LISBON DATA SCIENCE ACADEMY

Reducing wrongful discharge

Predicting wrongful discharge at the Hazel and Bazel Hospital

Prepared for:
The Hazel and Bazel Hospital

Prepared by:
Yash Pandya
Lisbon Data Science's *Awkward Problem Solutions™*

5 of February 2022

Table Of Contents

1. Client requirements

1.1 Summary

The client, The Hazel and Bazel Hospital (HBH), has had multiple instances of patients being wrongfully medically discharged, resulting in severe health consequences for the patient. There have been allegations made that these wrongful discharges may be targeting specific subsets of the population based on gender, ethnicity, race and insurance status.

HBH management has deemed it necessary to carry out an investigation into these allegations, and implement an additional service into HBH's hospital patient record management system. Specifically, this additional service will be integrated into HBH's internal system, and will use available patient data to predict the likelihood of the patient being prematurely discharged. To aid with the investigation and implementation, HBH has made available ≈ 80,000 partially anonymized patient records dating back to 2012.

As such, the primary objectives of this project are to:

1. Investigate the provided data and determine if there is evidence of discrimination based on gender, ethnicity or age when dismissing patients from care.
2. Determine if the Medical Specialty the patient is under the care of, the admission source, or their insurance status is a potential source of this discrimination
3. To identify patients who are likely to return to the hospital in less than 30 days

1.2 Requirements clarifications

Additional clarifications were requested from Dr Agnes Crumplebottom. In addition to some clarifications regarding input/imputation irregularities, identification of missing data and insurance status, Dr Crumplebottom provided the following key information:

1. A clearer target was outlined, specifically that at least "50% of the patients readmitted should actually be sick" - (i.e. if 100 patients are predicted to be readmitted, at least 50 of them should be sick). This is particularly challenging as the dataset provided is quite unbalanced and machine learning models may not allow us to achieve that target while significantly reducing wrongful discharges. In practice, a combination of precision and ROC-AUC scores will be the metric used to assess performance for this target - allowing for focus on true and false positive rates. The imbalance nature of the dataset suggests that the 50% true positive rate desired may be a challenge to attain. Discrepancies between subgroups should be minimized, although different medical specialties may still have differences as they are linked to conditions requiring frequent hospital visits.
2. Although initial correspondence stated that this data covered the years since 2012, and may have had alterations over time, no time stamp data was available. It is reasonable to assume that admissions are registered sequentially and we will attempt to use this property (the admission id), but it will limit analysis and modeling.
3. The IT team at HBH provided an expected structure of the REST API for integration into their system. This will allow for the implementation of some measures to ensure it remains robust. However, considering the nature of the data, and the fact that a lot of it seems to be manually entered, there will be a requirement for an extended testing phase, and we encourage the implementation of improved record keeping practices to ensure stability and reproducibility in the future.

2. Dataset analysis

2.1 General analysis

An initial analysis of the provided data was carried out using the field descriptions provided by Dr Crumplebottom to examine structure, distribution and quality of the data available. Some key aspects of the data are highlighted here. See annex 5.1 for additional description of the provided dataset, information about data types (information about processed data is in the table), and links to interactive profile reports generated based on initial and processed data.

The data contains unique admission and patient identifiers (although multiple patients appear several times throughout the data - as expected). Patient demographic data (summarized on the right), namely gender, age and race in addition to patient insurance status are included. The data shows slightly more female than male patients were admitted to the hospital during the relevant time frame (Fig. 1A), and that the hospital caters primarily to older patients (Fig. 1B). The race category has been included with redundant descriptors (such as 'white' and 'caucasian' being separate categories), although a rapid processing reveals that the patients are primarily white, with a sizable black minority and very few patients of other races (Fig. 1C). While multiple payer codes are provided in the data set (see annex 5.1 reports for additional information), our primary interest is in the 'SP' code for 'self payers' (uninsured patients), although a large number of patients have an unknown insurance status (Fig. 1D).

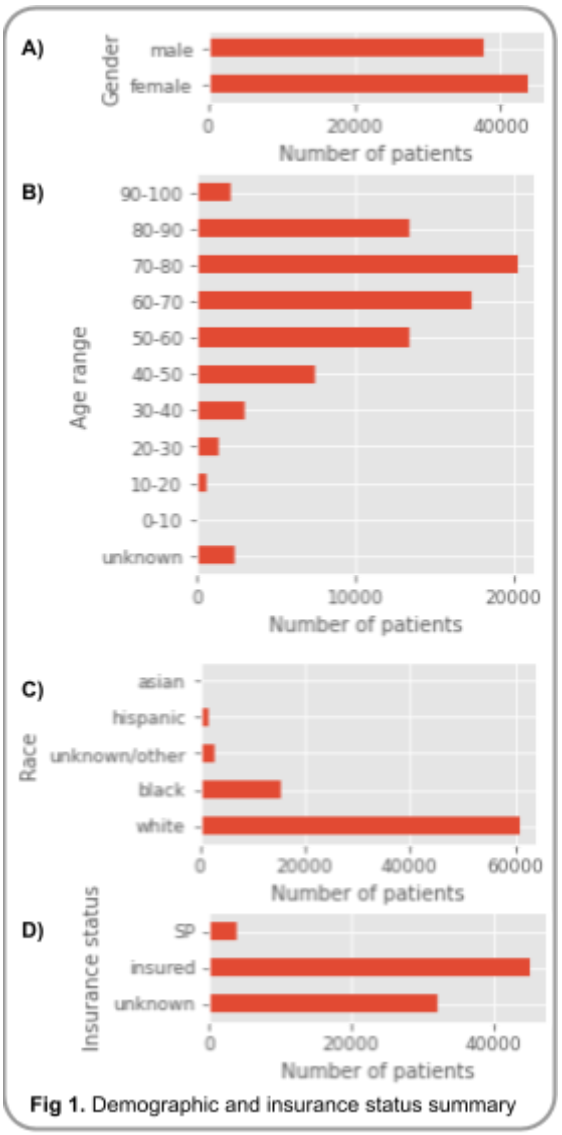


Fig 1. Demographic and insurance status summary

The dataset includes information pertaining to the medical reasons for the patient being admitted to the hospital and some information relating to their medical history and hospital stay (summarized overleaf and in annex 5.1). Unfortunately, weight data is largely missing for most patients. Although the dataset provides for the use of 8 distinct codes for the admission type and 26 for the admission source, these can be condensed and reveal that the majority of patients admitted arrive for emergency care (Fig. 2A and B).

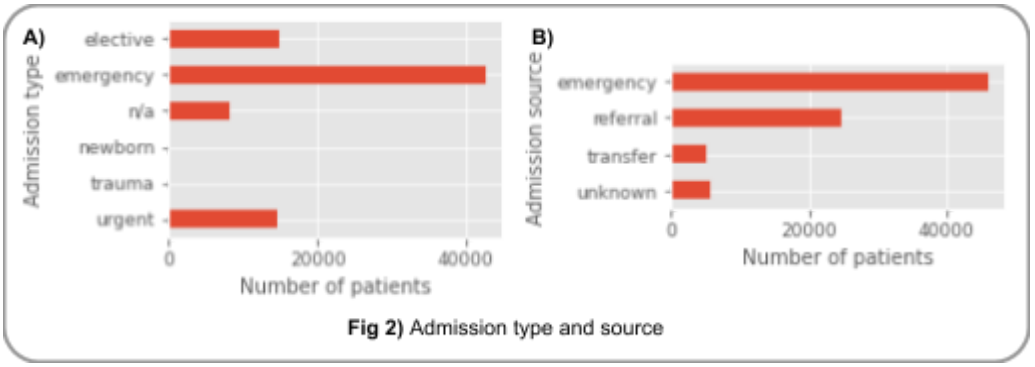


Fig 2) Admission type and source

Clinical information provided (Annex 5.1 Fig 1) includes the specialty they are visiting (the top 5 specialties visited are shown - of note is that for many patients, the specialty is unknown) and the length of their hospital stay, which shows that the most common length of stay is 2-3 days, while very few patients stay for 14 days. In the dataset provided longer stays are not found, probably due to these patients being transferred to long term care facilities. Surprisingly for a hospital specializing in the treatment of diabetes, the hospital specialties with the highest number of admissions are not endocrinology, but include general practice, internal medicine and cardiology. The number of previous hospital visits by the patient (as outpatient, inpatient or emergency visits during the previous year are included, and the strong negative skews show that the vast majority of patients have very few, or no hospital visits per year, but that there are some outliers who are regularly at the hospital.

A large number of patients do not undergo medical procedures during their visit. There also appears to be a multimodal distribution regarding the number of laboratory procedures (i.e. tests) which the patients carry out, with a large number taking very few or no tests, and another peak appearing in patients who take approximately 40 tests. Finally, the amount of medications which were administered during the patients stay indicates a high degree of polypharmacy, which is explained by the fact that a large portion of the patients are elderly, and may require multiple medications on a regular basis, in addition to those being used to treat their current illness. Several clinical features were strongly linked (such as length of stay and number of procedures undergone), suggesting several features are strongly linked (see annex 5.1 Fig 2 for complete correlation data of numerical variables)

The diagnosis, as well as 2 secondary diagnoses are provided using the ICD9 codes. These fields have a high cardinality), reflecting the diversity in illnesses which patients can present with. Interestingly, the most common diagnostic codes as a primary diagnosis appear to be related primarily with circulatory, respiratory and digestive disorders, although there are endocrine disorders (including diabetes) amongst the most common secondary diagnoses.

