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)	The value of systematic detection of physical child abuse at Emergency Rooms; a cross-sectional diagnostic accuracy study
AUTHORS	Sittig, Judith; Uiterwaal, Cuno; Moons, Karel; Russel, Ingrid; Nievelstein, Rutger Jan; Nieuwenhuis, Edward; van de Putte, Elise

VERSION 1 - REVIEW

REVIEWER	Marion Bailhache CHU de Bordeaux, Pole de pediatrie, F-33000 Bordeaux, France, Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
REVIEW RETURNED	17-Dec-2015

GENERAL COMMENTS	Judith S Sittig et al have estimated the accuracy of the CHAIN-ER to detect physical abuse among children aged 0-7 years attending the ER because of physical injury. The number of quality criteria of their diagnostic accuracy study is high. The accuracy of their reference standard is not determined but no gold standard is available for child abuse. They have thus used expert panel. The reference standard was not systematic but all suspected cases and a 15% random sample of the children with negative screen received it. The independence of reference standard and index test was unclear. In the introduction, line 3, the first reference (Gilbert et al) appears to be inverted with the second reference (Woodman et al). The secondary outcome of the study is not immediately obvious in the introduction, line 3, page 6. The definition by the authors to mean “need for help from social services” is indeed in the methods, page 7. In the methods, concerning the index test: who completed the checklist and when during the consultation (beginning of the consultation, after clinical examination or after all medical checks were completed)? In the methods, page 7, line 16, the definition of “need for help from social services” is unclear and seems to be very large, which might explain partially why the expert panel inter-rater agreement is very low in the results. In the methods, concerning the reference standard, line 5, page 8, what the authors mean by “shortly”? Could they specify when, at most after the ER visits, could the telephone interviews of parents/caregivers be made? Likewise (line 9, page 8) when, after the ER visits, were the healthcare professionals questioned about the risk factors for child abuse? And (line 12, page 8) when di the nurses check for registrations at the Child Protection Services? Line 15, page 8, were the nurses aware of the checklist outcome when they checked the electronic patient file for additional clinically relevant information after 6 months? Likewise were the healthcare professionals informed of the checklist outcome when they were

	questioned (line 9, page 8)? In the figure 2, define the abbreviations CAP, GP, CPS, VAS, Y/N to understand the figure without the need for the text. Were CAP (line 4, page 8) and CAAT (line 7, page 7) the same team? In the results, very few caregivers did not consent for evaluation of data. But the percentage seems unbalanced (10.7% vs 3.9%). Line 17, page 10, could the authors specify if the screening test was positive or negative for the one child to have possibly inflicted injury as a clinical outcome? Line 19, page 10, some parents/caregivers had no CAP interview (n=192). Could the authors specify the repartition of these parents/caregivers according to the results of index test? Likewise, line 21, page 10, could the authors specify the repartition of the 35 parents/caregivers who did not give permission for researchers to contact other healthcare professionals according to the results of the index test? Line 8, page 11, only 49 children of the 112 children with a positive screen were seen by CAP, why? In the discussion, line 24, page 11, the full agreement about inflicted injury between the expert panel and CAPs concerned less than half of the children with the positive screen. Line 28, page 11, even if many quality criteria are filled, all QUADAS criteria for diagnostic accuracy studies are not fulfilled. The criterion 3: is the reference standard likely to correctly classify the target condition? is unclear. Line 1, page 12, Louwers et al assessed the diagnostic accuracy of a very similar test, the Escape instrument (one of the six questions is clearly different). But their study had indeed less quality criteria. However, in other circumstances of children attending at Emergency room for any physical injury, other studies of diagnostic of child abuse are already used the same reference standard procedure for children with positive and negative screens. For example: Valvano et al. Does bruising help determining which fractures are caused by abuse. Child maltreatment 2009, 14(4):376-381. Line 11, page 12, panel members were blinded to the results of the checklist, but the persons which collected additional information (CAP, nurses, health professionals?) for their assessment were not all blinded of the checklist outcome. So was additional information not influenced by the checklist outcome? In addition, had panel members the same information for all the children? For example, had all the children radiological images to detect old or new fractures? The prevalence of physical abuse among children aged 0-7 years admitted to ER for any physical injury in the study is indeed very low compared with other studies (Woodman et al). However the study was not designed to estimate this prevalence. The prevalence is estimated relatively low even among the other studies (1%), which would represent only 42 cases among 4178, and 6 cases among the sample of 645 children. And 25 parents/caregivers refused evaluation of data among the random sample of 645 children with negative screen and were excluded. Assessment of all children or a more important sample by a valid and reliable test would be probably necessary to estimate correctly the prevalence. Line 22, page 12, the authors raise differences in measurement, setting and methodology to explain this difference of prevalence. Would they be a little clearer about these differences of measurements and methodology? Line 23, page 12, the authors suspect that most studies have overestimated the prevalence of physical abuse. On what arguments? In the literature, studies have shown that the cases of physical abuse seem to be underestimated. For example concerning the Pediatric Abusive Head Trauma, Jenny C et al have

