American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-8, Issue-11, pp-09-19
www.ajer.org
Research Paper Open Access

Improving Model Predictive Power by Integrating Structured and Unstructured Data: A Case Study of the Vaccine Adverse Event Reporting System (VAERS)

Taiwo Ajani[1], George Habek[2]
[1]*Computer Information Systems and Technology, Ferrum College, VA, USA*
[2]*Poole College of Management, North Carolina State University, NC, USA*
Corresponding Author: Taiwo Ajani

ABSTRACT : *We mined data from the Vaccine Adverse Event Reporting System (VAERS) and built two sets of algorithms from 2017 records. This resulted in a total of four models (structured data with- and without- texts for each of decision tree and logistic regression respectively). We scored the models using VAERS data from 2018 records to show the extent of separation with respect to predictive power. Results showed that models that integrated unstructured data (texts) had a significantly improved predictive power with cumulative lift increasing almost by a whole point (2.6) compared to same model sets that were applied to structured data (1.8) at the 20th percentile. Other indices such as captured responses and misclassification rate for the validation dataset indicated significantly noticeable superiority of integrated data over structured data (only) in the sets of models compared. Results suggest the value adding potentials of unstructured data is important to changing current approach in healthcare management. Integrated data proved to be more valuable in terms of predictive modeling, and providing actionable business results to aid in optimizing patients' healthcare.*
KEYWORDS:*Structured and Unstructured data, Data mining, Predictive modeling, Health analytics, Medical Informatics*

Date of Submission: 27-10-2019 Date of acceptance: 15-11-2019

I. INTRODUCTION

Electronic health data have become very valuable as mineable information sources for analytics and informatics. According to Abhyankar et al. (2014), one of their benefits is that they are repositories of historical data that would otherwise be time and cost prohibitive for individual researchers to collect. Indeed, the adoption of big data technologies by the healthcare and medical domain continue to open up benefits and potentials for improving health quality and outcomes. However, what is perceived as "value" or "benefit" depends on the organization's strategic goals for adopting big data and associated techniques. Data types, formats, awareness on available data resources and the analytic literacy of the organization, are pivotal to deriving such values and benefits. For instance, the academic and practitioner-oriented literatures highly characterize big data opportunities for organizations (Clarke, 2016), yet, high hopes neither guarantee nor translate to gaining of actual value (Gunther et al., 2017). We argue that in order to advance theories in big data value realization, there is a need for studies that demonstrate practical business value using structured and unstructured data especially in the medical and healthcare domain.

Gandomi and Haider (2015) highlighted the fact that predictive analytics often deals mostly with structured data to the exclusion of other forms of analytics applied to unstructured data that constitutes 95% of big data. From our industry and academic experience, the situation has not drastically changed from 2015. Random searches in big data and predictive analytics are most likely to yield literature focusing on structured data (Saini & Kohli, 2018; Alkhatib, Talaei-Khoel & Ghapanchi, 2015; Raghupathi & Raghupathi, 2014; Koh & Tan, 2005). Several challenges persist despite the promises of data mining in healthcare. According to Ajani and Habek (2018), data often exist in dissimilar technology platforms. For the untrained, it is often complicated to infer knowledge from complex heterogeneous patient data sources. For the expert, it is often difficult to leverage the patient/data correlations in longitudinal records and explain concepts to other domain stakeholders. These

factors may further exacerbate the already computationally difficult task of analyzing- and deriving value from-clinical data (Ajani & Habek, 2018).

Although analytical techniques continue to improve, several business and industry domains are still far from experiencing the values and benefits of integrated data. The utilization of integrated data is particularly essential to saving costs, improving healthcare quality and patients' outcomes (Scheurwegs, 2016; Abhyankar et al., 2014). In addition to structured data, collected clinical data often include unstructured (textual) responses from patients. Narrative notes may contain wealth of information with details and nuances that could provide clinical contexts (Abhyankar et al., 2014). Hence, the use of structured and unstructured data has become an active discipline of research in Medical Informatics. Several results have questioned the accuracy of structured administrative data such as the International Classification of Diseases, 9th edition, Clinical Modification (ICD-9) for identifying specific patient population with indications that ICD-9 codes did not perform as well in identifying complicated conditions due to issues such as underreporting or lack of granular codes (Abhyankar et al., 2014; Kern et al., 2006; Zhan et al., 2009; Floyd et al., 2012). In a study conducted to identify a cohort of ICU patients that received dialysis, Abhyankar and his colleagues (2014) found that combining data from clinical notes with multiple structured sources identified a larger set of patients who potentially underwent dialysis compared to any individual source. According to the authors, this approach increased confidence in the available data. Scheurwegs (2016) investigated whether integration of heterogeneous data sources improves prediction strength relative to using data types in isolation and concluded that models using multiple electronic health record data sources systematically outperform models that used data sources in isolation.

The scenario: The Vaccine Adverse Event Reporting System (VAERS) was created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) to receive reports about adverse events that may be associated with vaccines. Doctors and other vaccine providers are encouraged to report adverse events. Our challenge is to establish drivers for a patient's condition being diagnosed as "serious" based on a set of attributes using structured data on the one hand; and integrated (structured and unstructured) data on the other. The objective is to compare the predictive power of structured versus integrated data using regression (R) and decision tree (DT) logarithmic models. In other words, there will be a total of four models using R and DT for each of structured and integrated data respectively.

II. RESEARCH METHODOLOGY

This research centered upon the following research questions: (1) Are there any differences in the value and predictive power of structured vs integrated data for predictive analytics? (2) What are the drivers of a patient being diagnosed as serious? (3) Which one of the four models is best suitable for capturing adverse events?

To answer these questions, we mined 2017 records from publicly available VAER database as explained below. The first phase involved data preparation, data merge and initial data exploration in SAS studio (Ajani & Habek, 2018). Merged data were imported into SAS Enterprise Miner to complete the analytic/predictive modeling phase of the study. Available Data included historical information and demographic information such as age and gender at the patient level (this information came from the ER notes); table containing symptom and vaccination information along with a field indicating if the event was serious or not (this information came from the hospital records). The tables were merged in SAS studio as described by Ajani and Habek (2018) in their study titled "A Machine Learning Approach to Optimizing Diabetes Healthcare Management Using SAS Analytic Suite". Figure 1 depicts a schematic representation of the available data resources while Figure 2 indicates the first ten (10) observations from SAS data output after merging the data in SAS studio.

