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)	Development and validation of a score to assess complexity of general internal medicine patients at hospital discharge: a prospective cohort study
AUTHORS	Liechti, Fabian; Beck, Thomas; Ruetsche, Adrian; Roumet, Marie; Limacher, Andreas; Tritschler, Tobias; Donzé, Jacques

VERSION 1 – REVIEW

REVIEWER	Schwab, Camille Hôpital Saint-Antoine, Pharmacie
REVIEW RETURNED	30-Jun-2020

GENERAL COMMENTS	Thank you for the opportunity to review this interesting paper. However, I have some remarks: In the introduction section, the authors present the Charlson Comorbidity Index, but nothing is said about the Patient Clinical Complexity level but “used in the Swiss DRG system to allocate reimbursement”. In the methods section, the authors used “an estimated proportion of complex patients of one fourth” for the calculation of the study size, whereas in the introduction they mentioned “up to one third of patients are estimated to be complex”. Why is there a difference between these two proportions? For the study outcomes, it is not clear how the outcome was assessed. Was it the treating physician’s or the trained study nurse? For the polypharmacy, why have the authors used ≥ 10 drugs, whereas the most commonly reported definition of polypharmacy is the numerical definition of ≥ 5 medications daily (Nashwa Masnoon et al., BMC Geriatr. 2017; 17: 230. doi: 10.1186/s12877-017-0621-2). For the variables the authors explained how they established a list of predictor variables with a survey study (is the “Delphi method” term appropriate?), that some variables were removed if they were strongly correlated to another variable, and that only the strongest univariate predictor was kept. On what criteria did the authors qualify the variable as the strongest? Have the authors conducted bivariate analysis between the “high risk “group and the “low risk group” to select variables? The missing data section pointed out the fact that the authors have not mentioned which data were collected. They wrote that the data were “collected retrospectively through the electronic health record of the hospital”, but was it the latest value, at discharge? We understand that it is the value at discharge when reading the results, but in the missing section, for “the second value of hemoglobin and creatinine” we don’t know what the first value was.

	In the statistical analysis section, the author wrote that “the reference point (cut-off) of each scoring system was chosen in order to make the frequency of patients in the high-risk category as close as possible to 30%”. Why have the authors not use the Youden index? Results section: The title of the table 1 is: “Baseline characteristics for all patients”. But in the three columns, the sum of each group is not equal to n, and the percentage not equal to 100: For example: age/overall: 579 + 437 + 322 + 537 + 873 = 1875 (≠ 1889) and 31 + 23 + 17 + 28 + 46 = 99 (≠100). Can the authors correct the table or explain the difference? Page 12 line 10-12, there is an error on the reference source. The authors mentioned the ROC curves, but it would be interesting to see them. In the discussion section, the authors wrote “In our setting, discharge planning processes for older patients may be better established”. I don’t understand the link between the sentence and the results. Have the patients had interventions during their hospital stay? In the limitations section it is written that “the included predictors are not modifiable…a patient will still be complex if receiving less imaging procedures to reduce costs”. So I assume it is the same here, a well establish discharge planning process will not turn a high risk patient to a low risk patient? To finish, would like to say that I very much appreciate the limitations section: the subjective outcome, the non-generalizability and the use of the PCA restricted to population-based studies are the three major limits of your study, but taken in this context, the PCA is, thus, a very interesting tool indeed.

REVIEWER	Cardona, Magnolia Bond University, Institute for Evidence Based Healthcare
REVIEW RETURNED	06-Sep-2020

GENERAL COMMENTS	Thank you for the opportunity to review your very important work of using routinely collected clinical and administrative data to enhance the profile of complex patients and certainty of prediction of their future health service use. This concept has enormous practical applications across health settings. The cohort design with full census of consecutive patients, documentation by a trained nurse, logistic regression with backward elimination, estimation of sensitivity, specificity, PPV, NPV and AUROC, score weighting according to coefficients, as well as reporting following TRIPOD statement were all appropriate. Below are a few queries and clarifications I thought the authors could address to improve the value of the article for application in routine care. Page 6, lines 39-41: A missing piece of information is the ultimate goal and anticipated usefulness of a proxy for physicians’ assessment on discharge (“valuable surrogate” why valuable; not just because subjective judgment is imperfect but what for if the patient is about to leave hospital). I can see it would be useful to flag patients who may need community support services post-discharge, or to predict those who will return to hospital and may benefit from intensive self-management training/counselling. The authors need to emphasise their vision and rationale for the calculation in terms of clinical implications, and I am not referring