	shown that in a retrospective study of 178 cases, 54 cases were missed (children seen by physicians after PAHT and the diagnosis was not recognized). Among these 54 cases, 10 were classified as accidental head trauma. More recently Adamsbaum et al have confirmed this underestimated (Adamsbaum et al. Abusive Head Trauma: judicial admissions highlight violent and repetitive shaking. Pediatrics 2010). The index test weas applied only for children with physical injury. However the difficulties of detected physical abuse are also the results that several children have not specify symptoms and no trauma is reported by caregivers. For example, frequent erroneous diagnoses made in cases of abusive head trauma are viral gastroenteritis or reflux because the principal symptom reported by caregivers and noted by physician is vomiting (Jenny C et al. Analysis of missed cases of abusive head trauma. JAMA 1999). Line 25, page 12, how do the authors explain the very low expert panel inter-rater agreement for injury caused by neglect? Line 14, page 13, the authors conclude that, where such checklists are not used yet, they do advise careful prior consideration of cost-effectiveness and clinical and societal implications before de novo implementation. Indeed the performance of screening test is not the only consideration before considering systematic screening of physical abuse at Emergency Rooms among children attending for injury. The effectiveness of interventions could be another consideration, for example. But these considerations concern as much settings where checklists are not used yet, as settings where checklists are already used.

REVIEWER	Ruth Gilbert UCL Institute of Child Health
REVIEW RETURNED	26-Dec-2015

GENERAL COMMENTS	The authors report a diagnostic accuracy study to evaluate a screening checklist to detect physical abuse among children who attend the emergency room for injury. The population is restricted to children less than 8 years old presenting to 4 hospitals around Utrecht. Major comments 1. The study is important. Checklists are widely used in developed countries to screen for child maltreatment despite weak evidence to support their accuracy. 2. The study is original and definitive. No accuracy study on screening for physical abuse has come close to this level of rigour apart from the study by Barry Pless, which though published in 1987 was actually conducted in 1976. Unbiased determination of the reference standard for both screen positive and screen negative children is a major challenge, given that physical abuse is often hidden and hard to confirm. The authors achieve this through careful, blinded, independent assessment by an expert panel of clinical records (not including the screen test results), paediatric assessment, questionnaires to GPs and youth doctors and follow up for 6 monhts through health care data. The levels of consent and follow up achieved are impressive. 3. The study involves 4 hospitals, which improves generalisabilty 4. The methods are clearly explained and the conclusions are justified by the results. The key problem of low prevalence of physical abuse and hence low predictive value is identified as the core problem.

	5. Unlike many studies on this topic, the authors include important secondary outcomes of injury related to neglect and need for social care intervention. These outcomes are important and more common than physical abuse, though, as their results show, not accurately detected by a checklist. . 6. Near the end of the discussion, the authors could refer to external evidence that corroborates their findings that physical abuse contributes relatively few cases of child maltreatment that receive child protection intervention and may distract clinical attention from detecting neglect and emotional abuse, which are far more common and, in the long-term, also damaging (eg there are many references that could be used to make this point. For example, routine national figures for England and Wales: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/469737/SFR41-2015_Text.pdf (fig H) and table D2.) Minor 7. Abstract:Conclusions; • Add – restricted TO children • Advise not advice 8. page 5 Line 10 – change minor attendees to children 9. Page 10 second paragraph of results. Please refer early on in this paragraph to the excellent figure 3. This makes it much easier to follow the calculations in this paragraph. 10. Line 26 page 12 – change irrelevant to inaccurate. Emphasise that a checklist is not sufficiently accurate and should not replace skilled assessment by a clinician.

REVIEWER	Professor Alison M Kemp Cardiff University UK
REVIEW RETURNED	07-Jan-2016

GENERAL COMMENTS	Thank you for asking me to review this paper. This is a well conducted study about a very important and current topic. There is a movement to develop decision tools to screen populations for possible abuse. It is not an easy area to research and the team are to be commended for their approach. As the authors state it is one of the few studies that explored the presence or absence of 'disease' in the population who were negative on their SPUTOVAMO. Whilst the numbers included in the study are significant. The results of the study are compromised by the low prevalence of the condition. This also affects the parameters chosen to describe the data , namely the ppv. The small numbers of physical abuse cases give extremely wide confidence intervals to the sensitivity calculations. The authors have worked extremely hard to collect data on a considerable number of children I have a number of curiosities and comments below There are a couple of typos in the abstract 'Subsequent assessment by child abuse experts can be safely restricted ….to…..children with positive screens at very low risk of missing cases of inflicted injury. Because of the high false positive rate we do advice advise careful prior consideration of cost- effectiveness and clinical and societal implications before de novo implementation. I wonder regarding terminology …Is the SPUTOVAMA really a

