PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Multicenter validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU
AUTHORS	Mao, Qingqing; Jay, Melissa; hoffman, jana; Calvert, Jake; Barton, Christopher; Shimabukuro, David; Shieh, Lisa; Chettipally, Uli; Fletcher, Grant; Kerem, Yaniv; Zhou, Yifan; Das, Ritankar

VERSION 1 – REVIEW

REVIEWER	Holger Fröhlich UCB Biosciences, Germany University of Bonn, Germany
REVIEW RETURNED	13-Jun-2017

GENERAL COMMENTS	General: This is an interesting paper describing the validation of a gradient boosted tree algorithm to predict sepsis on the basis of large scale EHR data. The paper is generally well written, and the algorithm is appropriately validated. However, I have a couple of comments and questions, which should be addressed from my point of view prior to publication. Major: 1. There is nothing stated about the public availability of the used data. In order to be able to replicate the conducted study, the employed data should be made available (e.g. on a dedicated web page). 2. Page 9: There is nothing said about feature construction. Was there any NLP preprocessing of the original EHR data? If yes, how was that done exactly? Was there any longitudinal information included or was all data within a 4 hour time window aggregated? How many features did the final data matrix contain? Please be more explicit and add a separate Section. 3. Apparently the authors treated sepsis prediction as a classification task. This implies that the actual time till sepsis diagnosis was neglected. Why didn't the authors consider a time-to-event modeling approach, which would be far more natural from my point of view? 4. Transfer learning: This is an excellent and nice idea. I assume there was an optimization over a discrete set of weights. How were these chosen? Please be more explicit. 5. Page 13: There is no interpretation of the final classifier. How many features were selected by the GBM? What was their relative importance? How was that relative importance affected by Transfer

	Learning? 6. Limitations: I think another limitation is that it is ultimately unclear, how reliable the detection of sepsis based on the employed ICD codes works. Moreover: EHR based data does not necessarily reflect the actual time point of a certain disease condition within a patient, but just when that condition was detected e.g. by a physician.

REVIEWER	Corey Chivers University of Pennsylvania Health System United States
REVIEW RETURNED	15-Jun-2017

GENERAL COMMENTS	The authors present a retrospective validation of a machine-learning algorithm for detecting sepsis, severe sepsis, and septic shock using 6 commonly collected vital signs. The study is the first, to my knowledge to evaluate the performance of a predictive algorithm to detect each of the levels of the sepsis progression, and to do so following patient encounters across multiple care settings (ED-wards-ICU). However, I have several concerns about the methodology, and in several instances a lack of detail that prevents the reader's ability to evaluate it, that would need to be addressed before I could recommend the publication of this work. 1. Variables that are part of the label definition are in the training set (eg three of the SIRS criteria + SBP). A machine learning algorithm that is trained using covariates that appear in the label definition would be expected to learn this mapping. This is particularly problematic when evaluating the performance at time-zero, when access to components of the definition allow the model to simply re-construct the label definition. 2. To that end, which observations were actually labeled positive for the purposes of model fitting? The hours leading up to clinical indications? What was done with the observations after the clinical indications were present? 3. The patient inclusion criteria limit the potential generalizability of the results in a real-time clinical setting. The positive cases were limited to those with onset > 7h from the start of the record, yet predictions are being made from time zero. In practice, one would not know in real-time whether a patient meets this exclusion or not. 4. It was hard to impossible to evaluate the model without seeing the model equation or how the model was actually fit. The authors reference the development of InSight in [1], however the model definition in this paper appears to be different than the current model (this paper makes no reference of gradient boosted trees). I was not able to discern the specifics of the model specification there either. For instance, it's not clear how the doublet, and triplet correlation (what the authors refer to as trend components) features are constructed. The notation is unclear. Given that the correlation between three covariates is a matrix, it's not clear how the resulting matrix is converted into features, or how these would differ from the pair-wise correlations.

