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)	Hospital Readmission Risk Prediction Based on Claims Data Available at Admission. A Pilot Study in Switzerland
AUTHORS	Brüngger, Beat; Blozik, Eva

VERSION 1 – REVIEW

REVIEWER	Faraz Faghri National Institutes of Health, USA, Department of Computer Science, University of Illinois at Urbana-Champaign, USA
REVIEW RETURNED	09-Jan-2019

GENERAL COMMENTS	In this paper, authors investigate the feasibility of using claims data to predict early readmission. They use claim data from Swiss health insurance of 138,222 adults. They extract features from claim data and utilize logistic regression to develop the predictive model. They achieve AUC of 0.61. For the following reason, the paper does not properly answer the initial question whether “claims data can predict early readmission”: Conclusion is not supported by the results and used methods. Authors only investigate one feature extraction and one predictive model. And that particular method achieves a low AUC of 0.61. Which is not a satisfactory nor practical accuracy for the clinical setting. In order to address the feasibility of using claim data for readmission prediction task, one should investigate a broad range of feature extraction as well as predictive model development techniques. The challenge is to show what machine learning technique on what feature selection method is going to give us high performance. As shown by the results of the paper, the currently used machine learning and feature selection techniques are not providing a strong performing model. I recommend redoing the paper with other techniques and show which one is providing a strong and practical predictive model. Major comments and questions: 1. In this paper, feature extraction is done according to existing literature. Some features were found in the Swiss claim data and some not. However, authors have not utilized the full power of their data. Authors need to perform a comprehensive data-driven analysis in order to both fully utilize their data as well as addressing the feasibility of their claim. They need to "let the data talk” instead of relying on manual feature extraction. For starters, I recommend constructing a high dimension feature matrix, then perform dimension reduction techniques and later apply various predictive models. Depending on the size and nature of their data, they might also apply Random Forest techniques to develop a

	predictive model first and then using feature eliminations try to increase the performance and find more valuable features. 2. Paper only describes the extracted features and not the full spectrum of available features in the dataset. I recommend having the full data dictionary in the supplementary materials. Again, since it is a feasibility study, feature selection should be broad and comprehensive, not limited by existing literature. 3. The method section is lacking in terms of data cleansing. There is no indication of how missing data were treated, whether the binning of continues data has been balanced, etc. 4. Paper lacks code which makes it hard to evaluate the results and possibility of replication. Authors need to include their Jupyter notebook (or any other alternative) for analysis. 5. Why the focus on the sole claim data? at the time of hospitalization, we have access the claim data as well as physician's assessment. Why not using the hospital assessment at the time with claim data as a prior? 6. Logistic regression model evaluation requires cross-validation. Without showing the cross-validation it is hard to asses overfit or potential bias in the train/test data selection. Minor comments: 1. The text is hard to read, especially the introduction. Long and complex sentence structures. I recommend a major edit. 2. Papers reviewed and cited in the introduction are at least five years old. There has been significant research done on the issue, especially with the rise of new machine learning techniques. I recommend authors to investigate more recent literature and methods as well.

REVIEWER	Tim Badgery-Parker Research Fellow, Menzies Centre for Health Policy, The University of Sydney, Australia
REVIEW RETURNED	30-Jan-2019

GENERAL COMMENTS	# General comments This paper aims to estimate the feasibility of using claims data available at admission to predict if a patient will have a subsequent readmission within 30 days. The authors conclude that the data available at admission are suitable for this purpose, but that poor model performance (both in their model and in other similar studies) makes this approach not worthwhile. First, it seems to me that poor model performance is to be expected. As the authors note, high readmission rates are used as an indicator of poor quality of care. They are only useful as such an indicator if they reflect what happens during the hospital stay — information that is explicitly excluded by the authors' use of only data available at admission. Second, I think the authors have handicapped their model performance by their methodological decisions. Many of these are common practice, but not best practice. The authors categorised continuous variables to account for non-linearity. This greatly reduces the information in the data. They would be better modelling continuous variables with restricted cubic splines or polynomial terms. The variable selection process was reasonable, but a better prediction model might have been derived from a penalised regression model such as lasso or elastic net.

