Author's response to reviews

Title: How well do clinical prediction rules perform in identifying serious infections in acutely ill children across an international network of urgent-access datasets?

Authors:

Jan Y Verbakel (jan.verbakel@med.kuleuven.be)
Ann Van den Bruel (ann.vandenbruel@phc.ox.ac.uk)
Mathew J Thompson (matthew.thompson@phc.ox.ac.uk)
Richard Stevens (richard.stevens@phc.ox.ac.uk)
Bert Aergeerts (Bert.Aertgeerts@med.kuleuven.be)
Rianne Oostenbrink (r.oostenbrink@erasmusmc.nl)
Henriette A Moll (h.a.moll@erasmusmc.nl)
Marjolein Y Berger (m.y.berger@umcg.nl)
Monica Lakhanpaul (m.lakhanpaul@ucl.ac.uk)
David Mant (david.mant@phc.ox.ac.uk)
Frank Buntinx (Frank.Buntinx@med.kuleuven.be)

Version: 2 Date: 11 August 2012

Author's response to reviews: see over
Dear Editor,

Responding to your invitation, we are re-submitting a completely revised version of the enclosed manuscript: “How well do clinical prediction rules perform in identifying serious infections in acutely ill children across an international network of urgent-access datasets?” for consideration by BMC Medicine.

We thank you and the referees for the useful comments we received concerning our previous submission (MS: 1355671679734568) and following intensive discussion among the authors and revisions to the content and presentation of the manuscript, we believe we have addressed all the queries by the referees, and now have a much improved and more clinically relevant paper.

Please find attached our answers to the queries and a complete overview of the revisions made.

We still strongly believe that this work is of interest to a wide international audience. Serious infections in children are responsible for important childhood morbidity and mortality around the world. The diagnostic challenge of serious infections lies in the potential rapid evolution, the non-specific presentation early in the course of the illness and the low incidence, especially in economically developed countries, where serious infection is relatively uncommon. Clinical prediction rules may improve their early recognition.

In most cases this diagnostic challenge is faced in ambulatory care settings. The difficulty of diagnosis in these settings is underlined by previous research published by the co-authors in The Lancet showing that (1) up to half of children with meningococcal disease are missed at first consultations with GPs in the UK and (2) the value of clinical features, serving as red flags for serious infection. Clinicians need to make quick decisions generally based solely on presenting clinical features. However, research data to inform these decisions are scarce and data to support which clinical prediction rule to choose are rare.

Our paper is unique, because it is a validation in multiple settings, multiple countries through a collaborative network and illustrates an innovative way of presenting the results. We present the validation results of 4 clinical prediction rules and 2 guidelines in 7 datasets, providing individual patient data of over 11,000 children. The findings of our paper are likely to have implications for modifying current clinical practice and national guidelines for assessing children with acute infections. We provide data to back up the diagnostic value of currently known guidelines for feverish children, and indicate which existing clinical prediction rules should now be incorporated into everyday practice.

None of the checklists (PRISMA, STARD, etc.) are applicable to this research paper, due to its unique methodology, and therefore, none are available as supplemental files in the submission. There was no formal protocol available to include in the submission.

The corresponding author states that he had full access to all the data in the study. All of the authors have given their consent for submission to BMC Medicine. Finally, we have no commercial associations with impact on this work.

JV undertook the translation, synopsis, recoding and data checking and the methodology and results of each step were thoroughly discussed with all primary study authors (AVdB, MB, RO, HM, MT, ML). All authors believe the manuscript represents honest work, and met the requirements for authorship.

Yours sincerely,

Dr Jan YJ Verbakel
Department of General Practice – KU Leuven
Kapucijnenvoer 33 – Blok J
3000 Leuven – Belgium
jan.verbakel@med.kuleuven.be
Revisions Report

Revisions to the manuscript are marked in “track changes”

- **Referee 1:**

  **Major compulsory revisions:**
  - **Comment 1:** "Figure 1b must be revised."
  
  **Response:** In accordance with the referee’s comment, comment 6 of referee 2 and comment 17 of referee 3 we have eliminated Figure 1 (ROC points estimates of the external validation results) to improve overall clarity throughout the paper.

