PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A scoping review of malaria forecasting: Past work and future directions
AUTHORS	Zinszer, Kate; Verman, Aman; Charland, Katia; Brewer, Timothy; Brownstein, John; Sun, Zhuoyu; Buckeridge, David

VERSION 1 - REVIEW

REVIEWER	Olivier J.T. Briët, PhD Senior Malariologist Swiss Tropical and Public Health Institute Switzerland I have no competing interests.
REVIEW RETURNED	06-Sep-2012

THE STUDY	Is the research question clearly defined? P2L6 “Objectives: The diversity of malaria forecasting methods has created difficulties in identifying the optimal predictors and methods that would provide the most accurate malaria forecasts. The objective of our review is to identify and assess methods, including predictors, used to forecast malaria.” I would not blame the diversity of methods for creating “difficulties”. Diversity is a good thing. As alternative, I suggest something along the lines of: ‘Objectives: There is a growing literature on malaria forecasting methods. The objective of our review is to identify and assess methods and predictors that forecast malaria well’ Similarly: P3L25 “The diversity in forecasting methods has prevented comparisons of forecasting results, making it difficult to identify the optimal predictors and methods for malaria forecasting.” The larger the diversity, the more comparisons can be made. The problem appears to be that many authors present only their method applied to their data, and do not compare it to other methods (on their data). P5L25 “Although significant malaria predictors have been identified in different settings, the diversity in forecasting methods has hampered comparisons of results, making it difficult to identify the optimal predictors and methods for malaria forecasting.” See my point under P3L25. The problem statement needs to be more accurate. Malaria forecasting methods are likely to vary in performance depending on setting and forecasting horizon. Thus, there might not be “optimal predictors and methods” that work for all settings. However, if researchers would compare multiple methods for their settings and horizons, a review such as this could help assess if there are some methods that do well in a majority of

	settings and can be part of an ensemble of recommended methods. Are the abstract/summary/key messages/limitations accurate? P4L8 “and the recommendations in the review, if followed, will lead to improvement in the quality and public health impact of malaria forecasting.” There is no evidence for this, so I suggest to remove this ‘strength’. P4L16 “The key limitation of this review is that potential details regarding methodological approaches in the studies may have been missed due to these details being excluded from the published manuscript.” This is an inherent limitation of any literature based review, and therefore a bit obvious. Suggest to change into: ‘A limitation of a literature review is that unpublished methods, if any, are omitted from this review. Do any supplemental documents e.g. a CONSORT checklist, contain information that should be better reported in the manuscript, or raise questions about the work? No, everything is fine. Editor, please rephrase this question.
RESULTS & CONCLUSIONS	Are the interpretation and conclusions warranted by and sufficiently derived from/focused on the data? P27L15 “Various direct measures were used to estimate forecasting error. Absolute measures, such as the mean absolute error, are relevant for measuring accuracy within a particular series but not across series because the magnitude of the mean absolute error depends on the scale of the data [51]. Percent errors, such as mean absolute percent error are scale-independent but are not recommended when the data involves small or 0 counts. In economics, a measure called mean absolute scaled error has been recommended as a forecast-accuracy metric for forecasting [51]. We recommend incorporating mean absolute scaled error into malaria forecast evaluation as this evaluation measure will facilitate comparison between studies, but we also recommend reporting mean absolute error as this metric allows an intuitive interpretation of the errors.” I agree with the authors that 0 counts and numbers between 1 and 0 give problems for percent error measures. However, percent error methods have the advantage that they scale variance: Malaria case counts are essentially Poisson distributed (or distributed according to a mixture of Poisson distributions). As such, the variance increases with the mean. An error of 10 on a mean of 100 is thus equivalent to an error of 1 on a mean of 10. Malaria is often very seasonal, and as such, has variable variance. The MASE proposed by Hyndman & Koehler, albeit a useful measure, does not appear to be dealing with this. Therefore, I would be reluctant to drop the percent error measure. The problem with 0 counts could be dealt with by data transformation, e.g. adding a small number such as 1, to all counts. The MASE is recommended by Hyndman & Koehler “for comparing forecast accuracy across series on different scales”. Coming back to my comment on P5L25, I think that a MASE to compare across data series (from different sites) could be useful, but that methods can also be compared by ranking their performance

