FDA

Data Mining and Applications for Pharmacovigilance

US Food and Drug Administration

December 17, 2015

Authors:

Katie Brochu kebrochu@wpi.edu ______________________

Nicholas Diaz nadiaz@wpi.edu ______________________

Sadie Gauthier sggauthier@wpi.edu ______________________

Erik Nadel einadel@wpi.edu ______________________

In Cooperation With:

Marni Hall, Marni.Hall@fda.hhs.gov
Director, RSS, OSE, CDER

Suranjan De, Suranjan.De@fda.hhs.gov
Deputy Director, RSS, OSE, CDER

Ellen Pinnow, Ellen.Pinnow@fda.hhs.gov
Regulatory Science Program Lead, RSS, OSE, CDER

Proposal Submitted to:

Professor Fred Looft, fjlooft@wpi.edu

Professor Brigitte Servatius, bservat@wpi.edu

Washington D.C. Project Center

This project proposal is submitted in partial fulfillment of the degree requirements of Worcester Polytechnic Institute. The views and opinions expressed herein are those of the authors and do not necessarily reflect the positions or opinions Worcester Polytechnic Institute.

Abstract

In order to ensure the safety and efficacy of post-market pharmaceutical products, the United States Food and Drug Administration relies on its pharmacovigilance efforts and input from the general public. The FDA receives submissions of adverse event reports from patients, health care practitioners and manufacturers. The FDA has started looking to the field of data mining to automate the search for safety signals. A training manual was created to introduce FDA employees to the concepts and applications of data mining techniques in pharmacovigilance.

Acknowledgements

We would like to thank the following people for their support and assistance in the completion of this project.

Fred Looft	Project Advisor
Brigitte Servatius	Project Advisor
Kent Rissmiller	Washington Project Center Director
Marni Hall	Director, RSS, OSE, CDER
Suranjan De	Deputy Director, RSS, OSE, CDER
Ellen Pinnow	Regulatory Science Lead, RSS, OSE, CDER
Carol Pamer	Regulatory Science Lead, RSS, OSE, CDER
Sanjay Sahoo	Operations Research Analyst, RSS, OSE, CDER

We would also like to thank the RSS staff for their support, hospitality and wealth of knowledge provided to us.

Readers interested in more information about this project or the raw data should contact Marni Hall at [Marni.Hall@fda.hhs.gov], Ellen Pinnow at [Ellen.Pinnow@fda.hhs.gov], or Suranjan De at [Suranjan.De@fda.hhs.gov]

Table of Contents

1. Introduction

In 1937, Elixir Sulfanilamide was responsible for over 100 deaths in 15 states across America (FDA, 2015a). Originally released in the form of a tablet, the pharmaceutical Sulfanilamide was highly effective at treating streptococcal infections. In June 1937, a salesman reported that states in the south had a great demand for the drug to be in liquid form (FDA, 2015a). After some experimentation, Harold Cole Watkins, the lead chemist and pharmacist of S.E. Massengill Co., found that Sulfanilamide could be dissolved into diethylene glycol, a substance used as an antifreeze (FDA, 2015a). Without having tested the new formula for toxicity, the company sent 633 shipments of the elixir, now a deadly poison, all over the country (FDA, 2015a). The poisonous nature of the drug was almost immediately noted by doctors and reported to the FDA. After initiating an investigation, the FDA found that the company was aware of the toxic nature of the drug, but had only suggested that its buyers return their shipments and did not indicate the urgency of the situation (FDA, 2015a). As a result, the FDA immediately tracked down all of the salesmen at S.E. Massengill Co., found out where the elixir had been sent, and managed to retrieve 234 of the 240 gallons of distributed elixir (FDA, 2015a). This is an early example of pharmacovigilance – "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems" (World Health Organization, 2015). The practice of pharmacovigilance is essential to identifying adverse events and ensuring issues are identified quickly.

The modern day practices of pharmacovigilance are similar to those of 1937, but work on a much larger scale. Just like in 1937, reports of drug-related health issues (adverse events) are still submitted to the FDA by doctors. Additionally, the general public now has the option to directly submit reports to the FDA if they suspect there to be an issue with a drug or product. But the bulk of modern reports come from the drug manufacturers themselves, who are required by law to report any adverse events of which they have knowledge. These reports are all submitted to the FDA and are filed into a large database known as FAERS (FDA Adverse Event Reporting System). As of 2015, FAERS contains over 11 million reports (Sanjay Sahoo, personal communication, November 13, 2015), and over a million more are added every year (FDA, 2015c). After being filed, the reports are examined by individual reviewers, called Safety

Evaluators, who are responsible for a defined class of drugs. The Safety Evaluators, who are familiar with the current labeling, known adverse events, and mechanism of actions of their drugs, read the reports, look for particular anomalies or issues relative to the normal product safety profile, and check the validity of the report. If the collection of reports is deemed significant due to anomalies or issues after this process, the drug and adverse event relationship is investigated more thoroughly and regulatory action may be taken. (Chen & Scarazzini, 2012)

As explained above, the FDA receives millions of reports about adverse drug events yearly. Processing and analyzing this amount of data is difficult to accomplish, as the number of reports outweigh the number of report reviewers (Safety Evaluators). Every month, each individual Safety Evaluator has to attempt to read and act upon an average of 3,417 adverse event reports (Sanjay Sahoo, personal communication, November 13, 2015). To assist with the analysis of the vast amounts of data, the FDA has begun deploying data mining techniques.

At the FDA, data mining is the practice of using various algorithms and statistical analyses to find patterns within sets of data (Suranjan De, personal communication, September 29, 2015). With data mining, the FDA can improve its report analysis process by automatically selecting the most relevant reports for review (reports that contain serious and unexpected adverse events) as well as allowing reviewers to view the information from all the reports received in an organized manner, instead of having to manually consider each one. However, because data mining is a relatively new practice at the FDA, not all of the employees are trained in data mining and many of them have mixed, unclear definitions of data mining.

Although not yet in routine use for most applications, data mining has been successfully applied by the FDA in past years (Duggirala et al., 2015). For example, in 2010 and 2011, data mining was retroactively used to identify warning signs that associated Fluzone ® with febrile seizures in young children (Duggirala et al., 2015). Researchers calculated an Empirical Bayesian Geometric Mean for each event, which is a value used to determine relevance of reports. This value was adjusted according to the various traits of each report. Next, values fitting within a specific confidence interval were marked for further investigation by reviewers, which may have led to the identification of the safety issue. Although, as this was a retroactive study, it is impossible to say with certainty whether the issue would have been identified, since the issue was already known when the data mining was conducted.

Due to data mining's novelty in pharmacovigilance, data mining results are not depended on alone; instead they are compared to the pharmaceutical knowledge of the FDA Safety Evaluators. As part of the process of testing and moving toward data mining, the FDA has applied its new data mining strategies to existing data in which safety issues had already been identified, to demonstrate the earlier identification of safety issues (Duggirala et al., 2015).

Having identified the potential of data mining in pharmacovigilance, the FDA is in critical need of more staff members with data mining knowledge. Currently, most of the FDA data mining projects are in a development phase and have yet to be integrated into routine operations. Educating employees with basic data mining practices would increase collaboration on data mining projects, allowing data mining experts to more readily collaborate with colleagues with expertise in other fields. With more involvement, projects would be able to proceed out of the development phase and into regular practice more rapidly than if each employee had to be informally educated about data mining by their colleagues.

This project assisted the FDA in conducting pharmacovigilance efforts more efficiently through the development of educational materials on data mining tools and applications. The developed materials were designed to provide clear and consistent definitions of data mining at the FDA and give staff members a clear understanding of how data mining could be applied at the FDA. Along with this knowledge, the educational material briefly discussed tools that were being used by the FDA as well as tools available in the industry at the time. Finally, the educational material pointed to further resources at the FDA that staff members could use to learn more about data mining.

2. Background

2.1 Pharmacovigilance and the FDA

2.1.1 Data Collection

The practice of pharmacovigilance is important because unexpected problems can arise after a drug is released onto the market. Problems that are undetected in the small sample sizes and limited patient demographics of clinical trials may arise once the drug is released into a larger and more diverse population. In addition, unanticipated drug interactions or patients' underlying health conditions could alter a drug's performance causing adverse reactions. In order to ensure the safety of the American public, the FDA collects data on these adverse events, allowing the agency to make informed decisions about what actions are needed to address the long term risks of each drug.

The data on adverse events is collected through the FDA Adverse Event Reporting System (FAERS), through a process outlined in Figure 1.

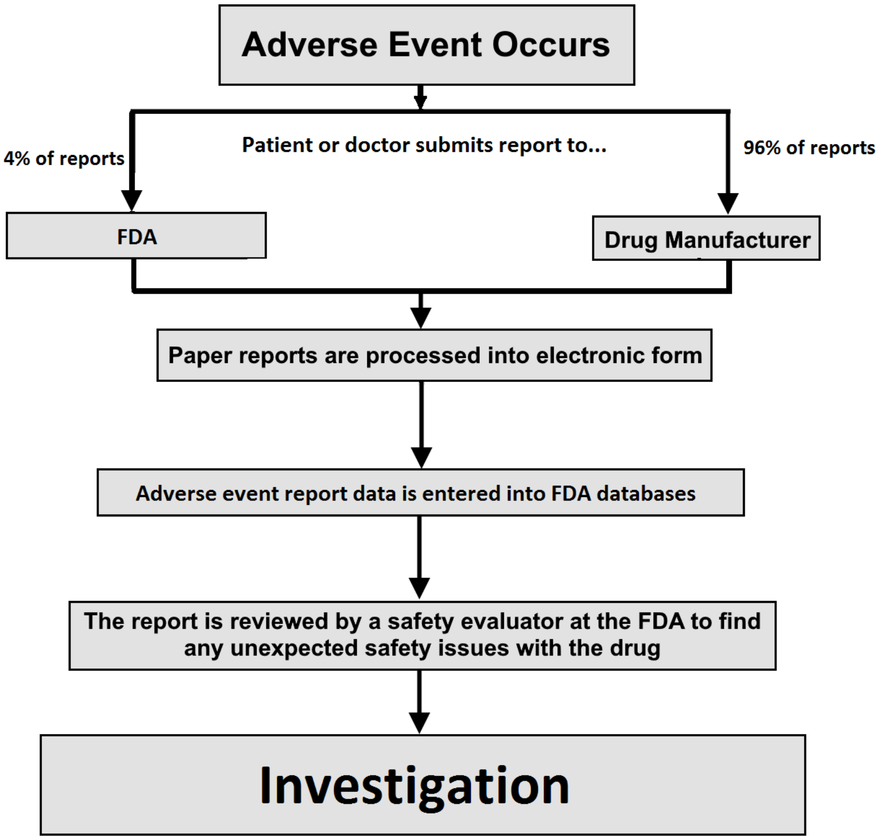

Adverse Event Occurs
Patient or doctor submits report to...
4% of reports
96% of reports
FDA
Drug Manufacturer
Paper reports are processed into electronic form
Adverse event report data is entered into FDA databases
The report is reviewed by a safety evaluator at the FDA to find any unexpected safety issues with the drug
Investigation

Figure 1: Flow chart of the adverse event monitoring process

As shown in Figure 1, doctors, consumers, and manufacturers report directly to this system, providing a detailed description of the adverse events that are believed to be linked to a specific drug. 96% of reports submitted to the FDA come from manufacturers, who receive reports from consumers and doctors. The other 4% are directly reported to the FDA by doctors and consumers (FDA, 2015b).

When a disproportionate amount of adverse events are reported to the FDA, a safety signal is generated. A safety signal is an indication of an abnormal number of adverse events compared to what would be expected with a certain product's use (USDHHS, FDA, CDER, CBER, 2005). The generated safety signal prompts the FDA to initiate an investigation, during which the analysis of patient demographics, length of exposure to the drug, current dosage and any past dosages, underlying health conditions, and the use of other medications are thoroughly examined (USDHHS, FDA, CDER, CBER, 2005). Following the investigation, the FDA recommends appropriate regulatory actions that may include a change in the labelling of the drug, updated communication of safety risks to the public, or the removal of the drug from the market (Fine, 2013).

2.1.2 Reporting Entities

The FDA encourages data input from different entities that may be affected by the activities of the agency such as consumers, nurses, sponsors, pharmacists/pharmacies, physicians, and third party payer. In this case, a sponsor is any individual, company, agency, institution, or organization other than the FDA to submit adverse events that may prompt the initiation of an FDA clinical investigation (USDHHS, FDA, 2015). The reporting entities are encouraged by the agency to use trained health care professionals to assist in reporting adverse events. When this reporting entity is a consumer, it is important for the agency to gain permission to contact the consumer's health care professional to obtain further information related to the patient and the patient's adverse event, as well as relevant medical records. When the reporting entity is a sponsor, the FDA recommends identifying factors that may suggest a causal relationship between the drug and adverse event such as absence of symptoms before taking the drug, consistency of the adverse event with known effects of other drugs or product of the same class, absence of other explanations for the event, and evidence from previous clinical trials and/or case studies (USDHHS, FDA, CDER, CBER, 2005).

2.1.3 Data Processing

Adverse event reports are electronically submitted to the FDA or submitted by paper and entered into the FAERS database. Electronic versions of these reports are saved in the FAERS database and made available for review by Safety Evaluators at the FDA (FDA, 2015b). Pharmaceutical companies and manufacturers are required to submit all of the adverse events that they receive to the FDA, which is known as mandatory reporting, and these reports are classified by the FDA as either expedited or non-expedited (FDA, 2015b). If a report is expedited, it represents a serious adverse event that was not expected (not in the drug's label). If a report is non-expedited, then it was serious, but expected; non-serious and expected; or non-serious but unexpected. All voluntary reports, usually from patients or health care professionals, are called direct reports. If an electronic submission is an expedited report, then the FDA has four days after receipt to process the report (FDA, 2015b). If the report is non-expedited, then the FDA has 30 days to process the report (FDA, 2015b). This processing includes verifying the validity of the case report and coding the data according to the ICH E2B guidelines.

The other form by which a report can be submitted to the FDA is paper. All paper cases are required to be processed by the FDA in 7 days (FDA, 2015b). Unlike electronic reports, all of the paper reports are sent to Landover, Maryland where the FDA has contracted a company dedicated to processing paper reports (FDA, 2015b). The overview of this process can be seen in Figure 2 while Figure 3 highlights the breakdown of the paper processing stages.

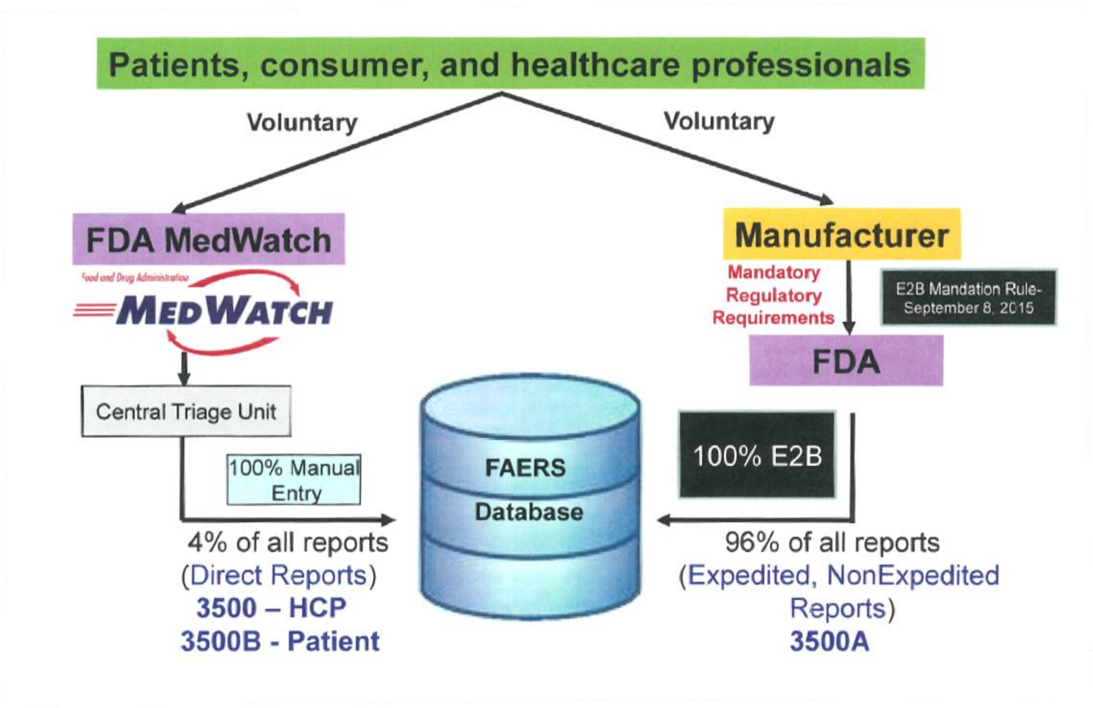


Figure 2: Overview of Report Processing Procedure in Landover, Maryland (FDA, 2015b)

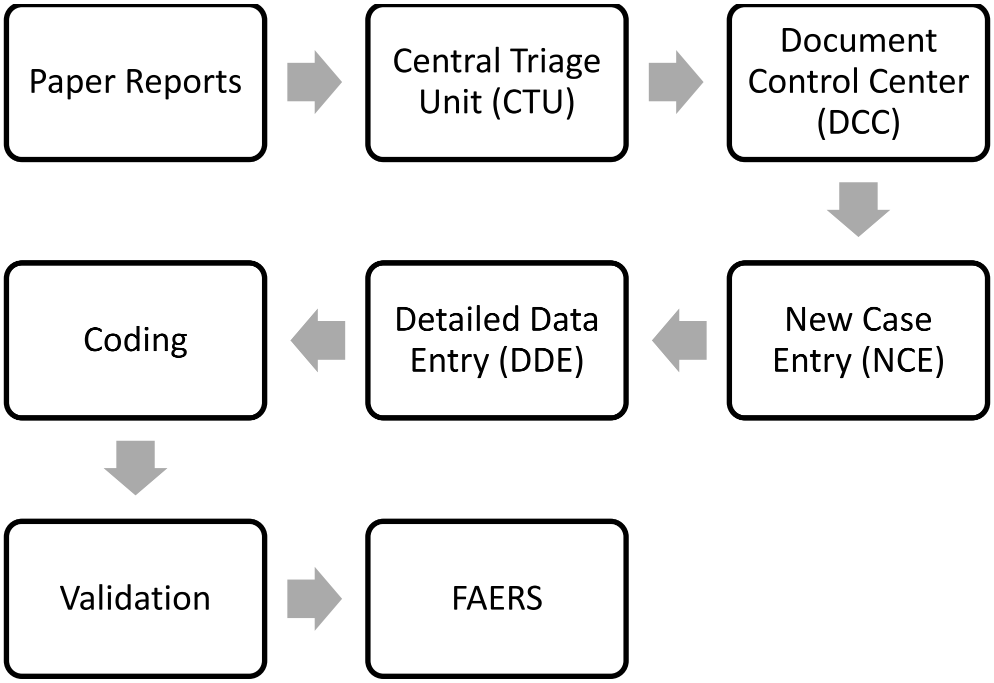

Paper Reports
Central Triage Unit (CTU)
Document Control Center (DCC)
New Case Entry (NCE)
Detailed Data Entry (DDE)
Coding
Validation
FAERS

Figure 3: Stages of processing paper reports at Landover (adapted from FDA, 2015b)

The first step in the processing of paper reports is within the Central Triage Unit (CTU) (FDA, 2015b). Here, “members of the CTU separate, review and sort each report according to the Center(s) responsibilities” (FDA, 2015b). For instance, if a report is received about an adverse event related to a post-market drug, the report is filed under the Center for Drug Evaluation and Research (CDER), whereas a report about a vaccine would be filed under the Center for Biologics Evaluation and Research (CBER). CDER and CBER are responsible for the post-market drugs and vaccines respectively, thus when the processing is completed, the reports will be distributed to Safety Evaluators at those centers for further evaluation. Additionally, in the CTU, the reports are condensed for accuracy and the completeness of the document is verified (FDA, 2015b). Reports are then date stamped for their date of receipt at Landover, photocopied or scanned if necessary, and then sent to the Document Control Center (DCC) (FDA, 2015b). At the DCC the reports are batched based on their case type (expedited, non-expedited, direct). A batch, consisting of 10 reports, receives a barcode sticker to make it easier to locate if an FDA Safety Evaluator requests to see the original, hardcopy report (FDA, 2015b).

The next step for the reports after leaving the DCC is Data Entry. This is currently arranged into two separate phases, New Case Entry (NCE), and Detail Data Entry (DDE) (FDA, 2015b). NCE enters the information in the first half of the report into the system and assigns a case number to each report within a specific batch (FDA, 2015b). The specific information that is entered at this point in the process “consists of initial or follow-up criteria (duplicate check), Sender Organization, FDA received date, form type (expedited, direct, E2B) and FDA center involved” (FDA, 2015b). Next in DDE, the information in the second half of the report is put into the system and the quality and accuracy of previously entered data is checked (FDA, 2015b). It is important to note that all of the information provided in the report is entered verbatim and no interpretations of the information are made by members of DDE (FDA, 2015b).

After all of the information in the report has been entered in the system, the report is sent to the coding team. This team consists of people that are familiar with the specific products and medical terms (FDA, 2015b). The coding team checks the quality and accuracy of the entered information by cross-checking the input information against dictionaries that exist in the FAERS database, in order to ensure that the correct terminology was used (FDA, 2015b). This is also the step where the paper reports are converted to the electronic E2B format (FDA, 2015b). Once in the proper format, the reports are sent to the validation section. This section “consists of

Physicians, pharmacists, nurses who are familiar and understand medical terminology and use the MedDRA dictionary" (FDA, 2015b). It is the job of validators to code the adverse events in the reports according to the E2B guidelines, and MedDRA (FDA, 2015b). This is the final step of the process at Landover, and from this point reports are put into the FAERS database, where they are later reviewed by Safety Evaluators.

In 2014 alone, Landover processed 130,295 paper reports and 10% of all reports submitted to the FDA were still being submitted in paper form (FDA, 2015b). This was very costly to the FDA since the department needed to hire enough people to handle the large amounts of reports, among other operating costs. To create a solution for this problem, the E2B Mandation was established, and as of September 8, 2015 all mandatory reports must be submitted electronically in E2B format (FDA, 2015b). Since mandatory reports are submitted by manufacturing and pharmaceutical companies, who are required to submit all adverse event reports they receive to the FDA, this drastically cut back on the amount of paper reports (FDA, 2015b). After the implementation of the E2B Mandation, the amount of paper submissions dropped from 10% to 4% (FDA, 2015b). This Mandation also led to the further standardization of reporting, as only three forms are accepted. The 3500 form is filled out by health care professionals, while the 3500B form is filled out by patients, and the 3500A form is filled out by manufacturers. 96% of reports filed after the Mandation of September 8th, 2015, use the 3500A form, while only 4% of submissions use the 3500 or 3500B forms (FDA, 2015b). These forms have been included in Appendix D for reference.

2.1.4 Security of Patient Information

A primary concern of the FDA is to maintain the privacy and security of a patient's information. This information includes human subject research and reports submitted by individual patients and practitioners. By establishing a standard to maintain this privacy, the FDA is accomplishing the goal of complying and conforming to any international definitions, laws and standards, as appropriate (USDHHS, FDA, CDER, CBER, 2005). The European Medicines and Heads of Medicines Agencies concur that confidentiality of patient data and documents containing this data should be guaranteed. Going further, these European agencies explain that tracking systems used for systematic documentation and records is an essential requirement for quality standards (European Medicines Agency and Heads of Medicines Agency, 2012).

2.1.4 Safety Signal Identification

Reports and records from consumers and other reporting entities form the foundational data set to identify, interpret, and develop plans to manage safety signals. If a safety signal exists, there are many ways to investigate the signal to determine if there is a potential safety risk. The FDA encourages sponsors to look at all of the various methods for safety signal investigation including, but not limited to, pharmacoepidemiologic studies, registries, and surveys. Pharmacoepidemiologic studies can be experimental or observational in nature and are designed to assess the risk associated with a particular drug and consist of protocols, control groups, and specific hypotheses. Although powerful, these studies can yield conflicting results if the investigators do not minimize bias and document possible errors (USDHHS, FDA, CDER, CBER, 2005). Registries, a second method of investigation, are organized systems for the collection, manipulation, and deliverance of information about individual persons who have been exposed to medical treatment and have a particular disease that predisposes them to this health-related event or previous exposure to substances or circumstances that are known or suspected to cause adverse health effects (The National Committee on Vital and Health Services, 2014). Registries are most useful for obtaining important data that may not be available in large, automated databases or is collected from multiple sources. Finally, surveys can be conducted at the very beginning of the marketing of the drug or when a sponsor wants to evaluate a signal from spontaneous case reports (USDHHS, FDA, CDER, CBER, 2005). All investigative methods may lead to important and credible findings.

2.1.5 Acting on Safety Signals

Based on the results of the safety signal investigation, the FDA can conduct further studies to characterize safety signals and establish whether these signals pose potential safety risks. If there is a potential safety risk, the FDA advises the sponsor to submit all safety data and the analysis methods performed. A complete submission contains all case reports: spontaneous (voluntary reports) and published (case studies and literature); background information for the adverse drug event and specific affected populations; associations made from pharmacoepidemiologic studies (such as odds ratios, relative risks); biologic and pharmacodynamic effects that were observed through preclinical studies; general marketing history of other similar products; and findings from controlled clinical trials. This submission makes it possible for the FDA to assess the level of causality between a particular drug and the associated adverse event (USDHHS, FDA, CDER, CBER, 2005). Additionally, the submission includes recommendations for investigating a specific signal of an adverse event through additional studies and proposes "risk minimization actions" (USDHHS, FDA, CDER, CBER, 2005, p.19).

In addition to the sponsor's submission, the FDA also provides its own assessment of the identified signal in question. The submissions to the FDA aim to formally assess the potential safety risk posed and take into account the sponsor provided information and any other relevant information known to the FDA.

There are specific points that the FDA considers while compiling the received submissions. These points include the strength of the relative risk of the adverse event associated with the specific drug, the consistency of findings from various data sources, biological plausibility, the gravity of the adverse event in regards to the condition being treated, whether further studies would be reasonable to pursue, the level of advantages the drug provides, and availability of other therapies (USDHHS, FDA, CDER, CBER, 2005).

For many drugs and products, manual pharmacovigilance activities are sufficient for risk assessment of drugs that are post market. However, in certain circumstances, a sponsor may be advised to create a pharmacovigilance plan focused on detecting and analyzing safety risks for a specific drug or one that is suspected of having additional monitoring needs (USDHHS, FDA, CDER, CBER, 2005). These plans are designed to enhance a sponsor's collection of safety information and describe necessary efforts beyond standard reporting, which can often be inconsistent. The pharmacovigilance plans are especially useful when a new drug is launched or when a health risk is identified during the marketing stage. When new data and data sources emerge, the FDA recommends that sponsors re-evaluate their pharmacovigilance plan and its effectiveness (USDHHS, FDA, CDER, CBER, 2005).

2.2 Data Mining and its Applications

2.2.1 Big Data and Data Quality

Big Data is a broad term that refers to large-volume, complex, growing data sets with multiple, autonomous sources (Xindong, Xingquan, Gong-Qing, & Wei, 2014). Data sets of this type have become more and more prevalent as use of the internet increases globally, providing companies and organizations with more information on consumers than ever before. The term "big" also invites quantification and gives difficulty in defining a concrete definition (Ward & Barker, 2013). Another definition is, "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" (Jacobs, 2009). For instance, the FAERS database contains approximately 1.68 terabytes of information (assuming 11,198,975 reports of size 150kb each) (Sanjay Sahoo, personal communication, December 4th, 2015). Although this data may not be as massive as some datasets in industry, the "tried-and-true" method of manually evaluating each report at the FDA is unable to keep up with the high volume of reports received.

When dealing with data approaching a size large enough to warrant an investigation of alternate methods, quality becomes an important factor in ensuring that the newly proposed methods perform just as well as the current "tried-and-true" methods. At the FDA, having quality data is essential to the accuracy of operations in the field of pharmacovigilance. There are three main risks that are assessed in determining the quality of big data: errors and inaccuracies in the data itself; sources and pedigree (background) of the data; and underlying purpose of data collection (Sukumar, Natarajan, & Ferrell, 2015). Errors and inaccuracies would include entry errors, missing data fields and errors from extracting and transforming data for analytics (Sukumar et al., 2015). In analyzing the source of the data, the source reveals the limitations and appropriateness to the type of analysis being performed (Sukumar et al., 2015). In other words, it is important to ensure that the source of the data is relevant to what you want to learn from your data. For example, if you tried to use baseball statistics to research cancer, you will likely not get any meaningful results. Further inaccuracies can be found in the source by analyzing the underlying purpose of the data collected by the source. This will affect the quality when the data collection procedure involves modifying data before storage(Sukumar et al., 2015). As an example, data collected with an attempt to preserve privacy could remove data critical to analysis, such as patient ages or genders, which would result in the inability to analyze the correlation among patient backgrounds.

