Evaluation of the Use of Child Abuse Consult Service

Emily Powers, Nir Neumark, and Dan Spakowicz

Abstract

In 2013, the Yale New Haven Hospital Pediatric Emergency Department (ED) introduced a tool to help providers identify physical child abuse in infants. This tool has a list of eight injuries which, if observed in a child less than one, instructs the provider to call the Detection, Assessment, Referral and Treatment (DART) child abuse team of physicians and a social worker (SW). To assess how the tool is working, the ED charts are currently being manually and retrospectively reviewed to identify which children have a qualifying injury and whether the DART service and social worker were called. We created custom SQL queries to the Electronic Medical Record (EMR) of Yale New Haven Hospital to retrieve all relevant infant emergency room visits one year before and after the implementation of the DART tool, and deposited the records in a custom SQL database. We next used Natural Language Processing (NLP) and machine learning methods on a manually curated set of records to create models for the identification of the eight injuries and whether the DART consult was called. These methods will facilitate the automated processing of our EMR database and allow for the large-scale assessment of the efficacy of the DART program.

Introduction

Each year an estimated 1.25 million children in the United States are victims of abuse or neglect[1]. This treatment has been cited as a primary cause of maladjustment and poor behavior in later years, as well as a decreased average life expectancy of over two decades[2,3]. Often the only intervention point in the pre-school years is when the infants are brought to emergency rooms; more than 80% of fatalities associated with abuse were under kindergarten age[1]. This visit represents a critical opportunity to evaluate injuries related to abuse directly or during observed during treatment of an unrelated illness.

The identification of abuse is complicated by at least three factors. First, infants and children are inherently clumsy and most if not all injuries can occur by legitimate accidents. Second, parents directly obfuscate cause of injury to avoid legal repercussions. Third, and related to the second point, situations that involve a decision whether to trust a parent narrative are prone to implicit biases, leading some races and classes to be less-heavily scrutinized than others[4]. To alleviate these three complications the DART tool was created that primes providers for warning injuries associated with abuse and mandates consultation with a specially trained team, thereby reducing the decision burden of the primary health care provider in ambiguous or emotionally-charged situations.

The eight warning injuries are based on the Cardiff Child Protection systematic review of abuse literature[5]. Included are (1) fractures of the long bone, ribs or skull; (2) intra-cranial injury (e.g. subdural, epidural or intraparenchymal hemorrhages); (3) burns; (4) solid organ injury by imaging or lab evidence; (5) bruising of the ear, head, neck or torso, including facial, scalp or forehead hematomas; (6) subconjunctival hemorrhage; (7) a torn frenulum; and (8) hemotympanum. Providers are mandated to call a DART consult when any of these eight injuries are observed and document the consultation in the provider note.

Since its implementation in 2013, assessment of the tool has been hampered by lack of structured data for either the eight indicator injuries or the DART consultation. We therefore set out to evaluate the DART program by capturing relevant EMRs and applying methods to learn these parameters from unstructured provider notes, using a manually-curated set of records as the training dataset.

Methods

The initial data set included all children less than 12 months of age with a visit to the emergency department and chief complaint or diagnostic code consistent with injury from April 2013 to June excepting a 3-month washout period from April 2014 - June 2014 after introduction of the intervention. Data were collected via manual chart review from the electronic health record (Epic) on demographics, exam findings, laboratory tests, imaging studies, DART and SW consults, and the ultimate determination of abuse by DART. Using primarily information contained in free text notes, each visit was manually classified as yes/no for each of the 8 injuries with further subdivision by location for the bruising category as well as yes/no for DART and SW consult and as intentional/accidental/indeterminate for ultimate determination of abuse.

Manually collated data were stored in a REDCap database hosted by Yale University in a single table format with some details extracted from notes with respect to injury mechanism but full text notes were not initially included.

In order to perform NLP analysis on the free text notes for automatic identification of injuries and consults, direct access to Clarity (Epic's storage relational database) was obtained. With the help of the Yale Joint Data and Analytics Team, an SQL query was written and executed on Clarity using Microsoft SQL Server 2017 to extract all free text notes associated with the encounters in question. Specific note type "ED Provider Note" served as the basis for NLP.

Analyses were performed using R v3.4[6] using the packages xtable[7], ggplot2[8], pROC[9], RColorBrewer[10], corrplot[11], randomForest[12], e1071[13], tm[14], SnowballC[15], and tidyverse[16]. Training and test data were randomly split 7:3 unless otherwise noted.

This study was approved by the Yale University Institutional Review Board: MODCR00000875.

