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.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Diagnostic markers of acute infections in infants aged 1 week to 3 months - a retrospective cohort study
AUTHORS	Hamiel, Uri; Bahat, Hilla; Kozer, Eran; Hamiel, Yotam; Ziv-Baran, Tomer; Goldman, Michael

VERSION 1 - REVIEW

REVIEWER	Nee, Patrick A. Whiston Hospital, Emergency Department
REVIEW RETURNED	07-Mar-2017

GENERAL COMMENTS	This is a retrospective review of a large sample (n=1039) of pyrexial young infants who had bloods drawn for inflammatory markers and cultures during a 2 year study interval. Some markers, singular or combined, were associated with serious or invasive bacterial infection. The tight age group was presumably chosen because of the great variability of lymphocyte count according to age. The study group were likely at the sicker end of the severity spectrum since bloods were drawn. The authors do not state whether the samples were drawn by venepuncture or heel prick. The authors found that CRP and absolute neutrophil count (ANC) were most closely associated with the outcome measures. Because CRP may not be available in their setting, the neutrophil to lymphocyte ratio (NLR) was said to have some utility in the early diagnosis of bacterial infection. Results are tabulated as Sn, Sp and LRs, and ROC curves are drawn, showing values for AUC of 0.73 and 0.82 for the combination of CRP and ANC in SBI and IBI respectively . The means by which these values were combined is not clear. The ROC curves for all parameters does not allow a clear diagnosis, but may add to the determination of prognosis. The authors suggest that some of their measured variables "may add to the suspicion for serious bacterial infection" The incidence of bacteraemia at 1.1% was much lower than that obtained in the adult ED population in a study cited by the present authors (ref 9). Duration of illness before samples were drawn is not reported. Blood cultures were reported as contaminated if more than one organism was isolated. This may not be sufficiently sensitive or specific to diagnose contamination. The present study addresses an important topic, the sample size is good and it is very well written. The downside is that it is a retrospective review, which weakens the conclusions somewhat

	There is little to be said on the presentation of the submission, which is very well written. However, the authors refer to procalcitonin (P 13/26) without further discussion (why isn't this used to determine the risk of SBI in this situation?) and they use the terms "predictive capability" and "predictive power" (Pp12/13). Given the retrospective nature of the present study they may wish to rephrase. I believe predictive power is a quality assigned to a model, which is not derived in the present study. Minor changes would enhance the quality of the paper; the authors should note that CRP is universally available in the UK setting. They should explain what they mean by the "combination" of co-variates CRP and ANC (what, exactly, did they do?). They should also discuss the utility of measures yielding an AUC of 0.73, for example. Is this sufficient to make a diagnosis? Do they recommend drawing bloods in all circumstances in the ED setting? Would heel prick samples suffice? Should infants with indicative tests receive antibiotics? In other words, some discussion on how these findings should be used in the management of pyrexial infants would be useful. Finally, given the variability of ANC and LC according to age, how do they propose to account for age when determining parameters such as NLR?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

Point #1:

"This is a retrospective review of a large sample (n=1039) of pyrexial young infants who had bloods drawn for inflammatory markers and cultures during a 2 year study interval. Some markers, singular or combined, were associated with serious or invasive bacterial infection. The tight age group was presumably chosen because of the great variability of lymphocyte count according to age. The study group were likely at the sicker end of the severity spectrum since bloods were drawn. The authors do not state whether the samples were drawn by venepuncture or heel prick."

Reply to Point #1: In our hospital, blood is drawn from all febrile infants under 3 months of age. We have clarified this in the manuscript (page 5, line 8 now reads: "Blood was drawn from all febrile infants who were admitted to the ED…"). All samples were drawn by venepuncture, we have added this to our manuscript (page 5, line 22 now reads: Samples were drawn by venepuncture")

Point #2:

The authors found that CRP and absolute neutrophil count (ANC) were most closely associated with the outcome measures. Because CRP may not be available in their setting, the neutrophil to lymphocyte ratio (NLR) was said to have some utility in the early diagnosis of bacterial infection. Results are tabulated as Sn, Sp and LRs, and ROC curves are drawn, showing values for AUC of 0.73 and 0.82 for the combination of CRP and ANC in SBI and IBI respectively. The means by which these values were combined is not clear.

