Reviewer 1:

I have reviewed this manuscript from statistical methodology point of view and identified several major points as listed below.

1. Instead of defining outcomes as A and B, authors may labelled them as primary and secondary outcome as it is confusing.

Author response: We agree and have updated the text accordingly.

1. There is a typo in writing number of samples for training set in the statistical analysis section. It is written as ’18.013’.  
     
   Author response: Corrected.
2. Recent theoretical developments (with software implementation) allow us to calculate minimum required sample size for model development and validation. For example, [Riley et al (2020)](https://www.bmj.com/content/bmj/368/bmj.m441.full.pdf) demonstrated how one should calculate the minimum required sample size for developing a clinical prediction model.  
     
   TP: Udfra 2021 artiklen nedenfor synes det at være sunde men ikke helt tidssvarende metoder til at finde ud af minimum sample size. Desuden er der jo ikke en fast grænse, og de foreslår at træne på “det hele” og ikke opdele det, mens Cross Validation ikke er nævnt!!!  
   Så jeg tror, at vi skal anderkende, at der findes metoder (og måske citere Riley, hvis det er det, som han ønsker?), og evt. putte en kommentar om sample size ind i teksten. Men jeg hører meget gerne jer, ikke mindst om der er en (nem) software implementering, Christian?  
     
   Forslag til svar: ?
3. Similarly, [Riley et al (2021)](https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.9025) showed how minimum sample size could be calculated for external validation of a clinical prediction model with a binary outcome. This will provide an idea about how much sample size we need when developing and validating a prediction model.  
     
   Potentielt forslag til svar, hvis vi ikke adopterer ideen (TP): Thank you for drawing attention to the above paper. In principle there is no lower limit to the sample size, only it may (largely) impact the subsequent model performance. But since we do not train the model on all data, but retain an “unconsidered” portion (and optimise this usage through temporal cross validation) for testing model performance - which is the part we report - we believe that our study complies with best practice of data modelling.  
   However, after evaluating the recommendations given in Riley et al. the inclusion of xxxx procedures with xx events confirm that we do indeed have enough data and live up to the standards given. We have included the above considerations and the provided results of the minimum sample size for validation in the methods section.
4. [All machine-learning methods are data hungry](https://bmcmedresmethodol.biomedcentral.com/track/pdf/10.1186/1471-2288-14-137.pdf) and should have been used when appropriate sample size is available so you need even more sample as these models test all possible non-linear relationships.  
     
   TP: The linked paper talks about stability, so possibly the worry is, that we “cherry picked” the model to show a good performance. I think that we should argue/show some stability numbers, and that since we also do the LR and see similar ROC shape simply below (also for the parsimonious models), we feel comfortable that this is not the case.  
     
   Author response : As this is closely related with 4. Please see our response above covering both points.
5. There is a strong need that machine-learning techniques adhere to established methodological standards already defined in prediction model research. Fair and neutral evaluations and comparisons against these existing prediction model approaches must be done. Authors have predominately reported model discrimination measures (such as AUROC, precision, and sensitivity etc.) in this paper. However, it is widely recommended to report both model discrimination and calibration statistics (such as Brier score, calibration-in-the-large, calibration slope etc.) as model performance measures. Machine learning models are not exempted from calibration measures. See [Riley et al](https://www.bmj.com/content/bmj/353/bmj.i3140.full.pdf) paper published in BMJ in 2016.   
     
   TP: Jeg kiggede på Riley 2016 og fandt:  
   “Calibration examines the agreement between predicted and observed risks, and can be quantified by measures such as the calibration slope and the expected/observed (E/O) statistic (box 1)”  
   Jeg tror simpelthen de gerne vil se, at fordelingen af scores er den samme i training/validation og test data. Dette mener jeg er et hårdt krav, idet “Weak overtraining is good” [Giles Louppes], men lad os da bare gøre det.  
     
   Forslag til svar: [TP: Mere af det samme ovenfor?] Author response: In our opinion the arguments by Riley in the mentioned paper can be debated as some authors have argued the “weak overtraining is good” Ref: Gilles Louppes xxxx However, we have added the distribution of training/validation and test data scores for interpretation in the appendix.
6. Prediction models based on machine learning are much more reliant on computers for implementation of the underlying model, which is often labelled as a black box. This inherent complexity behind machine learning obfuscates interpretation or face-value assessment of the prediction model algorithm. Therefore, the machine learning algorithms should be available to others for independent validation or to implement it into the clinical workflow. Therefore, authors should add analytic codes as a supplementary file.  
     
   TP: Fint - eller link til GitHub vil jeg mene er mere professionelt. Vi kan også svare med en kommentar omkring SHAP-værdier og at vi af samme grund også laver en LR-model.  
     
   Forslag til svar (TP):  Author response: We agree with the reviewer that the underlying model is often a “black box”. However, this is why we present the detailed SHAP-Analysis evaluating the contribution of the individual variables on the risk-score. Furthermore, as suggested by the reviewer we have included a link to GiTHub in the methods section for supplying the analytical code.
7. Prognostic models are typically evaluated with measures of accuracy that do not address clinical consequences. Decision-analytic techniques allow assessment of clinical outcomes but often require collection of additional information and may be cumbersome to apply to models that yield a continuous result. [Decision curve analysis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2577036/pdf/nihms74373.pdf) identified the range of threshold probabilities in which a model was of value, the magnitude of benefit, and which of several models was optimal. Decision curve analysis is a suitable method for evaluating alternative prognostic strategies that has advantages over other commonly used measures and techniques. Therefore, if useful, the decision curve analysis should be explored.  
     
   TP: Clinical consequences?!? That the weaker patients gets are more intense care…?  
   Jeg tror, at referee mener, at man bør finde frem til det sted på sin kurve, hvor et outcome (e.g. sparede antal sygedage) er bedst muligt. Men i dette tilfælde ligger de 20% fast, så det er ikke “useful”.  
     
   Forslag til svar (TP): We agree with the reviewer that clinical relevance is of major importance in prognostic studies. However, as described in our paper, , we consider the clinical capacity for “augmented treatment” to be about 20% [page 14, line 39-45], why decision curve analysis does not really apply. We also comment on the robustness of the conclusions for other capacity values using thresholds of 25 and 30% [Supplemental material Table 3 and 4], in line with your suggestions.
8. Authors mentioned that cases of missing values in the logistic regression model were handled by imputing missing values with the median of present values. It is not the recommended approach in the literature as multiple imputation should have been used to handle missing data in logistic regression.  
     
   TP: Jeg tror, at det der menes med “multiple imputations” er at følge PDF’en (i stedet for at bruge median) og gøre det mange gange for at teste, hvor stor variation i resultatet man får. Det kan vi godt gøre, og jeg vil faktisk give referee ret i, at dette er rimeligt.  
   Jeg fandt også følgende, som synes at underbygge ideen (men Robin’s rule?!?):  
     
   https://www.bmj.com/content/338/bmj.b2393#:~:text=Multiple%20imputation%20is%20a%20general,obtained%20from%20each%20of%20them.  
   The first stage is to create multiple copies of the dataset, with the missing values replaced by imputed values. These are sampled from their predictive distribution based on the observed data—thus multiple imputation is based on a bayesian approach. The imputation procedure must fully account for all uncertainty in predicting the missing values by injecting appropriate variability into the multiple imputed values; we can never know the true values of the missing data.  
   The second stage is to use standard statistical methods to fit the model of interest to each of the imputed datasets. Estimated associations in each of the imputed datasets will differ because of the variation introduced in the imputation of the missing values, and they are only useful when averaged together to give overall estimated associations. Standard errors are calculated using Rubin’s rules,[**16**](https://www.bmj.com/content/338/bmj.b2393#ref-16) which take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values. Valid inferences are obtained because we are averaging over the distribution of the missing data given the observed data.

Forslag til svar (TP): Author response: Thank you for pointing this out. We have now included the multiple imputations method as suggested,, but with little variation - probably due to most main variables being without missing data.

TODO: Adapt Multiple Imputations!

1. What do you mean by the sentence that all variables were then normalised? Pleas elaborate.  
     
   Forslag til svar (TP) Author response: The normalisation procedure is described in [page X, line Y] and has been slightly updated for clarity. Essentially, it is a Robust Scalar normalisation, moving all averages to zero, and scaling all Standard Deviations to be unity. This improves the performance of the Logistic Regression (the BDT is invariant), which points to some variables getting an improved distribution for the LR to extract information.
2. I really do not understand the logic of constructing a model based only on age (Age), and a machine-learning model based on all variables except for age to explore the influence of increasing age. You main multivariable model with age variable can provide the required information. I really hate to fit all possible scenarios rather prefer to fit clinically hypotheses driven models.  
     
    Author response: We agree with the reviewer and have moved these two models to the supplementary material table 2 and removed the curves from figure 1 and supplemental material 4. However, we believe that these two models illustrate that age by itself is a poor risk criterion in an easily understandable way. Consequently, we have kept our comments on the results of these models in the results and discussion. We hope this is acceptable.   
     
   TODO: Quickly check if any “young” patients were given high risk scores by the All-but-Age model.
3. Table 1 should have been organised according to the training and test data.  
     
   TP: Det skulle være til at fixe :-)  
     
   Author repsonse: Done.

Reviewer 2:

There are some revisions that need to be addressed as follows:  
  
TP: Helt generelt, så tror jeg, at Reviewer 2 har skimmet vores paper og så givet det et sæt standard statistiske spørgsmål. Det er måske hans “business model”... men vi kan heldigvis svare nemt på det meste!

1. Dataset information provided by the authors is not enough. The authors need to mention every detail of the dataset used in this study.  
     
   Author response: We have provided information on the sources and method for datacollection on page 7 L19-30. Details on the individual variables are found in table 1 and Supplemental Material table 1. We do not believe that a more detailed description of the Danish National Database on Reimbursed Prescriptions (DNDRP) is suitable for our paper and it can be found in the mentioned reference. We have also elaborated on which variables were included in machine-learning model on page 9 as suggested by reviewer 1. We hope this is acceptable.  
   Forslag til svar:
2. Are there any Null or NaN values present in the dataset? If yes, how do authors deal with that values?  
     
   Forslag til svar (TP)Author response: There are indeed though they are rare, and we have made imputations for these originally using the median value. Their impact happens to be small, which was found by applying Multiple Imputations (i.e. from the variable PDFs) and seeing the impact of repeating this randomised process, as suggested by reviewer 1.  
     
   TODO: Check if we explicitly write the fraction of NaN values, and specify their nature/type/variable.
3. What kind of pre-processing techniques are used for this dataset before using machine learning models?  
     
   Author response: The variables were all normalised using a Robust Scalar method, yielding a mean of zero and a Standard Deviation of unity, which improves the Logistic Regression, pointing to some variables getting an improved distribution for the LR to extract information.
4. What kind of dataset was used in this study? How many features were considered?  
     
   TP: Did he read the text? Are we not specific enough?  
     
   Author response: As mentioned in the methods section on page 10 L13 and table 1 we used 33 prospectively recorded variables of which 7 were continuous. We have rephrased the methods section to clarify this.
5. The experimental testbed is required for the readers and other researchers? What is the experimental testbed for this study?  
     
   Author response: We are uncertain about what is meant by “the experimental testbed”. If the question is with regards to the objective of the study this was to “develop an improved machine-learning model for preoperative prediction of “medical” complications resulting in prolonged length of stay or readmissions”, as written on page 6-7 L56-10
6. There are some statistical parameters used in this study to evaluate the performance of the model. The authors need to mention mathematical equations for this study.  
     
     
     
   Forslag til svar: ? Author response: We have supplied the statistical code as suggested by reviewer 1, we hope this is satisfactory.

It would be more helpful for the readers if the authors add one graphical workflow of the methodology used for this study.

Author response: We agree and have added such a workflow as Supplementary material 1 as it would otherwise have taken the total number of tables and illustrations in the finished manuscript up to 6. We hope that this is acceptable.