A Framework for Creating Structured Data from Mixed Free-Text and Categorical Data: Repurposing Emergency Medical Services Data for the Assessment of Child Maltreatment Risk

Nicholas Pondel BS, MLS(ASCP)

Entering Class of 2016

Culminating Project for

Master of Public Health

With a Specialization in Biomedical Informatics

Committee Members

Courtney Hebert MD, MS

Rebecca Andridge PhD

David Kline PhD

The Ohio State University College of Public Health and

College of Medicine's Department of Biomedical Informatics

Committee Review: July 11th, 2019

Submitted: July 14th, 2019

Abstract

Background: Child maltreatment is a massive problem in the United States. Better understanding of the risk factors involved, and better screening tools are needed. Together, these may help improve reporting and intervention rates for maltreated children. Emergency Medical Service (EMS) providers have a unique perspective on a patient's case and can offer new insight into social and environmental risk factors for maltreatment. Natural Language Processing techniques offer an effective method for using the unstructured, mixed data types generated by EMS in the field.

Objective: The purpose of this project was to develop an analysis framework using R to create structured, binary data from mixed free-text and categorical data generated by EMS providers providing routine care to pediatric patients.

Methods: We used a unique method for de-duplicating and combining records in R, natural language processing for word frequency analysis, manual review by a physician panel to develop risk factor variables, and a custom R script to search the text data for these variables to generate a structured dataset. All computation for this project was done using open-source software and the script code created will be made available at "github.com/npondel/mph_culmproj".

Results: We were able to successfully create a pipeline which transforms unstructured text data into a structured set for NLP. This was followed by string-matching for the variable keywords generated from manual review. The string-matching code generates binary values for manually created variables which will be used in a classifier model to assess maltreatment risk in pediatric patients. The scope of this project encompassed the development of the R tools used to clean and process the CDF data, as well as coding used to generate variables from it. Classifier models developed from these data will be published in a future paper by the team at NCH. The Center for Injury Research and Policy at Nationwide Children's Hospital (NCH) will utilize and iterate this pipeline for their ongoing projects with the Columbus Department of Fire (CDF).

Conclusion: By creating a unique framework of Natural Language Processing and human review, we were able to repurpose routinely collected Emergency Medical Services (EMS) data and extract new features from it. This will provide unique insight into the public health problem of child maltreatment. The framework developed is reusable for the NCH team as new data is received from the CDF. Similar methods could be applied to other clinical data sets to classify cases and help automate the extraction of useable information.

Introduction

Through a relationship with the Columbus Department of Fire, Dr. Julie Leonard's team at CIRP received EMS run data from 2011-2015 in order to examine risk factors for a wide variety of public health issues. By combining this EMS data with that from the NCH emergency department, there is the possibility of linking risk factors with maltreatment reports. While the structured data from the CDF was useful for some studies, to address the issue of child maltreatment risk we wanted to make use of the entire dataset, including the free-text narrative data. The reasoning was that EMS providers are in a unique situation to observe the child in their home/school/family environment and therefore may be able to pick up more social and environmental risk factors by analyzing this narrative data. Barriers to this analysis were numerous, including the overall unstructured nature of the free-text data coupled with the untidiness of data collection in the field by CDF. In order to properly utilize these data in risk factor models, it became clear we needed to clean and process it in a unique way. This challenge became the concept and motivation behind my culminating project presented here. First, we will review some background literature on child maltreatment and natural language processing. Following this is a detailed description of the data utilized in our pipeline and methods used to develop it.

Literature Review

Child Maltreatment

Child maltreatment is a massive problem here in the United States. According to the World Health Organization (WHO), child maltreatment can be defined as "physical and emotional mistreatment, sexual abuse, neglect, and negligent treatment of children as well as to their commercial or other exploitation" (Butchart, et. al. 2014). In 2017 approximately 2.3 million cases of suspected maltreatment were referred to child protective services, with over 673,000 cases confirmed (D.H.H.S 2017). Reports of suspected cases are submitted by a wide variety of individuals, including educators, law enforcement, and healthcare practitioners. Those on the front lines of emergency medicine are some of the first professionals who may recognize the problem for a child. In 2017, healthcare personnel made up the 4th largest group of report sources (9.6%) behind social services (11.7%), law enforcement (18.3%), and educators (19.4%) (D.H.H.S 2017).

Recognizing and reporting maltreatment can be difficult for many individuals, even experienced professionals. One survey of pediatric physicians found that only 24% of injuries in which medical personnel suspected child maltreatment to be involved were reported to child protective services (Flaherty et al 2008). Another study was able to show that in fatal cases of child maltreatment there was a significant correlation with

higher frequency of EMS use (Shenoi, Rohit P., et al. 2017). Unfortunately, there is little data available on the frequency of EMS use by child maltreatment victims in non-fatal cases. Still, emergency medical personnel, especially those working the pre-hospital setting, are in a unique position to observe the home and family environment surrounding a potential maltreatment case.