2.1The Data

We utilized two sets of VAERS files from 2017 (historical data) to build the model, and 2018 to score the model (see the table below). Each of these data files consisted of three parts: (1) VAERS data that has most of the structured attributes; (2) VAERS data which contains the unstructured symptom text field; (3) VAERS data which contains the structured vaccine information.Fig. 1 shows different types of data that were available for this research. Diamond shape represents numerical information; Rectangular shape represents categorical attributes deemed to be acceptable for analysis.

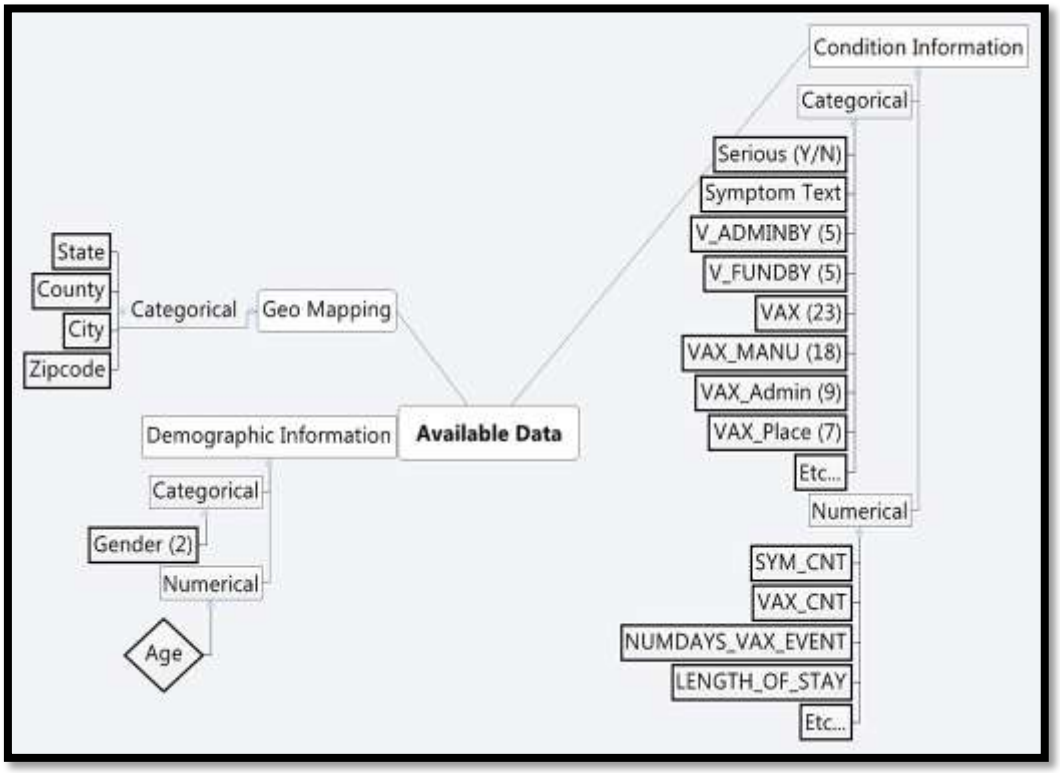

Fig.1. Available Data

ID	Patient Age	Gender	SYMPTOM_TEXT	SYMP CT	IMMUN CNT	VAX CNT	CITY	STATE	ZIP	VAX ADMIN	VAX PLACE	VAX MANU	VAX	Length Of Stay	NUM_Days VAX_Event	Serious Condition
1	4	M	Extreme swelling on vaccinated leg. Red and hot to touch. Pain in walking. Irritability. Gro	10	1	1	North Reading	MA	01889	Intramuscular	Left Leg	UNKNOWN MANUFACTURER	INFLUENZA	0	0	0
2	0.2	F	Started having seizures an hour after vaccines. Had EEG'S MRI and long term monitoring a	2	1	1	Andover	MA	01899	Oral	Unknown	MERCK & CO. INC.	ROTAVIRUS	4	0	1
3	28	F	Chronic urticaria, intense itching of the skin on scalp, ears, face, back, arms and legs. Tak	2	6	2	Lynn	MA	01901	Intramuscular	Left Arm	NOVARTIS VACCINES AND DIAGNOST	INFLUENZA	0	10	0
4	43	M	Itchy eyes at first then numb upper lip and swollen eyes and cheeks; tried next day after ov	1	3	1	Lynn	MA	01902	Intramuscular	Left Arm	SANOFI PASTEUR	INFLUENZA	0	0	0
5	59	M	Rapid heart rate when moving even just a little bit. Slowly subsides upon rest.	3	3	1	Lynn	MA	01904	N/A	Unknown	UNKNOWN MANUFACTURER	INFLUENZA	0	2	0
6	75	F	Redness after 2 days, clear after 3 days.	8	2	2	Lynn	MA	01905	Intramuscular	Left Arm	MERCK & CO. INC.	PNEUMONIA	0	1	0
7	3	F	Mild papular rash in left arm appeared right after injection site.	5	1	1	Saugus	MA	01906	Intramuscular	Left Arm	SANOFI PASTEUR	INFLUENZA	0	1	0
8	58	F	Persistent sharp pain radiating from administration site to shoulder.	4	1	1	Swampscott	MA	01907	Intramuscular	Right Arm	NOVARTIS VACCINES AND DIAGNOST	INFLUENZA	0	0	0
9	72	F	Patient was having pain from shoulder to elbow - sore and painful - started 3 1/2 weeks afte	4	1	1	Nahant	MA	01908	Intramuscular	Right Arm	SANOFI PASTEUR	INFLUENZA	0	28	0
10	8	F	This spontaneous report as received from a nurse refers to an 8 year old female patient wit	4	7	5	Amesbury	MA	01913	Intramuscular	Unknown	MERCK & CO. INC.	PNEUMONIA	0	1	0

Fig. 2. The first ten (10) observations of final merged dataset at the patient level containing the following information: 1. Demographic information such as age and gender; 2. Vaccine information; 3. Unstructured ER notes; 4. Geographical information.