Additional characteristics include whether or not the patient has a prosthesis or is on diuretics (both of which were a very small portion of patients), patient blood type, whether or not the patient required a blood transfusion, including over 10% of patients and whether or not the patient was prescribed insulin, or new medication for diabetes management, resulting in a change in their current medication. These last features, shown in annex 5.1 Fig 3 reveal that the vast majority of patients at the Hazel and Bazel hospital do have diabetes and that they are frequently prescribed new medications, with an often changing regimen.

The final aspects of the database pertain to the patient's departure from the hospital. 29 discharge codes are available, although several are very rarely used and others are unmapped. The codes can consequently be compressed into a few categories, showing that the vast majority of patients are discharged home, while many others are transferred into another healthcare service or to home care services (Fig. 5A). The final aspect of the dataset is the one which will be key for this project: readmission. This field informs whether a patient has been readmitted within 30 days of their discharge and can act as an indicator of a wrongful discharge. In the United States, in 2018, readmission rates were approximately 14%, placing HBH well below average with approximately 10% (Fig. 5B), although with readmissions costing a large sum, and disproportionately impacting patients with diabetes, it is important to minimize this value as much as possible.

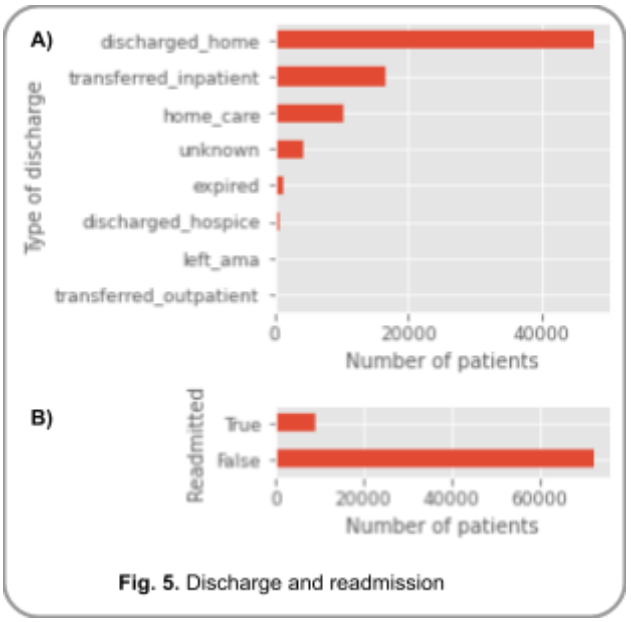


Fig. 5. Discharge and readmission

2.2 Business questions analysis

Considering 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 carrying out any discrimination based on these 4 sensitive features, the readmission data was thoroughly interrogated to look for discrepancies which may be indicative of discrimination. The overall readmission for the provided dataset is approximately 11%. There do not seem to be differences between different groups, although there does appear to be an increase in readmissions among hispanic individuals and in the 20-30 year old age range. In the case of patient insurance status, the large number of patients with missing information makes this conclusion somewhat weaker (5.2 Fig 1).

Next, the two potential sources of discrimination (the admission sources) and medical services the patients attended were both assessed. The admission sources were grouped as above (emergency, referral, transfer and unknown/other) based on the field descriptions provided and total readmission rates for each source were examined. Subsequently records from each admission source were selected and the readmission rates pertaining to individuals based on race, gender, age and insurance status were examined. Of particular interest, it appears that, among patients who are transferred from other healthcare services, hispanic patients have higher readmission rates, and in all groups except those referred to the hospital, 20-30 year olds have much higher readmission rates. Results for this analysis and a description of how data was grouped can be found in nnex 5.2 Fig 2.

To examine the medical specialties, the data was first filtered to exclude specialties which had less than 100 entries - with those with less being grouped into the category of ‘other’. Additionally, records where the medical specialty was unknown (accounting for about half of admissions) were also excluded. Once processed, this left 24 medical specialties with a varying range of readmission rates. This analysis reveals that oncology, nephrology, vascular surgery and rehabilitative medicine most commonly result in readmissions. A similar approach to that for admission sources was then used with records for each medical specialty being selected and readmission rates linked to race, gender, age and insurance status was examined. Due to the reduced representation of races other than black and white in the dataset, it is important to note that many groups were not represented in every specialty. Furthermore, certain specialties are linked to demographic characteristics (such as gender being linked to urology or gynecology). Of note, however, the following specialties have increased readmission rates among certain groups of the population: physical medicine and rehabilitation is 4 times more likely to readmit women than men. While there are some differences between readmission rates in specialties linked to race, it is difficult to draw valid conclusions here as races other than black and white are very poorly represented. In any case, notable results show that black patients are more likely to be readmitted following a stay at physical medicine or at obs/gynecology (data is summarized in annex 5.2 Fig 3-6)

Specialties also continue to show a large increase in readmission rates among young adults such as nephrology, radiology and emergency, while discrepancies which impact older populations appear to be linked to specialties that are linked to age related diseases, such as psychiatry and physical medicine and rehabilitation. There are several specialties which have discrepancies based on patient insurance status, with SP (self paying patients) having higher readmission rates. These specialties include cardiothoracic surgery and urology.

This thresholding approach allows for the identification of raised readmission rates, but it should be noted that, in many cases, although the readmission rates are low, there are significant differences between several subgroups clearly visible on the plotted data. These should also be investigated.

2.3 Conclusions and Recommendations

HBH has a readmission rate which is slightly lower than published average values. Amongst the admission sources for the patients, there is no clear difference in readmission values, although a key role is played here by the impact of a large fraction of patients arriving for emergencies, potentially masking the effects of other admission sources.

A similar problem is presented with respect to race, as only white and black patients are present in sufficient numbers to draw conclusions, with other minorities accounting for a small fraction of patients. In any case, it is clear that HBH needs to further investigate the causes behind increased readmission of minorities in pediatrics, several surgery related specialties and hematology/oncology, where black patients are twice as likely to be readmitted as white patients.

Furthermore, it is apparent from the findings that young adults arriving at the hospital for emergency care, or being transferred from other facilities have a high risk of readmission compared to other age groups, a difference which is reflected when considering all the patients as a whole. It is important to ensure that these differences are not the result of young adults being discharged prematurely.

Lastly, it would be advisable for HBH management to engage in continuous monitoring of this data, and improve the record keeping system such that the large number of “unknown” or missing values are filled. This would allow for a clearer analysis and an easier identification of potential sources of wrongful discharge by allowing for more accurate assessment of patients prior to discharge. The missing status of patient health data is a potential indicator of quality of care, whereas the missing information linked to patient insurance status may influence accounting.

3. Modelling

3.1 Model expected outcomes overview

The initial objective was to design and implement a system which would predict patient readmission within 30 days. The target is to ensure that at least 50% of readmissions are patients who would truly require further care, with variation of no more than 10% between race, gender and age groups, and no more than 5% between different medical specialties.

The dataset provided was unbalanced and consequently it proved to be challenging to reach the accuracy desired. In particular, and as explained in section 3.3, in order to reach the 50% precision requested, the threshold at which patients would be readmitted would be so high as to dramatically reduce the number of true positives which were actually identified. Consequently, it was deemed a necessary tradeoff to reduce precision and ensure that at least half of patients who would be wrongful discharges would be correctly identified. However, this does also mean a significant amount of patients would be incorrectly identified as wrongful discharges.

It is expected that the model will not yield discrepancies based on gender, admission sources or patient insurance status. With respect to race, our analysis has revealed that hispanic patients are frequently discharged early, but are also a very small minority of the patients at HBH. The current model may have moderate discrimination against this group, and will need to be improved if this patient group increases in proportion. Additionally, the data reflects a clear bias for young adults being prematurely discharged, and our system also predicts that they are more likely to be. Lastly, certain specialties are more likely to result in readmission, however, these specialties are linked to frequent patient visits such as rehabilitative medicine and oncology, consequently, this aspect is not likely to be particularly problematic.

3.2 Model specifications

The implemented model consists of a pipeline culminating with a Random Forest Classifier, with several preprocessing steps on the data, summarized as follows:

1. 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 (similar results were achieved with random undersampling, and with alternative classifiers such as BalancedRandomForest and RUSBoost which implement undersampling methodology).
2. Numerical variables ('time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses', 'hemoglobin_level'):
    a. Missing values are imputed with the median of the column
    b. Values are scaled using a robust scaling method to reduce the impact of outliers (minimum-maximum scaling gave similar results)
3. Binary variables ('has_prosthesis', 'blood_transfusion', 'diuretics', 'insulin', 'change', 'diabetesMed'):
    a. Are encoded as True/False boolean values
4. Ordinal categorical variables ('max_glu_serum', 'A1Cresult', 'age', 'weight')
    a. These were converted to categories based on those available, and arranged in their natural order
    b. The order was then used to encode a value based on their position in the order (e.g. for age 0-10, the value was 1, for 10-20, it was 2 etc).

    c. Values which were not included in the ordered categories (such as missing data '?' values) were encoded as -1.
5. A feature selection step was added at this stage to allow for easy iteration and optimization - can be excluded if needed (see annex 5.3 for details of selected features and justification).
6. Categorical variables were each processed individually and one hot encoded prior to fitting the classifier (see annex 5.3 for details):
7. A Random Forest Classifier was fitted after a grid search was used to select parameters. Optimal parameters were a maximum depth of 10, balanced class weights, 'gini' criterion for splits, and no bootstrapping of samples. The primary reasons for selecting this approach was to:
    a. Minimize risk of overfitting
    b. Reduce susceptibility to the effect of outliers (there were several fields were this was extremely necessary)

The model generated was evaluated based on a combination of F1 score, precision scores and ROC-AUC scores and the differences in F1 score between categories within the identified target variables. Specialties with at least 100 patients admitted were selected - although low, some of these had very high readmission rates, perhaps indicating a link to treatment of severe/rare diseases and consequently acting as good indicators for patients who may require repeated visits.

