A Review of FAERS Data for Estimation of Risk Prone Populations

Bin Lu
Volgenau School of Engineering
George Mason University
Fairfax, USA
blu4@masonlive.gmu.edu

Chris Mynatt
Volgenau School of Engineering
George Mason University
Fairfax, USA
cmynatt@masonlive.gmu.edu

Srija Ummaneni
Volgenau School of Engineering
George Mason University
Fairfax, USA
summanen@masonlive.gmu.edu

Victoria Golden
Volgenau School of Engineering
George Mason University
Fairfax, USA
vgolden@masonlive.gmu.edu

***Abstract*—As new drugs have been developed to address disease, the occurrence of drug interactions increased. Several countries, including the United States, have developed reporting systems to track adverse drug reactions in order to survey the drug market. The U.S. Food and Drug Administration publicizes submitted reports of adverse drug reactions. There is currently no deterministic methodology formally approved by the FDA for medical practitioners in which to formally characterize a patient as being high risk for an adverse reaction. This paper aimed to perform categorization and characterization of populations affected by adverse drug reactions through the use of descriptive statistics and a machine learning classification model. It was found that adult to elderly females were the most susceptible to an adverse drug reaction, and experienced a low death rate. Best case training of a machine learning classification model only resulted in an accuracy of 0.35, severely limiting its application in production, but still provides reaction outcomes with greater accuracy than baseline. The results presented demonstrate the potential utility and limitations in application of adverse drug report data when determining risk prone patient populations.**

***Index Terms*—Adverse Drug Reactions, FAERS, ICH ICSR, Medical Conditions**

I. Introduction

With the development of society and the continuous improvement of medical standards, more and more diseases can be treated by drugs, and people's quality of life has also been significantly improved. However, the drug itself has two characteristics: on the one hand, it can exert therapeutic effects to treat diseases, on the other hand, the drug itself has certain side effects, and may lead to adverse drug reactions. According to incomplete statistics, among patients who have been hospitalized, approximately 10% to 2% have drug adverse reactions [1]. Because of the dual nature of drugs, people's dependence on drugs and the side effects of the drugs themselves have become a problem.

The term "adverse drug reaction" refers to a harmful drug that occurs in the normal use of a qualified drug, such as toxicity reaction, allergic reaction, triad effects (deformity, mutation, cancer, etc.), residual reaction, etc [2]. The variable nature of drug reactions is caused by individual differentiation in drug tolerance. When drugs enter the market, clinical trials are needed to reduce the occurrence of adverse reactions. However, most clinical trials exclude children, the elderly, pregnant women and other infected groups from the test population [3], so the results of clinical trials do not always accurately represent all patient cohorts. In general, most adverse drug reactions are mild and tolerable, but adverse reactions to certain drugs can cause disability or death. In recent years, countries have been committed to continuously strengthening the management of the safe use of drugs; but, the incidence of phytotoxicity has increased. Drug safety issues are a global concern, as an increasing number of people become disabled or die each year due to adverse drug reactions. Presently, many countries have established laws, regulations and quality standards for drug supervision; however, the situation facing drug safety is still serious. At the same time, the consequences of adverse drug reactions are often very serious. According to the World Health Organization (WHO), nearly six to twelve percent of the annual inpatients are hospitalized because of adverse drug reactions [4], and about 50% of all deaths worldwide die from adverse drug reactions [5]. In the United States, adverse drug reactions are among the top 10 leading causes of death [6]. There are approximately 2 million patients who are exacerbated by adverse drug reactions each year, and about 110,000 of them die from adverse drug reactions, resulting in a direct economic loss of about $75 billion [7]. Not only in the United States, same events happened in other countries. For example, According to statistics, in 2015, China's Adverse Drug Reaction Monitoring Network received 1.39 million copies of the Report on Adverse Drug Reactions/Events, of which 28.2% of the reports of serious adverse events and newly reported adverse events accounted for 28.2% of the total number of reports. , an increase of 2.5% compared to 2014 [8]. Increasing adverse events are like an alarm that reminds us of the importance of drug safety monitoring.

Within the history of drug development, people's understanding of pharmacology was not profound because of the limited levels of knowledge at that time. Even currently, imperfections in the regulatory system comprise one of the most important reasons patients may encounter adverse drug reactions. Arguably, the most famous example is the Thalidomide event that occurred in the 1960s where this drug was used as an anti-nausea medication for pregnant women; However, it was later discovered that thousands of severe birth defects were attributed to the use of this drug [9]. The occurrence of

this incident prompted multiple nations to begin to recognize the importance of regulating drug safety.

Due to the harmfulness and seriousness of adverse drug reaction events, we need to take measures to strictly control the safety of listed drugs. Once there is a highly suspected adverse reaction signal, it must be reported to the relevant units for research, analysis and management to reduce the hidden dangers of drug use. . Of course, to monitor drug safety, it is necessary to collect reports of adverse drug reaction events, using uniform formats and standard reporting and storage. Many hospitals and related research institutions at home and abroad have collected and organized reports of adverse drug reactions. For example, the World Health Organization (WHO) established the WHO Collaborating Centre for International Drug Monitoring, the Uppsala Monitoring Centre (UMC) [10]. In addition, many countries have established their own spontaneous reporting systems for adverse reactions, such as the US Food and Drug Administration (FDA) established the FDA Adverse Events Reporting System (FAERS) and the Eudra Vigilance database of the European Medicines Evaluation Agency. These spontaneous reporting systems are widely used in the analysis and research of adverse reaction events, in which the amount of data is large, and the more standardized FAERS database has been widely recognized in drug safety monitoring. Therefore, this paper will also conduct data mining on the FAERS database to identify suspected adverse reaction signals and provide them to relevant institutions for management to reduce all aspects of the adverse effects of adverse drug reactions.