Reply to Point #2:

We used multivariate logistic regression to evaluate the probability that a child has an SBI/IBI. The following is now stated on page 7, line 7: “The multivariate logistic regression included the infection markers studied, and the probability calculated was the basis for the ROC curve analysis.“

Point #3:

The ROC curves for all parameters does not allow a clear diagnosis, but may add to the determination of prognosis. The authors suggest that some of their measured variables “may add to the suspicion for serious bacterial infection”. The incidence of bacteraemia at 1.1% was much lower than that obtained in the adult ED population in a study cited by the present authors (ref 9). Duration of illness before samples were drawn is not reported. Blood cultures were reported as contaminated if more than one organism was isolated. This may not be sufficiently sensitive or specific to diagnose contamination.

Reply to Point #3: Culture results were reviewed by a pediatric infectious specialist, and were deemed as either true or contaminated. However, if more than one organism was isolated, the culture was deemed as contaminated and excluded from the final analysis. Our goal was to avoid misclassification bias. We have clarified this in the manuscript (page 6, line 16 now reads: “Blood cultures were considered contaminated by pathogens and by the clinical course of patients, following review of a pediatric infectious specialist”).

Point #4: The present study addresses an important topic; the sample size is good and it is very well written. The downside is that it is a retrospective review, which weakens the conclusions somewhat There is little to be said on the presentation of the submission, which is very well written. However, the authors refer to procalcitonin (P13/26) without further discussion (why isn’t this used to determine the risk of SBI in this situation?)

Reply to Point #4: In this study, we aimed to examine commonly available diagnostic markers. We did not examine procalcitonin, as it was not readily available in our medical center at the time of the study. We have clarified this in the manuscript (page 14, line 5 now reads: “Our study did not examine procalcitonin, since our aim was to study commonly available diagnostic markers.”)

Point #5: and they use the terms “predictive capability” and “predictive power” (Pp12/13). Given the retrospective nature of the present study they may wish to rephrase. I believe predictive power is a quality assigned to a model, which is not derived in the present study.

Reply to Point #5: We have rephrased “predictive power” to “discriminatory ability” (page 4 line 23) and throughout the manuscript (page2, line 16; page 2, line 19; page 2, line 21; page 9, line 3; page 9, line 7; page 11, line 15; page 14, line 13; page 15 line 9)

Point #6 Minor changes would enhance the quality of the paper; the authors should note that CRP is universally available in the UK setting. They should explain what they mean by the “combination” of co-variates CRP and ANC (what, exactly, did they do?).

Reply to Point #6: We used multivariate logistic regression to evaluate the probability that a child has an SBI/IBI. This is now stated on page 7, line 7: "The multivariate logistic regression included the infection markers studied, and the probability calculated was the basis for the ROC curve analysis."

Point #7: They should also discuss the utility of measures yielding an AUC of 0.73, for example. Is this sufficient to make a diagnosis? Do they recommend drawing bloods in all circumstances in the ED setting? Would heel prick samples suffice? Should infants with indicative tests receive antibiotics? In other words, some discussion on how these findings should be used in the management of pyrexial infants would be useful.

Reply to Point #7: We have addressed the issues mentioned (page 14, line 15 now reads: "We recommend drawing blood for all febrile infants aged 3 months or less, and suggest using the cutoff values we determined, as well as other available ones, to aid in the management of febrile infants. The specificity of the markers studied is not sufficient to rule out bacterial infections. However, due to the reasonably high sensitivity, we recommend antibiotic use for all patients with one or more tests indicative of a possible bacterial infection, as well as for ill-looking patients.")

Point #8: Finally, given the variability of ANC and LC according to age, how do they propose to account for age when determining parameters such as NLR?