Results and Discussion

A set of 675 ED provider notes were extracted from Epic and associated with manually collected and categorized REDCap data. If any one of the index injuries was described in the note, the encounter was classified as fitting the tool (fits_matrix variable) and strict adherence to the tool would lead to a DART and SW consultation being called. The observations of each injury ranged from 117 occurrences in the most frequent (htma_facial) to a single occurrence for least frequent three injuries (hemotympanum, ribfx, and br_neck) (Figure 2). Slightly over 50% of the cases observed one or more of the injuries. However, a DART consultation was called in only 64% of cases where an injury was observed (Figure 2, top row). Indeed, in no cases was the observation of an injury perfectly correlated with a DART consult. The injury most likely to lead to a DART consultation was a skull fracture (85% correspondence with DART) and the lowest was hemotympanum, wherein the single observed case did not lead to a DART consultation. Notably, the least predictive injuries of DART were also the most common (htma_facial and htma_frontal) and likely to occur by accident; this suggests that providers are interpreting these injuries as not warranting DART involvement.

We next sought to predict these variables using NLP on the unstructured provider notes. Each note was transformed to lowercase, and then cleaned of numbers, stop words, punctuation and whitespace before being tokenized. Words observed fewer than five times in the corpus were discarded and the remaining set were binarized. Given the low frequency of many of the injuries in the manually-curated dataset we eschewed estimating error on records excluded from training and instead used an ensemble of bootstrapped decision tree classifiers with error-rate estimations from the out-of-bag records. The observation of hemotympanum was excluded due to its large number of missing values and only a single observation record that did not lead to prediction of a DART consultation.

The estimated error of each variable ranged from 0.002 for predicting burns to 0.17 for facial hematomas (Figure 3). Interestingly, the hematomas had the highest error estimations, whereas bruises had relatively lower error estimations despite being phenotypically similar. The wide variation error rates may pose a problem to the goal of automated annotation of some predictor variables, however it is unclear if this will significantly affect prediction of the DART consultation. Since no variables were perfectly predictive of DART and their observation frequencies varied, it is difficult to predict the relative contribution of each variable to whether a DART consultation was called. Moreover, the injuries are correlated to varying degrees, suggesting that only some of the predictors may be needed to predict DART (Figure 4). For example, skull fractures and boggy hematomas were found to be significantly positively correlated, which is expected given that the description “boggy” implies more severe swelling and thus greater likelihood of fracture. Similarly, skull fractures were correlated with intracranial injury. It is therefore possible that future iterations of the DART tool could instruct providers to call for a consultation if significant head injury is observed, of any type.

An unexpected correlation was the association of a subconjunctival hemorrhage

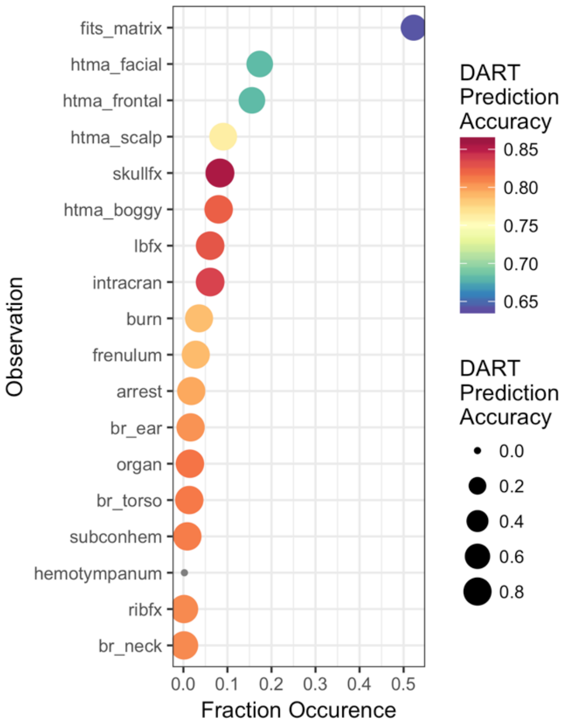

Figure 1. The fraction of patient visits for which the clinical observations were recorded and indicated in the manually-annotated free text notes. Colors and point sizes indicate accuracy of that variable when independently predicting whether a DART consult was called, using different scales.