	based on perhaps more appropriate non-scale free measures or percent errors based methods, and comparing rankings across sites.
GENERAL COMMENTS	This paper was a pleasure to read and review. I highly recommend publication. However, I have a few suggestions for improvement, hence the "minor revision" recommendation. P65L58 “We excluded studies that provided only … mortality predictions” Why was the one study found on mortality predictions excluded? If mortality due to malaria can be forecasted well, this is a much more important public health outcome than uncomplicated cases. P16L18 “Forecasting methods” Just like was done in the preceding paragraph, this section could benefit from quantification of the proportion of studies that used certain methods, after statements. E.g.: ‘Several of the studies (X%) used a GLM’. Particularly, it would be helpful to have a summary table, like Table 2, of methods used across studies. What proportion uses GLM? What proportion uses ARIMA? The main conclusion of this work, the fact that not enough researchers compare performance of several methods, should be illustrated clearly by showing how few of the studies compared several methods, and which methods they compared. P17L27 “The models included in the review were of the basic form, GM(1,1), which implies that this is a univariate model (malaria counts only) and the solution is the result of solving a single differential equation.” Perhaps make it clear that these are the four Chinese models. P22L46 “There should also be continued effort to develop new methods although common forecasting metrics are essential as they will help determine the optimal approach with existing and future methods.” What is meant by “metrics”? ‘Measures’? P23L3 “One should explore non-climate predictors as well as different forecasting approaches based upon the same data.” Do the authors mean univariate timeseries with “the same data”? P23L8 “Advantages and disadvantages of forecasting methods” This long section has some overlap with “Forecasting methods”. I suggest to remove overlap and integrate these sections. The authors go into details of auto-correlation in residuals and pre-whitening, but these are more relevant for regression and correlation analysis, and not very relevant to forecasting per se. As long as the method performs well on out of sample data, these things appear less relevant (a black-box view, like with neural networks, might be appropriate?). I would suggest to shorten this discussion on advantages and disadvantages, and focus on suitability of methods depending on the data time series. E.g. that SARIMA models do not work well on short series. P25L47 “Additionally, neural networks have a greater susceptibility to overfitting [45] and several thousand observations are typically required to fit a neural network with confidence [46].” This might merit some discussion that malaria time series rarely contain several hundred observations, let alone several thousand. P26L32 “Additionally, the findings derived from forecasting models

	based upon clinical confirmation of malaria are likely subject to error, due to the poor specificity of clinical case definitions for malaria." This sentence seems outside the scope of the paper. Malaria researchers know that lab-confirmed data are better than clinical case data, but sometimes the latter is all one can work with. P28L11 "Applying different forecasting methods to the same data, exploring the predictive ability of non-environmental variables, and using common forecast accuracy metrics will allow malaria researchers to compare and improve models and methods, and lead to the improvement in the quality and public health impact of malaria forecasting." Optional: Perhaps is it good to point out that if public health systems are reactive to forecasts using transmission reducing interventions, this in turn might weaken forecast accuracy, if transmission reducing interventions are not used as covariate in the forecasting model. Throughout (Perhaps an issue with the type setting): The seperation of superscript reference numbers by normal script commas is a bit unusual.

REVIEWER	DR. RAM RUP SARKAR SENIOR SCIENTIST Chemical Engineering Sciences (CEPD) CSIR-National Chemical Laboratory No competing interest.
REVIEW RETURNED	21-Sep-2012

