Validity of clinical severity scores for respiratory syncytial virus: a systematic review

Drafted for submission to the Journal of Infectious Diseases

Validity of clinical severity scores for respiratory syncytial virus: a systematic review

Type: Major Article

Abstract word count: 276

Full-text word count: 4046

Authors: Zakariya Sheikh[1], Ellie Potter[1], You Li[2], Rachel A Cohen[3], Gaël Dos Santos[4], Louis Bont[5], Harish Nair[6*] on behalf of PROMISE investigators

*Corresponding author: Harish.Nair@ed.ac.uk

Affiliations:

1. Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK.
2. School of Public Health, Nanjing Medical University, Nanjing, China.
3. GSK, Rockville, Maryland, USA.
4. GSK, Wavre, Belgium.
5. Department of Pediatrics, University Medical Center Utrecht, Utrecht, The Netherlands.
6. Usher Institute, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK.

Abstract

Background

Respiratory syncytial virus (RSV) is a widespread respiratory pathogen, and RSV-related acute lower respiratory tract infections are the most common cause of respiratory hospitalisation in children under two. Over the last two decades, a number of severity scores have been proposed to quantify disease severity for RSV in children yet there remains no overall consensus on the most clinically useful score.

Methods

We conducted a systematic review of English-language publications in peer-reviewed journals published since January 2000 assessing the validity of severity scores for children (≤24 months) with RSV and/or bronchiolitis, and identified the most promising scores. For included articles, (i) validity data were extracted, (ii) quality of reporting assessed using the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis checklist, and (iii) quality assessed using the Prediction model study Risk Of Bias Assessment Tool. To guide the assessment of the validity data, standardised cut-offs were employed, and an explicit definition of what we required to determine a score was sufficiently validated.

Results

Our searches identified 8,541 results, of which 1,779 were excluded as duplicates. After title and abstract screening, 6,670 references were excluded. Following full-text screening & snowballing 32 articles, including 31 scores, were included. The most frequently assessed scores were the modified Tal score and Wang Bronchiolitis Severity Score; none of the scores were found to be sufficiently validated according to our definition. The reporting and/or design of all the included studies was poor. The best validated score was the BROSJOD score, and a number of other promising scores were identified.

Conclusions

No scores were found to be sufficiently validated. Further work is warranted to validate the existing scores, ideally in much larger datasets.

Keywords: RSV, severity score, systematic review, validity

Lay Summary [online only]

Respiratory syncytial virus or RSV causes mild, 'cold-like' symptoms in older children and adults. In young children RSV is a common cause of lung infections like pneumonia and bronchiolitis. Scientists do not agree on the best way to define infant RSV severity. There are different methods for healthcare providers to assign RSV severity scores and scientists use mathematical techniques to evaluate a score's validity, to see how well it works.

We reviewed scientific articles for RSV or bronchiolitis severity scores for children under two years old. We looked at databases of scientific articles to find articles on this topic written in English and published from 1 January 2000 to 15 August 2023. We removed duplicates, then two people reviewed each article against the same list of criteria, to decide if we should include it. We then used standard checklists to determine the article's quality, and recorded the article's validity data.

Our searches found 8,541 results, of which 1,779 were duplicates and 6,670 were excluded; 32 articles were included with information on 31 severity scores. We did not find any fully validated RSV severity score for infants under two years old. The BROSJOD score had the best validity, and there were other promising scores.

Introduction

Respiratory syncytial virus (RSV) is a common respiratory infection; it is estimated that by the age of two years most children will have experienced at least one RSV infection [1]. While the vast majority of RSV infections in infants are self-limiting and non-serious, presenting only with generic symptoms of a mild upper respiratory tract infection (e.g. cough, runny nose), a fraction of infants, will develop an acute lower respiratory tract infection, most commonly presenting as bronchiolitis or less commonly as pneumonia. We previously estimated that in 2019, there were 33.0 million cases of RSV-related acute lower respiratory tract infections in children younger than 5, which resulted in 3.6 million hospital admissions, and 101,400 RSV-attributable overall deaths [2]. As such, RSV-related acute lower respiratory tract infections are the most common cause of respiratory hospitalisations in children aged below 5 years. Notably the vast majority of RSV-related acute lower respiratory tract infections occur in low-income countries.

Over the last two decades, a number of different scoring systems have been proposed to quantify disease severity of RSV in children to aide in clinical decision-making, and serve as outcome measure/clinical endpoint for clinical trials of vaccines and therapeutics. There are many ways to assess the usefulness of these scores; this primarily consists of assessing their validity (face, discriminative, construct, criterion), reliability, responsiveness and utility [3-4].

A major review of severity scores, published more than a decade ago but still oft-cited, found all of the paediatric dyspnoea scores to be insufficiently evaluated across all domains [3]. The literature base was re-examined in a systematic review & meta-analysis published in 2017, a review published in 2018 and most recently in a rapid review published in 2020 specifically looking to identify scores for resource-limited settings [5-7]. All of these similarly found the severity scores to have been insufficiently validated.