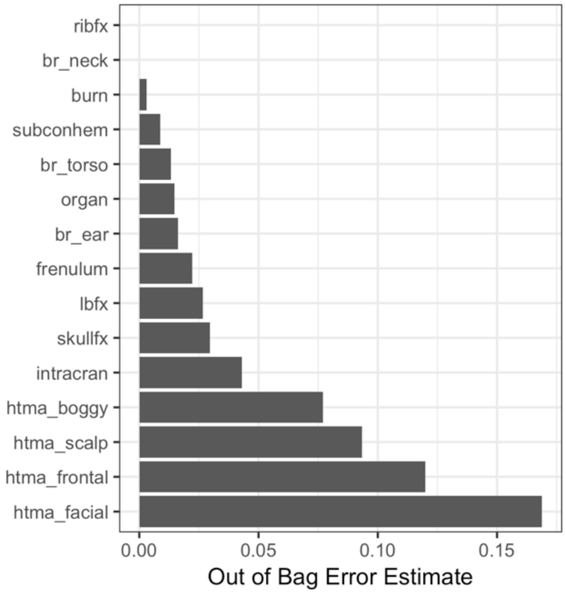

Figure 2. Error estimates for a random forest model for each of the injury variables.

with solid organ injury (Figure 4). Given that subconjunctival hemorrhages are clearly diagnosed by visual inspection, this association may be useful to ensure that providers check for solid organ injury whenever such eye injuries are observed. Similarly, bruised necks were correlated with bruised torsos; bruising on the more visible neck (in some infants) should prompt close inspection of the torso. No significant negative correlations were observed.

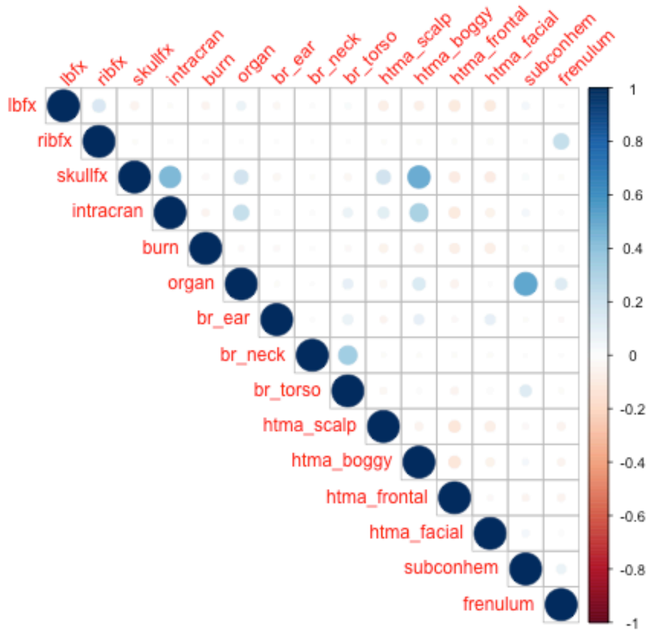

Figure 3. Injury correlations in the manually-curated dataset.

A random forest model for the prediction of a DART consultation using the manually-curated, structured data, as opposed to the NLP methods used for the injury models, showed contributions of relatively few injuries (Figure 5). The most influence both on accuracy and node purity was the presence of a skull fracture, followed by a long-bone fracture and intracranial injury. The remaining injuries had relatively little effect, particularly the facial and frontal hematomas. This suggests that despite the relatively high out-of-bag error estimate for predicting these variables (Figure 3) their overall effect on whether the DART consultation can be inferred may be relatively minor.

While the skull and long-bone fractures were predicted with a low estimated error of approximately 3% using NLP, the model did not include the presence of a radiology report or its associated note, both of which are available in the EMR. It is possible these errors could be further reduced by including such information.

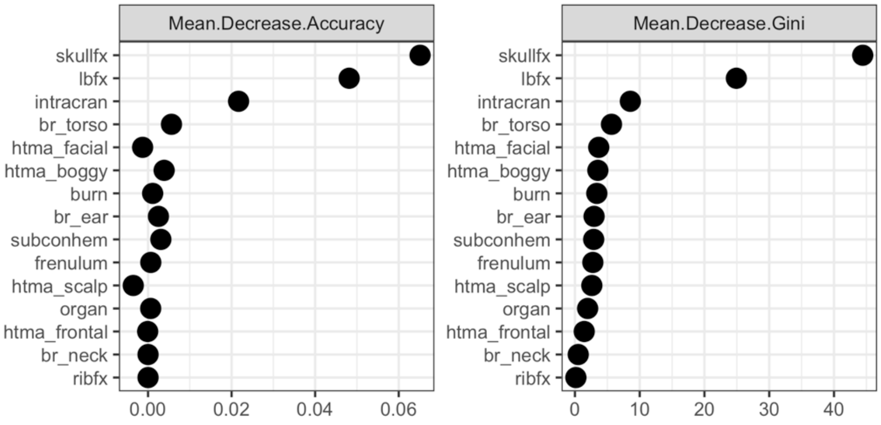

Figure 4. Variable importances for a random forest model of DART consultations.