  **Minor compulsory revisions:**
  - **Comment 2:** Abstract, Background, last sentence: "validate" for "validation"
  - **Comment 3:** Abstract, Methods, second sentence: "when necessary" for "where necessary".
  - **Comment 4:** Introduction, first paragraph: species names must be not capitalized (influenzae, pneumoniae).
  - **Comment 5:** Methods, first paragraph: In reference to the excellent Van den Bruel paper (Lancet 2010), it is probably better to refer to it as a “systematic review” rather than a “meta-analysis”.
  - **Comment 6:** Methods, second paragraph: Studies were “identified” by the systematic review, rather than “included” in it.

  **Response to comment 2 to 6:** We agree with the referee’s comments and have edited the wording accordingly.

  **Discretionary Revisions:**
  - **Comment 7:** Despite limitations of using datasets are mentioned in the "Discussion", authors could remark the difference with extensive prospective validation and impact analysis as the appropriate methods to test the real usefulness of these tools.

  **Response:** We have now added a paragraph in the discussion section (p. 14): "This illustrates the clear need to perform extensive prospective validation and impact analysis of clinical prediction rules prior to clinical implementation”

- **Referee 2:**

  **Major compulsory revisions:**
  - **Comment 1:** There is no flow sheet describing which databases were excluded. The authors state that "a recent meta-analysis identified 7 clinical trials", where in fact, that meta-analysis included 30 trials. Please add a flow diagram so that it is clear to readers that of the 30 eligible studies, only a small number were included. Also, revise the statement above (in abstract and Background section) that implies that there were only 6 eligible studies.

  **Response:** The referee is right to point out that there was no flowchart available. We have created one, describing the process of dataset inclusion and have provided this as a figure in the manuscript (Figure 1: Flowchart of dataset inclusion).

  - **Comment 2:** This is a validation study. There is no need for the label “individual patient data analysis.” It can easily be confused with an individual patient data meta-analysis. Please delete this from the title.

  **Response:** We have adjusted the title accordingly.

  - **Comment 3:** I understand the importance of looking at purely clinical decision rules (i.e., no labs are included as part of the decision rule). But in common parlance, many “clinical” decision rules include lab values. It is important that the authors discuss the importance (and the limitations) of looking at purely clinical decision rules. 1-month-old febrile infants, for example, will almost always have some form of
laboratory assessment in the US. Thus, is the accuracy of purely clinical decision rules relevant? I think so, but this needs to be clearly discussed. And in the discussion, it would be nice to compare the accuracy of purely clinical versus those that clinical + lab decision rules.

**Response:** We agree with the referee that many decision rules used in clinical practice, especially in emergency departments, include laboratory markers. From the point of view of first-contact consultations (e.g., primary care, urgent-access ambulatory care) where a fast decision is warranted (whether to refer or not, whether to hospitalise or not) laboratory values are rarely available in most European (and many other international settings), so prediction rules based solely on clinical signs and symptoms are typically all we have at our disposal. These prediction rules often tend to have low specificities, resulting in a high rate of false positives, which could be remedied by additional laboratory or point-of-care testing. In terms of ruling out serious infection in children, these rules usually have a high sensitivity, which is the primary aim when evaluating the possibly seriously ill child. We have limited our analyses to prediction rules, based on signs and symptoms only, but, as suggested by the referee, we mention the usefulness of additional testing to further improve diagnostic precision and reduce the number of children falsely flagged as potentially having a serious infection in the discussion section (clinical implications p. 15):

“Clinicians should be aware that none of the prediction rules provide perfect discrimination, and it is perhaps unrealistic to expect such clinical rules to provide this. Residual uncertainty may be further improved by conducting more detailed clinical assessments, repeating the assessment after a period of time, using additional testing (e.g. urine or blood tests), and in most cases providing appropriate safety netting advice for children sent home detailing instructions on when to seek further care.”

**Comment 4:** There also needs to be a discussion of whether it is appropriate to validate the decision rules in the available datasets. Are all the datasets composed of children who are at risk for all 4 conditions of interest (serious infection, pneumonia, meningitis)? It wouldn’t make sense to validate the pneumonia rule in the study of children referred for pneumonia. Only 3 studies (24, 25, 26) were children referred to because of fever. Is it appropriate to assess a rule for SBI in young children in afebrile older children? I would prefer that the authors focus on 1 or 2 questions in this paper (validation of clinical SBI rules in datasets of children suspected of having SBI) and to exclude the info regarding the meningitis decision rule and the pneumonia decision rules in different papers.

