PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	CADDIE2 - Evaluation of a clinical decision-support system for early detection of systemic inflammatory response syndrome in pediatric intensive care: study protocol for a diagnostic study
AUTHORS	Wulff, Antje; Montag, Sara; Steiner, Bianca; Marschollek, Michael; Beerbaum, Philipp; Karch, André; Jack, Thomas

VERSION 1 - REVIEW

REVIEWER	Zhivko Zhelev University of Exeter, UK
REVIEW RETURNED	04-Feb-2019

GENERAL COMMENTS	This is an interesting and, according to the authors, clinically relevant study that may help PICU physicians with early detection of SIRS and prompt initiation of evidence-based treatment. Although the protocol provides a comprehensive description of the background, rationale and the methods of the study, I feel that the following aspects need further clarification and detail. Outcomes and choice of study design As the authors point out, the study is not intended to evaluate user acceptability, or indeed the impact of implementing the system on patient outcomes. The study can only show if the system is accurate enough when compared to current practice (routine clinical assessment) against the reference standard of a chart review conducted by two independent blinded experts. It would be helpful for the readers to understand: • Why test accuracy and not more clinically relevant outcomes (e.g. patient outcomes assessed through an RCT) have been selected for the focus of this evaluation, especially given the challenges related to timing of SIRS' detection (see below)? • What is the long term plan for further evaluation(s) necessary to inform implementation of the system in clinical practice? Patient selection and recruitment • Why discharge within 12 hours is used as an exclusion criterion? It may be obvious to clinical experts, but still need some justification. • Recruitment: "As soon as 97 patients suffering from one or more SIRS episodes as well as 137 patients suffering from one or none SIRS episodes have been identified, the recruitment will be terminated early on." These two groups seem to overlap! Is this correct? Index test • It would be helpful to have a brief description of the CDSS system and its intended use, and the results from its previous

	evaluation(s). The authors refer to their previous work, but a brief summary would help readers understand better the proposed study without having to search for and read additional papers. • Physicians will be aware of the study's objectives and, as a result, their behaviour may change, e.g. they may be more vigilant towards episodes of SIRS, more meticulous in recording results etc. If this assumption is correct, how much of an impact this could have on test accuracy estimates and how do the authors plan to deal with it (e.g. compare SIRS' detection rate during the study to some historical period)? • The same could be asked about CADDIE2 – if it relies on routinely recorded data, is the quality and availability of such data similar before and during the study? Is it likely to decline again once the study has been completed? Reference standard • Although a chart review by two independent blinded clinicians is a more robust reference standard, compared to routinely recorded ICD codes, it may still introduce bias if, for example, important information is missing from the records. What potential sources of bias exist with the proposed reference standard and what is the authors' plan for dealing with them? • The assessment of diagnostic accuracy is based on the assumption that the target condition does not change in the interval between the application of the index test and the reference standard (either because they are determined at the same time, or because the interval is short enough to preclude such a change). When this condition is violated, there is a risk of bias in the Flow and timing domain (see QUADAS-2) and the results may not be valid (e.g. SIRS may be absent at the time when the index test was applied, but develop later on when the reference standard was measured, and vice versa). My understanding is that in this specific context timing is even more crucial. As the authors put it, "In pediatric septic shock patients, every hour without appropriate treatment was associated with an increased risk of death by 40% [9]. Conclusively, early recognition, evaluation and treatment of pediatric SIRS and sepsis are vital for improved survival" This begs the following questions: o Is the proposed 'acceptable' interval between index test and reference standard—within shift—sensitive enough to enable prompt clinical intervention? From the description of the target condition (which informs my understanding of the situation!) onset of SIRS in the beginning of the shift should be detected as soon as possible, as the risk of complications and death is increasing by hour. If CADDIE2 detects this at the end of the shift, this will be of little help, as the chances of the patient's surviving will have been significantly reduced. o In the above situation, where an acceptable interval is "within shift", SIRS detected by the reference standard at the beginning of the shift and by the index test at the end of the shift will still result in a true positive result, despite the loss of opportunity to provide timely treatment to the patient. o If the above is correct, then the accuracy estimates would not be very useful clinically, as they do not guarantee timely detection of SIRS. Statistical analysis and outcomes • Can you please clarify what exactly you mean by "sensitivity and specificity on the level of patients" and "sensitivity and specificity on the level of intensive care days". If by the first you mean that a