Based on this situation, analyzing the metadata of drug adverse reactions to estimate risk prone populations has great significance. Firstly, it can promote clinical rational use of drugs. Analyzation of metadata from adverse drug reaction reports to identify the applicable population and high-risk groups of different drugs carries multiple benefits, as it also helps to promote good communication with drug manufacturers, jointly publish drug safety information, improve drug specifications, identify potential safety hazards, and provide risk warnings. This helps medical practitioners to rationally choose drugs and minimize the occurrence of adverse reactions. In addition, the analysis and evaluation of drug reactions provides a basis for the FDA to rectify or phase out drugs from the market [11]. Government agencies can report on adverse drug reactions based on analysis reports and give safety recommendations. For drugs with serious adverse reactions, policymakers must determine whether it is necessary to suspend the medicine's distribution. In scenarios where no alternative medicine is available, the presence of serious adverse reactions could promote the development of new alternatives.

II. Literature Review

After a long period of development and progress, the drug adverse reaction reporting system has become more sophisticated and mature, allowing much research to be done in this area. Peng [12] and her associates used data mining methods to study adverse reactions described in the FDA Adverse Event Reporting System (FAERS), and analyzed the distribution of sex and age in the report. They found that among patients with adverse reactions in the digestive system, 40 to 75 years old patients accounted for the majority, and most of the adverse reactions occurred within half a year of medication. Motiur and Rahman [13] also used data from FAERS but focused on analyzing the difference between brand name and generic antiepileptics, finding that the generic formulation of lamotrigine has a higher risk to cause adverse reactions. Additionally, Yip et al. proposed that association of HLA and carbamazepine hypersensitivity as an adverse drug reaction demonstrates sensitivity regarding ethnicity and phenotype specificity [14]. But, it was found that most research activities carried out on adverse drug effects focused on drug classification, and effects of drug interactions related to specific adverse reactions, but limited progress has been made regarding generalized demographics who are at increased risk of susceptibility to having an adverse drug reaction, to include determination of age groups, sex, and weight. As such, the authors suggest that categorization and identification of demographics and patient cohorts who report adverse drug effects to FAERS may be of value.

Outside the scope of the FDA's AER system, significant research has been done on global populations with respect to characterization of adverse drug reaction prone population traits. General population characterizations assumed to be at higher risk are: pregnant women, women, ethnicity, disease states, polypharmacy, and age related individuals to include elderly and children [3][15]. Additionally, two of the characterized groups, pregnant women and children, tend to be underrepresented with clinical trial results due to ethical standards. However, while adverse drug reactions are reported for all aforementioned characterizations within the FAERS dataset, this type of analysis has not yet been performed on the FAERS dataset.

Although most researchers believe that FAERS data is very important for studying disease characteristics [16], some scholars still suggest data collected from FAERS has some limitations: there is data to suggest that adverse drug reaction reports are underreported when compared to spontaneous reporting via alternative mediums. The Drug Safety Research Unit in Southampton, UK, determined that the median under-reporting rate of ADRs across 12 nations was 94% (with an interquartile range of 82-98%) [17]. Their findings suggest that there may be significant under reporting in many nations, which can lead to a lack of insight regarding affected demographics and the types and volume of adverse drug reactions for which they may encounter. The reason why this happened is that FAERS uses a voluntary reporting system, and adverse events and medication error reports are usually submitted voluntarily by health care professionals and consumers, so there are a large number of unsubmitted adverse drug events. Unfortunately, this means the FDA cannot collect all reports of all adverse events or medication errors for a drug [18]. In addition, at the time of submission, the FDA does not require proof of causality between adverse events and drugs, and the report does not usually include details of adverse events. Since we do not

intend to use this data to calculate the incidence of adverse events or medication errors in the US population, these issues will have no impact on our findings.

The importance of determination of high risk demographics of patients is well understood from an insurance perspective, but is often nebulous in definition when being assigned by a practitioner to a patient. This is partially due to the qualitative nature of a typical high-risk patient diagnosis [19][20]. Typically, identification of a high-risk patient is determined by a combination of multiple variables, including age, weight, number and type of medications, and reported symptoms [21]. However, while efforts exist to expand current knowledge regarding identification of high-risk patients, there is currently no deterministic methodology formally approved by the FDA for medical practitioners in which to formally characterize a patient as being generally high risk especially when considered within the context of patients that may experience adverse drug reactions [22][23].

As such, the authors suggest the following research question to be addressed:

1. Within the United States, what characterizes the typical patient for which an adverse drug reaction report has been submitted to the United States Food & Drug Administration?

III. Objectives

The current objective of this research is to determine a general profile of a U.S. based patient for which an adverse drug reaction report has been generated and submitted to the FAERS. The project has the following sub-objectives:

1. Analyze relationships between individual physiological metadata and patient reactions.

2. Identify differentiation in trends between male or female populations when compared to the general populous.

3. Evaluate FAERS reports with geospatial information to review trends in patient demographics between varying nations.

4. Create a distributable model that estimates drug reaction fatality based on patient characteristics through the use of unsupervised machine learning.

Results from this study may prove valuable to the healthcare industry when determining high risk patients, as utilization for this data could affect the prescribing of higher risk approved drugs to higher risk demographics of patients. We hope that this effort proves significant, as adverse drug reactions comprise the fourth leading cause of death in the United States, and costs individuals and the industry billions of dollars annually [24][25][26].