Natural Language Processing

This project is unique in its use of EMS data to assess maltreatment risk. The data received from the CDF was quite unstructured. Numerous cases had multiple rows of data entered and many cases had missing values for the categorical variables. Due to the unstructured nature of the EMS data collected in the field, a unique data analysis method was needed to extract useful features from this free-text data mixed with structured data points. Natural language processing is a domain of informatics that has been applied to many information retrieval and medical knowledge problems. It remains an evolving field and will no doubt become increasingly useful in the future as we enter further into the era of ever-increasing data sources demanding new and advanced analysis techniques. The prospect of repurposing routine EMS care data to offer improved public health through maltreatment risk prediction is an exciting one. NLP offers a unique and powerful option to better utilize this data from its "natural" form.

Before electronic medical records (EMR)s, patient's charts were filled with notes and narratives documenting progress throughout their stay. As EMRs began to gain widespread adoption, clinical notes endured in a digital format. Natural language processing (NLP) is a broad term for any sort of computerized system which can extract machine-readable information from text data. NLP systems have existed since the early 1950's within the context of library science and information retrieval (Nadkarni et al 2011). These sorts of systems see widespread use today in modern search engines and library systems. Within medicine, NLP-generated data can be used in models to predict outcomes, aid in decision making, and assist in automated screening for diseases and conditions. It can also provide new insights for researchers by extracting information from text which is too noisy and time consuming to read manually. Various methods and frameworks for NLP have been developed by researchers and clinicians to fit their own needs. Here we will examine a select few examples of its use in the medical field.

NLP has been used to assist in the generation of inpatient problem lists (Meystre et al, 2006). In this study, researchers sought to solve the problem of poorly updated problem lists in patient's charts using an NLP-based method of list generation. For the creation of the vocabulary in their system, the team started with the Unified Medical Language System (UMLS) Meta-thesaurus and paired it down the list to 2500 concepts. The

ULMS meta-thesaurus is a vast collection of multiple medical vocabularies including CPT, SNOMED-CT, ICD-10, LOINC, and more. The custom subset of 2500 concepts was manually mapped to a list of the 80 most common items used in the institution's problem lists. For the actual searching and processing, the UMLS MetaMap Transfer Application was used along with a negation detection system called NegEx 2.

Evaluation of the researchers' pipeline was accomplished by creating a reference standard via manual chart review. A team of physicians reviewed 160 clinical documents and classified each according to the 80-item problem list. For a thorough comparison, the researchers collected results for manual physician review alone, NLP alone, and a third method where physicians made selections and were then presented with the NLP results and given the option to change their own selections. In terms of recall (sensitivity) the NLP system alone provided the best results while the physician assisted with NLP provided the highest precision (specificity). This particular NLP system was eventually implemented at the institution to generate problem lists in an automated fashion, leaving physicians to focus on reviewing them for accuracy. The server applications used in their project are typically applied to a wide scope of clinical data. While they are impressive in their breadth of capabilities, they also would be far too wide in scope for application to our particular challenges in this project. That being said, the overall principle of manually creating a subset of the UMLS meta-thesaurus mapped to the institution-specific concepts desired is a unique way to customize an NLP system.

NLP is useful as another tool to aid in the detection of post-surgical complications (Murff HJ et al, 2011). For this study, the authors sought to improve upon post-operative complication detection in their hospital. Their study was part of larger quality initiative, and thus they had access to a wealth of surgical notes curated by trained nurses to compare an NLP model against. The current standard practice was to assess post-operative complication data using billing codes as the metric. Using an NLP system, the authors thought they could improve the ability to correctly identify any of five key post-operative complications. These targets included acute renal failure, sepsis, deep vein thrombosis, myocardial infarction, and pneumonia. The study was performed at a Veterans Affairs hospital using a surgical population made up of primarily older males (a typical VA population). Even so, the quality of comparison data made it a perfect chance to try an NLP based system. The authors chose a more off-the-shelf approach than some others. They utilized the SNOMED-CT vocabulary processed using the "Multi-Threaded Clinical Vocabulary Server" (Elkin et al 2006). This application was developed at the Mayo Clinic and published for research use on free-text searching. The authors did manually add search terms for various concepts and abbreviations they believed to be common to their institution but possibly not others.