2.2SAS Studio Analytical Data Preparation Process

The analytical data preparation process can be very time consuming and is essential in ensuring the analytics conducted is not only accurate but answers the specific business question desired. The purpose is to establish drivers of a patient going to the ER that would have a serious condition and therefore likely to be admitted to the hospital. Therefore, we ask "What is driving a patient entering the ER to have a serious condition?" In addition, "What is the probability of that event happening?"We merged these three files into one master file upon which analytics was performed. Basically, each file has data for each record or observation. Level refers to the hierarchy of the data. It is important to note that these files may not be at the same level (i.e. patient, claim, etc.). In our case, all three files are already at the patient level, therefore no roll-up is necessary. If we were to conduct a roll-up exercise, an aggregation or summarization exercise would need to occur. Basically, the hierarchy level would need to be established. A key tool very effective for analytical data preparation is SAS Studio.

2.3 SAS Studio Data Preparation Flow

The output is divided into 5 major steps. In the first step, we created a folder in SAS studio and the Library (This is basically a Libname statement). Hence, the three (3) csv data files (labelled data, text and vaccine) were imported to create respective SAS datasets. (Note that there are several formats that can be

accessed and imported in addition to access engines to databases such as Excel, CSV, Tab, DB2, Oracle, SQL Server, etc.). Next, we merged the three datasets using VAERS_ID performing an inner join on the dataset. There are approximately fifty-two (52) thousand unique patient records and approximately 20 variables. Since the goal of this study was to establish causes for a patient entering the ER having a serious condition, it was determined to use the following business rule:***If ER_VISIT = "Y" then SERIOUS = 1 ELSE SERIOUS = 0.***

The above SAS script creates a new column called ER_VISIT. Note that "Y" means "Yes" i.e. if the patient visits the ER, then we categorize condition as serious represented by 1 in the ER_VISIT column otherwise non-serious condition is represented by 0 within the same column.

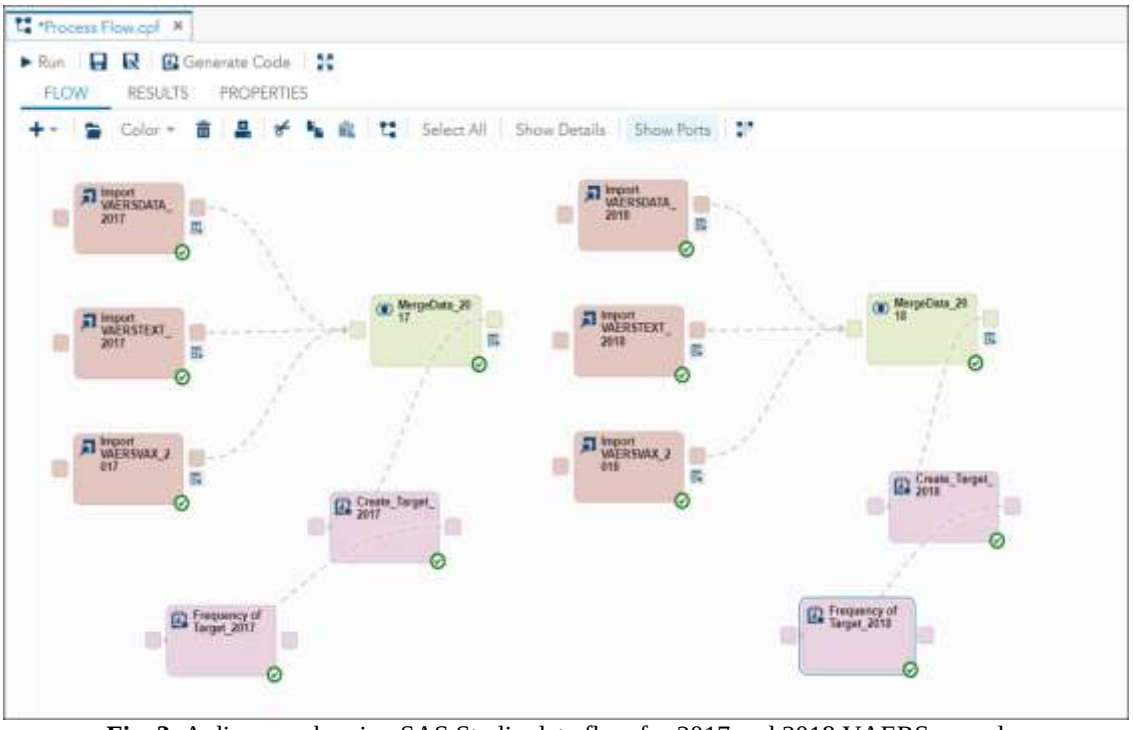

Fig. 3. A diagram showing SAS Studio data flow for 2017 and 2018 VAERS records.

Subsequently, we performed one-way frequency for the newly created variable to assess the theoretical soundness for predictive modeling. Based upon experience, it is a best practice to have 80% 0's and 20% 1's for a binary target (Y). Our results yielded about 88% 0's and 12% 1's – which is deemed acceptable. We would not wish to have a 1:1 split or more 1's than 0's as that would not theoretically ensure a strong predictive model. The goal of the modeling process is to establish the drivers that increase the percentage of 1's and optimize those results.

III. RESEARCH ANALYSES

3.1 Data Mining & Predictive Modeling

The data mining processes was completed in SAS Enterprise Miner (EM) and it included the analysis of structured data as previously demonstrated and described (Ajani and Habek, 2018). In addition, due to the advanced nature of this study which incorporates text analysis; we describe the text mining process as performed using the SAS EM text mining tools. Furthermore, we compare the predictive power of structured and unstructured data. Figure 10 shows the final data flow process for the analysis. SAS Text Miner within the Enterprise Miner software was used to analyze the results. SAS Text Miner discovers information buried in collections of text. By automatically reading text data and delivering algorithms for rigorous, advanced analyses, the solution makes it possible to grasp future trends and act on new opportunities more precisely and with less risk. It includes advanced linguistic capabilities within the core data mining solution of SAS Enterprise Miner so you can easily extend text insights into structured data mining and predictive analysis. For the practitioner, the software saves money and resources by automating the time-consuming tasks of reading and comprehending electronic text. By consolidating structured (quantitative) data sources with text-based (unstructured) information in a common environment, you gain a more accurate, complete view of your data. Comparison analysis performed using both types of data produces descriptive and predictive models that enables the researcher to spot opportunities and accurately recognize trends, leading to fact-based, prioritized actions.