This lack of a validated severity score is significantly impacting on clinical trials; a 2015 meeting of key academic, commercial & regulatory stakeholders in RSV vaccine development identified the lack of "clinically meaningful and reproducible indicators" as the biggest challenge to RSV vaccine development [8]. The lack of consensus was similarly expressed in a recent review of RSV vaccines [9].

Given that it has been almost three years since the last review was conducted, we sought to re-examine the literature base to identify and report on efforts to validate clinical severity scores for use in children (≤24 months) with RSV and/or bronchiolitis, and synthesise the data to report on the criterion-concurrent and construct validity of the identified severity scores, as well as the included parameters of these scores. Based on this, we identified the most promising scores.

Methods

Three online medical literature databases, MEDLINE, Embase and Global Health, were searched using the Ovid platform in June 2022 for English-language publications published in peer-reviewed journals since January 2000 on the validity of severity scores for children with RSV or bronchiolitis. The search strategies for each database can be found in *Annex 1;* they were adapted from a recent systematic review on biomarkers for disease severity in RSV [10].

A severity score was defined as a tool used to quantify disease severity over the course of the illness; as such single-purpose models, such as models designed to only predict hospital admission, were excluded.

Covidence was used to identify and automatically exclude duplicates [11]. After removing duplicates, we screened the titles and abstracts of the articles for relevance using pre-defined inclusion/exclusion criteria (see *Table 1*). The inclusion/exclusion criteria were similarly adapted from the aforementioned biomarkers review [10].

For the remaining included papers, their full-text was acquired, and subsequently screened for relevance. The reference lists of papers identified for inclusion, as well as 3 previous reviews, were examined to identify additional relevant references (i.e., snowballing) [3,6,7].

Data from the included studies were extracted into a standardised spreadsheet [12]. The World Bank's income level classification scheme was used to categorise the economies of the countries [13]. Data were simultaneously separately collected on the parameters included in each score (e.g., presence of fever). Additionally, score names were standardised.

Given the widely observed poor quality of publications reporting prediction models, as well as specifically for severity scores for RSV, we employed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis checklist (TRIPOD), a 23-item checklist to quantify the quality of reporting [5, 14-16]. The related Prediction model Risk Of Bias Assessment Tool (PROBAST) was also employed to assess the risk of bias of included studies [14, 17]. For the included studies, the TRIPOD and PROBAST checklists were both assessed.

Each of the above mentioned steps were conducted independently by two reviewers (EP & ZS); any uncertainty was resolved through consultation with a senior researcher (HN). We updated the searches up to 15 August, 2023.

Given the heterogeneous nature of the included studies and the small amount of data on each severity score, only a narrative synthesis was made and a meta-analysis was not conducted. The review was registered with PROSPERO (CRD42022343781).

Quality assessment of validity of identified scores

Using the data extracted from the included studies, we assessed each of the identified scores for their face, construct (discriminative & convergent) and criterion-concurrent validity. We found, similarly to the 2014 review, a wide range of different uses of these terms and so have explicitly specified how we categorised and assessed the validity data (see *Supplementary Table 1*) [3].

To guide our assessment, the same cut-offs as proposed by Hakizimana et al in their rapid review were used [7]. For area under the receiver-operator characteristic curve (AUROC), a score of <0.5 was classified as poor, 0.5-0.7 low, 0.7-0.9 moderate & >0.9 high, and for Spearman's correlation coefficient we took 0–0.19 as very weak, 0.2–0.39 weak, 0.4-0.59 moderate, 0.6–0.79 strong, and 0.8–1 as a very strong correlation. As Hakizimana et al didn't specify cut-offs for Pearson's correlation coefficient; we used <0.1 as negligible, 0.1-0.4 weak, 0.4-0.7 moderate, 0.7-0.9 as strong & >0.9 as very strong. For other measures we made a subjective assessment informed by the above cut-offs. We considered a p-value ≤0.01 as constituting statistical significance.

We considered a score to be sufficiently validated if at least two external validation studies with a low risk of bias rating (as assessed by PROBAST) had assessed the criterion-concurrent, convergent and/or discriminative validity for at least two separate outcomes each, and that performed at least moderately for each outcome. To identify promising scores (i.e. scores that are currently insufficiently validated), we made a subjective assessment based on the scores that were deemed that most likely could be sufficiently validated.

Results

Descriptive statistics

Initial searches produced 7,391 results (see *Figure 1*) of which 59 articles were identified for full-text screening after title and abstract screening. Of these, 24 were included. Our updated search yielded 1,150 results of which 30 articles were identified for full-text screening after title and abstract screening. Of these, 6 were included.
Two additional relevant articles were identified through snowballing. As such, overall 32 articles were included, comprising 31 unique scores (see *Supplementary Table 2*) [18-49]. The vast majority of the included studies used a prospective design (n=27), most commonly a cohort study (n=22) and the remaining 5 studies used either a purely retrospective design (n=4) or combination of retrospective and prospective design (n=1).

Four studies developed a new score, of which one included external validation in the same publication; the remaining 28 studies validated existing scores. Eight studies were multi-centre studies. Twenty-five studies used data collected in secondary care, including three studies which also made use of data from the community; the remaining six studies used data collected in tertiary care, including one which also made use of data from the community.