	diagnostic tool, it is variously described as a check list and seems to have the function of highlighting cases that should be investigated further before a 'diagnosis' of physical abuse can be made. A diagnosis of child abuse is unlikely to be made from the features within the SPUTOVAMO without further investigation. If indeed it is a diagnostic test we are talking a very high rate of false diagnosis which is very worrying indeed. Would a clinical decision rule be a better term to use? A clinical decision rule (CDR) is a clinical tool that quantifies the individual contributions that various components of the history, physical examination, and basic laboratory results make toward the diagnosis, prognosis, or likely response to treatment in a patient.5 Jul 2000 n mUsers' Guides to the Medical Literature - JAMA jama.jamanetwork.com/article.aspx?articleid=192850 Bearing in mind the very low prevalence of confirmed abuse , was a power calculation done prior to the study and if so this should be included Further clarification of the excluded children is needed. The figure suggests that 6723 children did not meet the inclusion criteria. Yet the exclusion criteria listed were 'Evident victims of physical child abuse (admitted by perpetrator at presentation), victims of (witnessed) traffic accidents and children who had died before arrival were excluded.' …surely there were not 6000 of these cases? Is the 11013 all attendees including medical attendances? It would be useful to know the coverage of all trauma cases …were all given the sputovamo. Should the flow chart start with all injury cases in children < 7 years rather than all cases…this would be a more useful figure ? Missing cases and data: What about the cases who did not consent or where there was missing parent data etc ? are these cases likely to have influenced the results ? there are a significant number where parenst did not allow further health care worker contact or could not be contacted themselves Of the missing cases 1/37 had probable abuse and whilst it says this would not affect the results we do not know whether this case had a positive or negative SPUTOVAMO. If negative this would indeed reduce the sensitivity? The objective of the SPUTOVAMO was to identify cases of physical abuse, yet secondary outcome measures of neglect and need for help are included. Yet the SPUTOVAMO would seem incompletely designed to identify these cases and indeed was not designed to do so . I am unsure therefore why these measures were included…this should be justified. This association is dismissed in the discussion and I wonder therefore whether it is worth including at all???...If it is then this needs to be justified What this study shows 'This diagnostic study to detect child abuse is the first study that meets all QUADAS criteria for diagnostic accuracy studies' this is not infact accurate as the study sets out to detect physical abuse as stated in the first line of the abstract . terminology needs to be consistent. The term child abuse covers a wide remit …emotional abuse, sexual abuse etc which are certainly not covered within the SPUTOVAMO. This would apply to the second bullet. In the results 0·27% ((5 + (1*4153/620))/4253) (95% CI 0·15 – 0·49) what does the underlined expression refer to Page 11 line 12 is there a word missing? 7 in children with a positive screen

	Page 5 line 25 missing word'Reference testing in checklist-negatives allows ….missing word?...to determine the negative predictive value of the checklist. None of the figures have a title Methods 'The checklist was a compulsory field in the electronic files of the medical records of all attendees.'I am unclear whether this is indeed all attendees or those with an injury ? Discussion There appears to be an association between postive sputovamo and later abuse …or at least a significant risk ratio. Physical abuse is rarely a one off event and is pervasive. Yet this is not mentioned in the discussion??? Does this warrant some discussion if included in the results. I have thoroughly enjoyed reviewing this study and look forward to seeing it in press.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Marion Bailhache
Institution and Country: CHU de Bordeaux, Pole de pediatrie, F-33000 Bordeaux, France; Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

Judith S Sittig et al have estimated the accuracy of the CHAIN-ER to detect physical abuse among children aged 0-7 years attending the ER because of physical injury. The number of quality criteria of their diagnostic accuracy study is high. The accuracy of their reference standard is not determined but no gold standard is available for child abuse. They have thus used expert panel. The reference standard was not systematic but all suspected cases and a 15% random sample of the children with negative screen received it.

- The independence of reference standard and index test was unclear.

JS: in the methods section (see 'expert panel' heading, page 8, lines 3-4) we described the independence of reference standard and index test by mentioning that the expert panel members (reference standard) were kept blinded for the SPUTOVAMO-R outcome (index test): 'Throughout this reference standard procedure, panel experts were kept blinded to the checklist result by deleting that information from steps 2, 3 and 4.' Moreover, the panel members did not exchange their diagnoses and made their assessments independently (see: 'members independently assessed whether the injury was inflicted (yes/no)…'page 8, line 5)

- In the introduction, line 3, the first reference (Gilbert et al) appears to be inverted with the second reference (Woodman et al).

We thank the reviewer for noticing this, indeed these two references were accidentally inverted.

- The secondary outcome of the study is not immediately obvious in the introduction, line 3, page 6. The definition by the authors to mean "need for help from social services" is indeed in the methods, page 7.

JS: in the introduction we explain that, although the checklist was originally developed to detect physical abuse, possibly other types of child abuse may be detected by the checklist as well ('Local versions were developed with other items such as evaluation of interactive behaviour of child and caregivers to suit the checklist for detection of other types of abuse'). Based on the reviewer's comment, we have now clarified this sentence by adding: 'such as neglect, or for detection of need for help in general.'
Indeed, in the methods section, we explain the definitions of all outcome measures more in detail. Here, we also added the explanation why we chose for these secondary outcomes: 'Because physical abuse contributes relatively few cases of child maltreatment that receive child protection intervention and may distract clinical attention from detecting neglect and emotional abuse, which are far more common and, in the long-term, also damaging(ref), we decided to include injury caused by neglect and need for help from social services as secondary outcomes.'(Page 6, lines 17-21)

- In the methods, concerning the index test: who completed the checklist and when during the consultation (beginning of the consultation, after clinical examination or after all medical checks were completed)?