2.2.2 KDD Process

The knowledge discovery in databases (KDD) process refers to, "the nontrivial extraction of implicit, previously unknown and potentially useful knowledge from data" (Verma, 2015). KDD is comprised of numerous steps: data preparation, searching of patterns, knowledge evaluation, and refinement, all repeated in multiple iterations (Fayyad, 1996). Knowledge in the definition of KDD means, “relationships and pattern between data elements” (Verma, 2015). The end goal of the KDD process is to extract high-level knowledge from low-level data. It is expected that the extracted knowledge will be beneficial to the user or task and understandable either immediately or after post processing. This knowledge comes from patterns found in data. A pattern discovered from the KDD process is considered to be knowledge if it exceeds an “interestingness threshold”, which is to say that the pattern provides new insight into the data set (Fayyad, 1996). This threshold is defined quantitatively or qualitatively depending on where the KDD process is being applied (Fayyad, 1996).

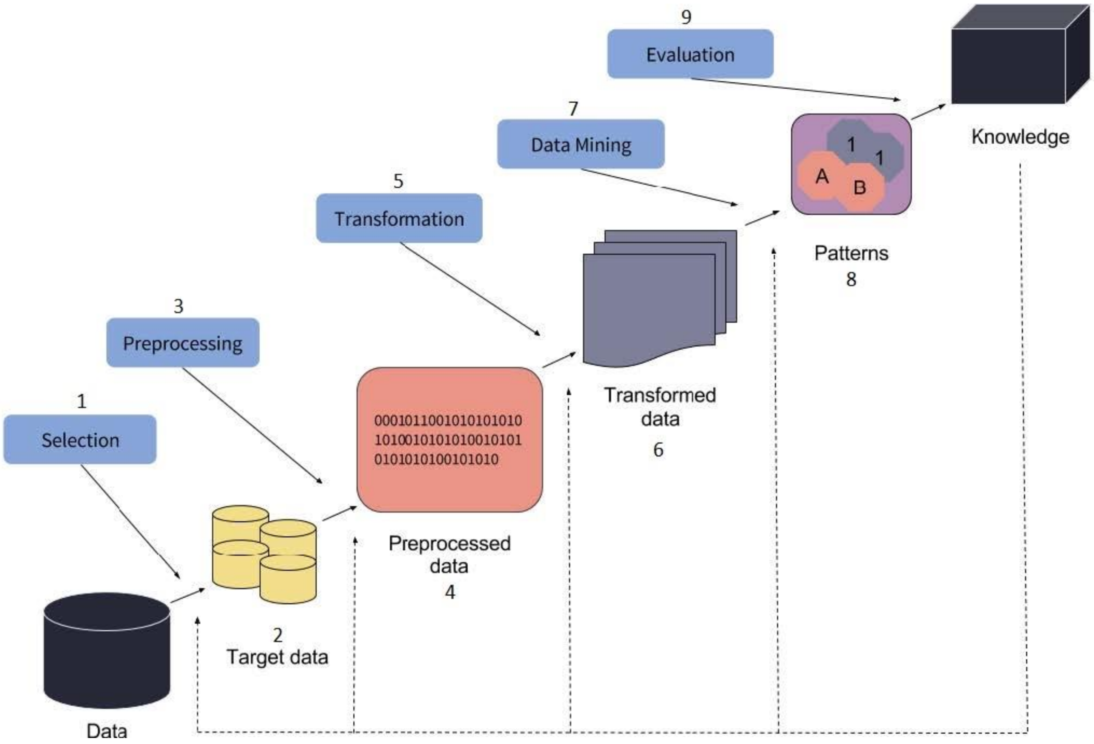


Figure 4: The KDD process diagram (Adapted from Fayyad, 1996, Figure 1)

There are nine steps in the KDD process as shown in Figure 4. The figure shows the flow of the KDD process and the dotted arrows represent the reversibility of transitions to each step. The reversibility is necessary to obtain the most accurate knowledge from the data because it is common to move backwards through the process and refine each step (Carolina Ruiz, personal communication, October 6, 2015). The first step of the process is to develop an understanding of the application domain, relevant prior knowledge, and identify the goal of the KDD process (Fayyad, 1996). An example of this first step would be researching the review process at the FDA and how data is collected and processed to understand the goal of applying KDD to safety signal detection.

Once an understanding of the application domain has been established, step two is the selection of a target data set that will be used in the remainder of the KDD process (Fayyad, 1996). Knowledge of the application domain is important for the second step in order to have insight on which data set could potentially be analyzed for new knowledge. The third step is to clean the target data set and pre-process the data. This step includes removing outliers, noise, and handling missing data fields (Fayyad, 1996). After cleanup, the fourth step involves data reduction and projection: finding useful features to represent the data according to the goal established in step two. With the data cleaned and fields chosen, the goals established in step one are matched with a particular data-mining task in step five (Fayyad, 1996).

The data-mining task is the method that will be used to mine the data. The task could be either predictive or descriptive (Fayyad, 1996). Either the goal is to calculate predictions from the target data, or the goal is to describe the target data from a different perspective. For example, the chosen data mining task could be to predict the amount of pens that will be used by employees given current data on usage. An example of a descriptive data-mining task would be disproportionality analysis, which is the analysis of events that occur disproportionally relative to a specific collection of events.

The next step is to choose the data mining methods that will accomplish the chosen task in the previous step. As an example, if the task chosen from step five was classification and the data set being used was particularly small, one might choose a method that was developed for analyzing small datasets over another method that might not work well with small datasets. The seventh step is the data mining itself: searching for patterns of interest in the target data using the methods you have selected (Fayyad, 1996). Once the data has been mined, the mined patterns must be interpreted with the previously mentioned possibility of returning to any of the previous steps for further iteration (Fayyad, 1996). The final step in the KDD process is to act upon the discovered knowledge. This includes using the knowledge directly, incorporating the knowledge into another system, or simply documenting the knowledge. As with any process, the final step also includes checking for and “resolving conflicts with previously known (or extracted) knowledge” (Fayyad, 1996).

2.2.3 Intro to Data Mining

The seventh step of the KDD process, data mining, is defined as “searching for patterns of interest in a particular representational form or a set of representations”(Fayyad, 1996). This definition is one of many definitions of data mining. Data mining can be defined differently for any discipline, due to the varying types of data among the various disciplines. In the field of pharmacovigilance at the FDA, data mining is the searching of patterns and clusters in sets of adverse event reports (Suranjan De, personal communication, September 29, 2015). The fundamental goal of pattern searching is similar across disciplines, but the term for ‘data’ is generic and defined differently among each discipline.

Just as mining implies a repetitive process of extracting from a substance, data mining involves the repeated application of data mining methods to extract knowledge from data. Most data mining methods are based on techniques from machine learning, pattern recognition, and statistical methods. Some of the categories of these data mining methods are classification, regression, and clustering (Fayyad, 1996). These data mining methods help achieve the goal of either developing a prediction or description from data.

Classification is used in the development of a prediction or description by mapping (classifying) data into predefined classes. Classification methods could be used for classifying trends or automated identification of entities. An example of classification being used as a predictor would be if a bank was trying to decide if future loan applicants would be eligible for a loan if they were to apply (Fayyad, 1996). The predefined classes allow easily interpretable data from the start of the data mining process because the user defines the classes. However, sometimes the classes to be mapped might not be able to be defined by the user.

For the case of undefinable classes, regression is designed to work within this limitation. A regression function builds the classes based on current data and creates a regression line that distinguishes between the classes (Fayyad, 1996). The more data the function receives to create the classes, the more accurate the function will be in predicting the class of new data. This is assuming the data input into the function is accurate.

When development of a prediction is not the previously decided goal of the KDD process, clustering can used as a descriptive task to represent data from new perspectives.

Clustering is a task where the goal is to identify finite sets of categories or clusters to describe data. Like a Venn diagram, categories can be mutually exclusive or consist of overlapping categories (Fayyad, 1996). An example of clustering would be identifying subclasses of events based on discovered overlapping criteria. With data in distinguished categories, the task of describing a target set of data becomes simpler with the ability to easily describe differences within the data.

Beyond clustering, regression, and classification, more methodologies exist for either creating predictions or descriptions from data. It is critical to understand that the quality of the descriptions or predictions is not defined by the method chosen, but by the quality of the chosen target data set. Inaccurate data will give inaccurate predictions or descriptions which is why it is important that the KDD process is able to be iterated in reverse to correct the data used in the function.

2.2.4 Statistical Data Mining Methods in Pharmacovigilance

In pharmacovigilance, data mining is primarily used as a descriptive task to uncover links, patterns, and similarities, allowing for clear analysis. This section covers four main statistical data mining algorithms useful in pharmacovigilance (PV): Proportional Reporting Ratio, Reporting Odds Ratio, Information Component, and Multi-item Gamma-Poisson Shrinker (Empirical Bayesian Geometric Mean) because they calculate signals of disproportional reporting (SDRs). The category of disproportionality analysis goes along with the other previously mentioned categories of data mining techniques (clustering, regression, classification, etc.). These algorithms were specifically developed to identify drug-associated adverse events based on disproportion, otherwise known as SDRs (Sakaeda, Tamon, Kadoyama, & Okuno, 2013).

Proportional Reporting Ratio (PRR) represents a direct measure of the strength of a safety signal. A PRR can be described as, “the ratio of the proportion of all reported cases of the event of interest among people exposed to a particular drug compared with the corresponding proportion among people exposed to all or several other drugs”(Rothman, Lanes, & Sacks, 2004). The breakdown of PRR can be seen in the equation:

$$PRR = \frac{(\frac{A}{A+B})}{(\frac{C}{C+D})} \tag{2.1}$$

where:

$A = Number\ of\ cases\ with\ suspect\ medicinal\ product\ P\ involving\ adverse\ event\ R$

$B = Number\ of\ cases\ with\ product\ P\ but\ not\ including\ R$

$C = Number\ of\ cases\ involving\ R, but\ not\ P$

$D = Number\ of\ cases\ not\ including\ R\ and\ P$

(Zorych, Madigan, Ryan, & Bate, 2013).

The main advantage of using PRR is that it is derived solely from spontaneous (voluntary) Adverse Drug Report (ADR) data and is simple to calculate and interpret. Another advantage is that the underreporting of adverse events will not influence PRR (Rothman et al., 2004). These advantages are important due to the dynamic nature of FAERS data and the need to recalculate frequently as more potential signals emerge. This algorithm also helps with avoiding biases caused by varying details in reports (Evans, Waller, & Davis, 2001). A limitation of PRR is that signals for a particular drug might reduce the magnitude of the PRR calculation for other signals of the same drug. This is due to the fact that some reports of a particular kind might appear more than others if the symptom is more common. (Evans et al., 2001).

Reporting odds ratio (ROR), is closely related to PRR with a few beneficial differences. Fundamentally, ROR is calculated in the same manner as PRR, but ROR accounts for bias and allows for relative risk assessment. The difference can be seen in the equation:

$$ROR = \frac{A/C}{B/D} \tag{2.2}$$

where the numerator is the ratio of the cases involving product P and R over cases involving R not including P (Zorych et al., 2013).

Information component (IC) is a component of IC temporal pattern discovery (ICTPD).

This algorithm is based on intra-personal comparison of risk periods and the preceding control period. ICTPD focuses on the exposure to a certain drug as seen in the equation (Zorych et al., 2013):

$$IC = \ log_2 \frac{A \times (A + B + C + D)}{(A + D) \times (A + B)} \tag{2.3}$$

ICTPD uses information from non-cases such as prescription information. The goal of this technique is to identify patterns in the associations between the prescription of a drug and the occurrence of a medical event.

Multi-item Gamma-Poisson Shrinker (MGPS) is calculated in a similar manner to PRR, but incorporates Bayesian "shrinkage" and stratification to produce scores where there is limited data and small number of cases (Hesha J. Duggirala et al., 2015). Bayesian "shrinkage" can be summarized as the improving of an estimate by combining the estimate with other information. Stratification is a procedure for mitigating effects of confounding by adjusting for associations between a drug and a variable and an event and the same variable (Almenoff et al., 2005). The differences in MGPS from PRR diminish the effect of outliers, reducing the number of false-positive safety signals. As a result, MGPS provides a more stable estimate of the relative reporting rate for a particular product.

The 4 algorithms developed to identify drug-associated adverse events were all developed to calculate signal scores (to assess whether a drug is associated with an adverse event or not). These algorithms are also known as signal detection algorithms (SDAs) (Rave Harpaz et al., 2013). With an identification of similar patterns, it would seem unnecessary to have developed multiple algorithms for the same goal. However, the difference lies in the scoring thresholds of the algorithm. Where the threshold is used to identify signals necessitating further review (Rave Harpaz et al., 2013). In choosing an algorithm, some have argued that the most important question is not which algorithm to use but what is the correct threshold (Balakin & Ekins, 2009). The benefit of using multiple algorithms is that one may catch a signal that the other does not. To summarize the differences between the algorithms, the algorithms that are frequentist detected a higher number of safety signals than the Bayesian based algorithms (Sakaeda et al., 2013). This comparison is only relative to a specific comparison of signals detected from handpicked drugs. It is noted by Bate and Evans that, "different algorithms have slightly different properties and consequently one might be preferable in a particular application" (Sakaeda et al., 2013).

Performance in these algorithms can be defined by sensitivity and specificity. In the context of mining adverse event signals, sensitivity is defined as, "the ability of a surveillance or reporting system to detect true health events, i.e. the ratio of the total number of health events detected by the system to the total number of true health events as determined by an independent and more complete means of ascertainment" (World Health Organization, 2015). Specificity is defined as, "a measure of how infrequently a system detects false positive health events, i.e. the number of individuals identified by the system as not being diseased divided by the total number of all person who do not have the disease" (World Health Organization, 2015). There is a tradeoff between these two performance traits. The more specific an algorithm is the lower the sensitivity and the slower the production of true signals of disproportionate reporting (SDR). The less specific an algorithm, the greater the sensitivity and faster production of true SDRs. When choosing an algorithm, these tradeoffs must be considered depending on the desired usage of a system. If a system is to be used to detect real-time signals, a less specific algorithm would need to be used. Otherwise, if a system is to be used as a passive monitoring system, a more specific algorithm would need to be used.

Data mining algorithms are becoming more frequently used as a supplement to traditional expert reviews of reports and to rapidly analyze the large volume of accumulated data (Rave Harpaz et al., 2013). New algorithms are constantly being researched to uncover new trends and associations in data or to improve upon existing algorithms. These algorithms could be routinely applied in order to monitor, prioritize, and identify undiscovered safety signals of adverse drug events that warrant further attention (Rave Harpaz et al., 2013). Given the role of data mining algorithms in PV, the Observational Medical Outcomes Partnership is aiming to identify the most reliable algorithms for analyzing large volumes of electronic healthcare data specifically for drug safety surveillance (Rave Harpaz et al., 2013).

2.2.5 Comparison of Statistical Data Mining Methods

Both internal (FDA) and external research that has been unable to determine a clearly superior method for data mining adverse event reports, (Hauben, Madigan, Gerrits, Walsh, & Van Puijenbroek, 2005). Though method comparison at the individual level is inconclusive, progress has been made in identifying the advantages, disadvantages and differences between the frequentist methods and the Bayesian methods (Rave Harpaz et al., 2013).

The group of frequentist methods consists of: Proportional Reporting Ratio (PRR), Reporting Odds Ratio (ROR), and Relative Reporting Ratio (RRR) The frequentist methods use ratios to find and estimate associations and are typically accompanied by hypothesis tests for independence (e.g. chi squared test, Fisher's test) and these tests are used as extra precautionary measures that take into account the sample size used while computing the association (R Harpaz et al., 2012). The group of Bayesian methods includes: Gamma-Poisson Shrinker (GPS), Multi-item Gamma-Poisson Shrinker (MGPS) and Bayesian Confidence Propagation Neural Network (BCPNN). Methods are categorized as Bayesian if the data mining method incorporates both the disproportionality measure, the measure of how much the drug-event combination occurs "disproportionally" compared to if there was no association between the drug and event, and sample size to "shrink" the disproportionality measure toward the baseline case of no association by an amount proportional to the variability of the measure (R Harpaz et al., 2012). This shrinkage is an attempt to account for the uncertainty of the disproportionality measure due to low-frequency reporting (R Harpaz et al., 2012). While there has been no consensus on which group is truly "superior," research has found differences between the two groups that can be important when deciding which type of method to use for different applications. Some general advantages and disadvantages of each group are highlighted in Table A. Note that DEA stands for drug-event association in the context of this table.

Table A: Comparison of Frequentists and Bayesian Methods (Balakin & Ekins, 2009), (Deshpande, Gogolak, & Smith, 2010), (Gravel), (Harpaz et al., 2012), (Hauben, Madigan, Gerrits, Walsh, & Van Puijenbroek, 2005), (Johnson, Guo, Gosink, Wang, & Hauben, 2012)

	Frequentist Methods	**Bayesian Methods**
Tend to highlight a greater number of DEAs	X	
Tend to highlight a greater variety of DEAs	X	
Tend to highlight DEAs earlier	X	
More computationally intensive		X
More sensitive to low-frequency of reports	X	
More intuitive computations	X	
Ability to sort associations along one single dimension		X
Address reporting biases or confounding		
May result in loss of credible signals		X
Lower impact of random fluctuations of relative reporting ratio (“shrinkage”)		X
Produce more false positives	X	
Produce more false negatives		X

This table summarizes the general trends of frequentist methods (e.g. PRR, ROR, IC) and Bayesian methods (e.g. BCPNN, MGPS).

In general, the frequentist group seems to highlight a greater number and variety of drug-event associations (DEAs) than the Bayesian group and tend to highlight these DEAs earlier as well (Hauben et al., 2005). These additional DEAs identified by the frequentist group are caused by confounding data or statistical noise especially at low-frequency reporting and thus require additional filtering by a Safety Evaluator (Hauben et al., 2005). For example, it has been observed that for low-frequency reporting, the frequentist methods are more prone to extreme values and therefore can generate more false positives (Hauben et al., 2005). As a result, frequentist group methods are often seen as more unstable due to the fact that the increase in the detection of signals is accompanied by an increase in the detection of noise (Johnson, Guo, Gosink, Wang, & Hauben, 2012). By comparison, the Bayesian methods group addresses the low-frequency reporting issue by adjusting the disproportionality measure to account for these low counts (Deshpande, Gogolak, & Smith, 2010). However, Bayesian methods have been shown to be less sensitive for detecting new signals in low-frequency reporting, implying that these methods can "overshrink" (Johnson et al., 2012). Which leads researchers to argue that Bayesian methods are "too conservative" and delay the detection of novel adverse drug events (R Harpaz et al., 2012). Research has shown that both the frequentist and Bayesian groups produce similar results for higher-frequency drug-event combinations (Johnson et al., 2012).

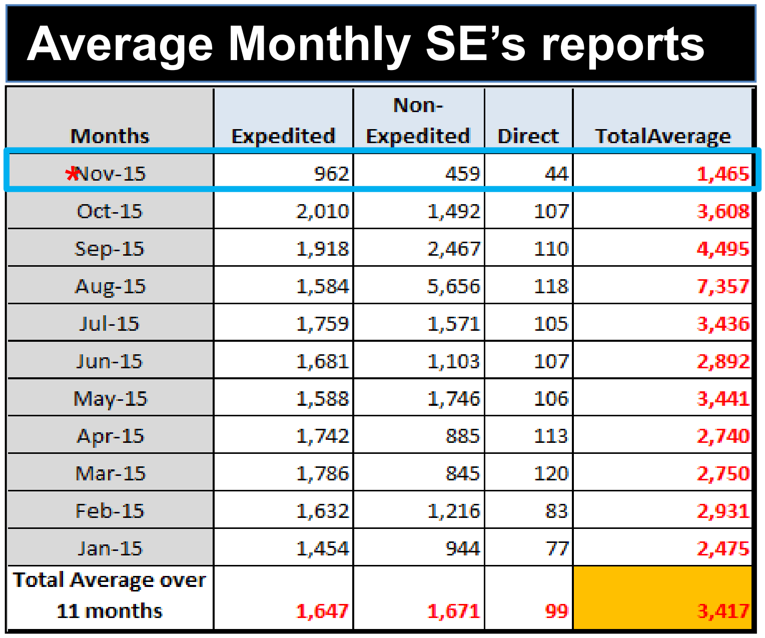
Though noisy, generally analysts prefer the frequentist methods over the Bayesian methods because they are more intuitive, easier to compute (Johnson et al., 2012) and less labor intensive (Balakin & Ekins, 2009). In the world of pharmacovigilance, this is understandable due to the fact that the Safety Evaluators, who are highly educated in many fields such as epidemiology, biology, pharmacology, etc., may not be knowledgeable in the field of data mining or statistics. It is tempting for analysts to focus solely on reducing the numbers of false-positive and false-negative signals. However, a balance between sensitivity and specificity is crucial for optimal signal detection (Balakin & Ekins, 2009). To improve optimality of signal detection, an analyst must know which data mining methods group to use in different scenarios they may face. For example, knowing the fact that frequentist methods tend to be unstable during low-frequency reporting indicates that in such a situation, an analyst would use a Bayesian method instead (Johnson et al., 2012).

As noted in previous sections, the FDA's intake of reports has been exponentially increasing over the years while the number of Safety Evaluators has remained relatively constant. Tables B and C illustrate the overwhelming number of reports that each Safety Evaluator must view.

Table B: Total Number of Reports (Sahoo, 2015).

Total Reports ***All Versions of Report**	**Total Reports for SE Review** ***Only latest version of the reports**
11,198,975	**8,435,279**
*Data as of 09 November 2015	

Table C: Average Monthly Safety Evaluator Reports (Sahoo, 2015).



Average Monthly SE's reports

Months	Expedited	Non-Expedited	Direct	TotalAverage
*Nov-15	962	459	44	1,465
Oct-15	2,010	1,492	107	3,608
Sep-15	1,918	2,467	110	4,495
Aug-15	1,584	5,656	118	7,357
Jul-15	1,759	1,571	105	3,436
Jun-15	1,681	1,103	107	2,892
May-15	1,588	1,746	106	3,441
Apr-15	1,742	885	113	2,740
Mar-15	1,786	845	120	2,750
Feb-15	1,632	1,216	83	2,931
Jan-15	1,454	944	77	2,475
Total Average over 11 months	1,647	1,671	99	3,417

**snapshot of 8 DPV Safety Evaluators reports data as of 13th Nov 2015 and also partial data presented for the month of Nov 2015*

Research suggests focusing more on specificity than sensitivity in data mining (Hauben et al., 2005). Though choosing one group to be "superior" has proven to be impossible, one conclusion can be made: both groups, Bayesian and frequentist, should be used in combination with additional filters when being implemented into the overall signal detection process (Balakin

& Ekins, 2009) since neither group addresses the reporting biases or confounding (R Harpaz et al., 2012).

2.2.6 Data Mining Score Interpretation Pitfalls

Though properly applied data mining is a proven method for identifying relationships and extracting information form large data sets, data mining is imperfect, and the methods have a few caveats. One of the most significant of these caveats is the reliability of the drug safety scores. There are many outside factors and biases that can affect the scoring process, that the resulting score is not completely reliable. For instance, a high score does not necessarily indicate a causal relationship between a drug and adverse event, while a low score does not rule out the possibility of a safety issue (Tonning, 2015). The determination of these scores can be influenced by a number of factors, one of which is drug publicity (Tonning, 2015). Publicity has the ability to draw a large amount of attention to adverse events, and cause a potentially unnecessary panic in the public. As an example, if a drug is linked to suicide, aggression, or violent tendencies in certain patients, this drug is likely to receive a disproportionate amount of media attention and thus more reports from concerned patients (Tonning, 2015). Violent or otherwise extreme symptoms may cause panic in the public concerning the safety of the drug, which could result in an exaggeration of the real risk. While these symptoms may have happened in only one patient or a small population of patients, they are not always representative of a much larger number of people that are taking the particular drug. The added publicity and increased reporting will result in a higher data mining score, but is not necessarily indicative of a serious safety issue.

Another factor that has the ability to influence a score is litigation (Tonning, 2015). The data mining algorithm used by the FDA removes these litigation cases, civil lawsuits filed against a pharmaceutical company due to the effects of a drug, but does not remove consumer reports that were influenced by publicity (Tonning, 2015). This introduces a level of bias into the resulting score, because cases with support are removed by the algorithm, while less reliable reports are kept in the system.

Other influential factors in a resulting data mining score are underlying diseases or conditions, as well as concomitant medications (Tonning, 2015). For example, if a drug is labelled as having negative effects on the liver, and the patient is unaware of an underlying liver condition, then the drug could cause a severe adverse event that the patient might believe was caused by the drug alone. If reported, this would influence the data mining score for the drug, but is not necessarily indicative of the drug’s safety. Along those same lines, concomitant medications could interact with the drug and cause adverse events that the patient may not have experienced had they not mixed medications.

Finally, underreporting of an adverse event could result in a lower data mining score, indicating that a safety signal is not present when this is not necessarily true. Data mining cannot mine data that is not present, so if patients are not reporting the adverse events they are experiencing, the safety signal cannot be identified as quickly.

2.3 Data Mining and the FDA

2.3.1 Data Mining and Pharmacovigilance

In pharmacovigilance, data mining is used to assist the evaluation process in several ways: prioritizing reports, analyzing drug-drug interactions, and evaluating both familiar and unfamiliar classes of drugs. Prioritizing safety reports to read is essential for a reviewer to pick out noise and locate reports that could be the source of a safety signal (Cindy Kortepeter, personal communication, November 17, 2015). Analysis of drug-drug interactions can help point out safety signals that might not be found if only performing data analysis on one specific drug. Evaluating a class of drugs, perhaps unfamiliar to a new Safety Evaluator, is useful in becoming aware of a class specific trend of Adverse Events (AEs) (Cindy Kortepeter, personal communication, November 17, 2015).

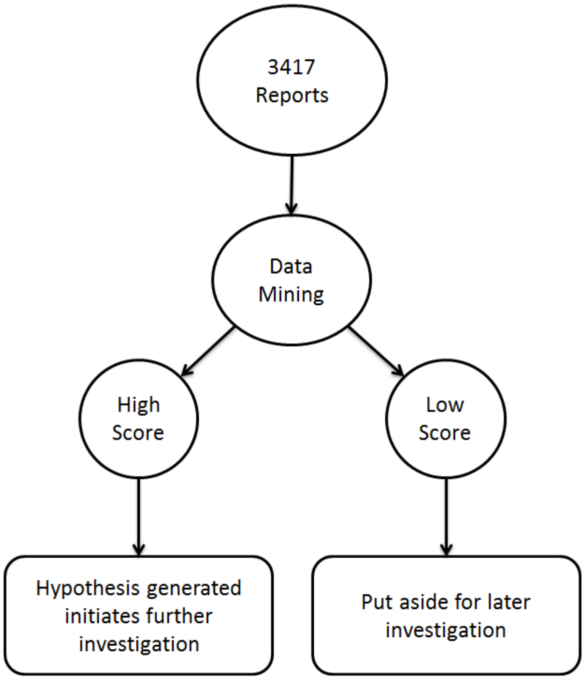


Figure 5: Hypothesis Generation

Data mining is needed in pharmacovigilance to analyze the increasing number of reports received, speed up the identification of potential safety issues, aid in hypothesis generation (as seen in Figure 5) (Tonning, 2015), and free personnel to devote more time to in-depth evaluation (Hesha J Duggirala et al., 2015) (Fine, 2013). Because the number of reports is growing exponentially, it is challenging for Safety Evaluators to view all of the reports within the mandated time constraints (Sanjay Sahoo, personal communication, November 13, 2015). Since not all of the reports are able to be read, reports that point out a potential safety signal might not be found (Sanjay Sahoo, personal communication, November 13, 2015). With data mining, these unviewed reports can be analyzed and used to form a basis that aids evaluators in creating a hypothesis of where potential safety signals might be. Additionally, by easing the amount of manual review that has to be conducted, data mining can give Safety Evaluators more time to focus their efforts on other time sensitive tasks. These benefits all contribute to giving personnel more time to dedicate to critical tasks such as investigating signals.

2.3.1 FAERS

The FDA Adverse Event Reporting System (FAERS) is the FDA’s post-market safety surveillance database (Tonning, 2015). The database contains information found in adverse event reports and medication error reports that are submitted to the FDA (Holloway, 2013). These reports can be submitted by paper or electronically (U.S. Food and Drug Administration, 2013) by manufacturers, patients, doctors, pharmacists, etc. (Tonning, 2015). Implemented in 2012, FAERS was designed to support the post-market safety surveillance program for drugs and therapeutic products. The database contains the validated and recoded information found in the Adverse Event Reporting System, the previous reporting database (Holloway, 2013). The number of reports submitted to the FDA and entered into FAERS has increased over recent years, see Figure 6. For example, in 2004, the total number of reports entered into FAERS was 422,307 and in 2013 the number of reports climbed to 1,178,306 (U.S. Food and Drug Administration, 2013). FAERS is structured in accordance to the international safety reporting guidelines issued by the International Conference on Harmonisation (ICH) (Holloway, 2013).

