Privacy and Bias analysis of MIMIC-III dataset

https://github.com/PrivacyAnalysis/Privacy_Analysis

Yujie Cai, Jiahui Tang, Xin Zeng

April 30, 2021

1 Introduction

Medical and health care are central to human beings. Medical datasets contain abundant information that professionals can utilize for their research. Once these datasets become available to the public, they draw more attention from the scholars of other domains, potentially generate more useful results, and maybe embed more problems while being used. To some extent, ethical issues will stem from inappropriate usage of the data. For instance, the collection and usage of detailed information about patients increases the risk of invading individuals' right of privacy. Later on, it will result in mistrust between patients and medical institutions, and thus hamper the progress of research. Another risk could arise from the inaccurate medical studies' findings, such as using a biased model to predict outcomes, as the resulting bias may create potential health disparities between groups of people. Therefore, we are incentivized to scrutinize those medical datasets first, to examine the need and effectiveness of de-identification, and further what kind of biases will be harmful if ignored when using the dataset.

From all of the publicly available medical datasets, we decide to analyze on Multiparameter Intelligent Monitoring in Intensive Care ver.3 (MIMIC - III). There are not many medical datasets containing person-level information that have been made publicly accessible (for example, in Kaggle.com, healthcare data are mostly aggregated at city or country -level). Among the ones with personal-level information, MIMIC is one of the most comprehensive ones documents patients' medical-related data as well as demographic information. In fact, it has records for over 40,000 patients who stayed in critical care units of Beth Israel Deaconess Medical Center between 2001 and 2012. We believe that analyzing such a huge database is somewhat illuminating for us who always think about how to deal with data and information properly. The coverage of wide and detailed information in MIMIC enables us to build models to do prediction. However, we shall consider the privacy protection of patients and be responsible for bias resulting from using data and building models. Then, we are able to make our project goal clearer: **we want to first analyze the privacy issues in the dataset. Central questions are 1)** what are the consequences and evaluation of already-implemented de-identification procedures? **2)** What are the consequences and evaluation of other de-identification procedures that we can apply on the dataset? **Secondly, we will go through the model-building process to predict mortality (a common prediction in the academic area). Central questions are 3)** what kind of bias should we take into account? **4)**How can we mitigate them and what are the consequences of the mitigation?

We hope to gain a better understanding of the dataset through critically answering these questions. If possible, we will share our thoughts on how to deal with this kind of medical dataset in general. We managed to achieve most of our goals. We concluded that there exists concerns of mis-representation of age variables and difficulties in conducting longitudinal studies, and there's trade off among accuracy, information, and privacy, with the fact that we could achieve higher K-anonymity and L-diversity by applying row suppression, adding synthetic records, column suppression (blurring), and generalization. In terms of modeling bias, historical and representation bias are the major concerns in this dataset, with attributes such as ethnicity and insurance type being the primary sources of bias. LightGBM model provides highest accuracy. In terms of mitigation of bias, two approaches are used to mitigate bias in the model, one is to tune with the model and the other is to change the threshold of classification. Further details will be provided in the following sections.

The report is organized in the following manner: section II presents the retrieval, usage, and composition of the dataset we process through this project; section III discusses the privacy and de-identification of the dataset, and attempts to answer question 1) and 2); section IV analyzes the embedded bias, and attempts to address question 3) and 4); section V serves as conclusion and the direction of future work.

2 Data retrieval, usage, and organization

Retrieval

The data is published at `https://mimic.physionet.org`. From the website, we know the database includes information such as demographics, vital sign measurements made at the bedside (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors:

- it is freely available to researchers worldwide
- it encompasses a diverse and very large population of ICU patients
- it contains high temporal resolution data including lab results, electronic documentation, and bedside monitor trends and waveforms.

To get the full access to the database, we need to go through a formal application process. we complete the CITI "Data or Specimens Only Research" course (available at `https://www.citiprogram.org/`) and submit an application with a school reference to the data provider.

Usage

First, we claim that we use this dataset fully academically, and specifically for our AC 221 final project. In our project, we are using only the subsets of the data since our project is a preliminary try-out of dealing with some hard questions. The detailed usage is under the Organization section.

Organization

In the user guide [1], we notice that the dataset has the structure in the graph:

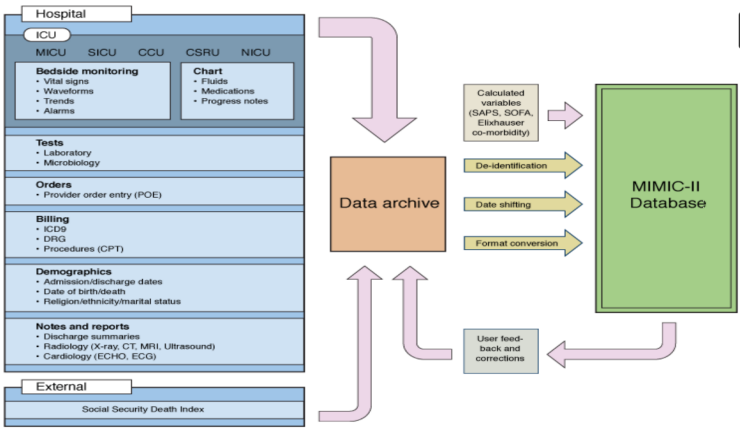


Figure 1: dataset content

We can see that the dataset contains abundant information, and go through de-identification, date shifting, format conversion. The demographics information is only a small part of the huge database. But already, analyzing them will give us much insight. A more detailed plot shows the relationship and linkage between the subsets of the dataset.