JS: The reviewer is right, this information was missing. We now added the sentence 'for every child an ER nurse or physician fills out the checklist directly after clinical examination' to this paragraph (page 6, line 7).

- In the methods, page 7, line 16, the definition of "need for help from social services" is unclear and seems to be very large, which might explain partially why the expert panel inter-rater agreement is very low in the results.

JS: Indeed, 'need for help from social services' is was intentionally not sharply defined as "any concern about the situation of the child that requires consultation of social services", because is it clinically such an important consequential outcome category. We had foreseen discussion about any stricter definition among particularly panel members, and we anticipated the same discussion with disseminating our findings on a strictly defined classification to the wider audience. We do agree with the reviewer that this choice left room for differences in clinical interpretation, which may have contributed to low panel agreement. It is therefore that, in our discussion, we mention that: 'Our finding of not being able to unequivocally diagnose injury due to neglect and need for help, renders prediction of this kind of injury with a checklist such as SPUTOVAMO-R inaccurate in young children' (Page 12, lines 9-10).

- In the methods, concerning the reference standard, line 5, page 8, what the authors mean by "shortly"? Could they specify when, at most after the ER visits, could the telephone interviews of parents/caregivers be made?

JS: The reviewer is right that this information was missing. To clarify the word 'shortly', we now added: 'i.e. within 2 weeks' (see page 7, line 12)

- Likewise (line 9, page 8) when, after the ER visits, were the healthcare professionals questioned about the risk factors for child abuse?

JS: The risk factors were collected 'within a month after the ER visit'. We added this information on page 7, line 17-18.

- And (line 12, page 8) when did the nurses check for registrations at the Child Protection Services?

JS: the nurses checked for registration at the CPS in the same period that they asked GPs and youth

doctors to fill out the questionnaires. Therefore, we now have added the sentence ‘in the same period’ (page 7, line 20).

- Line 15, page 8, were the nurses aware of the checklist outcome when they checked the electronic patient file for additional clinically relevant information after 6 months?

JS: Because the same research nurses performed steps 3 and 4, they were indeed aware of the checklist outcome when they checked the electronic patient file after 6 months. To clarify this we now added: ‘the same research nurses as mentioned at step 3’ (page 7, line 23).

- Likewise were the healthcare professionals informed of the checklist outcome when they were questioned (line 9, page 8)?

JS: We did not inform the healthcare professionals (GPs and youth doctors) about the checklist outcome when we requested to fill out the questionnaire. They could get informed about real clinical concerns in between the ER visit and questioning. However, even in such cases, the panel members were kept strictly blinded for checklist outcome by deleting this information (as mentioned at the ‘expert panel’ heading, page 8, line 3-4).

- In the figure 2, define the abbreviations CAP, GP, CPS, VAS, Y/N to understand the figure without the need for the text.

JS: to better understand figure 2, we now added these abbreviations to this figure.

- Were CAP (line 4, page 8) and CAAT (line 7, page 7) the same team?

JS: CAP stands for Child Abuse Paediatrician and CAAT for Child Abuse Assessment Team. The CAAT consists of paediatricians and other healthcare professionals specialized in child abuse. We added this explanation where we mention the CAAT and we now include a reference that describes the CAAT in detail (page 6, lines 10-11)..

- In the results, very few caregivers did not consent for evaluation of data. But the percentage seems unbalanced (10.7% vs 3.9%).
- Line 17, page 10, could the authors specify if the screening test was positive or negative for the one child to have possibly inflicted injury as a clinical outcome?

JS: the reviewer is right that only few caregivers refused consent, but more so among positive checklist outcome (10.7%) than negative checklist outcomes (3.9%). We assessed the clinical outcome of these 37 cases and found one clinical suspicion of physical abuse among the positive and none among the negative checklist outcomes. We now added the following information to the manuscript: …’we assessed the clinical outcome and found one child (with a positive screen) to have possibly inflicted injury’, see page 9, line 28, and page 10, line 1).

- Line 19, page 10, some parents/caregivers had no CAP interview (n=192). Could the authors specify the repartition of these parents/caregivers according to the results of index test?

JS: Of the 192 children without a CAP interview, 63 had a positive checklist and 129 a negative checklist outcome. We now have added this information to the manuscript (see page 10, line 2).

- Likewise, line 21, page 10, could the authors specify the repartition of the 35 parents/caregivers who did not give permission for researchers to contact other healthcare professionals according to the results of the index test?

JS: Of the 35 parents who did not give permission to contact other health care professionals, 8 had a positive and 27 a negative checklist outcome. This information is now added to the manuscript (see page 10, line 4).

- Line 8, page 11, only 49 children of the 112 children with a positive screen were seen by CAP, why?

