AC 221 Final Project: Privacy Protection in Studies about Coronavirus disease (COVID-19)

Jordan Turley, Qiuyang Yin, Qiang Fei

May 4, 2020

1 Abstract

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Since the first identified case in December 2019, it spread quickly among people and as of 3 April 2020, there are more than 1 million confirmed cases of COVID-19 worldwide. People in many countries are suffering from the disease because their family, friends and even themselves might get infected, and also because their regular life is affected. In such an unprecedented and tough situation, there is lots of effort going to studying and controlling this spreading disease academically. While the number of confirmed cases in different regions is reported to the public every day in most countries, papers and articles about different aspects of COVID-19 are also published on an almost daily basis. Although the information is published to contribute to the current urgent situation, some ethical issues may also emerge within this quick process. In this project, we want to study the potential ethical problems that can be identified in published studies about COVID-19 with a focus on the protection of an individual's privacy.

2 Introduction

In this project, our main focus will be to study the privacy issues raised by the use of datasets related to COVID-19. Our study will be divided into two parts. The first part will be concerned with checking and comparing how published datasets for some countries protect individuals' privacy in addition to providing enough information for their citizens and the world to study. The second part will be about literature reviews. Lots of papers and articles are published about COVID-19 and data and information about patients at different scales are used in most of them. We want to analyze some of these works to identify how they deal with privacy issues while ensuring that they transfer well-supported conclusions. After these two steps, we anticipate that we might also find other interesting and related topics for study, such as other ethical problems that might occur in these related studies,

3 Project Details

After some research, we find that only China, South Korea, and India have published information about individual leveled patient information, compared to other countries where only the total numbers are published, or at the state or county level in the US. We thus decide to perform detailed investigation on how the protection of an individual's privacy might be an issue for published datasets about COVID-19 in these countries, and how well these data can be used in terms of making predictions.

3.1 China's individual data

The dataset is obtained from a published research paper named Epidemiological data from the COVID-19 outbreak, real-time case information(Xu et al. Scientific Data 7:106, 2020) [4]. It contains confirmed cases until March 9th 2020, and contains more than 40 thousand data points, which is around half of all confirmed cases in China. This is a good amount of data for studying the spread of COVID-19.

After investigating the dataset, we find lots of missing values for variables such as age, gender and also outcomes. It is not surprising though because individual leveled data are mostly not published officially but rather gathered from news papers, where individual cases are reported and described. Thus, we want to do some EDA first to investigate what this dataset might suggest to us.

3.1.1 Exploratory Data Analysis

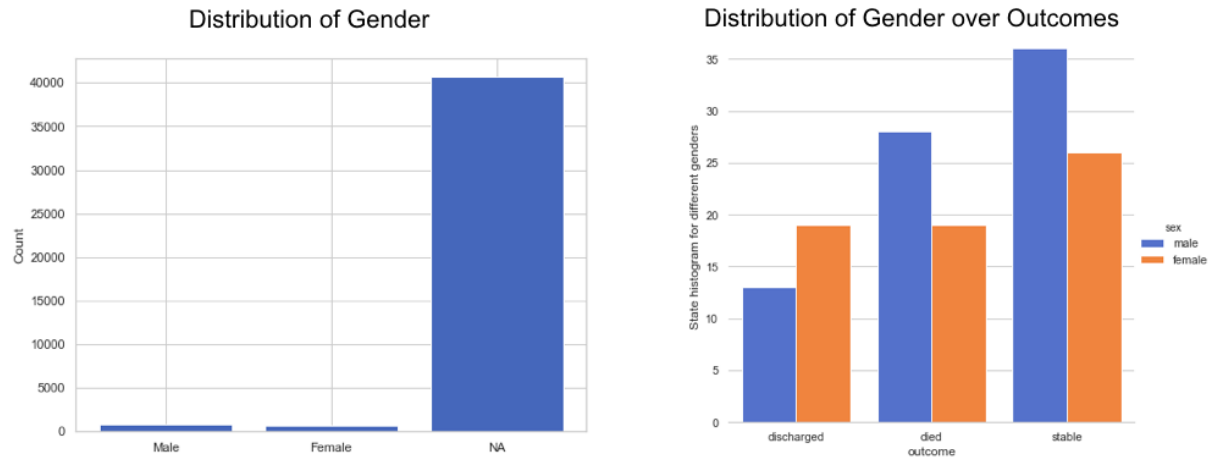

Figure 1: Distribution of Gender and Distribution of Patient Status based on Gender

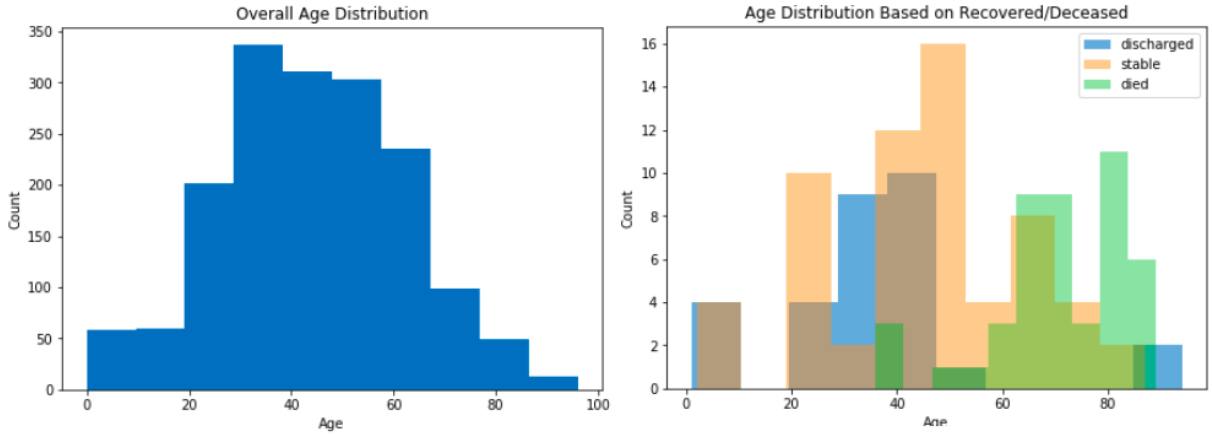

Figure 2: Overall Distribution of Age and Distribution for Recovered and Deceased Individuals

In Figure 1 and Figure 2, we can see that we do not have many non-NA data points for either gender or age, and even less data presents that also have non-NA outcomes. However, if not compared to the scale of the whole dataset, the existing data still show some trends, for example, in the relationship between outcome

and age. some analysis could still be done using this dataset. Also on the other side, we still have lots of data for some other variables such as locations, so this dataset is still worth studying.

3.1.2 Anonymity and Privacy

Next, we analyze the k-anonymity of the dataset. This dataset has one unique identifier which is ID, and also some quasi-identifiers including:

1. age (containing both years and ranges)
2. sex
3. city/province (There are around 300 NA values for cities, and no NA value for province)
4. admin_id (Administrative unit ID of the lowest level available for the case reported)

In theory, such information can be used to match an individual in some other dataset and thus the individual can be uniquely identified. Below, we calculate the k-anonymity of the dataset. In this case, it is basically how easy a person can be identified if someone else knows both such information and that this person is confirmed COVID-19 positive.

After checking the original dataset, we find that the initial k-anonymity of this dataset is 1. This is not too surprising as this datasets was not processed for anonymity before published . However, we do see that there are only 27 rows that are completely unique, and only a handful that are below 5-anonymous.