GENERAL COMMENTS	This is a timely review article that aims to assess the current state of the malaria modelling field, predominantly focusing on statistical, mathematical and machine learning methods as well as models to identify and assess methods, including predictors, used to forecast malaria.. While I agree with the broad aims of the paper and I believe that this sort of review will be useful, I feel that the authors need to improve the manuscript further in some key areas before a final decision can be taken. My main concern with this article is how the "major" malaria models/methods are selected. In principle, I'm OK with a brief focus on the key models/methods selected, but I would like to know how the authors decide the key word combination and truncation. This should be elaborated further. Moreover, I feel some other options/combinations should be included to widen the scope of the review. If this is not taken seriously the results/interpretation could be misleading as the authors may have ignored some interesting studies in this direction. At the moment, the article feels very incomplete in terms of its scope. The authors should throw more light on different deterministic and stochastic models or should explain what's the motivation for excluding such models? Or is this somehow motivated by the author's findings when the systematic review was undertaken? If stochastic models have been deliberately omitted, this may give a misleading bias to the findings of this paper. A figure depicting the time evolution/hierarchial structure of different methods/models could have been interesting for the readers to

	easily understand the scope of the review. Table 1 requires editing and a more summarised version may be presented in the main paper, while the details can go to the supplementary information. The article needs more discussion on the malaria predictors, covariates and measures, there choice, outcome and impact on forecasting methods so as to clarify the broad objective of the authors. Overall, I therefore think that this paper has some merit and would be a useful contribution to the field, but I think needs further work before acceptance could be recommended. My main concern is how thorough a review of the literature this really is. I would prefer to see this sort of article really pull out the generic concepts, structures and differences between the methods and models. The focus on the current models is OK, but should probably widen the scope to give a more comprehensive overview

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
Olivier J.T. Briët, PhD
Senior Malariologist
Swiss Tropical and Public Health Institute Switzerland

Comment #1:
P2L6 “Objectives: The diversity of malaria forecasting methods has created difficulties in identifying the optimal predictors and methods that would provide the most accurate malaria forecasts. The objective of our review is to identify and assess methods, including predictors, used to forecast malaria.”
I would not blame the diversity of methods for creating “difficulties”. Diversity is a good thing. As alternative, I suggest something along the lines of: ‘Objectives: There is a growing literature on malaria forecasting methods. The objective of our review is to identify and assess methods and predictors that forecast malaria well’

Response #1:
We agree. Diversity in methods is not the problem, the problem is due to the difficulties in drawing comparisons between the studies. This difficulty is due, in part, to a diversity of forecasting accuracy measures used as well as due to a lack of comparison made between studies. Also, as the Reviewer mentions, there will not be one model or method that suits all settings. We have made the appropriate changes in the manuscript.

Comment #2:
Similarly:
P3L25 “The diversity in forecasting methods has prevented comparisons of forecasting results, making it difficult to identify the optimal predictors and methods for malaria forecasting.”
The larger the diversity, the more comparisons can be made. The problem appears to be that many authors present only their method applied to their data, and do not compare it to other methods (on their data).

Response #2:
Please see Response #1.

Comment #3:
P5L25 “Although significant malaria predictors have been identified in different settings, the diversity

in forecasting methods has hampered comparisons of results, making it difficult to identify the optimal predictors and methods for malaria forecasting."
See my point under P3L25. The problem statement needs to be more accurate. Malaria forecasting methods are likely to vary in performance depending on setting and forecasting horizon. Thus, there might not be "optimal predictors and methods" that work for all settings. However, if researchers would compare multiple methods for their settings and horizons, a review such as this could help assess if there are some methods that do well in a majority of settings and can be part of an ensemble of recommended methods.

Response #3:
Please see Response #1.

Comment #4:
P4L8 "and the recommendations in the review, if followed, will lead to improvement in the quality and public health impact of malaria forecasting."
There is no evidence for this, so I suggest to remove this 'strength'.

Response #4:
This strength has been reworded and contains a reasonable assumption, in our opinion: "...and the recommendations in the review, if followed, should lead to improvement in the quality of malaria forecasting".

Comment #5:
P4L16 "The key limitation of this review is that potential details regarding methodological approaches in the studies may have been missed due to these details being excluded from the published manuscript."
This is an inherent limitation of any literature based review, and therefore a bit obvious. Suggest to change into: 'A limitation of a literature review is that unpublished methods, if any, are omitted from this review.

Response #5:
Fair point and the text has been changed.