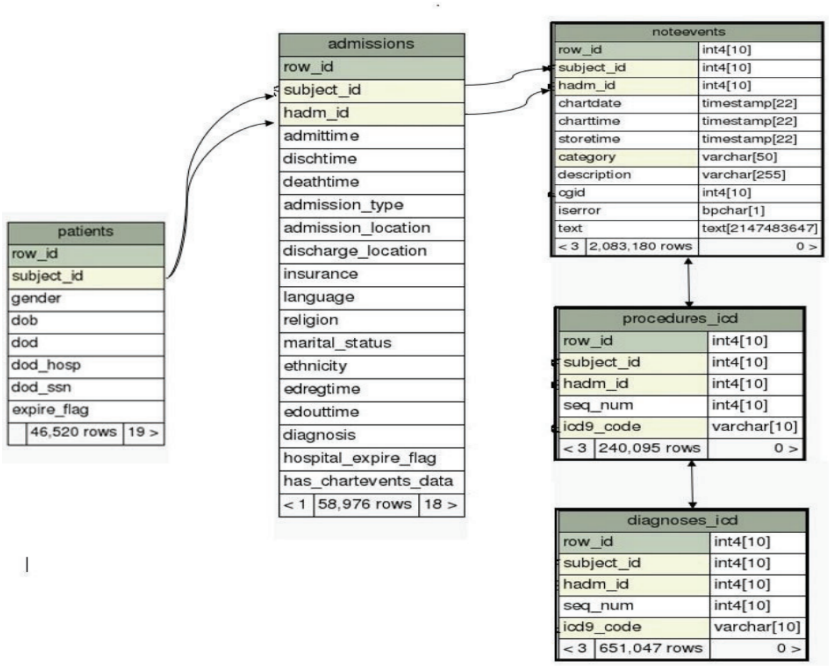


Figure 2: dataset schema

Using the schema [2] we shall be able to find which subsets of data will go into our later analysis on privacy and bias. Referring to the user guide, MIMIC datasets can be roughly divided into two kinds: instructional and records. The instructional works like legend, linking the code of diagnosis, drugs, etc. to professional nouns. Records are what we care about here. They record patients' data and usually begin with different ids, such as *row_id* and *subject_id*. The records can be divided further: static record, meaning each row is for one patient; time series record, meaning the records correspond to a single patient in different time intervals (like laboratory tests for different days).

We have several criteria of choosing our analyzable datasets. First, the dataset must contain enough personal or demographic information linking to the individual patient. Then, these information cannot be too specific, for example, the prescription text or notes the doctor made before or in the process of ICU admission (although de-identified). Our first criterion helps us omit dataset such as SERVICES and TRANSFERS. Our second criterion excludes the usage of NOTEEVENTS, and any time series related LABEVENTS, or MICROBIOLOGYEVENTS. In the end, we will combine ADMISSION, PATIENTS, and ICUSTAYS. As for linkage, we will use unique SUBJECT_ID to merge these datasets.

3 Privacy and De-identification

Central Question (1)

what are the consequences and evaluation of already-implemented de-identification procedures?

We are acknowledged that from the website, several de-identification has been done to protect patients' identities. From the user guide [1], we know:

- all dates were shifted into future
- the shifting of a given patient was uniform, preserving the time gaps
- The day of the week and the season of the year were preserved
- the shifting was randomly assigned
- All HIPAA-defined types of personal health information (PHI) were removed from the text, plus care-giver and hospital-specific identifiers.

We agree that the removal of HIPAA-defined PHI follows the conventional principle of dealing with healthcare related information. Admittedly it will hinder some kind of over-detailed analysis (for example, distribution of some variable by region based on zip codes), but we think this method effectively protects patients and will not

hinder our modeling and analysis.

The date shift is a particular feature in this dataset. The example is shown in the graph.

	SUBJECT_ID	ADMITTIME	DISCHTIME
0	22	2196-04-09 12:26:00	2196-04-10 15:54:00
1	23	2153-09-03 07:15:00	2153-09-08 19:10:00
2	23	2157-10-18 19:34:00	2157-10-25 14:00:00

Figure 3: date shift example

Here the year cannot be identified and reverse-engineered to any present time we know, because the shifting amount is randomly assigned to each patient but preserving time gaps for a given patient. We are not provided with a specific distribution of shifting amounts, so we are not able to perform reverse-engineering. In some ways, the date shift is an effective means of privacy protection since the time stamp contains so much information. However, by observing the distribution of age variables, we identified two concerns with respect to the shifting. Some outliers inevitably occur, because elder people with random, big shifts will result in extreme shifts. They will be 200+ years-old. This provides some inconvenience for users to pre-process data. Also, the analysis of any temporal, longitudinal studies involving changes in patient care practices over time cannot be supported by the fully de-identified data.

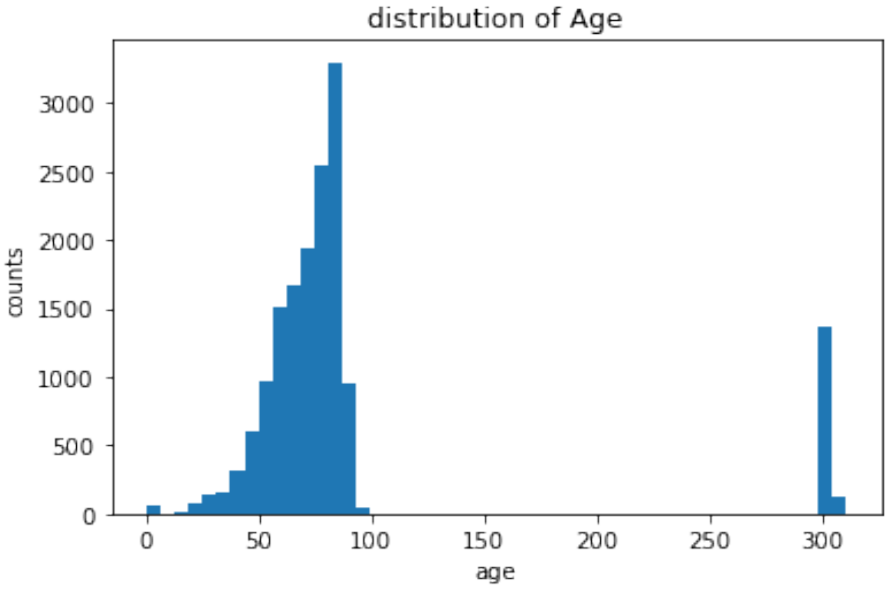


Figure 4: age distribution

To sum up, the already existing implementation of de-identification efficiently protects patients' privacy but it results in i) mis-representation of some people's age ii) difficulties in longitudinal studies.