The most frequently used scores were the modified Tal (mTal) score and Wang Bronchiolitis Severity Score (WBSS) each of which was used in five studies. Four studies used Bronchiolitis Score of Sant Joan de Déu (BROSJOD) and Wood- Downes-Ferrés score (WDF); three studies used the Global Respiratory Severity Score (GRSS). The Bronchiolitis Severity Score (BSS), Escala de Severidad de la Bronquiolitis Aguda (ESBA), Freire model, modified Respiratory Index Score (mRIS), and modified modified Wood's clinical asthma score (mWCAS) were each used in two studies. The remaining 21 scores were only evaluated once. Although Raita et al. [42] claimed to use the Freire model – a model developed by Freire et al. [30] - they excluded one of the parameters included in the original Freire model, so we considered it as a separate score and referred to it as the modified Freire model (mFreire).

Most commonly discriminative validity was assessed (n=24). Sixteen studies assessed convergent validity and 4 criterion-concurrent validity.

Seven papers used data from Spain, five from the United States, four from Israel, two each from Australia, France, Singapore and Turkey and one each from Canada, Colombia, Egypt, India, Ireland, Japan, New Zealand, Portugal,

and the United Kingdom. The vast majority of the included data were from high-income countries (n=28); only three studies used data from upper middle-income countries (Turkey [n=2] & Colombia), and two from a lower middle-income country (Egypt, India). No included papers used data from any low-income country.

Severity score components

For 27 of the scores, we were able to identify the parameters used; however, we were unable to identify all of the parameters used in the four machine learning models proposed by Raita et al. [42] – the authors only mention the 15 most important predictors. There was significant variation in the parameters used by each severity score model. After grouping synonymous terms (e.g. respiratory rate & respiratory frequency), 52 unique parameters were included in the scores (see *Supplementary Table 3*).

The mean number of parameters in each score was 5 (range 3-10). Most commonly included was respiratory rate (n=21); the next most common parameters included retractions (n=13), oxygen saturation (n=12), wheezing (n=11), and heart rate (n=6). The majority of parameters were used ≤3 times (n=41).

Discriminative Validity

Twenty-four of the studies assessed the discriminative validity of the scores, mostly by assessing their ability to discriminate between those discharged or admitted to the hospital, and between those admitted to the paediatric intensive care unit (PICU) and those hospitalised but not admitted to the PICU . The WBSS and BROSJOD were assessed in five papers, the WDF and mTal score in three papers, and the WDF, ESBA, Freire GRSS, mRIS and mWCAS in two papers; the remaining 14 scores were only evaluated once.

Anıl et al. [20] reported that hospitalised patients had significantly higher WBSS than those discharged, as assessed by an odds ratio (OR). There were significant differences between those classified as mild, moderate and severe (according to the WBSS) and a control group, for the pulse rate, respiratory rate and oxygen saturation. They also reported significant differences in the pH & pCO_2 between those with a severe WBSS score compared to the control, and mild & moderate bronchiolitis severity group. De Rose et al. [26] reported high discriminative validity of the WBSS, as assessed by the AUROC, at predicting the need for respiratory support. They additionally reported statistically significant higher median WBSS in those needing respiratory support, and those on nasal continuous positive airway pressure versus those on high-flow nasal cannula. Kubota et al. [37] found that the WBSS had a

moderate discriminative validity at differentiating among those hospitalised who required respiratory support. They additionally reported that the median WBSS score among those hospitalised who required respiratory support was modestly statistically significantly higher. Jacob et al. [35] reported that the WBSS was moderately associated with nasogastric tube feeding according to its OR, but this result was not statistically significant (i.e. $p>0.01$). They also reported that the WBSS did not significantly predict desaturation days during hospitalisation. Somech et al. [49] reported statistically significant differences in the mean WBSS among those who were ambulatory, hospitalised and admitted to the PICU.

Balaguer et al. [21] found that the BROSJOD score had a moderate validity, as assessed by its volume under the surface (VUS), at discriminating by expert classification at admission, and a high validity after 24 and 48 hours. They also found statistically significant associations between the score & hospital length of stay (LOS), PICU LOS and need for invasive mechanical ventilation; however, they found no association with need for non-invasive ventilation. Broadly consistent with these findings, Ricart et al. [44] found large statistically significant differences in the mean LOS, days of oxygen therapy, days of nasogastric tube feeding and maximum mean fraction of inspired oxygen (F_IO_2) among those with a more severe BROSJOD score. There were also large statistically significant differences in the percentage of those with a more severe BROSJOD score who were admitted to the PICU or required ventilation. Also, Rodriguez-Gonzalez et al. [46] reported that the BROSJOD score had a moderate ability at discriminating by need for respiratory support, but did not significantly correlate with PICU admission. Granda et al. [34] reported that the BROSJOD score had moderate ability at predicting of any admission, need for supplemental oxygen, PICU admission within the next 48 hours or death.