JS: Unfortunately, the CAP could not interview the parents of 192 (of the 720) children, because of refusal or repeatedly being unreachable (see 'Results', page 10, lines 2-4). Of the 112 children with a positive screen, parents of 63 children were not interviewed by the CAP (see page 10, line 2)..

- In the discussion, line 24, page 11, the full agreement about inflicted injury between the expert panel and CAPs concerned less than half of the children with the positive screen.

JS: Indeed, the reviewer is right.

- Line 28, page 11, even if many quality criteria are filled, all QUADAS criteria for diagnostic accuracy studies are not fulfilled. The criterion 3: is the reference standard likely to correctly classify the target condition? is unclear.

JS: The reviewer correctly refers to QUADAS speaking of a likelihood for the reference standard to correctly classify the target. Reviewer will agree that our panel reference standard, like virtually all reference standards, did not provide absolute certainty, we have used it for its highest possible likelihood of doing so.

- Line 1, page 12, Louwers et al assessed the diagnostic accuracy of a very similar test, the Escape instrument (one of the six questions is clearly different). But their study had indeed less quality criteria. However, in other circumstances of children attending at Emergency room for any physical injury, other studies of diagnostic of child abuse are already used the same reference standard procedure for children with positive and negative screens. For example: Valvano et al. Does bruising help determining which fractures are caused by abuse. Child maltreatment 2009, 14(4):376-381.

JS: We thank the reviewer for this information. We agree that other researchers in fact support our use of an expert team as the optimal reference standard. (In the study of Valvano et al., participants were only the cases that underwent the reference test).

- Line 11, page 12, panel members were blinded to the results of the checklist, but the persons which collected additional information (CAP, nurses, health professionals?) for their assessment were not all blinded of the checklist outcome. So was additional information not influenced by the checklist outcome?

JS: We like to refer to our answer on the first question of this reviewer.
We did not inform the healthcare professionals (GPs and youth doctors) about the checklist outcome when we requested to fill out the questionnaire. They could get informed about real clinical concerns in between the ER visit and questioning. However, even in such cases when the professional referred to these concerns, the panel members were kept strictly blinded for checklist outcome by deleting this information (as mentioned at the 'expert panel' heading, page 8, line 3-4).
The research nurses who finally composed the paper expert panel file, deleted all information from the expert panel file that could suggest the checklist result. See 'expert panel' heading, page 8, line 4.
As we fully agree with the reviewer that independence of checklist outcome and added information from the final diagnostic process was of the utmost importance, we firmly believe that we have done everything in our power to achieve that.

- In addition, had panel members the same information for all the children? For example, had all the children radiological images to detect old or new fractures?

JS: We did strive for paper files with the same information for all children, but occasionally files were incomplete because parents did not consent on all data collection (see 'Results', page 10, line 2-6). In such cases, we requested judgment based on all available information (see 'Expert panel', page 8, line 13-14). Radiological images were only present if they were part of usual care. Reviewer will agree that standard radiological imaging for research purposes only, would be both unfeasible and unethical. Retrospectively, a child abuse expert radiologist did evaluate all images made (see step 1, 'reference standard procedure', page 7, line 7-10).

- The prevalence of physical abuse among children aged 0-7 years admitted to ER for any physical injury in the study is indeed very low compared with other studies (Woodman et al). However the study was not designed to estimate this prevalence. The prevalence is estimated relatively low even among the other studies (1%), which would represent only 42 cases among 4178, and 6 cases among the sample of 645 children. And 25 parents/caregivers refused evaluation of data among the random sample of 645 children with negative screen and were excluded. Assessment of all children or a more important sample by a valid and reliable test would be probably necessary to estimate correctly the prevalence.

JS: Indeed our estimated prevalence was low, but we disagree with the reviewer about the possible reasons for it. Our study was exactly designed to estimate the true abuse prevalence in this ER population and we uphold that we could not have done much more to improve our estimate, certainly in comparison to previous studies on the subject. We are not exactly sure about what the reviewer means by "… a more important sample by a valid and reliable test…". We challenge the reviewer's suggestion that a more valid and reliable test than our reference test can be found.

- Line 22, page 12, the authors raise differences in measurement, setting and methodology to explain this difference of prevalence. Would they be a little clearer about these differences of measurements and methodology?

JS: The prevalence estimated by Woodman et al. is based on 66 different studies. Expectedly, there are differences in measurement, setting and methodology among all these studies. To emphasize the likelihood of such differences to be part of the overall estimates, we added the sentence '…which is estimated based on the data of 66 studies' (page 12, line 5)..

- Line 23, page 12, the authors suspect that most studies have overestimated the prevalence of physical abuse. On what arguments? In the literature, studies have shown that the cases of physical abuse seem to be underestimated. For example concerning the Pediatric Abusive Head Trauma, Jenny C et al have shown that in a retrospective study of 178 cases, 54 cases were missed (children seen by physicians after PAHT and the diagnosis was not recognized). Among these 54 cases, 10 were classified as accidental head trauma. More recently Adamsbaum et al have confirmed this underestimated (Adamsbaum et al. Abusive Head Trauma: judicial admissions highlight violent and repetitive shaking. Pediatrics 2010).