Both billing code data and the NLP system were compared against the nursing chart review as a gold standard. In their results, the authors found that the billing code data matching had poor sensitivity across all complications except myocardial infarction. Specificity metrics using billing code data were already very high across all complications. The NLP classification system significantly improved sensitivity metrics across 3 of the complications. Among the remaining ones: myocardial infarction was already well classified using billing code data and pulmonary embolism / deep vein thrombosis were not well classified by either data source. These findings are interesting, suggesting possibly that certain outcomes are easier to predict due to their unambiguity (a myocardial infarction). It also may suggest that it is not always possible to improve classifications where diagnosis itself is more difficult or more nuanced. This study had strength in its large population and high-quality comparison data. While it used rather conventional NLP systems, it showed they could outperform the current standard for analytics in many metrics.

In our last example, NLP can be used to help identify individuals at risk for complications such as suicide after discharge (Perlis, R. H et al 2016). The system developed by researchers at Boston's Mass General and the Brigham and Women's Hospital used an interesting method of NLP to achieve a unique and simple system. They sought to improve metrics for death by suicide post-discharge among the general population at a large medical center. The authors stated that prior research showed a clear correlation where those attempting suicide were more likely to have seen a primary care physician recently. Their goal was to create a simple NLP system that could stratify individuals into risk categories with higher strata referred for follow up with mental health professionals.

Demographics and physician notes data were pulled from the databases of both hospitals involved and matched with patients records in the Massachusetts Dept. of Public Health records for death certificates. The authors considered both death by suicide and accidental deaths in their analysis to account for misclassification. The most interesting part of the study was the method of NLP configuration. The authors created a custom list of 3000 common English words and assigned a "valence" score to each one. This scoring system was a continuous scale used to subjectively measure the positive or negative connotations of the word in common usage. From this scoring method, the NLP system assigned an overall valence score to each document by aggregating all the words matching the custom list. This accounts for both the strength of the valence and the frequency of the word's use in a document. Using this unique NLP system, the researchers were able to improve a baseline comparison regression model using only structured demographics data. The improvement was small; however, the model did show a significantly improved fit with the NLP additions. In their final, NLP-assisted model, the upper 50% of risk scores correctly identified 82.1% of suicides in the study population. This suggests

that a mode like this could be beneficial for referring individuals for follow-up. Much in the same vein, the models to be fit from the data which was generated in this culminating project are intended to be a tool in screening for follow-up by assessing overall risk for maltreatment.

Summary

As we can see, Natural Language Processing methods can be applied to a wide variety of clinical problems using a wide variety of methods. Though all the methods presented here are unique and interesting, none of them are quite ideal for the dataset we are utilizing in this project. First, they were all developed for use on purely free-text narrative data alone, not a mixed dataset of narratives with categorical data. Additionally, neither of the two pre-built NLP methods examined above have been designed or tested for use on EMS data. The novel concept of valence scoring is very interesting, but the method described in the previous paper is more appropriate for psychological risk factors rather than social and environmental ones. However, this manually created scoring method for an NLP system did inspire the use of custom-created word variables in our own project.

Information generated from NLP can provide a closer look at the providers thinking. It can help show the clinical perspective for a patient's case. It can also provide information which individuals have not entered into structured data fields. The volume of free-text data generation in medicine is unlikely to change any time soon. Clinicians are only human and often want to describe what they observe in their own words. Writing free-text can feel more natural compared to checking boxes and selecting from drop-down menus. Therefore, we need to continue to examine new methods of data analysis to unlock the potentially useful information trapped in free-text data. In this project, I will present one method that was developed to solve a specific case of analyzing unstructured, mixed data types. My end-goal of a structured dataset will provide the starting point for further research into possible classifier models by the team at NCH. It is my hope that this project will provide an example for others to iterate on and apply similar methods to their own mixed-data problems.

Methods

Data

Computations for this project were performed using the R statistical programming language (R Core Team 2019). R is a high-level coding language primarily used for the analysis and visualization of data. It can be implemented across nearly every modern operating system including unix-based server systems, as well as desktop MacOS and Windows. The R language is often preferred by biomedical researchers for its relative ease

of use (compared to more intricate languages like Python). It functions mainly as a tool for writing scripts for replicable data analysis. Another factor contributing to its popularity is the free and open-source license, allowing for powerful expansion by a large online community of amateur and academic package developers.

This culminating project was completed as part of a maltreatment risk factor identification project approved by the IRB at Nationwide Children's Hospital. Data for this study was acquired from the Columbus Fire Department for the years of 2011-2015 for all patients with an age of less than 18yrs old. The dataset contained all information gathered by EMS professionals during a run (individual patient case). EMS records were linked with data from the Nationwide Children's Hospital emergency room by probabilistic record linkage using the R packages "RecordLinkage" and "EpiWeights" to match patients based on name, gender, and date of birth. Records with partial matches based on this system were manually reviewed. Manual review systematically established a match through examining date of birth, followed by the date of the encounter, and finally the patient's address. In total, 920 children were included in the set of outcome variables, having both CFD data as well as a match in the NCH emergency room data. The emergency room dataset contained important information on maltreatment reports filed by ER staff. Due to the sensitive nature of maltreatment cases, only report data was available to us. Case review and confirmation data is not made available for research use. For the purposes of the outcome variables, any report of suspected maltreatment was treated as a maltreatment case.