**Response:** The referee is right to point out the importance of applying clinical prediction rules to the appropriate populations. Concerning the disease-specific rules (pneumonia- and meningitis-rules) we focused only on children with or without the target condition, leaving all children with outcomes other than the target condition in the analysis with the risk of deflating the rule’s performance in the validation datasets. We should clarify that all datasets used in the validation analyses included children who were at risk for the target condition of the clinical prediction rule at hand.

Some rules were specifically designed to detect a target condition in a certain age group (Yale Observation Scale in children aged 3 to 36 months, NICE feverish illness guideline in children under 5 years of age). We have therefore performed sensitivity analyses on the validation datasets, comparing the results of that particular age group to the entire range of age of the dataset at hand, to ensure the robustness of these results, but also report results for the age group the rule was intended for. (Results section last paragraph & Appendix 4)

In first-contact consultations, where we focus mainly on the acute illness of the child, the authors strongly believe that the priority is detecting whether a child has any serious infection and whether the child should be referred or hospitalised, rather than attempting to come to a firm final diagnosis. This requires tests with high sensitivity and/or low negative likelihood ratio. Therefore, we believe that a rule that encompasses all SBI and effectively rules out these conditions is of great value in these settings. We certainly acknowledge that for those children referred to hospital, and those admitted to inpatient beds, the focus changes to confirming or ruling in particular diagnoses, but this lays out with the focus of this paper.

**Comment 5:** The authors do a great job of calculating post-test probabilities and likelihood ratios. But in the abstract and discussion, there are ignored. The focus is on sensitivity. If the sensitivity of the test is what the authors believe is of primary importance (I do not), why do they include the likelihood ratios and probability values? This is important because it may influence the conclusions. Looking at the post-test probabilities, it appears that none of the decision rules consistently rule out SBI. Yet, in the 1st and 2nd paragraph of the discussion and the abstract, the authors, based on sensitivity of >90%, seem to provide support to the NICE, NHG, and 5-stage decision tree guidelines.

**Response:** We agree with the referee’s comment and have included the percentage of false positives (e.g. children identified by the prediction rules who would be referred or admitted) and added negative
likelihood ratios (LR-) to strengthen the results and conclusions throughout the paper. A negative likelihood ratio less than 0.2 indicate a strong rule out value.[1]

- **Comment 6**: Figure 1 is not helpful. Please eliminate it.

Response: We have eliminated the figure to maintain overall clarity throughout the paper.

- **Comment 7**: The authors mention they did sensitivity analysis, but this is not well explained in the methods, and the results are no included, and no discussion of it included.

Response: In accordance with the referee's comments, we have now updated the sensitivity analyses and added a description in the methods, results and discussion section, e.g. in the methods section (p. 7): "We performed the following sensitivity analyses. Firstly, when a CPR was specifically designed for a certain age group (e.g. YOS for children aged 3 to 36 months & NICE guidelines for children up to 5 years of age), we compared performance in target age group and the entire age range of the dataset at hand. Second, when one or more variables of the original prediction rule were missing, we examined performance in those datasets with no missing variables, to avoid biasing results on the amount of missing variables."

**Discretionary Revisions:**

- **Comment 8**: Delete the sentence "approximations were used when necessary" from the abstract. It just causes confusion.

Response: We agree with the referee's comment and have adjusted the abstract accordingly.

- **Comment 9**: Quality of written English.

Response: Native English speakers reviewed the text in detail and adjusted the paper according to the suggestions by reviewer 1.

- **Referee 3:**

**Major issues:**

- **Comment A**: The point of validation of prediction rules is to assure clinicians and policy planners that their performance under the conditions of the initial derivation will hold up in actual clinical practice and in a variety of clinical settings and populations. Many factors may alter the performance of such instruments, not the least of which is variability in how practitioners interpret and apply the clinical variables in the course of patient care. One important limitation of retrospective approaches to validation is that it entirely sidesteps the effects of that variability. One published guideline for evaluating the strength of evidence supporting the use of prediction rules considers that retrospective validations, even if done on independent data sets, are not sufficient to move a rule beyond the level represented by its initial derivation. Compounding this limitation of your approach, you have elected to draw on databases that originated from studies involving the use of different rules, with the result that you have had to use "approximations" and analogies to equate the variables included in the original rules and guidelines with data points in the respective databases. Finally, you have included validation data from the targeted databases even when substantial numbers of predictors were not collected, even by "approximation". These factors serve to decrease confidence that the results of your study are predictive of the potential performance of the respective rules and guidelines in actual practice. The reviewer has serious concerns regarding whether the methodology you have employed is capable of illuminating the practical usefulness of these guides.