Bueno-Campaña et al. [22] found that a high WDF was moderately correlated with the need for respiratory support as assessed by its relative risk. Granda et al. [34] found the WDF to have a moderate discriminative ability for predicting for a range of relevant outcomes. Similarly, Rivas-Juesas et al. [45] reported that the WDF & ESBA at admission both had a moderate ability at discriminating between those classified as severe and non-severe. They also found the mean WDF & ESBA score at admission in the severe and non-severe group to be statistically significantly higher. However, Ramos-Fernández et al. [43] reported that the ESBA score at admission only had a poor ability at discriminating by admission to the PICU, but that the highest ESBA was highly discriminative.

Caserta et al. [23] reported a high discriminative validity of the GRSS, as assessed by its AUROC, at predicting admission and similar results when a sub-group analysis was conducted in those ≤3 & 3-10 months. Unfortunately,

however, they didn't report the CIs. They also found statistically significant difference in mean GRSS among those admitted to the PICU and those hospitalised but not admitted to the PICU. When externally validated by Kubota et al. [37], they found that the GRSS (as well as the WBSS) had a moderate discriminative validity at differentiating among those hospitalised who required respiratory support. They additionally reported that the median GRSS (and WBSS) score among those hospitalised who required respiratory support was modestly statistically significantly higher. Similarly, De Rose et al. [26] reported a strong discriminative validity of the GRSS at predicting the need for respiratory support; however they also found that the median GRSS of those needing nasal continuous positive airway pressure versus high-flow nasal cannula were statistically insignificant.

McCallum et al. [39] reported the mTal had a low-moderate discriminative ability as measured by the point estimate of the AUROC at predicting oxygen need at 12 hours and 24 hours; however, the confidence intervals (CIs) of the AUROCs are so wide, we ignored their results. When externally validated by Golan-Tripto et al. [33] it was found that it had overall a moderate discriminative validity at differentiating based on need for oxygen support and hospital LOS ≥72 hours. Notably, the discriminative validity for oxygen support (but not hospital LOS) was statistically significantly higher among those with greater experience. Similarly Granda et al. [34] found mTal to have a moderate ability for predicting for a range of relevant outcomes.

Chong et al. [24] reported that the mRIS, a modified version of the Tal score (albeit different from the modified Tal score [mTal]) had a fair ability at discriminating between those who required non-invasive respiratory support, but a poor ability at discriminating by admission, intravenous hydration and LOS≥ 2 days. Another publication [25] using a subset of the same dataset similarly reported a poor ability of the mRIS at discriminating by admission.

Freire et al. [30] reported that their model had a moderate ability at discriminating among those hospitalised who required escalated care and those who didn't; the performance was similar when internally validated using bootstrap validation. External validation by Granda et al. [34] similarly found moderate ability of the Freire's for predicting for a range of relevant outcomes. When a modified version of Freire's model was evaluated by Raita et al. [42], it was found to have a low ability at discriminating by positive pressure ventilation and intensive treatment use. Raita et al. [42] also reported validity data for the 4 machine learning models they developed; all of the models had moderate discriminative ability at discriminating by positive pressure ventilation use and intensive treatment use.

Duarte-Dorado et al. [28] reported statistically significant, albeit modest, differences in median mWCAS among patients at admission and discharge, and those hospitalised who required admission to the PICU. Granta et al. [34] reported that the mWCAS, as assessed by AUROC, had a moderate ability at differentiating for a range of relevant outcomes.

Abbate et al. [18] reported a statistically significant weak correlation between the Modified Wang Bronchiolitis Severity Score and LOS. Amat et al. [19] reported that the Wainwright severity score on admission had a moderate association with hospitalisation (assessed using an unadjusted OR) and that those admitted to the PICU had a statistically significantly higher severity score compared to those hospitalised but not admitted to the PICU. Univariate analysis also identified a correlation with need for intensive care (but the magnitude was not reported) but not with LOS. De Rose et al. [26] reported a strong discriminative validity of the KRS at predicting the need for respiratory support. Destino et al. [28] reported a low discriminative ability, as assessed by its AUROC, for the Children's Hospital of Wisconsin Respiratory score (CHWRS) and RDAI at predicting admission. Garcia- Mauriño et al. [32] reported fair discriminative validity of the Clinical Disease Severity Score (CDSS) at predicting admission, need for oxygen, need for positive pressure ventilation and, PICU admission. Granda et al. [34] reported that the RSS, RCS, RS, and BRAS had moderate ability at differentiating for a range of outcomes with no significant difference between the different scores. Krishna et al. [36] reported a statistically significant association between the BSS and the type of respiratory support as well as significant differences in the heart rate and oxygen saturation between those classified as mild or moderate based on the BSS score. Özkaya et al. [41] reported that mBSS, a modified version of the WBSS, was moderately associated with admission, as assessed by the AUROC.

Convergent Validity

17 studies assessed convergent validity; only the mTal, BROSJOD, WBSS & GRSS score were assessed more than once.

El Basha et al. [29] found a strong correlation, as measured by the Spearman's correlation coefficient, between the mTal & the duration of oxygen therapy; the correlation was statistically significantly stronger in term infants compared to pre-term infants. Golan-Tripto et al. [33] found the mTal to moderately correlate with duration of oxygen therapy, and hospital LOS, but also reported significant variation by clinical severity. However, McCallum et al. [39] reported only a weak correlation between the mTal score and hospital LOS.

