Data Quality report – Initial findings.

1. Overview

In this report the findings will be outlined based on the initial cleaned dataset (covid19-cdc-20206773-cleaned_V1.csv) that has been produced after a series of data preparation techniques. It will provide a summary of the data and describe a number of data quality issues that were observed and how they will be dealt with. The appendix will provide some background for this dataset, including a range of feature summaries, and bar charts used for visualization of the data.

From the initial analysis of the data it was evident that the data had no continuous features, only categorical features were present. The data had a total of 12 features, 4 of which were date data. This date data will be analysed as categorical data for the remainder of this data quality report and data quality plan. The remaining 8 features were all text based data. A large amount of data appeared to be missing from the dataset. Missing data was labelled as 'missing' in the case of the textual data, and 'NaT' for the missing datetime data. As well as this a large proportion of the data was labelled as 'unknown', this will be further investigated within the data quality plan and extra steps will be taken to ensure that no relevant features or rows are dropped without close analysis of the data itself.

2. Summary

A number of tests were carried out in order to test the logical integrity of the data. Due to the fact that the dataset is made up predominantly of categorical features there is very little room for logical errors as the majority of the data is either yes or no entries, and simplistic features relating to age group, gender, race/ethnicity etc. A significant number of failures of the date type data were found, where some patients were entered as probable cases, while also having date entries for positive specimens being recorded and vice versa. These logical errors will be further addressed in the logical integrity section.

A number of changes are recommended for the categorical values within this dataset. A large proportion of the data is missing and therefore further investigation of these entries is required, in some cases it may be recommended to drop the feature entirely due to such large proportions of missing values. For example features such as pos_spec_dt, icu_yn and medcond_yn all have roughly 70% of their values missing. This would usually require dropping these features, this will be investigated further. For those features that are missing less than 30% of their features, data imputation will be considered, if this is not a feasible option then the rows may have to be dropped.

On top of this a large proportion of entries are described as unknown. On checking the case report form it was noticed that 'unknown' was an optional choice for a number of the data entries, therefore this must be considered when deciding whether or not to drop this feature. While the data entry may be unknown, a patient may have chosen the value due to the fact they don’t fit into a predefined category, and as such it should not be dismissed in the same fashion as a missing value, but rather considered as a valid value.

Regarding the datetime features, due to the fact that these are being treated as categorical data and not continuous, it is recommended that they be converted to a more practical categorical breakdown other than by day. As the date data are in daily format, this results in the datetime features having very high cardinality which can make the visualizations of the data quite messy, it would be more appropriate for the analysis if these dates were categorized by month.

A large number (855) duplicate rows were discovered. On closer inspection of these rows it was discovered that a very large proportion of these rows were missing values across the majority of the features, due to this fact these rows were dropped from the dataset as they provided very little useful information.

3. Review logical Integrity of the data

The data seems to hold up logically at first glance, however there are a few errors that could arise due to the datetime data and logical errors surrounding hospital and ICU admissions. Tests were carried out to check the logical integrity of the data. They are as listed below:

Test 1: Check if any entries have been to ICU but not to the hospital

- Result: 1 case found.
- It will be assumed that this case required immediate medical attention and was admitted to ICU immediately.

Test 2: Check if any probable cases have a corresponding date of first positive specimen collection.

- Result: 239 cases found.
- This must be checked with the domain expert as there are quite a large number of rows affected and this may be to do with the dataset not yet being updated. Therefore no action will be taken on the affected rows.

Test 3: Check if any laboratory confirmed cases do not have a corresponding date of first positive specimen collection.

- Result: 5897 cases found.
- This must be checked with the domain expert as there are quite a large number of rows affected and this may be to do with the dataset not yet being updated. Therefore no action will be taken on the affected rows. This logical error leads more cause to investigate the feature Pos_spec_dt which has a very large percentage of missing values.

4. Review Categorical Features

There are a total of 12 categorical features in the original data set, four of which are date type data, but as explained previously these will be addressed as categorical features for the purpose of this data quality report and plan. The date type data originally were recorded by day, but for our analysis these entries will be converted to monthly entries in order to reduce the cardinality of these features. All other features are written text data, including the target feature 'death_yn' which records 'Yes' or 'No' dependant on whether the patient died as a result of contracting covid-19.