By examining some of the unique records, we found that they all have a unique age. We thus did some generalization though converting ages to age ranges. This change also serves as a cleaning-up step as entries for age exist both in forms of integers and in forms of ranges. We then checked and found that we cannot increase anonymity by changing the city to NA for these values. And the variables cannot really be further generalized in this case. Thus, we need to drop some records.

However, instead of deleting rows with combinations of all quasi-identifiers that are not seen for enough times in the dataset, we don't use age-range and gender as quasi-identifiers in this case. On the one side, both are some variable that would provide some information about the patient, and we do want to include this information here. On the other side, we believe that it will not be a problem if we don't further address these two variables. The reason why they make some key indices unique is that there are too many missing data in the dataset. As it is highly unlikely that people could tell whether a specific case is counted as NA or with the exact given information, these quasi-identifiers are not so uniquely identifiable as they appear to now. Thus, we proceed by checking how anonymity is achieved with quasi-identifiers excluding age and gender, and we delete the remaining rows.

	k = 1 samples	K = 2 samples	K = 3 samples	K = 4 samples	K = 5 samples
Original Data Set	27	825	2	93	0
After Generalization	3	561	3	137	0
Not looking at age/ gender	2	148	1	78	0

Figure 3: Number of Records not Achieving Given Anonymity in each Case

A summary of number of rows not qualified under each k-anonymity at each step is shown in Figure 3. Deleting remaining rows would not make too much change as we have around 40,000 data points in total and here we only lose less than a thousand data points. However, we still want to check if trends are preserved for variables where not many non-NA values exist.

3.1.3 Examination

At this step, we want to check the relation between some variables and the outcomes before and after we increase anonymity is still maintained.

We did some examination on the dataset achieving 5-anonymity, and trends are generally preserved, as shown in the case here considering the relationship between age and outcome.

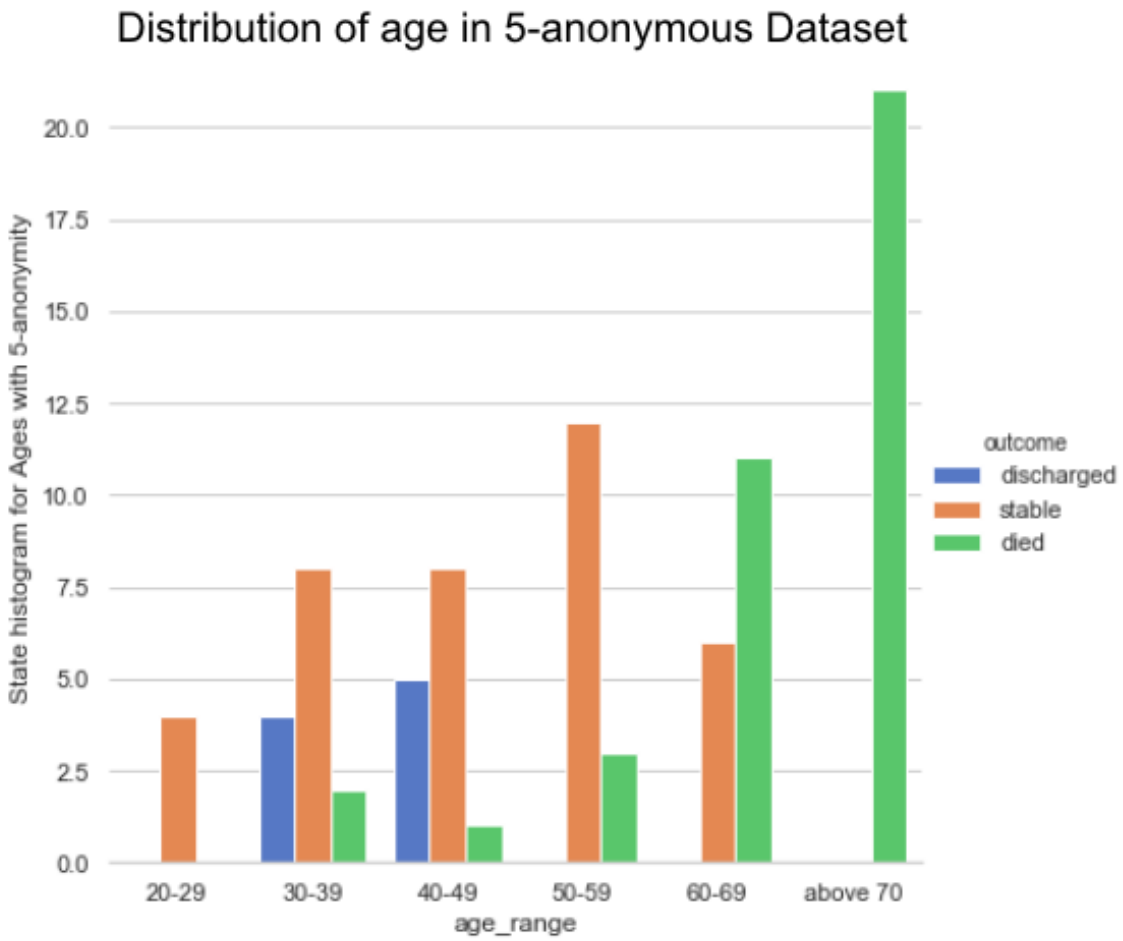

Figure 4: Age vs. Outcome in 5-anonymous Dataset

As we can see in Figure 4, the majority of people who died because of COVID-19 fall in a higher range of ages, which is similar to the situation in the original dataset.

3.1.4 Conclusion for data from China

Thus, we can conclude that privacy is generally not an issue in data from China due to the scale of the dataset and the prevalence of NA values in some important quasi-identifiers.

We are not making predictions in this case because we have very limited amount of known outcomes from the dataset, and it is almost unlikely that we can correctly predict outcomes. Cases using predictions are shown below.

3.2 South Korea's Individual Dataset

The dataset is obtained from a public Kaggle dataset named Data Science for COVID-19 (DS4C). Its data originally come from KCDC (Korea Centers for Disease Control Prevention) and local government, and thus to become an official source dataset. The DS4C team makes a structure dataset based on KCDC's report and make it publicly on Kaggle. So far, the dataset is updated on April 30th and the next update would be June 1st. Till now the total number of confirmed cases in the whole dataset is 10,674, with 8,114 recovered and 236 deceased.

Since we are interested in the privacy protect in the dataset, we mainly focus on the individual-level dataset. The patient information dataset consists of 3,326 patients, with 1,637 recovered, 1,622 still isolated and 67 deceased. Unlike China's dataset, we notice far fewer missing rows and more features in the dataset. Besides demographic information like sex, age and birth year, it also contains features like the specific date of confirmation and release (confirmed date, released date, deceased date) and other useful features like infection order and contact number. Due to the huge information it provides, the dataset should be useful.

