars with more than six years’ experience in paediatrics as well as having experience in respiratory paediatrics. Definitions of the respiratory noises to be assessed were agreed on before the start of the study. Further than this, we did not undertake to train the clinicians to high levels of agreement, so that estimates of reproducibility were reflective of the everyday clinical environment. Each examination took place for one minute during a time when the infant was quiet and settled. The right upper zone of the anterior chest was the only site examined, so that a direct comparison could be made with acoustic analysis. The clinicians were given an evaluation form which required assessment of the presence or absence of wheeze, rattles, and crackles.

Table 1  Patient characteristics

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (months) (mean (SD))</td>
<td>5.1 (3.3)</td>
<td>69</td>
</tr>
<tr>
<td>Weight (kg) (mean (SD))</td>
<td>6.8 (3.5)</td>
<td>69</td>
</tr>
<tr>
<td>Per cent male</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>Diagnosis (n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bronchiolitis</td>
<td>51/88</td>
<td></td>
</tr>
<tr>
<td>RSV</td>
<td>1/88</td>
<td></td>
</tr>
<tr>
<td>Virus unknown</td>
<td>36/88</td>
<td></td>
</tr>
<tr>
<td>Required oxygen</td>
<td>66/88</td>
<td></td>
</tr>
<tr>
<td>Required nasogastric/IV feeding</td>
<td>49/88</td>
<td></td>
</tr>
<tr>
<td>Required IV antibiotics</td>
<td>9/88</td>
<td></td>
</tr>
<tr>
<td>Required bronchodilators</td>
<td>22/88</td>
<td></td>
</tr>
<tr>
<td>Average length hospital stay</td>
<td>5.7 days</td>
<td></td>
</tr>
</tbody>
</table>

Other diagnoses (n)
- Pneumonia/atelectasis 3
- Cerebral palsy 2
- BPD 1
- Asthma 3
- VATER anomaly 1
- Cystic fibrosis 1
- Whooping cough 1

BPD, bronchopulmonary dysplasia; IV, intravenous; RSV, respiratory syncytial virus; VATER, vertebral/anal/tracheal/esophageal/renal syndrome.

Acoustic analysis

Recordings of the infant’s respiratory sounds were undertaken by HE using a contact sensor (Siemens EMT 25C) placed over the right upper zone of the infant’s chest, anteriorly, attached to the skin using a double sided adhesive ring. The recording on each occasion lasted for approximately one minute of tidal breathing, with the baby quiet or asleep. Recordings were made in a quiet room with background noise kept to a minimum. Specialised equipment for sound recording and analysis was used (RALE, Respiratory Acoustics Laboratory Environment). The signals were filtered, amplified, and transmitted to an IBM compatible personal computer. The sound was analysed using a fast Fourier transformation technique. A waveform (amplitude/time), a power spectrum (intensity/frequency), and a sonogram (frequency/time, with intensity as a colour scale), were displayed on the computer screen and each of these visual patterns could be assessed for the identification of wheeze, rattles, and crackles.
Classification of sounds
The classical wheeze is a high pitched whistling sound with a musical quality, characterised acoustically by a sinusoidal waveform with distinct peaks in the power spectrum display.7 Crackles are defined as intermittent sounds of short duration with a characteristic sharp, sudden deflection followed by rapidly dampened wave deflections.8 Rattles (referred to in some parts of the United Kingdom as “ruttles”) are much lower in pitch than a wheeze, with a continuous non-musical quality, and may reflect excessive secretions in the airways (fig 1). Rattles are often mistaken for wheeze,10 but in acoustic terms, this noise is quite distinct, with an irregular non-sinusoidal waveform and diffuse peaks in the power spectrum.

Sample size calculation
Sample size was calculated using Nquery Advisor v4.0.11 For the validity comparison between stethoscope examination and acoustic analysis, for a sample size of 96 a two sided 95% confidence interval (CI) for the $\kappa$ statistic will extend approximately 0.2 from the observed value of $\kappa$, assuming the true value of $\kappa$ to be between 0.1 and 0.4 and the proportion of successes to be 0.5.

Data analysis
The degree of observer agreement was represented using the $\kappa$ statistic with 95% CI using the StatsDirect statistical package.12 The $\kappa$ statistic can range from −1 to +1. A negative value implies that the proportion of agreement made by chance is greater than the proportion of observed agreement—that is, the agreement is less than would have been expected by chance. Positive values ranging from 0 to <0.2 indicate poor agreement, >0.2 to 0.4 fair agreement, >0.4 to 0.6 moderate agreement, >0.6 to 0.8 good agreement, and >0.8 to 1 very good agreement.13

Validity
Construct validity was assessed by determining convergent and discriminant validity (n = 100). Convergent validity shows the agreement between the two different methods in measuring the same sound. Discriminant validity is represented by the ability of the test to discriminate between different sounds.