	5. From what I could gather from [1], the correlation and trend variables are computed as ± |median of the trends|, and similarly for correlations. I am concerned that there may have been data leakage into the test set by computing these medians on the entire data set. 6. Also wrt the correlation features, I’m concerned that values here merely encode the sampling frequency of the underlying variable. On the wards, a q4 sampling frequency would lead to one observation per 5 hour window (again, going from [1] as the requisite details of the current implementation were not present). In this regime, the correlations between covariates would be 1. Only in a higher sampling regime (suspected deterioration in wards, or patient transfer to an ICU) would these correlations become less extreme. 7. The authors suggest that their model is robust to missing data, however missing at random is does not seem to me to be a particularly realistic missingness model. Related to point 6, there is informed presence bias in clinical time-series data. Specifically, the sampling frequency is a surrogate for clinician judgement of current patient acuity (with the extreme example being ICU-v-Ward sampling frequencies). A more realistic missingness model would at a minimum contain some temporal auto-correlation of missingness. The authors state that “These results are useful in estimating InSight’s performance in institutions or specific units where measurements may be taken less frequently or have reduced availability”. This statement is probably more true of the transfer learning experiments than of their approach to modeling missing data. Without the requisite detail on how the model was actually implemented, I cannot recommend this manuscript for publication. [1] J.S. Calvert et al. / Computers in Biology and Medicine 74 (2016) 69–73

REVIEWER	Kathryn Colborn University of Colorado Denver, Colorado School of Public Health, Department of Biostatistics and Informatics, USA
REVIEW RETURNED	21-Jun-2017

GENERAL COMMENTS	Response to “Validation of a machine learning algorithm for the prediction and detection of sepsis using only vital sign data”. The authors present a prediction model for sepsis that was developed using machine learning, specifically, cross-validation of a gradient tree boosting algorithm. The model was developed using data from UCSF Medical Center and was validated in the MIMIC-III data from BIDMC. The authors’ definitions for the three outcomes, sepsis, severe sepsis and septic shock seem well-justified. I have provided specific comments and suggestions below. • The three outcomes are rare in this data set. I would suggest providing the sensitivity achieved by cross-validation, which I expect is very low (you’ve just shown AUROC for fixed sensitivities). If it is low, I would suggest using sampling techniques, such as

	upsampling, downsampling or synthetic minority oversampling technique (SMOTE: see Chawla, et al., Journal of Artificial Intelligence Research, 2002). Providing the sensitivities and AUROC from various sampling methods might be interesting to readers, as the imbalanced learning problem is common. • I spent more time than I wanted to attempting to understand what “Sensitivity fixed at 0.80” meant. I don’t see examples of this in the literature. Perhaps more information on this with citations would be helpful to the reader. I was hoping to see a confusion matrix from the cross-validation results; I suggest including it somewhere. • It would be helpful to justify your choice of 4-fold cross-validation. I suggest using the TRIPOD method of developing a prediction model, supported by many major journals (see: http://annals.org/aim/article/2088549/transparent-reporting-multivariable-prediction-model-individual-prognosis-diagnosis-tripod-tripod). TRIPOD suggests using a temporal split of the data, and if that is not possible, justifying a random split. I don’t understand why you chose 4-fold, rather than more commonly used 10-fold or 5-fold, so please make this clearer. TRIPOD also suggests including a checklist of recommended items along with your journal submission. • In table 2, I think it would be easier to read if you included row names for sepsis, severe sepsis and septic shock, rather than colors.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: This is an interesting paper describing the validation of a gradient boosted tree algorithm to predict sepsis on the basis of large scale EHR data. The paper is generally well written, and the algorithm is appropriately validated. However, I have a couple of comments and questions, which should be addressed from my point of view prior to publication.

1. There is nothing stated about the public availability of the used data. In order to be able to replicate the conducted study, the employed data should be made available (e.g. on a dedicated web page).

Response: We have clarified the availability of data in both the methods section and in our data sharing statement. The MIMIC-III dataset is publicly available, and we have provided a website from which more information can be acquired. However, the UCSF data cannot be released publicly due to institutional policy.

2. Page 9: There is nothing said about feature construction. Was there any NLP preprocessing of the original EHR data? If yes, how was that done exactly? Was there any longitudinal information included or was all data within a 4 hour time window aggregated? How many features did the final data matrix contain? Please be more explicit and add a separate Section.