Central Question (2)

What are the consequences and evaluation of other de-identification procedures that we can apply on the dataset?

The dataset we used for analysis has 116,426 rows of records (over 40,000 patients but some patients can be admitted into ICU for several times and thus have multiple records). However, the dataset starts with 1-anonymity, meaning all rows contain unique information in some columns. As learned in this course, we want to apply de-identification techniques, such as row suppression, adding synthetic records, column suppression (blurring), and generalization to raise the k-anonymity measurement of the dataset and achieve a higher privacy level.

First, we create a list of quasi-identifiers. The working definition of quasi-identifiers is pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create an unique identifier. When combined, a quasi identifier list could become personally identifying information. Our list contains 15 variables:

Table 1: quasi-identifiers list

Name of Quasi Identifier	Meaning
ADMISSION_TYPE	Type of admission to the hospital (eg. emergency)
ADMISSION_LOCATION	Location of admission to the hospital (eg. emergency room)
DISCHARGE_LOCATION	Location of discharge from the hospital (eg. home healthcare or death)
INSURANCE	Insurance coverage (eg. Medicare, self-pay)
LANGUAGE	Language spoken (eg. English)
RELIGION	Religion (eg. Catholic)
MARITAL_STATUS	Marital status (eg. single)
ETHNICITY	Ethinic group belongs to (eg. white, black, hispanic
DIAGNOSIS	Diagnosis (eg. acute coronary syndrome)
GENDER	Gender (male, female)
FIRST_CAREUNIT	First care unit been to(eg. cardiac surgery (CSRU) care units)
LAST_CAREUNIT	Last careunit been to (eg. cardiac surgery (CSRU) care units)
FIRST_WARDID	First wardid served (eg. wardid #1)
LAST_WARDID	Last wardid served (eg. wardid #1)
LOS	Length of stay in CU, in day

The list doesn't include any timestamp as all the dates in MIMIC-3 dataset has been moved to future dates, and the overall distribution is not preserved. Then, we could use the quasi-identifiers from our list to analyze the implementation of different de-identification procedures.

K-anonymity

K-anonymity means k rows have the same information and thus cannot be uniquely identified. A dataset achieves K-anonymity if the information for each individual contained in the dataset cannot be distinguished from at least K-1 individuals whose information also appear in the dataset [3]. Using computer programming (please see appendix of code), we achieve 3- and 5-anonymity by dropping rows or adding synthetic records. To achieve 3-anonymity by dropping rows, only 3265 (3%) records of the original dataset remain; if adding synthetic records, we have to add 219,899 (2x of the original) records. To achieve 5-anonymity, only 1850 (1.5%) records of the original dataset remain; if adding synthetic records, we have to add 442,714 (3.5x of the original) records.

Simply applying row suppression and records adding gives us the feeling of the central consequence of de-identification: the tradeoffs among accuracy, information and privacy. We have to remove a huge amount of data in order to realize a certain level of anonymity, and we also have to "fake" a huge amount of records to achieve this goal. The row suppression method significantly reduces the size of dataset, and may impact the accuracy of statistical metrics (eg. means, variance of certain variables). Adding synthetic records is nearly impossible if we cannot have a highly-automatic algorithm. Duplicating the dataset is time and energy consuming and maybe gives rise to accuracy problems as well.

Blurring and Generalization

Another way to de-identify a dataset is to apply column selection and engineering by algorithms called blurring and generalization. That is, we can delete columns and group the content of columns. Here. We take a look at all variables, and check to see whether we should generalize for numerical variables, or delete some of the categorical variables, to achieve a higher level of K-anonymity.

It turns out almost all variables in quasi-list categorical columns have only 0-1 values with less than 5 counts of the same data records. It means looking at the columns individually, all of them achieved a good level of anonymity, except for two variables: LANGUAGE and DIAGNOSIS. And we also find one numerical column

that could be further generalized, which is LOS. Therefore, we conducted a naive search to see what we could do on each of these three variables, and how many records we need to drop to reach 5-anonymity. If we apply blurring on the DIAGNOSIS column, we need to drop 106,281 (half of the dataset) records; if we apply blurring on the LANGUAGE column, we need to remove 114,541 records. But if we generalize the LOS column, we need to remove 99,014 records, which is the least number of drops.

Our naive search algorithm can actually take two arguments: number of columns and the actual column we want to remove. Sometimes from a purely technical procedure, we can use the searching algorithm towards a selection with minimal distortion of the original dataset. However, this specific selection might lose columns that we need for future analysis. For example, if INSURANCE is not included in this selection, but some sociology and business researchers want to determine the relationship between a variable and the insurance coverage. They will not be able to do so, and they have no means to link this dataset to the other.

Another problem is that we question on the generalizability of the procedure. We can hardly think of a highly general approach or a piece of code that can be encapsulated then apply to other datasets. Even for this particular MIMIC-3 dataset, it is hard to use a universal approach to satisfy all groups of researchers. Although Angiuli and Waldo (2016) [4] suggest that there are promising futures of the combination of suppression and generalization, we think this problem needs further research from practice.

L-diversity

L-diversity is an extension to the K-anonymity model that is used to preserve privacy in data sets by reducing the granularity of a data representation, using techniques such as generalizations, blurring, and suppression. A dataset achieves L-diversity satisfying the condition that any given record in the dataset could be mapped onto at least L-1 other records in the data. The L-diversity model adds the promotion of intra-group diversity and protects corresponding sensitive values in the anonymization mechanism, especially when the sensitive values within a subgroup exhibit homogeneity [5]. In our dataset, we identified that the sensitive information is DIAGNOSIS, as patients do not want to disclose their diagnosis result and health conditions to others.