JS: Based on our findings, we suspect overestimations of prevalences of physical abuse in other studies. However, the reviewer is right that there is also literature on possible underestimation. These studies estimate the number of missed cases among substantiated cases. This is not the same as determining the prevalence of true cases (i.e. diagnosis of physical child abuse in our study). To determine the prevalence of child abuse cases, an accepted reference test must be available. We

used, as mentioned before, in our opinion the best possible reference test to diagnose cases of physical child abuse. Differences in prevalences could be explained by differences in the use of a proper reference test.

- The index test was applied only for children with physical injury. However the difficulties of detected physical abuse are also the results that several children have not specify symptoms and no trauma is reported by caregivers. For example, frequent erroneous diagnoses made in cases of abusive head trauma are viral gastroenteritis or reflux because the principal symptom reported by caregivers and noted by physician is vomiting (Jenny C et al. Analysis of missed cases of abusive head trauma. JAMA 1999).

JS: We agree with the reviewer that there may be wider ramifications. Therefore, we did also include the patients with injuries that became clear during the ER visit, even when this injury was not the initial symptom. To clarify this, we now have added 'When the initial symptom was not an injury, but when the trauma became clear during the ER visit, the child was also included' (page 5, line 17-18).

- Line 25, page 12, how do the authors explain the very low expert panel inter-rater agreement for injury caused by neglect?

JS: Although we did not ask the panel members for their reasons to score a case positive or negative for injury due to neglect, the reviewer will surely agree that it is the element of subjectivity in this measure that causes low inter-rater agreement.

- Line 14, page 13, the authors conclude that, where such checklists are not used yet, they do advise careful prior consideration of cost-effectiveness and clinical and societal implications before de novo implementation. Indeed the performance of screening test is not the only consideration before considering systematic screening of physical abuse at Emergency Rooms among children attending for injury. The effectiveness of interventions could be another consideration, for example. But these considerations concern as much settings where checklists are not used yet, as settings where checklists are already used.

JS: We very much agree with this statement, both concerning considerations about implementation of instruments and about effectiveness of interventions.

Reviewer: 2
Reviewer Name: Ruth Gilbert
Institution and Country: UCL Institute of Child Health, UK Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

The authors report a diagnostic accuracy study to evaluate a screening checklist to detect physical abuse among children who attend the emergency room for injury. The population is restricted to children less than 8 years old presenting to 4 hospitals around Utrecht.

Major comments
1. The study is important. Checklists are widely used in developed countries to screen for child maltreatment despite weak evidence to support their accuracy.
2. The study is original and definitive. No accuracy study on screening for physical abuse has come close to this level of rigour apart from the study by Barry Pless, which though published in 1987 was actually conducted in 1976. Unbiased determination of the reference standard for both screen positive and screen negative children is a major challenge, given that physical abuse is often hidden and hard

to confirm. The authors achieve this through careful, blinded, independent assessment by an expert panel of clinical records (not including the screen test results), paediatric assessment, questionnaires to GPs and youth doctors and follow up for 6 monhts through health care data. The levels of consent and follow up achieved are impressive.
3. The study involves 4 hospitals, which improves generalisabilty
4. The methods are clearly explained and the conclusions are justified by the results. The key problem of low prevalence of physical abuse and hence low predictive value is identified as the core problem.
5. Unlike many studies on this topic, the authors include important secondary outcomes of injury related to neglect and need for social care intervention. These outcomes are important and more common than physical abuse, though, as their results show, not accurately detected by a checklist. .

6. Near the end of the discussion, the authors could refer to external evidence that corroborates their findings that physical abuse contributes relatively few cases of child maltreatment that receive child protection intervention and may distract clinical attention from detecting neglect and emotional abuse, which are far more common and, in the long-term, also damaging (eg there are many references that could be used to make this point. For example, routine national figures for England and Wales: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/469737/SFR41-2015_Text.pdf (fig H) and table D2.)

JS: We thank the reviewer for this information and now refer to it in the manuscript, see page 6, lines 17-21.

Minor
7. Abstract:Conclusions;
• Add – restricted TO children
JS: We have added this.

• Advise not advice
JS: We have changed this.

8. page 5 Line 10 – change minor attendees to children
JS: We have changed this.

9. Page 10 second paragraph of results. Please refer early on in this paragraph to the excellent figure 3. This makes it much easier to follow the calculations in this paragraph.
JS: We added a reference to figure 3 in this paragraph as well.

10. Line 26 page 12 – change irrelevant to inaccurate.
JS: Done

Emphasise that a checklist is not sufficiently accurate and should not replace skilled assessment by a clinician.
JS: We added this sentence to this paragraph (page 12, line 11).

