PEER REVIEW HISTORY

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

ARTICLE DETAILS

TITLE (PROVISIONAL)	Prediction of early unplanned intensive care unit readmission in a UK tertiary-care hospital: A cross-sectional machine learning approach
AUTHORS	Desautels, Thomas; Das, Ritankar; Calvert, Jake; Trivedi, Monica; Summers, Charlotte; Wales, David; Ercole, Ari

VERSION 1 - REVIEW

REVIEWER	Jeffrey Che-Hung Tsai Department of Emergency Medicine, Taichung Veterans General Hospital. Taiwan
REVIEW RETURNED	01-May-2017

GENERAL COMMENTS	This study is well-designed, and has many advantages. This study has good results for contributory conclusions, and the conclusion is suitably concordant with the aim of the study. The authors provide good example of using machine learning for constructing a prediction model, as well as utilization and processing data of electronic patient record. The work of this study do provide a good example and suggestion of using transfer learning from a big source data to target data in individual institution. An useful example of transfer learning can emphasize the value of big data, such as MIMIC-III in this study.

REVIEWER	Geert Meyfroidt University Hospitals Leuven, Belgium
REVIEW RETURNED	02-May-2017

GENERAL COMMENTS	With great interest I have read and reviewed the manuscript by Desautels and co-authors. They describe the development of a Machine Learning model to predict early (<48h) ICU readmission. This is a very relevant challenge in the ICU. The complexity of this type of research is often underestimated. I would like to congratulate the authors on the clear way they have described the research. Nevertheless, I have some comments and questions as well, which I hope they are able to resolve. Major comments: ------------------- 1. I think the title should change to 'Prediction of early unplanned intensive care readmission... " 2. Table 1 should contain more information than it does now: diagnoses, duration of mechanical ventilation, type of ICU, ... You do not need to mention median as well as mean: if the data are distributed normally mean is sufficient, if not, use the median. 3. The median length of stay is more than 18 days? This must be a

	very specific subgroup of very ill patients, because this is not the usual casemix of many ICU's. What makes these patients stay this long in the ICU? 4. Although I think am familiar with a lot of Machine Learning methods and terminology (at least for a clinician...), it is unclear to me what 'Gradient tree boosting' exactly is? I suppose it is a variant of Random Forests, but I'm not sure... and the brief explanation is still a bit vague. Could the authors explain a bit better what this technique does and what the rationale was for using it? Are these black box models, or is there a way to know which are the most important predictors for early readmission? As a clinician, I would probably be more interested in a model that can explain to me why I should pay attention to a patient at high risk of readmission, before I discharge him. 5. I am not sure whether I understand the data pre-processing on page 6, final paragraph. Does this mean that a maximum of 500 hours of monitoring data was used? What was the original resolution with which these time series were stored? Waveform data? 1-minute data? Does this mean that all time series were summarized into 1h averages? If that is the case, a lot of dynamics is lost, of course, but some trends should still be in these summarized data. How was this assessed? Or was there no assessment of dynamics (standard deviation, amplitude variability, entropy, shape analysis, Fourier, cepstral coefficients, ...)?? 6. Not all patients had the same ICU length of stay (LOS). How was this accounted for? A longer LOS will automatically generate more values that might be aberrant. Was there a correction for LOS? I might not fully have grasped the concept of concatenation of values (page 7). 7. The central hypothesis of this research seems to be the use of clinical variables to predict early readmission. In my practice, such early readmissions are in a lot of cases actually too early discharges... Early discharge is usually not driven by the patient's characteristics (usually, we know when a patient is still vulnerable when he is being discharged), but rather by external factors, in many cases a push on ICU beds, especially when these beds are scarce (especially in the UK, which has one of the lowest ICU beds/population rates in Europe). It's difficult to take this into account, of course, but as surrogate predictors, relatively easy parameters such as the day in the week the the patients are being discharged (for instance, on weekdays the push on ICU beds is usually higher than in weekends), or seasonal variations (Flu months versus summer months). You could increase the performance of your models by including these factors. 8. The aROC's of the models are certainly not impressive, but performance of a model should not only be evaluated by its aROC. The authors should report on the calibration of the model as well, preferably using calibration plots, but also Hosmer-Lemeshow goodness-of-fit and Brier score should be calculated. In addition, I think reclassification indexes such as the net reclassification improvement (NRI) of the model, compared to SWIFT, and decision curve analysis should be done to evaluate the clinical benefit of using the models. 9. The discussion is now focussed on the promises and possibilities of the predictor, but does not address the inherent weaknesses of this study, and how these will be addressed in future studies by this or other research groups.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
No points to address.