Overall, the dataset achieved only 1-diversity, meaning that the k-diverse group contains 1 different value. It is perhaps because as we inspect the dataset, we found out that there are too many missing values, making a large number of records belonging to the same class in each field. Besides, we could also see that there are many distinct text values with different diagnosis notes in the DIAGNOSIS column, with 1 number of records. We can do further calculations to show that 11.212% of the total data archives 3-diversity and 4.64% of total data achieves 5-diversity. In general, the MIMIC-3 dataset could work on a DIAGNOSIS column and blurring on this column, such as mapping distinct diagnosis with less than 3 number of occurrences to be 'OTHERS'. In this way, the dataset could achieve a higher level of L-diversity and protect individuals for their sensitive information.

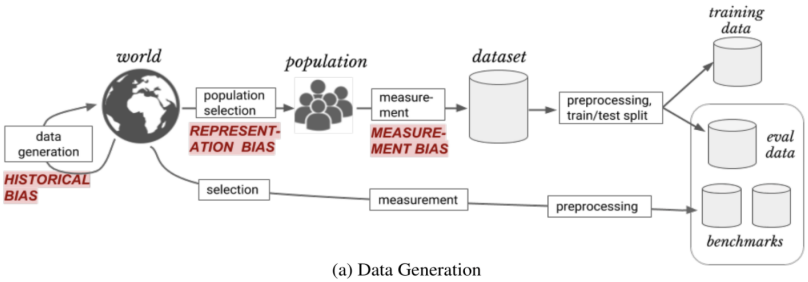
To summarize, L-diversity in general, is a stronger method to protect sensitive attributes and can be an extension to k-anonymity. K-anonymous dataset could still suffer from homogeneity and background knowledge attacks. To protect privacy from these attacks, l-diversity ensures intra-class diversity within sensitive fields. The larger the L, the higher the diversity gets.Hence, with higher levels of l-diversity, sensitive attributes are no longer distinguishable and privacy is protected even without knowing the knowledge that adversary possesses.

4 Bias

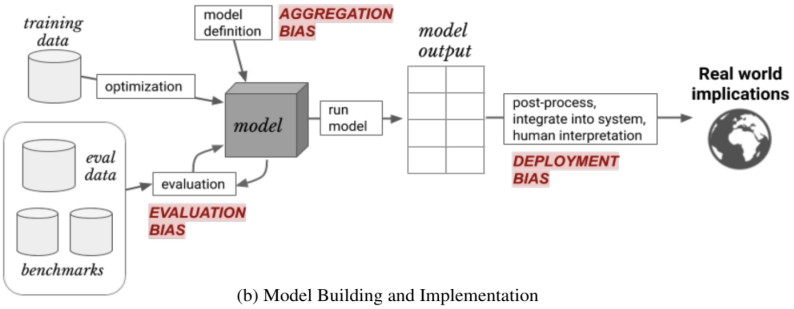
Central Question (3)

what kind of bias we shall take into account?

There are lots of sources of bias if we use a dataset to build a model and answer a question. From the graph provided by Suresh and Guttag (2020) [6] we can see that data generation and model building/implantation will result in bias. In data generation, we will have historical bias which is almost a social norm; we will also have representation bias and measurement bias. In modeling, we will have aggregation bias, evaluation bias when using metrics. Finally, we will have a deployment bias when implementing our model to real world.



(a) Data Generation



(b) Model Building and Implementation

Figure 5: bias in data and modeling

In the light of this, we will divide our experiment into two parts, analyzing the bias (historical and representation) in the dataset and in the (mostly evaluation bias) modeling procedure. In the modeling part, we also try two kinds of in-processing [7] bias-mitigation methods.

Data Generation

Lots of the bias embedded in data generation can be discovered through exploratory data analysis (EDA). Therefore, we provide some EDA to understand the data composition.

Univariate. As we mentioned before, age has certain outliers resulting from random shifts of dates. As we can see from the left distribution of age, most patients are above 50 years old. For gender, the dataset has almost balanced records for males and females.

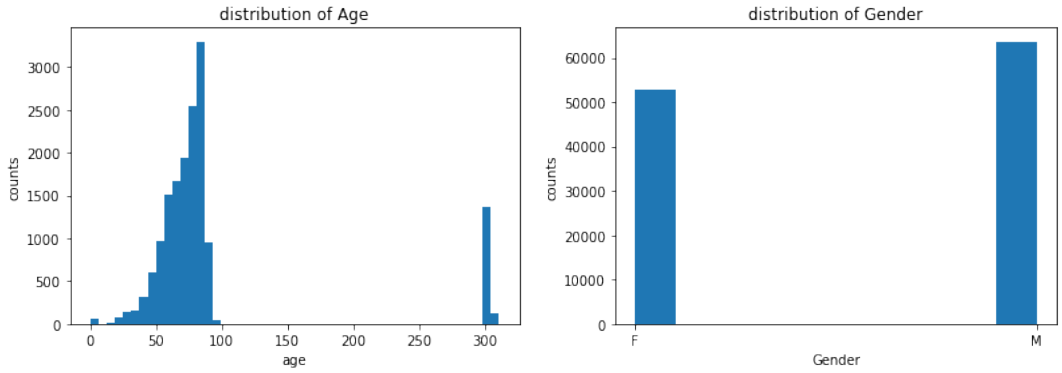


Figure 6: age and gender distribution

For ethnicity, we can see 66.01% of patients are white people, with 17.44% being blackAfrican American. Hispanic or latino, asian, and other very specific ethinic groups followed by.

ETHINICITY	COUNTS	% TOTAL RECORDS
WHITE	76855	66.01%
BLACK/AFRICAN AMERICAN	20304	17.44%
UNKNOWN/NOT SPECIFIED	5683	4.88%
HISPANIC OR LATINO	3138	2.70%
OTHER	2518	2.16%
ASIAN	2073	1.78%
UNABLE TO OBTAIN	946	0.81%
	111517	95.78%
HISPANIC/LATINO - PUERTO RICAN	818	0.70%
PATIENT DECLINED TO ANSWER	747	0.64%
ASIAN - CHINESE	409	0.35%
WHITE - RUSSIAN	352	0.30%

Figure 7: ethnicity distribution

Another interesting attribute to look at is the insurance type. We can see that 53.76% of patients enroll in Medicare, a federally funded plan for people who are older than 65 years old and young people with certain diseases or disabilities. 29.80% of patients have private insurance.