Response: We thank the referee for this comment and acknowledge this is an important item to address. We agree that prospective validation of existing prediction rules is the gold standard to evaluate their diagnostic performance. However, we believe the variability in how practitioners interpret and apply these rules is not limited to prospective validation. Whilst we agree with the referee that our results should be considered with caution, we do not agree our analyses imitate split sample or bootstrap validations (level 4 clinical decision rules), as suggested by McGinn et al. [2], which indeed only consider internal validation or precision. The datasets used for this validation study all originally were designed to predict serious infections, using standardised clinical registration forms, collecting relevant information for diagnostic studies, at first-contact consultations.
These original data were used for our analyses, which we believe to highly contrast with retrospective studies selecting patients based on e.g. diagnostic registries or positive cultures, retrieving data from random patient records.

Additionally, our analyses include a prospective validation of the Yale Observation Scale in three datasets (mentioned in the results section), enforcing the robustness of these findings.

Finally, the NICE and NHG feverish illness guidelines never underwent formal prospective validation prior to implementation in the UK or The Netherlands, and were consensus-based in combination with assessment of the available evidence (which was scarce). The results of our paper provide evidence for the performance of the guidelines in relevant settings, and add significantly to the confidence in continuing to use and promote these guidelines. Although it would be theoretically preferable to perform prospective validation, the reality of clinical practice and research in this particular clinical area means that this is logistically extremely difficult. However, we agree that our results, based on these retrospective analyses, should be interpreted with caution, and have adjusted the wording of results and discussion sections accordingly.

We support the referee’s assertion that approximations could limit our results. Therefore, we elaborated on sensitivity analyses to avoid biased results, based on the amount of required variables. However, variation in how variables or clinical features are used is part of routine care, and the authors feel that this strengthens the generalizability.[3] For example, clinical prediction rules to detect meningitis often include “meningeal irritation”, which can be measured through different clinical tests, e.g. Kernig’s sign, Brudzinski’s test, Tripod sign or nuchal rigidity. This reflects clinical practice and the variability in how practitioners measure and apply a clinical variable in the course of patient care.

Furthermore, in comparison with the previous version of our paper, we tightened our inclusion criteria on the number of predictors that should be collected, in order to perform validation analyses, e.g. the validation results of the meningitis rule, now only include datasets where all predictor variables were recorded, and for the feverish guidelines, consisting of multiple items, we only considered datasets for the validation analyses if less than 1/3 of all variables were not collected.

- **Comment B**: Your methods specify inclusion of data corresponding to infants and children between the ages of 0 and 18 years. The average ages of subjects in the databases you have used are 2-5 years old in all but one of the studies. There are at least 3 distinct age groups recognized in the paediatric emergency literature encompassed by these ranges. Prediction instruments will be expected to perform differently across these age groups and, indeed, different rules, including some considered in your study, have been developed for use in only one of these ranges. For example, the Yale Observation Score (your reference #15) was developed for use in the very young age group, i.e. < 3 months of age. In fact, it is in the very young, 0-3 months, and young, 3-24 or 36 months, that prediction rules of this sort are considered potentially valuable to avoid substantial numbers of unnecessary workups for serious bacterial infections. By the time a child has reached their 3rd birthday, the need for such instruments is largely dissipated and routine clinical observation and evaluation is likely to be as or more accurate than any such rule, as well as more efficient. You have lumped together data drawn from subjects across your entire range of age eligibility. This is not valid and you will have to break your performance data down into appropriate age subgroups for it to become clinically meaningful. Specifically, inclusion of data on older children into late adolescence will tend to inflate the apparent accuracy of the rules you are testing by including subjects for whom such a rule is unnecessary.