Reply to Point #8: In an effort to account for the variability of the blood count according to age, we analyzed 2 age groups separately. The mean NLR in the medical literature is 0.63-0.91 for the 1-week to 1-month age group, and 0.52-0.63 for the 1-month to 3-months age group. We suggest that an NLR of >2 should raise suspicion for an SBI. In another analysis, we assessed an "adjusted NLR ratio", as the calculated NLR divided by the mean NLR for age subgroups. There was no statistical difference between the NLR and the "adjusted NLR ratio" in the assessment of SBI. This is now mentioned on page 6, line 10: "An age-adjusted NLR ratio was also created, by dividing NLR by a mean NLR based on the medical literature,[15] according to age groups (1-2 weeks, 2 weeks-1 month, 1 month and older)", and on page 8, line 23 in the results section: "There was no statistically significant difference in the assessment of SBI between the unadjusted NLR ratio and the adjusted for age NLR ratio."

VERSION 2 – REVIEW

REVIEWER	Peter Watson University of Cambridge UK
REVIEW RETURNED	14-Jul-2017

GENERAL COMMENTS	Diagnostic markers of acute infections in infants aged 1 week to 3 months - a retrospective cohort study. bmjopen-2017-018092 Please find a few comments below which I hope might improve the results. In brief I don't see the need for the CHAID analyses or thresholding of the NLR and other predictors as given in Tables 2

	and 3 since they are not seemingly thresholded in the logistic regressions reported on pages 18-19 and I assume can be measured more informatively as a raw, rather than as a dichotomised, score. I would like to see the sensitivities and specificities reported for the best fitting models to IBI and SBI which are those involving CRP and either of ANC and NLR especially given one of the reported sensitivities and specificities in these tables for each of the inferior models are low performing around chance (50%) or below. I believe these changes would give a more clear and coherent account. Area under ROC curves is a standard approach to assessing the predictive accuracy of logistic regressions although measures such as total correctly classified might be more intuitive to clinicians. One also has to mention that the classification rates will be optimistic when the same set of data is used both to construct the logistic regression decision rule and to classify. CHAID (CARTS) also presented in this paper is an alternative approach - I am not sure, however, what CHAID adds to the logistic regression results? Aren't they both asking the same question - namely how good are various predictors at classifying SBI/IBI? Couldn't you simply use the logistic regression with the raw scores to then conclude by advocating the use of the classifiers (CRP with either ANC or NRL) in diagnostic testing for SBI and IBI based upon the areas under the ROC curve (as you do in the discussion on page 20, lines 11-13). It could be added that some of the areas under the ROC curve for the best fitting models presented at the bottom of page 18 and top of page 19 are deemed 'acceptable' for ANC or NLR added to CRP for SBI and 'excellent' for IBI (lines 5-11 on page 19) using the thresholds suggested by Hosmer and Lemeshow (2000). Hosmer and Lemeshow (2000) suggest areas under the ROC curve of 0.70 to 0.80 are 'acceptable', 0.80 to 0.90 'excellent' and 0.9 or above 'outstanding'. They point out an area under the ROC curve of 0.50 suggests no discrimination between the outcome groups as this corresponds to chance e.g. simply tossing a coin to decide group membership. Looking at Tables 2 and 3 on pages 27-28, however, most of the thresholded values have either a sensitivity or specificity of only around 50% or below (chance or below) which might suggest sizeable preponderences of false negative or false positive results. These low values of sensitivity and specificity may be related to low prevalences of SBI in various groups with smaller prevalences, for example, perhaps making these less prevalent bacterial infections harder to predict. The fit measures in Tables 2 and 3 appear to be for single predictors so do not look at the best fitting two predictor models (according to the AUC criteria on page 19, lines 2-4 and lines 9-11) for SBI and IBI namely CRP with ANC or CRP with NRL. It would be interesting and more reassuring, therefore, for the clinician to see the specificities and sensitivities for these joint models. You could simply do a logistic regression with CRP and ANC and CRP and NRL in and obtain the sensitivities (e.g. P(SBI=-| predicted to be a '-')) and specificities (e.g. P(SBI=+| predicted to be a '+')). I am also not clear why you are thresholding the predictors in Tables 2 and 3. I assume the thresholds used in Tables 2 and 3 came from CHAID? Couldn't you simply do a logistic regression with the raw NLR, CRP and other scores and work out sensitivities and