Figure 5 shows the first and last three lines of the dataset:

Out[3]:

	0	1	2	3323	3324	3325
patient_id	1000000001	1000000002	1000000003	7000000011	7000000012	7000000013
global_num	2	5	6	NaN	NaN	NaN
sex	male	male	male	male	female	female
birth_year	1964	1987	1964	NaN	NaN	NaN
age	50s	30s	50s	30s	20s	10s
country	Korea	Korea	Korea	Korea	Korea	China
province	Seoul	Seoul	Seoul	Jeju-do	Jeju-do	Jeju-do
city	Gangseo-gu	Jungnang-gu	Jongno-gu	Jeju-do	Jeju-do	Jeju-do
disease	NaN	NaN	NaN	NaN	NaN	NaN
infection_case	overseas inflow	overseas inflow	contact with patient	contact with patient	overseas inflow	overseas inflow
infection_order	1	1	2	NaN	NaN	NaN
infected_by	NaN	NaN	2.002e+09	7e+09	NaN	NaN
contact_number	75	31	17	5	9	6
symptom_onset_date	2020-01-22	NaN	NaN	NaN	NaN	NaN
confirmed_date	2020-01-23	2020-01-30	2020-01-30	2020-04-03	2020-04-03	2020-04-14
released_date	2020-02-05	2020-03-02	2020-02-19	NaN	NaN	NaN
deceased_date	NaN	NaN	NaN	NaN	NaN	NaN
state	released	released	released	isolated	isolated	isolated

Figure 5: A glimpse at South Korea dataset

3.2.1 Exploratory Data Analysis

We want to first do some exploratory data analysis to get a general idea of the distribution of variables before further investigation.

First we look at the distribution of age. The distribution is shown in Figure 6.

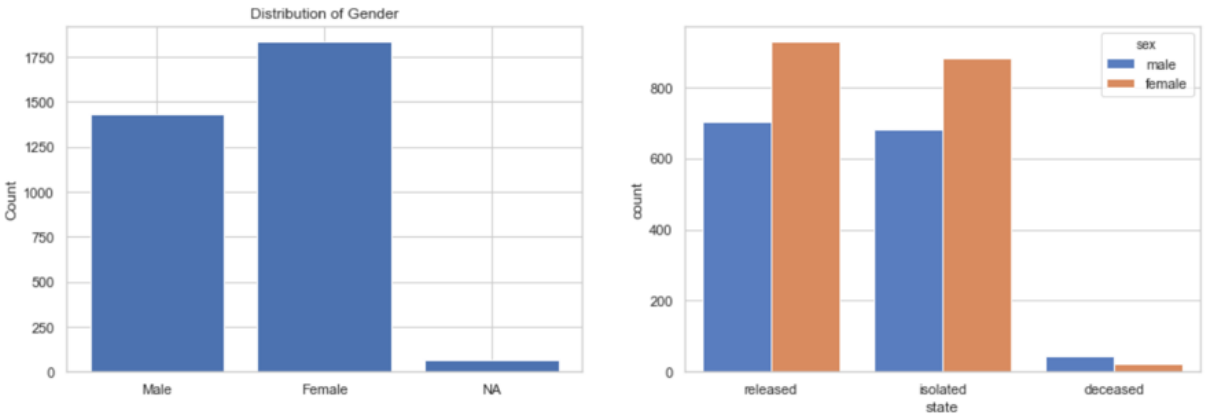

Figure 6: Distribution of Gender and Distribution of Patient Status based on Gender

We see that the number of male and female is generally comparable in the south korea dataset, which a slightly more female than male. Compared to China's and india's dataset, only around 2 percent of observations have gender missing. We also see the patients' current status based on gender, and see that the percentage of individuals that have survived and died is higher for men than women in South Korea.

Next we look at the distribution of `age`. In south korea's dataset, not only we have the age period information (i.e., the variable age), but also the birth year information (i.e. the variable birth year). In fact the age variable is redundant since we can calculate the actual age of patients directly from the birth year. Here we give the distribution of the actual age calculate by birth year, see Figure 7:

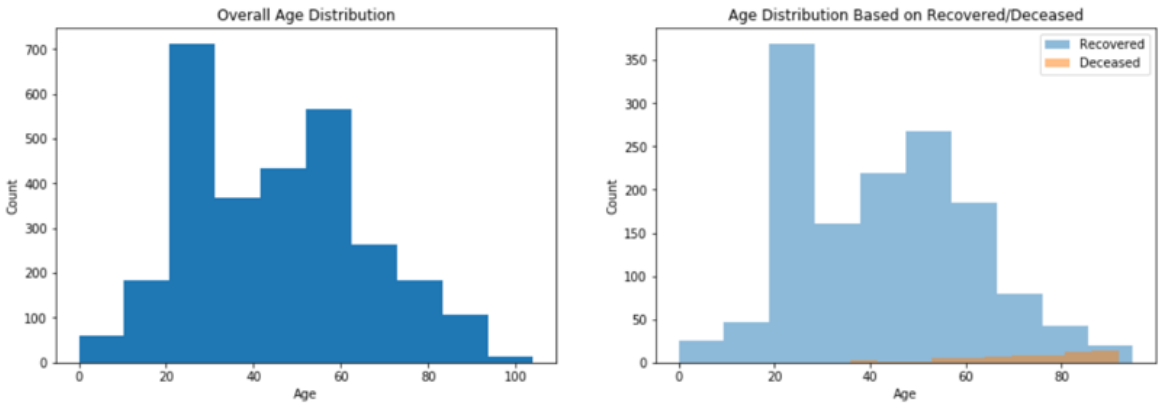

Figure 7: Overall Distribution of Age and Distribution for Recovered and Deceased Individuals

We see the overall distribution of age is symmetric, indicating that most of our observations have both young and elder people. For the right plot, we compare the age distribution for recovered individuals and deceased individuals, and observe the in south korea the majority of died patients are elder people. It seems the virus is more deadly for elder people in South Korea.

Next we take a look at the region where the dataset covers:

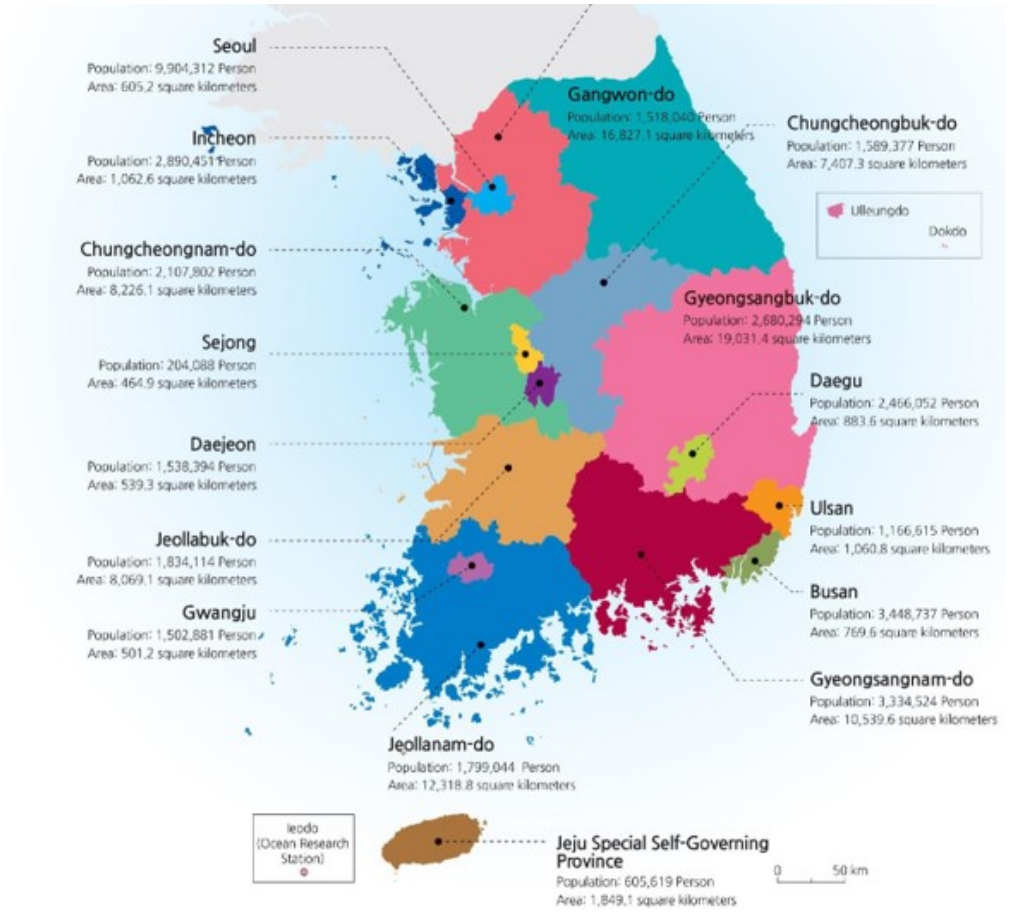

Figure 8: South Korea region map DS4C data description

Region includes province and city. Because all the region names are in Korean, we here are not providing a specific distribution. Instead we find the following information to be useful. In Korean, different suffix represents different divisions. i.e.

- -do: province
- -si: city
- -gun: county
- -gu: district

Finally we investigate on other features like Infection Order and Number of contacts.

Infection order is the order of infection, where 1 represents not infected by other patient in South Korea, 2 represents infected by 1st-order patient and so forth. The information should be useful while there seems to only be 30 data points among 3,300 patients that are not missing.

Number of contacts means how many patients the patient has contacted before isolated. It is orally reported by patient themselves. 0 means no contacts with other patients, 1 means contacting (and possibly infecting) 1 other person and so forth. The distribution of Num be of contacts is shwon in Figure 9:

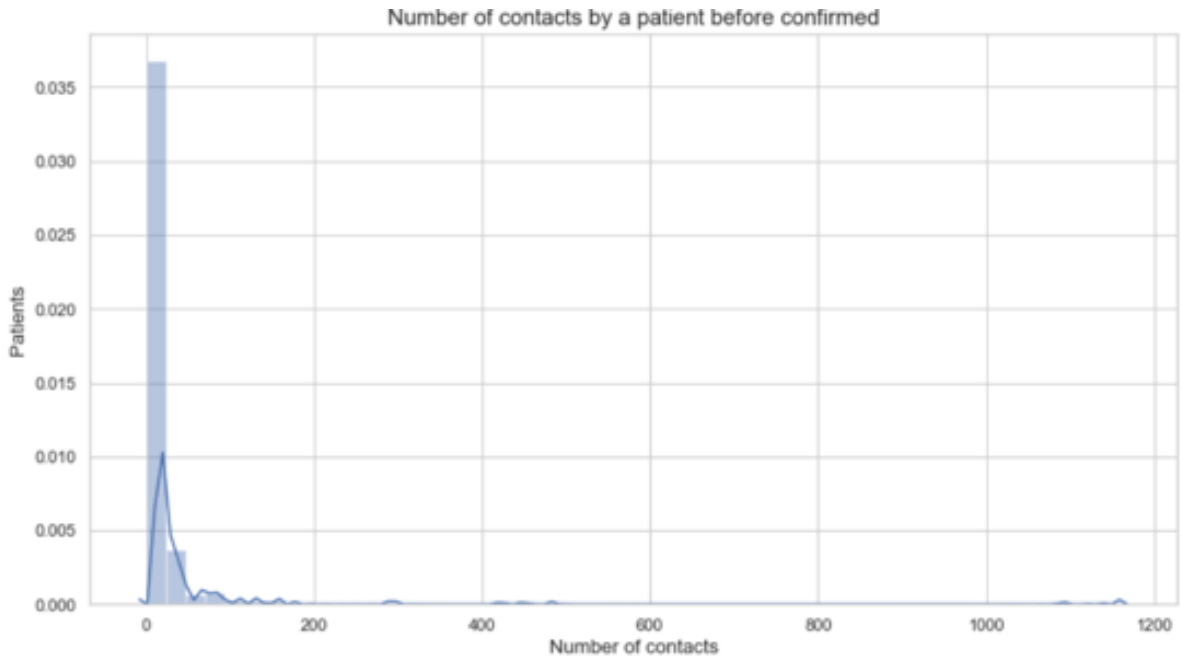

Figure 9: Distribution of number of contacts

The distribution is left-skewed, which means that majority of patients have less than 50 contacts. However we notice a few pretty large outliers that is greater than 1,000. These patients are often called super virus container. We will use the number of contacts in the further model.

3.2.2 Anonymity and Privacy

Next, we analyze the k-anonymity of the dataset. This dataset has one unique identifier which is patient ID, and also some quasi-identifiers including:

1. age(which is actually age period)
2. birth year
3. sex
4. city
5. province
6. country

One could potentially argue that some other features (eg. confirm date, number of contacts) could also be quasi-identifiers. However, the date that the virus was confirmed for an individual could only be linked to other datasets that this dataset is likely pulling from. One could potentially be linked to a hospital's database, but not everyone goes to the hospital after being confirmed. If you are generally a healthy person, you may be told to quarantine in your house instead of going to the hospital, which would make it difficult to identify an individual based on the date. We decided to exclude date from the quasi-identifiers.

Other features like Number of contacts is also not considered as quasi-identifiers because it is not supposed to be publicly known. It is hard for hackers to link those information with other databases.

While whether to include date as quasi-identifers is controversial, we notice that including it would likely not change the analysis much as there are a significant number of individuals confirmed each day.

After checking the original dataset, we find that the initial k-anonymity of South Korea dataset is also 1. This is not too surprising as this dataset was not processed for anonymity before published . The number of records not achieving given anonymity is shown as Figure 10

K = 1 samples	K = 2 samples	K = 3 samples	K = 4 samples	K = 5 samples
1687	262	64	32	14

Figure 10: Number of Records not Achieving Given Anonymity

Considering we only have around 3,300 samples, more than half of samples are unique, which indicate that the dataset could be unsafe without generalization.

We here have come up with two potential generalization:

- Generalize birth year to age period (10s, 20s, 30s, etc)
- Generalize city to province

The Number of Records not achieving given Anonymity after two steps generalization is shown in the Figure 11:

methods	K = 1 samples	K = 2 samples	K = 3 samples	K = 4 samples	K = 5 samples
Original	1687	262	64	32	14
Age period	567	217	118	71	34
Age period + province	52	29	29	21	14

Figure 11: Number of Records not Achieving Given Anonymity under different generalization

We can see that after first step generalization of age period we reduce unique samples by 70 percent, and by applying second step generalization we further reduces the number of unique samples by 90 percent, where we only ends up with 52 unique samples. We can then perform suppression by simply deleting rows to satisfy k-anonymity.

Often achieving k-anonymity is an "art". We can perform suppression in any of the stage: in the original dataset, after first generalization of age period and after two steps. Different methods incur different datasets, and will perform differently in a given task. Here we examine a given task in section 3.2.3 to compare the performance for different methods to achieve 3-anonymity.

To sum up, at this stage we get three datasets:

- Original dataset
- 3-anonymity dataset achieved by age period generalization + suppression
- 3-anonymity dataset achieved by age period generalization + province + suppression

We use these datasets for the further investigation.

3.2.3 Prediction

At this section, we want to compare methods to achieve k-anonymity for a designed task.

Instead of checking relationship between some variables in the China's dataset, since we have much more information in South Korea's dataset, we can perform far more meaning predictions. Two possible tasks are predicting deaths (classification) and predicting recovering time (regression).