3.3 Analysis of expected outcomes based on training set

Based on the implemented model and supplied training set, the model has a precision rate of just under 20% and a recall rate of slightly over 60%, and correctly identifies more than 60% of the cases of potential wrongful discharge (See annex 5.3 Fig 2). Calculation of geometric means to optimize the threshold for true positive rates did not provide a difference from the default value (0.49 was the indicator received, whereas default values are 0.5), and so the value was left as is (5.3 Fig 2).

The model also has an accuracy of slightly under 70%, and the ROC curve indicates a significant tradeoff between precision and recall. In order to achieve the target of 50% of true predictions being correct, the threshold for a patient being predicted as readmitted would be so high as to mean less than 5% of cases of wrongful discharge would be identified (5.3 Fig 3). This would not solve the problem facing management, consequently, selecting a higher recall rate allows for more patients to be identified. Considering most patients are insured, identifying them for additional care would not be a significant financial burden, whereas wrongful discharge lawsuits can result in heavy financial and reputational damage.

Comparison of scores generated for separate subgroups and specialties reveals that this model does not result in a large discrepancy between different genders, insurance status, admission sources or races, but it does still disproportionately affect a few specialities - often those with a lower number of patients. Most concerning is the impact on 20-30 year olds which we were unable to mitigate without significantly compromising performance in other areas.

3.4 Alternatives considered

Alternative approaches were explored during the initial stages of the project. The options considered and the reasons for them eventually being discarded are briefly explained below:

1. Decision tree classifier with grid search:
   This model was examined to allow for a high degree of interpretability, potentially granting additional insights, however, it had poorer performance when implemented as part of the pipeline used, quite possibly due to a high degree of overfitting to the training set.
2. KNN classifier:

An easily implemented, rapid algorithm which also allows for a high degree of interpretability. While initial testing with a small subset of the data was promising, when transitioning to the full data set, the primary flaw of this algorithm - struggles with larger datasets became very apparent and results suffered.

3. Gradient boosting classifiers:
   The results were poor in the absence of additional approaches to handle unbalanced datasets. It also proved easier to tune the parameters for a random forest classifier, which gave similar results independent of being trained on over/undersampled data or the true dataset, consequently the gradient boosting optionns, including those with random undersampling at each iteration, were discarded.

3.5 Known issues and risks

The implementation applied here has several potential issues and risks, affecting the way features are processed, and the modeling approach selected.

Regarding feature engineering and selection steps, the compression of various categories into smaller groups (i.e, turning many categories into a few), while helping to reduce dimensionality of the data, may have resulted in the loss of resolution and, possibly, a loss of signal linked to specific categorical variables. Furthermore, there are over 10,000 ICD-9 diagnostic codes, and over 700 of these were present in the data. While the approach employed means that there is room for any new diagnostic code to be included into the broader categories, this approach has definitely resulted in a significant loss of signal and resolution. A potential option to explore would be the imputation of ICD9 diagnoses using the codes, and treatment of the text using vectorization approaches and principal component analysis to minimize dimensionality.

It is important to note that the inclusion of sensitive features specified for investigation, gender, race, age, insurance status and admission source were included within the training of the model. While this may allow for groups facing unfair treatment to be identified, ideally, these would not be factors involved in determining patient suitability for discharge.

Using the Random Forest Classifier approach, while minimizing overfitting, does so at the cost of interpretability. The random selection of features at each step makes it challenging to clearly identify features which are heavily contributing. The reduction in variance afforded by averaging multiple predictions also comes at the cost of increased bias, potentially contributing to the results obtained. Lastly, this approach takes significantly longer to train, a potential issue if the dataset expands much more, or if it becomes necessary to retrain frequently.

4. Model Deployment

4.1 Deployment specifications

The developed model has been deployed behind an HTTP server on heroku following the steps outlined below:

1. Model serialization and deserialization
The developed pipeline was serialized using the python pickle library (part of python core) to serialize columns and datatypes, and the joblib library to serialize the pipeline. In the app, the model is deserialized for use using the same libraries (see code for details)

2. HTTP framework selection and database creation
Flask (http://flask.pocoo.org/) is the HTTP micro-framework selected for this deployment primarily because this project requires only two endpoints. Heroku postgresql was the database used for data storage in this implementation via the peewee library. The database allows for storage of the admission ID (must be a unique field), as well as the data received, the prediction returned and the true result (can be null).

3. Endpoints

1. **/predict:** The predict endpoint does the following steps:
    a. Convert the request into a dictionary for further processing
    b. Call validation functions that inspect the provided data, ensuring the presence of admission and patient IDs, and checking the validity of other provided fields. Failure to pass any of these tests results in the return of a 422 error, and an error message explaining the source of the error. A schematic showing the expected structure of a request payload is shown on overleaf.
    c. Return the prediction and store the data. If the admission ID already exists, it will return an error informing that the data was not stored.
2. **/update:** This endpoint allows for true labels to be added to the data for analysis and potential retraining of the model if required. This endpoint also checks inputs for validity.

4. Application structure
The application consists of the pipeline, data types and column files outputted from pickling (explained above), in addition to the following files and components:

1. Heroku.yml and dockerfiles which allow for the implementation on the heroku platform
2. A requirements.txt file which allows for replication on other systems if required
3. 2 python files:
    a. The app.py file which contains all the code pertaining to database setup, model unpickling and implementation of endpoints
    b. A unit test file which contains a variety of functions designed to examine the request input validity - these functions are called at various points in the app.py file
4. A custom_transformer package containing pipeline elements necessary for the model. These elements are all based on elements available from the sci-kit learn library:
    a. A preprocessor file (preprocessor.py) containing functions which clean and treat the data, as well as some minor processing of categorical variables
    b. A combination imputer/scaler for numerical features (custom_impute_scale.py)
    c. A one hot encoder (custom_onehot_encoder.py)

d. An ordinal encoder to process features which have a natural order in their categories such as age, containing the mapping for the relevant features from this dataset (custom_ordinal_encoder.py)
e. A feature selection element (featureselector.py) which takes a list of features to be included.

4.2 Known issues and risks

There are several areas of potential weakness in the current deployment setup, affecting various areas, these are briefly outlined below:

1. The methodology used in the preprocessing stage is designed to make use of unknown and new data, by assigning many categorical values to the 'unknown' group. Consequently, the system is not protected from unusual entries, such as "race":"blue". This could be addressed by implementing improved recording systems - such as the use of drop down lists with well defined categories, or by increasing stringency on data validity. This can be altered as required.
2. The functions to look for data validity have been tested thoroughly, however, it is impossible to explore all possibilities and this means that the two following issues may arise:
    a. Inability to generate a prediction with what would be considered a valid request
    b. Predictions being generated for what should be considered an invalid request
3. The implemented transformers would require significant reworking if new features were to become available, and in their current state are dependent on being run in a very specific order.
4. The heroku database currently being used has a limit of 10,000 entries - so will need to be periodically reset. Failure to do so may result in data being lost or the app crashing. Additionally, the system is currently set up to accept unique admission id values. While it will generate a prediction for duplicate values, it will not store these values, so imputation errors can result in data loss.

Prediction payload:


```
{
  "admission_id": 0,
  "patient_id": 0,
  "race": "string",
  "gender": "string",
  "age": "string",
  "weight": "string",
  "admission_type_code": 0,
  "discharge_disposition_code": 0,
  "admission_source_code": 0,
  "time_in_hospital": 0,
  "payer_code": "string",
  "medical_specialty": "string",
  "has_prosthesis": true,
  "complete_vaccination_status": "string",
  "num_lab_procedures": 0,
  "num_procedures": 0,
  "num_medications": 0,
  "number_outpatient": 0,
  "number_emergency": 0,
  "number_inpatient": 0,
  "diag_1": "string",
  "diag_2": "string",
  "diag_3": "string",
  "number_diagnoses": 0,
  "blood_type": "string",
  "hemoglobin_level": 0,
  "blood_transfusion": true,
  "max_glu_serum": "string",
  "A1Cresult": "string",
  "diuretics": "string",
  "insulin": "string",
  "change": "string",
  "diabetesMed": "string"
}
```


Update payload:


```
{
  "admission_id": 0,
  "readmitted": "string"
}
```

5. Annexes

5.1 Dataset technical analysis

Prior to any manipulation of the datatset, a profile report was generated from the provided file. The software details shown below will allow for a rapid replication.