	positive result by CADDIE2 and the reference standard is considered 'true positive', regardless of their timing (e.g. at any time during the patient's stay), I struggle to see the clinical use of such results! As I said earlier, this will violate one of the basic principles of test accuracy evaluation, namely that the condition status should not change in this interval (cross-sectional design). • In the statistical analysis, the authors state that "For the primary outcome measure, sensitivity and specificity will be determined together with Wald confidence intervals. The comparison will be carried out by comparing the lower bound of the confidence interval with the null hypothesis." I'm not sure what exactly is meant here by 'null hypothesis'. Can you please clarify? Subgroup analysis: Do the authors plan to conduct any subgroup analysis? For instance: • SIRS, sepsis and organ dysfunction/failure are often mentioned separately, e.g. "It is documented whether the patient suffered from SIRS, sepsis or organ dysfunction". I'm not familiar with the clinical area, but wonder if a subgroup analysis should be done to look at the accuracy of the system to identify each subcategory, if clinically relevant? • What about patient's age? Can this impact on the accuracy of the system or the routine assessment? • Timing of assessment, in terms of specific shifts (e.g. weekend vs. weekday)? There is plenty of research showing that this contributes to the quality of care and could potentially have impact on the accuracy of clinical assessment. • More generally, identifying the factors that could modify the accuracy of the system could, potentially, help better understand of its performance and would provide important information for future evaluations (e.g. systematic reviews). I recommend that the authors address the above issues before the protocol is published in BMJ Open.

REVIEWER	Ivan Cabezas Universidad de San Buenaventura - Cali, Colombia
REVIEW RETURNED	04-Mar-2019

GENERAL COMMENTS	Study goals are clearly stated, interesting and useful for the on-subject researching community and practitioners. Chosen outcome measures are classic and suited for the purpose. Pediatric patients are treated properly and enough and clear information is provided to parents. Presented discussion is based on study planning characteristics and aspects. Specific research conclusions would be addressed based on study obtained results. From the reviewer perspective, authors are following best-practices for study planning. Please provide more details on the study timeline. (Protocol papers should report planned or ongoing studies. The dates of the study should be included in the manuscript.)

REVIEWER	Joe Carcillo University of Pittsburgh, USA
REVIEW RETURNED	05-Mar-2019

GENERAL COMMENTS	Very nice study Good Luck

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer #1

Overall Comment: This is an interesting and, according to the authors, clinically relevant study that may help PICU physicians with early detection of SIRS and prompt initiation of evidence-based treatment. Although the protocol provides a comprehensive description of the background, rationale and the methods of the study, I feel that the following aspects need further clarification and detail.

Response: We thank reviewer 1 for the thorough review and the valuable suggestions to improve our manuscript. The multiple specific changes to the manuscript are described below and are marked in the revised manuscript.

Comment 1: Outcomes and choice of study design

As the authors point out, the study is not intended to evaluate user acceptability, or indeed the impact of implementing the system on patient outcomes. The study can only show if the system is accurate enough when compared to current practice (routine clinical assessment) against the reference standard of a chart review conducted by two independent blinded experts. It would be helpful for the readers to understand:

• Why test accuracy and not more clinically relevant outcomes (e.g. patient outcomes assessed through an RCT) have been selected for the focus of this evaluation, especially given the challenges related to timing of SIRS' detection (see below)?