Anıl et al. [20] reported that WBSS moderately correlated with hospital LOS whereas DeRose et al. [26] reported only a very weak correlation between WBSS (as well as KRS) and LOS. Jacob et al. [35] reported that the WBSS was the greatest predictor of hospital LOS however a quantitative measure of its predicative ability was not reported; regardless this finding was overall insignificant (i.e. $p>0.01$).

Caserta et al. [22] found the GRSS to be moderately correlated with hospital LOS whereas DeRose et al. [26] found them to be very weakly correlated.

Balaguer et al. [21] also reported that Wood Downe's score strongly correlated with the BROSJOD score at admission, 24 hours and 48 hours. They also reported that it significantly correlated with hospital and PICU LOS although the magnitude was not reported. Rodriguez-Gonzalez et al. [46] found the BROSJOD score to be moderately correlated with hospital LOS and duration of respiratory support, but to not correlate with PICU LOS.

Abbate et al. [18] reported a significant weak correlation coefficient between the Modified Wang Bronchiolitis Severity Score and LOS. Amat et al. [19] reported that the initial Wainwright severity score was not significantly correlated with hospital LOS on univariate analysis. Destino et al. [27] found both the CHWRS & RDAI at admission to not correlate with LOS. Duarte-Dorado et al. [28] found the mWCAS and Tal score to be strongly correlated at both admission and discharge. Marguet et al. [38] found the CAS to be only weakly correlated with hospital LOS. Rivas-Juesas et al. [45] found the ESBA & WDF scores to be weakly correlated with each other. Siraj et al. [48] reported that the BSS was not correlated with hospital LOS, weight-adjusted high-flow nasal canula flow rate or duration of high-flow nasal canula therapy. McGinley et al. [40] reported that the ReSVinet score was positively correlated with PICU admission, mechanical ventilation, hospitalization and respiratory support requirement; however did not numerically report the magnitude of the association.

Criterion-concurrent Validity

Only 4 studies assessed criterion-concurrent validity. Balaguer et al. [21] reported a strong correlation, unusually assessed via the Kappa index, between the BROSJOD score & expert opinion at admission, 24 hours and 48 hours. Gal et al. [31] reported that the mRDAI was correlated with $PtcCO_2$; this correlation remained after controlling for $PvCO_2$ and weight. Shete et al. [47] reported the mTal score to be strongly correlated with oxygen saturation.

Krishna et al. [36] reported that the BSS was significantly associated with the Lung Ultrasound Score but did not report the magnitude.

TRIPOD: Quality of reporting

The quality of reporting of the included papers, as assessed by the TRIPOD score of the included articles, was poor; the mean TRIPOD score was 52% (see *Supplementary Table 2* for overall TRIPOD scores, and *Annex 2* for detailed TRIPOD scores). The reporting of model calibration, information around missing data, and summary characteristics of candidate predictors/score parameters was particularly poor.

PROBAST: Risk of Bias & Applicability

The overall risk of bias & applicability classifications, as assessed using the PROBAST framework, for each included paper is listed in *Supplementary Table 4 (*see *Annex 3* for detailed PROBAST scores*)*. All of the included papers had either serious methodological issues, most commonly in their analysis, or a poor quality of reporting so that a judgement of the quality couldn't be made. The major methodological issues were small sample sizes, specifically with the datasets including few participants with the outcomes being predicted for, and as noted above, lack of sufficient reporting of calibration measures, quantity of missing data and, procedures for missing data.

Discussion

We identified 31 unique scores from 32 articles and found that none of the identified scores were sufficiently validated. Across all three domains, the most promising score was the BROSJOD score, however it does require further validation. The mTal score was the next best validated score. It is relevant to note the high degree of similarity in the parameters in these two scores. The methodological quality of all the included studies and the quality of reporting, systematically assessed using the PROBAST and TRIPOD checklists, respectively, was poor. The most commonly used score, the RDAI score, had very weak discriminative ability (borderline poor) and only weak convergent-criterion validity; we do not recommend further effort being taken to validate this score or its use.

Our finding that there is no sufficiently validated score is consistent with all of the previous reviews. The most promising scores we identified, namely BROSJOD & mTal, were similarly identified by Hakizimana et al. [7]; they, however, also concluded that the Tal score and the Liverpool Infant Bronchiolitis Severity Score (LIBSS) (see below) were promising. In comparison to Bekhof, Reimink and Brand's [3], and Rodríguez-Martínez, Sossa-Briceño and Nino's [6] review we included far fewer papers (and scores); the former included 60 articles (36 scores) and latter included 77 articles (32 scores) whereas, as mentioned above, we included 31 articles (32 scores). This was primarily due to our more stringent inclusion criteria and our specific focus only on validity data rather than data reporting on the responsiveness, usability or reliability of the scores. In contrast, however, we included more than three times the number of papers included by Hakizimana et al.'s rapid review [7] and Luarte-Martínez et al.'s systematic review [5]. Our findings on the geographic distribution of the data sources used to validate these scores concurs with the findings of Hakizimana et al. [7], namely that the vast majority of these validation efforts were conducted in high-income countries. However, the best validated scores identified above seem feasible to implement in low-resources settings.