Furthermore, the analysis does not seem to account for clustering of patients within hospitals. Use of readmission rates as a quality indicator requires that hospitals have different rates, so hospital should be included in the model.

The authors do mention in the limitations section that categorising is not the best approach and suggest that other models such as random forest might have better performance. I think that even within logistic regression they could get better prediction if they model continuous variables as continuous and account for hospital clustering.

In addition, they appear to have estimated AUC by first dichotomising the modelled probability into a binary prediction then comparing that with the true value. They should use the modelled probability directly in estimating the AUC, which is supposed to be a summary statistic accounting for all possible dichotomisation thresholds.. If they calculate AUC properly, they will probably see a substantial increase.

# Specific comments

## Abstract

1 PCG needs to be explained in the abstract.

## Methods

2 Patients can change health insurers once per year — how did the analysis deal with people entering or leaving cohort? Are the variables on previous use of health services available for someone who enters the cohort during the study period?

3 Variables on prior use of health services are not clearly defined. This appears to be use in the year before the admission, but this is not explicitly stated. If it is a year, was there data for a full year before admission for all patients?

4 line 112 (literature search): predictors were grouped into 5 categories but only 4 are listed, while additional table 1 appears to list 7 categories

5 Although commonly done, categorising continuous variables loses a lot of information. If nonlinearity is a concern, consider splines.

6 The model really should include hospital (I would include as random intercept) to account for clustering of patients within hospitals.

7 You should not dichotomise the predicted probability to estimate the AUC; use the actual predicted value from the model.

8 Sensitivity, specificity, PPV, and NPV are not suitable for assessing model performance. AUC (c statistic) is appropriate as measure of discrimination. You might also consider examining calibration with appropriate plot.

	9 You say you set threshold to balance sensitivity and specificity. It is not clear if you then used the same threshold in all the subgroup analyses, or reset it to balance sensitivity and specificity in the subgroups. ## Results 10 You comment that best PPV was in subgroup MDC kidney and urinary tract. But PPV depends on prevalence. Examining table 1 would tell you, before you ever calculated PPV, that this subgroup would have the highest PPV. As I said above (8), don't use PPV/NPV to assess the model. ## Discussion 11 line 278: "principally" -- I think you mean "in principle" 12 line 332: Surely any index admission after 2014 would have 0 for prior inpatient costs, not just those towards the end of the 3 year period. But if every index episode has data for a full year before, the overall distribution of inpatient costs should be correct. Anyway, inpatient costs were not included in the model, so this part of discussion seems to be irrelevant. 13 line 299: you cite PREADM authors that performance is likely to be low because most readmissions are not avoidable. The logic of this is not clear to me. Just because a readmission is unavoidable does not necessarily make it hard to predict from patient data. Or am I missing something?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Faraz Faghri

Institution and Country: National Institutes of Health, USA, Department of Computer Science, University of Illinois at Urbana-Champaign, USA

In this paper, authors investigate the feasibility of using claims data to predict early readmission. They use claim data from Swiss health insurance of 138,222 adults. They extract features from claim data and utilize logistic regression to develop the predictive model. They achieve AUC of 0.61.

For the following reason, the paper does not properly answer the initial question whether "claims data can predict early readmission": Conclusion is not supported by the results and used methods. Authors only investigate one feature extraction and one predictive model. And that particular method achieves a low AUC of 0.61. Which is not a satisfactory nor practical accuracy for the clinical setting. In order to address the feasibility of using claim data for readmission prediction task, one should investigate a broad range of feature extraction as well as predictive model development techniques. The challenge is to show what machine learning technique on what feature selection method is going to give us high performance. As shown by the results of the paper, the currently used machine learning and feature selection techniques are not providing a strong performing model. I recommend redoing the paper with other techniques and show which one is providing a strong and practical predictive model.