Response: We constructed our features from raw vital sign data, and did not use any unstructured (free-text) data in these experiments, and thus did not require NLP preprocessing. No preprocessing was performed prior to data extraction. Data were extracted from patient EHRs using custom structured query language queries written in-house. To anonymise the data, patient ages were subjected to a random jitter, and all variable dates and times were subjected to patient-specific random offsets to further ensure anonymity. The patient identifiers were then stripped, and the set entity / attribute / value data was then converted into flat files. We have attempted to clarify our feature construction in the manuscript by adding the following subsection entitled “Feature Construction” to our methods section:

We minimally processed raw vital sign data to generate features. Following EHR data extraction and imputation as described above, we obtained three hourly values for each of the six vital sign measurement channels from that hour, the hour prior, and two hours prior. We also calculated two difference values between the current hour and the prior hour, and between the prior hour and the hour before that. We concatenated these five values from each vital sign into a causal feature vector x with 30 elements (five values from each of six measurement channels).

3. Apparently the authors treated sepsis prediction as a classification task. This implies that the actual time till sepsis diagnosis was neglected. Why didn't the authors consider a time-to-event modeling approach, which would be far more natural from my point of view?

Response: We agree that a time-to-event modeling approach does seem more natural, from a prospective implementation standpoint. However, we performed a literature review of similar retrospective sepsis prediction systems, and found only one paper [18] which treated sepsis prediction as a time-to-event task, rather than a classification task [28-31].,,, We therefore determined that a classification approach represents current best practice for retrospective analysis of sepsis data, and importantly it allows for easy comparison against other sepsis prediction methods currently available in the literature.

We have added the following to the limitations section:

It is important to note that we designed the study as a classification task rather than a time-to-event modeling experiment, because the former is significantly more common in the literature [28-31]. The alternative would not allow for the use of an established, standard set of performance metrics such as AUROC and specificity without custom modification, and would make it more difficult to compare the present study to prior work in the field.

4. Transfer learning: This is an excellent and nice idea. I assume there was an optimization over a discrete set of weights. How were these chosen? Please be more explicit.

Response: We have clarified our optimization method with the following statement:.

Variable amounts of UCSF training data were incrementally added to the MIMIC-III training data set, and the resulting model was then validated on the separate UCSF test data set. Specifically, we left 50% of the UCSF patients as test data, and we randomly selected different fractions of the remaining UCSF data and combined them with the entire MIMIC-III data set as the training data. For each fraction used, we trained 100 models with different random relative weights on the UCSF and MIMIC-III training data. Then, the mean and standard deviation of AUROC values for each of these models were calculated on 20 randomly sampled sets, and the model with highest mean AUROC value among these 100 was used.

5. Page 13: There is no interpretation of the final classifier. How many features were selected by the GBM? What was their relative importance? How was that relative importance affected by Transfer Learning?

Response: We have added statements to our results section specifying which features contributed most significantly to model accuracy for both transfer learning and non-transfer learning experiments:

We ranked feature importance for the classifiers developed in this experiment, and determined that systolic blood pressure at the time of prediction was consistently the most important feature in making accurate model predictions. The relative importance of other features varied significantly based on the specific prediction task.

Feature importance was quite stable across transfer learning experiments, with systolic blood pressure measurements consistently playing an important role. Systolic blood pressure at two hours before onset, at time of onset, and at one hour before onset, in that order, were the most important features for accurate prediction in all tasks. Heart rate and diastolic blood pressure at time of onset were consistently the fourth and fifth most important features, though order of importance of the two features varied between tasks.

6. Limitations: I think another limitation is that it is ultimately unclear, how reliable the detection of sepsis based on the employed ICD codes works. Moreover: EHR based data does not necessarily reflect the actual time point of a certain disease condition within a patient, but just when that condition was detected e.g. by a physician.

Response: We agree with the reviewer, and accordingly we have added statements to the limitations section addressing both of the above concerns. Previous work has been done to validate the use of ICD codes in retrospectively identifying sepsis, and we have included a reference addressing the accuracy of ICD code use in identifying sepsis [26].

The gold standard references we use to determine sepsis, severe sepsis and septic shock rely on ICD-9 codes from the hospital database; this standard potentially limits our ability to capture all septic patients in the dataset, should any have been undiagnosed or improperly recorded. The administrative coding procedures may vary by hospital and do not always precisely reproduce results from manual chart review for sepsis diagnosis, although ICD-9 codes have been previously validated for accuracy in the detection of severe sepsis [26]. The vital sign measurements abstracted from the EHR are basic measurements routinely collected from all patients regardless of diagnosis and independent of physician judgement, and therefore this input to InSight is not dependent on the time of clinical diagnosis. However, the ordering of laboratory tests is contingent on physician suspicion, and the timing of these inputs may reflect clinician judgement rather than true onset time, potentially limiting the accuracy of our analysis.