IV. Methods

Data utilized for this project was sourced from the FDA Adverse Event Reporting System (FAERS), a computerized information database containing event reports, medication error reports, and product quality complaints resulting in adverse events to be submitted to the FDA. This data is utilized by the FDA currently within their postmarketing safety surveillance program, which aims to monitor and identify correlations between all approved drugs and reported adverse drug events or medication errors [27].

FAERS data is made available to the general public through multiple ingest mechanisms. For the purpose of conducting this project, FAERS data will be sourced through the use of openFDA which is a government sponsored project which provides easy access to a wide variety of resources including source code, reports, and relevant government sourced datasets. OpenFDA provides FAERS data, stored temporally in standardized JSON format from the first quarter of 2004 to the second quarter of 2019. The FAERS data sourced via openFDA is updated quarterly, and is minimally altered from source format in order to fit a standardized JSON structure. The FAERS dataset has slowly increased in volume with time. Individual year report counts crossed 0.5 million starting in 2010, with each subsequent year having more than the last [28].

FAERS data can be queried via openFDA using their published API; and, adverse event reports are stored in JSON structure. For the purposes of this research, the FAERS data will be consumed via the openFDA API and stored in a relational database to allow for a static dataset as well as facilitating data classification and organization. The data utilized for analysis will include all FAERS data reports received by the FDA from the 2016-2018 calendar years (n = 3,718,300). In native format, adverse event reports use the ICH E2B/M2 version 2.1 standard. This standard, established in February of 2001 by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, standardizes data elements for the transmission of ICSRs regardless of source or destination. This eases interoperability between regulatory authorities, pharmaceutical companies, investigators, and international monitoring personnel. As shown in Figure 1, the ICH M2 relational view of the E2B data elements helps to show relationships between the elements within various sections of the E2B document.

Each FAERS record is comprised of three major sections, consisting of general information (metadata) regarding the report and its creation (Elements A.*), patient information (Elements B.*), and the products taken and reactions reported during the time the event was experienced and recorded by the practitioner (Elements B.4.*). As shown in Figure 1 above many of the E2B elements within the ICH ICSR relational diagram may have many sub-elements, many of which are particularly related to drug information, reactions, and relevant past drug history. However, unfortunately, there are limitations to what may be accomplished by reviewing the data contained within the FAERS data. The FDA outlines four major limitations to be taken into consideration when utilizing the dataset [27]. These include:

1. Duplicate or incomplete reports may be present.

2.Drugs suggested as causal factors for adverse reactions are not definitive and cannot be proven on an individual report basis.

3. Report content has not been verified by the FDA.

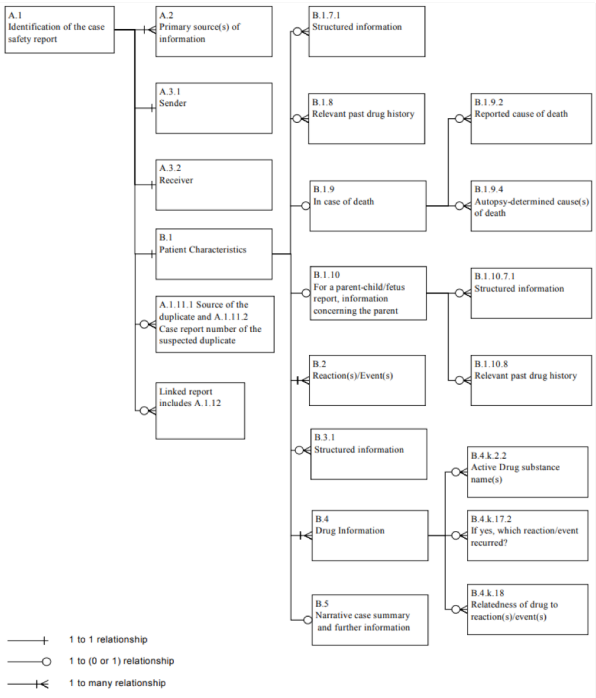

Fig. 1. M2 Relational View of E2B Data Elements [29]

4. Incidence rates may not be established by the reports due to incomplete data.

However, while the provided limitations do prevent the analysis of many drug-related venues of interest, this data is still relevant for the determination of characteristics relating to patient demographics and health. In order to address the objectives outlined for this effort, and due to the combination of the previously mentioned limitations and the potentially expansive nature of any given FAERS report, the following fields (Table 1) were identified as important within the JSON structure of the openFDA sourced FAERS report.

In native JSON format, the FAERS dataset utilized in this effort was greater than 78 GB. Application of traditional tools on this dataset, such as dataframes and loading of JSON into variables, were determined to be infeasible due to the volume of the dataset, so alternative methods were investigated. As such, for the purposes of this effort, a SQLite database was used for data storage. SQLite was chosen because of its ability to be a self-contained, serverless, zero-configuration, transactional SQL database engine [30]. This tool allows for a filesystem based solution that can easily be passed across platforms and negates the need for a separate server process along with associated costs. Prior to insertion of FAERS records into the SQLite database, a custom python script was utilized to selectively pull out fields from the raw JSON FAERS data from January 1, 2016 to December 31, 2018 and store only the information relevant to the objective of this project (Table 1). Information parsed using the script was entered into a SQLite 'Patients' table. It was during this stage

TABLE I
FIELDS PARSED FROM FAERS REPORTS FOR ANALYSIS.