Insurance Type		% TOTAL RECORDS
Medicare	62588	53.76%
Private	34694	29.80%
Medicaid	15640	13.43%
Government	2762	2.37%
Self Pay	742	0.64%
	116426	100.00%

Figure 8: insurance distribution

Multivariate. The mortality rate is balanced across gender.

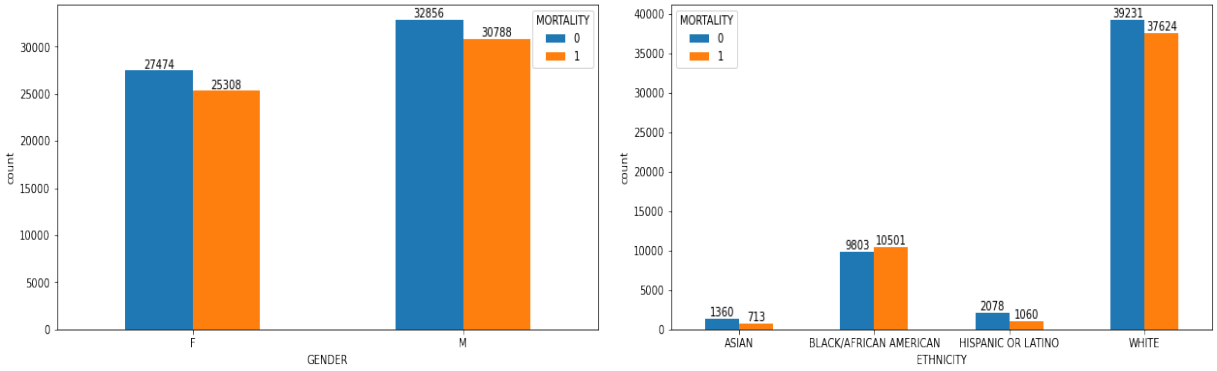


Figure 9: gender and ethnicity with mortality

The mortality rate for black/african american ethic groups is highest in four identified, common ethic groups in the United States.

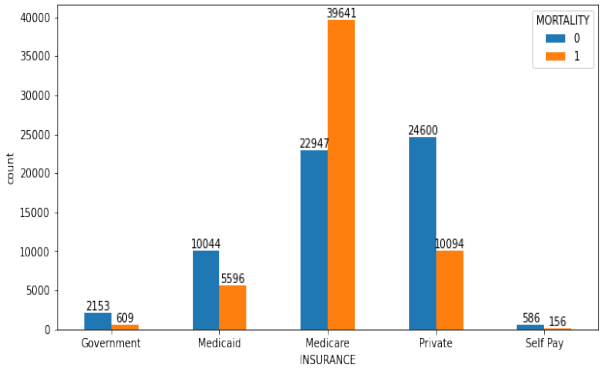


Figure 10: insurance with mortality

The mortality rate of patients who enroll in Medicare is higher. But we cannot draw immediate conclusions here because based on the eligibility of Medicare, latent causes such as age, underlying disease, and income shall

be considered into the reason for having a higher mortality rate.

Bias. Here we discuss the bias in ethnicity and insurance. Although the ethnicity appears to be very imbalanced with over 60% of patients are white patients, it is actually close to the ethnicity breakdown of the U.S. population. The 2019 distribution of the U.S. population by race and ethnicity shows 60.1% are white, 18.5% are hispanic, 12.2% are black, and 5,6% are asian. For our dataset, 66% are white and 17% are black, can be thought as close to the actual distribution. Therefore, the majority of white patients reflects historical bias in the decomposition of the country's population. However, in our dataset, Hispanic/ latino and asian are less than 3% and this does not match the actual breakdown. Therefore, we claim that there exists representation bias in ethnicity as well.

For insurance type, U.S. census website provides that private health insurance coverage was more prevalent than public coverage, covering 68.0% and 34.1% of people. However, in our dataset, Medicare is a public coverage and covers 53% of people. Private insurance covers 29.8% people. It suggests that the insurance breakdown in our dataset is not representative to the U.S. population.

Cautiously, we claim that there exists historical and representation bias in ethnicity as well as insurance type. The biases have to be carefully recognized and handled because they may be transmitted to the later modeling phrase. In our opinion, it is worth noting that black population, although achieving close representation compared to the actual distribution, still has higher mortality rates. We are curious if some unseen bias will propagate into modelling, causing models to falsely classify mortality. We will be able to explore in the modeling section.

Modeling

One of our goals is to go through the model-building process to predict mortality. After having the pre-modeling bias identified, we enter the modeling part. The target is mortality, a binary indicator with "1" being death, "0" otherwise. For the modeling codes, please see the appendix.

Data Preprocessing for Modeling. We conduct several data preprocessing steps for modeling. Since there are some variables, like DEATHTIME, DOD and DOD_HOSP, that have exactly the same meaning as our response variable - mortality, it's necessary for us to drop them. We also drop meaningless variables for our tasks, like SUBJECT_ID. Noticing that variables relating to time are accurate to the second, we only keep the information of year to facilitate our feature engineering process. Moreover, to conduct modeling tasks, we then implement dummy encoding to convert categorical variables. And we fill null values with 0.

Accuracy Comparison Among Multiple Models. We implemented Logistic Regression, Logistic Regression with L2 penalty, Random Forest and LightGBM models to compare the accuracy across multiple models. We observe that more sophisticated models tend to achieve higher accuracy. And LightGBM performs the best in both training and testing datasets.