	to the statistical value as presented “benefit gains relative to the prediction complexity” of p9, l41-43). P7, line 39-40: The primary outcome should be stated as “predictive accuracy of the PCA against the treating physician’s judgment as the gold standard” P8, lines 9-10: the authors need to justify how the other “readily available potential predictors” made it to the list. Were they based on evidence from other studies? Consultation with experts? Personal clinical experience? P8, line 18: I wonder since the final score cannot be estimated until the end of hospitalisation due to the inclusion of procedures during hospitalisation and test results at admission and discharge, perhaps the word “predicting” would be better replaced with the word “representing” or “denoting” complexity. Likewise for p16, lines 18-20 where 6 diagnoses are “predictor” of complexity. And p16, and line 44. They are not predicting; they are “flagging” “identifying” or “indicating” complexity. P8, lines 48-54: Please report the numbers/proportions missing for estimation of PCA. Have the authors considered that the second-value variables which required imputations and assumption should be excluded from analysis if not routinely available? If the objective of the score is to use routinely collected and readily available data, then why should the second value be suggested as part of the PCA if not usually performed or documented at the two time points? Just use the value available on admission or the one that is more complete at discharge to make the score more translatable. P9, lines 34-36. At this point I believe authors need to state why they chose a validation sample size that was three times smaller than the derivation sample and collected only for one month instead of the same length as the derivation sample. The value of Table 1 is diminished by presenting overall results. Instead, it is important and more informative to show the distribution of all characteristics by complex/non-complex for derivation and validation cohorts side by side. P12 I could not find references to Table 2 in the text describing the crucial table contents. P14, lines 3-7. For the benefit of non-statisticians among your readership, remind us how you arrived at 24 points as the high risk cut-off level (out of the theoretical max 81 points). Was that the point of best sensitivity and specificity when compared with physicians’ judgment? Discussion While I see value in examining the predictive ability of your risk indicator list, and comparing with CCI and PCCL, I do not agree that clinician assessment was a good choice of “gold standard” or that “there is no better standard reference” for the obvious reason of subjectivity and unreliability. I think adding DRG to the clinical judgment would have been a more robust and pragmatic comparator or even the DRG alone due to its integration of objective clinical parameters and resource

	utilisation. in that sense, the current AUROC curves comparing objective vs. subjective may not tell the full story. There is still a knowledge gap the authors could fill. I believe if the DRG data are available, the additional comparison of your new PCA score with DRG (alone or in combination with physician's assessment) would really enhance the value of your analysis to inform practice change. Finally, the concluding paragraph does not do justice to all the preceding good work. It needs a better punchline with more detail in relation to implications for practice to be a meaningful take-home message as the readers are not only statisticians or scientists, but more likely clinicians.

REVIEWER	Tang, Terence Trillium Health Partners, General Internal Medicine
REVIEW RETURNED	08-Sep-2020

GENERAL COMMENTS	* I do not have sufficient statistical expertise to review the statistical component of this manuscript. Further statistical review is required. * The authors developed a prediction score (using variables typically available at hospital discharge) to predict "complex patients". They used treating physician's subjective assessment as gold standard for "complex patients" and derived a score that seemed to perform better than Charlson (CCI) or PCCL. This work is interesting, and the question is of significance. MAJOR comments: 1. The definition of complexity used in this study is not clear. The authors clearly state that it is different from multimorbidity, but have not articulated a clear explanation of the term "complex patient". Some references were made to increased utilization of healthcare resources as a component of complexity, and also psychosocial factors. 2. The "gold standard" of complex patient is the subjective opinion of the treating physician. This may be of concern if the concept of complexity has not been clearly articulated and it is unclear whether this understanding is consistent between different treating physicians for different patients. Do we know whether different physicians treating the same patient will agree on whether a patient is complex? Perhaps, rather than simply using the treating physician's subjective opinion, using opinions from two physicians (e.g. a secothrough chart review) may help understand the performance of this "gold standard". I see this as a MAJOR issue for this paper. 3. Some limitations have not been sufficiently emphasized, especially issues with generalizability. This study has taken place in a single centre, and I have concerns about generalizability. I am worried that some of the variables in the final model are health system (or even hospital) specific and it will be difficult to know whether this model holds true even at a different hospital in Switzerland or another health system (e.g. Canada). Examples of concern include elective admission (the prevalence or definition of elective admission may be very sensitive to local context), malnutrition (may not be consistently coded in health administrative data at a different hospital or health system), measurement of costs and nursing workload may also be quite sensitive to a particular model of care and health system context.

	The issue with generalizability of this single-centre study has not been sufficiently emphasized. MINOR COMMENTS: 1. Suggest adding the phrase "at discharge" somewhere. For example: "Development and validation of a new tool to assess inpatient complexity at hospital discharge: The Patient Complexity Assessment (PCA) score". Readers may initially be under the impression that the score can predict complexity and resource utilization at admission. 2. Readers may also be interested in knowing how the PCA score correlates with certain outcomes (e.g. death, re-admission, and resource utilization). 3. I assume that treating physicians were asked to classify each patient as either complex or not complex (binary variable), hence the way the analysis is structured (i.e. having a cut-point of 24 for PCA score). This methodological details/rationale deserves explicit explanation.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Camille Schwab
Institution and Country: Service de Pharmacie, Hôpital Saint Antoine, AP-HP.Sorbonne Université, Paris
Please state any competing interests or state ‘None declared’: None declared

Comments to the Author
Thank you for the opportunity to review this interesting paper. However, I have some remarks:
In the introduction section, the authors present the Charlson Comorbidity Index, but nothing is said about the Patient Clinical Complexity level but “used in the Swiss DRG system to allocate reimbursement”.
Thank you for your remark. We added the following paragraph in the introduction:
“The patient clinical complexity level (PCCL) is calculated for each treatment episode to measure the cumulative effect of complications and comorbidities in a patient. The PCCL ranges from 0 (no complication or comorbidity) to 4 (very severe complication or comorbidity), according to a complex machine learning algorithm. [13,14] Identification of complex patients at discharge could help to identify those, who would profit from more intense follow-up, e.g. by general practitioners or social workers, although effectiveness of such interventions would have to be proven first.”