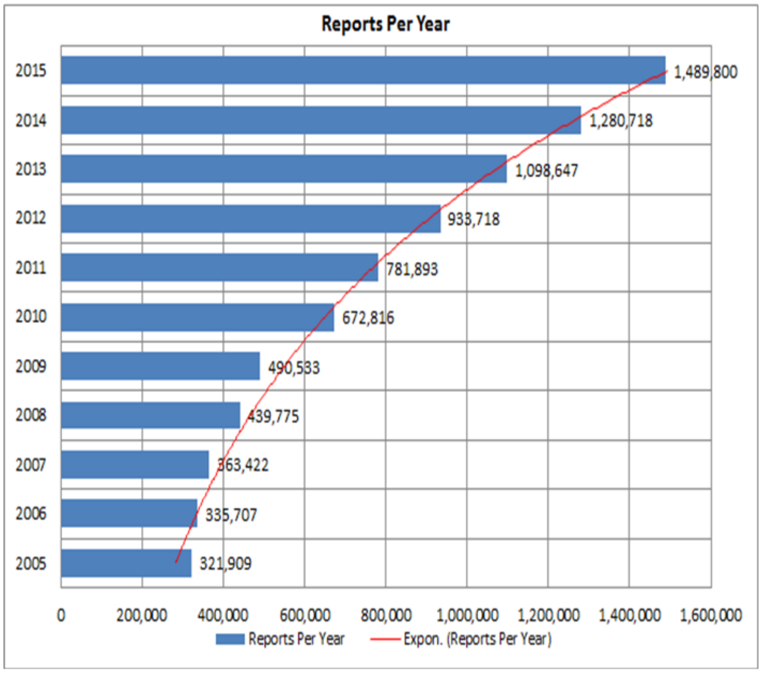


Figure 6: Number of Reports Submitted to the FDA Yearly (Sahoo, 2015).

2.3.2 Data Formatting of Individual Case Safety Reports

To make the process of data mining in pharmacovigilance easier, it is important that all of the data reported through individual case safety reports (ICSRs) is in a standardized format. This standardization allows for the easy exchange of data between reporting sources, regulatory authorities, pharmaceutical companies, and clinical investigators (Brolund, CDER, ICH M2 EWG, 2000). This ease of data exchange is especially important in s sensitive cases, since it would require additional time and effort for each of the recipients to translate the data into their own separate version of reporting, before reviewing it. This need for standardization prompted the International Conference on Harmonisation (ICH) to establish the E2B (R3) reporting guidelines, which are currently used by Japan, the European Union (EU) and the United States. All of the complying regulatory agencies in these areas require reporters to submit the same types of information, which is then translated into a corresponding string of numbers, or an object identifier (OID), that is unique to each entry (Brolund, CDER, ICH M2 EWG, 2000). Additionally, the guidelines eliminate the issues that result from language barriers, since the information in the report does not need to be translated from one language to another.

An important aspect of the standardization of data is the minimum information requirement for submitted ICSR. To be considered valid, all reports must at least include an identifiable patient, an identifiable reporter, an adverse event or reaction, and finally one suspect or interacting drug (Brolund, CDER, ICH M2 EWG, 2000). There is also administrative information that is required so that the ICSR can be properly processed by the agency to which it was submitted. Some of these administrative information requirements include the type of report, the sender's organization, and the sender's safety report unique identifier so that the case can be updated if another report is submitted by the same person about the same drug in the future (Brolund, CDER, ICH M2 EWG, 2000). These requirements make it easier to compare the data reported to multiple agencies, since all of the reports include the same minimum information.

Another form of standardization is coding of the input data. When the coding process begins, all of the non-medical terms remain in their original forms while the medical terms are run through the *Medical Dictionary for Regulatory Activities* (MedDRA).The MedDRA dictionary was developed by the ICH and includes "medical terminology used to classify adverse event information associated with the use of biopharmaceuticals and other medical products (e.g. medical devices and vaccines)" (Brolund, CDER, ICH M2 EWG, 2000). These MedDRA terms are also standardized under the E2B (R3) format to allow for the easy exchange of data among agencies. The MedDRA coding applies to all medical aspects of the report including, but not limited to, adverse events, medical history, and indications for drug use (Brolund, CDER, ICH M2 EWG, 2000). The coding works by using a hierarchy of medical terms throughout five different levels ranging from very specific to general. These levels from lowest to highest term specificity include System Organ Classes (SOC), High Level Group Terms (HLGT), High Level

Terms (HLT), Preferred Terms (PT), and Lowest Level Terms (LLT) (ICH Secretariat, 2013). The ICH uses MedDRA to classify the medical events, which are reported at different level of the hierarchy, into the LLT (Brolund, CDER, ICH M2 EWG, 2000). This is useful in pharmacovigilance because there are many different terms allocated to the same events or drugs. By translating all of the different terms into one, it is easier for reviewers to identify correlations that may have been missed had the reports been submitted using a variety of terms. MedDRA is also a multilingual program, further enhancing the ability to exchange data among international agencies.

After all of the medical terms in the ICSR have been converted in MedDRA, the process of translating all the textual input terms into an OID begins. The first step is to translate all of the input information into *eXtensibleMarkup Language* (XML). XML is a form of *Standard Generalized Markup Language* (SGML), which is a standard "designed to describe the structure and content of electronic documents between business entities that need information to be available for extended periods of time (archived)" (Brolund, CDER, ICH M2 EWG, 2000). The input data is translated into XML because XML contains schema, which can then be manipulated, stored, and indexed (Brolund, CDER, ICH M2 EWG, 2000). Another important facet of XML is its parser, Unicode, which provides a unique number code for each input character (Brolund, CDER, ICH M2 EWG, 2000). OIDs are generated from specific sequences of numbers that identify a unique term. These OIDs, which are registered by the ICH, can then be used internationally because they are now standardized. The coding process of input data is represented by Figure 7 below.

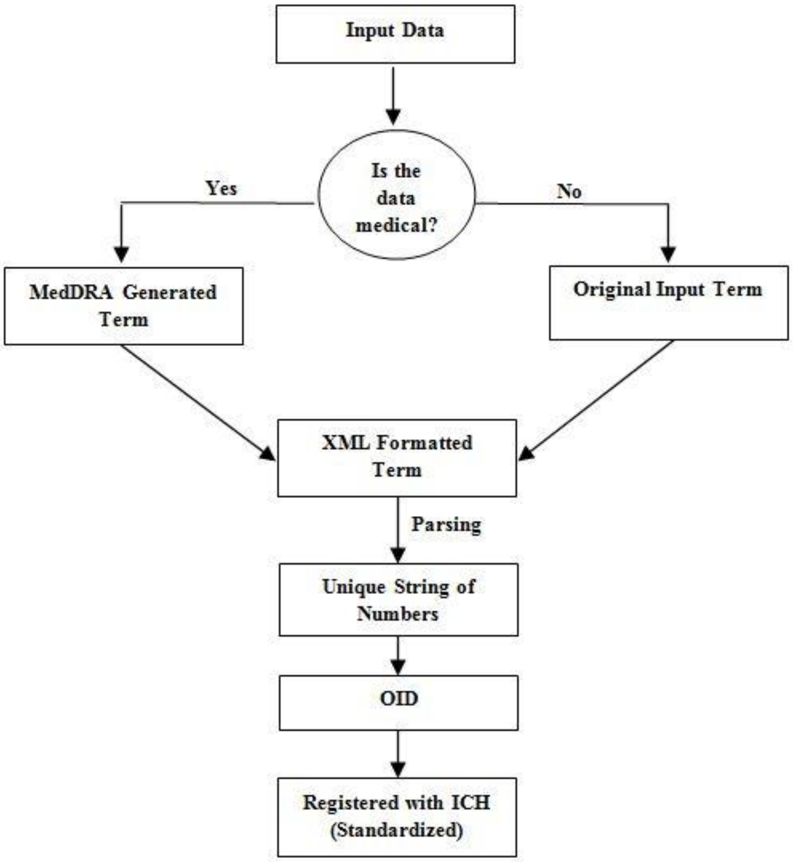


Figure 7: XML Flow Chart

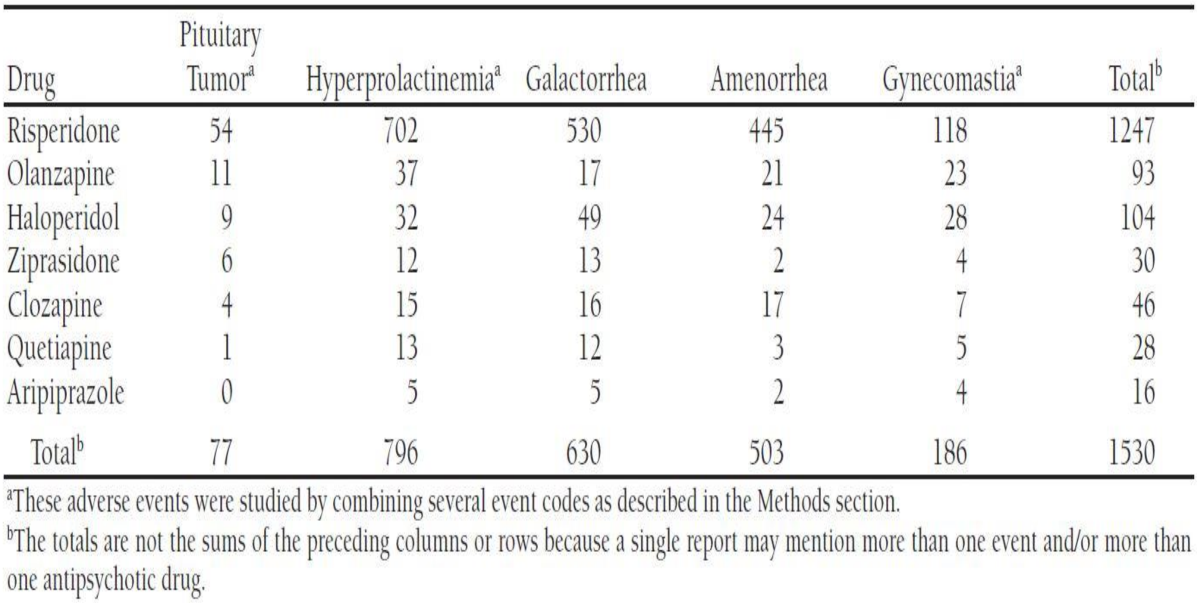
2.3.3 Case Study

The study of atypical antipsychotics and pituitary tumors serves as an example of data mining being used in pharmacovigilance for a retrospective study. Initially, the team conducting the following study identified that there was a high number of reports in the FDA's Adverse Event Reporting System (AERS), indicating that an atypical antipsychotic called risperidone was associated with pituitary tumors (Szarfman, Tonning, Levine, & Doraiswamy, 2006). Risperidone, which is "a potent dopamine D_2-receptor antagonist antipsychotic," was expected to cause an elevation in patients' prolactin levels, as this is a known side effect of many antipsychotics, but risperidone seemed to have a higher frequency of this event than other newer drugs in the same class (Szarfman et al., 2006). These elevated prolactin levels, also known as hyperprolactinemia, "can be asymptomatic or can result in symptoms such as galactorrhea, menstrual changes, infertility, and gynecomastia" (Szarfman et al., 2006). Knowing these effects of hyperprolactinemia, the study's objective was to compare the disproportionality of reporting of hyperprolactinemia, galactorrhea, and pituitary tumors among seven commonly used antipsychotic drugs (Szarfman et al., 2006). This meant to show whether or not risperidone was more strongly associated with the adverse events than the other antipsychotic drugs, though the result would not be able to definitively prove causality or non-causality.

The data used for this study came from the Adverse Event Reporting System (AERS), which has been recently updated and is now known as FAERS. At the time of this study, the database contained approximately 2.5 million adverse event reports, and was receiving around 1000 reports daily (Szarfman et al., 2006). The Multi-item Gamma Poisson Shrinker (MGPS) was applied to the 2.5 million reports, due to its ability to stratify the data and shrink it to reduce "the potential for great volatility of reporting ratio values due to fluctuations in numerators and denominators when reports for a particular drug-event combination are small" (Szarfman et al., 2006). This stratification and shrinkage allows for a more precise estimate and results in Empiric Bayes Geometric Mean (EBGM) values that represent the relationship between a specific drug-adverse event combination and reporting (Szarfman et al., 2006). This MGPS algorithm "is a reporting ratio disproportionality method designed to enable regulatory reviewers to efficiently search for potential drug safety problems in very large databases" (Szarfman et al., 2006). MGPS, when applied, searches through and analyzes the entirety of the database "taking into account all drugs and all adverse events" (Szarfman et al., 2006). In this case, MGPS was applied to all reports from January 1968 to May 2005, and seven drugs were analyzed. These drugs included aripiprazole, clozapine, olanzapine, quetiapine, risperidone, ziprasidone, and haloperidol all of which are atypical antipsychotics, with the exception of haloperidol which was used to exemplify older typical antipsychotics as a point of comparison (Szarfman et al., 2006). From here the adverse event reports listing one of these drugs as a "suspect drug," a possible cause of an adverse event were analyzed by the team (Szarfman et al., 2006). Any concomitant drugs that could be considered "suspect drugs" were analyzed as well to ensure that the report was not confounded (Szarfman et al., 2006).

Next, the MedDRA terminology and corresponding codes were determined for future analysis. The codes indicated the specific adverse events of interest, in this case those events related to elevated prolactin levels. Adverse event codes for pituitary tumors, hyperprolactinemia, and gynecomastia were combined, while galactorrhea and amenorrhea were not combined (Szarfman et al., 2006). The logic for the combined and individual adverse event codes is that MedDRA terminology is subjective and it is difficult to predict exactly how events will be coded, so the term combinations could allow for more relevant reports to be detected (Szarfman et al., 2006). As a result of the MedDRA code analysis, “the number of unique reports containing at least one of the five adverse event codes studied was more than 10-fold higher with risperidone than with haloperidol or olanzapine and more than 25-fold higher than with clozapine, ziprasidone, quetiapine, and aripiprazole” (Szarfman et al., 2006). There were a total of 77 pituitary tumor reports throughout the seven antipsychotics and risperidone was associated with 54 of those reports (Szarfman et al., 2006). The following Table D represents the frequency of the specific adverse events that were analyzed in relation to each of the antipsychotics, and it can be seen that risperidone has a much higher rate of all of the adverse events than the other drugs.

Table D: Frequency of Adverse Event Reports by Antipsychotic Drug (Szarfman et al., 2006)



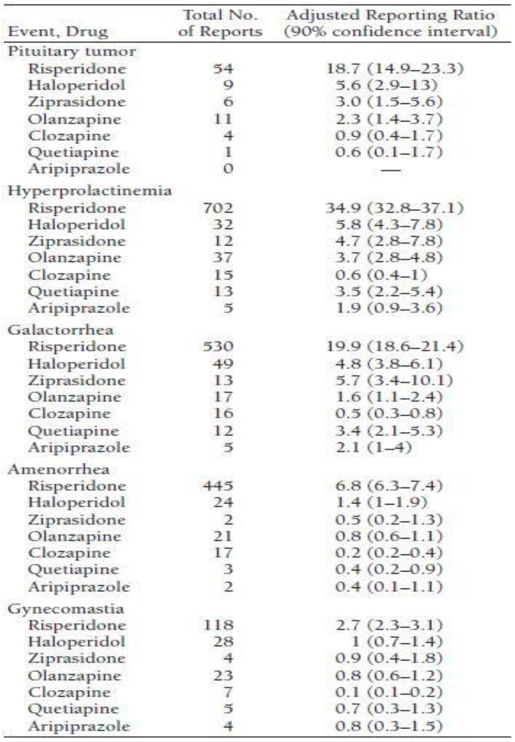
Drug	Pituitary Tumor[a]	Hyperprolactinemia[a]	Galactorrhea	Amenorrhea	Gynecomastia[a]	Total[b]
Risperidone	54	702	530	445	118	1247
Olanzapine	11	37	17	21	23	93
Haloperidol	9	32	49	24	28	104
Ziprasidone	6	12	13	2	4	30
Clozapine	4	15	16	17	7	46
Quetiapine	1	13	12	3	5	28
Aripiprazole	0	5	5	2	4	16
Total[b]	77	796	630	503	186	1530

[a] These adverse events were studied by combining several event codes as described in the Methods section.

[b] The totals are not the sums of the preceding columns or rows because a single report may mention more than one event and/or more than one antipsychotic drug.

The calculated EBGM values illustrate the same idea. The following Table E shows the adjusted reporting ratio, which is the EBGM value in this case, and each drug-event combination. It is important to note that any EBGM value above 2 is considered to represent a safety signal (Tonning, 2015).

Table E: Adjusted Reporting Rations for the Antipsychotic Drug-Event Combinations (Szarfman et al., 2006)



Event, Drug	Total No. of Reports	Adjusted Reporting Ratio (90% confidence interval)
Pituitary tumor		
Risperidone	54	18.7 (14.9–23.3)
Haloperidol	9	5.6 (2.9–13)
Ziprasidone	6	3.0 (1.5–5.6)
Olanzapine	11	2.3 (1.4–3.7)
Clozapine	4	0.9 (0.4–1.7)
Quetiapine	1	0.6 (0.1–1.7)
Aripiprazole	0	—
Hyperprolactinemia		
Risperidone	702	34.9 (32.8–37.1)
Haloperidol	32	5.8 (4.3–7.8)
Ziprasidone	12	4.7 (2.8–7.8)
Olanzapine	37	3.7 (2.8–4.8)
Clozapine	15	0.6 (0.4–1)
Quetiapine	13	3.5 (2.2–5.4)
Aripiprazole	5	1.9 (0.9–3.6)
Galactorrhea		
Risperidone	530	19.9 (18.6–21.4)
Haloperidol	49	4.8 (3.8–6.1)
Ziprasidone	13	5.7 (3.4–10.1)
Olanzapine	17	1.6 (1.1–2.4)
Clozapine	16	0.5 (0.3–0.8)
Quetiapine	12	3.4 (2.1–5.3)
Aripiprazole	5	2.1 (1–4)
Amenorrhea		
Risperidone	445	6.8 (6.3–7.4)
Haloperidol	24	1.4 (1–1.9)
Ziprasidone	2	0.5 (0.2–1.3)
Olanzapine	21	0.8 (0.6–1.1)
Clozapine	17	0.2 (0.2–0.4)
Quetiapine	3	0.4 (0.2–0.9)
Aripiprazole	2	0.4 (0.1–1.1)
Gynecomastia		
Risperidone	118	2.7 (2.3–3.1)
Haloperidol	28	1 (0.7–1.4)
Ziprasidone	4	0.9 (0.4–1.8)
Olanzapine	23	0.8 (0.6–1.2)
Clozapine	7	0.1 (0.1–0.2)
Quetiapine	5	0.7 (0.3–1.3)
Aripiprazole	4	0.8 (0.3–1.5)

Again, risperidone shows a stronger relationship between each of the events and reporting as it has a much higher EBGM value than all of the other antipsychotic drugs that were analyzed.

The results of this study showed that risperidone is very strongly related to reporting of the five adverse events in comparison to the six other antipsychotic drugs, but as mentioned before this does not prove causality. This study merely used data mining to support the hypothesis that risperidone caused a higher than expected elevation in prolactin levels. It was used in a supportive role in this case, and the FDA would like to be able to move away from using it solely to support hypotheses and instead utilize it to predict safety signals (Marni Hall PhD., personal communication, 2015).

2.3.4 Specific Examples of Data Mining in Pharmacovigilance

In recent years, the pharmaceutical industry has had numerous success stories involving data mining. A 2004 study conducted by the Institute for Safe Medication Practices' senior scientist of drug safety and policy (U.S. Food and Drug Administration, 2015), Thomas Moore, concluded that selective serotonin reuptake inhibitors (SSRI) antidepressant drugs were related to an increase in violent tendencies. The study applied specific data mining methods (PRR) to the public FAERS database with the adverse events of interest being homicide, homicidal ideation, physical assault, physical abuse and/or violence related symptoms. The FDA became aware of the study and proceeded to conduct a thorough analysis of the adverse events themselves. This analysis included the comparison of the data mining analyses of data from Moore's study and FDA data as well as the calculation of PRR and EBGM scores. Similar to Moore's study, the analysis conducted by the FDA concluded that both the PRR and EBGM methods resulted in antidepressant and ADHD drugs generally having scores greater than 2 and thus generating a strong association between these drugs and adverse drug events related to violence (Tonning, 2015).

Continuing with pharmacovigilance efforts, the FDA and pharmaceutical companies have been using data mining techniques on older data to see if the introduction of data mining methods would have resulted in finding these safety signals earlier. Using traditional, manual methods in 2007, Sprint Fidelis®, an electric cable or lead that connects a defibrillator to a patient's heart, was found to have associations with inappropriate shock events and lead fracture events which resulted in the voluntary market withdrawal of the product. However, when a retrospective analysis using data mining methods was done on the Manufacturer and User Facility Device Experience database, it was determined that safety signals associated with a cardiac defibrillator that was implantable could have been detected as early as March of 2006. Moreover, the investigation of these safety signals and the overall associations would have occurred sooner which would have saved time, money and possibly lives (Duggirala et al., 2015).

The pharmaceutical industry is an international industry with each country having its own databases. In 2008, the Center for Food Safety and Applied Nutrition (CFSAN), which is a division of the FDA, launched an investigation into the proposed association between destruction of the liver (hepatotoxicity) and the dietary supplement Hydroxycut® (Aloi, 2008). This proposed association was the result of data mining on the CFSAN's Adverse Experience Reporting System (CAERS) database. This division of the FDA had only begun exploring the idea of data mining in 2001, thus the CFSAN decided to first use many different data mining methods in order to get as much information as possible. The data methods included disproportionality analysis, Empirical Bayesian Geometric Mean (EBGM), an application of Adverse Event (AE) Outlier Methods and Brute Force (an exhaustive trial and error method)

(Chirtel, 2009). Once all of the information had been gathered from the different data mining methods, a statistical evaluation was conducted on the data mining results (Aloi, 2008). The working Consumer Safety Officer, Brenda K. Aloi (2008), concluded in the investigation report that:

> The data demonstrated that 13.24% (18 of 136) of the adverse events experienced by the Hydroxycut consumers were hepatobiliary disorder, while only 3.28% (306 of 9333) of the consumers who took dietary supplements other than Hydroxycut experienced hepatobiliary disorders. In other words, consumers who take Hydroxycut would have approximately 4 times the risk (RR-13.24/3.28=4.04) of experiencing hepatobiliary disorder as compared with consumers who take dietary supplements other than Hydroxycut. (Aloi, 2008, p. 1)

The report urged further investigation based on the strong, proposed association (Aloi, 2008). Soon after, Hydroxycut® was voluntarily recalled from the market in May 2009 (Duggirala et al., 2015).

2.4 Learning Outcomes

A learning outcome is a statement designed by an educator to describe exactly what their student will be able to do after participating in an educational course or activity. Learning outcomes are used commonly throughout education to provide clear goals for educational materials, which allows for more effective assessments of student success, more focus in the course design, and a simple way to present the course to interested parties. An action described by a learning outcome should have three traits - it must be observable, measurable, and performed by the student. Constructing well thought out learning outcomes is an important element of designing any educational materials, as it allows for better focus and clarity in objectives which helps in conveying the subject matter (Phillips, 1994).

2.5 Adult Learning

2.5.1 Central Concepts

In order to create effective educational materials for the Food and Drug Administration, we need to be versed in the practice of educating adults, known as andragogy. Adults react differently to educational approaches than children, thus adult education must be approached in different ways. Our project goal is to provide educational materials to deliver the fundamentals of data mining to an audience who may already be very well educated in other areas. This requires a slightly different methodology than that which many of us are accustomed to from our education as children and young adults.

To start understanding andragogy it proves useful to have an organized foundation in pedagogy, the practice of teaching to children. Teaching children is a process that begins with the question "What will they be taught?" (Northwest Center for Public Health Practice, 2014) As described in *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development* by Malcolm Knowles in 2014, pedagogical education tends to focus on the teacher as opposed to the student. The learner is seen by the teacher as a dependent personality, as someone who needs the teacher in order to learn. The learner need not know what applications the material they are learning may have in their lives, but only needs to know that they need to learn what the teacher teaches in order to pass and be promoted. (Knowles, 2014, p. 62)

For children learning the fundamentals of their future education, these principles work quite well. A young child has very little understanding of how learning to subtract may benefit them in their lives, but must learn it anyway in order to form a basis that will support them in future mathematical work. Even without knowing how it may help, the child gains a mathematical foundation that will eventually allow for success in many fields.

When teaching adults, the process begins not with the question, "What will they be taught?", but rather, "What do they need to learn?" (Northwest Center for Public Health Practice, 2014) Adults are task oriented learners and are most interested in learning something new if learning it will provide clear benefits to them in their lives. They are particularly motivated by life events and will be most eager to learn when circumstances arise requiring it, such as new regulations in their workplace or new technological standards that need to be met. Adults need to be given more independence because they are accustomed to making their own decisions with their own logic. If adults are placed into a situation where their independence is removed, they may have difficulties accepting the material presented to them. As such, this independence should be preserved by carefully presenting material in such a way that the learner is always able to understand how it connects with their work and goals. Adults are also a much more diverse group than children, with many different life experiences and prior educational successes. It is important to not only consider this prior knowledge when constructing materials in a respectful way, but also to allow for learners to use this knowledge, sharing it and applying it to the novel situations you are presenting them. (Knowles, Holton, Swanson, 2014). Adult education is all about ideas of independence, centered on allowing learners to choose to learn for themselves, according to their own values.

2.5.2 Methodologies for Adult Learning

The first step in the development of any educational material, for adults or otherwise, is to identify the target audience. Working with adult learners can prove difficult as individuals often come from a wide variety of fields and experience. However, it is very important to develop a general sense of who you intend to teach and what learning needs they have. This calls back to the question of "What do they need to learn?" Understanding why someone would take your course and what they intend to get out of taking your course is paramount to providing them with satisfactory materials (Northwest Center for Public Health Practice, 2014).

The second step is the development of learning outcomes for the materials. A learning outcome is a statement designed by an educator to describe exactly what their student will be able to do after participating in an educational course or activity. Learning outcomes are used commonly throughout education to provide clear goals for educational materials, which allows for more effective assessments of student success, more focus in the course design and a simple way to present the course to interested parties. An action described by a learning outcome should have three traits - it must be observable, measurable, and performed by the student (Phillips, 1994). Constructing well thought out learning outcomes will clearly define the tasks the materials will educate learners on and will establish the context and reasoning to motivate adult learners.

Once learning outcomes are developed, actual course materials can be created to suit each outcome. Each outcome can be classified as either knowledge based, skill based, or attitude based. Each type of learning outcome has different material and activity types that are best suited. Knowledge based outcomes involve the understanding of concepts and abstract patterns, such as learning the definition of data mining. Skill based outcomes center on practical applications, such as learning how to use a data mining tool. Attitude based outcomes are about emotion, such as learning a mindset with which to approach data mining problems. Each type is best approached in a different way (see Table F below) and classification of outcomes can provide valuable guidance in choosing what form learning materials will take (Northwest Center for Public Health Practice, 2014).

Table F, Learning Mediums sorted by best suited Outcome Style

Knowledge	Skill	Attitude
Lectures	Role Plays	Value Clarification
Brainstorms	Simulations	Nominal Group Process
Discussions	Teach Backs	Consensus Seeking

Lastly, when building the final materials, it is important to take into consideration a few final facets of adult learning. As mentioned previously, adults tend to learn best through distinct tasks, meaning they often benefit greatly from interactivity in their education. Interactivity will aid in connecting the material presented to the tasks it is intended to benefit, and will also aid in information retention, as seen in Figure 8 below.