Response: Thank you for raising this point. Testing the accuracy of the system is just the first step. Of course, the ultimate goal is to improve the patient's chances in the end. Many studies focus on directly proving the effect of a CDSS implementation by evaluating the length of stay, mortality, or the success of therapies without testing the accuracy of the algorithm first. Our more conservative approach firstly aims at assessing the diagnostic accuracy of the system before targeting on demonstration of effects on clinical outcomes because we think that any improvement for the single patient only is possible if the algorithm is able to accurately detect the disease, resulting in very high sensitivity and specificity needed. Only then, it may be integrated in the daily clinical routine, offering a realistic chance to prove the effect of CDSS to improve treatments and outcomes. Furthermore, when proving good accuracy results of the CDSS first, the conduction of more complex trials based on the CDSS as well as the real-time implementation needed for any RCT as proposed can be argued better. To clarify our approach and to give a hint on upcoming evaluations as the evaluation of patient outcomes, we added a paragraph to the discussion section:

Manuscript, discussion: “Moreover, this will allow the conduction of future studies as for example the evaluation of patient outcomes, the user acceptability, or the real-time performance of the system.“

Comment 2: Outcomes and choice of study design

• What is the long term plan for further evaluation(s) necessary to inform implementation of the system in clinical practice?

Response: Thanks for your interest in our future work. We would like to refer you to our reply to comment 1.

Furthermore, in case of sufficient accuracy of the system, two different ways of creating patient benefits are taken into account. One of our goals is the implementation and integration of the CDSS algorithms into the existing patient data management system (PDMS). One major benefit would be the direct visualization of new results in the PDMS and a potential acceleration of the diagnostic process. Here, further evaluations on the visual interface, on the implementation as real-time system and the required interfaces and on the user acceptability will be needed. A cooperation with the medisite® GmbH already exists to quickly work on a transfer of positive results to their PDMS.

The second long-term goal is the possible use of the CDSS for the labelling of big datasets of intensive care patients. Thereby, we can generate possible new insights about the contributing factors for SIRS and possible relevant consequences of this entity for the outcome of intensive care patients by implementing data-driven algorithms. Of course, these algorithms again will be evaluated according to their accuracy and effects on patient outcomes. We would like to refer to the modifications of the manuscript presented in the reply of comment 1.

Comment 3: Patient selection and recruitment

• Why discharge within 12 hours is used as an exclusion criterion? It may be obvious to clinical experts, but still need some justification.

Response: Thank you for this question. A patient who develops SIRS will not be discharged from the intensive care unit after 12 hours. Therefore, these patients are extremely unlikely to have SIRS and will cause an unnecessary increase in workload for the manual chart review. The reviewer is right that this might not be clear for all readers, so we added an explanation to the section patient population and eligibility criteria.

Manuscript, patient population and eligibility criteria: “Patients will be recruited continuously and included, if a positive consent is available, and their length of stay exceeds twelve hours because any patient developing SIRS will not be discharged earlier.”

Comment 4: Patient selection and recruitment

• Recruitment: “As soon as 97 patients suffering from one or more SIRS episodes as well as 137 patients suffering from one or none SIRS episodes have been identified, the recruitment will be terminated early on.” These two groups seem to overlap! Is this correct?

Response: This is correct, and needs to be this way because of the definition of specificity on a patient level in this specific situation (sensitivity is straight forward, and only based on those who develop at least one SIRS episode; sample size calculations defines then n=97). Since for the primary analysis of our study we aggregate SIRS information within one patient (presenting the most conservative way of calculating diagnostic accuracy), the CDSS can cause a false-positive signal even in patients who experience a SIRS if SIRS diagnosis by the CDSS is not in the same hour as the

SIRS diagnosis by the gold standard. A standard example would be a patient diagnosed with SIRS by the gold standard at day two (missed by the CDSS), for whom the CDSS also defines a SIRS episode at day seven (which is not confirmed by the gold standard). This patient would contribute one false-negative observation to the calculation of sensitivity as well as one false-positive observation to the calculation of specificity. This approach seems unusual, but in reality gives the most conservative estimates of sensitivity and specificity. If we exclude those with at least one SIRS from the estimation of specificity, we will overestimate the true clinical specificity because those predisposed to SIRS are most likely the ones in which additional false-positive alerts occur (potentially leading to alert fatigue). Moreover, we are sure that an estimation of diagnostic accuracy on the patient level is considerably more clinically important (and gives a more realistic picture when compared to classic diagnostic tests) than an analysis purely based on time intervals since a time interval analysis again heavily overestimates the true patient-level diagnostic accuracy. We rewrote the respective paragraphs to make this clearer. It now reads:

Manuscript, statistical analysis and sample size calculation: "For analyzing the primary outcome measure, the assessment is carried out on the patient level. This is challenging since the assessment is not cross-sectional (as e.g. if the unit of assessment would be an hour respectively a shift) but needs to incorporate the complex longitudinal course of potential assessments within one patient. It is, however, the clinically most meaningful and is the most conservative approach for estimating the diagnostic accuracy if conducted correctly. In our case the entire period of stay is considered, and information are aggregated on the patient level. Every person contributes (given that a correct diagnosis is restricted to the period of an hour respectively a shift) parts of its period of stay to the calculation of specificity independently of if the gold standard recorded a SIRS at some point since everybody will have periods without SIRS diagnosis (which need to be classified as well correctly by the CDSS). This leads to situations, which cannot be represented in only one cell of a contingency table. The classical four cases are amended by a new case, which occurs because the CDSS should not only correctly assess the occurrence of a SIRS event in general but with a correct timing (e.g. SIRS event is identified within the correct hour respectively shift). For example, this fifth case prevents that alert firings on day 30 of the intensive care stay will be evaluated as true positives if the gold standard reports a SIRS episode on day 2. Here, the CDSS did not identify the SIRS episode within the correct hour respectively shift. Thus, this case is used for the determination of both false positives (day 30, contributing to specificity) and false negatives (day 2, contributing to sensitivity). Hence, the fifth case (false positive and false negative) can be defined as follows: the gold standard reports at least one SIRS episode, and the CDSS detect SIRS episodes but (at least one) not within the same hour respectively shift. All other cases are defined as usual (e.g. false positive: the gold standard reports no SIRS episode but the CDSS detects one or multiple SIRS episodes).

Based on the different cases, the sensitivity and the specificity will be determined independently."

Comment 5: Index test

- It would be helpful to have a brief description of the CDSS system and its intended use, and the results from its previous evaluation(s). The authors refer to their previous work, but a brief summary would help readers understand better the proposed study without having to search for and read additional papers.

Response: Many thanks for the suggestion. We already included the results of the first evaluation in our manuscript (see section preceding studies: "Furthermore, we conducted a proof-of-concept study focusing on the technical practicability of the CDSS, yielding at promising results for both the technical infrastructure and the accuracy of the system (sensitivity of 1.00, specificity of 0.94) [17]."). Furthermore, we added the two main characteristics of our system (rule-based, interoperable) to make it clearer to the reader. We think that any other technical summarization of the system might exploit the objectives and the length of our manuscript.

Manuscript, introduction: “In our previous work, we designed a rule-based and interoperable CDSS for the detection of SIRS in pediatric intensive care [17]. The CDSS is able to retrieve and evaluate dynamic facts as routinely and automatically measured parameters from the bedside monitors to detect SIRS episodes.”

Comment 6: Index test

• Physicians will be aware of the study’s objectives and, as a result, their behaviour may change, e.g. they may be more vigilant towards episodes of SIRS, more meticulous in recording results etc. If this assumption is correct, how much of an impact this could have on test accuracy estimates and how do the authors plan to deal with it (e.g. compare SIRS’ detection rate during the study to some historical period)?

Response: Thank you for this important point. This will not affect the main analysis of our study since CDSS results (which are purely based on automatic data collection within the hospital information system – totally independent of physicians’ diagnoses and/or decisions) will only be compared to the gold standard of a posteriori chart review. However, for the secondary endpoint - in which CDSS results and the gold standard are compared to real-time diagnoses of physicians - this might indeed be an issue. We tried to minimize this effect by applying a run-in period of one month before the start of the study, in which physicians already documented SIRS status by shift although the study was not yet started. We will, nevertheless, indeed compare how the implementation of the documentation scheme and the study affected observed SIRS incidence on the study unit. We added the following statement

to the text:

Manuscript, statistical analysis and sample size calculation: “SIRS prevalence and incidence will be monitored throughout the pilot phase and the main phase of the study, and will be compared to pre-study values in order to estimate the risk of a training effect on physicians’ real-time diagnoses caused by knowledge about the aims of this study.”