In the methods section, the authors used “an estimated proportion of complex patients of one fourth” for the calculation of the study size, whereas in the introduction they mentioned “up to one third of patients are estimated to be complex”. Why is there a difference between these two proportions?
Thank you for your comment. We referred to the study of R Grant et al. where 26.2% of patients were identified as complex. The manuscript has been corrected accordingly in the introduction: “One fourth of patients are estimated to be complex in the primary care setting, while this proportion is not well known in the hospital setting.[1-3]”

For the study outcomes, it is not clear how the outcome was assessed. Was it the treating physician’s or the trained study nurse?
The outcome was assessed by the treating physician. The manuscript has been complemented in the following way: “The resident (or supervising consultant) was asked by a trained study nurse to assess at time of discharge the level of complexity of the entire hospital stay of her/his patient (complex or

not-complex).

For the polypharmacy, why have the authors used ≥ 10 drugs, whereas the most commonly reported definition of polypharmacy is the numerical definition of ≥ 5 medications daily (Nashwa Masnoon et al., BMC Geriatr. 2017; 17: 230).
We agree that the cut-off is often set at ≥ 5 medications in the outpatient setting. No universal definition of polypharmacy was established at the time of study design (2016). There is large heterogeneity in the definition of polypharmacy, as also mentioned in the publication of Nashwa Masnoon that you cited. Numerical definitions range up to 11 or more medications. The proportion of medical inpatients with polypharmacy is high and usually includes addition of acute treatment to the chronic medication, e.g. antibiotics, analgesia, diuretics, electrolyte substitution, prevention of venous thrombosis. In our study, polypharmacy was defined as ≥ 10 medications. This corresponds to the report of Masnoon et al. which write in the chapter “Numerical definitions of polypharmacy incorporating a duration of therapy or healthcare setting”: “Polypharmacy definitions incorporating a healthcare setting included the use of five or more medications at hospital discharge, and the use of 10 or more medications during hospital stay.” This reference is therefore included in the manuscript. Regarding the proportion of patients with «polypharmacy» in the derivation and validation cohort (ranging between 35% and 62%) we think, the cut-off is well justified.

For the variables the authors explained how they established a list of predictor variables with a survey study (is the “Delphi method” term appropriate?), that some variables were removed if they were strongly correlated to another variable, and that only the strongest univariate predictor was kept. On what criteria did the authors qualify the variable as the strongest? Have the authors conducted bivariate analysis between the “high risk “group and the “low risk group” to select variables?
Thank you to give us the opportunity to clarify how we derived the prediction variables. Candidate predictors are never exhaustive and most studies are using a priori predictors that are readily available. It is therefore a usual approach to include a broad range of candidate predictors obtained from different resources such as patient demographics or clinical history before identifying the important predictors using statistical methods (Royston P, Moons KG et al., 2009 Prognosis and prognostic research: developing a prognostic model. BMJ 338: b604). The study of the 111 physicians cited in our manuscript was not designed to search for predictors, but was used for the current study in order to extend the list of potential predictors and not to miss possibly important predictors.
We did not run bivariate analysis before multivariate analysis. In case of strong correlation between two variables, the strongest univariate predictor was kept. By strongest predictor, we mean the predictor with the highest odds ratio, as mentioned in the methods section.
To avoid confusion we rephrased this section as follows: “Candidate predictor variables have been selected based on a previous survey among general internists in the hospital setting which asked them to identify factors that contribute to patient complexity,[4] and on a selection of readily available potential predictors to have a broad spectrum of candidate predictors. Variables that were not routinely collected were removed (i.e. variables with more than 25% missing data, such as aspartate amino transferase, C-reactive protein, and albumin at discharge). Collinearity between variables was assessed using Pearson correlation coefficients. In case of strong correlation ($r > 0.7$), only the strongest univariate predictor was kept.”

The missing data section pointed out the fact that the authors have not mentioned which data were collected. They wrote that the data were “collected retrospectively through the electronic health record of the hospital”, but was it the latest value, at discharge? We understand that it is the value at discharge when reading the results, but in the missing section, for “the second value of hemoglobin and creatinine” we don’t know what the first value was.
Lab values included the first available at admission and the last available before discharge. We now adjusted the manuscript as follow: (“(…) laboratory values (hemoglobin, leucocyte count and

thrombocyte count, serum sodium and creatinine) at admission (first lab values at admission) and discharge (last lab values before discharge) (…)"

In the statistical analysis section, the author wrote that "the reference point (cut-off) of each scoring system was chosen in order to make the frequency of patients in the high-risk category as close as possible to 30%". Why have the authors not use the Youden index?
Thank you for your comment. We aimed for development of a new scoring system using the derivation cohort. The Youden index can be used to derive a cut-off for best dichotomous discrimination. However, in the present study, the proportion of complex patients was predefined, according to the observed proportion of complex patients, to compare the 3 scoring systems (PCA score, Charlson score, PCCL) using the validation cohort.

Results section:
The title of the table 1 is: "Baseline characteristics for all patients". But in the three columns, the sum of each group is not equal to n, and the percentage not equal to 100:
For example: age/overall: 579 + 437 + 322 + 537 + 873 = 1875 (≠ 1889) and 31 + 23 + 17 + 28 + 46 = 99 (≠100). Can the authors correct the table or explain the difference?
The number of missing values of subcategories were omitted in the table 1. We have added these values in the revised manuscript in order to make the add-up transparent.

Page 12 line 10-12, there is an error on the reference source.
We apologize for the inconvenience, the reference has been corrected.

The authors mentioned the ROC curves, but it would be interesting to see them.
ROC curves have been added as supplementary material.