3.2 The Text Mining Process

Once data were merged in SAS studio, we opened the Enterprise Miner and imported our data. In addition to connecting our algorithm nodes (tree for without text to data partition, and regression to input node) as described by Ajani & Habek (2018), we also connected the first text tool (i.e. text parsing node) to the input node. Data partition was used to split our data into training, and validation in the ratio 7:3 respectively. Subsequently, the text mining process began as shown below:

3.3 Text Parsing

The first part of the text mining process was parsing the document collection using the SEMMA tool shown in Figure 4.

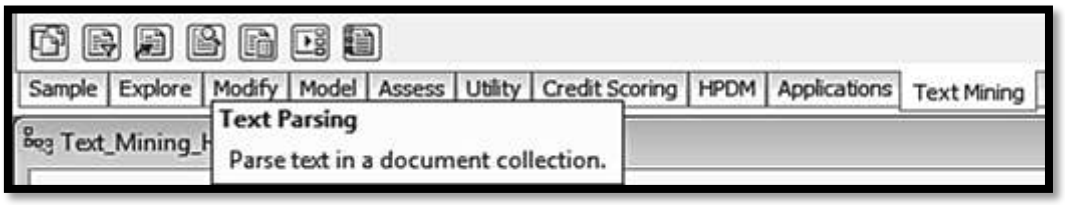

Fig. 4. SAS Text Miner showing text parsing node

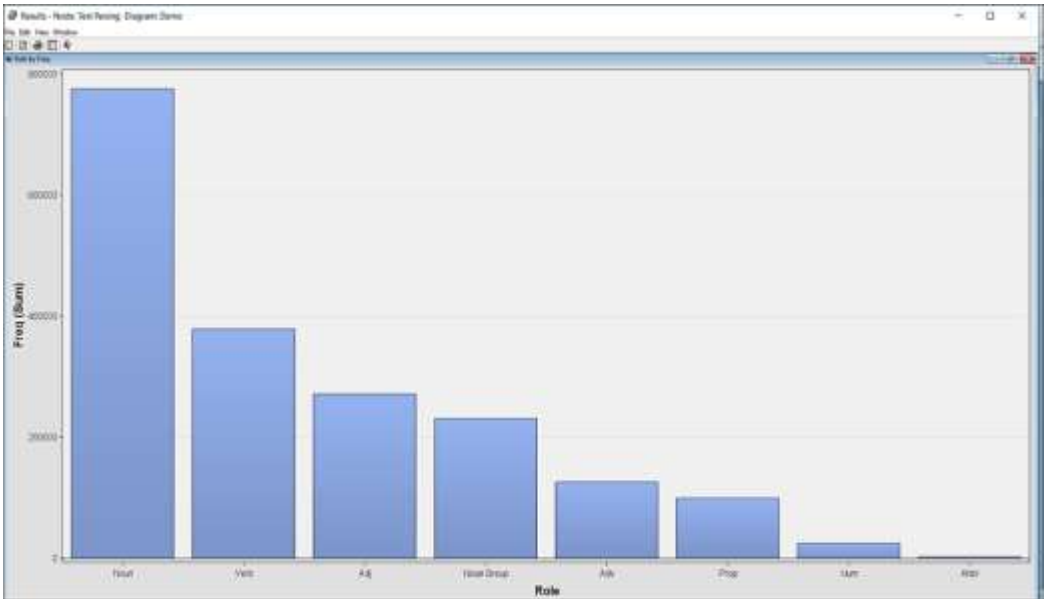

Fig. 5. SAS Text Miner Output showing text parsing results

There are several properties that come into play here: (1) The variable that will be parsed – in our example – the symptom text field; (2) Different parts of speech and noun groups are defaulted to "Yes"; (3) Find entities in parsing if desired such as names and addresses; (4) Ignore elements such as parts of speech (i.e., prepositions, pronouns), entities (i.e., names, addresses), attributes (i.e., numbers, punctuation); (5) Stem terms such as "swollen", "swelled", and "swelling" can be rolled up to "swell"; (6) Have a synonym list used in the analysis where you feed a table of terms that are commonly used in the industry; (7) Have a start/stop list where you can feed a table of terms to use by force or remove from the analysis. Figure 5 shows SAS output for the text parsing procedure.

3.4 Text Filtering

The second part of the process deals with filtering, which seeks to reduce the number of terms or documents included in a text analysis. Although there are many properties within this node as well, one key element is to assess the interactive filter viewer. Figure 6 and 7 illustrate the Enterprise Miner flow and filtering results:

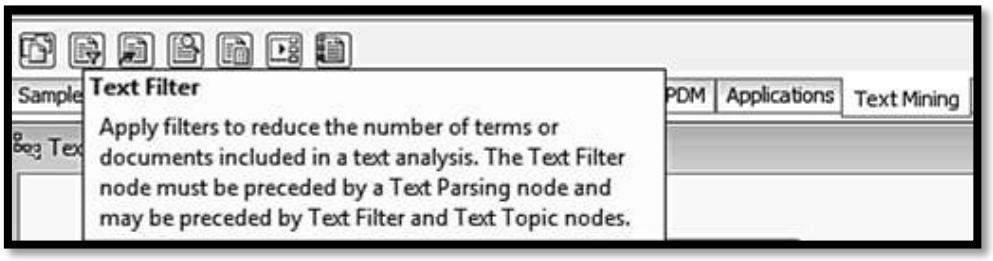

Fig. 6. SAS Text Miner Output showing the text filter node

Figure 7 also shows the synonym lists towards the bottom and the corresponding documents for a specific word searched for – seizure, for example. There are corresponding information displayed in the synonym list table including: The term or phrase; Frequency of occurrence for the term or phrase; Whether you wish to keep the term/phrase; The statistical weight applied for further analysis; The role of the term and; The term's attribute.