**Response**: The referee is right to point out the distinct age groups in children and the expected difference in performance of clinical prediction rules in these age groups. However, in the clinical experience of the authors (all of whom are either primary care clinicians or paediatricians) we strongly disagree with the referee’s statement that there is no value in clinical prediction rules in children over 3 years of age. When a rule was specifically designed to detect a target condition in a certain age group (Yale Observation Scale in children aged 3 to 36 months, NICE feverish illness guideline in children under 5 years of age), we performed sensitivity analyses on the validation datasets, comparing the results of that particular age group to the entire range of age of the dataset at hand, to ensure the robustness of these validation results. (Appendix 4) We added a description of these analyses in the methods, results and discussion section. Although some of these rules are specifically designed for the very young infant, we didn’t find any difference when validating these rules in a broader age range. This is similar to the recommendations of the NICE guidelines, which included the Yale items in the traffic light system, extrapolating the use of the items up to the age of 5 years.[4]

**Subsidiary issues:**

- **Comment 3**: Under Methods, please make clear that this was a retrospective study in which the performance of selected rules and guides was assessed by applying them to data collected in connection with different instruments. Your assertion regarding specificity is erroneous. An instrument with high
sensitivity, if the specificity is comparably low, will fail to have impact on the likelihood of the condition being considered.

**Response:** We agree with the referee’s comment on the retrospective nature of the analyses and have adjusted the methods section, by eliminating the statement on specificity.

- **Comment 4:** Under Results: Here and elsewhere you dwell almost entirely on observed sensitivity of the different instruments. In fact, in the body of the manuscript you have reported specificities and also likelihood ratios for positive and negative assessments for all of the rules you tested, and have included a figure that illustrates the potential impact of the rules on pre-test-probability. Please revise the abstract in such a way that reflects the potential impact. In some places, the results summary appears to encompass contradictory statements, such as where you describe the sensitivity of the NICE guideline. Please revise for clarity.

**Response:** We have now included the percentage of false positives (e.g. children identified by the prediction rules that would be referred or admitted) and likelihood ratios (LR-) to strengthen the results and conclusions throughout the paper. We have adjusted the results section and the abstract accordingly. The results summary on the NICE guidelines has been adjusted to avoid confusion and maintain clarity.

- **Comment 5:** P. 4, Par. 2. Somewhere in the manuscript, please summarize the NICE and NHG guidelines in detail, as you have the 7 prediction rules.

**Response:** In accordance with the referee’s suggestion, we have now included a summary of the NICE and NHG guidelines in Appendix 1 (Details of the clinical prediction rules and guidelines identified in the systematic review).

- **Comment 6:** P. 5, Par. 1 Your assertion regarding the Yale Observation Scale is misleading. You imply that it has been “validated” in multiple studies. In fact, studies that directly examined its performance in the population for which it was designed rejected it as clinically unacceptable rather than validating it. A number of the other studies you cite in support of this assertion used the score for purposes of objectifying the stratification of subjects in studies of performance of procalcitonin and other biomarkers. They were neither designed for assessment of, nor did they report, performance data for the YOS itself. Please be more judicious in your use of the term ‘validate’ here and elsewhere in the manuscript. Also, considering that the YOS is commonly considered a failed rule for non-research purposes, please explain why you felt it pertinent to include it in your own study.

**Response:** We agree with the referee’s comment and have adjusted our assertion regarding the Yale observation Scale. We have included the following description in the discussion section (p. 14): “The YOS was initially developed to identify serious illness in febrile children aged 3 to 36 months, but subsequently discarded, based on three prospective validation studies (of which only 1 validated it in the intended age group),[5-7] and used to stratify patients in another 5 studies, evaluating the performance of inflammatory markers (e.g. procalcitonin, C-reactive protein), with discouraging results, concerning the use of the YOS in clinical practice.[8-12] Bang et al. found slightly better performance of the YOS to predict bacteraemia in febrile children, when a high prevalence of bacteraemia (28% in this study) is to be expected.[13] Due to the often-low prevalence of ambulatory care settings, these results do not support the widespread use of the YOS.”

Most of these studies consider the YOS as a failed rule for non-research purposes: We included the YOS in our analyses because it was identified by the systematic review. An additional reason was that the NICE guideline used several of the Yale items in the traffic light system of “alarm features” Our results confirm the overall low sensitivities of the Yale Observation Scale, although some of the clinical features it contains (colour, hydration status) could be useful red flags to rule in serious infections.

- **Comment 7:** P. 5, Par 2 and 3. Please be explicit regarding the nature of your access to the databases of the studies you have used for your retrospective study.

**Response:** Direct access to the raw data of each dataset was granted and the datasets were made available to the primary study author. We have formally stated this in the methods section: "Direct access to raw data of each dataset was granted and key characteristics of each of the datasets were extracted.”

- **Comment 8:** P. 6, Item number 3: This approach appears entirely arbitrary. Unless the missing predictors were of minor consequence in determining the performance of the respective rules, you cannot safely systematically ignore their absence from the data used in your study. Such a modification of the
rules, depending on how they are constructed, would likely tend to lower observed sensitivity and raise observed specificity. Please address these issues.