Reproduction

Analysis started	2022-02-01 17:22:29.832680
Analysis finished	2022-02-01 17:23:27.469446
Duration	57.64 seconds
Software version	pandas-profiling v3.1.0
Download configuration	config.json

An interactive html report is available and can be downloaded at:
https://drive.google.com/file/d/1SSookY72Dl4wggDs2saegGhA5VKpO1le/view?usp=sharing

The file should be saved and opened using any web browser and all files will be available until at least April 30th 2022.

Subsequently, minor cleaning and processing of the data was performed namely:

1. Assignment of data types to columns
    a. In categorical features containing missing or '?', an additional category was created titled 'unknown'. This decision was made considering that failure to collect patient data within certain specialties etc may reflect poor quality of care and assist with prediction of premature discharge.
2. Assignment (and compression) of categories a rapid technical summary of the data was generated (results below for each field provided)
    a. Redundant categories within race were combined
    b. Admission and discharge codes were combined to form larger categories - see model specifications and code for details on exact combination methodology.
    c. Diagnoses were grouped by their primary category according to CDC classifications - see model specifications for these categories. This resulted in over 700 unique diagnoses being compressed into 17 categories.

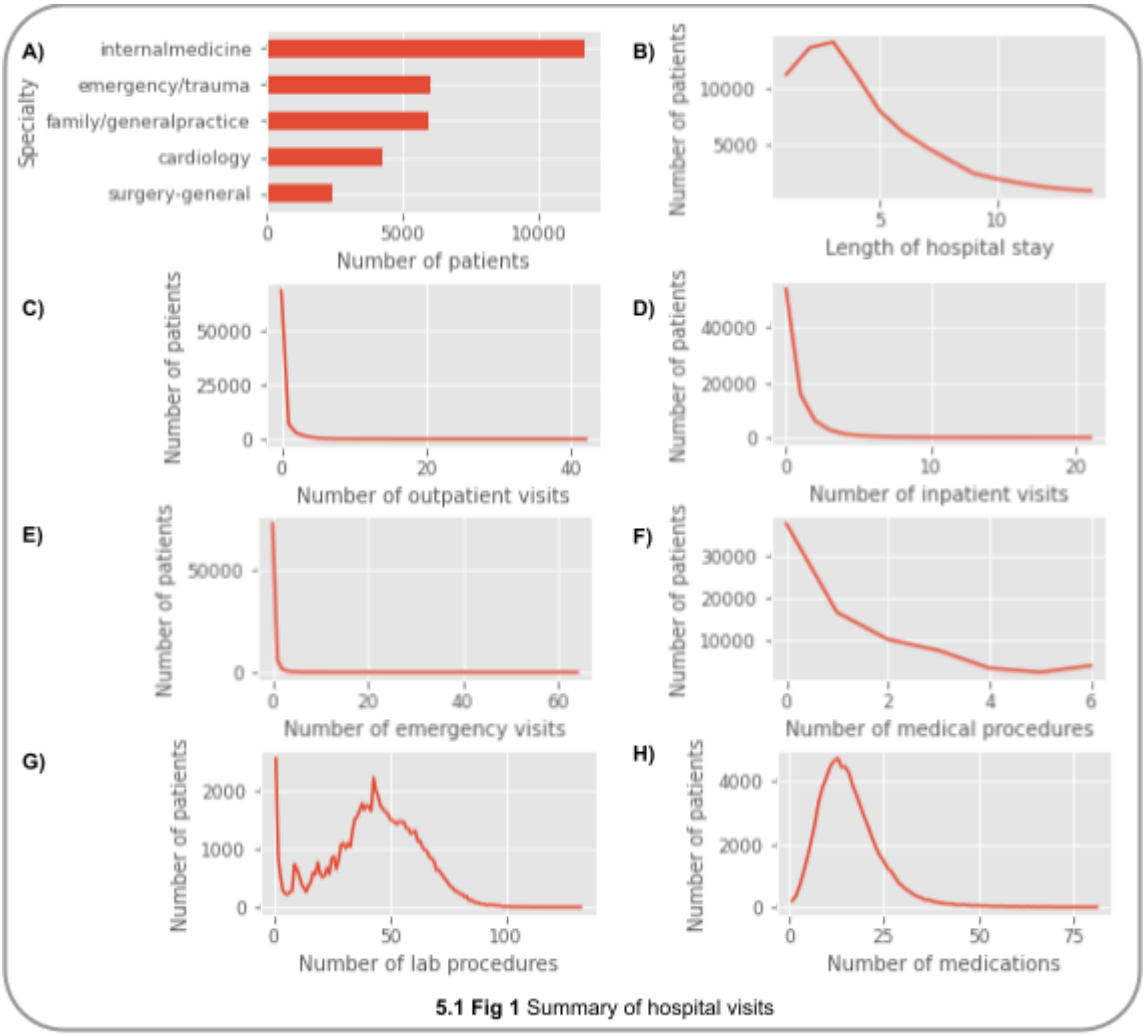
A profile report of the processed data was also generated and is available at:
https://drive.google.com/file/d/1JmzQmHlj3aktfVJbZ57gahxF85aKnImv/view?usp=sharing

The table below shows a brief description of each of the fields provided after preprocessing (N.B, patient_id was used as an identifier and is only included here for completeness)

Name: patient_id, dtype: float64 count 8.141200e+04 mean 1.086395e+08 std 7.732453e+07 min 1.980000e+02 25% 4.683906e+07 50% 9.083437e+07	Name: race, dtype: object count 81412 unique 5 top white freq 60873	Name: gender, dtype: object count 81412 unique 3 top female freq 43719

75% 1.751117e+08 max 3.790052e+08		
Name: age, dtype: object count 81412 unique 11 top 70-80 freq 20261	Name: weight, dtype: object count 81412 unique 10 top unknown freq 78913	Name: admission_type_code, dtype: object count 80250 unique 6 top emergency freq 42562
Name: discharge_disposition_code, dtype: object count 81412 unique 7 top discharged_home freq 47854	Name: admission_source_code, dtype: object count 81412 unique 4 top emergency freq 45942	Name: time_in_hospital, dtype: float64 count 81412.000000 mean 4.395924 std 2.975844 min 1.000000 25% 2.000000 50% 4.000000 75% 6.000000 max 14.000000
Name: payer_code, dtype: object count 81412 unique 3 top insured freq 45131	Name: medical_specialty, dtype: object count 81412 unique 25 top unknown freq 40020	Name: has_prosthesis, dtype: object count 81412 unique 2 top False freq 80550
Name: complete_vaccination_status, dtype: object count 81412 unique 2 top True freq 67434	Name: num_lab_procedures, dtype: float64 count 79919.000000 mean 43.071197 std 19.630405 min 1.000000 25% 32.000000 50% 44.000000 75% 57.000000 max 132.000000	Name: num_procedures, dtype: float64 count 81412.000000 mean 1.341768 std 1.708465 min 0.000000 25% 0.000000 50% 1.000000 75% 2.000000 max 6.000000
Name: num_medications, dtype: float64 count 78734.000000 mean 16.024424 std 8.107235 min 1.000000 25% 10.000000 50% 15.000000 75% 20.000000 max 81.000000	Name: number_outpatient, dtype: float64 count 81412.000000 mean 0.370953 std 1.278538 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 42.000000	Name: number_emergency, dtype: float64 count 81412.000000 mean 0.197588 std 0.881290 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 64.000000
Name: number_inpatient, dtype: float64 count 81412.000000 mean 0.637793	Name: diag_1, dtype: object count 81412 unique 17 top circulatory	Name: diag_2, dtype: object count 81412 unique 17 top circulatory

std 1.265472 min 0.000000 25% 0.000000 50% 0.000000 75% 1.000000 max 21.000000	freq 24193	freq 24666
Name: diag_3, dtype: object count 81412 unique 17 top circulatory freq 23979	Name: number_diagnoses, dtype: float64 count 81412.000000 mean 7.421965 std 1.931480 min 1.000000 25% 6.000000 50% 8.000000 75% 9.000000 max 16.000000	Name: blood_type, dtype: object count 81412 unique 8 top O+ freq 32053
Name: hemoglobin_level, dtype: float64 count 81412.000000 mean 14.192328 std 1.060000 min 10.500000 25% 13.400000 50% 14.100000 75% 15.000000 max 18.600000	Name: blood_transfusion, dtype: object count 81412 unique 2 top False freq 71697	Name: max_glu_serum, dtype: object count 81412 unique 4 top unknown freq 77159
Name: A1Cresult, dtype: object count 81412 unique 4 top unknown freq 67807	Name: diuretics, dtype: object count 81412 unique 2 top False freq 79893	Name: insulin, dtype: object count 81412 unique 2 top True freq 44360
Name: change, dtype: object count 81412 unique 2 top False freq 43772	Name: diabetesMed, dtype: object count 81412 unique 2 top True freq 62718	Name: readmitted, dtype: object count 81412 unique 2 top False freq 72340