Terms

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
serious condition	1	1	☐	0.0	Noun Group	Alpha
serious confusion	1	1	☐	0.0	Noun Group	Alpha
serious cri	1	1	☐	0.0	Noun Group	Alpha
serious criterion	172	116	☑	0.054	Noun Group	Alpha
serious diarrhea	1	1	☐	0.0	Noun Group	Alpha
serious due	1	1	☐	0.0	Noun Group	Alpha
serious effect	1	1	☐	0.0	Noun Group	Alpha
serious event	3	3	☐	0.0	Noun Group	Alpha
serious illness	2	2	☐	0.0	Noun Group	Alpha
serious issue	1	1	☐	0.0	Noun Group	Alpha
serious life	1	1	☐	0.0	Noun Group	Alpha
serious muscle	2	2	☐	0.0	Noun Group	Alpha
serious pain	1	1	☐	0.0	Noun Group	Alpha
serious problem	1	1	☐	0.0	Noun Group	Alpha
serious reaction	14	11	☑	0.117	Noun Group	Alpha
serious sequela	2	2	☐	0.0	Noun Group	Alpha
serious side	4	2	☐	0.0	Noun Group	Alpha
serious spontaneous report	1	1	☐	0.0	Noun Group	Alpha
serious swelling	1	1	☐	0.0	Noun Group	Alpha
serious symptom	1	1	☐	0.0	Noun Group	Alpha

Fig. 7. SAS Text Miner Output showing text filter results

For further investigation of a specific term, such as seizure, you can right click on the term and view concept linking. Figure 8 (below) reveals our concept link map. We first noticed an association with the term "serious reaction" as well as a "+" indicating that you can expand on that term for further associations. It is indicated that it occurs a total of 11 times in 11 documents. Next, we expanded on "swollen joint" which was found to be associated with "still suffering", "insignificant" and "joint". If you recall, our business question is to establish drivers for a condition being diagnosed as serious from notes as the patients enter the emergency room. Therefore, all these associations can have a deep analytical impact on the outcome.

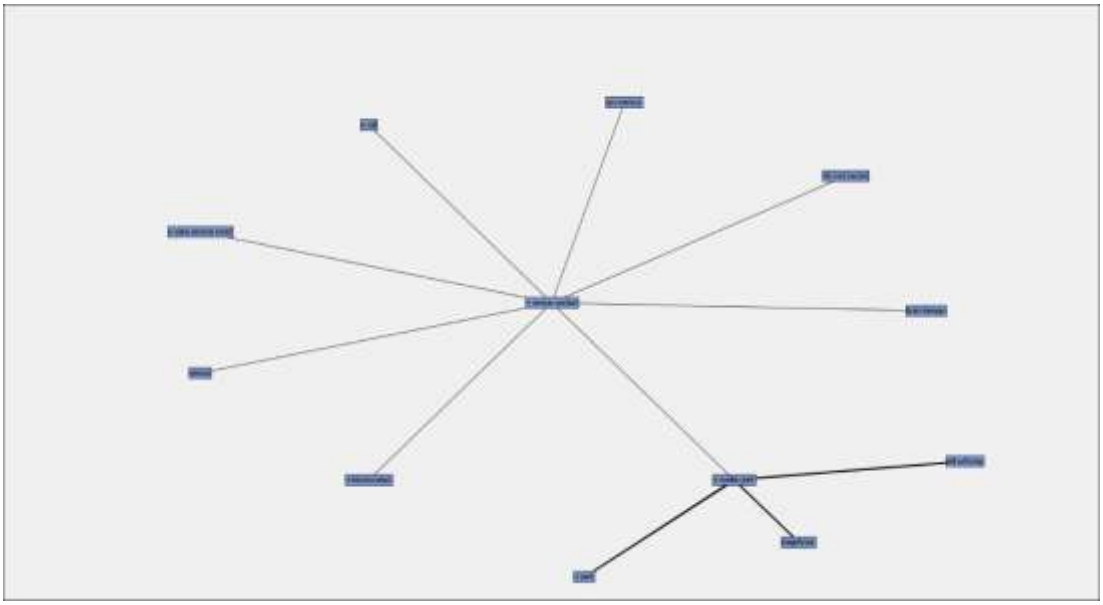

Fig. 8. SAS Text Miner Output (Text Filter) showing the Concept Link Map Results

3.5 Text Topic

This final step in the process is similar to cluster analysis, with one main exception – namely, the singular value decomposition (SVD) coordinates that are created in the cluster process are then converted to variables containing key words or phrases form the SVD coordinates. Hence, the interpretation is much clearer to the user. For this reason, a best practice is to feed the topics into successive modeling nodes for further analysis instead of the cluster node. Figure 9 illustrate the Enterprise Miner flow and text topic results:

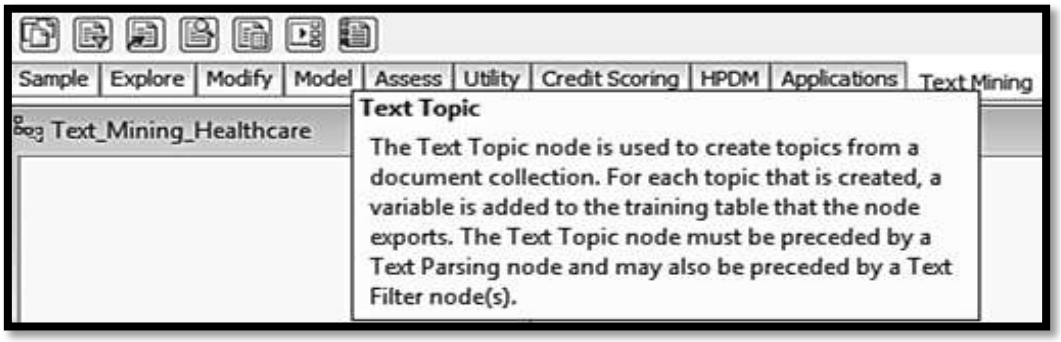

Fig. 9. SAS Text Miner Output showing text topic node

The results here include the terms in each topic, number of terms, and the number of documents they occur in. Again, this is similar to the cluster node with the SVD information actually being interpreted now into meaningful topics which will deem to be helpful in subsequent model development.