During the course of our searches, an additional promising score the LIBSS was identified, , but unfortunately no studies evaluating its validity met our inclusion criteria. The LIBSS was developed as a part of a PhD dissertation based on a comprehensive literature review, consultations with stakeholders, Delphi exercise and usability assessment, and then subsequently validated in a multicentre (n=11) prospective cohort study but no peer-reviewed full-text article reporting on the results of the validation study was identified [50].

There are some limitations of this review. The major limitations of our review were the restriction of included papers to only those published in English and not searching the grey literature; this likely means that some relevant papers may not have been included.

Further research is required to externally validate the BROSJOD, mTal, & LIBSS scores, ideally in low-income countries, and in primary care settings. The study designs should be guided by the PROBAST checklist or other similar tools, and report their findings in accordance with the TRIPOD checklist or other similar tools to ensure the studies are both well designed and communicated. Given that there are a number of promising scores, the scientific community should initially focus on validating or improving these scores and only, if necessary, work on proposing new scores. Additionally, ideally when assessing the validity of these scores, it would be useful if analyses were also done with a threshold on the time of the outcome assessment (e.g. discriminative validity of a score at predicting ICU admission within 24 hours of taking the score), as the course of the disease is not always linear and may lead to systematic under estimation or overestimation of the actual validity of the score.

References

1. Borchers AT, Chang C, Gershwin ME, Gershwin LJ. Respiratory Syncytial Virus—A Comprehensive Review. Clinical Reviews in Allergy & Immunology [Internet]. **2013**; 45(3):331–379. Available from: https://link.springer.com/article/10.1007%2Fs12016-013-8368-9
2. Li Y, Wang X, Blau DM, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. Lancet (London, England) [Internet]. **2022**; 399(10340):2047–2064. Available from: https://pubmed.ncbi.nlm.nih.gov/35598608/
3. Bekhof J, Reimink R, Brand PLP. Systematic review: Insufficient validation of clinical scores for the assessment of acute dyspnoea in wheezing children. Paediatric Respiratory Reviews. 2014; 15(1):98–112.
4. Streinmer DL. A Checklist for Evaluating the Usefulness of Rating Scales. The Canadian Journal of Psychiatry. **1993**; 38(2):140–148.
5. Luarte-Martínez S, Rodríguez-Núñez I, Astudillo P, Manterola C. Psychometric properties of scales used for grading the severity of bronchial obstruction in pediatrics: A systematic review and meta-analysis. Archivos Argentinos de Pediatria. **2017**; 115(03).
6. Rodríguez-Martínez CE, Sossa-Briceño MP, Nino G. Systematic review of instruments aimed at evaluating the severity of bronchiolitis. Paediatr Respir Rev [Internet]. **2018**; 25:43–57. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5557708/
7. Hakizimana B, Saint G, Miert C van, Cartledge P. Can a Respiratory Severity Score Accurately Assess Respiratory Distress in Children with Bronchiolitis in a Resource-Limited Setting? J Trop Pediatr [Internet]. **2020** [cited 2022 Feb 23]; 66(2):234–243. Available from: https://academic.oup.com/tropej/article/66/2/234/5570307#:~:text=Though%20scores%20should%20measure%20what
8. Roberts JN, Graham BS, Karron RA, et al. Challenges and opportunities in RSV vaccine development: Meeting report from FDA/NIH workshop. Vaccine [Internet]. **2016** [cited 2022 Feb 23]; 34(41):4843–4849. Available from: https://pubmed.ncbi.nlm.nih.gov/27566900/
9. Mazur NI, Higgins D, Nunes MC, et al. The respiratory syncytial virus vaccine landscape: lessons from the graveyard and promising candidates. Lancet Infect Dis. [Internet]. **2018**; 18(10):e295–e311. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1473309918302925

10. Öner D, Drysdale SB, McPherson C, et al. Biomarkers for Disease Severity in Children Infected With Respiratory Syncytial Virus: A Systematic Literature Review. J Infect Dis. [Internet]. **2020** [cited 2022 Feb 23]; 222(Suppl 7):S648–S657. Available from: https://pubmed.ncbi.nlm.nih.gov/32794555/
11. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available from: www.covidence.org.
12. Sheikh Z, Potter E, Li Y, et al. Validity of clinical severity scores for respiratory syncytial virus: a systematic review - data extraction sheet [Internet]. Edinburgh DataShare. University of Edinburgh; 2023 [cited 2023 Mar 24]. Available from: https://datashare.ed.ac.uk/handle/10283/4804
13. World Bank. World Bank Country and Lending Groups – World Bank Data Help Desk [Internet]. Worldbank.org. 2022. Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
14. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. **2019**; 170(1):51.
15. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Ann Intern Med. **2015**; 162(1):55.
16. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. **2015**; 162(1):W1.
17. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. **2019**; 170(1):W1.
18. Abbate F, Depietri G, Tinelli C, et al. Impact of the publication of the Italian guidelines for bronchiolitis on the management of hospitalized children in Pisa, Italy. Pediatr Pulmonol. **2023**; 58(8):2267–2274.
19. Amat F, Henquell C, Verdan M, Roszyk L, Mulliez A, Labbé A. Predicting the severity of acute bronchiolitis in infants: Should we use a clinical score or a biomarker? J Med Virol. **2014**]; 86(11):1944–1952.
20. Anıl M, Göç Z, Avcı R, et al. B-type natriuretic peptide is a useful biomarker predicting disease severity in children with isolated bronchiolitis in the emergency department. Turk J Pediatr. **2017**; 59(5):561.