	specificities based upon their unthresholded values as appears to be the case in Table 4 and its results on lines 1-4 of page 19 plus the IBI AUC results on lines 6-11 on page 19 which do not refer to any thresholds thus tying in the results from these logistic regression models with their sensitivities and specificities? Page 29, Table 4. You could add the areas under the curve reported here are for SBI - I think from page 19, line 4 it is for SBI? Why no table giving the areas under the curve for IBI as well? Do we need to keep any of the ROC curves figures? I am not convinced that these are adding anything to the results concerning areas under the ROC curve reported on pages 18-19.

REVIEWER	Patrick Nee Faculty of Education, Health and Community, Liverpool John Moores University, Liverpool, United Kingdom
REVIEW RETURNED	25-Jul-2017

GENERAL COMMENTS	No further comments.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1:

Point #1:
In brief I don't see the need for the CHAID analyses or thresholding of the NLR and other predictors as given in Tables 2 and 3 since they are not seemingly thresholded in the logistic regressions reported on pages 18-19 and I assume can be measured more informatively as a raw, rather than as a dichotomised, score. I would like to see the sensitivities and specificities reported for the best fitting models to IBI and SBI which are those involving CRP and either of ANC and NLR especially given one of the reported sensitivities and specificities in these tables for each of the inferior models are low performing around chance (50%) or below. I believe these changes would give a more clear and coherent account.

Area under ROC curves is a standard approach to assessing the predictive accuracy of logistic regressions although measures such as total correctly classified might be more intuitive to clinicians. One also has to mention that the classification rates will be optimistic when the same set of data is used both to construct the logistic regression decision rule and to classify. CHAID (CARTS) also presented in this paper is an alternative approach - I am not sure, however, what CHAID adds to the logistic regression results? Aren't they both asking the same question - namely how good are various predictors at classifying SBI/IBI? Couldn't you simply use the logistic regression with the raw scores to then conclude by advocating the use of the classifiers (CRP with either ANC or NRL) in diagnostic testing for SBI and IBI based upon the areas under the ROC curve (as you do in the discussion on page 20, lines 11-13).

It could be added that some of the areas under the ROC curve for the best fitting models presented at the bottom of page 18 and top of page 19 are deemed 'acceptable' for ANC or NLR added to CRP for SBI and 'excellent' for IBI (lines 5-11 on page 19) using the thresholds suggested by Hosmer and Lemeshow (2000). Hosmer and Lemeshow (2000) suggest areas under the ROC curve of 0.70 to 0.80 are 'acceptable', 0.80 to 0.90 'excellent' and 0.9 or above 'outstanding'. They point out an area under the ROC curve of 0.50 suggests no discrimination between the outcome groups as this corresponds to chance e.g. simply tossing a coin to decide group membership.

Looking at Tables 2 and 3 on pages 27-28, however, most of the thresholded values have either a sensitivity or specificity of only around 50% or below (chance or below) which might suggest sizeable preponderences of false negative or false positive results. These low values of sensitivity and specificity may be related to low prevalences of SBI in various groups with smaller prevalences, for example, perhaps making these less prevalent bacterial infections harder to predict. The fit measures in Tables 2 and 3 appear to be for single predictors so do not look at the best fitting two predictor models (according to the AUC criteria on page 19, lines 2-4 and lines 9-11) for SBI and IBI namely CRP with ANC or CRP with NRL.

It would be interesting and more reassuring, therefore, for the clinician to see the specificities and sensitivities for these joint models. You could simply do a logistic regression with CRP and ANC and CRP and NRL in and obtain the sensitivities (e.g. P(SBI=-| predicted to be a '-')) and specificities (e.g. P(SBI=+| predicted to be a '+')).