Reviewer 2: The authors present a retrospective validation of a machine-learning algorithm for detecting sepsis, severe sepsis, and septic shock using 6 commonly collected vital signs. The study is the first, to my knowledge to evaluate the performance of a predictive algorithm to detect each of the levels of the sepsis progression, and to do so following patient encounters across multiple care settings (ED-wards-ICU). However, I have several concerns about the methodology, and in several instances a lack of detail that prevents the reader's ability to evaluate it, that would need to be addressed before I could recommend the publication of this work.

1. Variables that are part of the label definition are in the training set (eg three of the SIRS criteria + SBP). A machine learning algorithm that is trained using covariates that appear in the label definition would be expected to learn this mapping. This is particularly problematic when evaluating the performance at time-zero, when access to components of the definition allow the model to simply re-construct the label definition.

Response: We agree with that re-construction of label definitions at onset presents a valid concern, and therefore we have performed additional experiments to test the algorithm's accuracy with all measurements included in label definitions removed. These results are included in our results section in Table 2.

2. To that end, which observations were actually labeled positive for the purposes of model fitting? The hours leading up to clinical indications? What was done with the observations after the clinical indications were present?

Response: In order to address the above questions, we have added the following statement to our methods section:

Training was performed separately for each distinct task and prediction window, and observations were accordingly labeled positive for model fitting for each specific prediction task. Patient measurements were not used after the onset of a positive clinical indication.

3. The patient inclusion criteria limit the potential generalizability of the results in a real-time clinical setting. The positive cases were limited to those with onset > 7h from the start of the record, yet predictions are being made from time zero. In practice, one would not know in real-time whether a patient meets this exclusion or not.

Response: While we agree that this exclusion criteria can not be readily generalized to a clinical setting, it was important in this retrospective model characterization, as it ensured enough test set data for 4-hour pre-onset prediction to be possible. We have clarified the rationale for this exclusion criteria in our methods section as follow:

In order to ensure enough data to accurately characterize sepsis predictions at four hours pre-onset, we further limited the study group to exclude patients whose septic condition onset time was within seven hours after the start of their record, which was either the time of admission to the hospital or the start of their ED visit; the latter was applicable only if the patient was admitted through the ED. A smaller window to sepsis onset time would have resulted in insufficient testing data to make 4-hour prediction possible in some cases, which would inappropriately affect performance metrics such as sensitivity and specificity.

We additionally clarified the limited generalizability of this exclusion criteria in our limitations section as follows:

Additionally, our inclusion criteria requiring at least seven hours of patient data preceding sepsis onset also limits generalizability to a clinical setting, where the predictor would receive data in real time and not based on these criteria.

4. It was hard to impossible to evaluate the model without seeing the model equation or how the model was actually fit. The authors reference the development of InSight in [1], however the model definition in this paper appears to be different than the current model (this paper makes no reference of gradient boosted trees). I was not able to discern the specifics of the model specification there either.
For instance, it’s not clear how the doublet, and triplet correlation (what the authors refer to as trend components) features are constructed. The notation is unclear. Given that the correlation between three covariates is a matrix, it’s not clear how the resulting matrix is converted into features, or how these would differ from the pair-wise correlations.

Response: We apologize for being unclear on this point. Our previous publication represents an older iteration of our sepsis prediction algorithm, and we have made many changes to the model since the publication by J.S. Calvert et al. We have clarified that our older publications do not describe our current model by including:

Our previous studies, performed on earlier versions of the model, have investigated InSight applied to individual sepsis standards such as the SIRS standard for sepsis, severe sepsis, and septic shock, on the MIMIC retrospective datasets…. However, this study, which evaluates a significantly improved algorithm, is the first to apply InSight to all three standard sepsis definitions simultaneously, and to validate the algorithm on a mixed ward population, including ED, ICU and floor wards from UCSF.

We have also made several additions to the manuscript to clarify our methods, particularly with regard to feature construction.

5. From what I could gather from [1], the correlation and trend variables are computed as ± |median of the trends|, and similarly for correlations. I am concerned that there may have been data leakage into the test set by computing these medians on the entire data set.