Comment #6
P27L15 "Various direct measures were used to estimate forecasting error. Absolute measures, such as the mean absolute error, are relevant for measuring accuracy within a particular series but not across series because the magnitude of the mean absolute error depends on the scale of the data [51]. Percent errors, such as mean absolute percent error are scale-independent but are not recommended when the data involves small or 0 counts. In economics, a measure called mean absolute scaled error has been recommended as a forecast-accuracy metric for forecasting [51]. We recommend incorporating mean absolute scaled error into malaria forecast evaluation as this evaluation measure will facilitate comparison between studies, but we also recommend reporting mean absolute error as this metric allows an intuitive interpretation of the errors."
I agree with the authors that 0 counts and numbers between 1 and 0 give problems for percent error measures. However, percent error methods have the advantage that they scale variance: Malaria case counts are essentially Poisson distributed (or distributed according to a mixture of Poisson distributions). As such, the variance increases with the mean. An error of 10 on a mean of 100 is thus equivalent to an error of 1 on a mean of 10. Malaria is often very seasonal, and as such, has variable variance. The MASE proposed by Hyndman & Koehler, albeit a useful measure, does not appear to be dealing with this. Therefore, I would be reluctant to drop the percent error measure. The problem with 0 counts could be dealt with by data transformation, e.g. adding a small number such as 1, to all counts. The MASE is recommended by Hyndman & Koehler "for comparing forecast accuracy across

series on different scales". Coming back to my comment on P5L25, I think that a MASE to compare across data series (from different sites) could be useful, but that methods can also be compared by ranking their performance based on perhaps more appropriate non-scale free measures or percent errors based methods, and comparing rankings across sites.

Response #6:
Scale variance is an important point that should be considered and the Reviewer is correct in that the MASE does not account for this. Please see the table below as an example of the performance for different measures of forecasting accuracy. The addition of a constant to replace 0 counts would offer a solution to one of current limitations of MAPE, as the Reviewer suggested. We also should have mentioned another limitation of the MAPE measure is that it places a heavier penalty on forecasts that exceed the observed counts versus those that are less than the observed (see forecast horizon 6 and 7 below). We believe that the MASE is a measure that is useful in allowing cross data series comparisons, despite its limitation of not accounting for scale variance. We have changed our recommendation to include MAPE as a measure of forecasting accuracy, given the importance of scale variance despite the penalty issue mentioned above. Both the MAPE and MASE could be used to compare across series and future work could examine how these two measures differ with respect to the ranking of methods. The appropriate changes have been made in the manuscript.

Forecast horizon	Observed	Predicted	MAEa	MSEb	MAPEc	MASEd
1	55	45	10	100	18.18	0.69
2	70	60	10	100	14.29	0.69
3	75	80	5	25	6.67	0.35
4	60	55	5	25	8.33	0.35
5	50	60	10	100	20	0.69
6	35	20	15	225	42.86	1.04
7	20	35	15	225	75	1.04
8	0	10	10	100	*	0.69
9	15	5	10	100	66.67	0.69
10	35	30	5	25	14.29	0.35
Average	102.5	9.5	29.59	0.66		

aMAE: Mean Absolute Error; bMSE: Mean Square Error; cMAPE: Mean Absolute Percent Error
dMASE: Mean Absolute Scaled Error

Comment #7:
This paper was a pleasure to read and review. I highly recommend publication. However, I have a few suggestions for improvement, hence the "minor revision" recommendation.

Response #7:
We appreciate your insightful comments and suggestions which have allowed us to improve the quality of this manuscript.

Comment #8:
P65L58 "We excluded studies that provided only … mortality predictions"
Why was the one study found on mortality predictions excluded? If mortality due to malaria can be

forecasted well, this is a much more important public health outcome than uncomplicated cases.

Response #8:
We agree that accurate forecasting of malaria-related mortality would be of significant public health benefit. Given that there are so few malaria mortality forecasting studies and the predictors of mortality would be different to that of malaria morbidity, we decided to exclude malaria mortality.

Comment #9:
P16L18 “Forecasting methods”
Just like was done in the preceding paragraph, this section could benefit from quantification of the proportion of studies that used certain methods, after statements. E.g.: ‘Several of the studies (X%) used a GLM’. Particularly, it would be helpful to have a summary table, like Table 2, of methods used across studies. What proportion uses GLM? What proportion uses ARIMA? The main conclusion of this work, the fact that not enough researchers compare performance of several methods, should be illustrated clearly by showing how few of the studies compared several methods, and which methods they compared.