Comment 7: Index test

• The same could be asked about CADDIE2 – if it relies on routinely recorded data, is the quality and availability of such data similar before and during the study? Is it likely to decline again once the study has been completed?

Response: Thanks. It is right that the system relies on routinely recorded data and the quality and availability of such data might differ. However, our system only relies on data which is automatically measured and transferred from the bedside monitors to the patient data management system. The routine workarounds for measurements will not be affected by our study as we do not include more or more precise measurements. Consequently, variabilities in the amount and quality of data cannot be leaded back to our trial because any manual manipulation is nearly impossible. We assume that our wording was not clear enough, which is why we included the following sentence:

Manuscript, introduction: “The CDSS is able to retrieve and evaluate dynamic facts as routinely and automatically measured parameters from the bedside monitors to detect SIRS episodes.”

Comment 8: Reference standard

- Although a chart review by two independent blinded clinicians is a more robust reference standard, compared to routinely recorded ICD codes, it may still introduce bias if, for example, important information is missing from the records. What potential sources of bias exist with the proposed reference standard and what is the authors' plan for dealing with them?

Response: It is correct that this might be a challenge. However, both reviewers are very experienced pediatric intensive care physicians who are able to discriminate unsound data. Furthermore, the reviewers have access to both the measurements per minute and the validated measurements of the attending nurses per hour. Consequently, if some data points are missing, the hourly measured value is a potential back up. Moreover, all other measurements from parameters not stated as SIRS criteria also can be reviewed by the reviewers during chart review to get an impression on the patient's condition. If the reviewer's results are inconclusive, a third reviewer will participate, the case will be discussed and the details will be published. We already included the involvement of a third reviewer into the section study objectives and diagnostic approaches ("In case of disagreement, a third clinician will be consulted."). For the other aspects, we modified the manuscript as follows:

Manuscript, patient population and eligibility criteria: "The reviewers who will perform the manual chart review for creating "gold standard" decisions are specialized pediatricians and very experienced in pediatric intensive care (working in this PICU for over three years), able to discriminate unsound and missing data."

Manuscript, study objectives and diagnostic approaches: "These comprise evaluating all patients' measurements, not restricted to the SIRS parameters, including additional values for vital signs validated hourly by the attending nurse."

Comment 9 and 10: Reference standard

The assessment of diagnostic accuracy is based on the assumption that the target condition does not change in the interval between the application of the index test and the reference standard (either because they are determined at the same time, or because the interval is short enough to preclude such a change). When this condition is violated, there is a risk of bias in the Flow and timing domain (see QUADAS-2) and the results may not be valid (e.g. SIRS may be absent at the time when the index test was applied, but develop later on when the reference standard was measured, and vice versa). My understanding is that in this specific context timing is even more crucial. As the authors put it, "In pediatric septic shock patients, every hour without appropriate treatment was associated with an increased risk of death by 40% [9]. Conclusively, early recognition, evaluation and treatment of pediatric SIRS and sepsis are vital for improved survival" This begs the following questions:

- Is the proposed 'acceptable' interval between index test and reference standard—within shift—sensitive enough to enable prompt clinical intervention? From the description of the target condition (which informs my understanding of the situation!) onset of SIRS in the beginning of the shift should be detected as soon as possible, as the risk of complications and death is increasing by hour. If CADDIE2 detects this at the end of the shift, this will be of little help, as the chances of the patient's surviving will have been significantly reduced.

- If the above is correct, then the accuracy estimates would not be very useful clinically, as they do not guarantee timely detection of SIRS.

Response: Thank you for this very important point. It is absolutely correct. The aim of the algorithm is to produce an alert as soon as the SIRS criteria are met. The system is able to analyze the data continuously and submit the alert as soon as the SIRS criteria are met. Of course, the aim of the algorithm is to produce an alert as soon as the criteria are met, not at the end of the shift. The system will label the exact time of onset of SIRS as will the reviewers. We apologize for the unclear wording in the first version of the manuscript.