21. Balaguer M, Alejandre C, Vila D, et al. Bronchiolitis Score of Sant Joan de Déu: BROSJOD Score, validation and usefulness. Pediatr Pulmonol. **2016**; 52(4):533–539.
22. Bueno-Campaña M, Sainz T, Alba M, et al. Lung ultrasound for prediction of respiratory support in infants with acute bronchiolitis: A cohort study. Pediatr Pulmonol. **2019**; 54(6):873–880.
23. Caserta MT, Qiu X, Tesini B, et al. Development of a Global Respiratory Severity Score for Respiratory Syncytial Virus Infection in Infants. J Infect Dis. **2017**; 215(5):750–756.
24. Chong S-L, Teoh OH, Nadkarni N, et al. The modified respiratory index score (RIS) guides resource allocation in acute bronchiolitis. Pediatr Pulmonol. **2017**; 52(7):954–961.
25. Chong S-L, Lai OF, Castillo L, et al. Nasal high-mobility group box 1 and caspase in bronchiolitis. Pediatr Pulmonol. **2018**; 53(12):1627–1632.
26. De Rose DU, Maddaloni C, Martini L, Braguglia A, Dotta A, Auriti C. Comparison of three clinical scoring tools for bronchiolitis to predict the need for respiratory support and length of stay in neonates and infants up to three months of age. Front Pediatr. **2023**; 11.
27. Destino L, Weisgerber MC, Soung P, et al. Validity of respiratory scores in bronchiolitis. Hosp Pediatr. **2012**; 2(4):202–209.
28. Duarte-Dorado DM, Madero-Orostegui DS, Rodriguez-Martinez CE, Nino G. Validation of a scale to assess the severity of bronchiolitis in a population of hospitalized infants. J Asthma. **2013**; 50(10):1056–1061.
29. El Basha NR, Marzouk H, Sherif MM, El Kholy AA. Prematurity, a significant predictor for worse outcome in viral bronchiolitis: a comparative study in infancy. J Egypt Public Health Assoc. **2019**; 94:15.
30. Freire G, Kuppermann N, Zemek R, et al. Predicting Escalated Care in Infants With Bronchiolitis. Pediatrics. **2018**; 142(3):e20174253.
31. Gal S, Riskin A, Chistyakov I, Shifman N, Srugo I, Kugelman A. Transcutaneous PCO_2 monitoring in infants hospitalized with viral bronchiolitis. Eur J Pediatr. **2014**; 174(3):319–324.
32. Garcia-Maurino C, Moore-Clingenpeel M, Thomas J, et al. Viral Load Dynamics and Clinical Disease Severity in Infants With Respiratory Syncytial Virus Infection. J Infect Dis. **2018**; 219(8):1207–1215.
33. Golan-Tripto I, Goldbart A, Akel K, Dizitzer Y, Novack V, Tal A. Modified Tal Score: Validated score for prediction of bronchiolitis severity. Pediatr Pulmonol. **2018**; 53(6):796–801.
34. Granda E, Urbano M, Andrés P, Corchete M, Alfredo Cano Garcinuño, Velasco R. Comparison of severity scales for acute bronchiolitis in real clinical practice. Eur J Pediatr. **2023**; 182(4):1619–1626.