Field name	**Field Description**
safetyreportid	The 8-digit Safety Report ID number, also known as the case report number or case ID. The first 7 digits (before the hyphen) identify an individual report and the last digit (after the hyphen) is a checksum. This field can be used to identify or find a specific adverse event report.
serious	Seriousness of the adverse event.
companynumb	Identifier for the company providing the report. This is self-assigned.
receiver	Name of the organization receiving the report. Because FDA received the report, the value is always FDA.
sender	Name of the organization sending the report. Because FDA is providing these reports to you, the value is always FDA-Public Use.
occurcountry	The name of the country where the event occurred.
receiptdate	Date that the most recent information in the report was received by FDA.
receiptdateformat	Encoding format of the transmission-date field. Always set to 102 (YYYYM-MDD).
transmissiondate	Date that the record was created. This may be earlier than the date the record was received by the FDA.
receivedate	Date that the report was first received by FDA. If this report has multiple versions, this will be the date the first version was received by FDA.
receivedateformat	Encoding format of the transmission-date field. Always set to 102 (YYYYM-MDD).
fulfillexpeditecriteria	Identifies expedited reports (those that were processed within 15 days). (Either 'yes' or 'no').
patientonsetage	Age of the patient when the event first occurred.
patientonsetageunit	The unit for the interval in the field patientonsetage. (Either 'Decade', 'Year', 'Month', 'Week', 'Day', or 'Hour').
patientsex	The sex of the patient. (Either 'Male', 'Female', or 'Unknown').
patientagegroup	Populated with Patient Age Group code. (Either 'neonate', 'infant', 'child', adolescent', 'adult', or 'elderly').
patientweight	The patient's weight, in kg (kilograms).
reactionoutcome	Outcome of the reaction in reactionmeddrapt at the time of the last observation. (Either 'Recovered/resolved', 'Recovering/resolving', 'Not recovered/not resolved', 'Recovered/resolved with sequelae (consequent health issues)', 'Fatal', or 'Unknown').
reactionmeddrapt	Patient reaction, as a MedDRA term. Note that these terms are encoded in British English. For instance, diarrhea is spelled diarrhea. MedDRA is a standardized medical terminology.
seriousnessdeath	Boolean if the adverse event resulted in death.
seriousnesslifethreatening	Boolean if the adverse event resulted in a life threatening condition.
seriousnesshospitalization	Boolean if the adverse event resulted in a hospitalization.
seriousnessdisabling	Boolean if the adverse event resulted in disability.
seriousnesscongenitalanomali	Boolean if the adverse event resulted in a congenital anomaly.
seriousnessother	Boolean if the adverse event resulted in some other serious condition.

that data preprocessing occurred.

Data reduction was performed through the use of attribute subset selection, as only metadata fields pertaining to patient characteristics were needed for analysis. Concept hierarchy generation was performed on geographic fields where necessary and all values were validated to ensure country codes were accurately represented as modern extant nations. Values were then transformed from ISO 3166-1 alpha-2 to ISO 3166-1 alpha-3 standard for increased compatibility with visualization tools. All date related fields were standardized to ISO 8601. According to the ICH ICSR Specification, the patient age group field is to only be filled out by the submitting party when patient age is unknown. As such, discretization was used on patient age where applicable to provide patient age groups. Most fields supplied by the ICH ICSR formatted FAERS records provide values as integers instead of the fields' meaning as a string. Data transformation was applied on all applicable ICH ICSR fields to replace the numeric value with its string definition for ease in understanding values without the need for lookup tables. Unfortunately, due to the multi-user submitted nature of the FAERS system, many records contained within the fields outlined in Table 1 initially contained blank or null values. Given the relatively voluminous size of the dataset (n = 3,718,300) and the desire to not introduce bias, it was determined that the most appropriate method would be to ignore empty values in the remaining columns instead of utilization of attribute mean or application of most probable values. After data preprocessing was complete, and only relevant fields were utilized, the database was reduced to approximately 738 MiB of patient attribute information. Unique patient information was identified by the unique safetyreportid and considered as an individual entity and primary key for the Patients table.

In order to analyze the data and produce the visualizations in the results section of the paper, the authors used Tableau Desktop software. Tableau is recognized as a data analyst visualization tool that is intuitive and data-driven. Use of Tableau was either through reading in a CSV file of the processed dataset or by linking to a SQL database (SQLite/MySQL). To aid in characterization of severity of adverse reaction in populations for which an adverse reaction is guaranteed to occur, a machine learning experiment was run to classify the dataset and create a distributable model that estimates drug reaction fatality based on patient characteristics through the use of unsupervised machine learning.

To achieve optimal performance, machine learning models often rely on significant time investment for algorithm selection, as well as hyperparameters that require extensive tuning [31]. Microsoft Azure Machine Learning Studio was chosen for this effort, as it exploits experiments performed on hundreds of datasets via probabilistic matrix factorization and utilizes an acquisition function from Bayesian optimization to guide the exploration of the space of possible machine learning pipelines [31]. As such, it provides a streamlined solution that automatically builds predictive machine learning pipelines for a user supplied dataset, automates the selection of data pre-processing methods, and allows the automated selection of algorithms and hyperparameter tuning. Instance selection was performed on the pre-processed FAERS dataset to aid in the accuracy in classification problems. The dataset was reduced , and the experiment only utilized records for which weight, sex, age, age group, and reaction outcome were present (n=24,411). To aid in classification, values were presented in their native ICH ICSR format which maintains FAERS record column values for the aforementioned fields as integers instead of strings. Input data was split into a training dataset and a holdout test dataset for validation of the model through the use of a train-validation split. Classes were balanced in the training data and a high cardinality feature detection test was performed, for which no high cardinality features were detected. The machine learning classification model was performed over 126 runs, using accuracy as the primary metric and automatic featurization.