Response: Correlation and trend variables for the model were not calculated in manner described above. Instead, we used a gradient boosting tree; therefore, there was no data leakage during testing. The inputs of the gradient boosting tree were causal, as described in the newly added feature construction section. We additionally hope that our statement above clarifying model improvement since previous publications addresses these concerns.

6. Also wrt the correlation features, I'm concerned that values here merely encode the sampling frequency of the underlying variable. On the wards, a q4 sampling frequency would lead to one observation per 5 hour window (again, going from [1] as the requisite details of the current implementation were not present). In this regime, the correlations between covariates would be 1. Only in a higher sampling regime (suspected deterioration in wards, or patient transfer to an ICU) would these correlations become less extreme.

Response: In order to ensure that correlation terms were based on a consistent sampling frequency, data from each channel (e.g., heart rate) were binned, using at most 2000 one-hour bins, beginning with the first recorded measurement for each channel. Values in each bin were averaged, yielding a single value. If a patient measurement was missing for a given hour, the empty bin was filled by imputing the most recent non-empty bin value. This binning and imputation method resulted in a consistent sampling frequency fed into the classifier, and therefore correlation features which accurately reflect meaningful correlations between variables. We have added the following statement in order to clarify this in our manuscript:

If a patient did not have a measurement in a given hour, the missing measurement was filled in using carry-forward imputation. This imputation method applied the patient's last measured value to the following hour (a causal procedure). In the case of multiple measurements within an hour, the mean was calculated and used in place of an individual measurement. Because patient data was standardized into single hourly measurements before being fed into the classifier, any information related to frequency of data collection was lost before predictions were made.

7. The authors suggest that their model is robust to missing data, however missing at random is does not seem to me to be a particularly realistic missingness model. Related to point 6, there is informed presence bias in clinical time-series data. Specifically, the sampling frequency is a surrogate for clinician judgement of current patient acuity (with the extreme example being ICU-v-Ward sampling frequencies). A more realistic missingness model would at a minimum contain some temporal auto-correlation of missingness. The authors state that "These results are useful in estimating InSight's performance in institutions or specific units where measurements may be taken less frequently or

have reduced availability". This statement is probably more true of the transfer learning experiments than of their approach to modeling missing data.

Response: We agree with the reviewer that our random deletion of data may not reflect a realistic clinical scenario. We have removed the line in question from our results section, and have additionally added the following statement to our limitations section:

Finally, our random deletion of data is not necessarily representative of data scarcity as it would occur in clinical settings where the rate of missing measurements would depend on the standard rate of data collection, which can vary widely, especially between the emergency department, general ward, and intensive care units.

Reviewer 3: The authors present a prediction model for sepsis that was developed using machine learning, specifically, cross-validation of a gradient tree boosting algorithm. The model was developed using data from UCSF Medical Center and was validated in the MIMIC-III data from BIDMC. The authors' definitions for the three outcomes, sepsis, severe sepsis and septic shock seem well-justified.

1. The three outcomes are rare in this data set. I would suggest providing the sensitivity achieved by cross-validation, which I expect is very low (you've just shown AUROC for fixed sensitivities). If it is low, I would suggest using sampling techniques, such as upsampling, downsampling or synthetic minority oversampling technique (SMOTE: see Chawla, et al., Journal of Artificial Intelligence Research, 2002). Providing the sensitivities and AUROC from various sampling methods might be interesting to readers, as the imbalanced learning problem is common.

Response: We have provided sensitivities achieved by ten-fold cross validation with a fixed specificity in Table 2 of our results section. Sensitivity remained at or above 0.98 with specificity fixed at 0.80.

Regarding positive outcome prevalence, we have also performed additional experiments and added relevant text as follows:

We also trained and validated the algorithm for each of the three gold standards for randomly selected, up- and down-sampled subpopulations with positive class prevalence between zero and one hundred percent.

Further, our experiments on applying InSight to up- and down-sampled sets showed that AUROC was largest when the set was chosen such that around half the patients met the gold standard. Moving lower on prevalence from 50% down to 0%, the AUROC values were only slightly lower while they dropped steeply when moving higher on prevalence from 50% up to 100% (a clinically unrealistic range).