Modeling Bias Investigation for Logistic Regression. We mainly use false positive rate and false negative rate as our metrics for evaluating the modeling bias across different groups. A false positive is an error in binary classification in which a test result incorrectly indicates the presence of a condition such as a disease when the disease is not present, while a false negative is the opposite error where the test result incorrectly fails to indicate the presence of a condition when it is present. False positive rate is the probability of falsely rejecting the null hypothesis for a particular test. The false positive rate is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification). Complementarily, the false negative rate is the proportion of positives which yield negative test outcomes with the test, i.e., the conditional probability of a negative test result given that the condition being looked for is present.

An unbiased model would achieve balanced values between false positive rate and false negative rate. An accurate model results in low values in both rates.

Our Logistic Regression Model achieves high accuracy and low bias across different groups of gender as we can see from the visualization plot. The false positive rate and false negative rate are well balanced for both male and female.

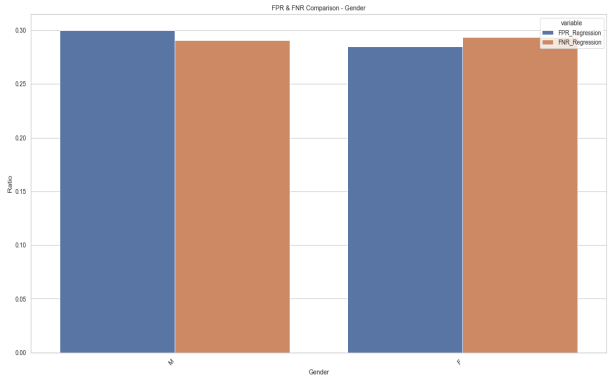


Figure 11: FNR and FPR in gender

However, large modeling biases exist across different insurance groups. The group of Medicare has contrary bias from other groups, with very high false positive rate and low false negative rate. Medicare is the federal health insurance program for people who are 65 or older, certain younger people with disabilities, people with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a transplant, sometimes called ESRD). We can see this is specifically for people with high risk of death, and that's why our model tends to wrongly classify people in this group as mortality.

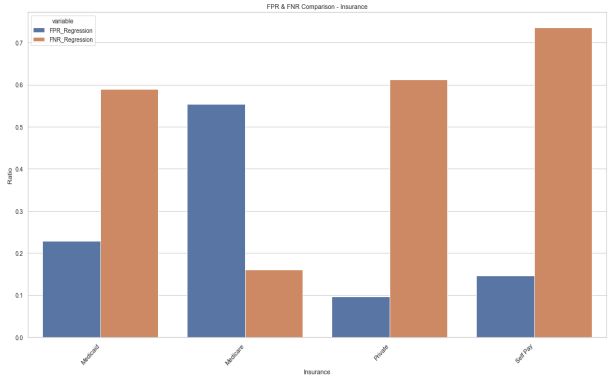


Figure 12: FNR and FPR in insurance

As for ethnicity, we observe similar performance as analyzed for insurance. Strong modeling bias exists for the majority groups in ethnicity. Moreover, groups traditionally regarded as higher risk of mortality in culture and society, like Black and African American, tend to have high false positive rates like the Medicare group in insurance as we analyzed before. However, White groups instead, have high false negative rates.

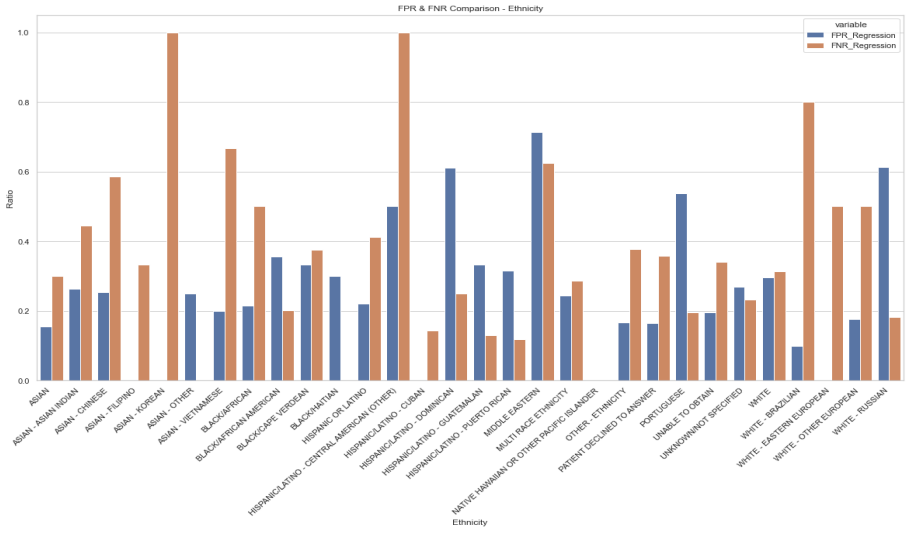


Figure 13: FNR and FPR in ethnicity

Central question(4)

How can we mitigate them and what are the consequences of the mitigation?

Modeling Bias Mitigation for Logistics Regression

In this part, we analyze two different modeling bias mitigation methods - implementing different models and using different thresholds for different groups. We want to mitigate our modeling bias while keeping the accuracy of our model at a relatively high level.

To mitigate the strong modeling bias in Logistic Regression, we firstly try using LightGBM model which achieves the highest accuracy in both training and testing data as we analyzed before. LightGBM is a gradient boosting framework that uses tree based learning algorithms. The model based on boosting tries to reduce the error in predictions by, for example, focusing on poor predictions and trying to model them better in the next iteration, and hence reduces bias.

LightGBM performs well across different groups of gender. It reduced both the false positive rate and false negative rate while keeping them balanced.