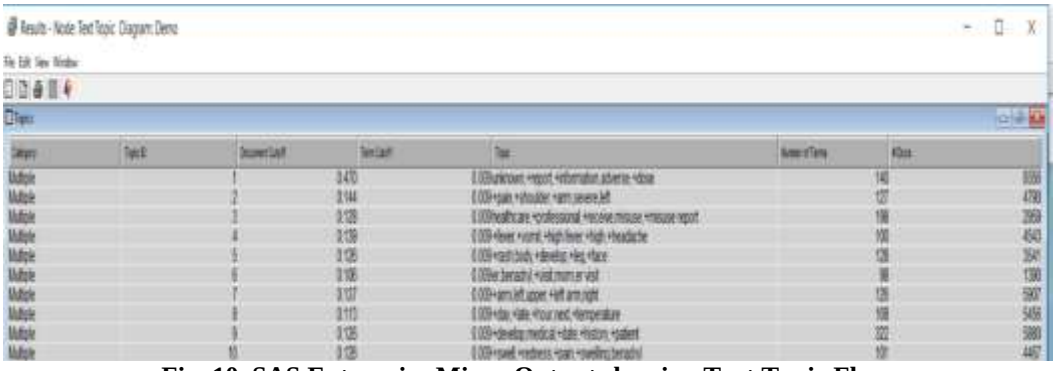

Fig. 10. SAS Enterprise Miner Output showing Text Topic Flow

3.6 Integration with the Enterprise Miner SEMMA Process

Now that we have completed the text analysis, comparing predictive models is the next step. In order to show the true value of applying text analytics using text miner, we compared two sets of different models: One using text miner results and the other assuming there is no unstructured data at all, namely, using the structured data only. Two (2) algorithms were compared: Decision tree (with/without unstructured data) and Logistic regression (with/without unstructured data). Each of the two algorithms were applied with and without text miner results (see Figure 10). Hence, a total of 4 models. The goal here is to show true separation with respect to predictive power of text mining. Again, recall that our model here is to establish drivers for a condition being diagnosed as "serious" based on a set of attributes, and develop the likelihood of a serious condition. Furthermore, once model is deemed to be the "winner", the next step is to score a new set of patients to determine the likelihood of a serious condition.

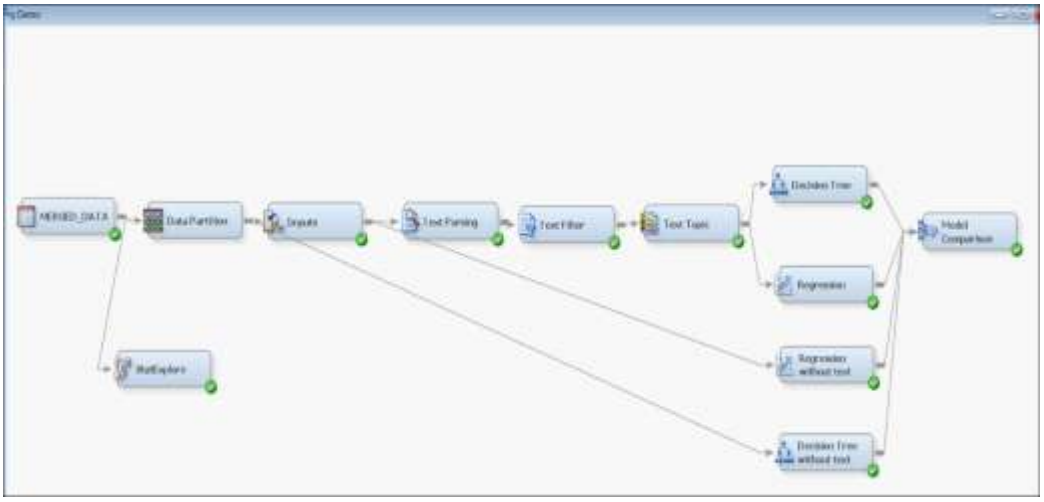

Fig. 11.The Final Predictive Modeling Flow

IV. RESULTS

4.1 Explaining the statistics

The fit statistics shown in Figure 12 below indicate that of the four models, the decision tree (with text) was selected as the "winning" or best model of this analytic process with misclassification rate that is the lowest of explored models. In theory, this would be a great selection to inform decision making. However, business rules often entails several factors. For instance, healthcare systems often harp on patient outcome, quality and safety as well as cost savings. Therefore, it is often necessary to have a holistic exploration of the output results. First we observed that the misclassification rates of the two algorithms are quite close (i.e. Tree and Regression - with and without text). However, a better improvement in misclassification rate is noticeable between regression (with text) – 0.183 and (without text) – 0.234. Fig. 12 illustrates the cumulative lift for the validation data set for both sets of models.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Tree	Tree	Decision Tr...	serious		0.181103
	Reg	Reg	Regression	serious		0.182851
	Tree2	Tree2	Decision Tr...	serious		0.214365
	Reg2	Reg2	Regression...	serious		0.234287

Fig. 12. Model Comparison Diagram

Figure 13 clearly indicates two sets of models. The top set of colors are those models incorporating the text information using the unstructured data. The bottom set of colors used the structured data only. It is important to note that just using the structured data alone provides adequate predictive power, where cumulative lifts are about 1.8 at the 20th percentile (regression – without text). However, when unstructured data is also applied, the cumulative lift almost increases by almost a whole point, approaching 2.6 (regression – with text), providing a significant improvement in predictive power. However, at the same 20th percentile, cumulative lift of the tree models with text (2.3) and without text (1.7) is about 0.6.