Data obtained from the CDF was divided into a number of individual datasets. We selected 7 of the 10 available in the data to use for our analysis. The datasets eliminated at this point included EMS given medications, EMS treatments, and previous known allergies. The remaining categories used in the project encompassed all other aspects of the EMS run: demographics, preexisting conditions, patient complaint, impressions, causes, symptoms, and the narrative.

Cleaning and Combining Data

In examining the data, it was immediately clear that there was little standardization as to what information should be captured in categorical variables like "causes" or "symptoms" and what should be included in the narrative (which is manual free-text). Different EMS providers have different habits in their data entry. There is a great deal of overlapping information in both text and categorical fields, as well as many cases where only categorical variables or only the narrative is entered. This fundamental issue presented unique challenges in extracting useable information from this dataset.

Early in the project, it was decided to utilize NLP to extract information from our datasets. We decided to compile all the available data into a text format and use word matching to identify cases with notable key words that may have a correlation with maltreatment cases. In order to develop these key word variables via manual review, we first needed a method of compiling all the data together into a text field to better examine it. Various methods were tested in an attempt to combine data by case ID, but we couldn't find a framework available in R for what we wanted to deduplicate rows and combine fields. We needed to combine the data in two dimensions as many cases had not only multiple data columns (narrative, demographics, etc.) but also multiple rows of entries for a single case due to the way the data is collected in the EMS system. We first needed to combine duplicate row entries, and then collapse the remaining variable columns into a single text field.

What we discovered and used in this process is a unique use of the "corpus" function in the text processing R package "quanteda". A corpus is a data type in R used by quanteda to hold a collection of text documents. The structure of a corpus dictates that it has two types of variables, an identifier and one or more text fields. The corpus data structure doesn't allow for duplicate identifiers and therefore appends a number to the ID when duplicates are found. We exploited this functionality to easily number our duplicate entries. We then pulled out this duplicate numeration from each case and used it as the "time" variable for reshaping using the common data manipulation tool "dplyr" and the "reshape" function. This process may make more sense when visualized (see figure 1 – Case Combining).

Word Counts

With all the data compiled into a text format, we used basic NLP through the tidytext R package to generate word counts. Counts of the most common single words, bigrams (pairs of two sequential words), and trigrams (triplets of sequential words) were extracted from the data. Using a script, the text data is first "tokenized" or segmented by a particular vocabulary type. The tidytext R package has easy to use functions for tokenization and word frequency analysis (see word_counts source code). Tokenization is adjusted according to the analysis of single words, bi-grams, or trigrams. After tokenization, the anti_join function from the dplyr package is used in order to remove any undesired "stop words" from the word counts. There is a built-in set of common English stop words in the tidytext package. These are words that are exceedingly common yet have little value in their syntax or meaning. Examples of stop words might include "the, of, and, or, there, he, she, they". In order to clean up the word counts manually beyond this, we created and iterated a custom list of stop words throughout the manual review process. Words removed in our custom set included common but not

meaningful words and phrases like “patient”, EMS squad identifiers, hospital names, medical tool specifications, and medication doses.

Variable Creation

These lists of word counts were submitted to a panel of physicians for manual review. Multiple times the lists were iterated to change the stop words. This allowed us to whittle down the word lists to only the most meaningful words and phrases. The review panel consisted primarily of an attending pediatric ER physician, a fellow in pediatric emergency medicine, and a research associate with experience in pediatric medical research. Following the stop words development, we began the process of identifying key categories of words that may have correlations with maltreatment risk. It is important to note that the creation of the word variables was performed blind to whether or not words and phrases came from maltreatment cases or not. This may help reduce bias and overfitting of our particular dataset when this framework is applied to future data from CDF. Each word variable would have one or more keywords associated with it. The presence of any of these key words in the case text would flag the variable as a match (assigned the value 1 with negatives being 0). Variables would not be mutually exclusive. In this way, we could use these variables as the inputs for classifier models. As often as time allowed for within their schedules, the physician team would meet to discuss and modify the stop words list which would be fed back into the word count generation for a new list of words. Eventually the review meeting scope was changed to define categories for the final models and classify available words from the list as key words for these categories. This regular review process made up the bulk of the time spent on the project over the past year.