V. Results

According to the records with marked sex, the FAERS data set exclusive to an occur country "USA" has a distribution of a 64 percent female population to 36 percent male population. The dataset has 11 percent of all records without a sex explicitly marked. This suggests a significant departure (p=0.009) from the United States population distribution of 51 percent female and 49 percent male. This comparison indicates males are 35 percent less likely to have an adverse drug reaction whereas females are 20 percent more likely to experience an adverse drug reaction. The global populations exclusive of the United States records have a population distribution of 55 percent female and 45 percent male.

The FAERS population consists of ages ranging from 0 years to greater than 100 years of age. Figure 2 shows the distribution of patient age in whole numbers by count overlayed with distribution of male versus female populations. Figure 3 supports the same distribution but groups ages by decade. The majority of individuals that were recorded with adverse drug reactions are aged 55 to 70 years of age. Male and female distribution showed a similar pattern, with relative peaks in adverse event reactions in the infant, adolescent, and adult age groups.

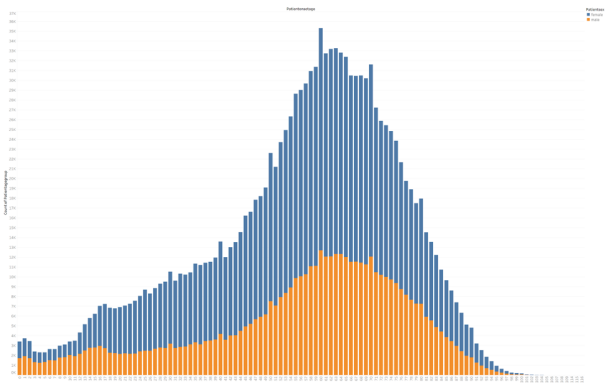

Fig. 2. Distribution of Sex by Count of Age in Population

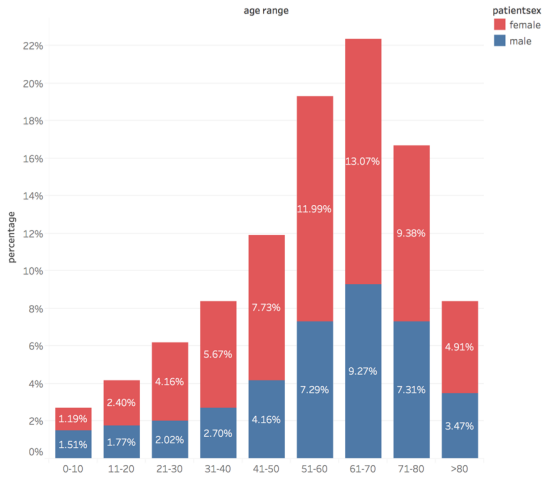

Fig. 3. Distribution of Sex by Count of Age in Population Grouped by Decade

The United States population has a mortality rate of 0.86 percent [32]. The FAERS data set exclusive to an occur country of "USA" has a mortality rate of 5 percent. Patients with an adverse drug reaction are 6.25 times more likely to die than the United States mortality rate. The global population data (all occurred countries exclusive of the United States) had a reported mortality rate higher than the US population of 11 percent.

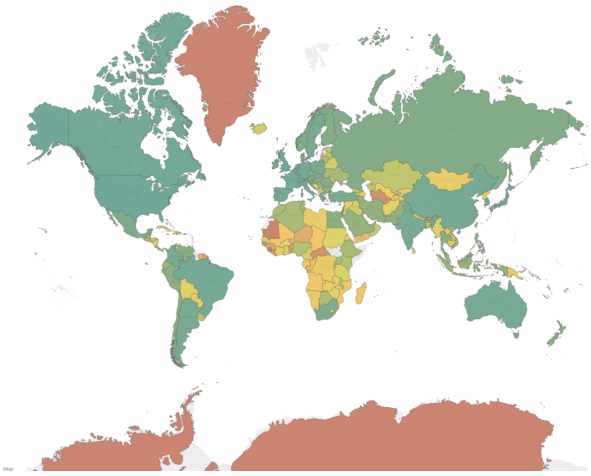

Fig. 4. Ranking by County by Occurrence of Reported Country

The United States is ranked number one for number of reports submitted by occurrence country. Figure 4 displays a global heatmap with green countries representing the rankings for most submitted records and red countries representing the rankings for fewest submitted records.

In order to better analyze the patient's age and recovery from adverse drug reactions, we divided the patients' age into several groups. As seen in Figure 5, we can see that no matter what age group of patients, most of them show two kinds of reaction outcome, which are not recovered and recovered. Very few patients have recovered but with sequelae. Among them,

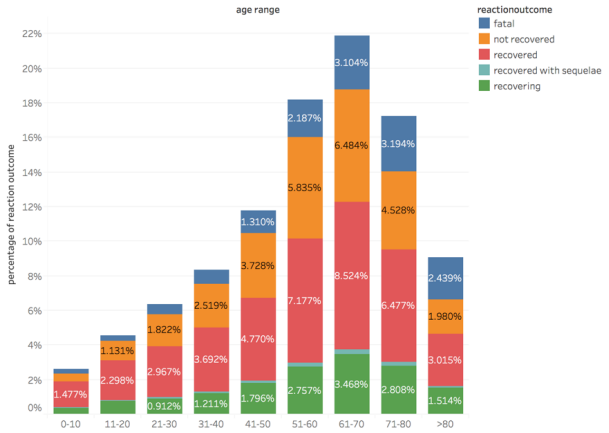

Fig. 5. Distribution of Reaction Outcome by Age Grouped by Decade

patients whose age between 61 and 70 has the highest rate of not recovered which is 6.484% and patients aged 51 to 60 who did not recover accounted for 5.835%, ranking second. In addition, with the gradual increase of age, the mortality rate caused by adverse drug reactions is also getting higher and higher: patients aged 71 to 80 takes 3.194% and patients aged 61 to 70 takes 3.104%.