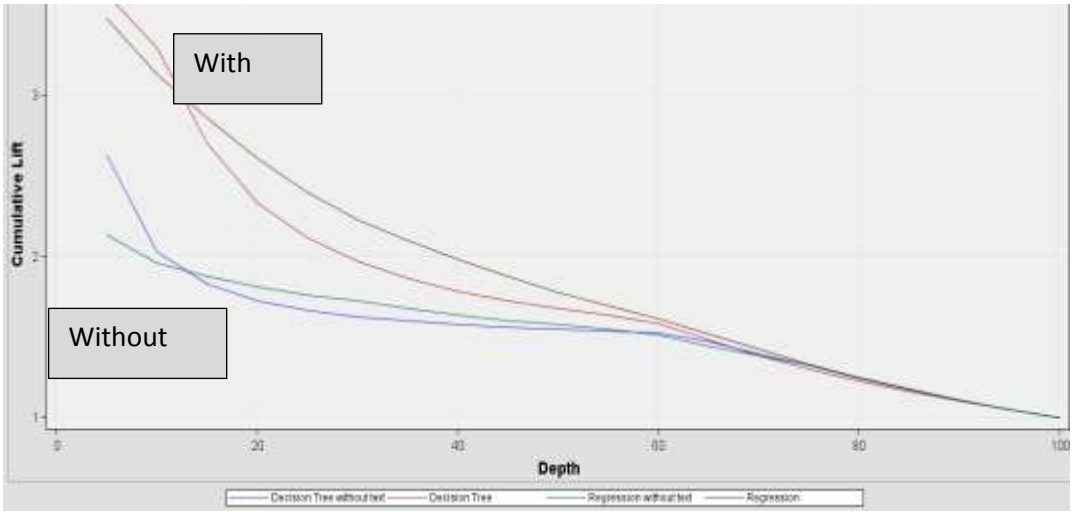

Fig. 13.SAS Enterprise Miner Output – Model Comparison Results – Cumulative Lift

Another valuable piece of output is to observe the Cumulative % Captured Response output (Fig. 14). We see a clear difference between the two sets of predictive models. This graph may lend itself to clearer interpretation than the previous one.

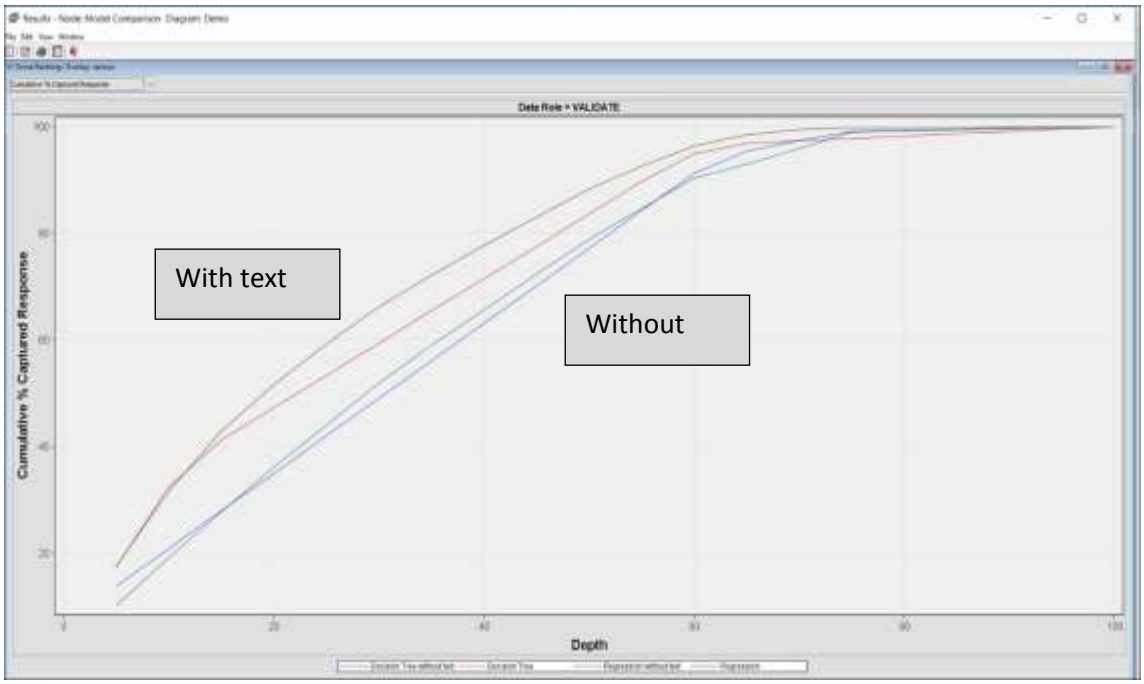

Fig. 14. SAS Enterprise Miner Output showing Model Comparison Results –Cumulative % Capture

Basically, the result depicts how much of the event, or in our case, patients likely with a serious condition, can be captured within a specific decile of a newly scored dataset. Our best practice focuses on the 20th decile for results. Notice that just using the structured data alone provides good results, with above 34% of the events being captured using the decision tree (34.5%) and regression (36.20%) - without text. However,

including the unstructured data yields about 52% and 46% of the events being captured by the regression and decision tree respectively. This is quite impressive with respect to predictive power improvement.

Let's now put these results into more of a real application. Suppose we have a new file of 1,000 patients that came to the ER within a certain period of time, and we wish to assign a likelihood to each patient as to the probability of their condition being serious. Based on our predictive model selected, we score those 1,000 patients and assign a probability for each patient/record accordingly. Figure 13 above would take the scored file and sort the records into the top 20% of the patients. Therefore, we take 20% of 1,000 which is 200 patients. The results above indicate that by using the text mining information with the structured data, we can capture about 52% of those 200 patients as having a serious condition (which may lead to being admitted to the hospital). Basically, we would capture 104 patients (using text miner) as opposed to about 72 patients (using structured data alone). The consequences here from a patient care perspective and specifically optimizing patient care can be very critical.

Finally, the results from a classification perspective for both sets of models is also important to note. Our goal in predictive modeling when we predict a binary event is to classify the scored patient correctly, and hence minimize the error as much as possible. Therefore, a common measure for accuracy is the misclassification rate for the validation dataset. Our goal is to have that close to zero as possible. For our example, the models using the structured data alone yield accuracy of about 79% or misclassification of about 21%. When incorporating the unstructured data, we achieved an accuracy of about 82% or misclassification of about 18%. Although, the difference may not seem to be that large, with respect to predicting a patient's condition to be serious, that difference can prove to be critical, especially when dealing with a large hospital or health system.

From a business approach, we chose the Regression algorithm as our model because of the huge amount of money that is potentially saved. Below is information about the causal factors (driver) of the likelihood that a patient will go to the ER with a serious condition: Age; No. of days from vaccine to adverse effect; State; Vaccine site (of the patient body); Gender; Unstructured information from the ER visit; Vaccine type; Adminby (pub, mil, pvt etc.); Fundby (pub, mil, pvt etc.).