Reviewer 2:

Major comments:
----
1. I think the title should change to 'Prediction of early unplanned intensive care readmission... "

We have changed the title to “Prediction of early unplanned intensive care unit readmission in a tertiary-care hospital: a cross-sectional machine learning approach.”

2. Table 1 should contain more information than it does now: diagnoses, duration of mechanical ventilation, type of ICU, ... You do not need to mention median as well as mean: if the data are distributed normally mean is sufficient, if not, use the median.

Under UK law, shared datasets must be the minimum required for the prosecution of the task at hand and the chances of re-identification should be minimal. Thus our data request from Cambridge University Hospitals was required to be as minimal as possible to train the model for the purposes of our study, an consequently some of this information is unavailable. In particular diagnostic codes were not exported as these carry a particular risk of identification in the event of rare diseases. We did not export the duration of ventilation as we did not have a prior consideration that this would be predictive given our predominantly emergency caseload so again this information is not available. However, we have added to Table 1 information on which ICU’s the patients first visited.

3. The median length of stay is more than 18 days? This must be a very specific subgroup of very ill patients, because this is not the usual casemix of many ICU's. What makes these patients stay this long in the ICU?

The casemix for our ICU consists of almost entirely emergency cases: Elective admissions (e.g. post-operative admissions) are negligible- these are handled other high-care areas in the hospital. As a result the number of ‘quick-turnaround’ cases is limited and our casemix is particularly severely unwell. In particular we are the regional centre for major trauma (with and without intracranial injury) and liver intensive care. We have added a paragraph in the ‘Data’ section to describe this and also discussed the possible limitations to external generalizability that result from this consideration in the Discussion.

4. Although I think am familiar with a lot of Machine Learning methods and terminology (at least for a clinician...), it is unclear to me what 'Gradient tree boosting' exactly is? I suppose it is a variant of Random Forests, but I'm not sure... and the brief explanation is still a bit vague. Could the authors explain a bit better what this technique does and what the rationale was for using it? Are these black box models, or is there a way to know which are the most important predictors for early readmission? As a clinician, I would probably be more interested in a model that can explain to me why I should pay attention to a patient at high risk of readmission, before I discharge him.

We have included the name of the ensemble algorithm used, AdaBoost, and citations for this method. AdaBoost is a gradient-based boosting method, that is, each boosting round adds a tree, where this tree is constructed to improve the classification of those examples which were incorrectly handled in the previous round of boosting. While ensembles of decision trees are somewhat difficult to interpret, it is possible to extract which features are most important to the ensemble as a whole. Unfortunately, this does not answer the clinician’s natural question, “why was this particular patient of concern?” In this respect, these models are somewhat black boxes.

5. I am not sure whether I understand the data pre-processing on page 6, final paragraph. Does this mean that a maximum of 500 hours of monitoring data was used? What was the original resolution with which these time series were stored? Waveform data? 1-minute data? Does this mean that all time series were summarized into 1h averages? If that is the case, a lot of dynamics is lost, of course,

but some trends should still be in these summarized data. How was this assessed? Or was there no assessment of dynamics (standard deviation, amplitude variability, entropy, shape analysis, Fourier, cepstral coefficients, ...)??

We have clarified that the data were non-waveform data, and the presentation of the exclusion and binning schemes.
Yes, at most 500 hours of data were stored per patient, and additionally, patients were discarded who had onset times after this period. Each event was originally logged with its own time stamp, but events for most channels (e.g., heart rate) were typically of approximately 1/hour frequency. All time series were indeed summarized into one-hour averages, matching common practice in UK ICUs (hourly, 'on-the-hour' data). In some cases, this may have lost some of the information associated with short-duration events with repeated measurements, such as transient hypotensive episodes. It is probable that incorporation of transient episodes might improve the prediction model but this could not be studied with this dataset. At the same time, higher resolution data is potentially subject to noise and artifact. We did not do a full assessment of these data properties for this paper. We have added a comment on these considerations in the discussion, and some comment in Methods > Data, paragraph 3.

6. Not all patients had the same ICU length of stay (LOS). How was this accounted for? A longer LOS will automatically generate more values that might be aberrant. Was there a correction for LOS? I might not fully have grasped the concept of concatenation of values (page 7).

The reviewer is correct; different patients have somewhat different data availability on the basis of their ICU LOS. However, because of our finite backward horizon (the present hour and the four previous) the only difference is data availability within these hours. Because we exclude patients with less data than this (relative to first vitals), most patients should have very similar data availability, up to effects from the imputation scheme. We comment on this in the last paragraph of Methods: Processing.