The SIRS scoring by the attending physician is just a second arm of this study to point out whether SIRS is recognized at all. During clinical routine, physicians are able to score patients at the end of their shifts only, and not with a correct timing. The implementation of a more specific documentation process (e.g. every time a SIRS is recognized) would completely change the clinical routine work and, therefore, the assessments would not represent the routine decision-making. There will be no comparison of the potential time delay of the doctor in recognizing SIRS. One aim of this study is to evaluate the accuracy of the CDSS and another aim is to show how often SIRS is detectable according to the definitions by the system but not clinically recognized by the clinician. This could be important because SIRS predisposes the patient for organ dysfunctions and failure and early recognition is crucial for the prevention. Hence, for this assessment, the shift is used as a reliable period. We changed all paragraphs that include the definition of the relevant time period into "within an hour respectively a shift".

Comment 11: Statistical analysis and outcomes

• Can you please clarify what exactly you mean by "sensitivity and specificity on the level of patients" and "sensitivity and specificity on the level of intensive care days". If by the first you mean that a positive result by CADDIE2 and the reference standard is considered 'true positive', regardless of their timing (e.g. at any time during the patient's stay), I struggle to see the clinical use of such results! As I said earlier, this will violate one of the basic principles of test accuracy evaluation, namely that the condition status should not change in this interval (cross-sectional design).

Response: Thank you for raising this important issue. In fact, our analysis chooses exactly the opposite approach for patient level diagnosis, and tries to be extremely conservative. I would like to refer you to our reply to comment 4. Please excuse that our wording was not good enough to get this point through. We re-phrased the relevant paragraphs substantially so that it is now clearer.

Comment 11: Statistical analysis and outcomes

• In the statistical analysis, the authors state that "For the primary outcome measure, sensitivity and specificity will be determined together with Wald confidence intervals. The comparison will be carried out by comparing the lower bound of the confidence interval with the null hypothesis." I'm not sure what exactly is meant here by 'null hypothesis'. Can you please clarify?

Response: Thank you for this comment. Unfortunately, the wording of the respective sentence was not good, given that the paragraph about sample size calculations is located behind the statistical analysis chapter. In general, the null hypothesis values for sensitivity and specificity (which were now added to the sentence) are the values against which we want to test the observed sensitivity and specificity in our study. We assume that the true sensitivity and specificity values are higher (which is reflected in the alternative hypotheses in the text). Therefore, the lower bound of the 95% confidence interval for the observed sensitivity and specificity needs to be higher than the pre-defined null hypotheses values to hold a type I error of 5%. We rephrased the paragraph accordingly:

Manuscript, statistical analysis and sample size calculation: “For the primary outcome measure, sensitivity and specificity will be determined together with Wald confidence intervals. The comparison will be carried out by comparing the lower bound of the confidence interval with the null hypothesis (which is, as described below in the sample size calculation paragraph, a sensitivity of 0.90, and a specificity of 0.80). If the lower bound of the 95% confidence intervals for sensitivity and specificity are both above the values of the pre-defined null hypotheses, we will reject the null hypotheses. For the secondary outcome measure, sensitivity and specificity will be determined together with confidence intervals based on general estimating equations. Additionally, for the secondary goal of comparing the results diagnostic accuracy of the CDSS to the one of routine decisions (both when evaluated against the gold standard), sensitivity and specificity values will be compared by means of McNemar tests and confidence intervals constructed based on general estimating equations.”

Comment 12: Subgroup analysis

• Do the authors plan to conduct any subgroup analysis? For instance: SIRS, sepsis and organ dysfunction/failure are often mentioned separately, e.g. “It is documented whether the patient suffered from SIRS, sepsis or organ dysfunction”. I’m not familiar with the clinical area, but wonder if a subgroup analysis should be done to look at the accuracy of the system to identify each subcategory, if clinically relevant?