The results of our up- and down-sampling experiments indicate that InSight is likely to only be slightly less effective (in AUROC terms) in settings with lower prevalence of sepsis, severe sepsis or septic shock, than UCSF or slightly more effective if the prevalence is higher than UCSF.

2. I spent more time than I wanted to attempting to understand what "Sensitivity fixed at 0.80" meant. I don't see examples of this in the literature. Perhaps more information on this with citations would be helpful to the reader. I was hoping to see a confusion matrix from the cross-validation results; I suggest including it somewhere.

Response: We have included the confusion matrices for our models both with and without gold standard definition measurements as supplemental material for this manuscript. We have also included the following statement in order to clarify the meaning of “sensitivity fixed at 0.80.”

In order to compare the specificities from each gold standard, we fixed sensitivities near 0.80; that is, we fixed a point on the ROC curve (i.e. set a specific threshold) after model development and tested algorithm performance under the chosen conditions in order to present data as consistently as possible. We similarly fixed specificities near 0.80 in order to compare sensitivities.

3. It would be helpful to justify your choice of 4-fold cross-validation. I suggest using the TRIPOD method of developing a prediction model, supported by many major journals (see: http://annals.org/aim/article/2088549/transparent-reporting-multivariable-prediction-model-individual-prognosis-diagnosis-tripod-tripod). TRIPOD suggests using a temporal split of the data, and if that is not possible, justifying a random split. I don’t understand why you chose 4-fold, rather than more commonly used 10-fold or 5-fold, so please make this clearer. TRIPOD also suggests including a checklist of recommended items along with your journal submission.

Response: We have rerun our experiments using 10-fold cross validation. Our results have been updated accordingly. Additionally, we have included a completed TRIPOD checklist with this revision.

4. In table 2, I think it would be easier to read if you included row names for sepsis, severe sepsis and septic shock, rather than colors.

Response: We have reformatted Table 2 of our results section as suggested.

We again sincerely thank the reviewers for their thorough review and their thoughtful comments, and are grateful for this opportunity to improve our manuscript. We hope we have addressed all concerns to your satisfaction.

VERSION 2 – REVIEW

REVIEWER	Holger Fröhlich University of Bonn, Germany
REVIEW RETURNED	26-Aug-2017

GENERAL COMMENTS	The authors have addressed all my concerns adequately.

REVIEWER	Corey Chivers Senior Data Scientist University of Pennsylvania Health System USA
REVIEW RETURNED	06-Sep-2017

GENERAL COMMENTS	The authors have done an admirable job of adequately addressing the majority of the reviewers comments. There are only three items, from my perspective, that should be further addressed prior to publication of this work.

	1. Reviewer 2, point 2: The clarification that training was performed separately for each distinct task is helpful for the reader, however, the authors should make clear the operational implications of doing so, perhaps in the discussion. Specifically, which model would be deployed in an operational setting? What could be said about that specific model with respect to the expected performance on each task and prediction window? If multiple models were deployed into the operational setting, how would they interact? 2. Reviewer 2, point 3: regarding the impact of the inclusion criteria, the authors added a statement about the limitation on generalizability. I would like to see a stronger statement about the expectation that performance in an operational setting is expected to be lower than the performance presented in this retrospective analysis. 3. Reviewer 2, point 6: The author's argument that the imputation method does not inject information related to frequency is erroneous. The carrying forward previous values has the effect of artificially reducing variance when the sampling frequency is low. Consider the following sequence of heart rates measured hourly: 65, 64, 66, 68, 63. The variance is 2.96. Now consider the same sequence with only every second measurement taken. With carry forward imputation the sequence is: 65, 65, 66, 66, 63, which has a variance of 1.2. The same logic applies to the difference operators in the feature construction of the model presented in this paper. This holds for sampling frequencies < q1 (eg q2 in this example). The relationship between sampling frequency and variance changes direction at frequencies > q1 when multiple measurements are averaged over the hour. On this end of the spectrum, increased sampling frequency reduces variance by smoothing intra-hour variation. I don't think that it is fatal that this is the case, but would like to see the authors acknowledge this. that there is information about sampling frequency encoded in the derived features of this model and that as such it is not a purely physiological model.