Reliability
Separate blind evaluations by different observers (between-observer agreement) and repeat assessments of the same patient by the same observer (within-observer agreement) were carried out. These were done in four consecutive subsets of participants.

- Between-observer agreement for the stethoscope was assessed by comparing the findings of clinicians AM and RG (n = 36). This analysis was included to compare our findings with previous studies.
- Between-observer agreement for the acoustic analysis was assessed by comparing the results of analysers HE and AS, of a subset of the breath sound recordings (n = 20).
- Within-observer agreement for the acoustic analysis instrument was assessed by re-recording each patient 5 minutes after the initial recording (n = 58). All of these recordings were analysed by the same analyser (HE).
- Within-observer agreement for repeat analysis was assessed by analysis of the same recording by the same analyser (HE) on two separate occasions, two months apart (n = 19).

RESULTS
In all, 102 infants were recruited, achieving our calculated sample size requirement. Acoustic and stethoscope data were available for 100 infants. Table 1 outlines the clinical characteristics of the patient group.

Validity
Convergent and discriminant validity are shown using 2×2 tables (table 2).

Table 2
Validity of acoustic analysis v stethoscope

<table>
<thead>
<tr>
<th></th>
<th>Wheeze</th>
<th>Rattles</th>
<th>Crackles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td><strong>A Convergent validity: acoustic analysis v stethoscope</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheeze</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>38</td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td>Yes</td>
<td>18</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>44</td>
<td>100</td>
</tr>
<tr>
<td>$\kappa$ = 0.07 (−0.13 to 0.26)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Wheeze v rattles</th>
<th>Wheeze v crackles</th>
<th>Rattles v crackles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>30</td>
<td>26</td>
<td>56</td>
</tr>
<tr>
<td>No</td>
<td>22</td>
<td>22</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>48</td>
<td>100</td>
</tr>
<tr>
<td>$\kappa$ = 0.01 (−0.18 to 0.21)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Wheeze v rattles</th>
<th>Wheeze v crackles</th>
<th>Rattles v crackles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>55</td>
<td>10</td>
<td>65</td>
</tr>
<tr>
<td>No</td>
<td>24</td>
<td>11</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>79</td>
<td>21</td>
<td>100</td>
</tr>
<tr>
<td>$\kappa$ = 0.16 (−0.02 to 0.33)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2×2 tables demonstrating convergent and discriminant validity; $\kappa$ (95% confidence interval) given for each comparison.
Convergent validity

Convergent validity is shown in table 2A. Of 100 infants, the number for whom there was agreement for identification of wheeze was 55 (κ = 0.07 (95% CI, −0.13 to 0.26)), for rattles it was 57 (κ = 0.11 (−0.05 to 0.27)), and for crackles it was 70 (κ = 0.36 (95% CI, 0.18 to 0.54)). The κ scores indicate poor agreement, except for crackles, for which agreement was fair.

Discriminant validity

To demonstrate discriminant validity, there should ideally be total disagreement between the techniques, and κ values for agreement should be negative. In 100 infants, for wheeze and rattles the stethoscope (table 2B) was able to distinguish between the two noises in 48 (κ = 0.01 (−0.18 to 0.21)), compared with acoustic analysis (table 2C), which was able to distinguish between wheeze and rattles in 34 (κ = 0.16 (−0.02 to 0.33)). For wheeze v crackles, the stethoscope was able to distinguish between the noises in 57 infants (κ = −0.1 (−0.27 to 0.07)) and acoustic analysis in 53 (κ = −0.03 (−0.19 to 0.16)). For rattles v crackles, the stethoscope was able to distinguish between the noises in 63 infants (κ = −0.24 (−0.4 to 0.07)) and acoustic analysis in 60 (κ = −0.13 (−0.2 to 0.05)). The negative values for κ imply that agreement is less than would have been expected by chance.

Reliability

The 2×2 classification tables for each comparison together with the κ statistics are shown in table 3.

### Between-observer agreement for stethoscope

These data are shown in table 3A. Of 36 infants, the number for whom there was agreement was 20 (κ = 0.18 (95% CI, −0.08 to 0.44)). This κ score represents poor agreement. The number of infants for whom there was agreement for rattles was 28 (κ = 0.53 (0.21 to 0.86)) and for crackles it was 31 (κ = 0.46 (0.14 to 0.79)). These κ scores represent moderate agreement.

### Between-observer agreement for acoustic analysis

These data are shown in table 3B. Of 20 infants, the number for whom there was agreement for wheeze was 12 (κ = 0.24 (−0.12 to 0.60)), for rattles it was 7 (κ = 0.22 (−0.18 to 0.63)), and for crackles it was 15 (κ = 0.44 (0.03 to 0.86)). Agreement was therefore poor for wheeze and rattles and moderate for crackles.