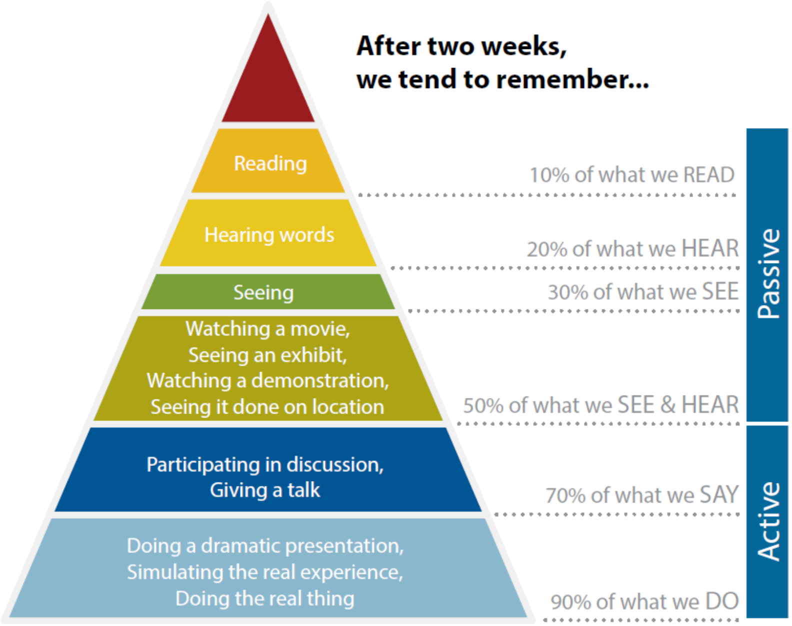


Figure 8: Information retention via the "Cone of Learning" (Northwest Center for Public Health Practice, 2014)

Essentially the "Cone of Learning" above shows that learners will retain information more readily if they engage with that information in more ways. This is especially important in teaching skills that may not see everyday use in the workplace, which may mean learners are less likely to need to revisit educational materials after using them initially.

Finally, it is imperative to avoid over-teaching when designing the final materials. Over-teaching refers to when you teach too much of a subject to a learner, overwhelming them and diluting the point of the exercise. This is best explained through simple analogy – when teaching someone to use a hammer, explaining the physics behind the force that needs to be applied can make the entire process far more complicated than merely demonstrating that it should be swung (Clawson, 2006). Avoiding over-teaching will be crucial in the teaching of data mining, as the field is highly technical, but most of deep, theoretical details will simply be unnecessary for the audience we are targeting. Adult education is all about teaching towards a distinct goal and unnecessary details will distract learners from their goals, hurting engagement and making the entire exercise less effective.

3. Methodology

The overall goal of this project was to assist the FDA to conduct pharmacovigilance efforts more efficiently through the development of educational materials on data mining concepts and applications, as well as providing a brief overview of which data mining tools in the market may be most applicable to the FDA's work. The success of the educational portion of the project depended on the completion of our objectives: study current pharmacovigilance strategies, assess the current data mining needs of the FDA and develop corresponding educational materials. We accomplished this in two stages, as depicted in Figure 9 below.

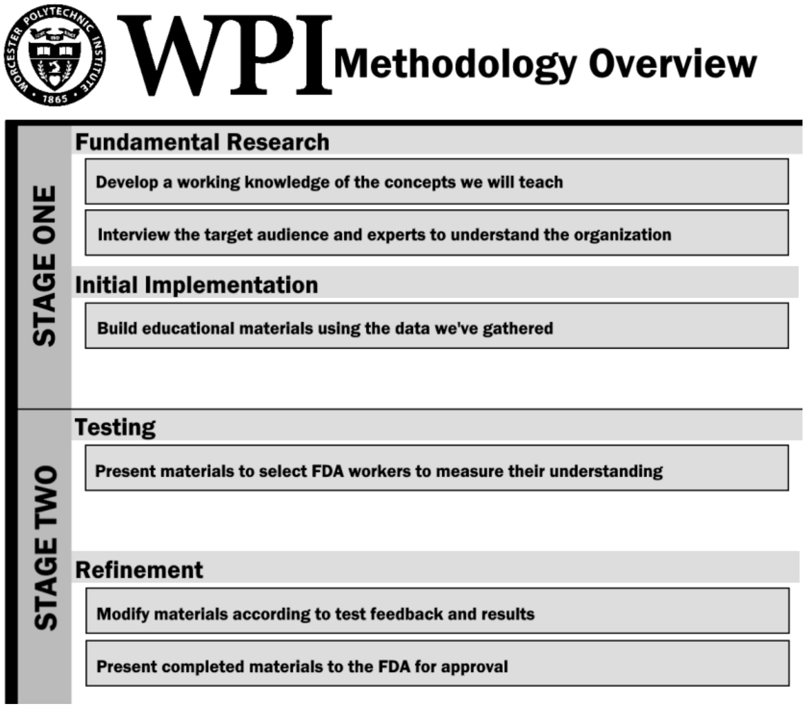

WPI Methodology Overview
STAGE ONE
Fundamental Research
Develop a working knowledge of the concepts we will teach
Interview the target audience and experts to understand the organization
Initial Implementation
Build educational materials using the data we've gathered
STAGE TWO
Testing
Present materials to select FDA workers to measure their understanding
Refinement
Modify materials according to test feedback and results
Present completed materials to the FDA for approval

Figure 9: Overview of our project methodology and goals

Stage one consisted of interviewing several experts and non-experts in the fields of data mining and pharmacovigilance. This helped to guide our attention to the areas that we needed to focus our research on. Stage two consisted of presenting our first draft educational materials to a subset of FDA employees, the target audience, after administering a pre-test. After reading through the draft a post-test was administered to measure the reader's reaction and achievement of specific learning outcomes. Based on the data from these tests, we then refined our materials into their final form.

The tool analysis portion of our work required basic research into data mining tools on the current market. We conducted this through a simple two phase methodology in which we conducted research using online and literary sources, and then arranged our findings into our educational materials.

3.1 Stage One

3.1.1 Fundamental Research

Stage one began by establishing a foundation from which we could generate initial prototypes of educational materials. This consisted of academic research in the fields of data mining and pharmacovigilance, followed by an assessment of the FDA’s data mining needs. The assessment was made up of semi-structured interviews with OSE employees who are experienced in pharmacovigilance, data mining and its applications at the FDA, and OSE employees who would be working with data mining tools. Our research also involved detailed investigation into the structure of the FDA to understand the needs of the people the materials are intended to aid. Lastly, we had a discussion with a training expert at the FDA to learn about the best methods we can use to educate employees.

The interviews with OSE data mining experts provided us with an understanding of the topics other FDA personnel would be expected to know about data mining, which skills they needed to develop, and the tasks they would be expected to perform. These expert opinions defined fundamental data mining knowledge relative to pharmacovigilance. Using this definition, we determined the specific concepts and learning outcomes that were incorporated into our testing and educational materials.

To best conduct the assessment of experts and target audience, we chose the format of a semi-structured interview. The flexible nature of semi-structured interviews provided an opportunity for greater exploration of ideas and the uncovering of the unexpected (Guest, 2006). We spoke with a convenience sample of available OSE staff members. This allowed us to speak to individuals who had time to engage with us, had interest in data mining and provided us with detailed answers and thoughtful insights. The few materials required were a pen, paper, and a computer to store all of the information. Interviews were not voice recorded.

We did, however, need a consent form to both inform participants and ensure that they were willingly engaging with us (see Appendix B).

3.3.2 Initial Implementation

After gathering all of the information from our fundamental research, we began the initial implementation of the educational materials. These educational materials were designed according to our specified learning outcomes (see Table G). These learning outcomes were designed according to input gathered during our interviews with FDA data mining and pharmacovigilance experts.

To conclude stage one, we constructed a first draft of our educational materials according to the information gathered during our research. These materials were in both slideshow and manual form and contained all of the information required to achieve the learning outcomes.

Table G: the learning outcomes for our educational materials

Learning Outcomes
Define data mining
Define data mining in context to the FDA
Identify methods and how they are used
Describe how data mining is used with FAERS data
Recognize prerequisite steps to perform effective data mining
List tools currently used at the FDA and what methods are implemented within those tools
List tools that could be potentially used
Recall where FDA data mining resources can be found

3.2 Stage Two

3.2.1 Testing

The second stage of our project began with preliminary testing of the target audience, to evaluate what knowledge they had regarding data mining. The questions for the pre-test were based on the learning outcomes created through interviews with experts. With the pre-test, we established the extent of data mining knowledge already known by the target population. This allowed us to then test the population for improvement after they had used our materials.

After pre-testing, our target audience read through the first draft of the manual that we had created. After reading through the entire manual, a post-test was administered to evaluate the efficiency and impact of the educational materials before finalizing them. This post-test not only measured the students' achievements of learning outcomes, but also allowed them to give their input on the format of the materials and the content covered. Readers were each given an hour to complete the pre-test, read through the manual, and complete the post-test.

Similarly, to stage one, we used a convenience sample. To ensure our findings were representative of the entire OSE staff population, the convenience sample consisted of selected OSE staff members with a range of data mining experience. The materials required and the necessary consent forms were the same as in stage one.

Test results for both pre- and post-test were evaluated using a standard percentage system. Each question within the tests corresponded to a learning outcome, and was worth an equal amount of points. Each question was graded by our team. For simple questions the grading was binary, correct if they wrote down what was expected, wrong otherwise. For our more complex questions (such as "How would you define data mining?") we looked for a list of key terms and concepts in the answer, and graded according to how many of these key items were present. These key items were all outlined in our answer key, which was written prior to the administration of any tests to avoid any bias. Once each test was graded, we calculated averages for both the pre- and post-tests and compared the overall test scores as well as scores of individual questions, to see quantitatively how much influence, both overall and for each learning outcome, our materials had on the learners.

3.2.2 Refinement

In this part of the process, we created the final draft of our educational materials. Revisions were made based on the results from the post-test and the learners' feedback about the content covered and the material format. Once we had the final draft that was better suited to the achievement of our learning outcomes, we presented the result to the OSE.

3.3 Industry Tool Investigation

3.3.1 Phase One: Tool Research

While we investigated the details of data mining and education, we worked on compiling a list of relevant data mining tools that may be of use to the FDA. These tools were chosen qualitatively, with a focus on pharmacovigilance efforts. The research into the tools was conducted primarily through investigation of print and online sources, with some supplementary guidance from our discussions with experts. We did not perform in-depth analyses of the tools, but rather searched for and identified the tools that had potential use for the FDA pharmacovigilance process.

3.3.2 Phase Two: Delivery of Tool Recommendations

Once we gathered a significant number of potential tools, they were placed into our educational materials. The tools were outlined in very basic terms, identifying why they were selected and recommending that further investigation be conducted by the agency.

3.4 Summary

Our methodology plan consisted of two stages: stage one, in which we conducted basic research to prepare a first draft of our materials, and stage two, in which we tested these materials with pre- and post-testing and refined them into a completed form. The stage one research consisted of a literature review, as well as semi-structured interviews to obtain guidance from data mining experts. Stage two testing consisted of presenting the materials to our target audience and testing their understanding afterwards, to compare to the pre-test results. Once completed, we had plenty of data to construct effective educational materials that were presented to the FDA.

Alongside our research into educational materials, we conducted a brief study into modern data mining tools that may be of use to the FDA, after which we sent the results of our research to the FDA for its consideration.

4. Data

This section discusses the data we received as a result of testing, explaining what data we received, and what information we learned from analyzing it.

4.1 Charted Data

The following charts show the results of our testing. Test subjects were given two identical tests, a pre-test prior to reading our data mining manual, and a post-test afterwards. The test subjects did not know that the tests were identical. These tests consisted of 8 questions, worth 6 points each, for a total of 48 points. The tests along with the grading rubric used to determine the final scores of test subject can be seen in Appendix G. The charts below show the average scores among all test subjects, broken down by question. For the data tables containing the original numbers, please refer to Appendix F.

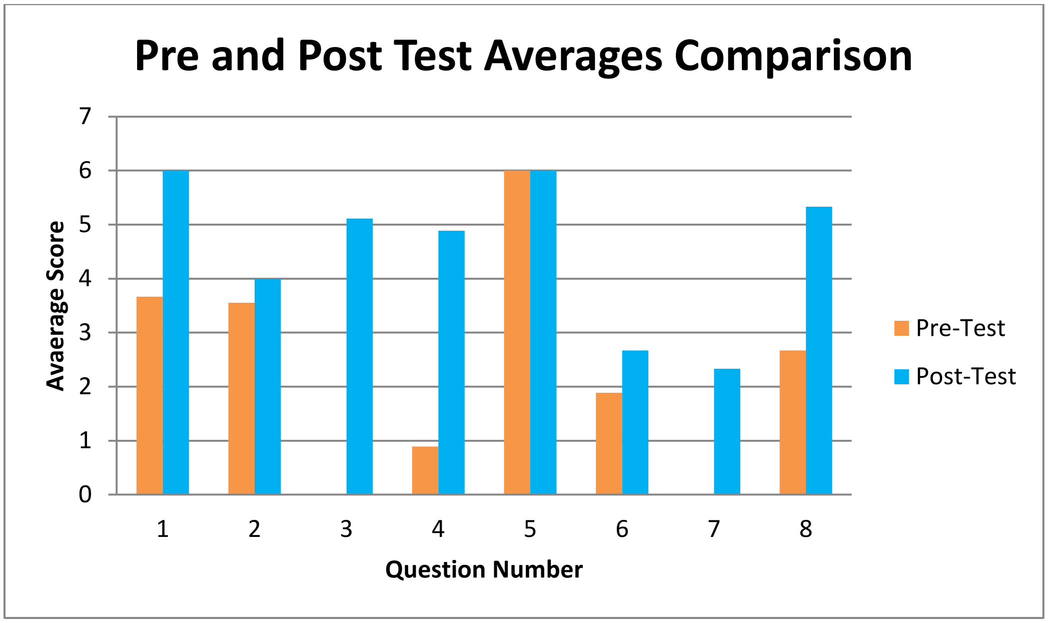


Figure 10: Average scores per question for both tests compared

We also recorded the time it took for each person to read the educational materials. The reading times ranged from 15 to 36 minutes, with an average of 25.11 minutes.

4.2 Interpretation of Data

Although we had only a small number of test subjects (9 in total), our data were still substantial enough to show some clear trends from which we could draw conclusions.

4.2.1 Pre-Test

The results of the pre-test illustrate that the prior data mining knowledge of our sample was not substantial – with the exception of our fifth question, scores were lower than expected. Stronger pre-test results were seen on questions one and five which covered the general definition of data mining and whether or not data mining could prove causation. However, the answers seen on these questions were disparate and unfocused, which was expected.

4.2.2 Post-Test

The post test results showed a significant increase in scores for each question. All areas of the manual produced an increase in score for their related learning outcomes, with some outcomes being understood better than others.

Excellent improvements were seen for questions 1, 3, 4, and 8, with each of those questions having a post-test average over 75%. Question 1 in particular showed great results, with every single test subject being able to perfectly define data mining after reading the manual, despite the discord seen in the pre-test. Questions 3 and 4 were related to technical aspects of data mining methods, and scored well despite lower scores in the pre-test. Question 8 asked subjects to provide the name of the internal FDA office that they could go to for more data mining information, and knowledge of that office more than doubled.

Not all questions showed improvement however, as results for questions two, six, and seven, were lackluster. Question two, which asked for a definition of data mining in the context of pharmacovigilance, showed slight improvement from the pre-test values, but definitions remained inconsistent and scores remained underneath the 75% mark. Questions 6 and 7 were about data mining tools, both the tools currently used by the FDA and those that might be used in the future. Scores for these questions remained below 50% even after the presentation of our manual, showing that the tools section was not conveying the information effectively.

The final question to consider is question 5. Question 5 was a true or false question about causation and correlation in data mining. In the pre-test and post-test, every test subject answered this question correctly. This could mean several things: that the correlation value of data mining is well known in the FDA, or that our question was worded in such a way as to make the answer obvious. Unfortunately, with such a small sample size we cannot make a fair determination one way or the other, meaning that question 5 cannot be considered while applying our data in the final revisions of the manual.

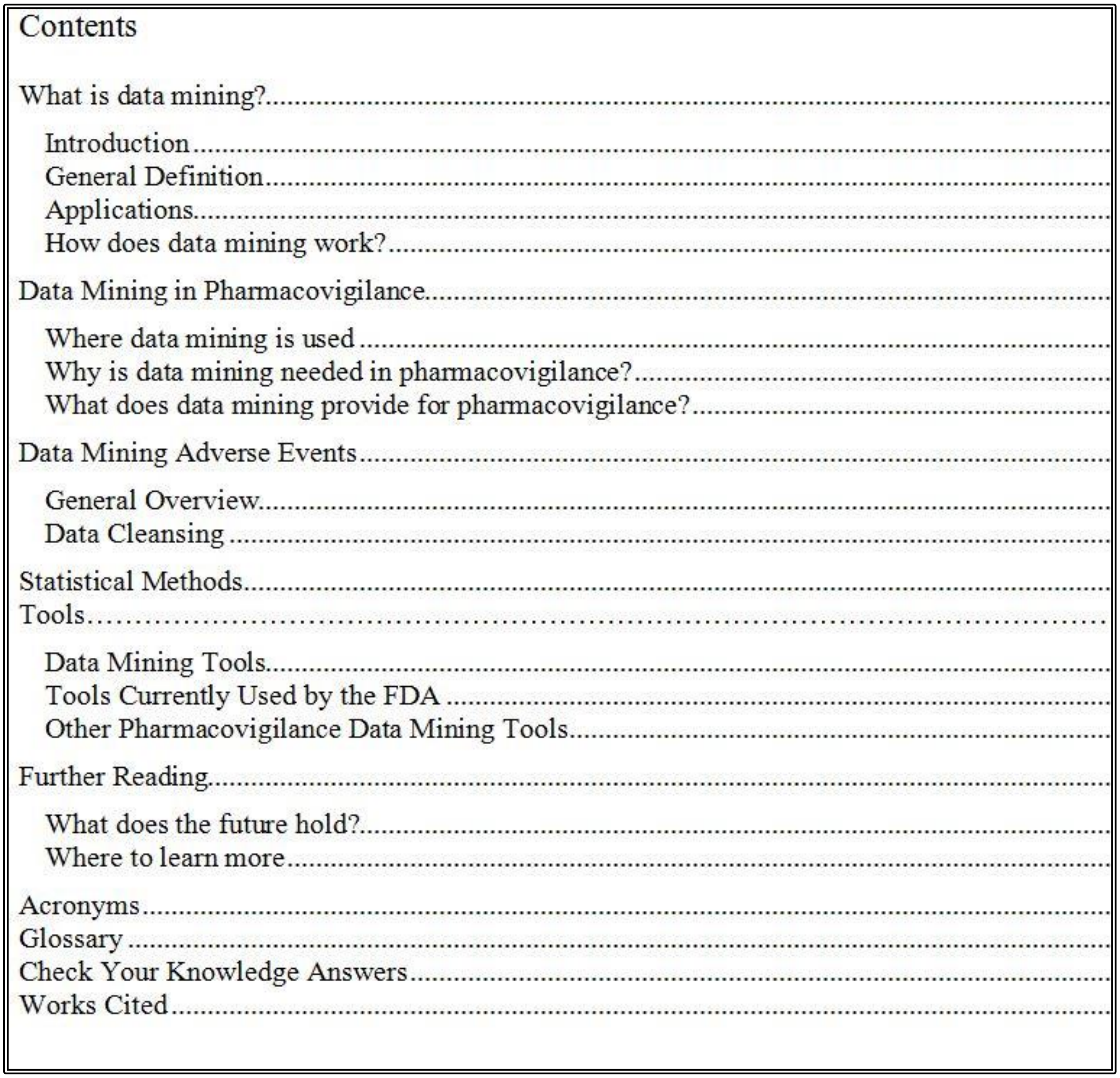
5. Results

5.1 Overview of our Deliverables

The primary item produced from this project was a manual, entitled *Data Mining for Pharmacovigilance* (refer to Appendix H). The manual was designed to educate FDA staff on all the specified learning outcomes (listed in Table G) in an easily accessible text format and can be used for convenient reference. Secondary deliverables included the results of an investigation into data mining tools that the FDA may wish to look into further. This investigation discussed several data mining tools that the FDA may be able to apply to pharmacovigilance efforts if implemented in the future. The tool investigation was written in the form of a brief list, containing summaries of each tool and their individual advantages and disadvantages. Lastly, the content of our manual was converted into an oral slide presentation and presented to the Office of Surveillance and Epidemiology staff. This slide presentation was a condensed version of the material and intended to provide a more general overview than the manual.

5.1.1 Data Mining Manual

Table H: The table of contents for our finished manual (see the full manual in Appendix H)



Contents

Section One: Definition of Data Mining

This first section of our manual describes data mining in a very general sense, introducing readers to the concept in a broad fashion before narrowing the focus to pharmacovigilance. Knowing that data mining is not a well-defined discipline, we chose the broadest definition as this project's general definition of data mining. This first section of the manual consists of three parts, with the first part offering a brief analogy and explanation of what data mining is and its formal definition. The second part uses real world examples to connect the definition to actual applications that the reader may be familiar with. The third part explains the basics of how data mining is performed outside of pharmacovigilance. The goal of the first section of the manual is to familiarize the reader with data mining in its most general form to allow for easier comprehension of the more complicated applications within pharmacovigilance. The learning outcome to be fulfilled in this section is "*Define data mining.*"

Section Two: Data mining use in pharmacovigilance

The second section of our manual moves past generic descriptions of data mining and focuses on the FDA and its pharmacovigilance needs. It begins by describing what data mining means within pharmacovigilance, touching on the determination of safety signals and data mining scores. This description is followed by a discussion of where data mining can be used for pharmacovigilance, specifying which areas of the field could benefit from the use of data mining techniques. We explain the problems with the current report evaluation system that make data mining appealing for the FDA going forward. Important from an adult learning perspective, this part of the section illustrates the difficulty of reviewing the increasing amount of reports, providing a motivation for readers to learn more about data mining. Finally, the section concludes with an explanation of how pharmacovigilance data mining is performed – similar to the third part of the Definition of Data Mining manual section, but focused fully on pharmacovigilance. The Data Mining Use in Pharmacovigilance section is intended to fulfill our second learning outcome, "*Define data mining in the context of the FDA.*"

Section Three: Data Mining Adverse Events

The third section of the manual is designed to introduce readers to the data that is being mined. The first part of this section is an overview of the FAERS database and its contents. Although FAERS is not the only set of data that the FDA wishes to mine, FAERS is accessible and reliable. The description of FAERS provides context for the following information, as well as the tools discussion, and provides a segue into the topic of data cleansing. The discussion of which explains the essential prerequisite step to data mining along with the importance of clean data. The data cleansing section also describes a few of the techniques in use to ensure effective data mining. This section of the manual is designed for two learning outcomes: *"Describe how data mining is used with FAERS data"* and *"Recognize prerequisite steps to perform effective data mining."*

Section Four: Data mining methods and their applications

This section consists of a detailed discussion of the two main groups of statistical data mining algorithms, Frequentist and Bayesian. These groups are both suited to different situations and therefore an understanding of each - specifically when and where to use each one, is very beneficial for the reader. The information in this section is supported by tables, graphs, and examples that break down the concepts and allow the reader to follow the comparison. The learning outcome for this section is *"Identify methods and how they are used."*

Section Five: Data Mining Tools at the FDA and beyond the FDA

Our penultimate section discusses various tools that implement data mining techniques. The section starts with a basic description of what a data mining tool is and then moves into a brief discussion of the tool currently being used by the FDA, Empirica® Signal. After touching upon the current implementation and limitations of Empirica® Signal, we introduce the list of tools produced by our tool investigation. Advantages and disadvantages of each tool are also provided, along with a general and brief description of the software. This section covers two learning outcomes: *"List tools currently used at the FDA and what methods are implemented within those tools"* and *"List tools that could potentially be used."*

Section Six: Further Reading

This final section of the manual is aimed toward the reader who wants to explore and learn more about data mining beyond the scope of our project. The section begins with a description of some of the advanced data mining techniques that are being developed today, which may be relevant to the FDA's work in the near future. The description is followed by an annotated list of links to papers and websites that might provide more information to the inquisitive reader. This section provides information for our last learning outcome, *"List locations of where FDA Data mining resources can be found."*

5.1.2 Tool Investigation

The investigation of data mining tools consists of a list of three different data mining tools that could be applied to pharmacovigilance. Each tool has a brief description and includes a short list of advantages and disadvantages. This investigation is designed to serve as a quick reference should the FDA wish to investigate alternative tools more thoroughly. It was ultimately presented to the FDA as a section of our manual, *Data Mining for Pharmacovigilance*.

5.1.3 Data Mining Presentation

The final deliverable was a presentation given to the staff of the Office of Surveillance and Epidemiology. This presentation consisted of an oral rendition of our data mining manual, accompanied by relevant visuals and slides. The content presented was not as in depth as the written version, but still described data mining in the same manner as our manual. Specifically, many of the more in depth descriptions of methods and their applications were cut from the oral version, in favor of a greater focus on simple analogies and summaries. This presentation was split in to six sections, with content based off of the sections in the manual:

1. Introduction
2. Data Mining
3. Data Mining in Pharmacovigilance
4. Statistical Methods
5. Data Mining
6. What does the future hold? / Conclusion

Each section is discussed for only a few minutes, with the complete presentation taking about 20-25 minutes. Time was also reserved afterwards for any questions that the audience had. This presentation was recorded and saved for future use at the FDA, in order to accommodate learners that may prefer non textual materials.

5.2 Looking Back at the Process

5.2.1 Stage One – Fundamental Research

In order to write the deliverables for our project, we were required to develop a fundamental background in data mining and an awareness of the current pharmacovigilance process at the FDA. Our first five weeks at the project site were focused on research and conducting interviews to establish our foundation in these topics. A large part of this research was a literature review, which involved reading several studies on data mining and data mining in pharmacovigilance. The part of this research that proved most difficult was consolidating the numerous definitions of data mining presented in current literature.

In researching the field of data mining, our first goal was to develop a clear, general definition of data mining, as well as a definition of data mining in the context of pharmacovigilance. We soon discovered that different sources often had very different definitions for data mining, depending on the reporting field. For instance, in the context of homeland security "data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets" (Seifert, 2004). Whereas in a financial context, "data mining is defined as the set of techniques that allows the exploration and analysis of data so that models, and logical schemes that are not explicit a priori, can be spotted within very large databases" (Rajola, 2014). While fundamentally similar, these definitions differ in many ways, from the use of analysis tools versus techniques, as well as the expected outcomes of the data mining process. We reviewed and compared many mixed definitions to find that the most important elements for our general definition were pattern and cluster discovery in large datasets. Using these elements, in collaboration with Dr. Suranjan De, we finalized a general definition of data mining in context of this project. This definition was reviewed by our sponsors at the FDA to ensure its validity.

Once we were able to define data mining, our next focus was on the data mining methodologies being used in pharmacovigilance. After speaking with our sponsors, it was determined that the scope of this project, in terms of algorithms, was to provide readers with an understanding of what the algorithms do and a general idea of when to use each, rather than the mathematical and technical aspects behind them. While researching the data mining algorithms, we came across many papers including *Novel Data-Mining Methodologies for Adverse Event Discovery and Analysis* (Harpaz et al., 2012), *The Role of Data Mining in Pharmacovigilance* (Hauben, Madigan, Gerrits, Walsh, & Van Puijenbroek, 2005) and *Multinomial Modeling and An Evaluation of Common Data-Mining Algorithms for Identifying Signals of Disproportionate Reporting in Pharmacovigilance Database* (Johnson, Guo, Gosink, Wang, & Hauben, 2012) that discussed similar methodologies for disproportionality analysis, thus providing a clear guide as to where we should focus our research. After conducting our research, we noted that these three articles mentioned that the data mining algorithms could be split up into two groups, each with different advantages and disadvantages. The advantages and disadvantages of the two method groups, frequentist and Bayesian, provided a way to explain the general idea of when to use each group.

Along with the research, we conducted interviews with various staff members in the OSE. With the aid of Ellen Pinnow, we were able to easily arrange interviews with individuals who had varying prior knowledge about data mining. Each interview had its own set of questions tailored to the specific person we were talking to and their skills (see Appendix A for example questions). The answers, focused on data mining and the FDA, made the importance of our project clearer, as all of the data mining experts we spoke with found it difficult to break data mining down into a simple, clear definition and commented on how difficult it is to explain just the fundamentals of the field. The use of open ended interviews allowed every staff member to explain his or her varying insights on the idea of data mining, and directed the direction of our research.

Overall, the fundamental research process was the most challenging aspect of our project due to our initial lack of data mining knowledge. Despite this, it was also the most important process because we were able to learn more about data mining and the FDA. OSE staff members were supportive of our team, often providing us with additional reading materials to be added to our research.

5.2.2 Stage One – Initial Implementation

The initial implementation of our materials was straight forward due to the clearly defined learning outcomes we established with our sponsor, Dr. Suranjan De. We were able to draft the learning materials rapidly, tailoring each section of the educational materials to specific learning outcomes. In addition, we used knowledge of adult learning techniques found in *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development* to present the information in a convenient way (Knowles, 2014). The main challenge we encountered during material development was determining the desired level of technicality of the materials to ensure their usefulness to the FDA. The goal for the manual was not to overwhelm the reader with technical information that is not relevant to his or her work at the FDA, but still provided an in-depth understanding of the fundamentals of data mining. Revisions were made by our advisors and sponsors, during this stage, to ensure that the manual was in the best format for testing.