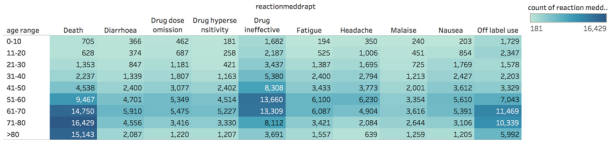

Fig. 6. Count of MedDRA by Age Grouped by Decade

Figure 6 shows the relationship between age and patients' MedDRA. It shows that more symptoms of adverse reactions appear if the patients are older, which is consistent with figure 5. There are 16,429 patients aged 71 to 80 and 14,750 patients aged 61-70 show the symptoms of death, which means elderly have more probability of death than young people.

There are thousands of symptoms of adverse drug reactions that may occur in patients, and we selected 10 of them for analysis. According to figure 7, we can find that drug ineffective, death and off label use are the top three reaction MedDRA. With the exception of death, women outnumber men in the remaining nine adverse drug reaction symptoms. Among them, the number of females and males who have a drug allergic reaction has the largest difference, the ratio is about 3.4 to 1.

With regards to the machine learning classification model, only low accuracy models were able to be created with the FAERS dataset from January 2016 to December 2019. The maximum classification model accuracy was achieved with a score of 0.35 and was performed with the StandardScaler-Wrapper, XGBoostClassifier algorithm. As seen in Figure 8, of the top K features, patient weight (importance=41225.024) was determined to have the highest global feature importance

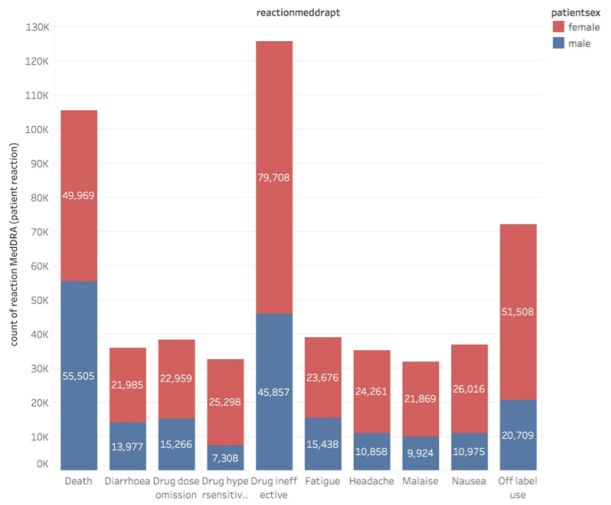

Fig. 7. Distribution of Sex by Count of Reaction MedDRA by MedDRA Reported

whereas patient sex was determined to have the least global feature importance (importance=442.187).

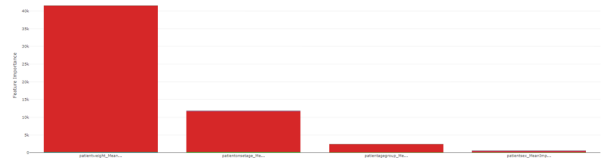

Fig. 8. Distribution of Top K Features by Global Importance in the Classification Model

Overall model performance was determined by the authors to be unsatisfactory, as can be seen in Figure 9 and Table 2. However, the classification model does demonstrate a greater chance to characterize severity of adverse reactions in populations for which an adverse reaction is guaranteed to occur.

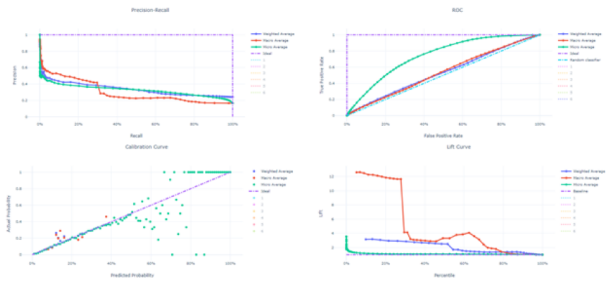

Fig. 9. Metrics Relating to Overall Classification Model Performance

VI. Discussion

Looking further into the distribution of male versus female reports, it was of note that the female population was skewed to encompass the vast majority of reported records as compared to the United States general population distribution. This result is statistically significant with p-value of 0.007. This

TABLE II
Run Metrics for the Chosen Classification Model.

Metric	Value
Norm macro recall	0.022314
Average precision score macro	0.21378
Precision score macro	0.34961
Recall score micro	0.34961
F1 score micro	0.34961
Balanced accuracy	0.18526
AUC weighted	0.57170
F1 score weighted	0.26015
Recall score weighted	0.34961
Precision score macro	0.42981
Weighted accuracy	0.46386
Precision score weighted	0.34415
F1 score macro	0.14557
AUC macro	0.59018
Average precision score weighted	0.29151
AUC micro	0.74297
Recall score macro	0.18526
Accuracy	0.34961
Average precision score micro	0.32689
Log loss	1.5130

leads the authors to believe females generally experience more complications when using pharmaceuticals. The authors note that numbers can be overrepresented due to females visiting the doctor more often than males as noted by historical studies [33].

Additionally, persons of ages ranging from 55 to 70 years of age were among the most prevalent with reported adverse drug reactions. This is likely due to the fact that older populations visit the doctor more frequently than populations ranging in ages of 1 to 64 [34]. Infants and individuals 65 and older are more likely to visit the doctor and have reported illnesses. The largest count by age was 60 years of age with 22,606 reports. This age does not fall into the classification described earlier.