String Matching

Development of the string-matching script began after an early draft of variables and key words was agreed upon. To accomplish the end result of a matrix of binary values assigned to cases and variables, we used a combination of functions from the R package “stringr” as well as classical computing functions like “grep”. Dplyr was also used again for data reshaping in examining means for variables to monitor quality and performance. Adjustments to the data source were needed in order to accommodate the need for searching and matching not only single words but bigrams and trigrams. The UNIX matching function “grep” works on the concept of text strings, which are continuous until separated by a space. Therefore, underscores were added to delineate bigrams and trigrams. We replaced all space characters in the dataset with underscores, and they were also added to the start and end of a line so that key words at the start or end would match precisely. Word variable lists were imported from sheets in a Microsoft Excel workbook into a “tibble” dataset in R using

both the “tibble” and “readxl” packages. The tibble data type closely resembles a traditional R data frame, however, it allows for nested lists. This allowed us to develop a function to search the text for each list of variables in succession.

We wrote a function to collapse each word variable list into a single vector OR statement for matching using the common UNIX function “grep”. Collapsing the words allowed grep to flag positive for any of the words being present in the text. Text case was ignored for this search. Using the “grepl” function returns logical variables of TRUE or FALSE for the presence of any of the words. These logical variables were simply converted to numeric values of “0” or “1” using a string replacement function from the package “stringr”. In order to search all word variables across the text dataset in a concise way, the “sapply” function from base R was utilized. This applies a function across a list of data and generates a data frame as output. In our case, it allowed for applying each variable’s word list across the text to generate a vector of binary values before moving on to the next list of words. Mean values for all variables were calculated in order to ensure no errors were made in the process. Bugs in the code or process were identified by extremely high mean values, indicated far too many matches to make sense. This is how we identified one issue with matching abbreviations such as “mi” for myocardial infarction. The grep function matches any instance of those two letters next to each other, which prompted us to develop the solution involving bracketing everything with underscores. This ensures that “_mi_observed_” in the dataset would flag positive while words containing the letters “mi” would not.

Results

This project produced a number of R scripts and a workflow for producing usable data from a previously unusable data source. This was accomplished through compiling the EMS mixed-type data into text-based data, analyzing word counts and frequencies, and finally the generation of binary variables through string matching. The final combined fields dataset contained 45693 unique encounters. The encounters included in this dataset were any case with any data entered in any of the 7 original datasets. Patients with multiple EMS cases in the set were maintained while duplicate entries for a single case where combined into the single text field for the case. The final text fields combined had a total of 43,275,350 characters with a mean character length per case of 947. At the time of this writing there were 35 variables developed from the manual review sessions with anywhere from 4 to 80 key words each (see figure 3 for example). The variables are still very much in development at this time. The string-matching script is performing well based on testing using the currently available variables and text data (see figure 4 for output example). By placing files into an input folder and running the R script, the NCH team can generate the output matrix of variables on the fly as they iterate the word lists into the future. This framework of review techniques and script tools created by this project will serve

the NCH team throughout this project and can also be adapted for future ones. By adjusting the parameters in the combining scripts, many mixed-type datasets could be transformed into a structured one and then fed into the string-matching script with its own set of word variables to match for. This tool will expand the team's capabilities to analyze new data and extract useful insights into public health issues. For a full diagram of the data flow through the pipeline, please see figure 2.

Public Health Impact

Through the effort of many individuals, we were able to develop a framework for the creation of not only clean text data but useable regression variables from previously unusable data. Much like the collaborative field of informatics, the team working with me on this project possessed diverse backgrounds in pediatric medicine, public health, research, and biostatistics. Through this project, we were able to apply informatics techniques to extract meaningful information from messy data. The developed framework opens up a world of possibilities for the future use of this data. While this particular study was performed on data received from CDF from 2011-2015, an ongoing agreement promises ongoing data flowing into the research team. The framework developed here will prove useful in the future as the NCH team develops new hypothesis for public health issues and new ways to utilize EMS data.

The broader public health implications of this project are significant. Child maltreatment is a massively under-reported issue in the United States. It's difficult to identify and even more difficult to get individuals to report on it. Automated screening tools may be one answer to the problem. Word variable lists could be tailored by institution and evaluated on other data types. With more development and testing, and a thorough statistical validation, methods like ours could be applied to more clinical datasets and for more use cases across the research spectrum. With so much unstructured clinical data generated on a regular basis, the challenge of useful analysis certainly exists for clinical and public health informatics. The era of big data helped push informatics to where it is today and will continue to present unique challenges that demand unique solutions.

Discussion of R Methods

As with any unfamiliar analytics tool, there are many challenges to navigate to achieve optimum results in R. One common challenge within R is ensuring proper data types are set for each object. For text mining applications like this project, it is essential to set the "strings.as.factors" argument to FALSE whenever creating new objects holding the text data. R has some common data types including numerical values, factors (categorical variables), and string data (free text). The default behavior is to read in any string-based data as a factor, which is helpful when dealing with more traditional datasets containing many categorical variables but

will cause issues when dealing with text data. The tibble package allows for the use of the tibble data type, which is a more forgiving data frame object. A tibble can not only support nested lists and data frames but also by default will not coerce string data into a factor variable instead of a character one. This can be beneficial when doing text processing applications in R.