5.2.3 Stage Two – Testing

The testing phase was essential for evaluating the effectiveness of our materials on each of the designated learning outcomes. Testing consisted of a pre and post-test, both identical, with each question specific to assessing a particular learning outcome. Although we only had nine testers, the data was still detailed enough for us to identify some basic trends. From the results, we were able to see which areas of the manual our testers understood, and which areas they were unable to retain. The areas where retention was weak were then looked at more closely during the refinement stage of our methodology. In addition to the quantitative test scores, we also asked testers for their opinions on the manual, and collected their suggestions and feedback for future refinement. Feedback and suggestions from test takers included fixing typos, good use of analogies, confusing mathematical definitions, and a suggested executive summary.

5.2.4 Stage Two – Refinement

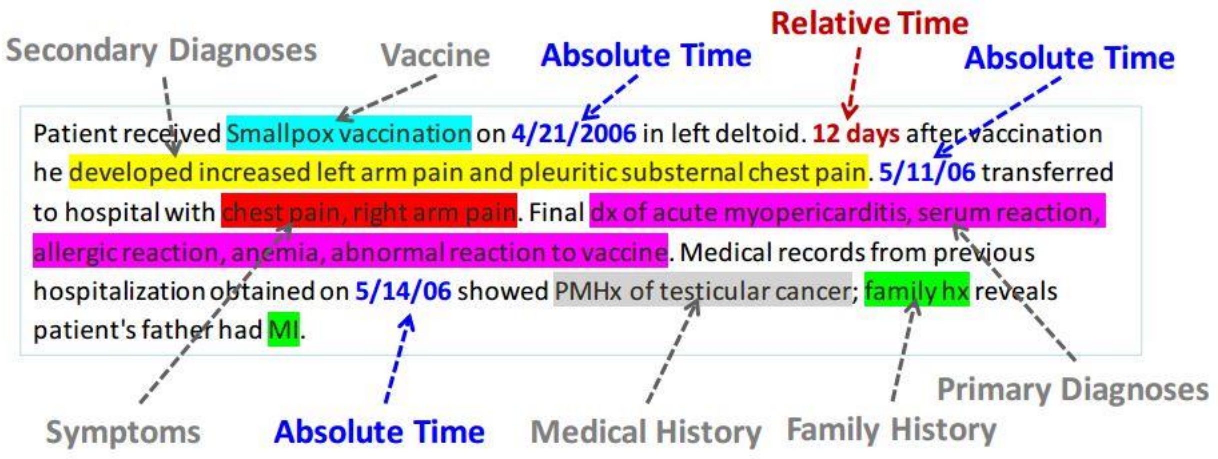
The refinement process was brief, as our initial testing results were positive and near to our goals. We worked to perfect our manual using the results of the tests, as well as revisions sent to us by our advisors and our sponsors. The central points of our manual were well received, and most of the revisions we received from sponsors and advisors were addressing small grammatical issues and the occasional factual inaccuracy, such as a failure to mention prescribing error adverse event reports or unintentionally implying that Empirica ® Signal is the only tool used at the FDA. After correcting these mistakes, we refined the materials according to which sections showed the least improvement in testing. Modifications were focused on the two lowest performing sections, Data Mining and Pharmacovigilance, and Data Mining Tools. These changes were generally the removal of confusing elements, such as an unclear chart about sensitivity and specificity and screenshots of data mining tools. These elements had caused confusion to at least one tester, and were not providing enough value to their respective sections. Other changes that were made tended to be slight rewordings of statements in the problem sections, with the goal of making certain concepts more prevalent to the reader. Although we would have liked to improve the tools section further, we did not have enough information to adequately add to it. Since test results were generally favorable, we chose not to do any major restructuring of the material during refinement, keeping it relatively close to our original version.

6. Conclusions and Recommendations

6.1 Technical Recommendations

6.1.1 Natural Language Processing

The FDA is on the cusp of big data and needs to start exploring more ways to manipulate and analyze all of the data. One avenue that the FDA could explore in the future is natural language processing (NLP). NLP is “an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech” (Chowdhury, 2003). This could be useful to the FDA because the adverse event reports contain textual narratives, which may have useful information that is being overlooked by the statistical methods of data mining. NLP would make it possible to extract and summarize key clinical features and time information from the textual narratives in safety reports (Botsis, 2015). Once these key features are extracted, they could be translated to the MedDRA dictionary preferred terms, thus allowing for query-based selection of reports (Botsis, 2015). Figure 11 illustrates the possible deconstruction of a textual narrative by using NLP. A project focused on applications of NLP in pharmacovigilance might provide for an interesting Computer Science/Mathematics MQP at Worcester Polytechnic Institute.



Patient received Smallpox vaccination on 4/21/2006 in left deltoid. 12 days after vaccination he developed increased left arm pain and pleuritic substernal chest pain. 5/11/06 transferred to hospital with chest pain, right arm pain. Final dx of acute myopericarditis, serum reaction, allergic reaction, anemia, abnormal reaction to vaccine. Medical records from previous hospitalization obtained on 5/14/06 showed PMHx of testicular cancer; family hx reveals patient's father had MI.

Figure 11: Information Retrieved from Narratives Through NLP (Botsis, 2015)

6.1.2 Image Processing

The FDA continues on the path to big data by collecting vast amounts of data in the form of adverse event reports, which allow the reporter to attach images to the report. Currently, this option is not commonly utilized by reporters. The few images the FDA does receive include images of pill containers, pills themselves, and adverse events such as rashes. Most of the focus on image submission is with generic drugs (Ellen Pinnow, personal communication, December 8th, 2015). By attaching an image of the generic drug bottle and/or pill, it may be easier for a Safety Evaluator reading the report to figure out which specific company manufactured this product or determine if the consumer was accidentally taking a different drug than they thought they were (Ellen Pinnow, personal communication, December 8th, 2015). If there was a push for patients to submit images of adverse events themselves, then the FDA could have the option of exploring a form of data mining called image processing. Intensively investigated in recent years (Ribeiro, 2009), content-based image retrieval (CBIR) methods aim to search for an image in the image database based on the appearance of the image instead of solely using a textual description (Lehmann, 2005). CBIR relies on "image processing algorithms to extract relevant characteristics (features) from the images" (Ribiero, 2009). Additionally, methods exist to automate the classification and categorization of medical images (Lehmann, 2005). Image processing could open the door for more complex analysis and categorization of adverse event reports based on the pictures that are attached to the report.

6.1.3 Drug-Drug Interactions

While the FDA has data mining tools which can perform drug comparison, the agency does not currently have a data mining tool for drug-drug interactions (Sanjay Sahoo, personal communication, November 17th, 2015). Moving forward, the FDA should consider investigating tools which find multi-item adverse drug event associations, which are "associations relating multiple drugs to possibly multiple adverse events," (Harpaz, 2010). While multi-item adverse drug event associations are rarely reported, they are extremely important as they could indicate a possible drug-drug interaction which may have gone unnoticed (Harpaz, 2010). The implementation of methods to find multi-item adverse drug event associations would be computationally expensive though (Harpaz, 2010). For example, if we have 10,000 unique drugs and adverse events, then the number of possible multi-item adverse drug event associations consisting of 2 drugs and 3 adverse events that needed to be examined would be approximately

10^{20} and further association statistics would need to be calculated (Harpaz, 2010). Similar to the statistical methods of data mining, “some of the associations discovered may be spurious (happening by chance), or due to confounding factors” (Harpaz, 2010). While these challenges exist, some researchers have claimed that they have successfully shown that multi-item adverse drug events are present and could be extracted from AERS (Harpaz, 2010).

6.2 Educational Recommendations

In efforts to further the education of employees on data mining, we would suggest that a member or members of the FDA’s Data Mining Council attend annual conferences on data mining. The members of the council who attend the conferences could compile the information relevant to data mining at the FDA and send the information out to other interested employees in an agency wide newsletter. This would provide other employees the opportunity to continue their education on data mining, without forcing them to travel or take time away from their own work. A few of the popular conferences on data mining are listed below, as well as a link to pages where more information on the conferences can be found.

- KDD – Knowledge Discovery and Data Mining
  http://kdd.org/conferences

- ICDE – International Conference on Data Engineering
  https://www.ieee.org/index.html?WT.mc_id=hpf_logo

- ICDM – IEEE International Conference on Data Mining
  https://www.ieee.org/index.html?WT.mc_id=hpf_logo

- SDM – SIAM International Conference on Data Mining
  http://www.siam.org/

Another option that we would recommend to the FDA is to train new OSE employees and Safety Evaluators about data mining during their initial orientation. The purpose of training is to encourage employees to become familiar with and use data mining methods in their everyday work. This training could be available online or presented at a training seminar. As the current Safety Evaluators are inundated with reports, it is challenging to review all of them manually and data mining could help them perform their analysis quicker and more efficiently. Since data mining can help to identify safety signals in the continuously growing number of reports, Safety Evaluators can save time that would have been spent acting upon these signals and allocate more time to in-depth analyses of the potential safety issue. If the Safety Evaluators were trained about the concept of data mining and how it can be useful to them, they would be more apt to use it in their day to day activities. On the other hand, new OSE employees would benefit from the training because even if it was not immediately relevant to their job, they would have the basic knowledge to understand what was going on if data mining ever came up or became relevant to their work in the future.

6.3 Conclusions

FDA employees who do not fully understand data mining do not trust the use of data mining methods in their everyday work. Without trust and understanding, data mining techniques cannot be used to their fullest extent to enhance employees' pharmacovigilance efforts. If data mining techniques were to be implemented more regularly, then this trust could be built over time. In order for these techniques to be implemented more regularly, the FDA employees need to have a better understanding of data mining. The manual produced by this project is the first step in establishing a preliminary level of comprehension from which trust can be built. With data mining being a constantly evolving field, further revisions of the manual will be necessary in a few years to ensure that the content is up-to-date with current research and applications.

6.4 Individual Team Reflections

Katie Brochu's Reflection

Having the opportunity to work with the FDA has provided me with interesting insight into the field of pharmacovigilance. I had not realized how much time and work the FDA puts in to ensure the safety of the public, and hopefully our project will help the agency with this effort. I think that this was a great experience for us to learn about team dynamics, and how to best utilize each of our respective strengths and compensate for our weaknesses. Another interesting aspect for me was to witness first-hand how the FDA operates, because as a Biomedical

Engineer, I expect to be working with it in the future. Overall this was a very rewarding experience, and I am very grateful to have had this opportunity.

Nicholas Diaz's Reflection

Working for the FDA over these past eight weeks has not been spectacularly easy. Waking up early every morning to ride public transportation for an hour so that I can sit in a cramped office at a government issue workstation is best described in one word: demoralizing. That said, the experience has been a good one. I have learned more about pharmacovigilance than I ever expected to, and have a much deeper understanding and respect for data mining. I have learned what the government is really like on the inside, and what it is like to really focus on a project day in and day out. Our work actually went quite smoothly – we are not, and have not always been perfect, but our team has managed to cooperate well and meet the goals we set for ourselves. Although the small details may have been unclear throughout our journey, we have been able to keep a steady course and make the educational materials that we set out to make, and I am proud of the results. Hopefully the FDA is happy as well, and our work benefits the OSE in years to come.

Sadie Gauthier's Reflection

Working with the FDA has been an incredible experience. I learned the importance of pharmacovigilance and the inspiring work the FDA does to keep the public safe. The OSE staff incorporated us into their ranks, teaching us about the field and about working for the government in general. I learned that in government agencies, everyone must work together to achieve a common mission. I also learned that government work can often be slow moving due to laws and regulations, but it is these laws and regulations that keep the agencies on track. It is amazing to think that just 14 weeks ago I had no idea what data mining was and now I have learned enough to create educational materials on the matter. From an educational stand point, it really shows just how far I have come. I am humbled and honored that our work will be used by the FDA to help the agency to further their goals of public safety.

Erik Nadel's Reflections

Working at the FDA has been a truly unique learning experience. I learned about the significance of pharmacovigilance and the difficulty of monitoring current drugs on the market. Developing these educational materials helped me understand the significance of condensing broad, complex topics into quickly readable forms. It is rewarding to know that the impact of our project will help the OSE staff become more involved in data mining projects working to improve the efficiency of pharmacovigilance operations. Besides the project, it was interesting to learn how the FDA's Office of Surveillance and Epidemiology operates and have the opportunity to attend conferences on the safety of various drugs.

7. Works Cited

Almenoff, J., Tonning, J., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., . . . LaCroix, K. (2005). Perspectives on the Use of Data Mining in Pharmacovigilance. *Drug Safety, 28*(11), 981-1007. doi: 10.2165/00002018-200528110-00002

Aloi, B. K. (2008). Investigation of Reports of Hepatotoxicity Attributed to Dietary Supplement; Hydroxycut—DFPG #09-09; ORA Concurrence #2008121001—High Priority—FACTS #1009567. Washington, D.C.: Department of Health and Human Services U.S. Food and Drug Administration.

Avorn, J., & Schneeweiss, S. (2009). Managing Drug-Risk Information - What to Do with All Those New Numbers. *New England Journal of Medicine, 361*(7), 647-649. doi: 10.1056/NEJMp0905466

Balakin, K. V., & Ekins, S. (2009). *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery*: Wiley.

Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011). Mining FDA drug labels using an unsupervised learning technique - topic modeling. *BMC Bioinformatics, 12*(S11).

Bizer, C., Boncz, P., Brodie, M., & Erling, O. (2012). The meaningful use of big data: four perspectives -- four challenges. *ACM SIGMOD Record, 40*(4), 56-60. doi: 10.1145/2094114.2094129

Botsis, T. (2015). Automating the Deconstruction & Summarization of Adverse Event Reports: U.S. Food and Drug Administration.

Brolund, G., CDER, & Group, I. M. E. W. (2001). *Electronic Transmission of Individual Case Safety Reports Message Specification.* Paper presented at the International Conference on Harmonisation of Technical Requirements of Pharmaceuticals for Human Use, Washington, D.C.

Brown, J. S., Kulldorff, M., Chan, K. A., Davis, R. L., Graham, D., Pettus, P. T., . . . Roblin, D. (2007). Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and drug safety, 16*(12), 1275-1284.

Candore, G., Juhlin, K., Manlik, K., Thakrar, B., Quarcoo, N., Seabroke, S., . . . Slattery, J. (2015). Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Safety, 38*, 577-587.

Cassidy, S. (2004). Learning Styles: An overview of theories, models, and measures. *Educational Psychology, 24*(4), 25.

Chen, M., & Scarazzini, L. (2012). Signal Management Best Practices for Divisions of Pharmacovigilance (pp. 34): Office of Surveillance and Epidemiology Center for Drug Evaluation and Research.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology, 37*(1), 51-89.

Clawson, J. (2006). Adult Learning Theory: It Matters *Teaching Management: A Field Guide for Professors, Consultants, and Corporate Trainers*: Cambridge University Press.

De Bello, T. C. (1990). Comparison of Eleven Major Learning Styles Models: Variables, Appropriate Populations, Validity of Instrumentation, and the Research Behind Them. *Journal of Reading, Writing, and Learning Disabilities International, 6*(3), 19.

Department of Health and Human Services, & Food and Drug Administration. (2015). Food and Drugs *Code of Federal Regulations* (Vol. 5). Washington, D.C.: FDA.

Deshpande, G., Gogolak, V., & Smith, S. W. (2010). Data Mining in Drug Safety. *Pharmaceutical Medicine - New Zealand, 24*(1), 1010-1021.

Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R. A., Baker, J. D., Ball, R., . . . Kass-Hout, T. (2015). Data Mining at FDA--White Paper. Retrieved October 30, 2015, from http://www.fda.gov/ScienceResearch/DataMiningatFDA/ (No longer accessible)

Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R. A., Baker, J. D., Ball, R., . . . Kass-Hout, T. (2015). Use of Data Mining at the Food and Drug Administration. Washington, D.C.: Journal of the American Medical Informatics Association Advance Access.

European Medicines Agency and Heads of Medicines Agencies. (2012). Guideline on Good Pharmacovigilance Practices (GVP) *Signal Management* (pp. 17). United Kingdom: European Medicines Agency.

Evans, S., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety, 10*(6), 483-486.

Evans, S. T. W. (2000). Pharmacovigilance: a Science or Fielding Emergencies? *Statistics in Medicine, 19*(23).

Fayyad, P.-S. (1996). From Data Mining to Knowledge Discovery: An Overview. In P.-S. Fayyad (Ed.), *Advances in Knowledge Discovery and Data Mining* (pp. 1-34). Menlo Park, CA: AAAI Press / The MIT Pres.

FDA. (2015). About FDA. 2015, from http://www.fda.gov/AboutFDA/

FDA. (2015). FDA Adverse Event Reporting System System Data Management Program (FAERS - DMP). Landover, Maryland: U.S. Food and Drug Administration.

FDA. (2015). [Untitled chart of FAERS reports from 2006-2015]. Silver Spring, MD: FDA.

FDA, & CDER. (2013). How Data Mining Supports the Drug Review Process: Q & A with Skip Francis, Office of Translational Sciences. Retrieved October 29, 2015

Fine, A. L. (2013). Introduction to Post-Marketing Drug Safety Surveillance: Pharmacovigilance in FDA/CDER. Washington, DC: U.S. Food and Drug Administration.

Gantz, J., & Reinsel, D. (2010). The Digital Universe Decade - Are You Ready? *IDC White Paper*. Framingham, Massachusetts: IDC.

Google. (2008). Google Flu Trends. Retrieved September 25, 2015, from https://www.google.org/flutrends/about/

Gravel, C. (Producer). (November 18, 2015). An Introduction to the Statistical Methods for Signal Detection in Pharmacovigilance. Retrieved from http://www.mclaughlincentre.ca/events/Parmaco2/Chris%20Gravel_UofO%20workshop%20May%2028-29%202009.pdf

Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods, 18*(1), 23.

Hallberg, P., & Sundstrom, A. (2009). Data Mining in Pharmacovigilance -- Detecting the Unexpected: the Role of Index of Suspicion of the Reporter. *Drug Safety, 32*.

Härmark, L., & van Grootheest, A. C. (2008). Pharmacovigilance: methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology, 64*(8), 743-752. doi: 10.1007/s00228-008-0475-9

Harpaz, R., Chase, H. S., & Friedman, C. (2010). Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics, 11*(Suppl 9), S7.

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of Pharmacovigilance Signal Detection Algorithms for the FDA Adverse Event Reporting System. *Clinical pharmacology and therapeutics, 93*(6), 10.1038/clpt.2013.1024. doi: 10.1038/clpt.2013.24

Harpaz, R., DuMouchel, W., Shah, N., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics, 91*(6), 37.

Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., & Van Puijenbroek, E. P. (2005). The Role of Data Mining in Pharmacovigilance. *Expert Opinion on Drug Safety, 4*(5), 929-948.

Holloway, C. (2013). Standard Database Descriptions: AERS.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM, 52*(8), 36-44.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews: Genetics, 13*.

Johnson, K., Guo, C., Gosink, M., Wang, V., & Hauben, M. (2012). Multinomial Modeling and An Evaluation of Common Data-Mining Algorithms for Identifying Signals of Disproportionate Reporting in Pharmacovigilance Databases. *Bioinformatics, 28*(23).

Knowles, M. S., Holton III, E. F., & Swanson, R. A. (2014). *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development*: Routledge.

Kolb, A. Y. (2005). The Kolb learning style inventory—version 3.1 2005 technical specifications. *Hay Resource Direct, 200*.

Kolb, D., & Kolb, A. (2012). Kolb's Learning Styles. In N. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1698-1703). New York, New York: Springer US.

Lehmann, T. M., Guld, M. O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., . . . Wein, B. B. (2005). Automatic Categorization of Medical Images for Content-Based Retrieval and Data Mining. *Computerized Medical Imaging and Graphics, 29*(2-3), 143-155.

Li, Q., Deleger, L., Lingren, T., Zhai, H., Kaiser, M., Stoutenborough, L., . . . Solti, I. (2013). Mining FDA drug labels for medical conditions. *BMC Medical Informatics and Decision Making, 13*(53).

Mason, H., & Wiggins, C. (2010). A Taxonomy of Data Science.   Retrieved September 19, 2015, from http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

Miguel de Almeida Vieira Lima, L., Goncalo Sales Craveiro Nunes, N., Goncalo Pires da Silva Dias, P., & Jorge Batel Marques, F. (2012). Implemented Data Mining and Signal Management Systems on Spontaneous Reporting Systems' Databases and their Availability to the Scientific Community-A Systematic Review. *Current drug safety, 7*(2), 170-175.

National Committee on Vital and Health Services. (2014). Frequently Asked Questions About Medical and Public Health Registries.   Retrieved October 30, 2015, from http://ncvhs.hhs.gov/9701138b.htm (No longer accessible)

Northwest Center for Public Health Practice. (2014). *Effective Adult Learning: A Toolkit for Teaching Adults* Retrieved from https://www.nwcphp.org/documents/training/Adult_Education_Toolkit.pdf

Organization, W. H. (2015). Retrieved December 2, 2015, from http://www.who.int/en/

Phillips, L. (1994). *The Continuing Education Guide: The CEU and Other Professional Development Criteria*: Louis Phillips.

Rajola, F. (2014). *Customer Relationship Management in the Financial Industry: Organizational Processes and Technology Innovation* (2 ed.): Springer Science & Business Media.

Ribeiro, M. X., Balan, A. G. R., Felipe, J. C., Traina, A. J. M., & Traina Jr, C. (2009). Mining Complex Data. In D. A. Zighed, S. Tsumoto, Z. W. Ras, & H. Hacid (Eds.), *Studies in Computational Intelligence* (Vol. 165, pp. 113-131): Springer Berlin Heidelberg.

Rothman, K. J., Lanes, S., & Sacks, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and drug safety, 13*(8), 519-523.

Sahoo, S. (2015). Support for DPV/DMEPA FAERS/FBIS Eco-System: Status Report (December 2014 to Today). Washington, D.C: U.S. Food and Drug Administration.

Sakaeda, T., Tamon, A., Kadoyama, K., & Okuno, Y. (2013). Data Mining of the Public Version of the FDA Adverse Event Reporting System. *International Journal of Medical Sciences, 10*(7), 796-803. doi: 10.7150/ijms.6048

Secretariat, I. (2013). Understanding MedDRA: The Medical Dictionary for Regulatory Activities. Washington, D.C.: IFPMAICH.

Seifert, J. W. (2004). Data Mining: An Overview. In D. Pegarkov (Ed.), *National Security Issues* (pp. 201-217). New York: Nova Science Publishers, Inc.

Sukumar, S. R., Natarajan, R., & Ferrell, R. K. (2015). Quality of Big Data in health care. *International Journal of Health Care Quality Assurance, 28*(6), 621-634. doi: doi:10.1108/IJHCQA-07-2014-0080

Szarfman, A., Machado, S., & O’Neill, R. (2002). Use of Screening Algorithms and Computer Systems to Efficiently Signal Higher-Than-Expected Combinations of Drugs and Events in the US FDA’s Spontaneous Reports Database. *Drug Safety, 25*(6), 381-392. doi: 10.2165/00002018-200225060-00001

Szarfman, A., Tonning, J. M., Levine, J. G., & Doraiswamy, P. M. (2006). Atypical Antipsychotics and Pituitary Tumors: A Pharmacovigilance Study. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 26*(6), 748-758.

Tatonetti, N. P., Denny, J., Murphy, S., Fernald, G., Krishnan, G., Castro, V., . . . Roden, D. (2011). Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology & Therapeutics, 90*(1), 133-142.

Tonning, J. (2015). Anonymized Data: Considerations for Data Mining. Washington, D.C.: U.S. Food and Drug Administration.

Trochim, W. M. K. (2006). Sampling Terminology. *Researched Methods Knowledge Base.* from http://www.socialresearchmethods.net/kb/sampterm.php

Tsumoto, S., & Hirano, S. (2010). Risk Mining in Medicine: Application of Data Mining to Medical Risk Management. *Fundamenta Informaticae, 98*(1), 107-121.

U.S. Department of Health and Human Services, & Food and Drug Administration. (2005). Guidance for Industry. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. Washington, DC: FDA.

U.S. Food and Drug Administration. (2013). Reports Received and Reports Entered into FAERS by Year. Retrieved November 4, 2015, from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm

U.S. Food and Drug Administration. (2015). 21st Century Cures Act Touted as a Cure, but Some Hesitant About Costs. *Adverse Event Reporting News, 11*(19), 8.

Verma, I. S. (2015). Knowledge Data Discovery and Its Issues. *Expansion, Impact and Challenges of IT & CS*, 88.

VIII, C. W. G. (2010). *Practical Aspects of Signal Detection in Pharmacovigilance*. Geneva: Council for International Organizations of Medical Science.

Vilar, S., Harpaz, R., Chase, H. S., Costanzi, S., Rabadan, R., & Friedman, C. (2011). Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *Journal of the American Medical Informatics Association, 18*(Supplement 1), i73-i80. doi: 10.1136/amiajnl-2011-000417

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.

Wilson, A. M., Thabane, L., & Holbrook, A. (2004). Application of Data Mining Techniques in Pharmacovigilance. *British Journal of Clinical Pharmacology, 57*(2).

Xindong, W., Xingquan, Z., Gong-Qing, W., & Wei, D. (2014). Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on, 26*(1), 97-107. doi: 10.1109/TKDE.2013.109

Yap, C. W., Cai, C. Z., Xue, Y., & Chen, Y. Z. (2004). Prediction of Torsade-Causing Potential of Drugs by Support Vector Machine Approach. *Toxicological Sciences, 79*(1), 170-177. doi: 10.1093/toxsci/kfh082

Zorych, I., Madigan, D., Ryan, P., & Bate, A. (2013). Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical methods in medical research, 22*(1), 39-56.

8.1 Appendix A: Example Questions for the Data Mining Experts

Question for individuals who are already experts on data mining:

- What would you consider basic knowledge for understanding data mining as it applies to pharmacovigilance?
    - How much theoretical knowledge should new Safety Evaluators in the Division of Pharmacovigilance (DPV) have?
    - What level of knowledge should other Office of Surveillance staff have?
- Based on your experience:
    - Describe the project you worked on that required data mining knowledge
    - What data mining tools were used?
    - Which models/algorithms have been most applicable to pharmacovigilance?
    - Which specific sets of data have you worked with?
- What other data mining tools do you know that may be applicable to pharmacovigilance?

8.2 Appendix B: Informed Consent Agreement

Informed Consent Agreement for Participation in a Research

Study Investigators: Katie Brochu, Nicholas Diaz, Sadie Gauthier,

Erik Nadel **Contact Information:** dc15-fda@wpi.edu

Title of Research Study: Data Mining Materials Assessment

Sponsor: FDA

Introduction

You are being asked to participate in a research study. Before you agree, however, you must be fully informed about the purpose of the study, the procedures to be followed, and any benefits, risks or discomfort that you may experience as a result of your participation. This form presents information about the study so that you may make a fully informed decision regarding your participation.

Purpose of the study: The purpose of this study is to create and evaluate educational materials to be used by other OSE employees. Using your feedback, we can improve the content of the materials and ensure that they are effective and easily taught.

Procedures to be followed: For this study, you will be asked to participate in one or more interviews about your current knowledge of data mining and its applications at the FDA. You might also be asked about evaluating drafted educational materials covering basic data mining topics. If you are chosen as a tester, you will complete a short test before and after being presented educational materials. These tests will be short and will ask basic questions covering basic data mining topics and applications.

Risks to study participants: There is no risk involved in this research. The only procedures involve only basic questions on the subject of data mining. No personal questions will be asked.

Benefits to research participants and others: There are no benefits given. This participation is done entirely on a voluntary basis.

Record keeping and confidentiality: The only data collected will be interview responses, test results, and feedback on educational materials. Records of your participation in this study will be held confidential so far as permitted by law. However, the study investigators, the sponsor or it's designee and, under certain circumstances, the Worcester Polytechnic Institute Institutional Review Board (WPI IRB) will be able to inspect and have access to confidential data that
identify you by name. Any publication or presentation of the data will not identify you.

For more information about this research or about the rights of research participants, or in case of research-related injury, contact: Researchers Katie Brochu, Erik Nadel, Sadie Gauthier, and Nicholas Diaz at dc15-fda@wpi.edu OR

Professor Kent Rissmiller, Tel. 508-831-5019, Email: kjr@wpi.edu) and the University Compliance Officer (Jon Bartelson, Tel. 508-831-5725, Email: jonb@wpi.edu

Your participation in this research is voluntary. Your refusal to participate will not result in any penalty to you or any loss of benefits to which you may otherwise be entitled. You may decide to stop participating in the research at any time without penalty or loss of other benefits. The project investigators retain the right to cancel or postpone the experimental procedures at any time they see fit.

By signing below, you acknowledge that you have been informed about and consent to be a participant in the study described above. Make sure that your questions are answered to your satisfaction before signing. You are entitled to retain a copy of this consent agreement.