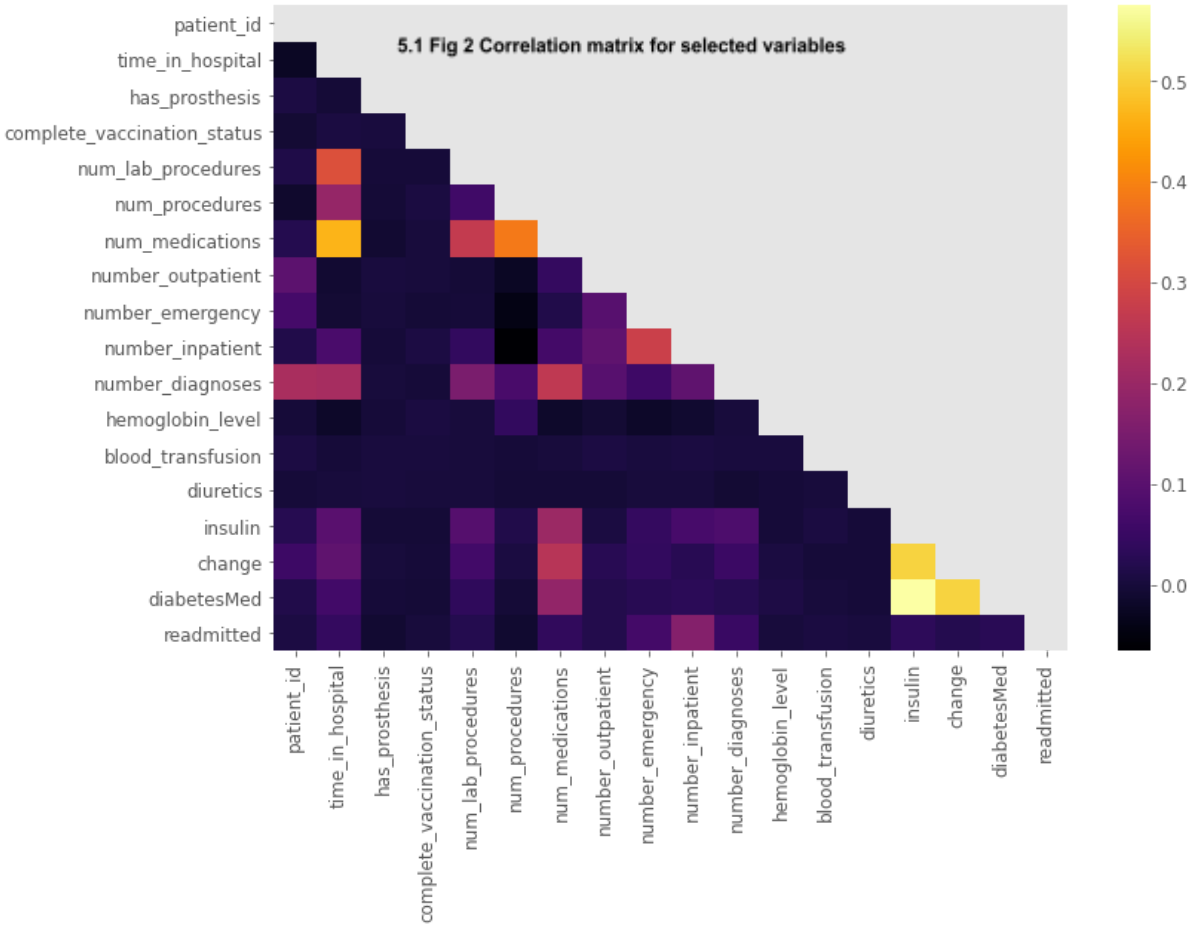


5.1 Fig 1 Summary of hospital visits

Numerical features were briefly explored and, a correlation matrix (overleaf) was produced to determine if any variable were strongly correlated to one another, or if they were duplicating information. Below are the top and bottom 10 correlation pairs and correlation coefficients:

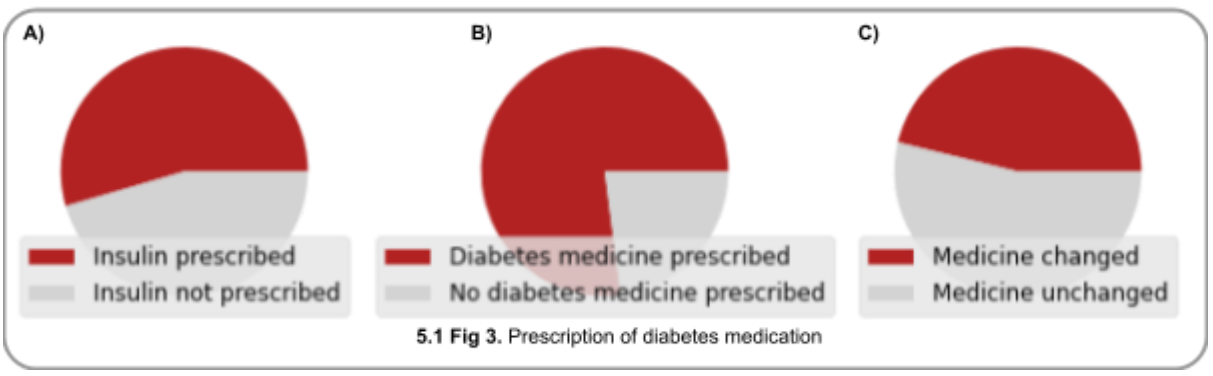
insulin	diabetesMed	0.574090
change	diabetesMed	0.506269
insulin	change	0.505491
time_in_hospital	num_medications	0.464316
num_procedures	num_medications	0.386588
time_in_hospital	num_lab_procedures	0.315046
number_emergency	number_inpatient	0.281446
num_lab_procedures	num_medications	0.268071
num_medications	number_diagnoses	0.261044
	change	0.247635

num_procedures	number_inpatient	-0.065887
	number_emergency	-0.041881
	number_outpatient	-0.024351
patient_id	time_in_hospital	-0.023890
number_emergency	hemoglobin_level	-0.020732
time_in_hospital	hemoglobin_level	-0.018590
num_medications	hemoglobin_level	-0.018230
patient_id	num_procedures	-0.014419
number_inpatient	hemoglobin_level	-0.011542
num_procedures	readmitted	-0.011276



Finally, considering the objective of this project being to predict readmission, correlation coefficients of the features with readmission were extracted (shown below)

number_inpatient	0.164781
number_emergency	0.065709
number_diagnoses	0.049817
time_in_hospital	0.042647
num_medications	0.039105
insulin	0.033144
diabetesMed	0.027206
change	0.019311
num_lab_procedures	0.019160
number_outpatient	0.017730
patient_id	0.008108
blood_transfusion	0.005350
diuretics	0.001942
hemoglobin_level	-0.000169
complete_vaccination_status	-0.000558
has_prosthesis	-0.008412
num_procedures	-0.011276