7. The central hypothesis of this research seems to be the use of clinical variables to predict early readmission. In my practice, such early readmissions are in a lot of cases actually too early discharges... Early discharge is usually not driven by the patient's characteristics (usually, we know when a patient is still vulnerable when he is being discharged), but rather by external factors, in many cases a push on ICU beds, especially when these beds are scarce (especially in the UK, which has one of the lowest ICU beds/population rates in Europe). It's difficult to take this into account, of course, but as surrogate predictors, relatively easy parameters such as the day in the week the the patients are being discharged (for instance, on weekdays the push on ICU beds is usually higher than in weekends), or seasonal variations (Flu months versus summer months). You could increase the performance of your models by including these factors.

The reviewer's comment that we know when a patient is still vulnerable seems very reasonable and certainly as clinicians we feel that we do. It is for that reason that we find that even our model predictive power achieved is remarkable since if this were the case, there should be very little predictability in early readmission at all.

We agree that there may well be other factors at play although and the reviewer is correct that the UK has very few ICU beds per capita. Whilst no UK ICU clinician would deliberately send a vulnerable patient to the ward, institutional factors have been shown in international studies to have an impact at least on mortality (and it would not be unreasonable to extrapolate this to ICU readmission). As a result of bed limitations, our ICUs (in common with many UK institutions) run not only at high acuity but at consistently high bed occupancy (>95%) and these cases are almost entirely emergencies. As such there is relatively little variation in ICU occupancy so this is unlikely to be a factor.

Of course occupancy is not the same as demand and the referee's idea of stratifying by winter months or day of the week is a very interesting one. We think that this would also need to be extended to some metric of ward bed availability as this is often a limiting factor for ICU discharge although how best to capture this is unclear. Ultimately day of week and time of year was deliberately excluded from our data-sharing agreement so it would not be possible to do this with our data. We further suspect that such stratification would reduce the available degrees of freedom for training (the dataset size available was already modest) and ultimately degrade confidence in the model further without a far larger set of patients. This would be an interesting area of research for a future project and we have commented on this in the discussion.

8. The aROC's of the models are certainly not impressive, but performance of a model should not only be evaluated by its aROC. The authors should report on the calibration of the model as well, preferably using calibration plots, but also Hosmer-Lemeshow goodness-of-fit and Brier score should be calculated. In addition, I think reclassification indexes such as the net reclassification improvement (NRI) of the model, compared to SWIFT, and decision curve analysis should be done to evaluate the clinical benefit of using the models.

As we are reporting aggregate results across ten cross-validation folds, for each weighting parameter value, calibration plots are not feasible in our presentation. However, we have calculated fold-averaged Brier scores and reported these in Table 2.

9. The discussion is now focused on the promises and possibilities of the predictor, but does not address the inherent weaknesses of this study, and how these will be addressed in future studies by this or other research groups.

There are, of course, limitations to this study, in particular questions of generalizability and optimization of the predictive covariates chosen. We have greatly expanded the discussion section describing the limitations of our study. One clear limitation is the modest prima facie predictive power of the model. Whilst this may be improved by different choices of explanatory features, we feel that the most striking conclusion is that even such a degree of predictability exists in a set of patients who have already been deemed ‘safe’ for down-transfer suggesting that there are features that are not systematically appreciated by the clinician. We have tried to emphasize this in the discussion and elsewhere in the text.

VERSION 2 – REVIEW

REVIEWER	Geert Meyfroidt Associate Professor, KU Leuven, Belgium Deputy Head of Clinics, University Hospitals Leuven, Belgium
REVIEW RETURNED	01-Jun-2017

GENERAL COMMENTS	I would like to thank the authors for the adaptations made to the manuscript, and to their answers to my questions, in particular to questions 1, 3, 7, 8 and 9. I still have some minor comments. Q2: on not disclosing any diagnostic information. I must acknowledge that I lack the knowledge or expertise to interprete the UK laws on privacy, I must say I find it hard to believe that it would be ethically unacceptable to disclose any diagnostic information other than the location of a patient, because this requirement would imply that most retrospective studies on critically ill patients would be impossible to perform. The problem of a very rare diagnosis where n=1 that could potentially lead to identifying a patient can easily be solved by reporting diagnostic categories rather than individual diagnoses, (e.g. TBI, SAH, stroke, postanoxic, trauma non-TBI, liver transplant, ...), and to label all those that do not fall under these categories as "other". Not all readers will be familiar with the case-mix of your units, although the now included description of that case-mix is some solution to this problem. I cannot ask the impossible from the authors, but it is certainly advisable to provide some diagnostic info. If not, this is a pity with regards to generalizability to other centers, and I would advise them