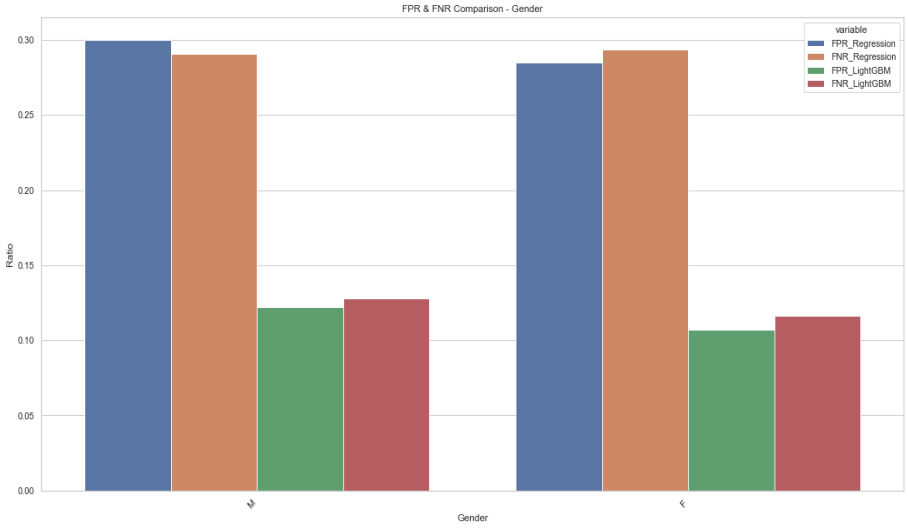


Figure 14: FNR and FPR in gender: mitigated

As for the variable insurance, LightGBM also does well in reducing the values of both the false positive rate and false negative rate. However, due to the different extent in the decrease of values, it might result in the worsening imbalance in some groups, like self pay.

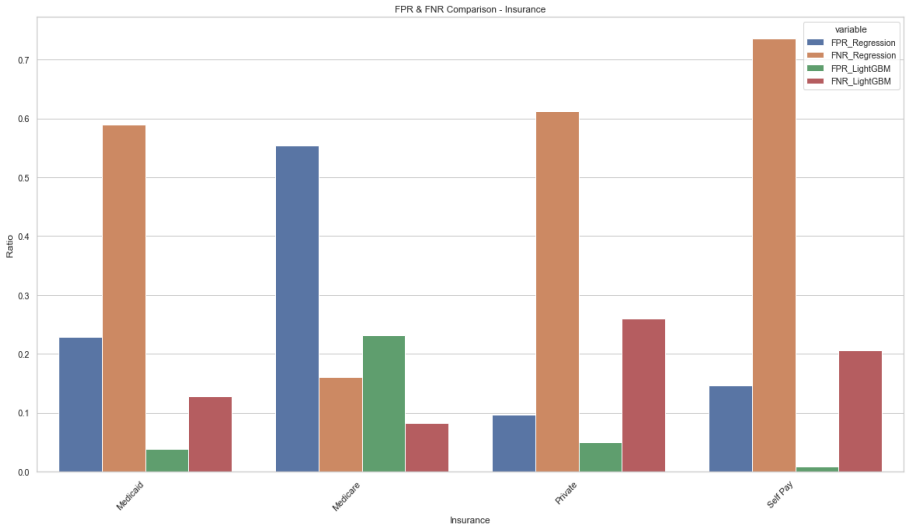


Figure 15: FNR and FPR in insurance: mitigated

As for ethnicity, LightGBM performs similarly as in insurance. For nearly all the groups, it reduces both the values of both the false positive rate and false negative rate. However, several groups also become more imbalanced, like Asia - Chinese, Black African and so on.

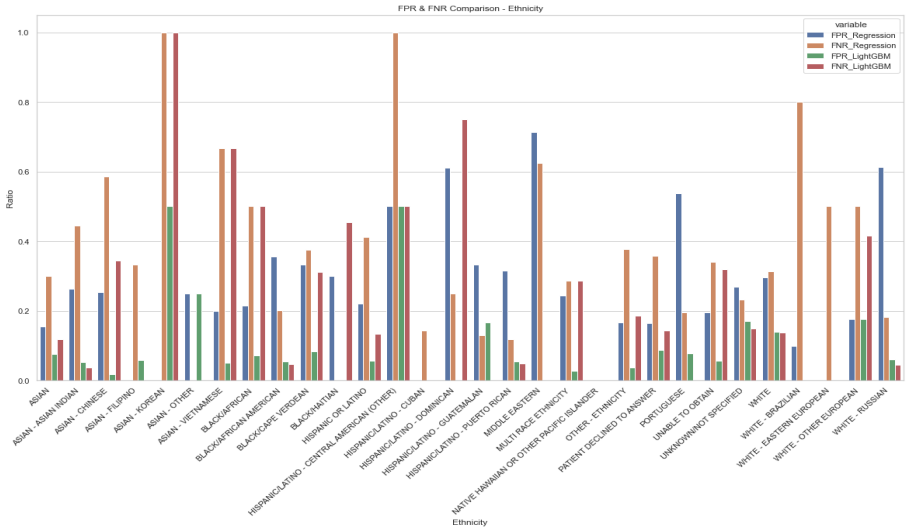


Figure 16: FNR and FPR in ethnicity: mitigated

Then, we try our second modeling bias mitigation method - using different thresholds for different groups. This can be used in any pair of groups we select. For instance, the visualization plot shows how the false positive rate and false negative rate between groups of White and Black and African American and the model accuracy change with different thresholds. When thresholds are between 0.1 and 0.3, the false negative rate is quite imbalanced between the two different groups. Moreover, we could achieve a perfect balance between these two groups by setting the threshold about 0.05 or 0.8. However, on the one hand, the threshold is very far away from 0.5. On the other hand, the accuracy is not high in these two points. And we sacrifice model accuracy to achieve the balance.