Compared to performing regression, it is hard for us to predict deaths directly on South Korea's dataset: South Korea has a really low death rate (2 percent)! . This can be mainly contributed to the fact that South Korea is doing a great job in controlling the disease. While it is a good news for South Korea's people, such low death rate makes machine learning algorithms more difficult without using re-sampling methods.

As a result, we here compare methods on another easier task: predicting recover time. Formally, the recover time is defined as:

$$\text{Recover Time} = \text{release date} - \text{confirmed date} \tag{1}$$

The model we are using follows a typical setting as follows:

- Linear Regression
- One hot encoding for category variables
- 80/20 train test split
- Fill the missing values with mean

This is a classic setting for the regression problem (although not necessary to be the best). We mainly evaluate model by using the R^2 on the test set.

The result for three methods on this regression task is:

methods	Number of samples	Train R2	Test R2
Original	1218	0.229	0.036
Age period + suppression	872	0.155	0.009
Age period + province + suppression	1136	0.164	0.101

Figure 12: Comparison between three datasets on predicting recover time

We can tell the Figure 12 a few things:

- After anonymity, due to suppression, the number of samples reduces. The more generalization the more samples we can keep (while losing more variables).
- Even in the original dataset, while the training R^2 is 0.229, the R^2 on the test set is not high, with only 0.036.
- By applying age period generalization plus suppression we get a lower R^2 on both training set and test set.

- By applying age period generalization, province generalization plus suppression, while we got a lower training R^2, we in fact get a higher Test R^2, which indicates that the new datasets in fact helps the linear regression to predict recover time.

This last point is a good example that more anonymity does not necessarily brings lower accuracy.

3.2.4 Conclusion for data from South Korea

Thus, we can make the following conclusions for the South Korea dataset:

- Privacy needs to be focused on since there is some many unique samples based on quasi-identifiers
- An easy way to fix k anonymity is by applying generalization of age and province and then doing the suppression.
- After the processing the 3-anonymous dataset, we actually do a better job in predicting recover time than the original dataset.

3.3 India's Individual Dataset

The dataset for India is published at [1]. This website is an unofficial interactive website with several visualizations tracking the Coronavirus outbreak in India. The data is continually updated using "state bulletins and official handles" and is "validated by a group of volunteers and published into a Google sheet and an API" from the website's FAQ section. Raw datasets in JSON and CSV format are published at [2].

3.3.1 Dataset

We used the `death_and_recovered` dataset in CSV format, since we intended to model the probability of survival from the given attributes. The dataset contains the following attributes, with `Sl_No` as a unique serial number for each individual:

- `Sl_No`
- `Date`
- `Age Bracket`
- `Gender`
- `Patient_Status`
- `City`
- `District`
- `State`
- `Statecode`
- `Notes`
- `Nationality`
- `Source_1`
- `Source_2`
- `Source_3`
- `Patient_Number`

3.3.2 Data Cleaning and Exploratory Data Analysis

One issue we encountered with this dataset is that the second half of the rows in the CSV file had an extra comma, causing issues with reading and parsing the data. To fix this, we simply dropped the extra comma and wrote the data to a new file for easy access later.

We found many of the columns above were not needed. These were `Source_1`, `Source_2`, `Source_3`, `Notes`, `Patient_Number`, and `State` as it is redundant with `Statecode`. After dropping these columns, we were left with `Sl_No`, `Date`, `Age Bracket`, `Gender`, `Patient_Status`, `City`, `District`, `Statecode`, and `Nationality`.

The first thing we looked at was how much of the data was missing, which is shown in Figure 13.

Attribute	% Missing
Date	0.876
Age Bracket	93.813
Gender	91.478
Patient_Status	1.27
City	97.636
District	75.135
Statecode	1.284
Nationality	98.161

Figure 13: Percent of Data Missing for Each Attribute

We see that we have the most data in the `Date`, `Patient_Status`, and `Statecode` columns, but the others all have upwards of 75% of data missing. We will see how this affects both anonymity and prediction accuracy in the next sections.

Next, we do some exploratory data analysis to see the distributions of the variables in the dataset, as well as how different variables interact with each other.

First, we look at the distribution of `Patient_Status`. In a subsequent section we will be creating a model to predict this variable, so it is important to view the distribution of the values. The distribution is shown in Figure 14.

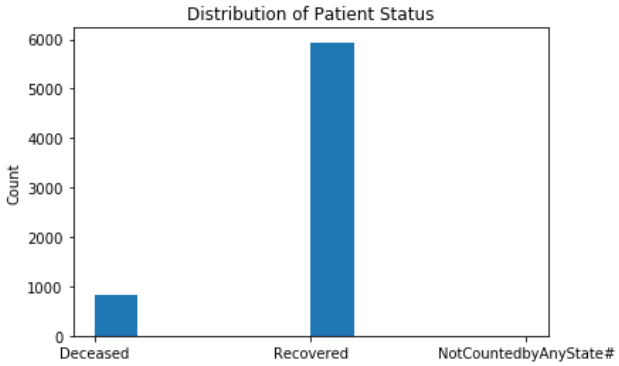

Figure 14: Distribution of `Patient_Status`

We see that this dataset is very imbalanced, as we have about six times as many recoveries as deaths. We would expect a distribution like this in most datasets that deal with a virus or sickness, as we would expect the majority of people that contract this virus to survive with a much smaller fraction dying from the virus. We have a strange third value of `NotCountedByAnyState#`. We did not know what this meant, and there were less than 5 rows with this value, so we just ignored them.

Next, we look at the distribution of confirmed cases in each state. This distribution is shown in Figure 15.

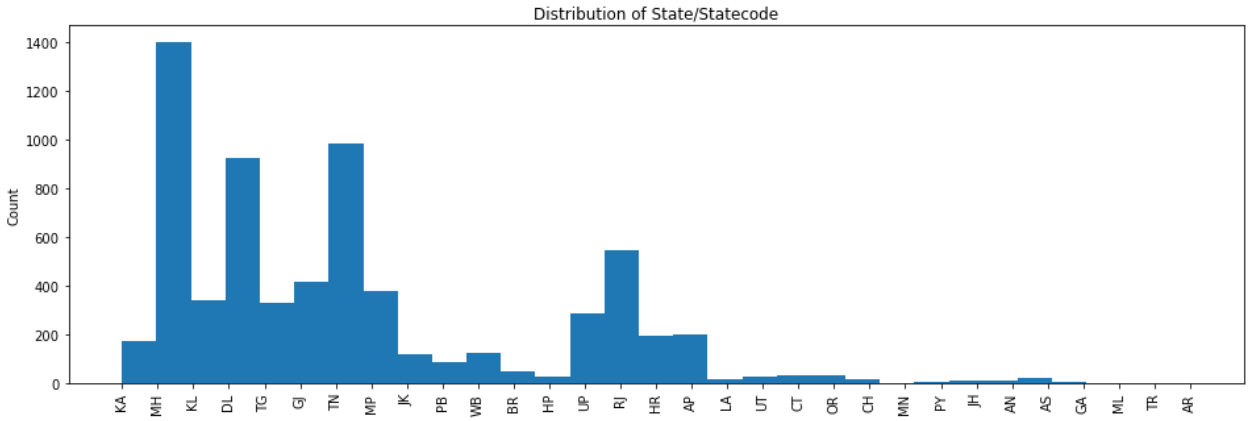

Figure 15: Distribution of Confirmed Cases per State

The state with the most confirmed cases is MH, or Maharashtra, which is the second most populated state in India. The most populated is UP, or Uttar Pradesh. We see near the middle of the plot, but it has much fewer confirmed cases than several of the other states. This could be due to several reasons. There could be less data available for this state, testing could be less widely available there than other states, or this state could have simply done a much better job containing the outbreak than other states.