I am also not clear why you are thresholding the predictors in Tables 2 and 3. I assume the thresholds used in Tables 2 and 3 came from CHAID? Couldn't you simply do a logistic regression with the raw NLR, CRP and other scores and work out sensitivities and specificities based upon their unthresholded values as appears to be the case in Table 4 and its results on lines 1-4 of page 19 plus the IBI AUC results on lines 6-11 on page 19 which do not refer to any thresholds thus tying in the results from these logistic regression models with their sensitivities and specificities?

Reply to Point #1:
We thank the reviewer for this comment. The aim of this study was to examine various biomarkers as discriminators for SBI/IBI, to be used as a practical tool for clinicians. Moreover, we investigated if a combination of biomarkers could serve as a significantly better discrimination tool.
We arbitrarily examined both commonly used cutoff values (when available, for example for CRP and WBC), and cutoff values that would be easy for the clinician to use (for example for NLR), and described their discrimination ability in Tables 2 and 3. We also used classification trees to identify additional non-intuitive cutoff values.

The study did not use multivariate logistic regression to identify predictors for SBI, or to build a prediction model, but only used multivariate logistic regression to evaluate the added value of the combination of biomarkers. The discriminative ability of the combination of the biomarkers was ultimately found to be similar to that of the single biomarkers.
In Table 1 we reported the crude data for young infants with and without SBI, which showed a statistically significant difference in biomarkers between infants with and without SBI; while Tables 2 and 3 added data on their discrimination ability for various cutoff values.

The cutoff values of CRP, ANC and NLR that we analyzed in this study, are easy for the clinician to use and implement. Since the combinations of the biomarkers showed similar discrimination ability to that of the single biomarkers, we do not believe that the use of a combination of markers rather than single markers is of priority, for this would be much more difficult for clinicians. We understand that this point was not clear enough in the manuscript, and we have included a number of additions in order to clarify it:

1. Page 12 second paragraph now reads: “Hosmer and Lemeshow suggest that areas under the ROC curve of 0.70 to 0.80 offer 'acceptable' discrimination, 0.80 to 0.90 'excellent' discrimination and 0.9 or above offer 'outstanding' discrimination.[18] Thus, in assessment of SBI, values of ANC (AUC 0.69) and CRP (AUC 0.71), along with the combinations of CRP with either ANC (AUC 0.73) or NLR (0.72), offer similarly ‘acceptable’ discriminative ability. In assessing IBI, values of CRP, ANC and NLR, as

well as the combination of CRP with either NLR or ANC, similarly offer ‘excellent’ or close to excellent discriminations. In the neonatal age group, all markers mentioned above meet the 'acceptable' criterion. Due to the ease of use of the single biomarkers compared to the combinations, and the similarity of their discriminative abilities, we recommend clinicians to use the markers separately rather than creating a combined score.”

2. Page 8 last paragraph: “Tables 2 and 3 show sensitivities, specificities and ratio values of WBC, CRP and NLR, for cutoff values that were arbitrarily chosen either due to their common use in clinical practice or to their ease of use (for example in the case of NLR), for the discrimination of SBI.”

3. Page 13, line 30: “In our search for non-intuitive cutoff values, we created a decision tree (Figure 4) that shows the added value of NLR to CRP in assessing febrile neonates.”

Point #2:
Page 29, Table 4. You could add the areas under the curve reported here are for SBI - I think from page 19, line 4 it is for SBI? Why no table giving the areas under the curve for IBI as well?
Reply to Point #2: We added AUCs for IBI, as suggested (Table 4)

Point #3:
Do we need to keep any of the ROC curves figures? I am not convinced that these are adding anything to the results concerning areas under the ROC curve reported on pages 18-19.
Reply to Point #3: The figures add only a graphic visualization, so we do not insist on their inclusion. We suggest perhaps to leave this matter to the discretion of the editor.