For the distribution of patients' age and recovery, although the proportion of children under the age of 10 who failed to recover from adverse drug reactions was small, there were still some. This may be related to children's dysgenesis, high water content in the body, and weak immunity. From the age group of 11 to 20 to the age group of 71-80, with the increase of age, mortality and non-recovery rate showed an increasing trend. Among them, the group aged 61 to 80 accounted for the largest proportion. In this case, we believe that this may be related to the fact that the elderly often suffer from a variety of diseases, a large number of medications and a long medication duration. In addition, as the elderly get older, kidney function gradually weakens, which is also one of the causes of higher mortality. All of this data and cases remind us that the hospital needs to strengthen the monitoring of adverse reactions in elderly patients.

From the relationship between sex and the symptoms of adverse drug reactions, most of the top ten symptoms that occur frequently can be gradually recovered such as fatigue, nausea and diarrhoea. Although the severity of these symptoms is relatively small, the number of these symptoms are particularly

high, probably because their clinical manifestations are more easily observed than the adverse reactions of the visceral part, and they are related to the direct feeling of the patient. The damage of other systems is more hidden and may not be easily detected by medical staff. The reason that more than three times as many women have symptoms of drug allergy may be that women are more sensitive to their physical condition during menstruation and during pregnancy. In addition, drug ineffective and off label use appear very often, but these kinds of reactions can be reduced if the doctor can be more cautious when prescribing medicine to patients, based on the patient's past cases, family history and other conditions.

It was shown in the results of the machine learning classification model that utilization of FAERS data within the chosen timeframe is insufficient for use in training a machine learning classification model. This may be in part attributed to the disparity between record count and complete records, as prior to pre-processing only 0.66% of records contained values for all of the five requisite fields of sex, weight, age, and reaction outcome. Even beyond this limitation, due to the user submitted nature of the reports, veracity of the dataset is unknown. While outliers may be removed during pre-processing, it is impossible to determine accuracy of FAERS reports when physiological attribute values remain within acceptable value ranges. Even so, the model in its current state could still prove useful for untrained medical personnel when attempting to determine the potential seriousness of an incoming patient for which an adverse reaction has occurred as it provides estimations with greater accuracy than random choice.

VII. FUTURE WORK

Much of the work relating to future efforts relates to expansion of the dataset, improving the veracity of the dataset, and evaluation of further pre-processing of the data. For example, the dataset had a large portion of records without a sex assignment. These records could be lessened in count by deriving conclusions of sex based on biologically unique characteristics. Examples of such would be records that had mention of a uterus would be classified as female and records that had mention of prostate would be classified as male. An attribute selection such as this would need to be evaluated to prevent bias towards a sex based upon prevalence of drug reactions related to sex specific organs. Another physiological attribute, weight, was largely absent from the dataset for all three years observed. Unfortunately, given the other patient attributes provided by the dataset, it is impossible to estimate an accurate weight for all records submitted.

VIII. CONCLUSION

Within the United States, the general categorization of individual for which an adverse drug reaction report was submitted between January 2016 and December 2018 is a female in their late 50's to early 60's. Of those individuals, they are most likely to either fully recover from an adverse drug reaction or persist with the same drug reaction. Likelihood of dying is increased for the general populous for any individual experiencing an adverse drug reaction, however the majority of adverse drug reactions do not result in death. A number of factors including environmental and biological can affect the outcome of taking a drug, so all conclusions based on this experiment are purely observational in nature.

As identification of higher risk populations was determined and statistically significant, the authors hope that the results from this study may prove valuable to the healthcare industry when determining high risk patients. Utilization of this data could have an effect on the prescription of higher risk approved drugs to the aforementioned higher risk demographics of patients and would hopefully reduce the number of lives impacted by adverse drug reactions, which comprises the fourth leading cause of death in the United States.

REFERENCES

[1] Kongkaew, C., Noyce, P. R., & Ashcroft, D. M. "Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies". Annals of Pharmacotherapy, 42(7–8), (2008): 1017–1025. https://doi.org/10.1345/aph.1L037

[2] Nahler M P G. adverse drug reaction (ADR)[M]// Dictionary of Pharmaceutical Medicine.Springer Vienna, 2009:3-4.

[3] Ahmad, Syed Rizwanuddin,M.D., M.P.H., "Adverse drug event monitoring at the food and drug administration: Your report can make a difference," Journal of General Internal Medicine, vol. 18, (1), pp. 57-60, 2003. Available: https://search-proquest-com.mutex.gmu.edu/docview/923381714?accountid=14541. DOI: http://dx.doi.org.mutex.gmu.edu/10.1046/j.1525-1497.2003.20130.x.

[4] Parameswaran Nair N, et al. "Hospitalization in Older Patients Due to Adverse Drug Reactions - the Need for a Prediction Tool." Clinical Interventions in Aging, vol. 11, Dove Medical Press, May 2016, pp. 497–505.

[5] Martin J. Global institutions: the World Health Organization (WHO)[J]. Bulletin of the World Health Organization, 2009, 87(6):484.

[6] Johnson, J., Bootman, J. Drug-related morbidity and mortality: a cost-of-illness model. Arch Intern Med. 1995:155:1949-56

[7] Ahmad S.R. Adverse drug event monitoring at the Food and Drug Administration[J] Journal of general internal medicine 2003,18(1), 57-60.

[8] China National Food and Drug Administration, National Annual Report on Adverse Drug Reaction Monitoring (2015) [M].http://www.sda.gov.cn/WS01/CL0844/158940.html.

[9] James H. Kim, Anthony R. Scialli. "Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease", Toxicological Sciences, Volume 122, Issue 1, July 2011, 1–6, https://doi.org/10.1093/toxsci/kfr088

[10] Surhone L M, Timpledon M T, Marseken S F. Uppsala Monitoring Centre[M]. Betascript Publishing, 2010.