Next, we look at the distribution of `Age Bracket`, which is just the age of each individual in years. It is important to keep in mind that age is missing for 93% of individuals. However, we still see some interesting trends, which we see in Figure 16.

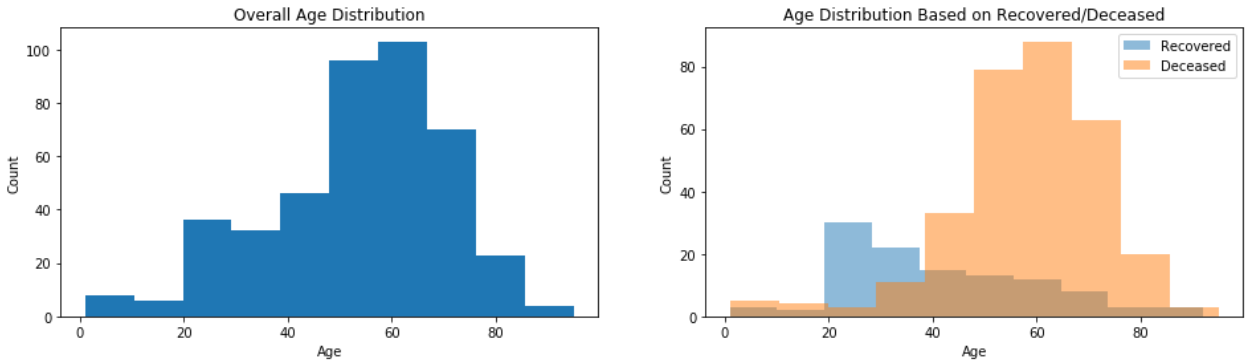

Figure 16: Overall Distribution of Age, and Distribution for Recovered and Deceased Individuals

We see the overall distribution of age is left-skewed, indicating that most of our observations that we have the data for are older individuals. In the plot on the right, we see that for the observations where we have age, the majority of these are individuals that died. We see that the virus does seem to be more deadly for older people.

Finally, we look at the distribution of gender, which we see in Figure 17.

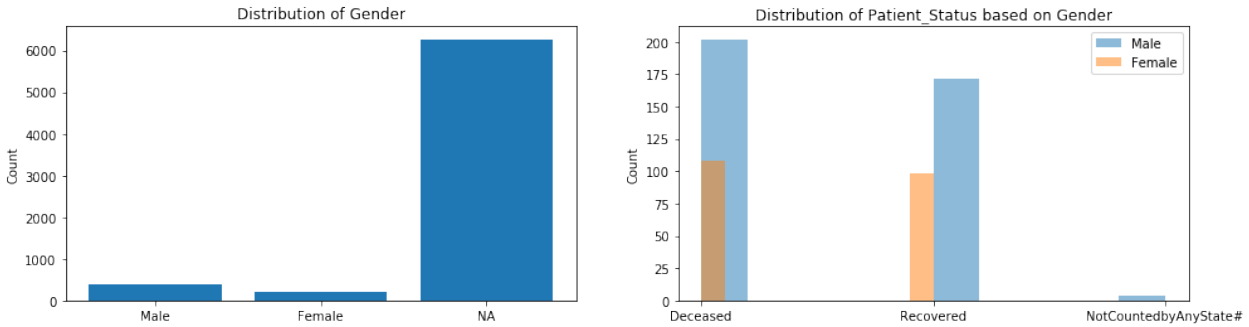

Figure 17: Distribution of Gender and Distribution of Patient Status based on Gender

We see that we have about twice as many men as women, but we see that the overwhelming majority of the observations have gender missing. We also look at the distribution of patient status based on gender, and see that the percentage of individuals that have survived and died are pretty similar for men and women.

3.3.3 Privacy and k-Anonymity

A dataset being k-anonymous means that for every row in the dataset, there are at least $k-1$ other rows that are indistinguishable from this row, with respect to some set of quasi-identifiers that could be used to identify this individual to another dataset.

With this dataset, we identified the following set of quasi-identifiers:

- `Age Bracket`
- `Gender`
- `City`
- `District`
- `Statecode`
- `Nationality`

One could argue that `Date` is a quasi-identifier. However, the date that the virus was confirmed for an individual could only be linked to other datasets that this dataset is likely pulling from. One could potentially be linked to a hospital's database, but not everyone goes to the hospital after being confirmed. If you are generally a healthy person, you may be told to quarantine in your house instead of going to the hospital, which would make it difficult to identify an individual based on the date. We decided to exclude date from the quasi-identifiers, but including it would likely not change the analysis much as there are a significant number of individuals confirmed each day, so there are lots of observations for each unique date.

Unsurprisingly, this dataset is initially 1-anonymous. Below are examples of some of the most and least anonymous rows in the dataset.

Age	Gender	City	District	Statecode	Nationality	*k*-Anonymity
NaN	NaN	NaN	NaN	MH	NaN	1078
NaN	NaN	NaN	NaN	TN	NaN	961
NaN	NaN	NaN	NaN	DL	NaN	921
NaN	NaN	NaN	NaN	RJ	NaN	488
NaN	NaN	NaN	NaN	TG	NaN	320
NaN	NaN	NaN	NaN	MP	NaN	304
NaN	NaN	NaN	NaN	UP	NaN	237
NaN	NaN	NaN	NaN	KA	NaN	151
NaN	NaN	NaN	Mumbai	MH	NaN	149
NaN	NaN	NaN	Kasaragod	KL	NaN	142

(a) Most Anonymous Observations

Age	Gender	City	District	Statecode	Nationality	*k*-Anonymity
85.0	M	Mumbai	Mumbai	MH	NaN	1
80.0	M	Mumbai	Mumbai	MH	NaN	1
86.0	F	Ghatkopar	Mumbai Suburban	MH	NaN	1
45.0	M	NaN	Buldana	MH	NaN	1
74.0	M	Hyderabad	Hyderabad	TG	NaN	1
67.0	M	Surat	Surat	GJ	NaN	1
85.0	F	Ahmadabad	Ahmadabad	GJ	NaN	1
70.0	M	Bhavnagar	Bhavnagar	GJ	NaN	1
46.0	F	Ahmadabad	Ahmadabad	GJ	NaN	1
47.0	M	Ahmadabad	Ahmadabad	GJ	NaN	1

(b) Least Anonymous Observations

Figure 18: Most and Least Anonymous Observations in India Dataset

As we can see, the sets of quasi-identifiers that are most anonymous are the ones that have a lot of missing data, and the least anonymous are the ones that have little missing data. This tells us that with respect to anonymity, missing data is a good thing. In Figure 19, we see the number of observations that are at or below 5-anonymous.

1-anonymous	2-anonymous	3-anonymous	4-anonymous	5-anonymous
417	56	24	16	12

Figure 19: Number of Observations at or below 5-Anonymous

We have almost 7000 observations in the dataset, but only about 500 of them are at or below 5-anonymous. Because of this, to make the dataset 3-, 4-, or 5-anonymous, we decided to use suppression and simply drop these rows. This will only result in a loss of about 7% of the dataset. The full dataset has 6853 observations, the 3-anonymous dataset has 6252, the 4-anonymous dataset has 6188, and the 5-anonymous dataset has 6128.