One of the downsides to implementing an analytics solution in R is the relatively high learning curve to using it. Unlike some other programming languages, R is not intended to be compiled into an executable for use by routine computer users. Running scripts developed by others is simple enough, but tweaking and customizing workflows for different data types demands an individual with experience in R. Despite the quirks and learning curve, R is a powerful scripting and processing tool for many types of analytics, including text data analysis. It can provide powerful automation and allows for repeatable data workflows. The open-source nature of the language allows for a huge community of package development to support working with all kinds of data. It is a language well worth the time invested in learning it as it will no doubt continue to be valuable in the statistics and informatics fields into the future.

Limitations

This project was performed retrospectively on clinical datasets not originally intended for research purposes. One of the primary goals of this paper is to describe a method to repurpose unstructured clinical data in a useful way. Many cases were missing values for variable fields. Some cases had much more text data than others, and thus, might be more likely to have key words identified in the string searching. Additionally, misspellings were fairly common in the dataset; however, this was not accounted for in the string matching as it would have been difficult or impossible to anticipate all the possible misspellings of the key words. Negation detection is a big problem in natural language processing and this project is no exception. The concept of negation detection involves advanced NLP systems which can look at words before and after the key word in order to determine the difference between the intended positive match (example: "lacerations are present") and a negative result which would flag as a match in more basic NLP systems (example: "lacerations are not present"). Negation is certainly an issue with this pipeline and is a possible future direction for improving this framework.

In looking at limitations of using this data in a classifier model, the only maltreatment metric available was whether or not a report was filed. Unfortunately, this is the best option we have as the maltreatment case review is extremely confidential information. Additionally, only EMS cases where a report was filed by an NCH staff member were included. Reports may have been filed by EMS personnel or other individuals in the child's

life such as school teachers. These would not have been identified in this study and thus maltreatment may be underreported here.

Conclusion

In this culminating project for my MPH degree program, I worked with a team at NCH to transform a previously unusable dataset into a structured one that can be used for regression analysis and other modeling. This was accomplished through a unique framework we developed which included combining multiple mixed data types from an unstructured, field-generated dataset into a clean, text-based dataset. From this, we were able to review various features such as word counts and frequencies to develop manually created categories for sorting words into sets of key words from which binary variables were derived. Future studies will be developed from this dataset to examine the various risk factors for child maltreatment found in EMS records. Reusable R scripts were created for the NCH team to allow them to process new data from the CDF in this format for future projects. Based on the results of studies performed on this data, new risk factors could be identified. Future directions for this kind of research may include automated screening tools utilizing NLP searching methods. These could be applied to routine clinical notes and EMS notes in real time. This form of clinical decision support has the potential make an impact on the reporting and referral for maltreatment by emergency medical providers.

Through future work on this data we've created, knowledge in the field of child maltreatment will be improved. Child maltreatment is huge public health challenge today and identifying it is half the battle. Maltreatment is incredibly underreported, even among medical professionals. Efforts must be made to improve screening in these settings where intervention is possible. EMS providers are uniquely positioned to observe and report maltreatment problems. With the framework developed in this project, we were able to extract useful features from EMS data, which can provide a unique perspective into the environmental factors surrounding a child maltreatment event. Through the work performed in this project we may be able to make an impact on the way child maltreatment is identified and reported within the emergency medicine setting.