REVIEWER	Kathryn Colborn University of Colorado Anschutz Medical Campus Colorado School of Public Health Department of Biostatistics and Informatics
REVIEW RETURNED	01-Sep-2017

GENERAL COMMENTS	The authors thoroughly addressed comments I provided on a previous review of this paper. This is an important contribution to the septic literature, and the methods are appropriately applied and important to publish in order to advance the field. I have no further recommendations.

VERSION 2 – AUTHOR RESPONSE

Reviewer 2:

1. The clarification that training was performed separately for each distinct task is helpful for the reader, however, the authors should make clear the operational implications of doing so, perhaps in the discussion. Specifically, which model would be deployed in an operational setting? What could be said about that specific model with respect to the expected performance on each task and prediction window? If multiple models were deployed into the operational setting, how would they interact?

Response: Depending on the specific needs of each deployment setting, any of the models described in the study could be independently deployed. Model choice could depend on existing workflow, the hospital's diagnostic and treatment procedures, and the specific outcomes which the hospital hopes to improve. We do not anticipate deploying multiple models simultaneously in a single location. We have added the following statement to our Discussion section, as recommended, in order to clarify the above:
The separate models trained for each gold standard and prediction window in this study further demonstrate the potential clinical utility of machine learning methods. In addition to training on a specific patient population, machine learning methods can allow for the development of prediction models which are tailored to a hospital's unique needs, data availability, and existing workflow practices. Any one of the models developed in this study could be independently deployed in a clinical setting; choice of model deployment would be contingent upon the needs of a particular hospital, and the expected tradeoff in performance for different model choices.

2. Regarding the impact of the inclusion criteria, the authors added a statement about the limitation on generalizability. I would like to see a stronger statement about the expectation that performance in an operational setting is expected to be lower than the performance presented in this retrospective analysis.

Response: We have amended our Limitations section in order to address the above comment:
Additionally, our inclusion criteria requiring at least seven hours of patient data preceding sepsis onset also limits generalizability to a clinical setting where the predictor would receive data in real time. Algorithm performance in a clinical setting may reasonably be expected to be lower than its retrospective performance in this study.

In order to estimate algorithm performance across a variety of clinical settings, we have also trained and tested the algorithm for severe sepsis detection on additional datasets from Stanford Medical Center (Stanford, CA), Oroville Hospital (Oroville, CA), Bakersfield Heart Hospital (Bakersfield, CA), and Cape Regional Medical Center (Cape May Court House, NJ). The algorithm consistently achieved an AUROC above 0.92 on these additional datasets; these strong results suggest potential strong performance in a variety future clinical deployments.

3. The author's argument that the imputation method does not inject information related to frequency is erroneous. The carrying forward previous values has the effect of artificially reducing variance when the sampling frequency is low. Consider the following sequence of heart rates measured hourly: 65, 64, 66, 68, 63. The variance is 2.96. Now consider the same sequence with only every second measurement taken. With carry forward imputation the sequence is: 65, 65, 66, 66, 63, which has a

variance of 1.2. The same logic applies to the difference operators in the feature construction of the model presented in this paper. This holds for sampling frequencies < q1 (eg q2 in this example). The relationship between sampling frequency and variance changes direction at frequencies > q1 when multiple measurements are averaged over the hour. On this end of the spectrum, increased sampling frequency reduces variance by smoothing intra-hour variation. I don't think that it is fatal that this is the case, but would like to see the authors acknowledge this. that there is information about sampling frequency encoded in the derived features of this model and that as such it is not a purely physiological model.

Response: We have removed the line in our Data Extraction and Imputation section stating that all information on data collection frequency was removed before predictions were generated. We have also added the following to our Limitations section:
While the imputation and averaging performed before feature construction eliminated some information about sampling frequency, these methods do not remove all non-physiological information inherent to our system. Further, imputation of the most recently available past measurement may artificially alter the rate of the temporal changes in patient vital signs that we incorporate into feature vectors, which may in turn affect risk predictions. Averaging multiple patient measurements may similarly remove informative variation in vital signs.

We again sincerely thank the reviewers for their time and their thoughtful comments, and are grateful for the opportunity to improve our work. We hope we have addressed all concerns to your satisfaction.

VERSION 3 – REVIEW

REVIEWER	Corey Chivers University of Pennsylvania Health System
REVIEW RETURNED	20-Oct-2017

GENERAL COMMENTS	The authors have address each of my remaining concerns.