Apart from the datetime features the remaining features are text based with a range of different cardinalities. For example 'age_group' is split into 10 different categories, while 'race_ethnicity_combined' is split into 9. Within all of the features except for 'cdc_case_earliest_dt', 'current_status' and 'death_yn' there are significant amount of values which are labelled as 'missing', these are the equivalent of a null value and therefore will be treated as such. There must be a distinction made between entries labelled as 'missing' and those labelled as 'unknown', missing values indicate that the data was missing from the case report form, while unknown values were provided as an option within the form. In any case these unknown values provide little informational content and therefore must be further investigated whether they are to be dropped or not as they may skew the results of the analysis at a later stage.

The most significant data quality issue that will need to be dealt with is the large proportion of missing values across the dataset. There are a number of features across the dataset that have more than 50% of values missing from them. These missing values will be further analysed in the data quality plan. It will be checked whether or not these features can be dropped, or perhaps if imputation is a better approach for replacing the missing values.

The modal value of 'race_ethnicity_combined' is 'unknown', and the second modal value for 'hosp_yn' is also 'unknown'. While these values were inputted by the patients filling out the case report form, they provide very little information regarding the features they correspond to. These unknown values will be investigated further, perhaps imputation could be a possible solution, and if not dropping these rows could be considered. This will all be dependent on the rows themselves and whether or not they are redundant in our analysis or not.

5. Actions to take

Two main actions will be taken, these are summarised below

- Missing Values:
    - All features with over 50% of values missing will be dropped form the dataset, this includes pos_spec_dt, icu_yn and medcond_yn.
    - Onset_dt has nearly 50% of its values missing, but it will be assumed that all values with missing entries in this feature are patients who did not experience any symptoms due to covid-19 and therefore they will be kept as they provide useful information for analysis.
    - 21% of hosp_yn values are missing, these missing values will be investigated and data imputation will be considered.

- Unknown Values
    - The rows containing unknown values will be further investigated and it will be decided whether or not these rows are redundant in this dataset.

- Investigate feature: cdc_report_dt
    - This feature is missing over 20% of its values and due to the fact that we categorized the datetime data entries into months, it is highly likely that a large percentage of cdc_report_dt dates will match their corresponding cdc_case_earliest_dt dates. As well as this it is recommended on the CDC website that researchers use cdc_case_earliest_dt for time based analyses.

6. References

[1] CDC description of COVID-19 Case Surveillance data
https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

7. Appendix

7.1 Feature Summaries

- ‘cdc_case_earliest’ – earliest available date for the record
- ‘cdc_report_dt’ – initial date case was reported
- ‘pos_spec_dt’ – date of first positive specimen collection
- ‘onset_dt’ – symptom onset date
- ‘current_status’ – Case status (laboratory confirmed/Probable case)
- ‘sex’ – gender (Male/Female/unknown/other)
- ‘race_ethnicity_combined’ – Race/ethnicity of patient
- ‘hosp_yn’ – Hospitalization status (yes/No)
- ‘icu_yn’ – ICU admission status (Yes/ No)
- ‘medcond_yn’ – Underlying medical conditions (Yes/No)

7.2 Categorical Features (Descriptive Statistics)

	count	unique	top	freq	mode	freq_ mode	%mode	2ndmode	Freq2nd mode	%2nd mode	%missing
cdc_case_earliest_dt	9145	13	2020-12	1928	2020-12	1928	0.21	2020-11	1571	0.17	0.00
cdc_report_dt	7545	11	2020-12	1418	2020-12	1418	0.19	2021-01	1337	0.18	17.50
pos_spec_dt	2829	11	2020-11	513	2020-11	513	0.18	2020-12	483	0.17	69.07
onset_dt	4985	12	2020-11	922	2020-11	922	0.18	2020-12	853	0.17	45.49
current_status	9145	2	Laboratory-confirmed case	8487	Laboratory-confirmed case	8487	0.93	Probable Case	658	0.07	0.00
sex	9135	3	Female	4761	Female	4761	0.52	Male	4304	0.47	0.11
age_group	9133	9	20 - 29 Years	1647	20 - 29 Years	1647	0.18	30 - 39 Years	1449	0.16	0.13
race_ethnicity_combined	9063	8	Unknown	3271	Unknown	3271	0.36	White, Non-Hispanic	3253	0.36	0.90
hosp_yn	7209	3	No	5095	No	5095	0.71	Unknown	1428	0.20	21.17
icu_yn	2376	3	Unknown	1295	Unknown	1295	0.55	No	1003	0.42	74.02
death_yn	9145	2	No	8831	No	8831	0.97	Yes	314	0.03	0.00
medcond_yn	2563	3	No	923	No	923	0.36	Yes	902	0.35	71.97

7.3 Bar Charts

An accompanying pdf will display all relevant bar charts of the categorical data.