Reviewer: 3
Reviewer Name: Professor Alison M Kemp
Institution and Country: Cardiff University, UK Please state any competing interests or state ‘None declared’: None

Please leave your comments for the authors below

Thank you for asking me to review this paper. This is a well conducted study about a very important and current topic. There is a movement to develop decision tools to screen populations for possible abuse. It is not an easy area to research and the team are to be commended for their approach. As the authors state it is one of the few studies that explored the presence or absence of 'disease' in the population who were negative on their SPUTOVAMO.
Whilst the numbers included in the study are significant. The results of the study are compromised by the low prevalence of the condition. This also affects the parameters chosen to describe the data , namely the ppv. The small numbers of physical abuse cases give extremely wide confidence intervals to the sensitivity calculations.
The authors have worked extremely hard to collect data on a considerable number of children I have a number of curiosities and comments below

- There are a couple of typos in the abstract 'Subsequent assessment by child abuse experts can be safely restricted ….to…..children with positive screens at very low risk of missing cases of inflicted injury. Because of the high false positive rate we do advice advise careful prior consideration of cost-effectiveness and clinical and societal implications before de novo implementation.
JS: We removed the typos.

- I wonder regarding terminology …Is the SPUTOVAMA really a diagnostic tool, it is variously described as a check list and seems to have the function of highlighting cases that should be investigated further before a 'diagnosis' of physical abuse can be made. A diagnosis of child abuse is unlikely to be made from the features within the SPUTOVAMO without further investigation. If indeed it is a diagnostic test we are talking a very high rate of false diagnosis which is very worrying indeed. Would a clinical decision rule be a better term to use?
A clinical decision rule (CDR) is a clinical tool that quantifies the individual contributions that various components of the history, physical examination, and basic laboratory results make toward the diagnosis, prognosis, or likely response to treatment in a patient.5 Jul 2000 n mUsers' Guides to the Medical Literature - JAMA
jama.jamanetwork.com/article.aspx?articleid=192850

JS: Although arguable, we consider the CHAIN-ER study as diagnostic study rather than for example a screening study, but it clearly contains elements of both diagnostic and screening research. We are aware of the fact that the checklist is only the start of the process of making the diagnosis physical child abuse. In fact, the diagnostic process starts with a patient with a certain complaint (injury), which makes the physician suspicious of him or her having a particular disorder (physical child abuse). However, CHAIN-ER is not designed as a CPR study. If designed as a CPR study, we should have examined how a particular type of injury (for example bruises), in combination with other findings could predict the probability of physical child abuse. SPUTOVAMO-R is used for children with all types of injuries, which makes it unlikely to be used as a CPR.

- Bearing in mind the very low prevalence of confirmed abuse, was a power calculation done prior to the study and if so this should be included

JS: We did an a-priori power consideration, the below is a translation of the original protocol text. As shown, our expectation of prevalence was higher than actually found in our study. For that reason we did post-hoc abandon the original plan to construct a diagnostic prediction rule, and performed the univariable analytical approach presented in the manuscript text.

We are certainly willing to include the below text in the manuscript but as it became less relevant given our study findings, we would like to leave that decision at the discretion of the editor.

Statistical power evaluation.

In the Wilhelmina Children's Hospital some 750 children with injury below age 6 years are seen annually. In the Mesos Medical Center, the Antonius Hospital and Diakonessen Hospital these numbers are some 1,000, 600, 500 children respectively. The total available study population thus amounts to 2850 children. In the Wilhelmina Children's Hospital. Diakonessen Hospital, and Antonius hospital some 1.7-2.2% of these are suspected based on current signaling (SPUTOVAMO), some 4% in Mesos Hospital likely resulting from recent optimization measures. At Chain ER initiation it was unclear whether suspicions were true cases of abuse. Previous Dutch research had shown that 75% of screen positives could not be verified as true abuse cases. This would for CHAIR ER imply that only some 30 children could be expected to get a certain abuse diagnosis annually. We did expect a somewhat higher yield partly because of the intensity of our procedure and our search for additional risk factors. With an inclusion over 15 months (expected number of patients 3560) we expected to detect some 70 (2%) true positives (by gold standard diagnosis). We intended to further diagnostically follow up on 1 out of every 5 screen negatives, amounting to a total 660 screen negatives. As the initial intent of CHAIN ER was to design a diagnostic prediction model, this number of endpoints allowed for accurate evaluation of screening test characteristics and of added value of various diagnostic phases (1). For that modelling it would enable 7 simultaneous predictors of outcome.

1. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat.Med. 1996;15:361-87.

- Further clarification of the excluded children is needed. The figure suggests that 6723 children did not meet the inclusion criteria. Yet the exclusion criteria listed were 'Evident victims of physical child abuse (admitted by perpetrator at presentation), victims of (witnessed) traffic accidents and children who had died before arrival were excluded.' …surely there were not 6000 of these cases?

JS: the total of 11013 ER visits were all ER visits of children under the age of 7 years, so both trauma cases and non trauma cases. To make this more clear, we added 'No injury, or…' to the box 'did not meet inclusion criteria' in the figure.

- Is the 11013 all attendees including medical attendances? It would be useful to know the coverage of all trauma cases …were all given the sputovamo.