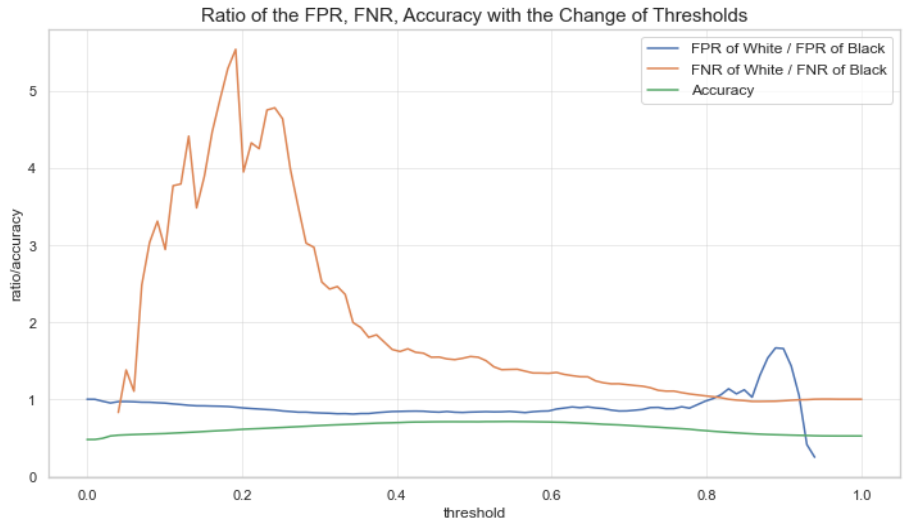


Figure 17: threshold adjustment

In summary, both methods have their advantages and disadvantages. As for models, it's advantageous in reducing the values of both false positive rate and false negative rates, thus improving model accuracy greatly. However, sometimes we might get a strong imbalance between groups. And it's hard to understand the sophisticated model. As for using different thresholds, we are able to achieve a balanced result between groups. It's also more flexible and transparent. We could choose any groups we want to adjust. However, for a dataset with lots of variables, if we want to achieve balance in many groups, we need to adjust one by one. Moreover, good thresholds for one group might not be applicable for others. Changing the threshold might influence other groups greatly. Moreover, sometimes we reduce the model accuracy by achieving the balance.

5 Conclusion and Future

Conclusion

MIMIC-III is a rich dataset to work with. Through the exploration of privacy issues and bias embedded in this dataset, we can now summarize the answer of our four central questions:

1) what are the consequences and evaluation of already-implemented de-identification procedures?

The existing methods of de-identification of date shifts and text removal are effective. However, we pose concerns of mis-representation of age variables and difficulties in conducting longitudinal studies.

2) What are the consequences and evaluation of other de-identification procedures that we can apply on the dataset?

We managed to apply row suppression, adding synthetic records, column suppression (blurring), and generalization to raise the k-anonymity measurement of the dataset and achieve a higher privacy level. However, we discover the trade off of accuracy, information, and privacy. Moreover, we argue that l-diversity is a better method than k-anonymity to de-identify the dataset.

3) What kind of bias should we take into account?

We shall consider historical and representation bias in MIMIC data generation. Specifically, attributes like ethnicity and insurance type are worth noticing because bias can be propagated into predicting mortality. When building models, we use FNR, FNP to assess the bias in modeling, and we are aware of evaluation bias if we use the same type of metrics. We care more about group fairness and accuracy of the algorithm. We discover that the most complex LightGBM model provides highest accuracy and the baseline logistic model has resulted in unbalanced FNR and FNP across ethnic groups and insurance types, but balanced rate across gender.

4) How can we mitigate them and what are the consequences of the mitigation?

We use two in-processing methods to try to mitigate the bias in modeling. First is to try on different algorithms (models) and we use LightGBM model to significantly reduce the FNR and FNP across gender, ethnicity, and insurance type. However, the drawback is that sometimes we overtune the balance of certain groups in ethnicity and insurance type. Another method we implement is to change the threshold of classification in logistic regression, specifically for reducing the imbalance of black and white group. We find that the threshold to be either too low or too high than the conventional 0.5, making this method less pragmatic.

Future Work

With the recent pandemic, there exists rich content and sources for medical data. And there is also a growing demand for analyzing medical datasets for prediction, time series analysis, trend and anomaly analysis. Datasets may seem anonymized but the level of privacy is undetermined, especially when we use a quasi list to evaluate the dataset on the basis of K-anonymity and L-diversity.

In the future, the work we've done could be formulated into a framework as a process to evaluate the de-identification, privacy, bias contained in the medical datasets. There is hope that we could be able to generalize our comments and findings on the de-identification process and potential biases to medical datasets.

Mitigation method we explored in the modeling session could also be further extended to handle data generation bias as well as other aspects of modeling bias. Besides, as we did an extensive discussion on technical bias, we could also further explore emerging bias in the future, which is the bias that arises in the way in which people interact and use the system and database. It could be more personalized level bias involving changing societal knowledge, population, or cultural values.

By setting a part of de-identification code as a reusable framework, the ultimate goal of this piece of work could be generalizing and developing a software package to improve privacy protection (K-anonymity de-identification) on a universal set of medical datasets. The light software package could contain features such as allowing the users to check for direct and quasi-identifiers, making the dataset conformed to K-anonymity using suppression, generalization, or both, customized to the user's demand.

Acknowledgement

Special thanks to Dr. Michael Smith, teaching fellow Santiago Romero-Brufau and Alex Cabral for providing inspiration and guidance throughout our final project. Thank Max Li for providing feedback during our final presentation.

References

[1] Clifford,G., Scott, D., a & Villarroel, M. (2012). *User Guide and Documentation for the MIMIC II Database.* Retrieved from
`https://mimic.physionet.org/archive/mimic-ii-guide.pdf`

[2] Nuthakki,S., Neela, S., Gichoya, J., & Purkayastha, S. (2019). *Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks.* Retrieved from
`https://arxiv.org/abs/1912.12397`

[3] L. Sweeney. k-anonymity (2002). *a model for protecting privacy. International Journal on Uncertainty.* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[4] Anugili, O., & Waldo, J. (2016). *Statistical tradeoffs between Generalization and Suppression in the De-Identification of Largs-Scale Data Sets.* IEEE 40th Annual Computer Software and Applications Conference.

[5] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam,M. (2007). *L-Diversity: Privacy beyond k-anonymity* [ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007), 52 pages. DOI=10.1145/1217299.1217302
`http://doi.acm.org/10.1145/1217299.1217302`

[6] Suresh, H. & Guttag, J. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. Retrieved from
`https://arxiv.org/abs/1901.10002`

[7] `http://aif360.mybluemix.net/resources`