How Accurate Is the Clinical Assessment of Acute Dyspnoea and Wheeze in Children?

Colin V E Powell

Arch Dis Child. 2015;100(9):813-815.

Assessing acutely unwell children with wheezing illness is a daily activity for an acute paediatrician. It is one of the most common presentations in unscheduled care.[1] Judgements are made clinically on the severity of acute dyspnoea and treatment commenced. The response to treatment and the need to escalate the treatment or the decision to discharge or admit to hospital will depend on those clinical judgements. Research evidences for the efficacy and effectiveness of treatments are also based on clinical outcomes. These need to be considered sufficiently robust and valid to represent a primary outcome of sufficient quality to demonstrate differences in treatment. In acutely wheezy children, there have been many attempts to validate clinical scores for use both clinically and for research purposes. Indeed there are over 35 severity scores for use in assessing acutely dyspnoeic wheezy children.[2]

Are these measurements and scores sufficiently accurate for use clinically and for use in research? There are three key features to consider: validity, reliability and utility. Each one of these qualities is composed of a number of features. 'Validity' comprises face validity, content validity, construct validity and criterion-concurrent validity. 'Reliability' is composed of measurement error, interobserver reliability, intraobserver reliability, internal consistency and responsiveness. Utility is made up of suitability, age span covered, ease of scoring, skills required to complete the score, the 'floor to ceiling effect' and interpretability.[2] It is clear that none of the scores used in this field have had full detailed development addressing all of these features.[2]

Thus, it is important to challenge their use in both research and clinical arenas and in the July 2015 edition, Bekhof *et al*[3]examined some of the clinical signs used to make up the components of wheeze severity scores. They challenged us to the notion that measurement error due to poor interobserver error is so large that it will impair the perception of any clinically relevant changes in dyspnoea after treatment in two-thirds of observations. This is really important and thus an accurate outcome for acute wheezing still remains a huge challenge for researcher and clinical health professional.

They clearly focused on two of the issues of reliability: measurement error and interobserver and intraobserver reliability. They also asked you to consider the concept of smallest detectable change (SDC) and minimally important change (MIC) and the effect on perception of clinically relevant changes in a child's condition.

It is worthwhile reading their definitions section thoughtfully. Repeated measurements show variations within and between assessors and variation within patients. The SE of the measurements represents the magnitude of measurement error. Reliability is the degree to which the measurement is free from measurement error. Intraobserver reliability is the variation within one observer and interobserver reliability refers to the variation between observers. The statistical parameters used to report reliability are the Cohen's Kappa score for dichotomous data and intraclass correlation coefficient for continuous data.

Measurements also need to be able to identify change in response to treatment. The SDC is the smallest, within-person change, which can be interpreted as real change above measurement error. Then, this needs to be considered in the light of the MIC, which is the MIC that is important to the patient or clinician. When MIC exceeds the SDC, the measurement has good clinical value because clinically relevant changes can be distinguished from measurement error. This is an important concept to understand.

Bekhof used a technique for validation called the 'visual anchor'-based MIC distribution to calculate the MIC of the dyspnoea score. Using two external criteria as anchors, the consultant paediatricians' judgement whether the sign was worse, had no change, was slightly improved or markedly improved (note not the patient or family member) and the change in respiratory rate before and after bronchodilators as measured by the nurse in the emergency department, the MIC was calculated.

Their results showed moderate to good intraobserver reliability for those clinical signs but poor interobserver reliability for the clinical assessment of dyspnoea severity scoring systems. Due to the variation within and between observers, the SDC exceeded the MIC in nearly 70% of observations. This in turn will obscure the detection of any clinically important improvement after treatment. Their conclusions are that this poor interobserver reliability of clinical dyspnoea in children limits its usefulness both clinically and in research and highlights the need for more objective measurements in these patients. They highlight that when a large study using a severity score as a primary outcome is reported, interobserver and intraobserver reliability of the score, when used in that study, should also be reported.

There are criticisms of the methods used here, which need to be considered. This work was all done using video and not a clinical scenario where one can palpate scalene muscle contractions, which is part of the Paediatric Respiratory Assessment Measure (PRAM)[2] and auscultate for wheeze, which is a component of many scores.[2] There were a limited number of patients, so the full range of severity was not represented, which is important when examining the 'floor to ceiling' effect of a score.

They examined an age range between 3 months and 7 years, but they did not have sample size to divide the study into different age categories; >90% were less than 4 years old but. Recent work on bronchiolitis scores in the younger infants has shown adequate inter-rater reliability[4] although poor construct validity. Indeed is it appropriate to expect a score to be validated across all ages? PRAM has been validated for use between age groups 2 and 17 years, and the Asthma Severity Score between '0' and 17 years.[2] Surely the validation process must be different for different ages where variability such as verbal ability and understanding, compliance of chest wall, pathology causing the acute wheezing and the ability to complete lung function tests will all have major impact on the performance and assessment of a score. Perhaps we need to have a score that includes age in the model or a modified approach for different age groups.

They used the judgement of consultant paediatricians as the gold standard and 'visual anchor' for the MIC; one might argue that the parent or indeed the subject with wheeze (given an appropriate developmental age) be the more important visual anchor. Finally, they do not mention the impact of training on the use of these scores, which could influence measurement error.

Nevertheless we are left with a huge challenge. The variable quality of wheeze severity scores makes full reporting of their qualities mandatory when used as outcomes in studies.[5,6] What is a minimal clinically significant difference in a score is extremely relevant, especially with the effect of measurement error and the SDC perhaps being too great for an observer to perceive a clinically relevant change MIC. When a score change is statistically significantly different when two treatments are compared, what does this clinically mean to the patient?[6]

It is easy for 'expert opinion' to suggest what might be in a score but it is much harder to validate a score examining the 13 qualities required for full validation.[2] International agreement among the clinical and research communities concerning the establishment of core outcomes for acute wheezing is essential for further large multicentre studies and to improve clinical assessment of wheezy children.

Among the 35 wheeze severity scores, we must agree on which one is most appropriate for which clinical situation and which age and severity and refine these further rather than inventing yet another one. We need to work further on agreeing on the most objective outcome to be measuring in studies in this population to complement a standardised severity score.

References

1. Jackson DJ, Sykes A, Mallia P, *et al*. Asthma exacerbations: origin, effect and prevention. *J Allergy Clin Immunol*2011;128:1165–74.

2. Bekhof J, Reimink R, Brand P. Systematic review: insufficient validation of clinical scores for the assessment of acute dyspnoea in wheezing children. *Paedatr Resp Rev* 2014;15:98–112.

3. Bekhof J, Reimink R, Doorenbos N, *et al*. Large observer variation of clinical assessment of dyspnoeic wheezing children. *Arch Dis Child* 2015;100:649–53.

4. Fernandes R, Plint A, Terwee C, *et al*. Validity of bronchiolitis outcome measure s. *Pediatrics* 2015;135:e1399–408.

5. Panickar J, Lakhanpaul M, Lambert PC, *et al*. Oral prednisolone for preschool children with acute virus-induced wheezing. *N Engl J Med* 2009;360:329–38.

6. Powell CVE, Kolamunnage-Dona R, Lowe J, *et al*. Magnesium sulphate in acute severe asthma in children (MAGNETIC): a randomised, placebo controlled trial. *Lancet Respir Med* 2013;1:301–8.

Provenance and peer review
Commissioned; internally peer reviewed.