Response #9:
Good point and an additional table has been added (Table 2) as well as numbers inserted in the text.

Comment #10:
P17L27 “The models included in the review were of the basic form, GM(1,1), which implies that this is a univariate model (malaria counts only) and the solution is the result of solving a single differential equation.”
Perhaps make it clear that these are the four Chinese models.

Response #10:
We added the term Grey to improve clarity of this sentence: “The Grey models included in the review...”.

Comment #11:
P22L46 “There should also be continued effort to develop new methods although common forecasting metrics are essential as they will help determine the optimal approach with existing and future methods.”
What is meant by “metrics”? ‘Measures’?

Response #11:
We used the term metric to mean a system of measurement but to lessen confusion, we have changed it to measures.

Comment #12:
P23L3 “One should explore non-climate predictors as well as different forecasting approaches based upon the same data.”
Do the authors mean univariate timeseries with “the same data”?

Response #12:
No, what we meant was the predictive value of non-climate predictors should be explored and potentially included in multivariate time series models, which is separate from the issue of applying

different forecasting approaches on the same data. This has been clarified in the text.

Comment #13:
P23L8 “Advantages and disadvantages of forecasting methods”
This long section has some overlap with “Forecasting methods”. I suggest to remove overlap and integrate these sections. The authors go into details of auto-correlation in residuals and pre-whitening, but these are more relevant for regression and correlation analysis, and not very relevant to forecasting per se. As long as the method performs well on out of sample data, these things appear less relevant (a black-box view, like with neural networks, might be appropriate?). I would suggest to shorten this discussion on advantages and disadvantages, and focus on suitability of methods depending on the data time series. E.g. that SARIMA models do not work well on short series.

Response #13:
Based upon the comment above, we have removed any generic method description from the Results (Forecasting methods) to the Discussion (Advantages and disadvantages of forecasting methods), to reduce overlap between the sections. Also, we have added a sentence in regards to the amount of data required for a SARIMA model. As stated in the “Differences between forecasting methods” section (previously “Advantages and disadvantages of forecasting methods”), we do feel that accounting for auto-correlation in the data is important for prediction, and verifying that it was properly dealt with through checking residuals and performing pre-whitening. It is important because if it is not properly accounted for, the standard errors will be incorrect and confidence intervals are an important piece of information needed regarding the precision of the expected value and expected range of values. Additionally, it could lead to biased estimates of effect of the different predictors in the model.

Comment #14:
P25L47 “Additionally, neural networks have a greater susceptibility to overfitting [45] and several thousand observations are typically required to fit a neural network with confidence [46].”
This might merit some discussion that malaria time series rarely contain several hundred observations, let alone several thousand.

Response #14:
We agree and a brief sentence has been added to the manuscript.

Comment #15:
P26L32 “Additionally, the findings derived from forecasting models based upon clinical confirmation of malaria are likely subject to error, due to the poor specificity of clinical case definitions for malaria.”
This sentence seems outside the scope of the paper. Malaria researchers know that lab-confirmed data are better than clinical case data, but sometimes the latter is all one can work with.

Response #15:
This is a valid point, although the intended purpose with this sentence was to caution when comparing forecast accuracy between studies (or series) if comparing clinical vs. laboratory confirmation of malaria, due to the poor specificity of clinical case definition. Taking into consideration of your suggestion, this sentence has been removed.

Comment #16:
P28L11 “Applying different forecasting methods to the same data, exploring the predictive ability of non-environmental variables, and using common forecast accuracy metrics will allow malaria researchers to compare and improve models and methods, and lead to the improvement in the quality

and public health impact of malaria forecasting."
Optional: Perhaps is it good to point out that if public health systems are reactive to forecasts using transmission reducing interventions, this in turn might weaken forecast accuracy, if transmission reducing interventions are not used as covariate in the forecasting model.

Response #16:
This is a great point and also related to our suggestion about exploring non-environmental predictors. This has been added to the manuscript.