We fully agree with the reviewer that the present study cannot answer the question whether claims data predict early readmissions in general. Our main goal was to evaluate whether claims data available at hospital admission are a promising basis for the development of a prediction tool for readmissions (please see also our response to editorial comment above). To our knowledge, this has internationally not been investigated before. The underlying argument for restricting to claims data available at hospitalisation is that these data are usually nationwide available, conform with a standardised format and are pre-existing without large additional efforts of data collection. We also agree that the results are disappointing and discourage to follow up on focusing on patient-level data available at the time of hospitalisation alone. However, we are confident that this knowledge is helpful for researchers and quality managers both from the local context and from other health systems with a national health insurance (please see our changes of title, abstract, and discussion section lines 359-362).

Major comments and questions:

1. In this paper, feature extraction is done according to existing literature. Some features were found in the Swiss claim data and some not. However, authors have not utilized the full power of their data. Authors need to perform a comprehensive data-driven analysis in order to both fully utilize their data as well as addressing the feasibility of their claim. They need to “let the data talk” instead of relying on manual feature extraction. For starters, I recommend constructing a high dimension feature matrix, then perform dimension reduction techniques and later apply various predictive models. Depending on the size and nature of their data, they might also apply Random Forest techniques to develop a predictive model first and then using feature eliminations try to increase the performance and find more valuable features.

   We agree that using a data-driven approach would be a promising alternative methodology. We consciously decided to use a theory-based approach and to base the pilot study on evidence from previously published experiences in the field for two reasons: 1) we aimed to benefit from international experiences and 2) we are convinced that a theory-based approach facilitates acceptance of such tools among healthcare providers. We added text to the manuscript to clarify these aspects (please see discussion section lines 370-375)

2. Paper only describes the extracted features and not the full spectrum of available features in the dataset. I recommend having the full data dictionary in the supplementary materials. Again, since it is a feasibility study, feature selection should be broad and comprehensive, not limited by existing literature.

   Please see our response to the comment above.

3. The method section is lacking in terms of data cleansing. There is no indication of how missing data were treated, whether the binning of continues data has been balanced, etc.

   We added text for clarification (please see methods section lines 166-167, 146-147).

4. Paper lacks code which makes it hard to evaluate the results and possibility of replication. Authors need to include their Jupyter notebook (or any other alternative) for analysis.

The manuscript and all results of statistical analyses have been reproducibly built using R and R Markdown. The code is available upon reasonable request.

5. Why the focus on the sole claim data? at the time of hospitalization, we have access the claim data as well as physician's assessment. Why not using the hospital assessment at the time with claim data as a prior?

   Currently, in Switzerland there is no uniform format nor a national database including clinical data collected at hospital stays. Clinical data would certainly add value to the database. The aim of the present study was to evaluate whether the use of existing data is promising.

6. Logistic regression model evaluation requires cross-validation. Without showing the cross-validation it is hard to asses overfit or potential bias in the train/test data selection.

   Based on the reviewer comment we added k-fold cross-validation (k = 20) in the estimation of the AUC and additionally report respective 95%-confidence intervals. Please see changes in methods section line 202, and results section tables 3 and 4.

Minor comments:

1. The text is hard to read, especially the introduction. Long and complex sentence structures. I recommend a major edit.

   We edited the text of the introduction section in order to make it easier to read.

2. Papers reviewed and cited in the introduction are at least five years old. There has been significant research done on the issue, especially with the rise of new machine learning techniques. I recommend authors to investigate more recent literature and methods as well.

   We agree that it takes some years until studies are included in systematic reviews. Given the fact that we based on international evidence derived from systematic reviews, it is logical that we were not able to include very novel articles in the literature review. However, we also included recently published articles comparing different prediction techniques (e.g. Artetxe, Arkaitz, Andoni Beristain, and Manuel Graña. 2018., reference number 12).

Reviewer: 2
Reviewer Name: Tim Badgery-Parker

Institution and Country: Research Fellow, Menzies Centre for Health Policy, The University of Sydney, Australia

This paper aims to estimate the feasibility of using claims data available at admission to predict if a patient will have a subsequent readmission within 30 days. The authors conclude that the data available at admission are suitable for this purpose, but that poor model performance (both in their model and in other similar studies) makes this approach not worthwhile.