Next, we would like to see how the distributions of variables changes after making the dataset more anonymous. We compare the distribution for the age and gender variables between the original and the 3-anonymous dataset in Figure 20.

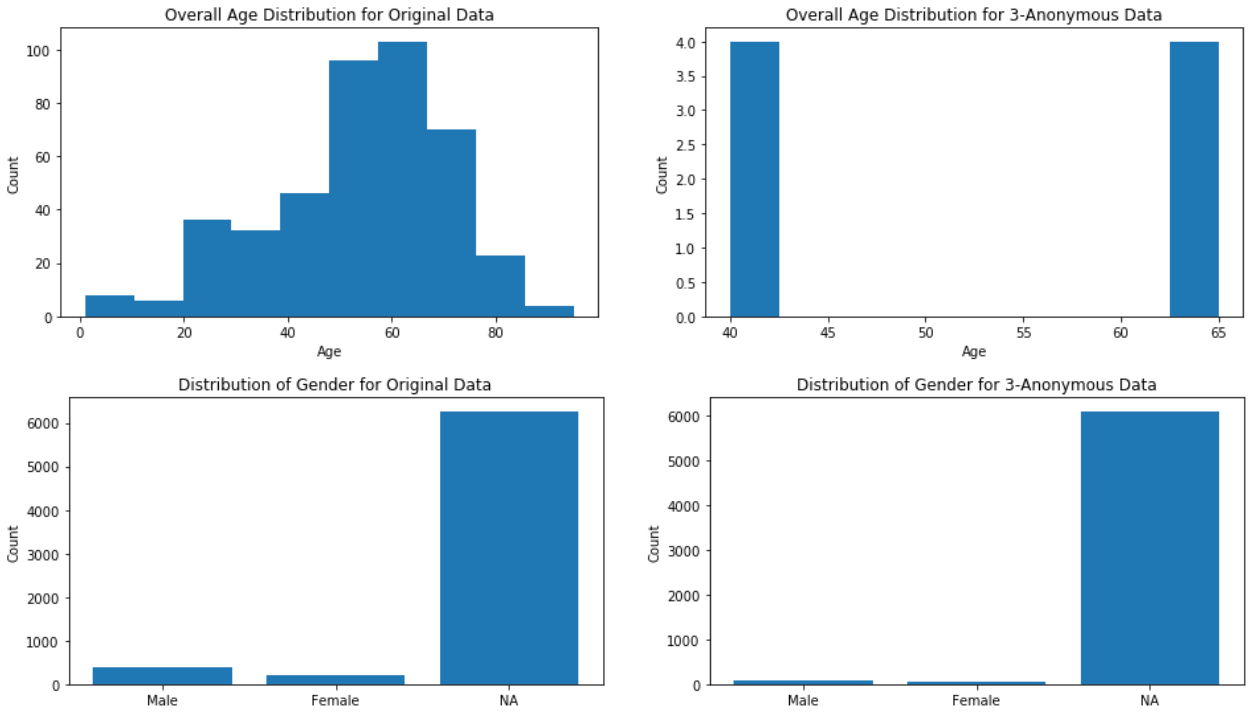

Figure 20: Comparison of Age and Gender Variables Between Original and 3-Anonymous Datasets

We still have about twice as many men as women in the 3-anonymous dataset and a lot of missing data, so this does not change much. However, in the 3-anonymous dataset, we see that we only have two distinct values of age across eight observations. Also, as we go to 4- and 5-anonimity, we completely lose all useful values of Age and are only left with missing values. We will see how this affects our analysis in the next section.

3.3.4 Prediction

We wanted to test how viable these datasets still were for some type of prediction. To evaluate the 3-, 4-, and 5-anonymous datasets versus the full dataset, we followed the following process. We first split the full dataset into a training and test set, with the test set size of 20%. We then anonymize the training set to 3-, 4-, and 5-anonymous levels. For the full training set and the anonymized training sets, we train a logistic regression model on the data and evaluate it on our earlier held-out test set.

By default, most Python libraries drop rows that contain missing data. This would result in nearly all of our rows being dropped, so we had to develop some method to fix this. For quantitative variables, we imputed the mean and created a new binary column indicating whether the original variable was missing or not. If the entire column was missing, like in the case of age with the 4- and 5-anonymous data, we just dropped the whole column. For categorical variables, we simply created a new category for missing data. This allowed us to use all of the data we had at hand to make predictions.

Since the response variable is so imbalanced, we look at the percentage of observations that are recoveries as a baseline. About 86% of the cases in the dataset are recoveries, so we will try to do better than this. In Figure 21, we see the accuracy and the confusion matrix for each of the four models.

Dataset	Train Accuracy	Test Accuracy
Full	95%	90.3%
3-Anonymous	96.2%	85.9%
4-Anonymous	96.2%	86.6%
5-Anonymous	96.2%	86.9%

Full	Pred. Died	Pred. Lived
True Died	68	110
True Lived	21	1155
3-Anon	Pred. Died	Pred. Lived
True Died	0	178
True Lived	13	1163
4-Anon	Pred. Died	Pred. Lived
True Died	0	178
True Lived	3	1173
5-Anon	Pred. Died	Pred. Lived
True Died	0	178
True Lived	0	1176

Figure 21: Performance Metrics for Logistic Regression on Four Datasets

We see that we do get upwards of 85% accuracy with all of the models, but it is more useful to look at the confusion matrices. With the full dataset, we do an okay job predicting survival or death. However, when we use the anonymized data, we start to predict more and more survivals, and for the 5-anonymous dataset, we only predict survivals. We clearly see a problem. If we wanted to use a dataset as training data for a model, in terms of accuracy, we would much rather have the full dataset than an anonymized dataset. This is in part due to the dataset being very imbalanced, but it is also due to the loss of data by making this dataset anonymous, since we see that we can do better with the exact same model by using the full training set for training. We could also potentially use date as a predictor, using something like the number of days since the first confirmed case. We would hope that our ability to treat a person gets better over time, so the probability of death would hopefully decrease as time goes on. This would be interesting to look at in future analyses.

3.3.5 Conclusion

This India dataset allowed for some interesting analysis both in terms of anonymity and prediction. Since this dataset is unofficial and is obtained using public records, there is a lot of missing data for most columns. This makes it easier to anonymize the dataset, since we only have a few hundred rows that we have to drop to achieve 3-anonymity or better. However, when we drop these rows, we drop a lot of the useful information contained in this dataset, like age. Because there is so much missing data, there are not many unique age values, so if we say that we have an individual that has a certain age, we can likely uniquely identify that individual in the dataset. As we make the dataset more anonymous, we lose analytic and predictive power. We only tried logistic regression, one of the simplest classification models, but we can clearly see the decrease in performance as we increase the anonymity of the dataset. With a more anonymous dataset, it will be harder to create a useful model. There are several factors involved with this that we could have tried, like data balancing using oversampling or undersampling, or using more complex models than only logistic regression, but we clearly see the difficulty in effectively modeling something using an anonymous dataset versus the full dataset.

4 Policy Review

The comparsion of different dataset is shown in Figure 22.