### Within-observer agreement for acoustic analysis

These data are shown in table 3C. Of 58 infants, the number for whom there was agreement for wheeze was 46 (κ = 0.57 (0.32 to 0.82)), for rattles it was 51 (κ = 0.59 (0.34 to 0.85)), and for crackles it was 47 (κ = 0.56 (0.30 to 0.82)). The κ scores indicate moderate agreement for each comparison.

### DISCUSSION

This study is the first to compare acoustic analysis with stethoscope examination in infants. One of the main findings was that the reliability of stethoscope examination of respiratory sounds in infants between two experienced observers was poor to moderate. Between-observer reliability using computerised acoustic analysis was also poor to moderate, but this may have reflected the relative inexperience of the second analyst used in the study. Within-observer reliability for all sounds using acoustic analysis was moderate to good. A perfect correlation would never occur, as changes in the subject’s clinical state—even as a result of a cough—may take place between recordings, but our findings do reflect good stability of the technique. Validity of computerised acoustic analysis was poor to fair, although it was difficult to draw clear conclusions owing to the unreliability of the stethoscope to which it was being compared.

A strength of this study was that rigorous methodology was used, in line with the Cochrane Methods Group recommendations for studies of diagnostic accuracy. In this

---
way most potential biases were overcome. A sample size was calculated for the comparison between the techniques and this sample was achieved. A problem with our study was that smaller sample sizes for the subsets were used for the reliability comparisons. These smaller subsets were unavoidable because of time and personnel restraints, and resulted in wider confidence intervals for these comparisons.

The $k$ values for agreement using the stethoscope in our study were similar to or lower than those reported previously in studies in adults,\textsuperscript{15} and also lower than the value in a previous study in which a comparison was made between computer analysis and subjective assessment.\textsuperscript{18} All the previous studies assessed only the presence or absence of one noise, wheeze. Ours is the only study so far to have investigated in detail the distinction between three different sounds, which—while reflecting clinical practice more accurately—added to the complexity of the study and probably reduced the apparent reliability of the two measurements.

Most of the infants included in this study had bronchiolitis, which is characteristically associated with a combination of more than one respiratory sound. Each participant therefore is likely to have had more than one noise identified during the minute-long study period. Other studies have investigated infants with bronchiolitis and recurrent wheeze using computerised acoustic analysis.\textsuperscript{14} \textsuperscript{17} “Complex repetitive waveforms” were described in some of these infants. These are likely to be similar to those noises recorded as “rattles” in our study. The presence of different sounds and the inconsistency in nomenclature adds confusion to the evaluation of an already challenging group of individuals in terms of their age and ability to cooperate.\textsuperscript{14}

In clinical practice, the stethoscope is widely used in the evaluation of young children with respiratory symptoms, given the paucity of more objective diagnostic alternatives. One of the most common diseases of childhood, asthma, is usually diagnosed on the basis of a history of recurrent wheeze, and epidemiological studies of the natural history and treatment of asthma are based on cohorts characterised by the presence or absence of wheeze. In bronchiolitis, oxygen requirement and the presence of crackles on examination of the chest have been shown to be the only clinical predictors of severity of the illness.\textsuperscript{19} The poor level of agreement for stethoscope findings in our study raises concern over the reliability of this method for the evaluating wheezing illness in this age group, and for the initial assessment of infants with bronchiolitis.

Our study shows how unreliable the stethoscope is in the assessment of respiratory sounds in infants. This has important implications for its use as a diagnostic tool for lung disorders in infants and confirms that it cannot be used as a gold standard. Because of the unreliability of the stethoscope, the validity of acoustic analysis could not be demonstrated, although it was able to discriminate between sounds well and showed good within-observer reliability.

Acoustic analysis does have a possible role, provided that validity and reliability can be improved. In addition to specific and targeted training, the development of simpler and lower cost instruments and computerised pattern recognition systems may reduce variability in the interpretation of respiratory sounds in infants.

**ACKNOWLEDGEMENTS**

We are grateful to Dr John Earis for his helpful comments on the manuscript and to the RLCH Endowment fund for providing funding for the equipment.

**Authors’ affiliations**

H E Elphick, R L Smyth, Institute of Child Health, Royal Liverpool Children’s Hospital, Liverpool, UK

G A Lancaster, Centre for Medical Statistics and Health Evaluation, University of Liverpool

A Solis, A Majumdar, R Gupta, Respiratory Unit, Royal Liverpool Children’s Hospital

**REFERENCES**


Validity and reliability of acoustic analysis of respiratory sounds in infants

H E Elphick, G A Lancaster, A Solis, A Majumdar, R Gupta and R L Smyth

Arch Dis Child 2004 89: 1059-1063
doi: 10.1136/adc.2003.046458

Updated information and services can be found at:
http://adc.bmj.com/content/89/11/1059

These include:

References
This article cites 14 articles, 5 of which you can access for free at:
http://adc.bmj.com/content/89/11/1059#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/