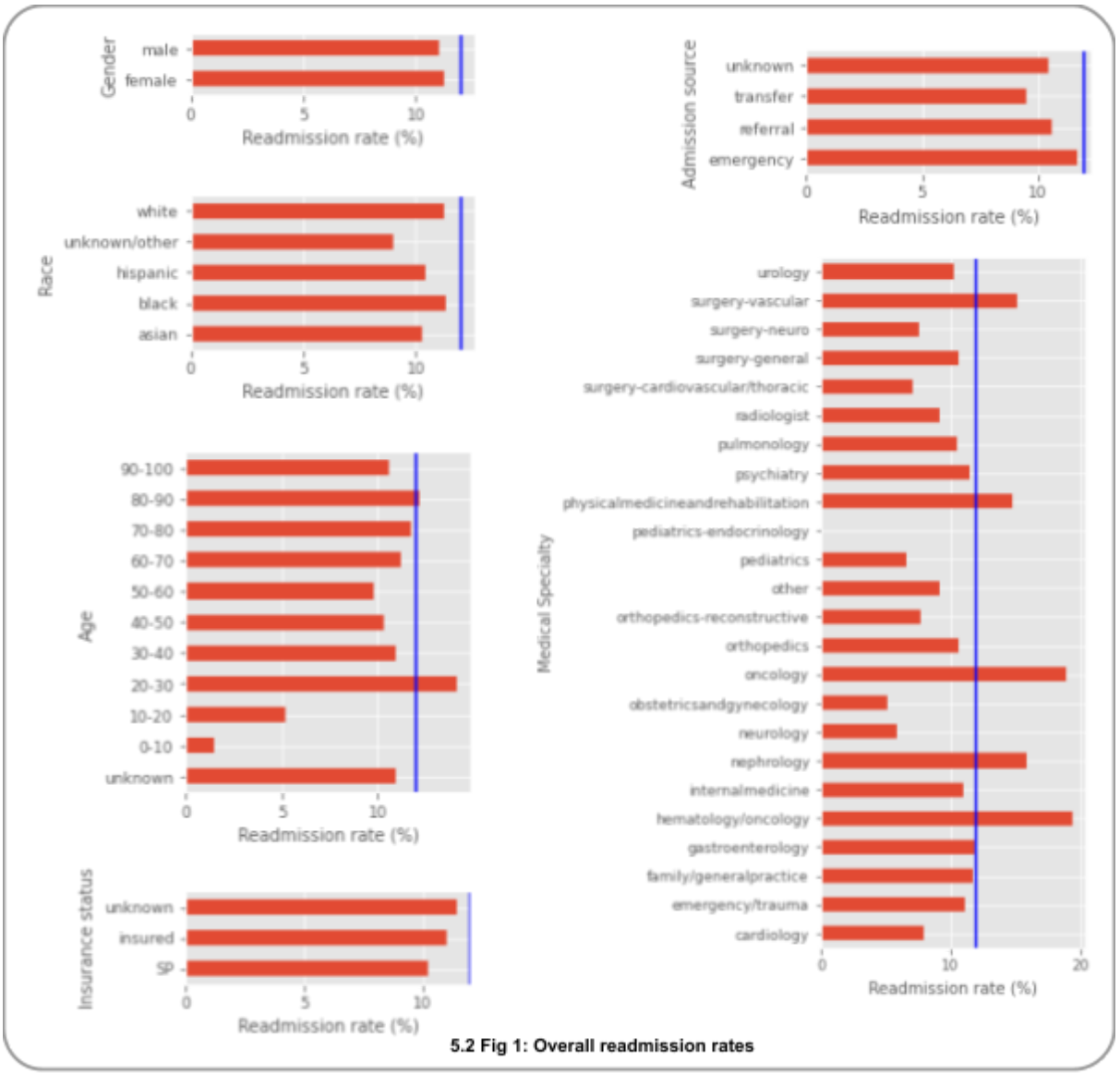


5.1 Fig 3. Prescription of diabetes medication

5.2 Business questions technical support

1) Analysis of overall discrimination

Categories within each of the sensitive characteristics (age, gender, insurance status and race were grouped and the readmission rates were plotted. The same approach was used for admission sources and medical specialties, as shown below). The specialties represented are those which had at least 100 admissions in the time period covered, all others were grouped in a category titled “other” and a significant portion were unknown (data for the latter two is not shown as it is not informative). Readmission rates for the whole data set were approximately 11%. To allow for a slight variation the threshold of 12% was selected to identify factors resulting in above average readmission rates. In the future, it would be beneficial to carry out a deeper analysis and examine odds ratios and risk, as well as address the massive quantity of missing data. However, the simple analysis shown here allows us to rapidly identify problematic situations. Of note, patients between 20 and 30 years of age, and patients visiting hematology/oncology, oncology, vascular surgery, physical rehabilitation and nephrology had higher readmission rates.

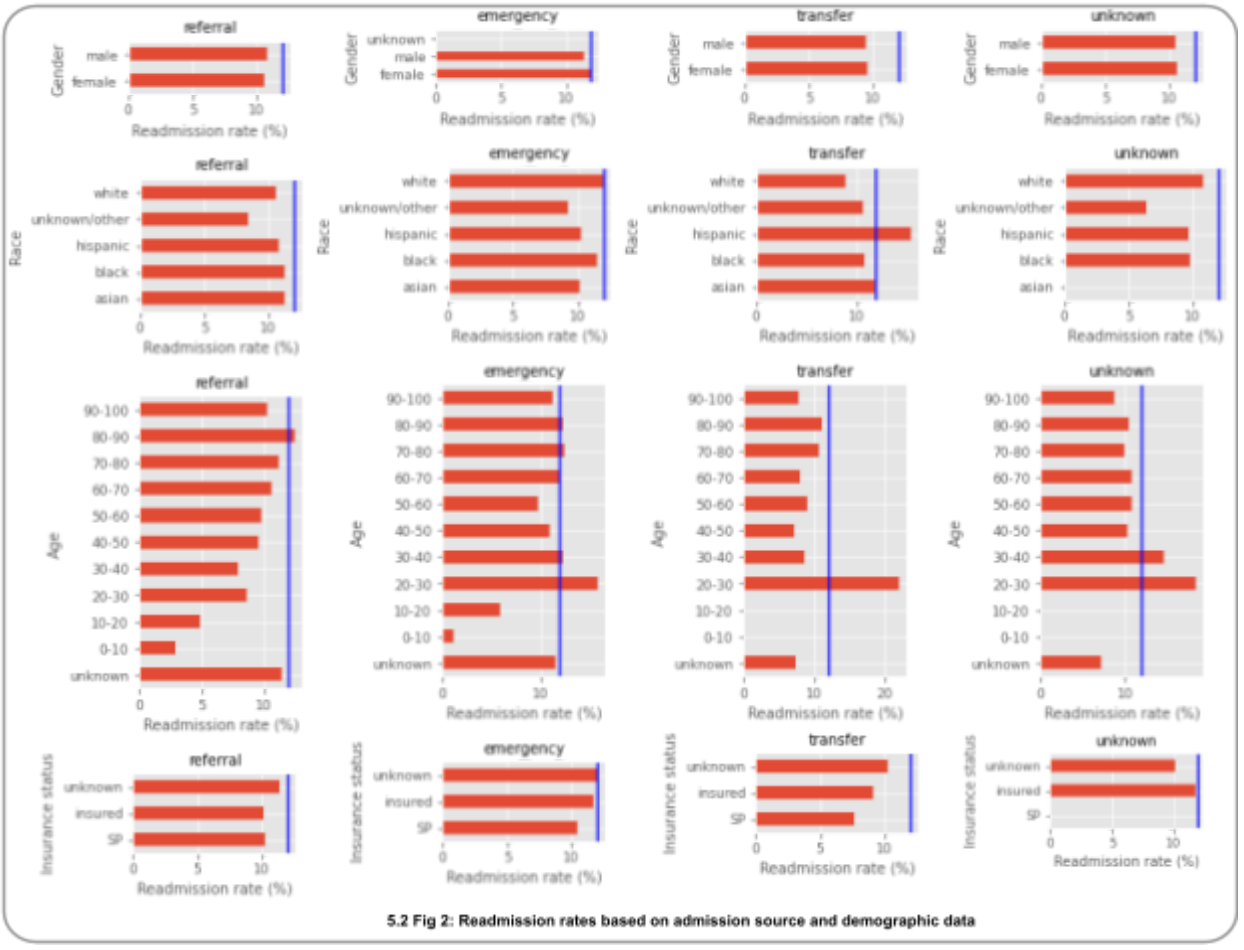


5.2 Fig 1: Overall readmission rates

2) Analysis of discrimination originating from admission sources

Admission sources were grouped into referrals, emergencies, transfers and unknown and the data was filtered to select each of these. Subsequently, divided into categories within the sensitive

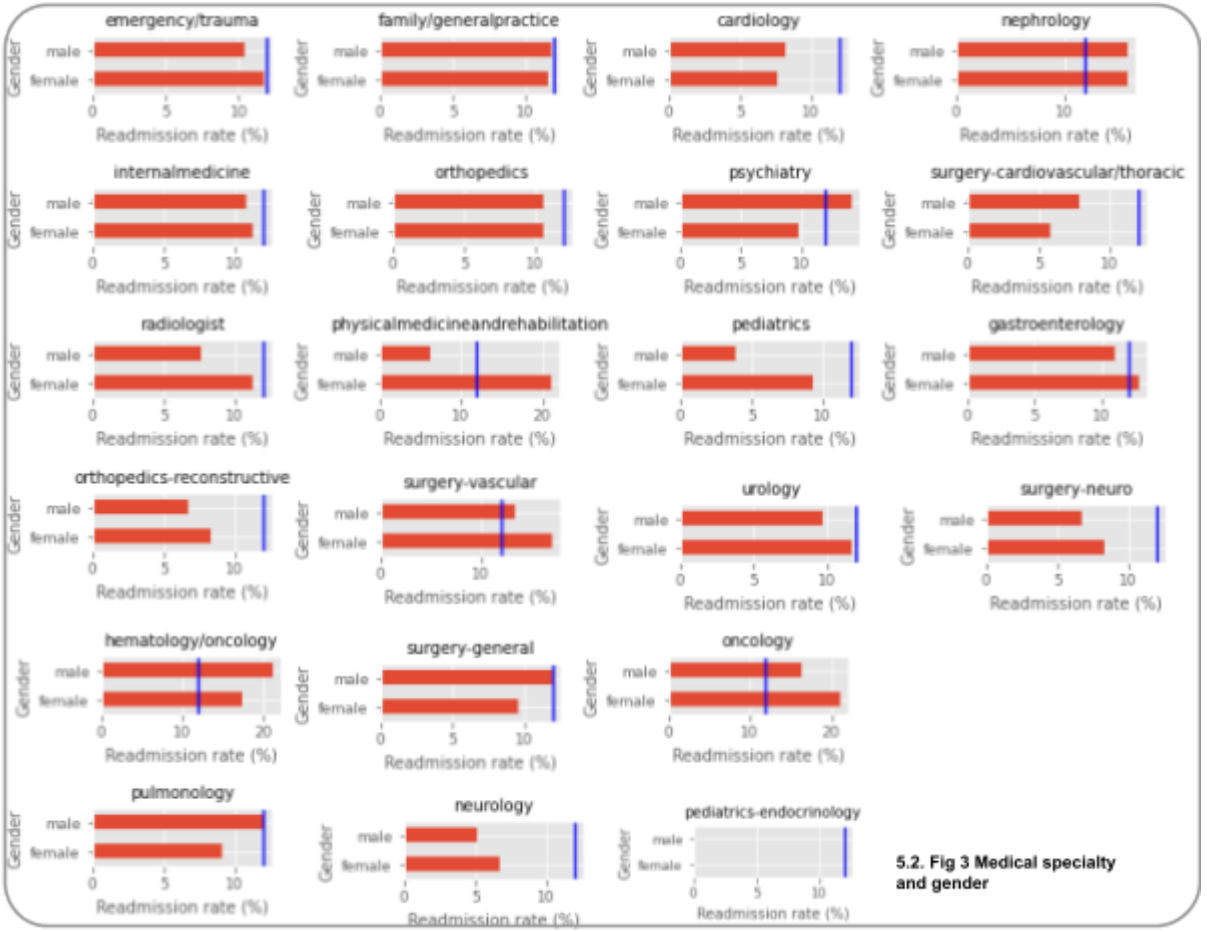
characteristics (age, gender, insurance status and race). Again, readmission rates above the threshold of 12% were deemed potentially problematic and further examined. In particular, although 20-30 year olds are a small portion of this dataset, the group consistently has higher readmission rates when having unreferred visits to the hospital. A similar tendency is observed for hispanic patients who are transferred to the hospital from other health care services, and they also show a significantly raised readmission rate.