In the discussion section, the authors wrote "In our setting, discharge planning processes for older patients may be better established". I don't understand the link between the sentence and the results. Have the patients had interventions during their hospital stay? In the limitations section it is written that "the included predictors are not modifiable…a patient will still be complex if receiving less imaging procedures to reduce costs". So I assume it is the same here, a well establish discharge planning process will not turn a high risk patient to a low risk patient?
Geriatric patients are often indiscriminately discharged to other institutions, e.g. geriatric rehabilitation centers, acute geriatric hospital or nursing homes. It is much more difficult to find such institutions for younger, complex patients. We assume therefore, that the discharge process for some younger patients can be more difficult for residents (e.g. including more contacts with externals) and therefore this may add to the perception that such patients are considered as complex. Moreover, as tertiary clinic for internal medicine, younger patients with only few diagnoses may be handled by specialized clinics or as outpatients. Nevertheless, a well-established discharge process will probably not turn a complex patient in a non-complex patient because more factors define patient complexity.
The manuscript has been adapted to respond to your concerns.

To finish, would like to say that I very much appreciate the limitations section: the subjective outcome, the non-generalizability and the use of the PCA restricted to population-based studies are the three major limits of your study, but taken in this context, the PCA is, thus, a very interesting tool indeed.
Thank you!

Reviewer: 2
Reviewer Name: Magnolia Cardona
Institution and Country: Institute for Evidence Based Healthcare, Bond University, QLD, Australia
Please state any competing interests or state 'None declared': None to declare

Comments to the Author
Thank you for the opportunity to review your very important work of using routinely collected clinical and administrative data to enhance the profile of complex patients and certainty of prediction of their future health service use.

This concept has enormous practical applications across health settings. The cohort design with full census of consecutive patients, documentation by a trained nurse, logistic regression with backward elimination, estimation of sensitivity, specificity, PPV, NPV and AUROC, score weighting according to coefficients, as well as reporting following TRIPOD statement were all appropriate.

Below are a few queries and clarifications I thought the authors could address to improve the value of the article for application in routine care.

Page 6, lines 39-41: A missing piece of information is the ultimate goal and anticipated usefulness of a proxy for physicians' assessment on discharge ("valuable surrogate" why valuable; not just because subjective judgment is imperfect but what for if the patient is about to leave hospital). I can see it would be useful to flag patients who may need community support services post-discharge, or to predict those who will return to hospital and may benefit from intensive self-management training/counselling. The authors need to emphasise their vision and rationale for the calculation in terms of clinical implications, and I am not referring to the statistical value as presented "benefit gains relative to the prediction complexity" of p9, l41-43).
Thank you for your valuable comment. A statement in the manuscript addresses now your concerns: "Identification of complex patients at discharge could help to identify those, who would profit from more intense follow-up, e.g. by general practitioners or social workers, although effectiveness of such interventions would have to be proven first."

P7, line 39-40: The primary outcome should be stated as "predictive accuracy of the PCA against the treating physician's judgment as the gold standard"

The manuscript has been changed accordingly.

P8, lines 9-10: the authors need to justify how the other "readily available potential predictors" made it to the list. Were they based on evidence from other studies? Consultation with experts? Personal clinical experience?
Thank you to give us the opportunity to clarify how we derived the prediction variables. Candidate predictors can never be exhaustive and most studies are using a priori predictors and predictors from the literature. Morevoer, to build a useful prediction tool, the predictors have to be measurable and readily available. It is therefore a usual approach to include a broad range of candidate predictors obtained from different resources such as patient demographics or clinical history before identifying the important predictors using statistical methods (Royston P, Moons KG et al., 2009 Prognosis and prognostic research: developing a prognostic model. BMJ 338: b604). The study of the 111 physicians cited in our manuscript was not designed to search for predictors, but was used as a source for potential predictors, as other published scientific literature.
Only the strongest predictor of those correlating was kept as mentioned in the methods section.
To avoid confusion we rephrased we rephrased this section as follows: "Candidate predictor variables have been selected based on a previous survey among general internists in the hospital setting which asked them to identify factors that contribute to patient complexity,[4] and on a selection of readily available potential predictors to have a broad spectrum of candidate predictors. Variables that were not routinely collected were removed (i.e. variables with more than 25% missing data, such as aspartate amino transferase, C-reactive protein, and albumin at discharge). Collinearity between variables was assessed using Pearson correlation coefficients. In case of strong correlation ($r > 0.7$),

only the strongest univariate predictor was kept."

P8, line 18: I wonder since the final score cannot be estimated until the end of hospitalisation due to the inclusion of procedures during hospitalisation and test results at admission and discharge, perhaps the word "predicting" would be better replaced with the word "representing" or "denoting" complexity. Likewise for p16, lines 18-20 where 6 diagnoses are "predictor" of complexity. And p16, and line 44. They are not predicting; they are "flagging" "identifying" or "indicating" complexity.
Thank you for questioning our unambiguous use of the term "predictor". In the predicting modelling, it is common to use the term "predictor", however we agree that it may be confusing to use this wording throughout the manuscript when in fact these variables are not used to predict a future outcome but identify a current state (complexity). The use of the word «predictor» or «predicting» has been reviewed throughout the manuscript and been replaced by «indicator» or «indicating» if appropriate.