First, it seems to me that poor model performance is to be expected. As the authors note, high readmission rates are used as an indicator of poor quality of care. They are only useful as such an indicator if they reflect what happens during the hospital stay — information that is explicitly excluded by the authors' use of only data available at admission.

Second, I think the authors have handicapped their model performance by their methodological decisions. Many of these are common practice, but not best practice. The authors categorised continuous variables to account for non-linearity. This greatly reduces the information in the data. They would be better modelling continuous variables with restricted cubic splines or polynomial terms. The variable selection process was reasonable, but a better prediction model might have been derived from a penalised regression model such as lasso or elastic net. Furthermore, the analysis does not seem to account for clustering of patients within hospitals. Use of readmission rates as a quality indicator requires that hospitals have different rates, so hospital should be included in the model.

The authors do mention in the limitations section that categorising is not the best approach and suggest that other models such as random forest might have better performance. I think that even within logistic regression they could get better prediction if they model continuous variables as continuous and account for hospital clustering.

In addition, they appear to have estimated AUC by first dichotomising the modelled probability into a binary prediction then comparing that with the true value. They should use the modelled probability directly in estimating the AUC, which is supposed to be a summary statistic accounting for all possible dichotomisation thresholds.. If they calculate AUC properly, they will probably see a substantial increase.

We thank the reviewer for his reflective review which helped us to improve the manuscript significantly. We respond to the specific comments raised below.

Specific comments

Abstract

1. PCG needs to be explained in the abstract.

   Done

Methods

2. Patients can change health insurers once per year — how did the analysis deal with people entering or leaving cohort? Are the variables on previous use of health services available for someone who enters the cohort during the study period?

   We restricted our study population to persons with at least one year of complete data prior to the index event (the index hospitalisation), as well as complete data during the maximal follow-up period (90 days). We extended the description of the data preparation process in the methods section (lines 162-164) to clarify this.
3. Variables on prior use of health services are not clearly defined. This appears to be use in the year before the admission, but this is not explicitly stated. If it is a year, was there data for a full year before admission for all patients?

   This is correct, prior use of health services was examined within one year prior to the index hospitalisation. Data for a full year before the index admission was available for all patients (see comment pt 2). We added a sentence to the methods section (lines 138-139) to clarify this.
4. line 112 (literature search): predictors were grouped into 5 categories but only 4 are listed, while additional table 1 appears to list 7 categories

Thank you for careful reading. This was a mistake which we corrected. The predictors were grouped into seven categories.

5. Although commonly done, categorising continuous variables loses a lot of information. If nonlinearity is a concern, consider splines.

   Based on the reviewer comment we added a sensitivity analysis including a model without categorisation of continuous variables. We centered, scaled and transformed (Yeo-Johnson transformation) all continuous variables. To account for non-linearity, we used restricted cubic splines (with 5 knots) for all these variables. This did not improve the model discrimination. The AUC in the model with the entire population was 0.60 (95%-CI: 0.60 - 0.61). Please see methods section lines 155-156 and discussion section lines 304-306.

6. The model really should include hospital (I would include as random intercept) to account for clustering of patients within hospitals.