	China	South Korea	India
Sample size	42,334	3,364	6,853
Death rate	19.84%	1.9% (67/3364)	12.1% (824/5938)
Missing data proportion (sex)	95%	1.9%	91%
Quasi-identifier	Age, gender, province, city, admin	Birth year, Age, gender, country, province, city	Age, gender, city, district, state, nationality
Original k-anonymity	1	1	1
Accuracy change after anonymity		Increase	Decrease

Figure 22: Summary of comparison of three datasets

In addition to our technical analysis of these datasets, we wanted to review the privacy-related and data-related policies in each of these three countries and see how this affected our interpretation of these datasets and the results we found.

In China, the information about confirmed cases are generally not publicly available. While the numbers of confirmed cases in different regions in China are reported in daily bases, other information such as gender and age of patients is not regularly reported. Such information can only be found through some reports published by official media. Though the number of deaths and discharges is also reported everyday, it is almost impossible to associate these outcomes to the confirmed individuals, and we thus have so many missing data in our dataset. This situation corresponds to the fact that almost all activities about COVID-19 are somewhat under organization of the Chinese government. Essential information and discovery is usually summarized before published to public. Researchers would need to either work on their own to gain patient information or request cooperating with the government, if their requests are approved.

In South Korea, all of the COVID-19 data come directly from the government specified agency (i.e. KCDC). Unlike the data for research in China and unofficial website data for the India, the data in South Korea has far fewer missing data. This is due to the fact that one of the purpose of South Korea government is to announce the information of COVID-19 "quickly and transparently" [3]. This can be achieved since south korea is smaller than China and US and thus easier to collect data. This fast and transparency help south Korea to become a place where the death rate is among the lowest throughout the world. However, privacy seems to be big issue for South Korea. Our analysis suggests that by applying simple generalization the KCDC can reduces huge number of unique samples and thus protect privacy.

In India, there seems to be pretty strict data protection policy. A 2019 Harvard Business Review article [5] outlines several aspects of India's data protection policies. The Supreme Court of India passed a bill in 2017 stating that privacy is an individual right of all citizens. In addition, a company must obtain explicit consent from an individual before collecting their data. Finally, it also states that the data provider is the owner of the data. This could explain why there is so much missing data in the dataset. With these strict standards, it is harder to find public record data at the individual-level relating to COVID-19. If an individual does not give consent for their age or location to be recorded, then there is no way for a third party website like CODIV-19 India to obtain this data. The requirement for a company to obtain explicit consent likely contributed to a lot of the missing data. If an individual is asked whether or not they would like for their data to be recorded, the simple response is no, since you don't know what this company may do with your data, and when asked explicitly and given the option, you may value your privacy. It seems like these strict standards contributed to a lot of the missing data found in the India dataset.

5 Future Work

A study like this on data relating to a new virus yields numerous directions of future work. One aspect would be to look for other individual-level datasets. In most countries the data is only available at very generalized levels like county, state, or country level, so there is little worry of anonymity. However, individual-level datasets have the opportunity for privacy breaches. HIPAA only requires the removal of directory information, like name, address, or SSN, but says nothing about other aspects of anonymity like k-anonymity. If there are individual-level datasets for other countries, it would be interesting to see the level of anonymity that these datasets preserve or what is required to reach 3-anonymity or more.

In [6], we saw Sweeney take a similar study a step further and actually reidentify individuals like then-governer of Massachusetts William Weld. It would be interesting to attempt this with these datasets in a non-malicious way. It is one thing to talk about theoretical privacy offered by a dataset but taking it to the next level and actually attempting to identify an individual based on their quasi-identifiers gives a new sense of importance to a study. This is one of the reasons Sweeney's paper was important, as she was able to identify not just anyone, but the governer of Massachusetts.

There are other methods of anonymization that we did not try in this study. In the India dataset, for example, we only use suppression since less than 10% of the data needed to be removed to achieve even 5-anonymity. However, this resulted in the loss of a lot of important data that potentially could have been preserved had we used something like generalization or a combination of the two. It would be interesting to see how many rows would have to be suppressed after doing some sort of generalization.

It would be interesting to try other models for prediction. In this study, we only used regression models, but there is a myriad of other models that could apply to this data and potentially be able to perform significantly better than the models we applied could. A model that can predict the likelihood of survival with high accuracy would be something very useful, but it isn't exactly sure clear it would be best applied. If an individual has a 99% chance of death, should you dedicate more resources to this individual as they need more help, or fewer resources since they will likely die anyway and it is important to save limited resources for other patients? On the other hand, if an individual has a 20% chance of death, should we just send them home because they will probably be okay anyway? Sending all low risk patients home could cause more of them to die because some of them actually needed care but our model said they didn't. It seems inhumane to deny an individual resources because they are probably going to die, but in a global pandemic where resources are running out quickly, hard decisions like this might have to be made. However, it also seems even less humane to let a computer make these decisions. A study of this would relate much more heavily to ethics than the technical details of statistical modeling and anonymity.

Finally, since COVID-19 is an ongoing pandemic, we may need to simply wait for more data. There are thousands of new confirmed cases each day which creates more and more data for analysis. At this point, the virus is still growing very quickly, so even in just a week we will have significantly more data than we have now that could yield new insights.

6 Conclusion

COVID-19 is the first worldwide pandemic in the current era of computing that we live in where computers are fast and data is abundant. This reason alone is enough to study datasets surrounding the virus. It is a lot harder to computationally study something like the Black Death, which occured in the 1300s, or even something more recent like flu outbreaks in the 1900s, since computers either did not exist or were very new, and data collection meant writing something in a book rather than storing something in a database.

With all of this data that is being generated, we must raise the question of privacy. This question is even more important during times of a worldwide crisis where there are significant trade offs to both sides. On one hand, privacy and anonymity is crucial to maintain as an individual that is confirmed to have the virus could be targeted by hate groups, discriminated against in future job or school applications, or just simply embarrassed that they got sick with the virus. On the other hand, there are thousands of people dying every day from this virus. This makes it even more important that researchers are able to obtain accurate and thorough data that they can study. The difference between finding a vaccine in the next month or in

two years could simply be the difference between one column being included in a dataset and a researcher discovering a trend. Where we draw the line is a really difficult question, and much more difficult than with other more 'harmless' data like an individual's social media data or browser history. It is important to ask these same questions for this type of data, but it does not carry the same weight as data relating to a current pandemic does, where people are suffering and dying every day.

In this study, we saw that the three datasets we studied were all HIPAA compliant, but were only 1-anonymous initially. Two of our datasets were riddled with missing data, which was good for anonymity and meant that we only had to drop a small percentage of the dataset, but meant that dropping these individuals dropped important information that could be used later for analysis. Removing this data has the potential to hinder future analysis as it changes trends that were present in the original dataset.

It will be interesting to see how data is handled relating to the virus in the coming weeks and months. Data-collecting agencies like testing labs must ask themselves important questions like how much data to release to both protect individual-level privacy but to give researchers enough data to be useful. Hopefully an optimal point can be found where both of these are preserved.

References

[1] COVID19 India - Coronavirus Outbreak in India. https://www.covid19india.org/. Accessed: April 26, 2020.

[2] COVID19 India API. https://api.covid19india.org/. Accessed: April 26, 2020.

[3] Data Science for COVID-19 (DS4C). https://www.kaggle.com/kimjihoo/coronavirusdataset. Accessed: April 26, 2020.

[4] Epidemiological data from the COVID-19 Outbreak, real-time case information. https://github.com/beoutbreakprepared/nCoV2019. Accessed: April 26, 2020.

[5] How India Plans to Protect Consumer Data. https://hbr.org/2019/12/how-india-plans-to-protect-consumer-data.

[6] L. Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.