P8, lines 48-54: Please report the numbers/proportions missing for estimation of PCA. Have the authors considered that the second-value variables which required imputations and assumption should be excluded from analysis if not routinely available? If the objective of the score is to use routinely collected and readily available data, then why should the second value be suggested as part of the PCA if not usually performed or documented at the two time points? Just use the value available on admission or the one that is more complete at discharge to make the score more translatable.
The authors have indeed considered excluding those values. However, as mentioned in the methods section, we had predefined, that only variables with more than 25% missing data were removed. The second value of hemoglobin and creatinine were missing in 18.6% and 19.3%, respectively. We assumed these values would be stable if not asked for a second time by the physicians in charge of the patient, which is a pragmatic approach to reflect in our views clinical reality. In addition, laboratory values measured before hospitalization may have been made available by outpatient physicians, e.g. family doctors, which may prevent hospitalists from further control of chronic anemia or chronic renal failure, while in acute aggravation, a laboratory follow-up will usually be part of good clinical practice. The second value of sodium and platelet count were missing in 16.1% and 18.6%, respectively. Likewise, we assumed in a pragmatic approach, these values were normal, if not asked for a second time by the physician in charge because it is clinical routine, to control these parameters until normal or near-normal values. We estimate that this approach is accurate in the clinical setting and enhances usability of the PCA score, because abnormal values or changes are included in the score. The manuscript has been adopted for better understanding.

P9, lines 34-36. At this point I believe authors need to state why they chose a validation sample size that was three times smaller than the derivation sample and collected only for one month instead of the same length as the derivation sample.
We understand your interrogation point. We preset the sample size of the derivation cohort to be 1,400 (based on sample size calculation). We are now clarifying the justification of sample size difference between the derivation and the validation cohort as follow: "We originally planned to consider around 35 variables in the prediction model. With an estimated proportion of complex patients of one fourth, we preset the sample size of the derivation cohort to be 1,400 (rule of thumb of 10 outcomes per variable tested). We predefined, that if more than 1,400 patients will be included during the study period of 6 months, we would use these patients to externally validate the prediction model." On February 16, 2017, the required number of participants for the derivation cohort was reached (n = 1,407). Hence, the 482 participants included after this date were representing the validation cohort. It is not required to use the same number of participants to externally validate the model. Moreover, when the number of patients is limitated, it is recommended to use a larger sample size for the derivation dataset in order to avoid overfitting which is not a problem for the validation (see e.g. E.W. Steyerberg, Clinical Prediction Models, Springer Science and Business Media, LLC, New York, 2009).

The value of Table 1 is diminished by presenting overall results. Instead, it is important and more informative to show the distribution of all characteristics by complex/non-complex for derivation and validation cohorts side by side.
We think that for better readability the final manuscript it is sufficient to show the overall patient characteristics. We added the distribution of derivation and validation cohorts in to the supplementary materials.

P12 I could not find references to Table 2 in the text describing the crucial table contents.
We apologize for the inconvenience. The reference has been corrected.

P14, lines 3-7. For the benefit of non-statisticians among your readership, remind us how you arrived at 24 points as the high risk cut-off level (out of the theoretical max 81 points). Was that the point of best sensitivity and specificity when compared with physicians' judgment?
The cut-off was chosen to approximate the frequency of observed complex patients in the PCA, Charlson score and PCCL (30%). We have stated in the methods section the following: "Applying PCA, CCI and PCCL, we calculated the score of each patient and split the patient sample into a high and a low risk group. The reference point (cut-off) of each scoring system was chosen in order to make the frequency of patients in the high-risk category as close as possible to 30% (i.e. approximating the frequency of observed complex patients)."

Discussion
While I see value in examining the predictive ability of your risk indicator list, and comparing with CCI and PCCL, I do not agree that clinician assessment was a good choice of "gold standard" or that "there is no better standard reference" for the obvious reason of subjectivity and unreliability.
I think adding DRG to the clinical judgment would have been a more robust and pragmatic comparator or even the DRG alone due to its integration of objective clinical parameters and resource utilisation. in that sense, the current AUROC curves comparing objective vs. subjective may not tell the full story. There is still a knowledge gap the authors could fill. I believe if the DRG data are available, the additional comparison of your new PCA score with DRG (alone or in combination with physician's assessment) would really enhance the value of your analysis to inform practice change.
Thank you for sharing your concerns. You correctly state, that there is no good choice of how to identify complex patients. Therefore, up to now, the treating physician's judgement is considered «gold standard» to identify complex patients. Simplified, the DRG only sums up diagnoses, while complexity is not limited to multimorbidity and depends on other aspects, as mentioned in the manuscript ("Complexity is not limited to multimorbidity and chronicity of disease but depends also on multiple other aspects, including psychological, social, economic and environmental factors.[1,2,5-7]"). Moreover, the PCCL is a measure based mainly on DRG data. To avoid confusion, we have added the following sentence in the introduction: "Generally, those patients using more resources, time and/or effort are regarded as complex patients, although no universal definition of patient complexity is available."

Finally, the concluding paragraph does not do justice to all the preceding good work. It needs a better punchline with more detail in relation to implications for practice to be a meaningful take-home message as the readers are not only statisticians or scientists, but more likely clinicians.
Thank you for making this point. We tried to remain factual and cautious. However, we have modified the final paragraph, which states now: "Thereby, the PCA score might improve the monitoring of resources distribution and coordination of care, e.g. by flagging complex patients to general practitioners or social workers for closer follow-up or low-threshold service."

Reviewer: 3
Reviewer Name: Terence Tang

Institution and Country: Department of Medicine, University of Toronto, CANADA
Please state any competing interests or state ‘None declared’: None declared

Comments to the Author
* I do not have sufficient statistical expertise to review the statistical component of this manuscript. Further statistical review is required. *

The authors developed a prediction score (using variables typically available at hospital discharge) to predict "complex patients". They used treating physician's subjective assessment as gold standard for "complex patients" and derived a score that seemed to perform better than Charlson (CCI) or PCCL. This work is interesting, and the question is of significance.

