**Replies to comments for LDSSA report 1**

**Overall impression of comments:**
The feedback gives the impression that there is a list of keywords which need to be mentioned/defined in certain sections. This is not pushing us to understand concepts and apply them, but rather to memorise and regurgitate, even when it is not necessarily contributing to the text. In other areas, the comments do not necessarily line up with the way I interpreted the instructions, even though I agree with what the comments were (essentially, this means extra work which could have been avoided).

**1.2**

1. **"For this section to pass, make sure to translate this into technical requirements".** I put this in as a subjective assessment (there was no technical requirement other than reducing the wrongful discharges). The requirement was that at least 50% true positive rate (in addition to discrimination metrics etc). My point here was that, because of the nature of the dataset, we might be able to achieve a 50% true positive rate, but that it may mean that the number of true cases identified would be too small to make a material difference for the hospital. This was not specified in technical requirements, so Im not sure what you are asking for.: **"There are two business requirements: 1. Minimizing wrongful discharges 2. While keeping at least 50% of the patients with readmitted decision actually be sick You have done a good job with translating the second one, and the first one is missing. The first one can be translated into an actual metrics, e.g. recall, precision, f1 score, etc. This is important so that the client can understand why you decide to optimize for a specific metrics in your ML model. and how it's backed by actual business requirements".** I understand - I have the other metrics mentioned in other parts of the report and I will include it here as well (i.e. recall/f1 to make a meaningful reduction in readmissions and balance with precision).
    a. Text has been altered to say: "A clearer target was outlined, specifically that wrongful discharges (leading to readmission within 30 days) should be minimized. This can be achieved with a high recall model, allowing for detection of most, if not all of the sick patients. However, a further requirement that at least "50% of the patients readmitted should actually be sick" has been added. This is particularly challenging as the dataset provided is quite unbalanced and machine learning models may not allow us to achieve the target of 50% precision while making a meaningful reduction in wrongful discharges. To achieve a balance between these two requirements, F1 and ROC-AUC scores will be the metrics used for model optimization"
2. **"there's a missing important metrics"**
    a. Text now includes recall and F1 scores (see above)

**2.1**

1. **"Can you please provide a reference here?"**
    a. Reference has been included (with hyperlink): (AHRQ statistical brief #278, 2021)

**2.2**

1. **"Instead of referring the reader to the annexes, include any relevant visualizations mentioned next to the text. This also helps the reader find the**

**info more easily if they don't have the time to read the entire section. Think of your client as a busy person who wants to find what they need as quickly as possible".** Hi, the guidelines for the structure of this section (on the google sheets document) says "refer to the annexes whenever you need to support your claim" so I put the figures in the annex and, as you can see from the annexes, there are several pages of them. Putting those figures in here would push me way over the limit. Including even the simplest figure which combines the analysis requested for race, gender, age, admission source, insurance and medical specialty, which is half a page (this section is 1 page). I understand and agree with your point, but it contradicts the guidelines and structure we were given. I don't see a reasonable way to address this, unless I need to rewrite this section entirely - which from your comment seems not to be the case (and which I don't want to do). **My point is that relevant plots, such as those which are showing discrimination within a group and deliver a story by themselves, can enrich this section a lot, specially for a non-technical reader. As also mentioned in the report structure, annexes are for things that are supplementary and not essential to the reader -- my advice is to show the ones which are relevant, and at the same time, be kind to the user and avoid distracting them too many times so they are forced to go back and forth (again, this is tailored for the business reader). So in this case, my suggestion was for you to choose 1 or 2 plots that are relevant and sustain your claims. You have very well presented plots on the annexes, of those you can use only the one(s) that you think deliver a message that is important to the client in a visual manner. The pages on the report structure are recommended, using up to 1.5 pages is ok. So you still have space to include 1 or 2 relevant plots (of course not all the plots you have on the annexes are necessary here!!). Does this make sense?** Ok, so in this case I will transfer the plots which use the simplest analysis and/or show an important result into this section and leave the more complicated subgroup panels in the annex. Thank you for the clarification.
    a. The section has been rewritten to allow for inclusion of a few specific plots without excessively exceeding the page limit (this was mainly cutting text and summarizing results where possible). Included plots in this section are total demographic readmission rates, readmission rates for race/age in transferred patients and selected medical specialties for age/gender/race. Also included are plots for specialties discriminating based on sex. The annexes have been maintained in the same state as previous submission for completeness. Current size of the section is now 1.5 pages - with less than 1 page text and 3 small figures. A second round of comments for this part would be nice (major vs minor comments?).
2. **"Please briefly explain how you measured/computed discrimination."**
    a. The first paragraph of the section has been reworked to include a description of the methodology applied. Additionally, the plots placed into the section have additional lines which illustrate the method applied. Text now reads: "The objectives of this investigation were to i) establish whether or not there was any discrimination based on patient gender, race, age or insurance status resulting in wrongful discharge, and to ii) determine if any admission sources or services were discriminating based on these 4 sensitive features. The technical specifications provided indicate discrepancies of up to 5% between subgroups, and up to 10% between medical specialties are

acceptable for the API predictions. For this analysis however, a more stringent 5% difference was used as an indicator for areas requiring further investigation. Readmission rates of the subgroups were compared with expected readmission rates (the rate if each group was equally represented). Additionally, the difference between the lowest and highest values was extracted. In the event this difference was beyond the specified threshold of 5%, it was considered that discrimination may be occuring"

**3.1**

1. **"Overall, it's quite good :-) The only thing lacking is to sustain your statements using actual results (complementing qualitative with a bit of quantitative)"**
    a. Values for precision/recall scores have been included where necessary. Additionally, when discussing discrimination scores, results from the precision test comparison have been included. Main body of text has had some minor alterations.

**3.2**

1. **"Good, just need to describe how you bulid your target feature and address the comments below"**
    a. Description of building the target is now specified, text reads: "The target for prediction was a Boolean value built based on the 'readmitted' value, with values of 'Yes' being replaced by True, and 'No' by False."
2. **"would move this discussion to 3.3. also, in the end it's not clear why you chose oversampling against the others"**
    a. This text has been moved to 3.4 ("Alternatives considered" and an explanation of why random oversampling was selected has been added. Text now reads: "1. Similar results were achieved with random undersampling, and with alternative classifiers such as BalancedRandomForest and RUSBoost which implement undersampling methodology. These were discarded in favor of oversampling in order to ensure that information was not lost from the majority class."
3. **"How did you obtain these? And what about the hyperparameters?"** I'm not sure I understand what you are asking me for in the first part here: I wrote that I used a grid search to select optimal parameters for the random forest. I know it was an error to call it parameters when I meant hyperparameters (and will correct it), but I'm not sure what you think is missing. **Hi Yash. You're right that you mention how you achieved this. Nonetheless, it is expected that you provide the information that will allow someone else to replicate your model. So it is lacking: - the values you tested for each hyperparameter - the train/val split or an explicit statement that you choose the minimum train error. - mention what is done regarding the remaining hyperparameters.** Thanks for the clarification, I will include the additional information.
    a. Bullet point for description of model has been updated and a table showing the parameter space has been included (see text). Considering that this is for a technical, but maybe not python audience, a link has been included to sklearn documentation of random forest so that exact definitions can be found (no idea what this is called in R for example). Text now reads: "A Random Forest Classifier, selected to minimize risk of overfitting, was fitted after a grid

search was used to select hyperparameters. The options tested in the grid search are listed in the table to the right. Optimal hyperparameters are in the final column. Unlisted parameters were used at their default values to economize on the time required to train the models (in summary, min_samples_leaf:1, min_weight_fraction_leaf:1, max_leaf_nodes=None, min_impurity_decrease=0.0, oob_score=False, verbose=0, ccp_alpha=0, max_samples=None,  for definitions of hyperparamters, see documentation here)."

4. **"Describe the data split you considered, as well as the threshold."**
    a. Test/Train split details have been added at point 2 of this list. Text now reads: "Data was sorted by admission_id (assumed to be sequential) and a test/train split, keeping 10% of entries for the test set was done. Random oversampling of the true class was used to generate a balanced training dataset."
    b. It has been specified that the default threshold (0.5) was maintained - justification is included in section 3.3. Text now reads: "...The threshold used to decide predictions was maintained at the default value (0.5) as calculation of geometric means yielded an optimal threshold that was extremely close to this value (0.49, see section 3.3 for further details)..."

**3.3.**

1. **"could you provide practical meaning? as you do for recall"**
    a. Text now reads: "...the model has a precision rate of just under 20% (20% of patients predicted for readmission are actually readmitted)..."
2. **"practical meaning?"**
    a. Text now reads: "...model also has an accuracy (the proportion of correct predictions) of slightly under 70%,..."
3. **"Refer fig.3 here instead of next sentence"**
    a. The reference has been moved as requested
4. **"Consolidate with quantitative results"**
    a. Results for a "discrimination analysis" using precision scores has been added. The text now reads: "Comparison of precision scores generated for separate subgroups reveals that this model does not result in a large discrepancy between different genders (Male: 19%, Female: 18%), insurance status (Insured: 18%, Uninsured: 16%), admission sources (all values ranging between 18-21%) or races (White: 18%, Black: 20%, other races had higher scores but were present in low numbers), but it does still disproportionately affect a few specialities (The maximum difference was of 28%). Most concerning is the impact on 20-30 year olds who, as expected based on exploratory data analysis, had significantly higher predicted readmission rates which we were unable to mitigate without significantly compromising performance in other areas (A summary of methodology and results is available in annex 5.3)"
    b. An additional summary of this analysis has been included in annex 5.3 with a brief description of methodology and full results.

**3.5**

1. **"- You should mention risks, or future problems, that could eventually compromise the performance of your model."**

a. A brief explanation of data drift and concept drift potential impact on the model has now been added as a future problem/risk. Additional text now reads: “Changes in the data over time (data drift), such as alterations in the demographics of patients coming to the hospital, or changes in the quality of staff/care (due to new methods, new staff, or new facilities for example) provided may result in impaired model performance. Additionally, as the definition of wrongful discharge is refined (concept drift), potentially accounting for different specialties or links between the medical reasons for initial admissions and readmissions, it is likely that the model will no longer be as useful.” A few other minor changes were made to ensure that the text remains as close to recommended length as possible.

**4.1**

1. **“- There's no information on the utility of the endpoints provided. If they were named something else, their function would be very unclear.”**
    a. A brief explanation of what the /predict endpoint does has been included. Text now reads: “/The predict endpoint allows for the user to input data regarding the admission in question (including patient ID, diagnosis, discharge disposition - see the schematic overleaf for structure of the expected payload), and generate a prediction for whether that patient will be readmitted to the hospital within the next 30 days. This endpoint goes through the following steps:...”
    b. The description of the update endpoint has been expanded. Text now reads: “/update: This endpoint allows for true labels (the actual readmission status) to be added to the database, alongside previously made predictions and details about the admission. This can be used for analysis and potential retraining of the model if required. This endpoint also checks inputs for validity, to ensure true labels correspond to an admission for which a prediction has been generated, and stored in the database.”
2. **“- Field validation of the data received is briefly mention and could be expanded”**
    a. Bullet point b) of the /predict endpoint details has been altered. Text now reads: “Call validation functions that inspect the resulting dictionary to ensure:
        i. The presence of all the required fields (even if they are null or empty)
        ii. The presence of numerical admission and patient IDs
        iii. That datatypes of columns are, or can be converted to, the expected datatype for use in the model (particularly important for numerical fields such as ‘num_medication’, ‘num_procedures’ etc)
        iv. Checking that strictly defined categorical/string fields (‘blood type’, ‘change’, ‘insulin’, ‘diuretics’) contain an expected value.

    Failure to pass any of these tests results in the return of a 422 error, and a message explaining the source.”

**4.2**

1. **“- It fails to mention risk associated with the deployment of app itself, such as resource limitations”**
    a. Limitations of using heroku postgresql were previously mentioned. This has been reworked and made much more explicit. The text now contains a list of

key limitations identified with using the Heroku Platform (price, power, portability) and now reads: “The API has been deployed using Heroku, which may bring several limitations in the future:

i. Running a high traffic API on Heroku can become expensive very quickly, as well as potentially impacting performance
ii. As Heroku provides the majority of infrastructure, there is the possibility of becoming “locked in”, as transitioning to alternative platforms becomes difficult and costly
iii. The heroku database currently being used has a limit of 10,000 entries - so data will periodically need to be exported and the database reset. Failure to do so may result in data being lost or the app crashing.

2. **“- Security is also not addressed, which is important in the context of the client and the type of information that is provided”**
    a. An additional point has been added discussing security. This reads: “Lastly, considering the sensitive nature of the data being stored, the deployed API has a major security flaw: weak access control (anybody with the URL can make requests in this case). This leaves the API vulnerable to attacks such as SQL Injection, resulting in access to sensitive data and Distributed Denial of Service (rendering the endpoints unusable with excess traffic). The integration of this API with the hospital IT systems could therefore introduce new security weaknesses into these systems.”