PhUSE US Connect 2018

Paper PP12

Big Data Paradox: Development of a Unique and Repeatable Analytical Process while Implementing a Robust Data Science Solution

Chris Hurley, MMS, Canton, MI, USA
Aditya Gadiko, MMS Raleigh, NC, USA
Srinivas Mukkala, MMS, Canton, MI, USA
Kulin Mehta, MMS, Austin, TX, USA

ABSTRACT

Who could have imagined that a few simple ASCII datasets would pose such a huge "big data" challenge? Therein lies the paradox, while the data is simple and there are relatively few variables to analyze, the volume is huge and presents a great processing challenge. This poster demonstrates a repeatable "big data" solution that utilizes FAERS data, provided quarterly by the FDA, to provide the answers to our post-marketing safety surveillance questions. The process, techniques, and tools used to handle and analyze such massive amounts of data, in a clinical setting, are demonstrated here, creatively enabling us to cut the Gordian knot and overcome a seemingly impossible challenge.

INTRODUCTION

Enormous amounts of data are being generated and made available by the second. However, data has value only when it is utilized appropriately. It is very difficult to handle large datasets using traditional methods and techniques, as the data may be big in volume, variety, veracity, and velocity. FAERS is an excellent example of static, structured and very high-volume data. The nature of the data offered simple and well-controlled inputs which were ideal for evaluating a focused proof of concept (PoC) iteration within a very robust data science methodology. The methodology was applied as FAERS data were analyzed and visualized with the goal of developing a post-marketing signal detection solution - validating the proof of concept at the same time.

What is FAERS?

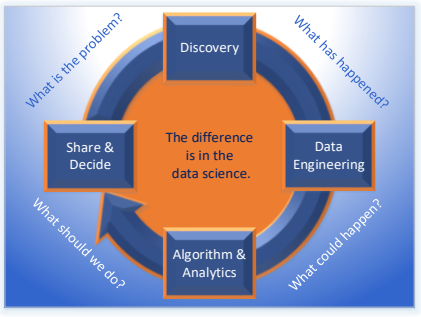
The FDA Adverse Event Reporting System (FAERS) is a database that contains spontaneous adverse event reports, medication error reports, and product quality complaints resulting in adverse events that were submitted to FDA. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products.*

PROCESS

The PoC employed a very robust **Data Science process.** The following four steps are at the core of this process:

1. **Discovery** – Determined where to leverage data for making better business decisions, defined the goals and objectives of the analytical approach and specified the deliverables to be shared with the stakeholders. The interaction with stakeholders provided an understanding of how to leverage the FAERS data and create value through signal detection analytics.

    - **Consulted with Internal PV stakeholders**: The discussion of analyses deliverables included a desire for a user-friendly interface and real-time interaction with the data. The analysis objectives and user needs were defined in a quantifiable manner.
    - **Data Sources**: The data source was confirmed, specifications for the data were obtained, an understanding of how to merge and organize the data was developed, and then work proceeded with the data.



2. **Data Engineering** – The FAERS data was not perfectly clean but it was useable and the data was prepared for analysis. These were the steps that were followed during this "engineering" phase of the PoC.

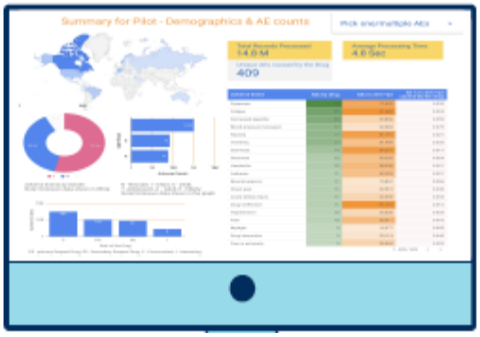
    - **Tools/Components for Stack**: Utilized a stack which could handle unstructured, semi-structured, and structured data. There was a clear path through the stack for the FAERS data processing. This option of using stack was taken to evaluate the processing of this large volume of data for the FAERS analyses.
    - **Loaded Data:** The data was loaded into the system for analysis and processed based on the category and type of the data. For instance, a combination of structured text and SAS data files.
    - **Frequency:** Static data was downloaded and combined to be made ready for the analyses. FAERS data is available quarterly from the FDA site. The ability to process FAERS data quarterly was considered and proven with the PoC.
    - **Data Cleaning:** The FAERS data is not perfectly clean. After data was loaded into the system, it needed to be cleaned. That is, a missing data search occurred and rules were created to populate the missing data and correct errors. It was observed that approximately 70 percent of the calendar time for this project was consumed during this stage of the methodology.

3. **Algorithms and Analytics** – After clean data was available and understood, a plan to analyze the data was developed. This was defined by the goals and objectives provided by the stakeholders in the discovery phase. The following artifacts of the static and dynamic analyses were modeled and developed to answer the business problems.

    - **Multiple Data Sources**: The FAERS data was joined with MedDRA and WHO drug dictionaries to standardize it and make it more useful for signal detection.
    - **Analysis**: The overarching business goal was to determine safety signals within and across quarters of FAERS data. This was done by creating a series of analytics that were easily presented into a series of dashboards.
    - **Developed Dashboards**: The dashboards were developed which included functionality for the pharmacovigilance (PV) stakeholders to interact and filter data in real time according to their needs. The visual nature of the heat maps, and other dynamic and static presentations (such as summary tables), gave the ability to easily find signals in the data and then compare with other quarters or similar drugs in the class.

4. **Share and Decide** – The analyses were completed and the results were shared with the stakeholders then feedback was given. This feedback enabled another iteration of the methodology where it was determined that further deliverables were needed to make the analyses easier on the stakeholder.

    - **PV Stakeholders Feedback**: Stakeholders provided feedback on the dashboards. Enhancements were implemented and changed the way the analysis ready data was handled to include features such as drill down (ability to view the lowest possible row within a visual), alter the way data is represented, provide the ability to export data when certain filtering is applied, and apply real-time user customization of the dashboards.



CONCLUSION

The data science methodology was fully exercised with the FAERS PoC. This is a robust process with many steps under each of the four major areas: Discovery, Data Engineering, Algorithm & Analytics, and Share & Decide. Typical flow goes in a circular nature. Analyses and answers can be shared and decisions for more follow-up can quickly be applied through the process to meet the new requirements and answer the new questions.

The proof of concept successfully demonstrated the methodology could:

- Produce a more efficient solution to meet the business need, in a relatively short time
- Quickly and efficiently ingest and prepare analyses data from very large raw data sets
- Apply an interactive, fast, and visually-rich web-based analytics solution
- Provide the stakeholders with meaningful visual analytics and trends of complex data prospectively to allow timely risk management decisions
- Allow us to elicit regular and in near real-time feedback
- Provide a consistent and repeatable process which can be applied to any “big data” project

REFERENCES

*More information on FAERS -
https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/

CONTACT INFORMATION

Comments and questions are valued and encouraged. Contact the authors at:

Author Name: Chris Hurley
Company: MMS Holdings Inc.
Address: 6880 Commerce Blvd
City / Postcode: Canton, MI 48187
Work Phone: +1 (734) 245-0469
Email: churley@mmsholdings.com
Web: www.mmsholdings.com