MAJOR comments:
1. The definition of complexity used in this study is not clear. The authors clearly state that it is different from multimorbidity, but have not articulated a clear explanation of the term "complex patient". Some references were made to increased utilization of healthcare resources as a component of complexity, and also psychosocial factors.
Thank you for pointing at an important issue of patient complexity. No clear definition of patient complexity exists. We have therefore clarified in the introduction part of the manuscript as follows: “Generally, those patients using more resources, time and/or effort are regarded as complex patients, although no universal definition of patient complexity is available. Complexity is not limited to multimorbidity and chronicity of disease but depends also on multiple other aspects, including psychological, social, economic and environmental factors.[1,2,5-7]”
This is in accordance with the description in the methods section which we have also specified for better understanding: “The primary outcome was the predictive accuracy of the PCA against the treating physician’s judgment as the gold standard true complexity of hospitalized patients based on the treating physician’s judgement to identify complex general internal medicine inpatients (discharging resident physician or supervising consultant if the resident physician’s assessment was absent). Complex patients were defined as those using more resources, time and/or effort while hospitalized. The outcome physician’s assessment was prospectively collected by a trained study nurse at time of patient’s discharge by asking the treating physician (discharging resident physician or supervising consultant if the resident physician’s assessment was absent) if the patient was “complex” or “not-complex”.”

2. The "gold standard" of complex patient is the subjective opinion of the treating physician. This may be of concern if the concept of complexity has not been clearly articulated and it is unclear whether this understanding is consistent between different treating physicians for different patients. Do we know whether different physicians treating the same patient will agree on whether a patient is complex? Perhaps, rather than simply using the treating physician's subjective opinion, using opinions from two physicians (e.g. a secothrough chart review) may help understand the performance of this "gold standard". I see this as a MAJOR issue for this paper.
We appreciate your comment. As mentioned above, we have to keep in mind that no better gold standard allow to capture a holistic measure of patient complexity than the caregivers themselves. Complex care is multidimensional. The idea of asking a second opinion based solely on a chart review is interesting, however, the chart review doesn’t allow to capture neither the time spend to care for the patient, nor organize investigations and multidisciplinary rounds, as only few examples. Therefore, we believe that no better gold standard exists.
Our choice is comforted by the fact that the proportion of complex patients is consistent in different cohorts: 26% in the study of Grant RW, et al. 2011 vs. 32% and 24% in our derivation and validation cohort respectively. Therefore, although the «gold standard» of «patient complexity» is the treating physician’s judgement, in a large cohort study such as the present one, including over 1,400 patients and a large number of different treating physicians, a certain degree of objectivity can be reached by

multiple subjective evaluation, if no objective definition exists. We think that our study can contribute to develop a better and more objective definition of «patient complexity» by identifying indicators or predictors of complexity as identified by a large number of individual judgements.

3. Some limitations have not been sufficiently emphasized, especially issues with generalizability. This study has taken place in a single centre, and I have concerns about generalizability. I am worried that some of the variables in the final model are health system (or even hospital) specific and it will be difficult to know whether this model holds true even at a different hospital in Switzerland or another health system (e.g. Canada). Examples of concern include elective admission (the prevalence or definition of elective admission may be very sensitive to local context), malnutrition (may not be consistently coded in health administrative data at a different hospital or health system), measurement of costs and nursing workload may also be quite sensitive to a particular model of care and health system context. The issue with generalizability of this single-centre study has not been sufficiently emphasized.
We agree that generalizability may not be given due to the nature of the study. It is common practice to develop a prediction model on a specific cohort, which has first to be validated in other settings. Therefore, we would strongly encourage validation of our PCA score in other locations. However, we are confident that most aspects will be transferable to other settings, e.g. costs and nursing workload are not measured as absolute values but as those above 75th percentile while for malnutrition we used the ICD10 coding system. We have emphasized this concern in the discussion under the limitation section as follows: “(…) the PCA score has been developed at a single tertiary hospital in Switzerland and therefore may not be generalizable to other settings, e.g. other health care systems. However, costs and nursing workload are not measured as absolute values but as those above the 75th percentile, making it transferable to other settings. Also, some patients may appear as complex in one setting, while they will be judged as non-complex in other settings (e.g. primary care vs. university hospital), nevertheless the proportion of complex patients in out setting was similar to the one in primary care.[1]”

MINOR COMMENTS:
1. Suggest adding the phrase "at discharge" somewhere. For example: "Development and validation of a new tool to assess inpatient complexity at hospital discharge: The Patient Complexity Assessment (PCA) score". Readers may initially be under the impression that the score can predict complexity and resource utilization at admission.
We have adapted the title accordingly: “Development and validation of a score to assess complexity of general internal medicine patients at hospital discharge - a prospective cohort study”.

2. Readers may also be interested in knowing how the PCA score correlates with certain outcomes (e.g. death, re-admission, and resource utilization).
We agree that these outcomes would be interesting to be followed; however, they were beyond the scale of the present study, which did not have the resources for long-term follow up after hospital discharge.

3. I assume that treating physicians were asked to classify each patient as either complex or not complex (binary variable), hence the way the analysis is structured (i.e. having a cut-point of 24 for PCA score). This methodological details/rationale deserves explicit explanation.
This methodologic aspect has been specified in the methods section:
“The primary outcome was the predictive accuracy of the PCA against the treating physician’s judgment as the gold standard to identify complex general internal medicine inpatients. Complex patients were defined as those using more resources, time and/or effort while hospitalized. The physician’s assessment was prospectively collected by a trained study nurse at time of patient’s discharge by asking the treating physician (discharging resident physician or supervising consultant if the resident physician’s assessment was absent) if the patient was “complex” or “not-complex”.”

“The reference point (cut-off) of each scoring system was chosen in order to make the frequency of patients in the high-risk category as close as possible to 30% (i.e. approximating the frequency of observed complex patients).”

VERSION 2 – REVIEW

REVIEWER	Schwab, Camille Hôpital Saint-Antoine, Pharmacie
REVIEW RETURNED	22-Dec-2020

GENERAL COMMENTS	The authors have answered all of my concerns. I congratulate them for this work.

REVIEWER	Cardona, Magnolia Bond University, Institute for Evidence Based Healthcare
REVIEW RETURNED	28-Dec-2020

GENERAL COMMENTS	Thanks for the opportunity to review your important work on the development of a discharge-time score for patient complexity to spot those patients potentially at risk of needing closer monitoring and referral to other services in the community. The authors appear to have thoroughly addressed previous reviewers' comments. The article is already of publication standard but I have proposed minor amendments before final release that would leave readers with fewer questions. The optimal utilisation of administrative and clinical data already collected to develop the score was an efficient way to go about it, the comparison with two other routinely used indices of complexity, the consultation on candidate predictors and the exclusion of variables which are often missing in routine care and the use of TRIPOD checklist for reporting are all very appropriate. These are all ingredients for a useful contribution to knowledge that changes practice. However, having a subjective judgment as gold standard for complexity is the weak link although this was probably a pragmatic decision. To reassure readers, on Page 7, lines 46-49, the authors could clarify how the clinician arrived at the decision complex vs. not complex (using CCI or PCCL? Or some other loose set of criteria?) 52 parameters from the medical record. Unless their aggregation and calculation is automated, this will pose burden on hospital administration staff or clinicians before deciding on severity level. Also, for extrapolation/adaptation in other health systems, the final 11 indicators may vary from one setting to another in different health systems and therefore a recommendation to conduct PCA validations in other countries could start with the full set and see if the final indicators of complexity are the same as in the Swiss system; I think this would warrant a mention in the discussion. Page 9, lines 5-9: the assumptions of normality or last observation carried forward or imputation for missing at random may be customary in development and validation studies. However, given the extensive list of original candidate variables, perhaps the newly developed patient complexity score could have used only variables that were complete in routine care (not just 75% complete), as indicated in the methods/protocol.

	Page 10, lines 30-34: Add whether the clinical judgment (or some other index) was the method used to determine that 563 patients were complex. Page 12, Table 2: Effectively, the 11 variables included in the final score development yielded moderate to very good predictive ability, and the conclusion of the PCA superiority against CCI and PCCL holds and is backed up by the findings. Despite the authors' attempts to address the inverse relationship between age and complexity, and the counterintuitive relationship between elective (rather than emergency/unplanned) admissions and complexity are still puzzling. It is still unclear to me (and likely other readers) whether this was a coding error or a special circumstance of the participating tertiary hospital. Perhaps it is worth expanding the comment in the discussion on page 15. I'd suggest the next step in this research program to confirm the hypothesis that complex patients are higher-level users of other services or have worst outcomes than non-complex patients is to check the eligible patient outcomes (hospital reutilisation, use of social services, admission to nursing home, other complication rates and post-discharge mortality) after 6-12 months in a future study to compare with the prediction of low-high complexity cut-off points.

REVIEWER	Tang, Terence Trillium Health Partners, General Internal Medicine
REVIEW RETURNED	30-Mar-2021

GENERAL COMMENTS	I have reviewed an earlier version of this submission. I believe that the authors have adequately addressed my comments and this article is fit for publication. I will be interested in seeing a validation of this score outside of the setting of this current study. This is a score that identify complex hospital patients at discharge, and I see the major utility, as the authors mention, as to monitor complexity in hospital and performing complexity related study of hospitalized patients retrospectively.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Dr. Camille Schwab, Hôpital Saint-Antoine, Institut Pierre Louis d'Epidemiologie et de Sante Publique
Comments to the Author:
The authors have answered all of my concerns. I congratulate them for this work.

> Thank you. We appreciate the recognition of the improvements of the manuscript based on your suggestions.

Reviewer: 2
Dr. Magnolia Cardona, Bond University, Gold Coast University Hospital
Comments to the Author:

Thanks for the opportunity to review your important work on the development of a discharge-time score for patient complexity to spot those patients potentially at risk of needing closer monitoring and referral to other services in the community. The authors appear to have thoroughly addressed previous reviewers' comments. The article is already of publication standard but I have proposed minor amendments before final release that would leave readers with fewer questions.

> Thank you for acknowledging that the previous comments have been thoroughly addressed and that the manuscript reach already now the standard of publication. We are happy to take into consideration your new minor comments.

The optimal utilisation of administrative and clinical data already collected to develop the score was an efficient way to go about it, the comparison with two other routinely used indices of complexity, the consultation on candidate predictors and the exclusion of variables which are often missing in routine care and the use of TRIPOD checklist for reporting are all very appropriate. These are all ingredients for a useful contribution to knowledge that changes practice. However, having a subjective judgment as gold standard for complexity is the weak link although this was probably a pragmatic decision. To reassure readers, on Page 7, lines 46-49, the authors could clarify how the clinician arrived at the decision complex vs. not complex (using CCI or PCCL? Or some other loose set of criteria?)