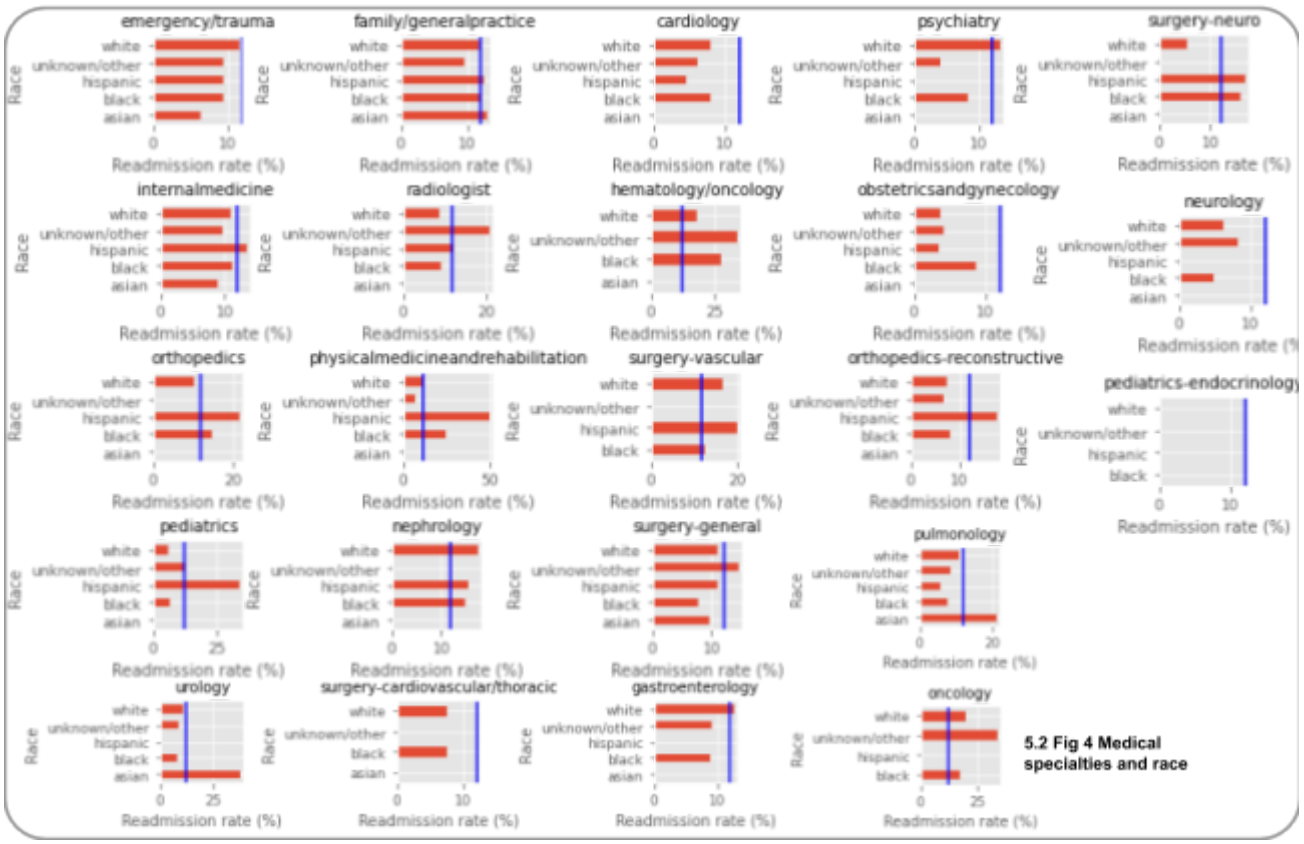
5.2 Fig 2: Readmission rates based on admission source and demographic data

3) Analysis of discrimination originating from different specialties

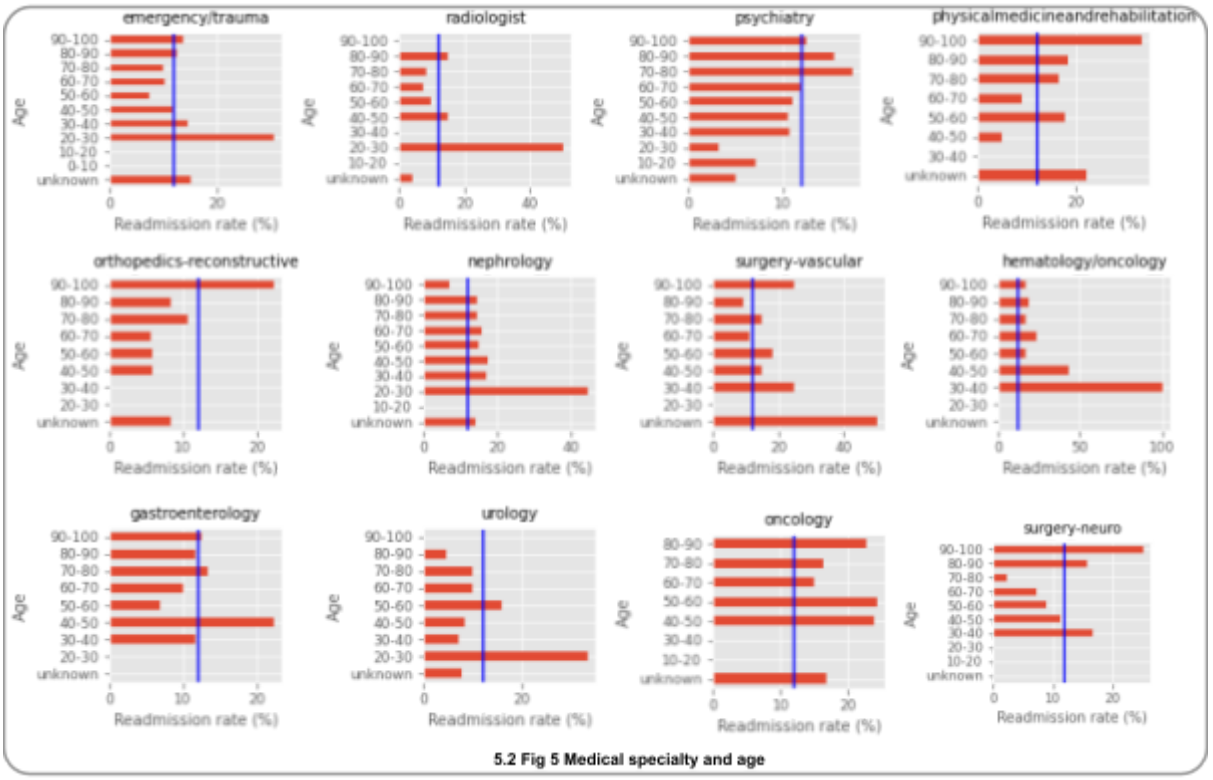
Specialties with at least 100 admissions present were analysed. The data was filtered to select each of these and divided into categories within the sensitive characteristics (age, gender, insurance status and race) as above. The 5 specialties identified as having higher readmission rates had similar results here, although physical medicine and rehabilitation had a large discrepancy between men and women. Data for race is summarized, but it is difficult to draw conclusion as races other than black and white have very low representation in the dataset. Data for age shows that certain specialties have differences in readmission based on age, possibly due to the nature of the illnesses they treat, selected data is shown for this field. Lastly, there are discrepancies linked to patient insurance status in several specialties which may require further investigation (Data shown on following pages).



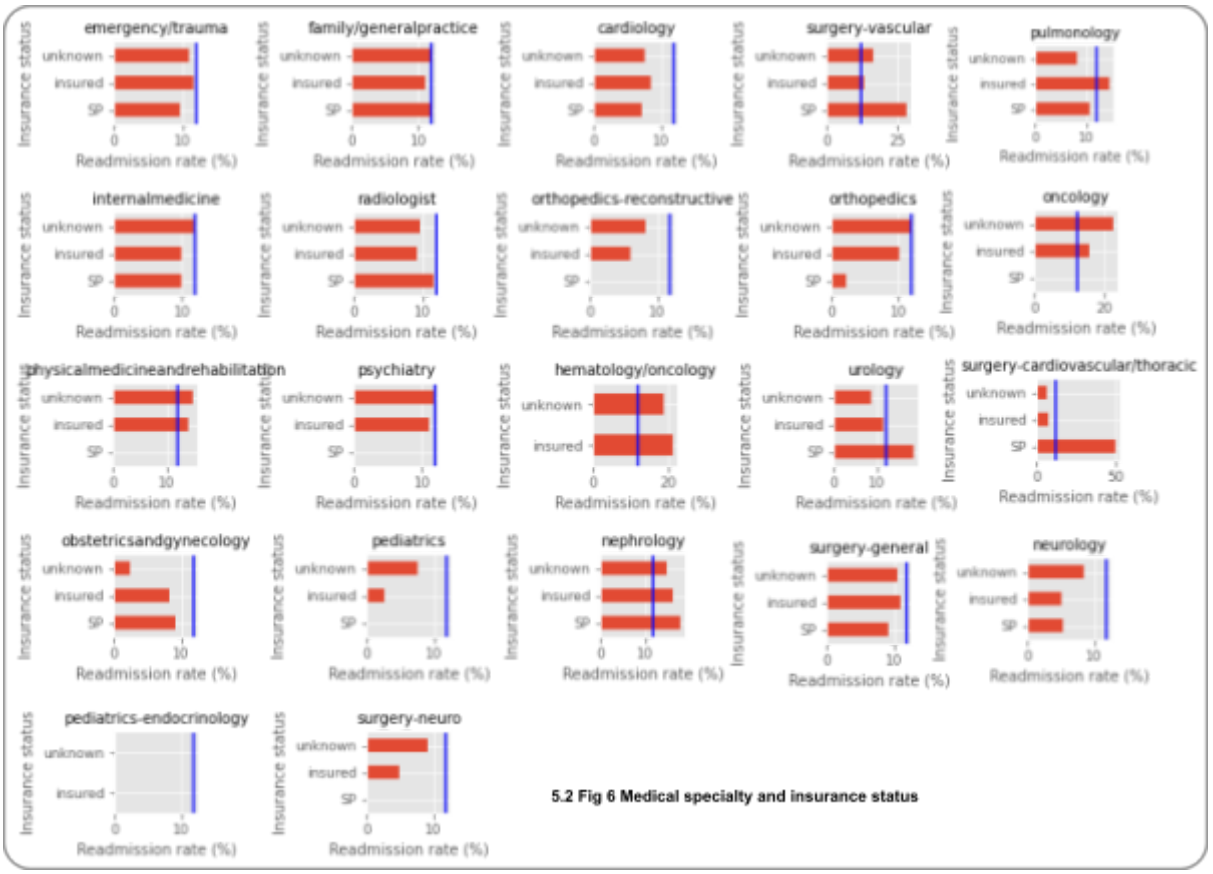
5.2. Fig 3 Medical specialty and gender



5.2 Fig 4 Medical specialties and race



5.2 Fig 5 Medical specialty and age



5.2 Fig 6 Medical specialty and insurance status

5.3 Model technical analysis

Feature selection

A Random Forest Plot with no features excluded was carried out on transformed data. Examining feature importance revealed particular importance for discharge disposition, various patient health indicators (number of hospital visits, medications etc, some diagnosis subgroups).

Some variables, such as medical specialties, whether or not the patient has a prosthesis or is on diuretics did not appear to have a role.

Furthermore, features with a large amount of missing data (specifically 'weight' were excluded). In categorical analysis, missing data was imputed as 'unknown' in its own category.

While patient ID appears to contribute, it is likely that this is due to a few patients who are frequently at the hospital.

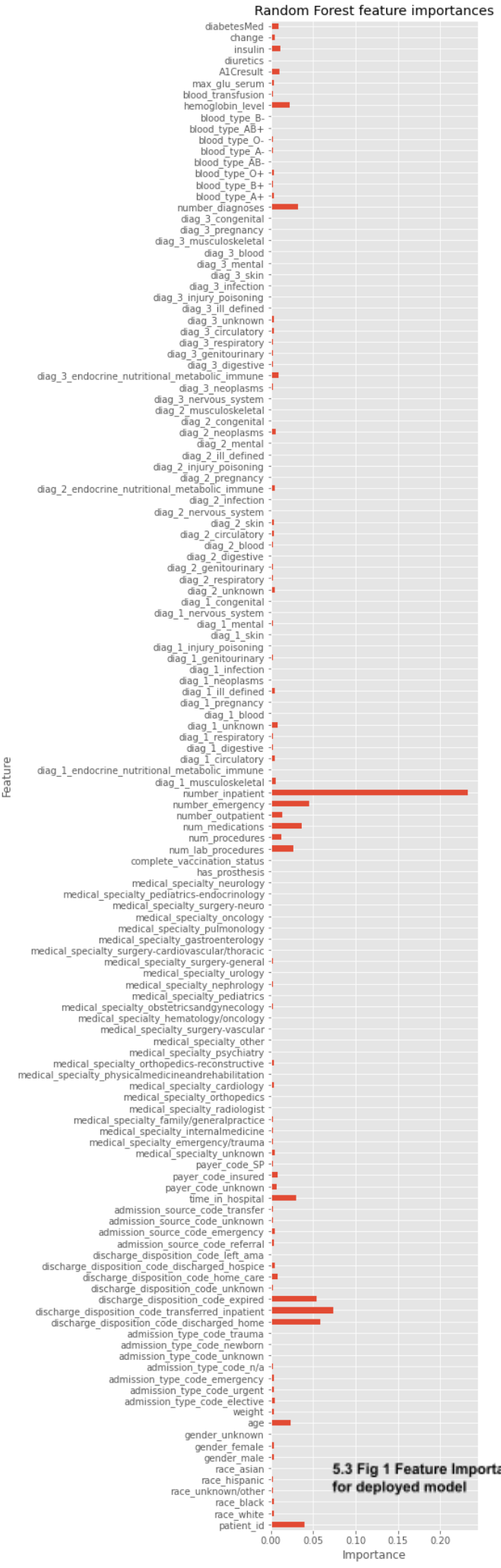
In conclusion, all features were used for model deployment except the following: weight, diuretics status, vaccination status, prosthesis status, medical specialty and patient ID.

Sensitive features were also included in the model - if they are potential reasons for wrongful discharge, it is reasonable to "correct" for that in a predictive system looking to prevent these incidents. Once root causes for this potential discrimination are solved, these features can be excluded.

Categorical encoding

Categorical variables were encoded as follows:

8. 'race': string entries were processed to automatically assign 'white', 'black', 'hispanic', 'asian' and 'other'
9. 'gender': left as is
10. 'admission_type_code': codes were grouped to assign either 'referral', 'transfer' or 'emergency', missing data was treated as 'unknown'
11. 'discharge_disposition_code': codes were grouped to assign either 'home', 'transfer' or 'home_care', 'hospice', 'left_ama' or 'expired', missing data was treated as 'unknown'
12. 'admission_source_code': codes were grouped to assign either 'emergency',



5.3 Fig 1 Feature Importance for deployed model

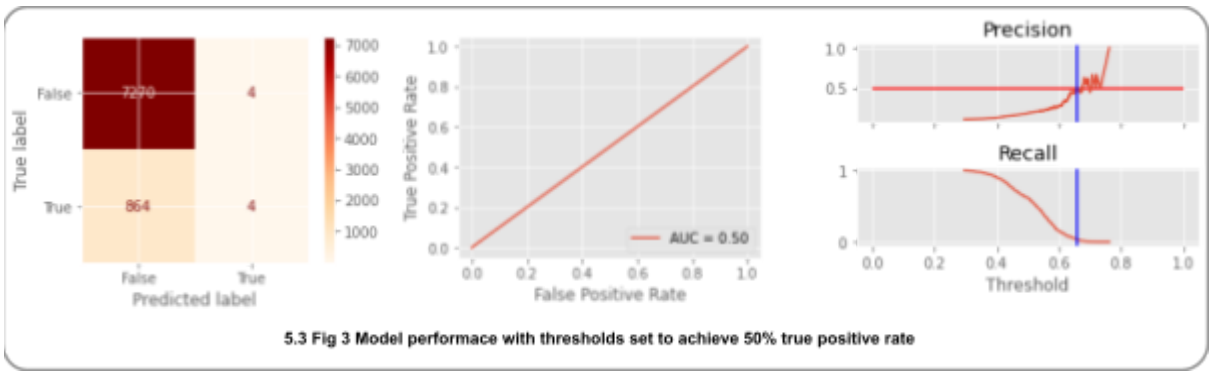
'elective', 'newborn', 'trauma' or 'urgent', missing data was treated as 'unknown'

13. 'payer_code': patients with codes other than SP were considered insured, missing data was classified as 'unknown'
14. 'medical_specialty': specialties with at least 100 admissions were retained, others were grouped into 'other', missing data was treated as 'unknown'
15. 'blood_type': left as is (imputation errors were found, with patient blood types changing - further clarification and correct values were not provided)
16. 'diag_1', 'diag_2' and 'diag_3': were grouped and assigned into one of 19 categories based on CDC classification of diagnostic codes. This data is parsed and available at: http://icd9.chrisendres.com/index.php?action=contents

Model performance

Model performance evaluated using ROC AUC and F1 scores, using current deployed system (top) and thresholds to attain 50% true positive rate (bottom). The latter figure clearly shows that the current mode, with thresholds set to attain a 50% true positive rate would not result in a reduction in wrongful discharge and consequently, this option was discarded. A precision-recall curve was generated but provided no additional insight, so is not included.

5.3 Fig 2 Summary of model results with provided training data



5.3 Fig 3 Model performace with thresholds set to achieve 50% true positive rate