[11] "Preventable Adverse Drug Reactions: A Focus on Drug Interactions", U.S. Food and Drug Administration, FDA, Available: https://www.fda.gov/drugs/drug-interactions-labeling/preventable-adverse-drug-reactions-focus-drug-interactions

[12] Yuan Peng, Yan Zhong, Fu Liu, "Detection and Analysis of Engel's Net Safety Signal Based on US Adverse Event Reporting System" Journal of drug epidemiology, vol. 27 (8), pp. 509-517, 2008.

[13] Rahman, Md. Motiur et al. "Comparison of Brand Versus Generic Antiepileptic Drug Adverse Event Reporting Rates in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS)." Epilepsy Research 135 (2017): 71–78. Web.

[14] V. L. Yip et al, "Genetics of Immune-Mediated Adverse Drug Reactions: a Comprehensive and Clinical Review," Clin. Rev. Allergy Immunol., vol. 48, (2-3), pp. 165-175, 2015. Available: https://search-proquest-com.mutex.gmu.edu/docview/1672261012?accountid=14541. DOI: http://dx.doi.org.mutex.gmu.edu/10.1007/s12016-014-8418-y.

[15] G. Kaufman, "Adverse drug reactions: classification, susceptibility and reporting," Nursing Standard (2014+), vol. 30, (50), pp. 53, 2016. Available: https://search-proquest-com.mutex.gmu.edu/docview/1814729924?accountid=14541. DOI: http://dx.doi.org.mutex.gmu.edu/10.7748/ns.2016.e10214.
[16] H Fang, et al. "Exploring the FDA Adverse Event Reporting System to Generate Hypotheses for Monitoring of Disease Characteristics." Clinical Pharmacology & Therapeutics, vol. 95, no. 5, Nature Publishing Group, Jan. 2014, pp. 496–98, doi:10.1038/clpt.2014.17.
[17] Hazell, L. & Shakir, S.A. "Under-Reporting of Adverse Drug Reactions". Drug-Safety (2006) 29: 385.
[18] Zhang Dandan, "FDA's Adverse Event Reporting System - How much do you know about FAERS?" Center for International Food, Drug, Policy and Law, Shenyang Pharmaceutical University, https://www.sohu.com/a/148778442_803087
[19] Braun, Lt et al. "Diagnostic Errors by Medical Students: Results of a Prospective Qualitative Study." Bmc Medical Education 17 (2017): n. pag. Web.
[20] Graber ML. The incidence of diagnostic error in medicine. BMJ quality & safety. 2013:bmjqs-2012-001615.
[21] Sankar A, Beattie WS, Wijeysundera DN. How can we identify the high-risk patient?. Curr Opin Crit Care. 2015;21(4):328–335. doi:10.1097/MCC.0000000000000216
[22] Bates DW, Miller EB, Cullen DJ, et al. Patient Risk Factors for Adverse Drug Events in Hospitalized Patients. Arch Intern Med. 1999;159(21):2553–2560. doi:10.1001/archinte.159.21.2553
[23] van den Bemt, P.M.L.A. et al. "Risk Factors for the Development of Adverse Drug Events in Hospitalized Patients." Pharmacy World and Science 22.2 (2000): 62–66. Web.
[24] Lazarou, J., Pomeranz, B.H. & Corey, P.N. "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies". JAMA 279 (1998): 1200–1205
[25] Classen, D.C., Pestotnik, S.L., Evans, R.S., Lloyd, J.F. & Burke, J.P. "Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality". JAMA 277 (1997): 301–306.
[26] Harpaz, R., DuMouchel, W., Shah, N.H., Madigan, D., Ryan, P. & Friedman, C. "Novel data-mining methodologies for adverse drug event discovery and analysis. Clin. Pharmacol". Ther. 91 (2012): 1010–1021.
[27] "Questions and Answers on FDA's Adverse Event Reporting System (FAERS)." U.S. Food and Drug Administration, FDA, www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers.
[28] U.S. Food & Drug Administration. "FDA Adverse Event Reporting System (FAERS) Public Dashboard". (2019) [Online] Available: https://fis.fda.gov/sense/app/d10be6bb-494e-4cd2-82e4-0135608ddc13/sheet/7a47a261-d58b-4203-a8aa-6d3021737452/state/analysis. [Accessed: 03- Nov- 2019]
[29] G. Brolund, "Electronic Transmission of Individual Case Safety Reports Message Specification", Admin.ich.org, 2019. [Online]. Available: https://admin.ich.org/sites/default/files/inline-files/ICH_ICSR_Specification_V2-3.pdf. [Accessed: 03- Nov- 2019].
[30] "About SQLite", Sqlite.org, 2019. [Online]. Available: https://www.sqlite.org/about.html. [Accessed: 03- Nov- 2019]
[31] N. Fusi, R. Sheth & M. Elibol. "Probabilistic Matrix Factorization for Automated Machine Learning", Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1705.05355.pdf. [Accessed 07-Dec-2019].
[32] "Explore Census Data", Data.census.gov, 2019. [Online]. Available: https://data.census.gov/cedsci/. [Accessed: 06- Dec- 2019]
[33] "NCHS Pressroom - 2001 News Release - Women Visit Doctor More Often than Men", Cdc.gov, 2019. [Online]. Available: https://www.cdc.gov/nchs/pressroom/01news/newstudy.htm. [Accessed: 06- Dec- 2019]
[34] "Products - Data Briefs - Number 161 - July 2014", Cdc.gov, 2019. [Online]. Available: https://www.cdc.gov/nchs/products/databriefs/db161.htm. [Accessed: 06- Dec- 2019]