____________________________ Date: ____________________

Study Participant
Signature

____

Study Participant Name (Please print)

______________________________________ Date: ___________________

Signature of Person who explained this study

8.3 Appendix C: List of Interviewees

Interviewees

Data Mining Experts

- Robert Ball, Deputy Director, OSE, CDER
- Henry “Skip” Francis, Supervisory Medical Officer, OTS, CDER
- Carol Pamer, Regulatory Science Program Lead, RSS, OSE, CDER
- Ana Szarfman, Medical Officer, OTS, CDER
- Joseph Tonning, Medical Officer, OTS, CDER

Adult Learning Experts

- Dorrie Ballman, Staff Development Specialist, OSE, CDER

Divisional Experts

- Gerald Dal Pan, Director, OSE, CDER
- Marni Hall, Director, RSS, OSE, CDER
- Suranjan De, Deputy Director, RSS, OSE, CDER

FAERS Experts

- Sanjay Sahoo, Operations Research Analyst, RSS, OSE, CDER
- John Quinn, Senior Program Manager, RSS, OSE, CDER

MedDRA Experts

- Sonja Brajovic, Medical Officer, RSS, OSE, CDER

Reviewers

- Cindy Kortepeter, Deputy Director, Division of Pharmacovigilance I, OSE, CDER

Technology Transfer Staff Members

- Chekesha Clingman, Special Assistant, OTS, CDER

- Shaniece Bowens, Project Manager, OTS, CDER
- Yolanda Mock Hawkins, Technology Transfer Specialist, OTS, CDER

8.4 Appendix D: MedWatch Forms

3500A Form: Pages 79-81

3500B Form: Pages 82-86

3500 Form: Pages 87-89

Reset Form

Form Approved: OMB No. 0910-0291, Expires: 6/30/2015
See OMB statement on reverse.

U.S. Department of Health and Human Services
Food and Drug Administration

MEDWATCH

FORM FDA 3500A (2/13)

For use by user-facilities,
importers, distributors and manufacturers
for MANDATORY reporting

Page 1 of ____

Mfr Report #

UF/Importer Report #

FDA Use Only

PLEASE TYPE OR USE BLACK INK

A. PATIENT INFORMATION

1. Patient Identifier

In confidence

2. Age at Time of Event:

or

Date of Birth:

3. Sex

□ Female

□ Male

4. Weight

____ lbs

or

____ kgs

B. ADVERSE EVENT OR PRODUCT PROBLEM

1. □ Adverse Event and/or □ Product Problem *(e.g., defects/malfunctions)*

2. Outcomes Attributed to Adverse Event
(Check all that apply)

□ Death: ____ *(mm/dd/yyyy)*

□ Disability or Permanent Damage

□ Life-threatening

□ Congenital Anomaly/Birth Defect

□ Hospitalization - initial or prolonged

□ Other Serious (Important Medical Events)

□ Required Intervention to Prevent Permanent Impairment/Damage (Devices)

3. Date of Event *(mm/dd/yyyy)*

4. Date of This Report *(mm/dd/yyyy)*

5. Describe Event or Problem

(Continue on page 3)

6. Relevant Tests/Laboratory Data, Including Dates

(Continue on page 3)

7. Other Relevant History, Including Preexisting Medical Conditions *(e.g., allergies, race, pregnancy, smoking and alcohol use, hepatic/renal dysfunction, etc.)*

(Continue on page 3)

Submission of a report does not constitute an admission that medical personnel, user facility, importer, distributor, manufacturer or product caused or contributed to the event.

C. SUSPECT PRODUCT(S)

1. Name *(Give labeled strength & mfr/labeler)*

#1

#2

2. Dose, Frequency & Route Used

#1

#2

3. Therapy Dates *(If unknown, give duration) from/to (or best estimate)*

#1

#2

4. Diagnosis for Use *(Indication)*

#1

#2

5. Event Abated After Use Stopped or Dose Reduced?

#1 □ Yes □ No □ Doesn't Apply

#2 □ Yes □ No □ Doesn't Apply

6. Lot #

#1

#2

7. Exp. Date

#1

#2

8. Event Reappeared After Reintroduction?

#1 □ Yes □ No □ Doesn't Apply

#2 □ Yes □ No □ Doesn't Apply

9. NDC# or Unique ID

10. Concomitant Medical Products and Therapy Dates *(Exclude treatment of event)*

(Continue on page 3)

D. SUSPECT MEDICAL DEVICE

1. Brand Name

2. Common Device Name

2b. Procode

3. Manufacturer Name, City and State

4. Model #

Lot #

Catalog #

Expiration Date *(mm/dd/yyyy)*

Serial #

Unique Identifier (UDI) #

5. Operator of Device

□ Health Professional

□ Lay User/Patient

□ Other:

6. If Implanted, Give Date *(mm/dd/yyyy)*

7. If Explanted, Give Date *(mm/dd/yyyy)*

8. Is this a Single-use Device that was Reprocessed and Reused on a Patient?

□ Yes □ No

9. If Yes to Item No. 8, Enter Name and Address of Reprocessor

10. Device Available for Evaluation? *(Do not send to FDA)*

□ Yes □ No □ Returned to Manufacturer on: ____ *(mm/dd/yyyy)*

11. Concomitant Medical Products and Therapy Dates *(Exclude treatment of event)*

(Continue on page 3)

E. INITIAL REPORTER

1. Name and Address

Phone #

Email Address

2. Health Professional?

□ Yes □ No

3. Occupation

4. Initial Reporter Also Sent Report to FDA

□ Yes □ No □ Unk.

Reset Form

MEDWATCH

FORM FDA 3500A (2/13) *(continued)*

Page 2 of

FDA USE ONLY

F. FOR USE BY USER FACILITY/IMPORTER *(Devices Only)*

1. **Check One**

□ User Facility □ Importer

2. **UF/Importer Report Number**

3. **User Facility or Importer Name/Address**

4. **Contact Person**

5. **Phone Number**

6. **Date User Facility or Importer Became Aware of Event** *(mm/dd/yyyy)*

7. **Type of Report**

□ Initial

□ Follow-up #

8. **Date of This Report** *(mm/dd/yyyy)*

9. **Approximate Age of Device**

10. **Event Problem Codes** *(Refer to coding manual)*

Patient Code – –

Device Code – –

11. **Report Sent to FDA?**

□ Yes *(mm/dd/yyyy)*

□ No

12. **Location Where Event Occurred**

□ Hospital

□ Home

□ Nursing Home

□ Outpatient Treatment Facility

□ Outpatient Diagnostic Facility

□ Ambulatory Surgical Facility

□ Other: *(Specify)*

13. **Report Sent to Manufacturer?**

□ Yes *(mm/dd/yyyy)*

□ No

14. **Manufacturer Name/Address**

G. ALL MANUFACTURERS

1. **Contact Office (and Manufacturing Site for Devices)**

Name

Address

Email Address

2. **Phone Number**

3. **Report Source** *(Check all that apply)*

□ Foreign

□ Study

□ Literature

□ Consumer

□ Health Professional

□ User Facility

□ Company Representative

□ Distributor

□ Other:

4. **Date Received by Manufacturer** *(mm/dd/yyyy)*

5.

(A)NDA #

IND #

BLA #

PMA/ 510(k) #

Combination Product □ Yes

Pre-1938 □ Yes

OTC Product □ Yes

6. **If IND, Give Protocol #**

7. **Type of Report** *(Check all that apply)*

□ 5-day □ 30-day

□ 7-day □ Periodic

□ 10-day □ Initial

□ 15-day □ Follow-up #

9. **Manufacturer Report Number**

8. **Adverse Event Term(s)**

H. DEVICE MANUFACTURERS ONLY

1. **Type of Reportable Event**

□ Death

□ Serious Injury

□ Malfunction

2. **If Follow-up, What Type?**

□ Correction

□ Additional Information

□ Response to FDA Request

□ Device Evaluation

3. **Device Evaluated by Manufacturer?**

□ Not Returned to Manufacturer

□ Yes □ Evaluation Summary Attached

□ No *(Attach page to explain why not)* or provide code:

4. **Device Manufacture Date** *(mm/yyyy)*

5. **Labeled for Single Use?**

□ Yes □ No

6. **Event Problem and Evaluation Codes** *(Refer to coding manual)*

Patient Code – –

Device Code – –

Method – – –

Results – – –

Conclusions – – –

7. **If Remedial Action Initiated, Check Type**

□ Recall □ Notification

□ Repair □ Inspection

□ Replace □ Patient Monitoring

□ Relabeling □ Modification/ Adjustment

□ Other:

8. **Usage of Device**

□ Initial Use of Device

□ Reuse

□ Unknown

9. **If action reported to FDA under 21 USC 360i(f), list correction/ removal reporting number:**

10. □ **Additional Manufacturer Narrative** and / or 11. □ **Corrected Data**

This section applies only to requirements of the Paperwork Reduction Act of 1995.

The public reporting burden for this collection of information has been estimated to average 66 minutes per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to:

Department of Health and Human Services
Food and Drug Administration
Office of Chief Information Officer
Paperwork Reduction Act (PRA) Staff
PRAStaff@fda.hhs.gov

Please DO NOT RETURN this form to the above PRA Staff email address.

OMB Statement: "An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number."

Reset Form Delete Page

(CONTINUATION PAGE)

For use by user-facilities,
importers, distributors, and manufacturers
for MANDATORY reporting

Page 3 of

MEDWATCH

FORM FDA 3500A (2/13) *(continued)*

B.5. **Describe Event or Problem** *(continued)*

Back to Item B.5

B.6. **Relevant Tests/Laboratory Data, Including Dates** *(continued)*

Back to Item B.6

B.7. **Other Relevant History, Including Preexisting Medical Conditions** *(e.g., allergies, race, pregnancy, smoking and alcohol use, hepatic/renal dysfunction, etc.) (continued)*

Back to Item B.7

Concomitant Medical Products and Therapy Dates *(Exclude treatment of event) (For continuation of C.10 and/or D.11; please distinguish)*

Back to Item C.10

Back to Item D.11

Other Remarks

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Food and Drug Administration

Form Approved: OMB No. 0910-0291
Expiration Date: 6/30/2015
(See PRA Statement below)

MEDWATCH Consumer Voluntary Reporting (FORM FDA 3500B)

When do I use this form?

- You were hurt or had a bad side effect (including new or worsening symptoms) after taking a drug or using a medical device or product.
- You used a drug, product, or medical device incorrectly which could have or led to unsafe use.
- You noticed a problem with the quality of the drug, product, or medical device.
- You had problems with how a drug worked after switching from one maker to another maker.

Don't use this form to report:

- Vaccines – report problems to the Vaccine Adverse Event Reporting System (VAERS)
- Investigational drugs or medical devices (those being studied, not yet approved) – report problems to your doctor or to the contact person listed in the clinical trial

Will the information I report be kept private?

The FDA recognizes that privacy is an important concern, so you should know:

- We ask only for the name and contact information of the person filling out the form in case we need more information. This information will not be given out to the public.
- Information about the problem may be shared with the company that makes the product to help them better understand the problem you are reporting, unless you request otherwise (see Section E).

What types of products should I use this form for?

- Drugs, including prescription or over-the-counter medicines, and biologics, such as human cells and tissues used for transplantation (for example, tendons, ligaments, and bone) and gene therapies
- Medical devices, including any health-related kit, test, tool, or piece of equipment (such as breast implants, pacemakers, diabetes glucose-test kits, hearing aids, breast pumps, and many others)
- Nutrition products, including vitamins and minerals, herbal remedies, infant formulas, and medical foods, such as those labeled for people with a specific disease or condition
- Cosmetics or make-up products
- Foods (including beverages and ingredients added to foods)

Are there specific instructions for filling out the form?

- Fill in as much information as possible and send in the report even if you do not have all the information.
- You can fill out this form yourself or have someone fill it out for you. If you need help, you may want to talk with your health professional.
- Feel free to include or attach an image. Please do not send the products to the FDA.

How will I know the FDA has received my form?

- You will receive a reply from the FDA after we receive your report. We will personally contact you only if we need additional information.
- Your report will become part of a database so that it can be reviewed and compared to other reports by an FDA safety evaluator who will determine what steps to take.

How can I contact the FDA if I have questions?

Toll-free line: 1-800-332-1088
www.fda.gov/reportinghelp
To report online: www.fda.gov/medwatch/report.htm

The information below applies only to requirements of the Paperwork Reduction Act of 1995.

The burden time for this collection of information is estimated to average 25 minutes per response, including the time to review instructions, search existing data sources, gather and maintain the data needed and complete and review the collection of information. Send comments regarding this burden estimate or any other aspect of this information collection, including suggestions for reducing this burden, to the address to the right:

OMB Statement: "An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB number."

Department of Health and Human Services
Food and Drug Administration
Office of Chief Information Officer
Paperwork Reduction Act (PRA) Staff
PRAStaff@fda.hhs.gov

DO NOT SEND YOUR COMPLETED FORM TO THIS PRA STAFF EMAIL ADDRESS.

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Food and Drug Administration

MEDWATCH Consumer Voluntary Reporting
(FORM FDA 3500B)

Form Approved: OMB No. 0910-0291
Expiration Date: 6/30/2015
(See PRA Statement on preceding general information page)

Section A – About the Problem

What kind of problem was it? *(Check all that apply)*

□ Were hurt or had a bad side effect *(including new or worsening symptoms)*

□ Used a product incorrectly which could have or led to a problem

□ Noticed a problem with the quality of the product

□ Had problems after switching from one product maker to another maker

Date the problem occurred *(mm/dd/yyyy)*

Did any of the following happen? *(Check all that apply)*

□ Hospitalization – admitted or stayed longer

□ Required help to prevent permanent harm *(for medical devices only)*

□ Disability or health problem

□ Birth defect

□ Life-threatening

□ Death *(Include date)*: ____________________

□ Other serious/important medical incident *(Please describe below)*

Tell us what happened and how it happened. *(Include as many details as possible)*

Continuation Page

List any relevant tests or laboratory data if you know them. *(Include dates)*

Continuation Page

For a problem with a product, including

- prescription or over-the-counter medicine
- biologics, such as human cells and tissues used for transplantation (for example, tendons, ligaments, and bone) and gene therapies
- nutrition products, such as vitamins and minerals, herbal remedies, infant formulas, and medical foods
- cosmetics or make-up products
- foods (including beverages and ingredients added to foods)

➡ **Go to Section B**

For a problem with a medical device, including

- any health-related test, tool, or piece of equipment
- health-related kits, such as glucose monitoring kits or blood pressure cuffs
- implants, such as breast implants, pacemakers, or catheters
- other consumer health products, such as contact lenses, hearing aids, and breast pumps

➡ **Go to Section C (Skip Section B)**

For more information, visit *http://www.fda.gov/MedWatch*

Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.

Section B – About the Products

Name of the product as it appears on the box, bottle, or package *(Include as many names as you see)*

Name of the company that makes the product

Expiration date *(mm/dd/yyyy)*	Lot number	NDC number

Strength *(for example, 250 mg per 500 mL or 1 g)*	Quantity *(for example, 2 pills, 2 puffs, or 1 teaspoon, etc.)*	Frequency *(for example, twice daily or at bedtime)*	How was it taken or used *(for example, by mouth, by injection, or on the skin)*?

Date the person first started taking or using the product *(mm/dd/yyyy)*: ____________

Date the person stopped taking or using the product *(mm/dd/yyyy)*: ____________

Why was the person using the product *(such as, what condition was it supposed to treat?)*

____

____

____

Did the problem stop after the person reduced the dose or stopped taking or using the product? □ Yes □ No

Did the problem return if the person started taking or using the product again?

□ Yes □ No □ Didn't restart

Do you still have the product in case we need to evaluate it? *(Do not send the product to FDA. We will contact you directly if we need it.)*

□ Yes □ No

➪ ***Go to Section D (Skip Section C)***

Section C – About the Medical Device

Name of medical device

Name of the company that makes the medical device

Other identifying information *(The model, catalog, lot, serial, or UDI number, and the expiration date, if you can locate them)*

____

____

____

Was someone operating the medical device when the problem occurred?

□ Yes

□ No

If yes, who was using it?

□ The person who had the problem

□ A health professional *(such as a doctor, nurse, or aide)*

□ Someone else *(Please explain who)*

____

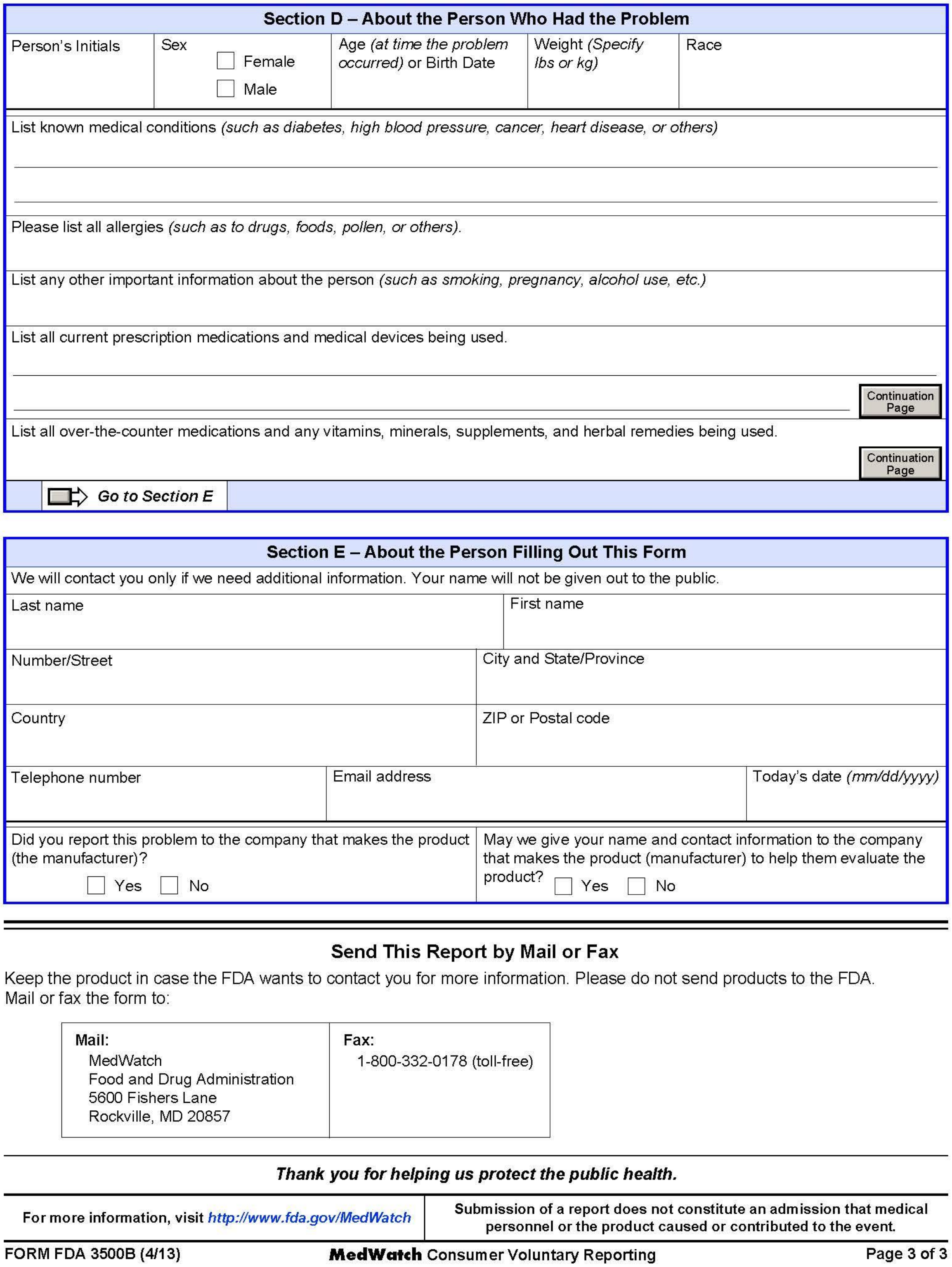
For implanted medical devices ONLY *(such as pacemakers, breast implants, etc.)*

Date the implant was put in *(mm/dd/yyyy)*	Date the implant was taken out *(If relevant) (mm/dd/yyyy)*

➪ ***Go to Section D***

For more information, visit http://www.fda.gov/MedWatch

Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.



Section D – About the Person Who Had the Problem

Person's Initials	Sex □ Female □ Male	Age *(at time the problem occurred)* or Birth Date	Weight *(Specify lbs or kg)*	Race

List known medical conditions *(such as diabetes, high blood pressure, cancer, heart disease, or others)*

Please list all allergies *(such as to drugs, foods, pollen, or others).*

List any other important information about the person *(such as smoking, pregnancy, alcohol use, etc.)*

List all current prescription medications and medical devices being used.

Continuation Page

List all over-the-counter medications and any vitamins, minerals, supplements, and herbal remedies being used.

Continuation Page

Go to Section E

Section E – About the Person Filling Out This Form

We will contact you only if we need additional information. Your name will not be given out to the public.

Last name	First name	
Number/Street	City and State/Province	
Country	ZIP or Postal code	
Telephone number	Email address	Today's date *(mm/dd/yyyy)*

Did you report this problem to the company that makes the product (the manufacturer)? □ Yes □ No	May we give your name and contact information to the company that makes the product (manufacturer) to help them evaluate the product? □ Yes □ No

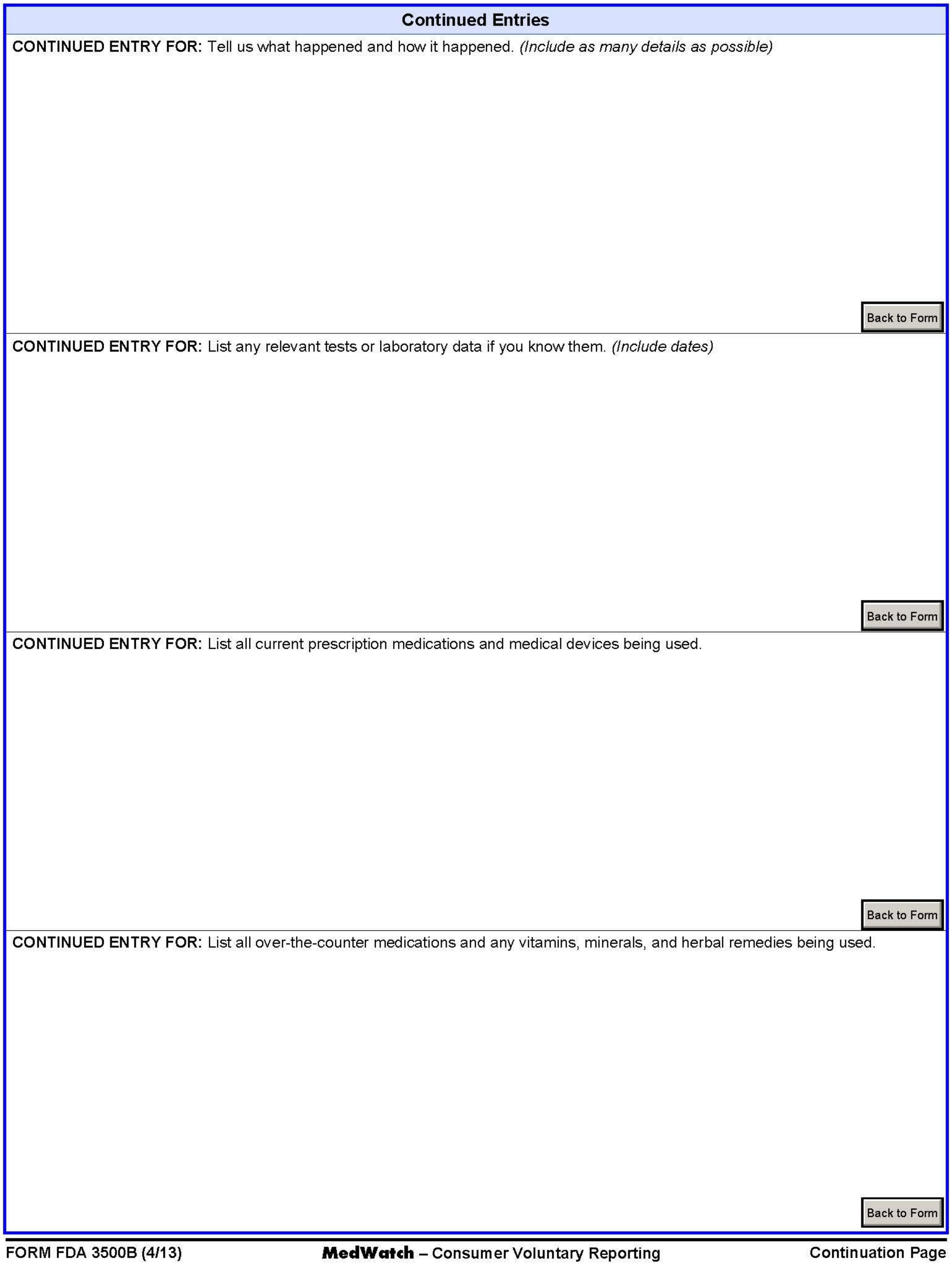
Send This Report by Mail or Fax

Keep the product in case the FDA wants to contact you for more information. Please do not send products to the FDA.
Mail or fax the form to:

Mail: MedWatch Food and Drug Administration 5600 Fishers Lane Rockville, MD 20857	**Fax:** 1-800-332-0178 (toll-free)

Thank you for helping us protect the public health.

For more information, visit *http://www.fda.gov/MedWatch*	**Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.**



Continued Entries

CONTINUED ENTRY FOR: Tell us what happened and how it happened. *(Include as many details as possible)*

Back to Form

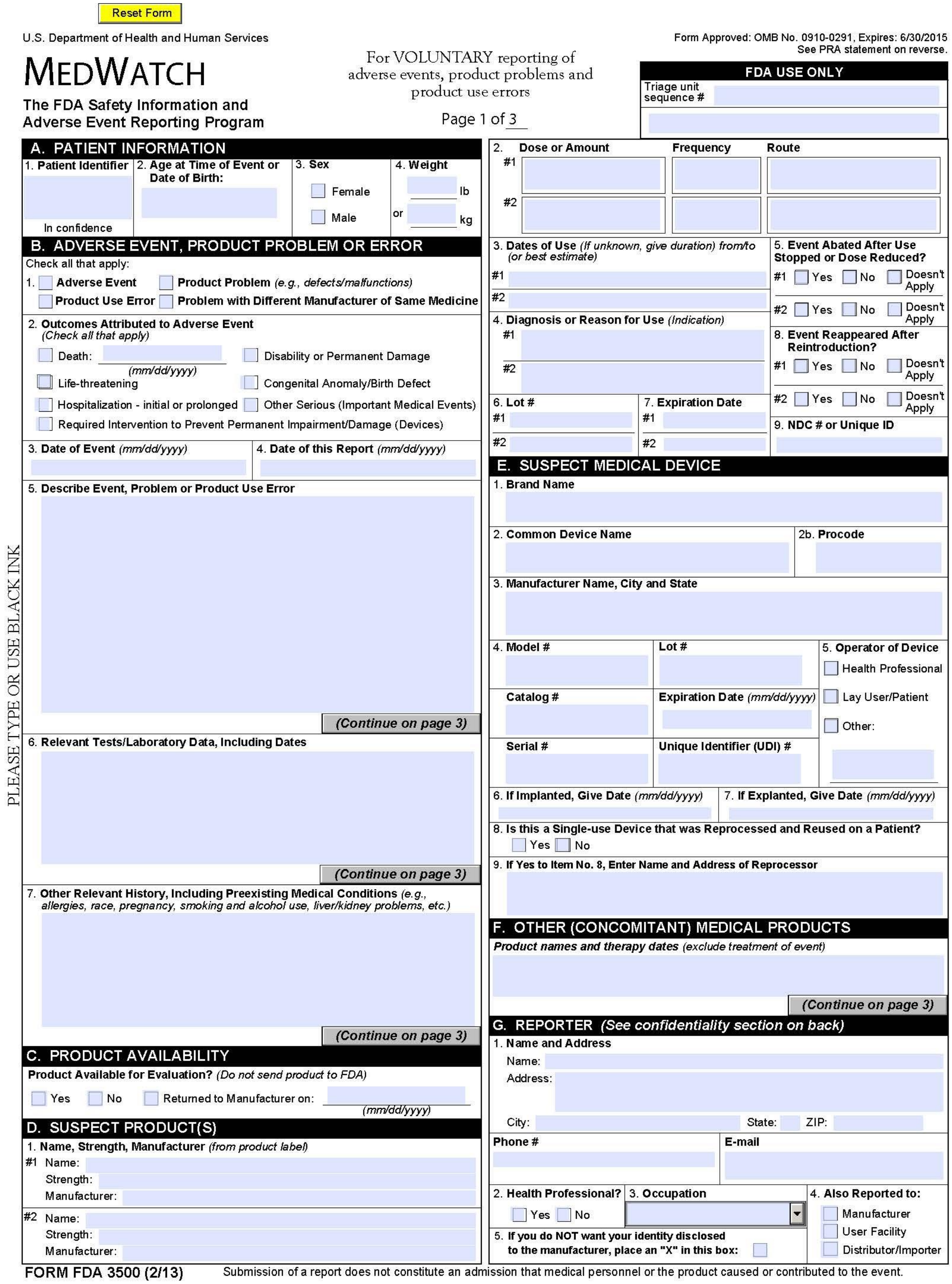
CONTINUED ENTRY FOR: List any relevant tests or laboratory data if you know them. *(Include dates)*

Back to Form

CONTINUED ENTRY FOR: List all current prescription medications and medical devices being used.