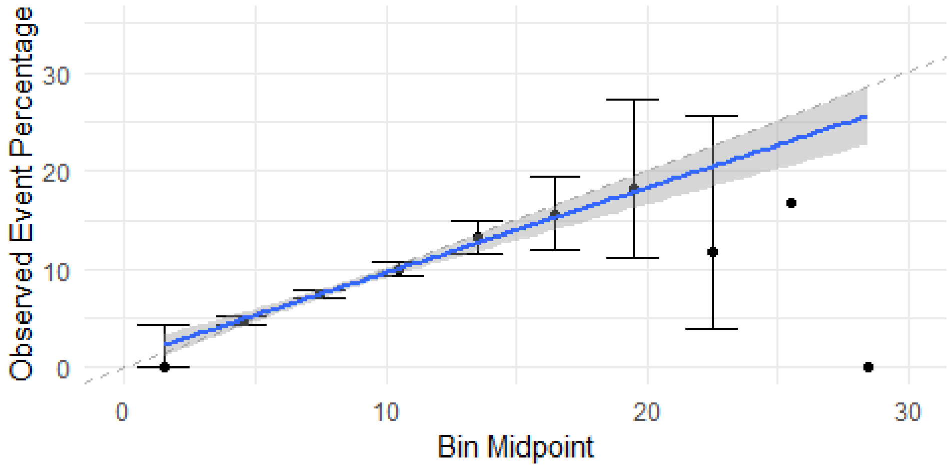
   As a sensitivity analysis we calculated a mixed-effects model with random intercepts per hospital. The model performance did increase slightly with this change to 0.63 (95%-CI: 0.61 - 0.64). Please see methods section lines 179-180 and discussion section 304-306. Since we focused on patient-level predictors we did not include these results into the manuscript. Although, we suggest to add detailed hospital-specific data to future research. Please see discussion section lines 381-387.

7. You should not dichotomise the predicted probability to estimate the AUC; use the actual predicted value from the model.

   We apologise for unclarity in the previous version of the manuscript. We did not dichotomise the predicted outcome to calculate AUC but must admit that the description was misleading. We rephrased the methods section accordingly (please see methods section line 202, and 209-213).

8. Sensitivity, specificity, PPV, and NPV are not suitable for assessing model performance. AUC (c statistic) is appropriate as measure of discrimination. You might also consider examining calibration with appropriate plot.

   According to the reviewer comment we added results on the calibration of the prediction model (please see methods section lines 203-207, and results section lines 285-287). The calibration plot showed a good calibration. We added only the intercept (calibration-in-the-large) and the slope of the weighted regression line (number of observations as weights) in the calibration plot. For the sake of conciseness of the manuscript we did not include the plot in the manuscript. Please see the plot below for reviewers only. However, if the editor or reviewers feel that this information is crucial to the reader we are of course willing to revise and include it.



Calibration plot, model with categorised continuous variables, weighted regression line (blue)

9. You say you set threshold to balance sensitivity and specificity. It is not clear if you then used the same threshold in all the subgroup analyses, or reset it to balance sensitivity and specificity in the subgroups.

   The cut point was calculated in each supgroup seperately in the previous version. According to the reviewer comment we do not report sensitivity and specificity for subgroups anymore in the present version (please see results section lines 288-291).

Results

10. You comment that best PPV was in subgroup MDC kidney and urinary tract. But PPV depends on prevalence. Examining table 1 would tell you, before you ever calculated PPV, that this subgroup would have the highest PPV. As I said above (8), don't use PPV/NPV to assess the model.

    Presentation of PPV in the previous version of the manuscript was misleading given PPV only shows the success of the prediction model, depending on the prevalence of the event in the sample. In the revised version of the manuscript we focus on reporting of PPV, NPV, sensitivity and specificity for the total study population (please see results section lines 288-291). As mentioned in the response to the reviewer comment above, instead, we added results of the calibration and k-fold cross-validation for the AUC estimates (please see methods section lines 202-207, and results section tables 3 and 4 and lines 285-287).

Discussion

11. line 278: "principally" – I think you mean "in principle"

    Done

12. line 332: Surely any index admission after 2014 would have 0 for prior inpatient costs, not just those towards the end of the 3 year period. But if every index episode has data for a full year before, the overall distribution of inpatient costs should be correct. Anyway, inpatient costs were not included in the model, so this part of discussion seems to be irrelevant.

    Patients with an index admission after 2014 not only have zero inpatient costs in the previous year. Also, the minimal time span without any inpatient care increases to up to three years for *patients at the end of the screening period. But since inpatient costs are not part of the final model we deleted this part from the discussion section as it is not relevant for the reader.*

13. line 299: you cite PREADM authors that performance is likely to be low because most readmissions are not avoidable. The logic of this is not clear to me. Just because a readmission is unavoidable does not necessarily make it hard to predict from patient data. Or am I missing something?

    We added text for clarification of this statement (please see discussion section lines 324-325).

VERSION 2 – REVIEW

REVIEWER	Tim Badgery-Parker Research Fellow, Menzies Centre for Health Policy, The University of Sydney, Australia
REVIEW RETURNED	05-Apr-2019

GENERAL COMMENTS	The authors have better explained and justified the modelling choices, and done some sensitivity analyses which apparently did not make much difference. Those models should probably be in the supplementary material. I still don't like hospital not being in the main model presented. Region is, but is that because of different patient factors in regions or different hospital factors in regions (assuming most patients go to a hospital in their region -- I don't know how mobile people are in Switzerland)? Without hospital in the model, this is impossible to know. I think you also need to discuss more about why the model is not useful, beyond just saying it has low AUC and PPV. What performance do you need before a model would be useful? (actually, as I write this I realise it is not even clear who is intended to use the model -- clinicians? managers? insurers?) You currently have NPV of 95%. Is that useful? Does it help a hospital to know who will probably not have a readmission?

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: Tim Badgery-Parker

Institution and Country: Research Fellow, Menzies Centre for Health Policy, The University of Sydney, Australia

The authors have better explained and justified the modelling choices, and done some sensitivity analyses which apparently did not make much difference. Those models should probably be in the supplementary material.

We added the results of the model without categorisation of the continuous variables to the supplementary material (please see supplementary appendix 3). In contrast, we think that the results of the mixed effects model with random intercepts per hospital should not be included in the manuscript. They were created as part of an exploratory analysis in response to the reviewer comment. Since we did not plan to analise the influence of the hospitals in the original design of this study, we had to re-extract the data with additional information on hospitals from the database, which slightly changed in the meantime due to progressing claims settlement in the observed time period. Furthermore, for a more thorough analysis the avaiable data on hospitals would require a refactoring. It is only avaiable on a too low granularity (sometimes whole hospitals, sometimes separate clinics), leading to very low patient counts. 39% of the identified hospitals/clinics have less than 500 patients in the study population, 30% less than 100 patients, and some have only one patient. The mixed effects model based on a restricted population. Patients hospitalised to a hospital/clinic with less than 500 (AUC = 0.63 (95%-CI: 0.61 - 0.64)) and 100 (AUC = 0.63 (95%-CI: 0.63 - 0.64)) patients in the study population were exlcuded.

I still don't like hospital not being in the main model presented. Region is, but is that because of different patient factors in regions or different hospital factors in regions (assuming most patients go to a hospital in their region – I don't know how mobile people are in Switzerland)? Without hospital in the model, this is impossible to know.

Unfortunately, as mentioned above, including hospitals in an appropiate way was not planned in the original study design and exceeds this projects possibilities and ressources. The region variable in the model is included primarily to account for cultural differences within the study population (there are large differencies in regional health care spending). But as you assume correctly, mobility with respect to hospitalisations is very limited in Switzerland (Huber 2015, vol. 48, fig. 3.13), especially in the large regional entities. We agree with you that adding information about the hospitals to the model could potentially improve its performance. As mentioned in the discussion (see lines 397-400), we suggest doing that in future reasearch.

I think you also need to discuss more about why the model is not useful, beyond just saying it has low AUC and PPV. What performance do you need before a model would be useful? (actually, as I write this I realise it is not even clear who is intended to use the model – clinicians? managers? insurers?) You currently have NPV of 95%. Is that useful? Does it help a hospital to know who will probably not have a readmission?

Thank you for this feedback. We clarified who the target audience of the developed model is (see lines 79-82). And we also sharpend our argument why we think the performance of our model is below a useful level (see lines 299, 339-344).

Bibliography

Huber, Kathrin. 2015. *Entwicklung Der Interkantonalen Patientenströme Im Übergang Zur Freien Spitalwahl: Analyse Der Stationären Akutsomatischen Spitalbehandlungen von 2010 Bis 2013*. Vol. 48. Obsan Dossier. Neuchâtel: Schweizerisches Gesundheitsobservatorium (Obsan).