Reviewer 2
Reviewer: DR. RAM RUP SARKAR
SENIOR SCIENTIST
Chemical Engineering Sciences (CEPD)
CSIR-National Chemical Laboratory
Dr. Homi Bhabha Road, Pune 411 008, Maharasthra, India

No competing interest.

This is a timely review article that aims to assess the current state of the malaria modelling field, predominantly focusing on statistical, mathematical and machine learning methods as well as models to identify and assess methods, including predictors, used to forecast malaria.. While I agree with the broad aims of the paper and I believe that this sort of review will be useful, I feel that the authors need to improve the manuscript further in some key areas before a final decision can be taken. My main concern with this article is how the "major" malaria models/methods are selected.

Comment #1:
In principle, I'm OK with a brief focus on the key models/methods selected, but I would like to know how the authors decide the key word combination and truncation. This should be elaborated further. Moreover, I feel some other options/combinations should be included to widen the scope of the review. If this is not taken seriously the results/interpretation could be misleading as the authors may have ignored some interesting studies in this direction. At the moment, the article feels very incomplete in terms of its scope.

Response #1:
As briefly mentioned in the manuscript, the search terms were created under the guidance of a librarian who specializes in systematic reviews. We tested a variety of approaches, such as the inclusion of prediction (versus prediction model or predictive model) and found that it was not specific enough and did not demonstrate enhanced sensitivity with respect to capturing temporal forecasting studies of malaria. We did not search based upon type of method and feel strongly that our search terms were sensitive and captured all studies that met our inclusion criteria as specified in the Methods (pages 5-7). As shown in Figure 1, we did capture 613 citations after removing duplicates.

We also searched the literature without using our prescribed search terms to ensure the inclusion of studies that met our criteria. We did not identify any outlying studies that were missed.

Comment #2:
The authors should throw more light on different deterministic and stochastic models or should explain what's the motivation for excluding such models? Or is this somehow motivated by the author's findings when the systematic review was undertaken? If stochastic models have been deliberately omitted, this may give a misleading bias to the findings of this paper.

Response #2:
As mentioned in Response #1, we did not exclude studies based upon methods. Each study was required to meet a list of criteria that were decided before the searches commenced. We are assuming that stochastic is referring to mathematical models that make stochastic predictions and after applying our inclusion criteria, there were no stochastic models included in our study.

Comment #3:
A figure depicting the time evolution/hierarchial structure of different methods/models could have been interesting for the readers to easily understand the scope of the review.

Response #3:
We agree that a figure depicting the time evolution of the different methods would be quite interesting for the readers although this is beyond the scope of the paper.

Comment #4:
Table 1 requires editing and a more summarised version may be presented in the main paper, while the details can go to the supplementary information.

Response #4:
We included a new table, Table 2, which provides a summary of the different studies according to method used.

Comment #5:
The article needs more discussion on the malaria predictors, covariates and measures, there choice, outcome and impact on forecasting methods so as to clarify the broad objective of the authors.

Response #5:
It is unclear what specifically the Reviewer would like to see modified in the manuscript. We feel that we have adequately described the predictors and provided two tables for the readers (Tables 3 & 4). The outcomes were briefly mentioned in the Methods (p5) and have been presented in Table 1. As mentioned in the manuscript, the impact of the methods and predictors on forecasting accuracy could not be accessed across the studies given the lack of a common forecast accuracy measures. On pages 21-22, we do describe forecasting results in a qualitative manner for those studies that compared methods based upon the same data.

Comment #6:
Overall, I therefore think that this paper has some merit and would be a useful contribution to the field, but I think needs further work before acceptance could be recommended. My main concern is how thorough a review of the literature this really is. I would prefer to see this sort of article really pull out the generic concepts, structures and differences between the methods and models. The focus on the current models is OK, but should probably widen the scope to give a more comprehensive overview.

Response #6:

Please see our Response #1, with respect to the thoroughness of the literature review. Again, it is difficult for us to understand how to improve the manuscript without specific examples or suggestions. We have a section in the Discussion (Differences between forecasting methods, starting on p23) devoted to describing the differences between the models, within which we also discuss the general structures of the models/approaches.