Response: Thank you for this important comment. We agree that subgroup analyses are a valuable tool for checking the consistency of the diagnostic accuracy and added a respective sentence to the manuscript:

Manuscript, statistical analysis and sample size calculation: “All analyses will be accompanied by secondary subgroup analyses, stratified e.g. by patients’ age, type of shift and clinical picture associated with SIRS detection (including SIRS, sepsis, severe sepsis and septic shock). Factors that might modify the diagnostic accuracy of the CDSS will thus be evaluated in an exploratory way, allowing a better understanding of potential limitations of the system.”

Comment 13: Subgroup analysis

• What about patient’s age? Can this impact on the accuracy of the system or the routine assessment?

Response: Thank you again. The SIRS criteria as well as the implemented algorithm are agedependent, which is why an analysis according to the age is valuable. We think that our algorithm is very robust against potential influencing age dependent problems, e.g. when babies start to cry and the heart rate goes up, as a sign of this restlessness but not as a sign for an ongoing SIRS. To proof this ability of the system, a subgroup analysis according to the age is needed. Consequently, we added the age to the paragraph (see above, comment 12).

Comment 14: Subgroup analysis

• Timing of assessment, in terms of specific shifts (e.g. weekend vs. weekday)? There is plenty of research showing that this contributes to the quality of care and could potentially have impact on the accuracy of clinical assessment.

Response: This is indeed true; however, it will not affect the main analysis where CDSS results will be compared to the gold standard of a posteriori chart reviews. For the sensitivity analysis, in which we compare CDSS ratings and gold standard ratings to real-time diagnoses of physicians on the ward, this will indeed be a potential issue. Generally the ICU is staffed at least with two experienced ICU doctors in all shifts (day-, late- and night shift) over the week as well as on the weekend, so it will be quite interesting whether differences in analysis can be found. We will present subgroup analyses by shift, both for the primary and secondary outcome. We added the type of shift to the paragraph (see above, comment 12).

Comment 14: Subgroup analysis

- More generally, identifying the factors that could modify the accuracy of the system could, potentially, help better understand of its performance and would provide important information for future evaluations (e.g. systematic reviews).

Response: We totally agree, and will look in detail into this. We added a sentence to the manuscript (see above, comment 12).

Response to Reviewer #2

Overall Comment: Study goals are clearly stated, interesting and useful for the on-subject researching community and practitioners Chosen outcome measures are classic and suited for the purpose. Pediatric patients are treated properly and enough and clear information is provided to parents. Presented discussion is based on study planning characteristics and aspects. Specific research conclusions would be addressed based on study obtained results. From the reviewer perspective, authors are following best-practices for study planning.

Response: We thank reviewer 2 for reviewing our manuscript and the accurate

recommendations. Thanks a lot for the positive feedback.

Comment 1: Please provide more details on the study timeline.(Protocol papers should report planned or ongoing studies. The dates of the study should be included in the manuscript.)

Response: Thanks. It is correct that the specific dates were missing. We included the dates for the different phases of the study in both the methods section of the abstract and the manuscript:

Manuscript, Methods and analysis: Personal briefings on the new routine documentation were carried out during this pilot phase (July 1, 2018, estimated duration: 1 month). A designated assistant physician will present the study to the patients, their parents or their legal guardians and ask for consent within the recruiting phase (August 1, 2018, estimated duration: 6 months). (…) The clinicians do not perform extensive analyses of documentations or reported data (assessment phase I, August 1, 2018, estimated duration: 6 months) (…) Later

on, two experienced clinicians will start with their weekly, extensive, blinded review and the definition of “gold standard” assessments per patient and per hour (assessment phase II, February 1, 2019, estimated duration: 3 months). (…) In the final analysis phase (May 1, 2019, estimated duration: 2 months), the diagnostic accuracy of the CDSS will be evaluated by comparing the assessments to the “gold standard” decisions from the experts (primary goal of the study).

Response to Reviewer #3

Overall Comment: Very nice study. Good Luck

Response: Many thanks to reviewer 3 for the positive feedback.