Response: The referee is right to point out that this approach could be judged as arbitrary. We included a sensitivity analyses when variables of the original prediction rule were missing, to avoid biasing the results on the amount of missing variables. (Appendix 4) Whenever more than one (clinical prediction rules) or more than two (fever guidelines) of the original variables were missing, a sensitivity analysis was deemed unsuitable and the study was not included in our main analysis. This was consensus-based, because the absence of 1 out of 6 variables for a clinical prediction rule and the absence of 1/3 of all variables for a fever guideline of up to 35 items, was judged to be acceptable and clinically sensible. For the meningitis and pneumonia rule (consisting of 2 or 3 variables), no missing variables were allowed. E.g. the sensitivity analysis on the meningitis rule, revealed different Areas under the curve when "nuchal rigidity" was not available, thus eliminating one dataset, available for validation.

We agree with the referee that a rule modification could lower the observed sensitivity and raise the observed specificity. However, this depends on how the rule is constructed, and apart from the Yale Observation Scale, which is a continuous score scoring 6 items, in all other rules and guidelines in these analyses is scored "if yes to any of these items". For the NICE feverish illness guidelines the items of the score are categorised in groups of 2 to 13 items. Per category they are scored "if yes to any of these items". Whenever an item is missing in rules scoring "if yes to any of these items", the sensitivity tends to be lower and the specificity tends to be higher. However our results, even after sensitivity analysis, maintain their robustness concerning the observed sensitivity and specificity. (Appendix 4)

- Comment 9: P. 7, Par. 3. The reviewer fails to follow the logic of this assertion. In high prevalence settings, a lower sensitivity would result in even more false negative assessments, increasing the clinical consequences of delayed diagnosis. A higher specificity, to be sure, would decrease the number of unnecessary workups. Are you saying that under such circumstances it is more important to lower costs than to save lives? In passing, you might take note of the fact that changes in prevalence do not alter performance characteristics of tests and diagnostic instruments unless increasing severity accompanies increased prevalence. In general, I advise that you concentrate more on measures of actual impact on probability, i.e. likelihood ratios, and observed post test probabilities of positive and negative assessments in describing your methods and reporting and interpreting your results.

Response: We have adjusted the paragraph according to the referee’s suggestions (elimination of the assertion on specificity). We confirm that the focus is more on sensitivity, rather than specificity to prevent missed diagnoses of serious illness (and save lives) in first contact settings as the main focus, instead of lowering the overall costs.

We have included the percentage of false positives (e.g. children identified by the prediction rules who would be referred or admitted) and added negative likelihood ratios (LR-) to strengthen the results and conclusions throughout the paper.

- Comment 10: P. 7, Par. 4: Please explain why and how your approach to the NICE and NHG related data differed from that used for the prediction rule related data.

Response: We agree with the referee’s comment and to avoid confusion, have eliminated this sentence to improve clarity concerning the NICE and NHG analyses. Our approach to the NICE and NHG guidelines did not differ from that used for the prediction rules.

- Comment 11: P. 12, Par 1-2: Presentation of quantitative results should be confined to the results section.

Response: We limited the results to the results section, although some assertions on the overall results remain.

- Comment 12: P. 12, Par 1: Regarding the assertion in the middle of this paragraph, your results are, strictly, confined to performance data on the rules and guides you considered as derived retrospectively from databases which were, by virtue of your approach, explicitly designed to test performance of different rules and guides. A sweeping statement such as this therefore requires an explicit elaboration of the embedded logic leading to its assertion.

Response: The referee is right to point out the need to clarify this assertion. We agree our results, based on these retrospective analyses, should be interpreted with caution, and have adjusted the discussion section accordingly with the appropriate wording and a more careful formulation (p. 12): "Assuming that our results reflect clinical practice, these simple rules which use clinical signs and symptoms, should perform similarly on new prospective data collection on similar populations. However, although 99 to
100% of all children testing negative on these rules can safely be sent home, specificities were low, which would result in high numbers of children flagged as potentially having a serious infection. Additional clinical assessment or additional testing or review at a later stage will be necessary to avoid inappropriate referrals or hospital admissions in this group of children.”

- **Comment 13:** P. 13, Par, 1: In line with an earlier comment, please revise the wording related to the YOS to reflect that, in fact, this rule was considered by published studies to have been invalidated and was subsequently rejected as a tool for informing clinical decision making.