>Thank you for giving us the opportunity to clarify how the outcome “complexity” was collected. Indeed, patient complexity is inherently difficult to define. Different models have been proposed to better characterize complex patients (e.g., Shippee ND et al., Cumulative complexity: a functional, patient-centered model of patient complexity can improve research and practice, J Clin Epidemiol 2012 Vol. 65 Issue 10 Pages 1041-51; Schaink A et al., A scoping review and thematic classification of patient complexity: offering a unifying framework, J Comorbidity 2012, 1-9). In addition, complexity may change over time, e.g. between admission and discharge. It is important also to highlight that there is no better gold standard available that the physician’s judgement to assess patient complexity. F/span>or this study, we pragmatically used the treating physician’s judgement to identify complex patients, which we defined as those “using more resources, time and/or effort while hospitalized”. Using this pragmatic approach, it was possible to categorize a large number of patients and use this classification to identify the factors contributing to patient complexity. This definition was already used in previous research (Grant RW et al., Defining patient complexity from the primary care physician's perspective: a cohort study, Ann Intern Med 2011;155:797-804 PubMed ; Martin C, Sturmberg J. Complex adaptive chronic care. J Eval Clin Pract 2008;15:571–7.) . Nonetheless, we have now specified in the methods section of the manuscript (Study outcome and predictor variables) that “The resident (or supervising consultant) was asked by a trained study nurse to assess at time of discharge the level of complexity of the entire hospital stay of her/his patient without providing any specific scoring system (complex or not-complex).”

52 parameters from the medical record. Unless their aggregation and calculation is automated, this will pose burden on hospital administration staff or clinicians before deciding on severity level. Also, for extrapolation/adaptation in other health systems, the final 11 indicators may vary from one setting to another in different health systems and therefore a recommendation to conduct PCA validations in other countries could start with the full set and see if the final indicators of complexity are the same as in the Swiss system; I think this would warrant a mention in the discussion.

>The general rule of score validation is indeed to test it broadly in different population to insure its generalizability, as we now clarified in the limitation section of the Discussion, as follow: “Therefore, in other health systems the final indicators may vary, which might be considered when validating the PCA score.”. It is however not a common practice to test again the entire list of parameter used in the

original (current) study, and therefore it would not be accurate to specify this step (Harrell FE, Jr., Lee KL, Mark DB (1996) Multivariable prognostic models: issuesn developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15: 361-387).

Page 9, lines 5-9: the assumptions of normality or last observation carried forward or imputation for missing at random may be customary in development and validation studies. However, given the extensive list of original candidate variables, perhaps the newly developed patient complexity score could have used only variables that were complete in routine care (not just 75% complete), as indicated in the methods/protocol.

>Thank you for your remark. For this non-interventional study, some variables were missing to a relevant degree. Nevertheless, we assumed, using these variables would countribute to a better understanding of patient complexity if included, which is hence represented in the final PCA score. We have acknowledged this issue in the manuscript as follows: "[…] imputation of missing data may have changed the outcome of the study. However, potential predictors with more than 25% missing data were excluded."

Page 10, lines 30-34: Add whether the clinical judgment (or some other index) was the method used to determine that 563 patients were complex.

>We have specified the reporting of patient complexity as suggested by you: "In the derivation cohort, 447 patients (31.8%) were clinically judged as complex, and 116 (24.1%) patients in the validation cohort."

Page 12, Table 2: Effectively, the 11 variables included in the final score development yielded moderate to very good predictive ability, and the conclusion of the PCA superiority against CCI and PCCL holds and is backed up by the findings. Despite the authors' attempts to address the inverse relationship between age and complexity, and the counterintuitive relationship between elective (rather than emergency/unplanned) admissions and complexity are still puzzling. It is still unclear to me (and likely other readers) whether this was a coding error or a special circumstance of the participating tertiary hospital. Perhaps it is worth expanding the comment in the discussion on page 15.

>A counterintuitive relationship doesn't mean it's wrong. Our findings show interestingly that what we could consider as relationship might not always be true. For example, younger patients with severe diseases may well be seen as more complex to care by the physicians since the pressure to have them back to a full recovery might be more present (from the patient itself, the family, and/or the physician point of view) than in patient who are not willing all the investigations and care because of their advanced age. We do not expect more coding errors than in any other study using administrative and clinical data, especially what concern the age of the patients. We have addressed your remark more precisely with adding the following paragraph: "The inverse relationship between age and complexity, and the relationship between elective admissions and complexity may therefore represent structural incentives to hospitalize complex younger patients which overburden outpatient care."

I'd suggest the next step in this research program to confim the hypothesis that complex patients are higher-level users of other services or have worst outcomes than non-complex patients is to check the eligible patient outcomes (hospital reutilisation, use of social services, admission to nursing home, other complication rates and post-discharge mortality) after 6-12 months in a future study to compare with the prediction of low-high complexity cut-off points.

> Thank you for this suggestion as a future research step. We appreciated your critical reading of our

manuscript, and we truly believe that your final minor suggestions help further improved the manuscript.

Reviewer: 3
Dr. Terence Tang, Trillium Health Partners
Comments to the Author:
I have reviewed an earlier version of this submission. I believe that the authors have adequately addressed my comments and this article is fit for publication. I will be interested in seeing a validation of this score outside of the setting of this current study. This is a score that identify complex hospital patients at discharge, and I see the major utility, as the authors mention, as to monitor complexity in hospital and performing complexity related study of hospitalized patients retrospectively.

> Thank you for acknowledging the quality of the revision thanks to your comments. As mentioned in the discussion, further validation of the score would be useful.