	to plan on reporting this in future studies. As a second comment, I still think you should not report both the mean and the median, in table 1. Q4. Thank you for this very clear clarification. I think this small paragraph should be in the manuscript, methodology section. Q5 and 6. I am still not entirely clear on the temporal relationship of the 500 hour of data, and outputs of the ML model. It might be my lack of understanding of the methodology, I admit. From page 33, line 32-33, and page 35, line 10, I conclude that 'prediction time' is the time of the first ICU discharge, and that only data from the 5 hours before ICU discharge are being used? Is this correct? What if the patient's ICU length of stay is > 500 hours (roughly 21 days)? Since only the first 500 bins of ICU data were used (assuming that measurements start at around ICU admission), this implies that for the longer staying patient no data close to ICU discharge has been used? And since the median LOS was 18.04, I assume that quite a proportion of patients stayed beyond 21 days? I would suggest to try and explain this in a more easy way, or even to include a graphical representation of the timing of the models inputs, ICU discharge, and the time of readmission, to clarify this.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1:
No points to address.

Reviewer 2:
-- Q2: On not disclosing any diagnostic information.
I must acknowledge that I lack the knowledge or expertise to interpret the UK laws on privacy, I must say I find it hard to believe that it would be ethically unacceptable to disclose any diagnostic information other than the location of a patient, because this requirement would imply that most retrospective studies on critically ill patients would be impossible to perform. The problem of a very rare diagnosis where n=1 that could potentially lead to identifying a patient can easily be solved by reporting diagnostic categories rather than individual diagnoses, (e.g. TBI, SAH, stroke, postanoxic, trauma non-TBI, liver transplant, ...), and to label all those that do not fall under these categories as "other". Not all readers will be familiar with the case-mix of your units, although the now included description of that case-mix is some solution to this problem. I cannot ask the impossible from the authors, but it is certainly advisable to provide some diagnostic info. If not, this is a pity with regards to generalizability to other centers, and I would advise them to plan on reporting this in future studies.
As a second comment, I still think you should not report both the mean and the median, in table 1.

++ We agree that it is regrettable that the data set is restricted and does not contain this information. We would like to note that this was not a legal issue, but a reluctance to include this information on the part of the Trust. In the future, we will certainly attempt to include this information.

-- Q4. Thank you for this very clear clarification. I think this small paragraph should be in the manuscript, methodology section.

++ We have added a further note on interpretability of the model to the corresponding section.

-- Q5 and 6. I am still not entirely clear on the temporal relationship of the 500 hour of data, and outputs of the ML model. It might be my lack of understanding of the methodology, I admit. From page 33, line 32-33, and page 35, line 10, I conclude that 'prediction time' is the time of the first ICU discharge, and that only data from the 5 hours before ICU discharge are being used? Is this correct?

What if the patient's ICU length of stay is > 500 hours (roughly 21 days)? Since only the first 500 bins of ICU data were used (assuming that measurements start at around ICU admission), this implies that for the longer staying patient no data close to ICU discharge has been used? And since the median LOS was 18.04, I assume that quite a proportion of patients stayed beyond 21 days? I would suggest to try and explain this in a more easy way, or even to include a graphical representation of the timing of the models inputs, ICU discharge, and the time of readmission, to clarify this.

++ First, the reviewer is correct in his understanding that only the representation of the last five hours pre-discharge is used in making the prediction. However, it is crucial to stress that this representation is the imputed set of measurement values of the last five hours before discharge; the carry-forward imputation scheme means that sparsely sampled values (such as labs) are available in their most recent values, even if those values were obtained previous to the five-hour window. The vector of inputs can be viewed as a representation of the patient's current and recent state.

However, the reviewer has misunderstood an important point; we apologize for any lack of clarity. Our Table 1 "length of stay" values refer to the total duration of the hospital encounter, not the first ICU encounter. We have clarified this both in the caption and the table's row labels. We had complete transfer and discharge histories, such that we could determine the patient's ultimate fate within the encounter (the gold standard), determine the time of first ICU down-transfer. Separately, we carried out the EPR data discretization and binning process up until 500 hours, creating the potential classifier inputs. We then combined these processing streams and eliminated from the study population all encounters in which prediction time would have occurred after 500 hours; every patient in the final study population (n = 2018) therefore both (1) had a first ICU discharge at 500 hours or before and, thus, (2) had discretized and imputed data available within the crucial five-hour window. We have also clarified this point in the final paragraph of the Methods > Processing section.

VERSION 3 – REVIEW

REVIEWER	Geert Meyfroidt Department and Laboratory of Intensive Care Medicine, University Hospitals Leuven and KU Leuven, Belgium
REVIEW RETURNED	11-Jul-2017

GENERAL COMMENTS	Congratulations on the manuscript. Thanks for taking into account most of the comments.