Back to Form

CONTINUED ENTRY FOR: List all over-the-counter medications and any vitamins, minerals, and herbal remedies being used.



Reset Form

U.S. Department of Health and Human Services

MEDWATCH

The FDA Safety Information and Adverse Event Reporting Program

For VOLUNTARY reporting of adverse events, product problems and product use errors

Page 1 of 3

Form Approved: OMB No. 0910-0291, Expires: 6/30/2015
See PRA statement on reverse.

FDA USE ONLY

Triage unit sequence #

PLEASE TYPE OR USE BLACK INK

A. PATIENT INFORMATION

1. **Patient Identifier** In confidence
2. **Age at Time of Event or Date of Birth:**
3. **Sex** □ Female □ Male
4. **Weight** lb or kg

B. ADVERSE EVENT, PRODUCT PROBLEM OR ERROR

Check all that apply:

1. □ **Adverse Event** □ **Product Problem** *(e.g., defects/malfunctions)*
□ **Product Use Error** □ **Problem with Different Manufacturer of Same Medicine**

2. **Outcomes Attributed to Adverse Event** *(Check all that apply)*

□ Death: (mm/dd/yyyy) □ Disability or Permanent Damage
□ Life-threatening □ Congenital Anomaly/Birth Defect
□ Hospitalization - initial or prolonged □ Other Serious (Important Medical Events)
□ Required Intervention to Prevent Permanent Impairment/Damage (Devices)

3. **Date of Event** *(mm/dd/yyyy)*
4. **Date of this Report** *(mm/dd/yyyy)*

5. **Describe Event, Problem or Product Use Error**

(Continue on page 3)

6. **Relevant Tests/Laboratory Data, Including Dates**

(Continue on page 3)

7. **Other Relevant History, Including Preexisting Medical Conditions** *(e.g., allergies, race, pregnancy, smoking and alcohol use, liver/kidney problems, etc.)*

(Continue on page 3)

C. PRODUCT AVAILABILITY

Product Available for Evaluation? *(Do not send product to FDA)*

□ Yes □ No □ Returned to Manufacturer on: (mm/dd/yyyy)

D. SUSPECT PRODUCT(S)

1. **Name, Strength, Manufacturer** *(from product label)*

#1 Name:
Strength:
Manufacturer:

#2 Name:
Strength:
Manufacturer:

2. **Dose or Amount** **Frequency** **Route**
#1
#2

3. **Dates of Use** *(If unknown, give duration) from/to (or best estimate)*
#1
#2

4. **Diagnosis or Reason for Use** *(Indication)*
#1
#2

5. **Event Abated After Use Stopped or Dose Reduced?**
#1 □ Yes □ No □ Doesn't Apply
#2 □ Yes □ No □ Doesn't Apply

6. **Lot #**
#1
#2

7. **Expiration Date**
#1
#2

8. **Event Reappeared After Reintroduction?**
#1 □ Yes □ No □ Doesn't Apply
#2 □ Yes □ No □ Doesn't Apply

9. **NDC # or Unique ID**

E. SUSPECT MEDICAL DEVICE

1. **Brand Name**

2. **Common Device Name**

2b. **Procode**

3. **Manufacturer Name, City and State**

4. **Model #** **Lot #**

Catalog # **Expiration Date** *(mm/dd/yyyy)*

Serial # **Unique Identifier (UDI) #**

5. **Operator of Device**
□ Health Professional
□ Lay User/Patient
□ Other:

6. **If Implanted, Give Date** *(mm/dd/yyyy)*
7. **If Explanted, Give Date** *(mm/dd/yyyy)*

8. **Is this a Single-use Device that was Reprocessed and Reused on a Patient?**
□ Yes □ No

9. **If Yes to Item No. 8, Enter Name and Address of Reprocessor**

F. OTHER (CONCOMITANT) MEDICAL PRODUCTS

Product names and therapy dates *(exclude treatment of event)*

(Continue on page 3)

G. REPORTER *(See confidentiality section on back)*

1. **Name and Address**

Name:
Address:

City: State: ZIP:

Phone # **E-mail**

2. **Health Professional?** □ Yes □ No
3. **Occupation**
4. **Also Reported to:**
□ Manufacturer
□ User Facility
□ Distributor/Importer

5. **If you do NOT want your identity disclosed to the manufacturer, place an "X" in this box:** □

FORM FDA 3500 (2/13)

Submission of a report does not constitute an admission that medical personnel or the product caused or contributed to the event.

Delete Page

ADVICE ABOUT VOLUNTARY REPORTING

Detailed instructions available at: http://www.fda.gov/medwatch/report/consumer/instruct.htm

Report adverse events, product problems or product use errors with:

- Medications *(drugs or biologics)*
- Medical devices *(including in-vitro diagnostics)*
- Combination products *(medication & medical devices)*
- Human cells, tissues, and cellular and tissue-based products
- Special nutritional products *(dietary supplements, medical foods, infant formulas)*
- Cosmetics
- Food *(including beverages and ingredients added to foods)*

Report product problems - quality, performance or safety concerns such as:

- Suspected counterfeit product
- Suspected contamination
- Questionable stability
- Defective components
- Poor packaging or labeling
- Therapeutic failures (product didn't work)

-Fold Here-

Report SERIOUS adverse events. An event is serious when the patient outcome is:

- Death
- Life-threatening
- Hospitalization - initial or prolonged
- Disability or permanent damage
- Congenital anomaly/birth defect
- Required intervention to prevent permanent impairment or damage (devices)
- Other serious (important medical events)

Report even if:

- You're not certain the product caused the event
- You don't have all the details

How to report:

- Just fill in the sections that apply to your report
- Use section D for all products except medical devices
- Attach additional pages if needed
- Use a separate form for each patient
- Report either to FDA or the manufacturer *(or both)*

Other methods of reporting:

- 1-800-FDA-0178 - To FAX report
- 1-800-FDA-1088 - To report by phone
- www.fda.gov/medwatch/report.htm - To report online

If your report involves a serious adverse event with a device and it occurred in a facility outside a doctor's office, that facility may be legally required to report to FDA and/or the manufacturer. Please notify the person in that facility who would handle such reporting.

-Fold Here-

If your report involves a serious adverse event with a vaccine, call 1-800-822-7967 to report.

Confidentiality: The patient's identity is held in strict confidence by FDA and protected to the fullest extent of the law. FDA will not disclose the reporter's identity in response to a request from the public, pursuant to the Freedom of Information Act. The reporter's identity, including the identity of a self-reporter, may be shared with the manufacturer unless requested otherwise.

The information in this box applies only to requirements of the Paperwork Reduction Act of 1995

The burden time for this collection of information has been estimated to average 36 minutes per response, including the time to review instructions, search existing data sources, gather and maintain the data needed, and complete and review the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to:

Department of Health and Human Services
Food and Drug Administration
Office of Chief Information Officer
Paperwork Reduction Act (PRA) Staff
PRAStaff@fda.hhs.gov

Please DO NOT
RETURN this form
to the PRA Staff e-mail
to the left.

OMB statement:
"An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number."

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Food and Drug Administration

Please Use Address Provided Below -- Fold in Thirds, Tape and Mail

DEPARTMENT OF
HEALTH & HUMAN SERVICES

Public Health Service
Food and Drug Administration
Rockville, MD 20857

Official Business
Penalty for Private Use $300

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES
OR APO/FPO

BUSINESS REPLY MAIL
FIRST CLASS MAIL PERMIT NO. 946 ROCKVILLE MD

POSTAGE WILL BE PAID BY FOOD AND DRUG ADMINISTRATION

MEDWATCH
The FDA Safety Information and Adverse Event Reporting Program
Food and Drug Administration
5600 Fishers Lane
Rockville, MD 20852-9787

Reset Form Delete Page

U.S. Department of Health and Human Services

MEDWATCH

The FDA Safety Information and
Adverse Event Reporting Program

(CONTINUATION PAGE)

**For VOLUNTARY reporting of
adverse events and product problems**

Page 3 of 3

B.5. **Describe Event or Problem** *(continued)*

Back to Form

B.6. **Relevant Tests/Laboratory Data, Including Dates** *(continued)*

Back to Form

B.7. **Other Relevant History, Including Preexisting Medical Conditions** *(e.g., allergies, race, pregnancy, smoking and alcohol use, hepatic/renal dysfunction, etc.) (continued)*

Back to Form

F. Concomitant Medical Products and Therapy Dates *(Exclude treatment of event) (continued)*

Back to Form

8.5 Appendix E: Learning Styles

Learning Styles

A fundamental part of any educational endeavor is understanding not only the subject matter one intends to teach, but the students that are going to be taught. This understanding can arise from knowing basic facts about the students, such as their prior knowledge of the subject, but it can also be derived from developing a deeper understanding of the students themselves. Namely, an understanding of the learning styles that each student uses to best understand the material presented.

Learning style refers to the manner in which a student learns best, according to their own individual personality and mindset. All models within the field of learning styles are designed to accurately represent the ways in which students learn, but they all seek to achieve the same general goal – categorizing individuals according to how they take in information, so that they can be taught more effectively (Cassidy, 2004).

There are many possible models that can be used to describe learning styles – as detailed by educational researcher Thomas De Bello in his 1990 paper on the subject, there are almost as many definitions of models as there are theorists in the field. Examples of models include *Gregorc's Style Delineator*, *Holzman and Klein's Leveller-Sharpener Styles*, *Pavios's Verbaliser-Visualiser Cognitive Style, Letteri's Learning Types* and plenty more. Research into learning styles is rather scattered, possibly as a result of the extensive empirical investigation used to develop most of these models. Many of these models are based off of their own data, gathered in their own ways, resulting in many differences between them even though the goal of the studies remains the same. While each model presents its own advantages and disadvantages, there has yet to be a single model that could be empirically chosen as the best option (Cassidy,
2004).

For the purposes of this project, we draw on *Kolb's Experiential Learning Model*. This model was first described in 1971 (D. Kolb, 2012) and has been revised periodically since, keeping it relevant in the field (A. Kolb, 2005). This model was chosen for that reason – it has been used for a long time and documented carefully, meaning that there are plenty of resources for us to refer to when implementing it into our material design process. Additionally, many studies have been done on the Kolb model that show validity in its analysis, through hard results in various tests. The model is not quite perfect - other studies focusing on the psychometric elements of the model have raised concerns over its validity and reliability (Cassidy, 2004). Despite these concerns, the Kolb model is expected to serve our purposes well through its detailed analysis of the entire learning process.

The Kolb Learning Styles are founded on the central concept of four learning modes. These four modes make up a learning cycle, shown below in Figure A51, that the student must move through to learn new skills. The modes represent how information is acquired and then understood – the Concrete Experience mode (CE) and Abstract Conceptualization mode (AC) are both points of the cycle where information is grasped by the learner. Both modes are followed in the cycle by modes in which this information is transformed and understood by the learner, the Reflective Observation mode (RO) and the Active Experimentation mode (AE). A student's learning style is based on which of the four modes they are best suited to – the idea is that some people will excel in certain areas of the cycle more so than others and thus shaping the way they learn best and the techniques which will be most effective in teaching them (A. Y. Kolb, 2005). The cyclical process described in the Kolb model is just as important as the styles derived from it. From an educational standpoint, it can be difficult to tailor a course or materials to be exactly what each student needs, as it is likely that multiple students will be learning from the same source, and they may all excel in different areas of the learning cycle. In addition to that, different students may be in different sections of the cycle at different times, with some grasping information quickly while others struggle. The Kolb model cannot solve these problems, but it does allow for a better understanding of them. Using the Kolb model it is possible to generalize a population of students to see where in the process they might work best, given their current stage of development in the topic. This sort of work has been done in fields such as nursing, to create materials better suited to a certain population (D. Kolb, 2012). Even if a population cannot be generalized, careful consideration of the Kolb model can help determine shortcomings in a course, allowing an educator to look at what areas their students are struggling with, and potential ways that the course or materials could be changed to help facilitate the grasping or transforming of information during those problem areas of the cycle.

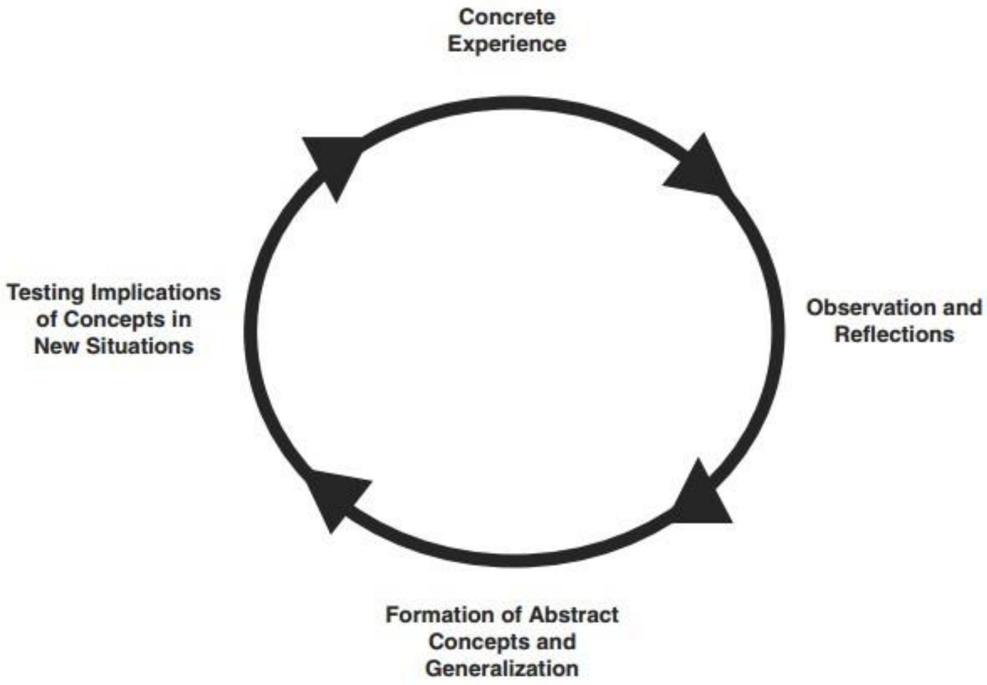

Concrete Experience
Observation and Reflections
Formation of Abstract Concepts and Generalization
Testing Implications of Concepts in New Situations

Figure A51: The learning cycle, as described by David Kolb. (A. Y. Kolb, 2005)

Kolb’s model defines four primary learning styles to consider when developing educational materials: diverging, assimilating, converging, and accommodating. Learners with diverging styles excel in the CE and RO modes and learn best when working in groups, listening to a variety of viewpoints, and getting individualized feedback. Those with assimilating styles are strong with the AC and RO modes and excel when presented with readings, lectures, and plenty of time to think about the concepts. The converging styled individual is best at AC and AE modes, making them most comfortable when experimenting with new ideas and trying practical applications for those ideas. Lastly, the individual with an accommodating style focuses on CE and AE modes and works best in teams by doing clearly outlined practical work which allows them to test different approaches (D. A. Kolb, 2005). While there are five other styles outlined by Kolb, these main four provide a clear overview of the areas most individuals will belong to as well as the educational methods which will be most effective. In theory, educational materials are most effective when they consider all aspects of the learning cycle and incorporate elements to support each style's needs. Table A5A, included below, illustrates how consideration of these styles might influence the medium of our educational materials. This ensures that all individuals will comfortably learn from what is presented.

Table A5A: Learning Styles and Proposed Mediums (D. A. Kolb, 2005)

Learning Style	Attributes of Style	Proposed Medium(s)
Diverging	Individuals work best when working in groups, being able to listen to a variety of viewpoints and getting individual feedback.	- Online tutorial - Interactive website - Interactive slideshow presentation
Assimilating	Individuals learn best from being presented readings and lectures and require plenty of time to absorb concepts.	- Printable resource packet - Interactive website - Online tutorial - Educational readings
Converging	Individuals maximize information retention by experimenting with new ideas and dealing with practical applications.	- Online tutorial - Interactive website
Accommodating	Individuals benefit from working in teams to focus on clearly outlined, practical work and test different approaches to the same problem.	- Interactive website - Online tutorial

Learning styles and the learning process can be very useful models to be familiar with when educating any group. Understanding why students may be succeeding or struggling in various areas allows for better materials to be produced, and for a better education to be delivered. The Kolb model serves this purpose well, detailing many different aspects of learners and the process through which they learn, allowing us to design our materials not just for the raw data to be presented, but for the individuals we intend to present it to.

8.6 Appendix F: Testing Data

What follows are tables describing the results of our testing. Each test consisted of 8 questions, which were graded equally. Each question was worth up to 6 points, with the complete test being worth 48 points total.

Table A6A: Pre and Post-Test Raw Data

PRE TEST										
Question:	1	2	3	4	5	6	7	8	Total	Percentage
Tester										
1	0	4	0	0	6	2	0	0	12	25
2	6	6	0	4	6	2	0	0	24	50
3	3	4	0	0	6	2	0	6	21	43.75
4	6	4	0	2	6	2	0	0	20	41.66667
5	3	2	0	0	6	2	0	0	13	27.08333
6	3	2	0	0	6	0	0	0	11	22.91667
7	3	2	0	2	6	3	0	6	22	45.83333
8	3	4	0	0	6	2	0	6	21	43.75
9	6	4	0	0	6	2	0	6	24	50
Averages:	3.666667	3.555556	0	0.888889	6	1.888889	0	2.666667	18.66667	38.88889

POST TEST										
Question:	1	2	3	4	5	6	7	8	Total	Percentage
Tester										
1	6	6	6	4	6	2	0	0	30	62.5
2	6	4	6	6	6	3	0	6	37	77.08333
3	6	4	6	4	6	4	6	6	42	87.5
4	6	4	6	6	6	4	0	6	38	79.16667
5	6	4	4	6	6	3	3	6	38	79.16667
6	6	2	0	6	6	2	0	6	28	58.33333
7	6	2	6	2	6	2	3	6	33	68.75
8	6	6	6	6	6	2	6	6	44	91.66667
9	6	4	6	4	6	2	3	6	37	77.08333
Averages:	6	4	5.111111	4.888889	6	2.666667	2.333333	5.333333	36.33333	75.69444

Table A6B: Score Improvement Raw Data

Improvement											
Question:	1	2	3	4	5	6	7	8	Total	Percentage	% improved
Tester											
1	6	2	6	4	0	0	0	0	18	37.5	250
2	0	-2	6	2	0	1	0	6	13	27.08333	154.1667
3	3	0	6	4	0	2	6	0	21	43.75	200
4	0	0	6	4	0	2	0	6	18	37.5	190
5	3	2	4	6	0	1	3	6	25	52.08333	292.3077
6	3	0	0	6	0	2	0	6	17	35.41667	254.5455
7	3	0	6	0	0	-1	3	0	11	22.91667	150
8	3	2	6	6	0	0	6	0	23	47.91667	209.5238
9	0	0	6	4	0	0	3	0	13	27.08333	154.1667
Averages:	2.33	0.44	5.11	4	0	0.77	2.33	2.66	17.66	36.81	206.08

8.7 Appendix G: Pre/Post Test and Grading Rubric

JIT Training Assessment (Pre/Post Test)

1. Please define data mining

____

____

2. How would data mining be applied to pharmacovigilance?

____

____

3. Please identify the types of data mining methods and an advantage of each type

____

____

4. What are the three prerequisite steps to effectively mine data?

____

____

5. Data mining proves causation, not correlation (T/F) ____
6. Name the tool currently used at the FDA and what methods are used by the tool

____

7. Name a tool that could be potentially used to at the FDA for data mining with one of its advantages

____

8. What office can be reached for further information on data mining?

____

Question	Grading Criteria	Possible Points	Points Earned
1	Finding patterns and/or clusters	3	
	Mentions the word “data” or “dataset” or “database”	3	
2	Mentions patterns	2	
	Mentions safety signals or “safety issues”	2	
	Mentions “adverse drug events” or “adverse drug reports”	2	
3	Identifies frequentist group	2	
	Identifies Bayesian group	2	
	One advantage specific to frequentist group	1	
	One advantage specific to Bayesian group	1	
4	De-duplication	2	
	Standardization or mentions “E2B” or “MedDRA”	2	
	Completeness Check	2	
5	False	6	
6	Empirica or Empirica Signal	2	
	MGPS (if only EBGM is written give half credit)	2	
	PRR and ROR (if only one of these is written give half credit)	2	
7	OpenVigil or QScan PostMarket or RLytics	3	
	One advantage associated with specified tool	3	
8	Office of Translational Sciences	6	
Total		**48**	

JIT Training Assessment Grading Rubric

8.8 Appendix H: Final Educational Materials

Data Mining for Pharmacovigilance

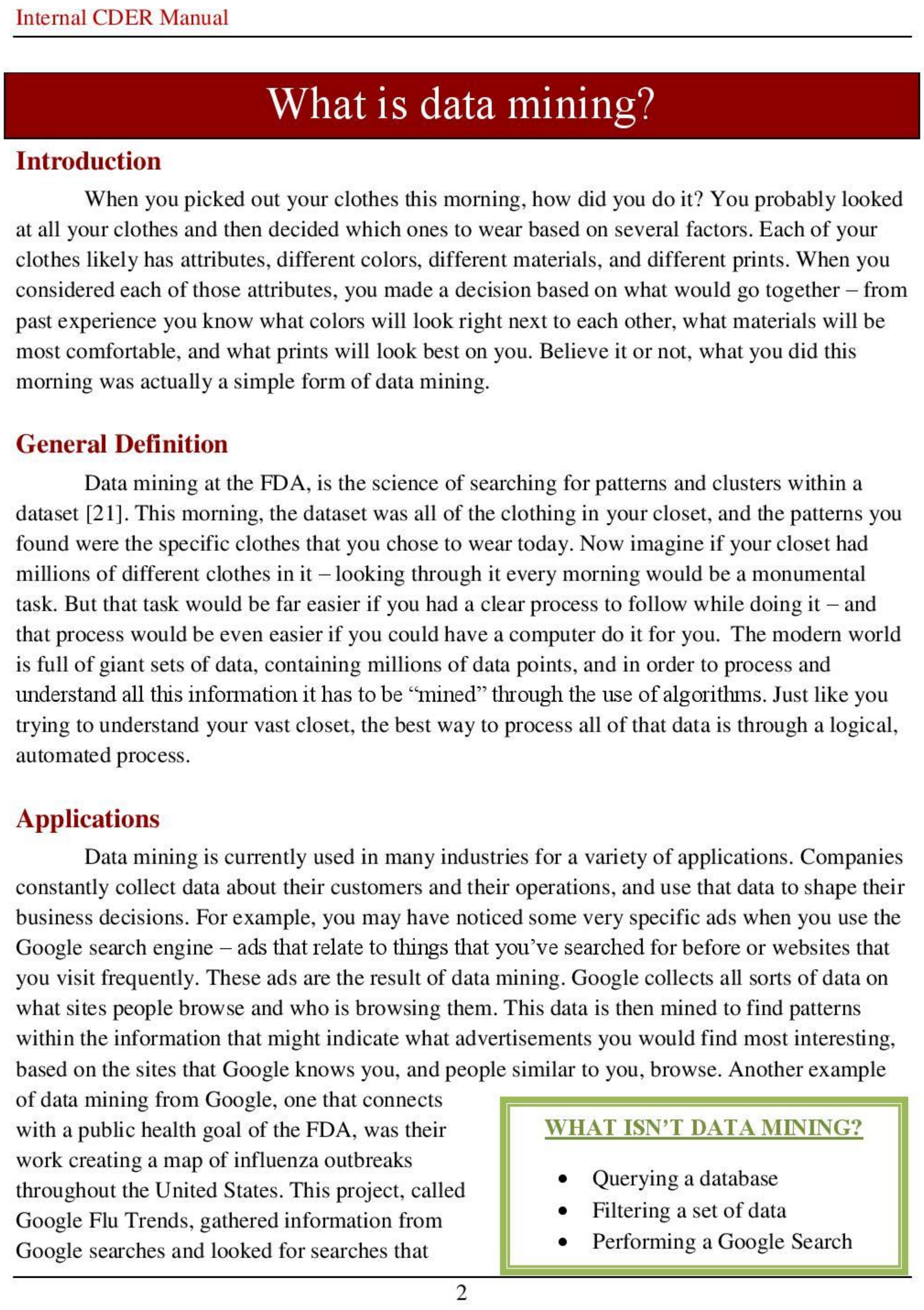
A Brief Overview

Office of Surveillance and Epidemiology, CDER, FDA

12/15/2015

This document was developed by a team of students from Worcester Polytechnic Institute working closely with the Office of Surveillance and Epidemiology. The views represented within this document are not necessarily those of the FDA.

Contents



What is data mining?

Introduction

When you picked out your clothes this morning, how did you do it? You probably looked at all your clothes and then decided which ones to wear based on several factors. Each of your clothes likely has attributes, different colors, different materials, and different prints. When you considered each of those attributes, you made a decision based on what would go together – from past experience you know what colors will look right next to each other, what materials will be most comfortable, and what prints will look best on you. Believe it or not, what you did this morning was actually a simple form of data mining.

General Definition

Data mining at the FDA, is the science of searching for patterns and clusters within a dataset [21]. This morning, the dataset was all of the clothing in your closet, and the patterns you found were the specific clothes that you chose to wear today. Now imagine if your closet had millions of different clothes in it – looking through it every morning would be a monumental task. But that task would be far easier if you had a clear process to follow while doing it – and that process would be even easier if you could have a computer do it for you. The modern world is full of giant sets of data, containing millions of data points, and in order to process and understand all this information it has to be "mined" through the use of algorithms. Just like you trying to understand your vast closet, the best way to process all of that data is through a logical, automated process.

Applications

Data mining is currently used in many industries for a variety of applications. Companies constantly collect data about their customers and their operations, and use that data to shape their business decisions. For example, you may have noticed some very specific ads when you use the Google search engine – ads that relate to things that you've searched for before or websites that you visit frequently. These ads are the result of data mining. Google collects all sorts of data on what sites people browse and who is browsing them. This data is then mined to find patterns within the information that might indicate what advertisements you would find most interesting, based on the sites that Google knows you, and people similar to you, browse. Another example of data mining from Google, one that connects with a public health goal of the FDA, was their work creating a map of influenza outbreaks throughout the United States. This project, called Google Flu Trends, gathered information from Google searches and looked for searches that

WHAT ISN'T DATA MINING?

- Querying a database
- Filtering a set of data
- Performing a Google Search

might indicate the flu – things like people searching for common influenza symptoms or remedies. By looking for these searches, and tracking the regions they came in from, Google was able to construct a live map showing the spread of the flu across the country [8]. This is a good example of how data mining can extend beyond simple advertisement, and make a significant contribution to pharmacovigilance when it is used in the right situations.

How does data mining work?

Data mining, though not universally well-defined, is a discipline heavily based upon mathematics and computer science, and the methods of performing it are focused in these academic areas. A lot of data mining is conducted through statistical analysis, where advanced mathematical ratios are used to create numerical scores that can provide quick, clear information to an observer about data sets that in most cases would have taken months to read through. The ratios used by these methods are generally designed to identify disproportionality, areas of the data that stand out from the rest. By cutting up the data and seeing what looks different from the rest, it's possible to start identifying the subtle patterns within it. As an example, say you have detailed data on the spending habits of 100,000 people, and you want to know what items are most popular for each age group. You can split the 100,000 people into different age demographics (child, teenager, young adult, etc.) and then analyze each of those demographics for various products. This analysis may bring you to a few conclusions – perhaps you will find that many more teenagers are interested in buying music than the general population, and that senior citizens purchase far fewer video games than other demographics. Performing a disproportionality analysis with data mining would yield numerical scores. The interpretation of these scores is important and would similarly yield that more teenagers are interested in buying music than the general population. Data mining scores provide a much easier option for interpreting all the sales data than attempting a manual review, saving a lot of time and effort, while still providing the knowledge needed.

Data Mining in Pharmacovigilance

In pharmacovigilance, data mining is defined as the science of looking for patterns and clusters among sets of Adverse Drug Reports (ADRs). In pharmacovigilance these patterns and clusters are evidence that can be used to identify new safety signals. The process for detecting the patterns and clusters, or safety signals, can be defined as a signal detection algorithm (SDA). A SDA primarily uses disproportionality to identify statistical associations between products and events. This disproportionality is represented as a score. A higher score indicates a higher disproportionality. The unexpectedly high reporting associations, "signal" the potential for a causal association between an adverse event and product [6]. These disproportionately reported combinations are referred to as the identified safety signals.

Where data mining is used

In pharmacovigilance, data mining is used to assist the evaluation process in several ways: prioritizing reports, analyzing drug-drug interactions, and evaluating both familiar and unfamiliar classes of drugs. Prioritizing safety reports to read is essential for a reviewer to pick out noise and locate reports that could be the source of a safety signal. Analysis of drug-drug interactions can help point out safety signals that might not be found if only performing data analysis on one specific drug. Evaluating a class of drugs, perhaps unfamiliar to a new Safety Evaluator, is useful in becoming aware of a class specific trend of Adverse Events (AEs).

CAUSALITY

It is important to note that data mining results only ever prove correlation – not causation. A calculated score might show a correlation between a drug and a disease process, but the score does not mean that the disease process is caused by the drug. The human element of the review process determines the causality through further investigation of the adverse events included in the correlated drug.

Why is data mining needed in pharmacovigilance?

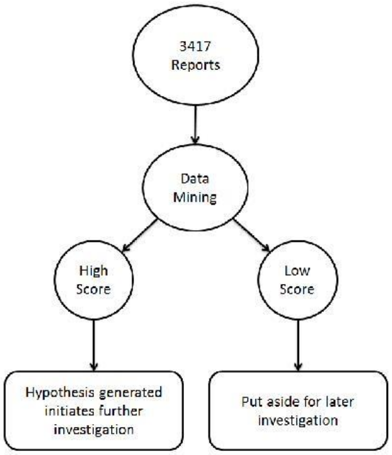

3417
Reports
Data
Mining
High
Score
Low
Score
Hypothesis generated initiates further investigation
Put aside for later investigation

Figure 1: Hypothesis Generation

Data mining is needed in pharmacovigilance because it can be used to analyze the increasing number of reports received, speed up the identification of potential safety issues, aid in hypothesis generation, and free personnel to devote more time to in depth evaluation [6]. Because the number of reports is growing exponentially, it is challenging for Safety Evaluators to view all of the reports within the mandated time constraints. Since it is challenging for all of reports are able to be read, reports that point out a potential safety signal might not be found. With data mining these unviewed reports can be analyzed and used to form a basis that aids evaluators in creating a hypothesis of where potential safety signals might be. Additionally, by easing the amount of manual review that has to be conducted, data mining can give Safety Evaluators more time to focus their efforts where they will be most useful. These benefits all contribute to giving personnel more time to dedicate to critical tasks such as investigating signals.

What does data mining provide for pharmacovigilance?

Frequentist and Bayesian data mining methods (See Methods Section) provide a measurement of disproportionality. Measurement of disproportionality is especially useful in pharmacovigilance because it can be used to measure the ratio of a specific adverse event compared to another drug or compared to all drugs. The goal of disproportionality is to highlight any event that occurs abnormally, otherwise known as a safety signal. Suppose an evaluator covering a drug, recently approved by the FDA, has received 1000 MedWatch reports, 50 of which are for rash. This means that 5% of the reports indicate that a rash occurred, but how can the reviewer know if that is important or just expected behavior? The simplest way to know is to find out how often rashes occur for all other drugs. If analyzing all the reports for other drugs shows that 5% of reports are rashes, no matter what drug is being looked at, then there is nothing disproportionate or abnormal about the new drug. This would mean the rashes are likely not of any significance. But if rashes only occur in 1% of the reports for all other drugs, then something may be wrong – the amount of rashes occurring is disproportionate to what is seen in other drugs. This indicates an area that might need to be reviewed more closely. Data mining can provide a score to show that disproportionality quickly and easily, making it far simpler for a reviewer to generate new hypotheses on what might be a safety risk.

SUMMARY

- In pharmacovigilance, data mining can be defined as the science of looking for patterns and clusters among sets of Adverse Drug Reports (ADRs) in order to detect safety signals.
- Data mining is used to assist the evaluation process in several ways: prioritizing reports, analyzing drug-drug interactions, and evaluating an unfamiliar class of drugs.
- Data mining can be used to measure disproportionality.
- High disproportionality may indicate a dangerous drug event that needs to be investigated.
- Data mining proves correlation, not causation.

Data Mining Adverse Events

General Overview

The FDA Adverse Event Reporting System (FAERS) is the FDA's post-market safety surveillance database [22]. The database contains information found in adverse event reports and medication error reports that are submitted to the FDA when a person believes they have suffered a complication from a drug they took [12]. These reports can be submitted by paper or electronically by manufacturers, patients, doctors, pharmacists, etc [7]. 4% of all submitted reports come directly to the FDA from healthcare professionals or patients. The other 96% come from manufacturers, who are required by law to submit any reports that they receive [7]. Healthcare professionals and patients will often submit their reports to the manufacturers instead of the FDA, and these reports must then be sent forward to the FDA in a timely manner [7]. As of 2015, FAERS contains over 11 million reports and over a million more are added every year [19].

Since 2006, the number of reports submitted to the FDA has been growing exponentially [19]. This trend can be seen in Figure 2 below. It is important to note that FAERS is not the only source of data the FDA utilizes but it is the most significant source of data that the FDA currently uses [14].

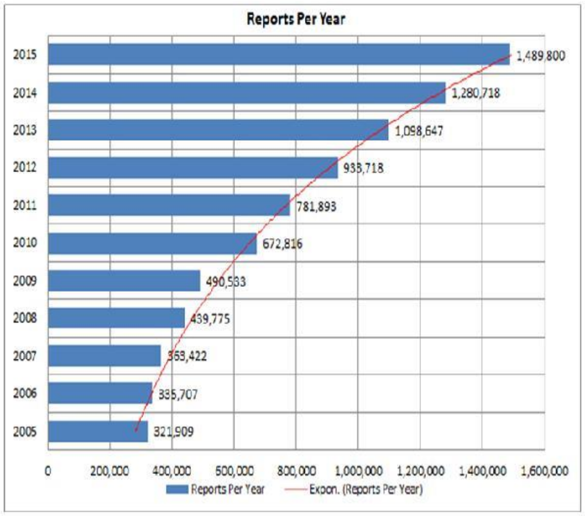


Figure 2: Number of Reports Submitted to the FDA Yearly [20]

Data Cleansing

Prior to being used in data mining, the data that is to be mined (in this case the report data within the FAERS database) must be "cleansed"; meaning that it must be organized and made clear for a computer to process [21]. This involves a few processes, such as making sure the proper fields of each report are filled in and that they are filled in with appropriate text [21]. An important aspect of data cleansing is the de-duplication of reports in the database. For instance, many times patients or healthcare professionals will report to both the FDA and the manufacturer, which results in duplicate reports when the manufacturer sends their copy to the FDA [19]. If these duplicate reports were used in data mining, the final score could be inaccurate because it would appear that multiple patients experienced the same adverse event, when really it was only one patient who sent in multiple reports [21]. In order to avoid this, FAERS data is filtered prior to data mining to identify and remove all instances of duplicate reports.

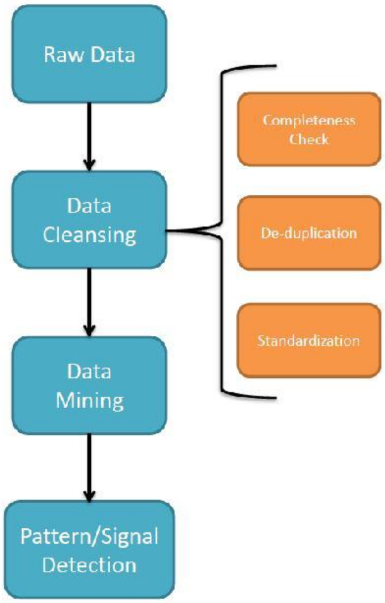


Figure 3: Diagram of Data Flow

Another aspect of data cleansing is the standardization of reports in both the input data and the textual narrative [19][21]. In order for a computer to be able to understand all of the information in reports, the reports have to be stored in a specific way – if one report has the patient's name in a field where another report has the patient's date of birth, the data will not be usable, as comparisons cannot be made. Standardization fixes this problem by outlining how reports should be structured in a very specific way, meaning that data mining programs can be built to know exactly where each type of data is [3]. Standardization also allows for the data to enter FAERS more readily – adverse event reports are filed according to international E2B guidelines, and every manufacturer that files reports is required to file them according to that standard, saving the FDA the trouble of converting reports into what the agency needs [3]. In addition to the report as a whole following a specific, clean standard, the adverse events within are described using the MedDRA dictionary [3]. The MedDRA dictionary classifies adverse events in a structured, hierarchical format that can be processed easier by computerized data mining tools [3]. Finally, the products have to standardized based on a dictionary managed by FDA.

CHECK YOUR KNOWLEDGE

1. What are three data preprocessing steps needed to perform effective data mining?

These cleansing processes are essential for effective data mining. If the data itself is inconsistent or inaccurate, the conclusions drawn from it will be unreliable [21]. As such, it's imperative to make sure that the data source you wish to analyze is organized and clear before you start trying to mine it.

SUMMARY

- FAERS is the FDA's post-market safety surveillance database.
- The number of reports submitted to the FDA is increasing exponentially.
- Data cleansing is an essential prerequisite to performing effective data mining.
- Without de-duplication, it would appear that multiple patients experienced the same event thus making report counts and data mining scores inaccurate.
- Data is standardized at the FDA according to E2B guidelines and utilizes the MedDRA and product dictionary.

Statistical Methods

There are two groups of data mining methods, the frequentist methods and the Bayesian methods [10]. These two groups consist of many individual data mining methods, which can be found in Table 1 below.

Table 1: Frequentist and Bayesian Method Groups [10]

Method Group	Individual Methods
Frequentist Methods	Proportional Reporting Ratio (PRR) Reporting Odds Ratio (ROR) Relative Reporting Ratio (RRR) Information Component (IC)
Bayesian Methods	Gamma-Poisson Shrinker (GPS) Multi-item Gamma-Poisson Shrinker (MGPS) Bayesian Confidence Propagation Neural Network (BCPNN) Empirical Bayesian Geometric Mean (EBGM)

The frequentist methods use a disproportionality measure which is “a comparison of observed counts of drug-adverse event pairs and the expected number based on other drugs in the database” to find and estimate associations in the data [4]. Bayesian methods also use the disproportionality measure in conjunction with “shrinkage.” Shrinkage is the process of adjusting a measure to account for uncertainty that is introduced into a disproportionality measure when a low amount of data is used to create it [10]. When there is only a small amount of data, rare events may seem more prevalent due to there being so little to compare them to – for example, if you only go outside twice in your life, and you are struck by lightning the second time, it would seem as if being struck by lightning is a common occurrence when in fact you were most likely incredibly unlucky. Shrinkage accounts for this by taking into account prior information about the data and using this to “shrink” the data, adjusting it towards a baseline in which no adverse event occurred [10]. A real life inspired example of the application of the two data mining groups is shown in Table 2.

Table 2: Real World Example of Frequentist and Bayesian Methods [23]

The Problem: I'm home and I can't remember where I left my cell phone. I call my cell phone and it starts to ring. Which area of my house should I search?	
Frequentist Approach	Bayesian Approach
I hear the phone ringing. My brain has the ability to figure out the direction that the sound is coming from. Therefore, when I hear the ringing, I infer the room of my house I must search.	I hear the phone ringing. My brain has the ability to figure out the direction that the sound is coming from. I also know the places in my house where I have left my cell phone in the past. Therefore, I combine my inferences using the ringing and my prior information about where my cell phone has been previously lost to choose the area of my house I must search.

CHECK YOUR KNOWLEDGE

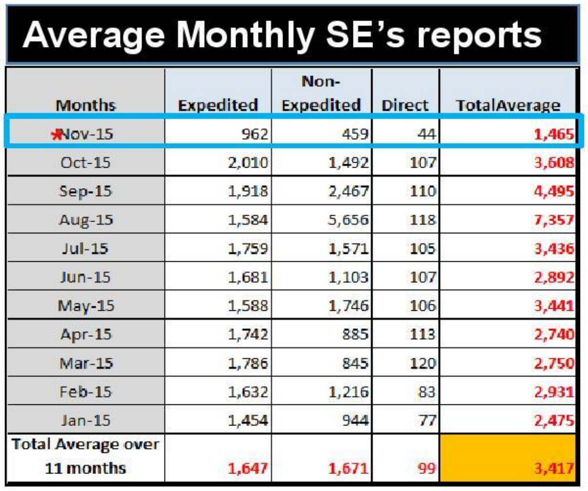
2. What is the main difference between frequentist methods and Bayesian methods?

As noted previously (see section; FAERS), the FDA's intake of reports has been exponentially increasing over the years, while the number of Safety Evaluators has remained relatively constant [11]. Tables 3 and 4 highlight the overwhelming number of reports that each Safety Evaluator must review as a result.

Table 3: Total Number of Reports [20]

Total Reports ***All Versions of Report**	**Total Reports for SE Review** ***Only latest version of the reports**
11,198,975	**8,435,279**
*Data as of 09 November 2015	

Table 4: Average Monthly Safety Evaluator Reports [20]



Average Monthly SE's reports

Months	Expedited	Non-Expedited	Direct	TotalAverage
*Nov-15	962	459	44	1,465
Oct-15	2,010	1,492	107	3,608
Sep-15	1,918	2,467	110	4,495
Aug-15	1,584	5,656	118	7,357
Jul-15	1,759	1,571	105	3,436
Jun-15	1,681	1,103	107	2,892
May-15	1,588	1,746	106	3,441
Apr-15	1,742	885	113	2,740
Mar-15	1,786	845	120	2,750
Feb-15	1,632	1,216	83	2,931
Jan-15	1,454	944	77	2,475
Total Average over 11 months	1,647	1,671	99	3,417

**snapshot of 8 DPV Safety Evaluators reports data as of 13th Nov 2015 and also partial data presented for the month of Nov 2015*

Since the FDA is becoming inundated with data (reports) relative to the resources (Safety Evaluators), research suggests focusing more on specificity than sensitivity in data mining [11]. In the context of mining adverse event signals, sensitivity is defined as "the ability of a surveillance or reporting system to detect true health events, i.e. the ratio of the total number of health events detected by the system to the total number of true health events as determined by an independent and more complete means of ascertainment" [24]. Specificity is defined as "a measure of how infrequently a system detects false positive health events, i.e. the number of individuals identified by the system as not being diseased divided by the total number of all persons who do not have the disease" [24]. The following Table 6 represents a comparison between frequentist and Bayesian Methods, summarizing the general trends of each.

Table 5: Comparison of Frequentist and Bayesian Methods [2][5][9][10][11][13]

	Frequentist Methods	**Bayesian Methods**
Tend to highlight a greater number of *DEAs	X	
Tend to highlight a greater variety of DEAs	X	
Tend to highlight DEAs earlier	X	
More computationally intensive		X
More sensitive to low-frequency of reports	X	
More intuitive computations	X	
May result in loss of credible signals		X
Lower impact of random fluctuations of relative reporting ratio ("shrinkage")		X
Produce more false positives (higher sensitivity)	X	
Produce more false negatives (higher specificity)		X

**DEA stands for drug-event association*

In general, the frequentist group highlights a greater number and variety of drug-event associations (DEAs) than the Bayesian group and tends to highlight these DEAs earlier as well [11]. However, it has been observed that for low-frequency reporting, the frequentist methods are more prone to generating false positives [11]. Bayesian methods compensate for this issue by adjusting the disproportionality measure to account for low reporting, but at the same time it loses sensitivity and possibly delays the detection of novel adverse events [5][13]. Each of the groups have different aspects that make them more suitable to certain signal detections. It is tempting for analysts to focus solely on reducing the numbers of false-positive and false-negative signals, but a balance of sensitivity and specificity is crucial for optimal signal detection [2].

DID YOU KNOW?

Research has shown that both frequentist and Bayesian data mining method groups produce similar results for higher-frequency drug-event combinations [13].

CHECK YOUR KNOWLEDGE

3. Consider a very small number of reports. Which data mining method group would be more sensitive and identify more false positives?

SUMMARY

- There are two groups of data mining methods: the frequentist and the Bayesian.
- There are pros and cons to both the frequentist and Bayesian methods and it's important to be able to identify when to use a specific data mining method group for each scenario.
- A balance between sensitivity and specificity is crucial for optimal signal detection.

Tools

Data Mining Tools

A data mining tool is defined as a software suite that utilizes data mining methods to allow a user to perform data mining quickly and easily. Data mining tools offer the benefits of a data mining method without requiring the user to manually input data and perform the calculations. Effective data mining is possible when less time is spent manually performing calculations and risking calculation errors.

Tools Currently Used by the FDA

Currently, the Center for Drug Evaluation and Research (CDER) uses a data mining tool called Empirica® Signal. Empirica® Signal is a tool that performs disproportionality analysis on adverse event data from the FAERS database. The primary method used by this tool is Multi-item Gamma-Poisson Shrinker (MGPS) which calculates an Empirical Bayesian Geometric Mean (EBGM) score. The tool then allows users to compare EBGM scores between drugs using several visual formats, and allows for the narrowing of the data set for focused analyses. Empirica® Signal also offers the capability to perform Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR) calculations on the database [17].

Methods used by Empirica® Signal

Empirica® Signal can use several methods to mine FAERS data.

Multi-item Gamma Poisson Shrinker (MGPS)
Proportional Reporting Ratio (PRR)
Reporting Odds Ratio (ROR)

Currently, the Empirica® Signal implementation at the FDA has some severe flaws that reduce the effectiveness of data mining. The first flaw is that the drug dictionary mapping is proprietary. Empirica® Signal uses an Oracle defined drug dictionary instead of the one established by the FDA. Using a proprietary dictionary slows down the analysis process because time has to be spent finding the corresponding FDA drug terminology. Another flaw is that the MedDRA coding version often lags behind the most recently released version. The downsides of outdated MedDRA coding are similar to the proprietary dictionary because of having to interpret the updated MedDRA coding from the results. It is important to assess the limitations of a data mining tool because they can cause inconvenience when using software.

Other Pharmacovigilance Data Mining Tools

The following are three examples of pharmacovigilance data mining tools that might be of interest to the FDA in the future. Below each description of the tool is a short list of its strengths and weaknesses. These examples are tools that are discussed in publications and used by pharmaceutical companies in analyzing data exposed from public FAERS data or other adverse event databases.

OpenVigil [16]

OpenVigil is a software package for analyzing pharmacovigilance data, with features including highly configurable search criteria filters and output filters. The analyses include disproportionality analysis, such as Proportional Reporting Ratio (PRR) calculations. Results can be viewed, sorted and filtered in the web browser or saved for further analyses in statistical software packages (such as SAS).

Strengths:

- Detection of drug-drug-interactions
- Adverse Event-profiles of drugs of same drug class

Weaknesses:

- Open source project (no guarantee of support)

Bayesian Confidence Propagation Neural Network [15]

Bayesian Confidence Propagation Neural Network (BCPNN) was developed by Orre and associates in cooperation with the World Health Organization. The BCPNN method is used as the measure of disproportionality. This tool includes a module for interactive visualization of a query which produces automatic presentations of results in graphical and table forms. No screenshot was provided of this software.

Strengths:

- Interactive query visualizations

Weaknesses:

- Limited to BCPNN for disproportionality analysis

RLytics [18]

RLytics is a software suite for the analysis of pharmacovigilance data developed by Real Life Sciences. The key features of this tool include the automated extraction of MedDRA coding from narratives in adverse event reports and the re-structuring of narratives for improved readability. There are no explicit mentions of disproportionality analysis, which means that it may not be an available option in this suite.

Strengths:

- Natural language processing capabilities
- Automated adverse event detection

Weaknesses:

- No explicit mention of disproportionality analysis

SUMMARY

- Data mining tools offer benefits of powerful methods without extensive manual data input.
- FDA currently uses Empirica® Signal for data mining analysis.
- Possible alternatives on the market include OpenVigil, BCPNN, and RLytics

Further Reading

What does the future hold?

Data science is a very complex, technical field, and has only existed for a relatively short time. As such, there are many exciting things in the field's future, and also many dissonant things in its present. The definition of data mining provided in this paper is one of many – defining data mining as the practice of analyzing datasets for patterns and clusters is the preferred definition within the Office of Surveillance and Epidemiology, but in further reading you may come across slightly different interpretations of what the field is. Through our research we have found that our definition, though very general, usually holds true – but be aware that other sources may incorporate different topics under the header of data mining or phrase their definitions slightly differently. In time, data science will likely become more focused, but as of right now the field is too young and too fast, with the potential of data mining techniques expanding rapidly.

The data mining techniques we describe here are primarily statistical, which is presently the most commonly used approach to mining in pharmacovigilance. But statistical approaches are not the only approach that can be used, and new techniques are being developed regularly. As mentioned previously, natural language processing (NLP) is an up and coming approach to processing narratives for pharmacovigilance information, and has potential to be incredibly useful for increasing the efficiency of report review. Unfortunately, NLP is not matured enough to be capable of reliably extracting accurate information from narratives. But the possibilities for data mining exceed even sophisticated NLP – for example, image processing could prove useful in many circumstances. It's currently possible for patients to attach images to their adverse event reports, but there is no way for those images to be understood through data mining. A very sophisticated data mining program might be able to look at an image in much the same way a doctor might, and be able to identify symptoms and possible causes using visual information. This sort of technique is far beyond present capabilities – but in the future it may be feasible.

Where to learn more

If you're interested in learning more about data mining, at both the FDA and beyond, these sources may be of interest:

Office of Translational Services – CDER Data Mining Group [Internal FDA Source]

http://inside.fda.gov:9003/CDER/OfficeofTranslationalSciences/CDERDataMiningGroup/ucm357628.htm

This is the homepage of the CDER Data mining group, which offers more information on data mining at the FDA, as well as specifics on the use of Empirica® Signal, and contact information for some of the data mining experts within the organization.

The Role of Data Mining in Pharmacovigilance

http://www.tandfonline.com/doi/pdf/10.1517/14740338.4.5.929

This paper outlines the purpose of data mining in pharmacovigilance and gives detailed mathematical explanations of the processes.

Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis

http://onlinelibrary.wiley.com/doi/10.1038/clpt.2012.50/pdf

This paper discusses new data mining methods and their use in pharmacovigilance, giving detailed examples and explanations of the process.

Acronyms

ADR – Adverse Drug Reports
AE – Adverse Event
BCPNN – Bayesian Confidence Propagation Neural Network
CDER – Center for Drug Evaluation and Research
DEA – Drug-Event Association
EB05 – Empirical Bayes (lower 95% confidence interval limit of EBGM)
EBGM – Empirical Bayesian Geometric Mean
FAERS – FDA Adverse Event Reporting System
GPS – Gamma-Poisson Shrinker
IC - Information Component
MGPS – Multi-item Gamma Poisson Shrinker
NLP – Natural Language Processing
PPV – Positive Predictive Value
PRR - Proportional Reporting Ratio
ROR - Reporting Odds Ratio
RRR - Relative Reporting Ratio
SDA – Signal Detection Algorithm
SMQ – Standard MedDRA queries

Glossary

Data Mining – the science of searching for patterns and clusters within a dataset

Direct Reports – Voluntary reports that are submitted directly to the FDA, usually by a patient or consumer

Disproportionality Measure – a comparison of observed counts of drug-adverse event pairs and the expected number based on other drugs in the database

FAERS – the FDA's post-market safety surveillance database reports in a structured, hierarchical format that can be processed by data mining

Indirect Reports – Reports submitted to the FDA by a manufacturer or healthcare professional

Pharmacovigilance - the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems [24]

Positive Predictive Value – the proportion of signals that turn out to be real safety issues

Sensitivity – The ability of a surveillance or reporting system to detect true health events, i.e. the ratio of the total number of health events detected by the system to the total number of true health events as determined by an independent and more complete means of ascertainment [24]

Shrinkage – the process of adjusting a measure to account for uncertainty that is introduced into a disproportionality measure when a low amount of data is used to create it

Specificity – A measure of how infrequently a system detects false positive health events, i.e. the number of individuals identified by the system as not being diseased divided by the total number of all persons who do not have the disease [24]

Tool – a software suite that utilizes data mining methods to allow a user to perform data mining quickly and easily

Check Your Knowledge Answers

1. Completeness check, De-duplication, Standardization
2. Bayesian methods use "shrinkage" to account for low report counts.
3. Frequentist methods

Works Cited

[1] Administration, U. S. F. a. D. (2013). Reports Received and Reports Entered into FAERS by Year. Retrieved from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070434.htm

[2] Balakin, K. V., & Ekins, S. (2009). *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery*: Wiley.

[3] Brolund, G., CDER, & Group, I. M. E. W. (2001). *Electronic Transmission of Individual Case Safety Reports Message Specification.* Paper presented at the International Conference on Harmonisation of Technical Requirements of Pharmaceuticals for Human Use, Washington, D.C.

[4] Candore, G., Juhlin, K., Manlik, K., Thakrar, B., Quarcoo, N., Seabroke, S., . . . Slattery, J. (2015). Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug safety, 38*, 577-587.

[5] Deshpande, G., Gogolak, V., & Smith, S. W. (2010). Data Mining in Drug Safety. *Pharmaceutical Medicine - New Zealand, 24*(1), 1010-1021.

[6] Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R. A., Baker, J. D., Ball, R., . . . Kass-Hout, T. (2015). Use of Data Mining at the Food and Drug Administration. Washington, D.C.: Journal of the American Medical Informatics Association Advance Access.

[7] FDA. (2015a). *FDA Adverse Event Reporting System System Data Management Program (FAERS - DMP).* Retrieved from Landover, Maryland:

[8] Google. (2008). Google Flu Trends. Retrieved from https://www.google.org/flutrends/about/

[9] Gravel, C. (Producer). An Introduction to the Statistical Methods for Signal Detection in Pharmacovigilance. Retrieved from http://www.mclaughlincentre.ca/events/Parmaco2/Chris%20Gravel_UofO%20workshop%20May%2028-29%202009.pdf

[10] Harpaz, R., DuMouchel, W., Shah, N., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics, 91*(6), 37.

[11] Hauben, M., Madigan, D., Gerrits, C. M., Walsh, L., & Van Puijenbroek, E. P. (2005). The Role of Data Mining in Pharmacovigilance. *Expert Opinion on Drug Safety, 4*(5), 929-948.

[12] Holloway, C. (2013). *Standard Database Descriptions: AERS*. Retrieved from http://inside.fda.gov:9003/downloads/cder/officeofsurveillanceandepidemiology/ucm353215.doc

[13] Johnson, K., Guo, C., Gosink, M., Wang, V., & Hauben, M. (2012). Multinomial Modeling and An Evaluation of Common Data-Mining Algorithms for Identifying Signals of Disproportionate Reporting in Pharmacovigilance Databases. *Bioinformatics, 28*(23).

[14] Marni Hall, personal communication, November 23, 2015.

[15] Miguel de Almeida Vieira Lima, L., Goncalo Sales Craveiro Nunes, N., Goncalo Pires da Silva Dias, P., & Jorge Batel Marques, F. (2012). Implemented Data Mining and Signal Management Systems on Spontaneous Reporting Systems' Databases and their Availability to the Scientific Community-A Systematic Review. *Current drug safety, 7*(2), 170-175.

[16] *OpenVigil*. Computer software. *Open Vigil*. Dr. Ruwen Böhm, 3 Mar. 2011. Web. <http://openvigil.sourceforge.net/>.

[17] Oracle. *Empirica Signal*. Program documentation. *Oracle*. Oracle, n.d. Web. <http://www.oracle.com/us/industries/life-sciences/empirica-signal-ds-396068.pdf>.

[18] Real Life Sciences. *RLytics Platform*. N.p., n.d. Internal.

[19] Sanjay Sahoo, personal communication, November 13, 2015.

[20] Sahoo, S. (2015). *Support for DPV/DMEPA FAERS/FBIS Eco-System: Status Report (December 2014 to Today)*. Retrieved from Washington, D.C

[21] Suranjan De, personal communication, October 29, 2015.

[22] Tonning, J. (2015). *Anonymized Data: Considerations for Data Mining*. Retrieved from Washington, D.C.: http://sharepoint.fda.gov/orgs/OIM/FDAScientificComputingBoard/_layouts/PowerPoint.aspx?PowerPointView=ReadingView&PresentationId=/orgs/OIM/FDAScientificComputingBoard/SC%20Day%202015/Breakout%20Session/Presentations/0908_2047W_Tonning.ppt&DefaultItemOpen=1&Source=http%3A%2F%2Fsharepoint%2Efda%2Egov%2Fsearch%2Fes%2FPages%2Fresults%2Easpx%3Fk%3Ddata%2520mining%26s%3DFDA%2520All%26start1%3D41

[23] Unknown. "Bayesian and Frequentist Reasoning in Plain English." *Cross Validated.* Stack Exchange Inc., 19 July 2010. Web. 30 Nov. 2015. http://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english

[24] World Health Organization (2015). Retrieved from http://www.who.int/en/