Figure 1 – Combining Case Data

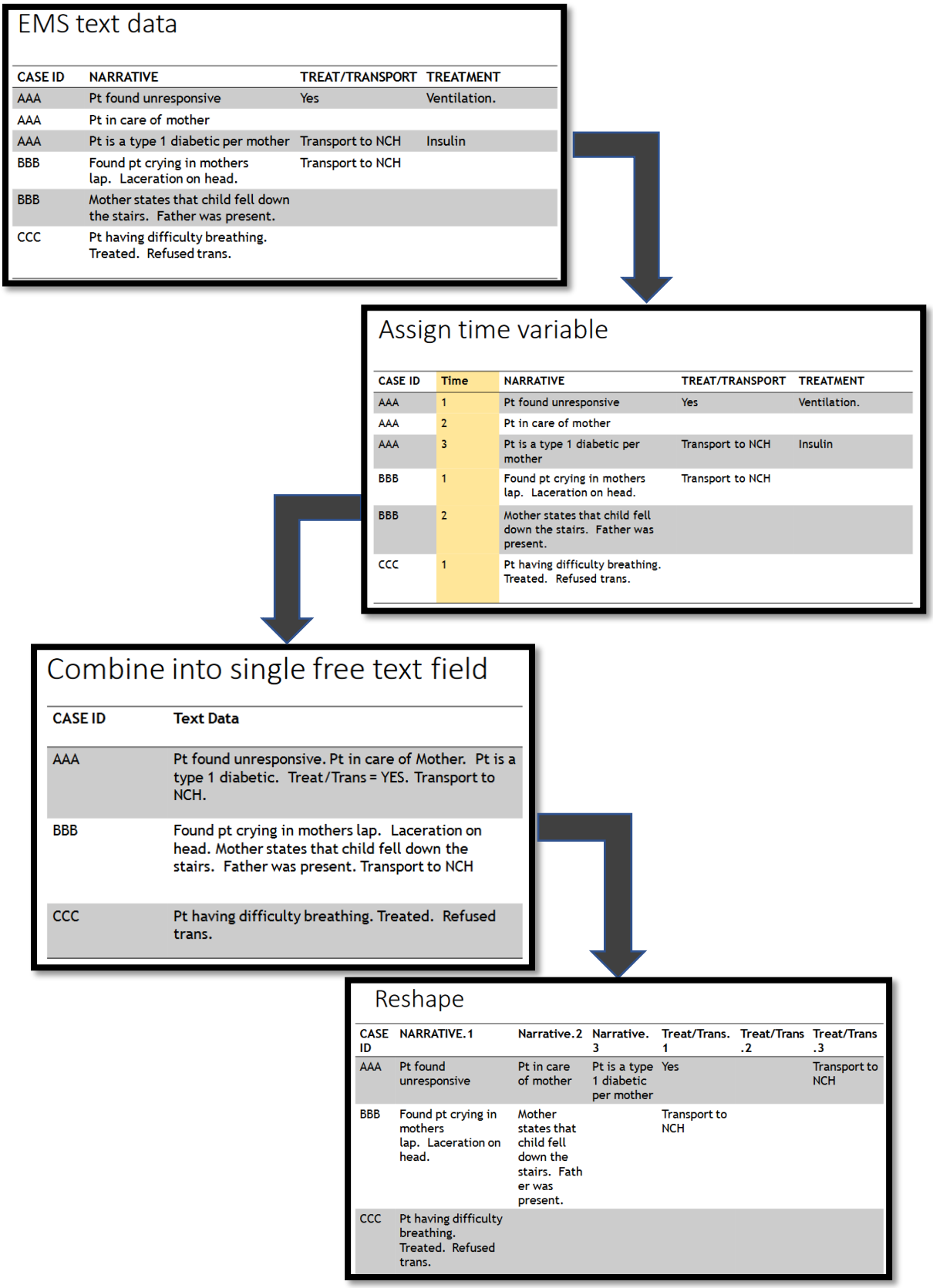

Figure 2 – Data Flow Chart

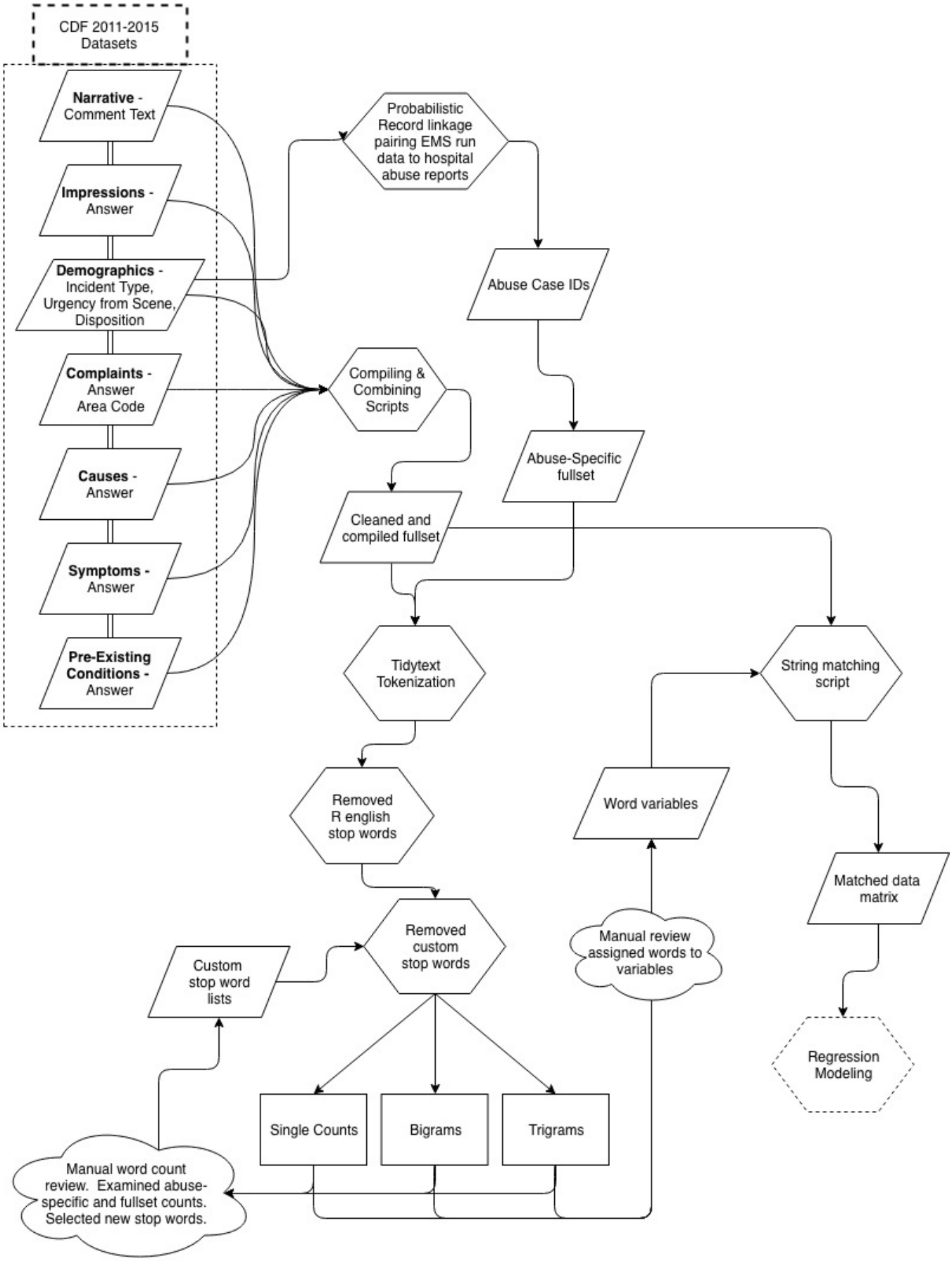

Figure 3 – Example of Word Variables and Keywords

Weapons	Violence	Drug Abuse	Non-Parent	Medical Neglect	Environmental
Gun	Beating	Pills	Boyfriend	Inhaler	Messy
Firearm	Brawl	Medicine_Cabinet	Girlfriend	Prescription	Dirty
Shot	Hit	Poisoning	Aunt	Asthma	Unclean
Knife	Fight	Ingestion	Grandma	Epipen	Smoking
Stab	Fell		Nanny		
GSW	Stairs		Cousin		

Figure 4 – Example of Binary Output Data

Case_ID	Gun	Medications	Non-Parent	Environmental	Medical Neglect
AB1251	0	0	1	1	0
CD1234	0	1	1	0	1
JF3098	1	0	1	1	0
GF2349	1	0	0	0	0

Citations

- Butchart, Alexander, et al. "Preventing Child Maltreatment: A Guide to Taking Action and Generating Evidence." *World Health Organization*, World Health Organization, 3 June 2014, www.who.int/violence_injury_prevention/publications/violence/child_maltreatment/en/.
- Elkin, Peter L. et al. "Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists". *Mayo Clinic Proceedings* Volume 81, Issue 6, Pages 741-748. Published June 2006.
- Emalee G. Flaherty, Robert D. Sege, Tammy Piazza Hurley. "Translating Child Abuse Research into Action". *Pediatrics* Sep 2008, 122 (Supplement 1) S1-S5; **DOI:** 10.1542/peds.2008-0715c