Response: We have adjusted the description of the YOS accordingly. (See: response to comment 6).

- **Comment 14:** P. 14, Par. 2: Why have you bolded the NICE, NHG and 5-stage decision tree in the text. This suggests that perhaps your actual focus had more to do with the first two criterion sets, even though you have provided less documentation of the methods you brought to bear on their assessment.

Response: The referee is right to point out that these items should not be bolded in the text, we have removed the bolding. See also our response to your comment 6. Additionally, a summary description of these guidelines is added to Appendix 1.

- **Comment 15:** P. 14, Par. 3: Regarding the YOS, it would be more forthright to, firstly, report the actual history of this rule, secondly, reinforce the rationale, if any, that informed your decision to assess it using inherently limited methods and, thirdly, discuss any potential differences you found in its performance compared to prior published reports.

Response: We have adjusted the description of the YOS as described in our response to your comment 6.

- **Comment 16:** P. 15, Par 1: The reviewer is not convinced that your study or those that preceded it support the notion that further research is required to “identify predictors” of SBIs. Rather, if anything, to the extent that your findings suggest potentially valuable performances, your results might support the assertion that further research is needed to prospectively validate the performance and impact of the more promising prediction instruments in real world clinical settings. In general, your discussion section would constitute an ideal place to summarize the process through which a highly supported prediction instrument must pass on its way to achieving a status that merits recommendations for use in decision-making and health care policy. This process is well established in the literature.

Response: We have added a paragraph in the discussion section on the usefulness of extensive prospective validation and impact analysis prior to implementation. For the process of which a clinical prediction rule must go through, we refer to prior work, published by some of our co-authors.[14-16]

- **Comment 17:** Legends to Figure 1: The references here should be integrated into the bibliography and the corresponding citation numbers used in tables and figures.

Response: In accordance with the referee’s comment, comment 1 of referee 1 and comment 6 of referee 2 we have eliminated Figure 1 (ROC points estimates of the external validation results) to maintain overall clarity throughout the paper.

- **Comment 18:** Table 1, Row 1: It is implied, but you do not actually state, that you only considered prediction instruments that were based exclusively on clinical, as opposed to laboratory or imaging, criteria. If this were not the case, you would have had to consider a large number of additional prediction criteria, particularly in connection with the very young and young age groups. If this inference is correct, the reviewer wonders why you would list recording of laboratory criteria within your inclusion criteria. Please clarify this point, starting in your abstract and methods sections and make suitable adjustments to the working of this table.

Response: We agree with the referee’s comment and have adjusted table 1 accordingly. Also, throughout the paper we now included several statements that the prediction rules were based exclusively on clinical signs in a more explicit manner. (See also: response to comment 3 by referee 2)

- **Comment 19:** Table 3: The right hand side of this “Table” needs to be carved out as a separate Figure.

Response: We support the referee’s assertion that Table 3 and 4 need to be as clear as possible, although we believe presenting the results in this novel format (dumbbell plots) is innovative and allows
us to compare sensitivity/specificity or likelihood ratio pairs with observed prevalence, pre- en post-test-probability. Ranging the datasets from low to high prevalence presents the plots of pre- en post-test-probabilities in a gliding scale with pre-test risk increasing from left to right. This approach is identical to the tables used in the systematic review by Van den Brul et al., published in the Lancet.[17]

- **Comment 20**: Table 4: In Row 1, pertaining to the NICE guidelines, you report a specificity of >1 (100%). This is impossible. Please scrupulously review all the date you have reported in your tables and appendices. The right hand side of this "Table" is actually a "Figure" which is undecipherable in the form you have presented it. It needs to be carved out as an independent Figure and not included in this Table.

Response: We agree Table 4 needs to be as clear as possible and we have added the "percentage" sign to each item of the title bar of the table for sensitivity and specificity. We chose to present the results in percentage (e.g. specificity of 85%) instead of point numbers (e.g. specificity of 0.85) in both the text and the tables.

- **Comment 21**: Appendix 1: Please provide comparable information regarding the criterion derived from the practice guidelines you considered (NICE, NHG)

Response: In accordance with the referee's suggestion, we have now included a summary of the NICE and NHG guidelines in Appendix 1 (Details of the clinical prediction rules and guidelines identified in the systematic review). See also our response to your comment 5.

References: