# Email from Dr Agnes Crumplebottom (Jan 25th)

Greetings,

It’s good to know who will be handling this mess. Your email was very thorough, if not extensive. At least you seem to have a solid understanding of our data and challenges. I’ve attempted to answer the main questions. Please find my answers below.

***Given the data at hand, what output/label do you expect? Do you want the verification model to state whether a patient should be discharged? Or do you want the model to classify the patient as a possible wrong discharge?***

As I’ve stated in my previous mail and in the data dictionary, the goal is to predict if a person is likely to be readmitted within 30 days. For the patients that would be readmitted in 30 days, we would like to identify them and keep them in our facilities further. This means that patients classified as ***readmitted = “Yes”*** should stay longer and patients classified as ***readmitted = “No”***show no indication that they require further assistance and can be safely discharged.

***How should we assess model performance? Which minimum values should be considered for the patient readmission classification and for the bias between specialties?***

This is a very technical question and, as the expert, it should be up to you to make the ultimate decision based on our business requirements: we cannot fail to provide care to a patient that needs it. This means we would like to minimize the number of wrong discharges that would lead patients to return in less than 30 days.

But we are not made of money so this should not be the only criteria. As a rule of thumb, consider that at least 50% of the patients identified for readmission should actually be sick.

The wrongful discharge rate, or readmission rate, should not vary more than 10 percentage points between sub-groups, and less than 5 percentage points between medical specialties.

***Information on patient insurance status is unclear. Could you provide more information on this?***

The record of a patient's insurance status should be able to be inferred from their Payer identifier information. With the ***SP*** (self-pay) code you will be able to identify the uninsured patients, with the remaining codes referring to different insurance providers.

***I understand you are concerned about imbalances regarding medical specialties. However, by looking at the data we find 71 specialties, many of which have a very low amount of data points.***

I trust your expertise to deal with this imbalance in the way you see fit, including defining a minimum number of observations, and justify it in your report.

***Regarding the dataset, I would like to clarify a few issues:***

***- do the null values in the num\_medications and num\_lab\_procedures variables indicate a zero/lack of medications/lab procedures or a real absence of data?***

***- do the "None" values in the complete\_vaccination\_status variable indicate a lack of information or lack of vaccines?***

***- finally, we have a lot of value equal to "?". Should these be considered as missing values?***

When no medication is provided to the patient, the value of *num\_medications* is zero (0). The same applies for *num\_lab\_procedures*.

Values such as “None”, “?” and similar in all non-numeric columns refer to information that was not possible to collect.

***Could you be more precise on what is a “service” exactly, as it does not appear explicitly in the data set.***

Consider as a service any of our medical specialties.

***​​For the race attribute, it looks like there are different descriptions for the same race groups. How should we consider these categories?***

We trust your analysis of the records. Being you the expert, we leave to you how to handle these seemingly similar descriptions. Please refer to this situation and explain the strategy you choose in your report.

***You mention how some things impact data over-time. Is it possible to add dates to the data to ensure we are allowing for this?***

Unfortunately, my team tells me that some incompetent IT guy lost the time stamp information for the data that you were given. Please see your report as an opportunity to highlight their wrongdoing.

***With the goal to evaluate the impact of staff in this situation is it possible to have at least a code that identifies the doctor responsible for the discharge or the doctor/team that was responsible to treat the patient?***

You can assume that “Physician specialty” represents the specialty (or department) responsible for discharging a patient and that this level of detail is sufficient for the task.

***Were there any changes in the hospital board, or some cuts in budget during the period of the observations you sent?***

The old goons from the board have been here since the time TVs were black and white. And even though they keep refusing additional funding for my much needed office renovation, there is no way I would let them cut down my budget.

***There seems to be some data inconsistencies, particularly regarding the diagnosis codes.***

Consider this as data imputation errors. With you being the expert, I expect a justification on how you decide to handle these cases.

***What are you expecting from the API response only a true or false answer? Can you please provide us with a few samples of the POST requests which will be made?***

Very good question! I will be asking the Technical team and we will get something together and share with you within a few days.

***Can you please provide us with a reference with a correspondence between the first 3 digits of ICD9 codes and the respective disease?***

According to one of our staff, who I like to call “annoying wikipedia”, the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) is based on the World Health Organization’s Ninth Revision, International Classification of Diseases (ICD-9) and is available from Center for Disease Control through their [FTP service](https://www.cdc.gov/nchs/icd/icd9cm.htm). The [wikipedia page](https://en.wikipedia.org/wiki/List_of_ICD-9_codes) might also be of use.

***Almost 96% of its values are missing for the weight feature, so it can’t be used to do a predictive model.***

This is a shame, and I expect to see this covered in your report.

***What medications are included in the DiabetesMed feature? Does this feature exclude insulin?***

The DiabetesMed refers to any diabetes medication, such as insulin, insulin stimulants, diuretics, and glucose uptake inhibitors.

***The patient\_id column is said to be unique per patient. However, I found out that some patients have multiple blood types. I am guessing that either patient\_id or blood\_type (or possibly other attributes) of these records have been wrongly entered or corrupted.***

This is a relevant question. **The *patient\_id* is indeed unique for each user**. So it seems you might have identified an issue with our data gathering methodology. I would suggest your IT department look into this issue. Again, as the expert, we leave it to you on how to handle this.

***Since data contains multiple patient visits for the same patients, may we consider only the first encounter for each patient to know whether or not they were readmitted in less than 30 days in order to prevent the bias of our model?***

As mentioned earlier, our data doesn’t have any time-related features. So I’d avoid making any assumptions that the visit outcomes from the same patient are correlated. Also, the model should be general enough such as it can make predictions on unseen patients.

***To improve record keeping, do you want our new system to restrict options to only those already in use for fields such as race, age, gender etc,? Are there any additional categories which are not in the database but are available options and can they please be added to the field descriptions?***

The world is a never-ending changing thing, so new categories may show up. We expect your solution to be flexible enough so that it does not crash when that happens.

***Why are more diagnoses associated with some patients (as per the number\_diagnoses column) than diag\_X (diag\_1, diag\_2 and diag\_3) columns? What are the other diagnoses?***

The diag\_1/2/3 variables correspond to primary, secondary and additional secondary diagnosis, respectively. These aim to identify the primary cause for hospitalization, so there might be other other diagnoses the doctor might find in the patient but we only record the three most important ones.

***There are several codes in the Discharge\_disposition\_code that I don’t understand. Can you please provide an explanation for them?***

I should also understand these codes but I don’t. I consulted my Chief Physician Dr. Susan Google, and she told me that these are standardized codes to identify the status of the patient at the end of their admission period in a health care facility.