JS: The reviewer is right that the 11013 attendees include all medical attendees aged under 7. SPUTOVAMO-R was indeed given to all attendees, as a requirement of the Dutch Health Care Inspectorate (as stated in the introduction). Conform the reviewers suggestion we clarified this also in the manuscript, (see 'checklist' heading, page 65-8)'The checklist was a compulsory field in the electronic files of the medical records of all attendees, regardless to the reason of ER attendance'.

- Should the flow chart start with all injury cases in children < 7 years rather than all cases…this would be a more useful figure?

JS: Based on previous questions about the total childhood population, we chose to start the flow chart with all children <7 years that attended the ER, because this indicates the percentage of trauma cases out of the total population of minor attendees. We are certainly willing to start the flow chart with 'eligible patients' if this indeed makes the figure clearer, but we would like to leave that decision at the discretion of the editor.

- Missing cases and data: What about the cases who did not consent or where there was missing parent data etc? are these cases likely to have influenced the results?

JS: When there was no consent for data evaluation at all (n=37), we could not achieve a final expert panel diagnosis. As explained in the discussion, we could only assess a clinical outcome for these cases and found one child (with a positive screen) to be possibly subjected to inflicted injury. The reviewer makes a very fair point here, but we are convinced that this one case will not have materially influenced our findings.

- there are a significant number where parents did not allow further health care worker contact or could not be contacted themselves Of the missing cases 1/37 had probable abuse and whilst it says this would not affect the results we do not know whether this case had a positive or negative SPUTOVAMO. If negative this would indeed reduce the sensitivity?

JS: This one case had a positive screen. This important comment was also put forward by the first reviewer, and we have now clarified this in the manuscript by adding this information (page 10, line 1)..

- The objective of the SPUTOVAMO was to identify cases of physical abuse, yet secondary outcome measures of neglect and need for help are included. Yet the SPUTOVAMO would seem incompletely designed to identify these cases and indeed was not designed to do so. I am unsure therefore why these measures were included…this should be justified. This association is dismissed in the discussion and I wonder therefore whether it is worth including at all???...If it is then this needs to be justified

JS: the reviewer is right that the argumentation on our decision to include the secondary outcomes was missing in the manuscript. We now explained why we choose for these outcomes by adding the following text to the manuscript (see 'outcome definition' heading. Page 6, line 17-21): 'Because physical abuse contributes relatively few cases of child maltreatment that receive child protection intervention and may distract clinical attention from detecting neglect and emotional abuse, which are far more common and, in the long-term, also damaging(ref), we decided to include injury caused by neglect and need for help from social services as secondary outcomes.'

- What this study shows
'This diagnostic study to detect child abuse is the first study that meets all QUADAS criteria for diagnostic accuracy studies' this is not infact accurate as the study sets out to detect physical abuse as stated in the first line of the abstract . terminology needs to be consistent. The term child abuse covers a wide remit …emotional abuse, sexual abuse etc which are certainly not covered within the SPUTOVAMO. This would apply to the second bullet.

JS: Here we added the word 'physical' where needed.

- In the results 0·27% ((5 + (1*4153/620))/4253) (95% CI 0·15 – 0·49) what does the underlined expression refer to?

JS: We see no underlining, and we are not sure if we understand the question fully.

- Page 11 line 12 is there a word missing? 7 in children with a positive screen

JS: Indeed, the word 'positive' was missing. We have added this (page 10, line 23).

- Page 5 line 25 missing word 'Reference testing in checklist-negatives allows ….missing word?...to determine the negative predictive value of the checklist.

JS: Here we added the word 'researchers'.

- None of the figures have a title

JS: We now added titles to all figures.

- Methods
'The checklist was a compulsory field in the electronic files of the medical records of all attendees.' I am unclear whether this is indeed all attendees or those with an injury?

JS: It was indeed a compulsory field in the electronic files of the medical records of all attendees, so not only for those with an injury. We clarified this by adding the sentence 'regardless to the reason for ER attendance' (page 6, line 6-7).

- Discussion
There appears to be an association between positive sputovamo and later abuse …or at least a significant risk ratio. Physical abuse is rarely a one off event and is pervasive. Yet this is not mentioned in the discussion??? Does this warrant some discussion if included in the results.

JS: We fully agree with the reviewer that physical abuse is rarely a one off event. However, our data do not give real evidence on this statement. The risk ratio we found in our study, only gives rise to speculations, rather than support the statement.

VERSION 2 – REVIEW

REVIEWER	Marion Bailhache CHU de Bordeaux, Pole de pediatrie, F-33000 Bordeaux, France; Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
REVIEW RETURNED	21-Feb-2016

GENERAL COMMENTS	The study is well designed and well conducted. Authors have answered all my questions.

REVIEWER	Ruth Gilbert University College London
REVIEW RETURNED	14-Feb-2016

GENERAL COMMENTS	The authors have carefully addressed the reviewers' comments in full. I have no further comments

REVIEWER	Alison Kemp Cardiff University UK
REVIEW RETURNED	15-Feb-2016

GENERAL COMMENTS	The manuscript has improved from the suggestions made by the reviewers and I would be happy to recommend it for publication.