4.2Business Value

The average medical risk of ER visit per patient is over $1,700 (Dahlen, 2019). The following table presents a picture of the dollar saving from using unstructured data when available in developing models for managing patient healthcare based on the patient data from this study. The number patient column indicate the total number of patients that our models identified are likely to go to the ER after vaccine administration. Each model below indicate different numbers. However, we established a cutoff (at the point where probability > 0.5) based on random chance theory which EM uses as cutoff to divide populations into 0s and 1s which in this case, the latter represent the event of serious condition that may require ER visit. Total cost per model was multiplied by $1,700.

Table 1: shows the number of patients at risk of developing a serious condition and the potential savings per model

Model	Number of patients	Costs	Cost savings $$
Decision Tree	608	1,033,600	
Decision Tree (w/text)	1,952	3,318,400	2,284,800
Regression	346	588,200	
Regression (w/text)	2,290	3,893,000	3,304,800

The goal of management is to improve patient care while minimizing costs. Any of the 4 predictive models can help identify patients at risk, the health system can establish policies that target such patient before they develop serious conditions that may result in ER visit, thereby leading to better patient outcome and cost savings. Furthermore, models using integrated data appear to perform better at identifying patients at risk adding even more value to the modeling process.

V. CONCLUSION

Each of the models clearly identified patients that could potentially benefit from interventional program such as follow-up health coaching programs based on the driver indicators. The above data and cost calculation is based on our case data of approximately 57,000 patients. Note that this data is from the HHS/CDC database with data contributed from health systems administering vaccines on several diseases across the United States in 2015. According to a CDC report on seasonal flu alone, based on reports of vaccination from survey respondents, more than 140 million people were vaccinated during July 2016 through May 2017 flu season among the U.S. population. Although it is difficult to know the approximate number of all vaccination administered in the US yearly because most vaccine data remain unreported, it is safe to assume that hundreds

of millions are administered in the US annually. Historical data that incorporate unstructured symptomatic patient response could lead to enormous cost saving. For instance, if incorporating unstructured data into model building for an approximate 60,000 data points leads to a ROI or cost saving of $3.3 million, extrapolating to about 60 million can potentially lead to a saving of $3.3 billion. This is tremendous and the value added by the combined use of unstructured data in conventional model building. The value is that better analytics and reporting helps keep vaccines safe. Each VAERS report provides valuable information that helps FDA & CDC make sure that vaccines are safe.

This study discusses a very important analytical exercise around unstructured data mining. This concept is often not utilized to the extent it should be. A large amount of information specific to healthcare tends to be in an unstructured format. Practitioners ask a lot of questions from patients, with increasing opportunities to capture responses to such data, they become recorded or stored where they can be retrieved. The incentive to retrieve such for analysis may not be apparent, we believe that exposing the value adding potentials of unstructured data is important to change the current practice. Our research strengthens other authors' (see above) conclusion that data integration significantly improves algorithm performance and adds value in multiple application in the health/medical field. While all of the models tested appear to benefit analytical modeling, the regression (despite the theoretical selection of the tree – with text) shows robust application in its ability to capture more of the potentially serious conditions. It also clearly identified conditions or drivers of a patient returning to the ER. Tapping into that information will prove to be more valuable in terms of predictive modeling and providing actionable business results to aid in optimizing patient care.

REFERENCES

[1]. Ajani T., & Habek G. 2018. Machine Learning as a Tool for Optimizing Diabetics Care Management Using SAS Analytic Suite. 2018 Proceedings of the Conference on Information Systems Applied Research ISSN: 2167-1508 Norfolk, Virginia v11 n 4810

[2]. Alkhatib, M.A., Talaei-Khoel, A & Ghapanchi, A. H. (2015). Analysis of Research in Healthcare Data Analytics. Australasian Conference on Information Systems, 2015 Sydney.

[3]. Abhyankar, S., Demner-Fushman, D., Callaghan, F.M., & McDonald, C.J. (2014). Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc. 2014; 21:801-807.

[4]. Clarke, R. (2016). Big data, big risks. Inform. Syst. J. 26(1), 77-90. http://dx.doi.org/10.1111/isj.12088

[5]. Dahlen, G. (2019). How Much Does an ER Visit Cost? Consumer Health ratings: Your Guide to Quality and Cost. Retrieved June 21, 2019 from: https://consumerhealthratings.com/how-much-does-er-visit-cost/

[6]. Floyd, J.S., Heckbert, S.R., Weiss, N.S., et al. (2012). Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. JAMA 2012; 307:1580–2 [PMC free article] [PubMed]

[7]. Gunther, W.A., Mehrizi, M.H.R., Huysman, M. & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. Journal of Strategic Information Systems 26(2017); 191-209.

[8]. Kern, E.F.O., Maney, M., Miller D.R., et al. (2006). Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 2006; 41:564–80 [PMC free article] [PubMed]

[9]. Koh, H.C. & Tan, G. (2005). Data Mining Applications in Healthcare. J Healthc Inf. Manag. 2005 Spring; 19(2):64-72.

[10]. Raghupathi, W. & Raghupathi, V. (2014). Big Data Analytics in healthcare: promise and potential. Health Information Science and Systems 2(1) page 3.

[11]. Saini S., & Kohli S. (2018) Healthcare Data Analysis Using R and MongoDB. In: Aggarwal V., Bhatnagar V., Mishra D. (eds.) Big Data Analytics. Advances in Intelligent Systems and Computing, vol. 654. Springer, Singapore.

[12]. Scheurwegs, E., Luyckx, K. Luyten, L et al. (2016). Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc. 2016; 23:e11-e1

[13]. Zhan, C., Elixhauser, E., Richards, C.L., et al. (2009). Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. Med Care 2009; 47:364–69 [PubMed].

Taiwo Ajani" Improving Model Predictive Power by Integrating Structured and Unstructured Data: A Case Study of the Vaccine Adverse Event Reporting System (VAERS)" American Journal of Engineering Research (AJER), vol. 8, no. 11, 2019, pp 09-19