- Meystre, S., & Haug, P. J. (2006). "Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation." Journal of Biomedical Informatics, 39(6), 589–599.
- Murff HJ, FitzHenry F, Matheny ME, et al. "Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing." *JAMA.* 2011;306(8):848–855. doi:10.1001/jama.2011.1204
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). "Natural language processing: An introduction." *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464
- Perlis, R. H., Roberson, A. M., Snapper, L. A., Castro, V. M., & McCoy, T. H. (2016). "Improving Prediction of Suicide and Accidental Death After Discharge from General Hospitals With Natural Language Processing". JAMA Psychiatry, 73 (10), 1064.
- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Shenoi, Rohit P., et al. "Previous Emergency Medical Services Use by Victims of Child Homicide." *Pediatric Emergency Care*, 27 Mar. 2017, p. 1., doi:10.1097/pec.0000000000001079.
- U.S. Department of Health & Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau. (2019). Child Maltreatment 2017. https://www.acf.hhs.gov/cb/research-data-technology/statistics-research/child-maltreatment

Other Computational Tools

- Code from this project is available online
    - Github.com/npondel/mph_culmproj
- Third-party R packages used include
    - TidyR
    - Plyr and Dplyr
    - Tidytext
    - Quanteda
    - Tibble
    - ReadXL
    - StringR
    - RecordLinkage
    - Epiweights