35. Jacob R, Bentur L, Brik R, Shavit I, Hakim F. Is capnometry helpful in children with bronchiolitis? Respir Med. **2016**; 113:37–41.
36. Krishna D, Khera D, Toteja N, et al. Point-of-Care Thoracic Ultrasound in Children with Bronchiolitis. Indian J Pediatr. **2022**; 89(11):1079–1085.
37. Kubota J, Hirano D, Okabe S, et al. Utility of the Global Respiratory Severity Score for predicting the need for respiratory support in infants with respiratory syncytial virus infection. PLoS One. **2021**;16(7):e0253532.
38. Marguet C, Lubrano M, Gueudin M, et al. In Very Young Infants Severity of Acute Bronchiolitis Depends On Carried Viruses. PLoS One. **2009**; 4(2):e4596.
39. McCallum GB, Morris PS, Wilson CC, et al. Severity scoring systems: are they internally valid, reliable and predictive of oxygen use in children with acute bronchiolitis? Pediatr Pulmonol. **2013**; 48(8):797–803.
40. McGinley JP, Lin GL, Öner D, et al. Clinical and Viral Factors Associated With Disease Severity and Subsequent Wheezing in Infants With Respiratory Syncytial Virus Infection. J Infect Dis. **2022**; 226(Supplement_1):S45–S54.
41. Özkaya AK, Yilmaz HL, Kendir ÖT, Gökay SS, Eyüboğlu İ. Lung Ultrasound Findings and Bronchiolitis Ultrasound Score for Predicting Hospital Admission in Children With Acute Bronchiolitis. Pediatr Emerg Care. **2018**; 36(3):p e135-e142.
42. Raita Y, Camargo CA, Macias CG, et al. Machine learning-based prediction of acute severity in infants hospitalized for bronchiolitis: a multicenter prospective study. Sci Rep. **2020**; 10(1):10979.
43. Ramos-Fernández JM, Piñero-Domínguez P, Abollo-López P, et al. Validation study of an acute bronchiolitis severity scale to determine admission to a paediatric intensive care unit. An Pediatr. **2018**; 89(2):104–110.
44. Ricart S, Marcos MA, Sarda M, et al. Clinical risk factors are more relevant than respiratory viruses in predicting bronchiolitis severity. Pediatr Pulmonol. **2012**; 48(5):456–463.
45. Rivas-Juesas C, Rius Peris JM, García AL, et al. A comparison of two clinical scores for bronchiolitis. A multicentre and prospective study conducted in hospitalised infants. Allergol Immunopathol. **2018**; 46(1):15–23
46. Rodriguez-Gonzalez M, Rodriguez-Campoy P, Estalella-Mendoza A, Castellano-Martinez A, Flores-Gonzalez JC. Characterization of Cardiopulmonary Interactions and Exploring Their Prognostic Value in Acute Bronchiolitis: A Prospective Cardiopulmonary Ultrasound Study. Tomography. **2022**; 8(1):142–157.

47. Shete S, Nagori G, Nagori P, Hamid M. Relation between pulse oximetry and clinical score in infants with acute bronchiolitis. Natl J Physiol Pharm Pharmacol. **2014**; 4(2):124.
48. Siraj S, Stark W, McKinley SD, Morrison JM, Sochet AA. The bronchiolitis severity score: An assessment of face validity, construct validity, and interobserver reliability. Pediatr Pulmonol. **2021**; 56(6).
49. Somech R, Tal G, Gilad E, Mandelberg A, Tal A, Dalal I. Epidemiologic, socioeconomic, and clinical factors associated with severity of respiratory syncytial virus infection in previously healthy infants. Clin Pediatr. **2006**; 45(7):621–627.
50. Miert C van. Measuring Clinical Severity in Infants with Bronchiolitis [Internet]. **2015** [cited 2022 Aug 20]. Available from: https://livrepository.liverpool.ac.uk/2037906/1/vanMiertCla_June2015_2037906.pdf

PROMISE Investigators:

Harish Nair, Harry Campbell, Richard Osei-Yeboah (University of Edinburgh); John Paget (NIVEL); Philippe Beutels (Universiteit Antwerpen); Anne Teirlinck (RIVM); Hanna Nohynek (THL); Louis Bont (University Medical Center Utrecht); Andrew Pollard (University of Oxford); Peter Openshaw (Imperial College London); You Li (Nanjing Medical University); Jeroen Aerssens, Gabriela Ispas (Janssen); Veena Kumar (Novavax); Tin Tin Htar, Elizabeth Begier, Jessica Atwell (Pfizer); Charlotte Vernhes, Rolf Kramer, Mathieu Bangert (Sanofi Pasteur); Gaël Dos Santos, Rachel Cohen, Theo Last (GSK); Bahar Ahani (AstraZeneca); Nuria Machin (TeamIT).

Author Contributions

HN conceived the idea and served as third person arbitrator. ZS & EP conducted the review. ZS authored the manuscript. RAC authored the lay summary. EP, YL, RAC, GDS, LB & HN commented critically on several drafts of the manuscript. PROMISE investigators reviewed the manuscript prior to submission.

Conflict of interest statement

None for ZS or EP.

YL has received funding from Wellcome Trust and GSK outside the submitted work. YL received personal fees from Pfizer.

RAC is an employee of the GSK group of companies, holds shares in the GSK group of companies, and has received other compensation from GSK outside the submitted work.

GDS is an employee of GSK group of companies and hold shares as part of his annual remunerations

LB has received funding through UMC Utrecht from AbbVie, Janssen, the Bill and Melinda Gates Foundation, Nutricia Danon, MeMed Diagnostics, GSK, Novavax, AstraZeneca, Sanofi, Ablynx, Bavaria Nordic, MabXience, Novavax and Pfizer.

HN has received funding from IMI, NIHR and Pfizer. HN received fees from GSK, AbbVie, AZ, Novavax, Sanofi, Merck and ReViral.

Funding statement

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101034339. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This publication only reflects the author's view and the JU is not responsible for any use that may be made of the information it contains herein.

Corresponding author contact information

Professor Harish Nair

566 Usher Institute
567 University of Edinburgh
568 Old Medical School
569 Teviot Place
570 Teviot Place,
571 Edinburgh
572 EH8 9AG, UK
573 Harish.Nair@ed.ac.uk