Prediction of the DART consult was attempted using a variety of methods. The error rate of a model using only the collective sum of the indicator injuries was 35% (fits_matrix, Figure 2; Negative Predictive Value (NPV), FIgure 6), driven by a high false positive rate and low false negative rate. Therefore, the positive predictive value (PPV) using the injury matrix was high, nearly 97%, however its NPV was very low relative to any other method. Support vector machines and random forest classifiers performed similarly when applied to the manually-curated set of structured data, with moderate NPVs and PPVs.

The optimal model was nearly obtained by merely searching for instances of the word "DART" in the provider notes, with the highest PPV and the second-highest NPV. The only model that performed better was a random forest on the notes, which would include the word "DART" among many others.

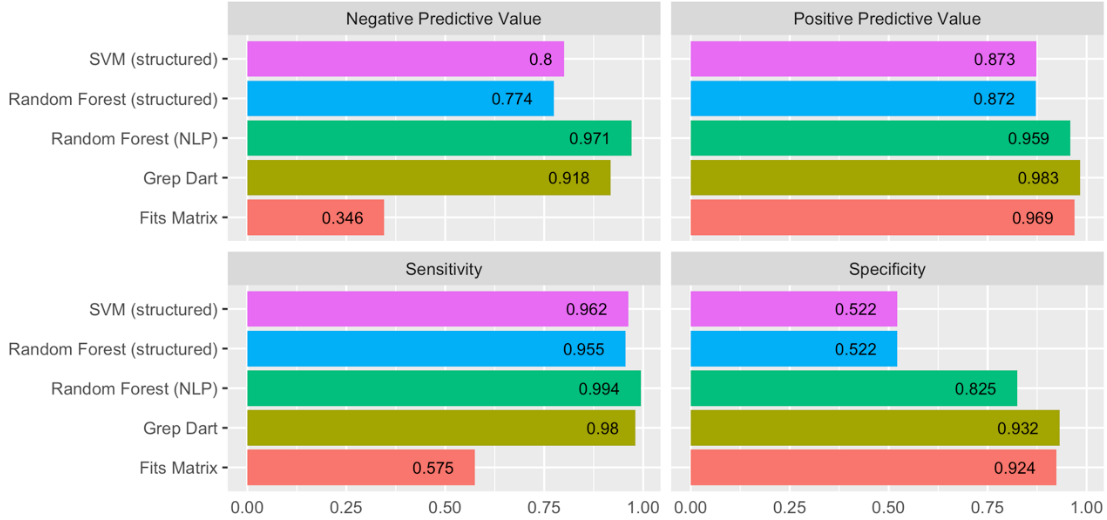

Figure 5. Model comparison for predicting whether a DART consultation was called.

This work demonstrates the capacity to build predictive models for the identification of injuries and consultations within unstructured notes. These models can be applied to the larger set of data that has been captured from querying electronic medical records to provide evidence for the assessment of the DART program and to determine its effect on identifying cases of child abuse.

Acknowledgements

The authors would like to thank Alexandra Signoriello, Kei Cheung, Robert McDougal, Cindy Brandt, Samah Fodeh, Jon Bates and the rest of the instructors of CBB750.

Author Contributions

EP conceived the project, generated the training dataset, SQL queries and database. NN and DS analyzed the data. EP and DS wrote the paper. For detailed code contributions please see https://github.com/mastoneq/thedartmatrix.

References

1. Sedlak, A. J. *et al.* Fourth national incidence study of child abuse and neglect (NIS-4). *Wash. DC US Dep. Health Hum. Serv. Retrieved July* **9,** 2010 (2010).

2. Felitti, V. J. *et al.* Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults. *Am. J. Prev. Med.* **14,** 245–258 (1998).

3. Brown, D. W. *et al.* Adverse Childhood Experiences and the Risk of Premature Mortality. *Am. J. Prev. Med.* **37,** 389–396 (2009).
4. Finkelhor, D. Epidemiological factors in the clinical identification of child sexual abuse. *Child Abuse Negl.* **17,** 67–70 (1993).
5. CORE INFO | Cardiff Child Protection Systematic Reviews.
6. Team, R. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012*. (ISBN 3-900051-07-0, 2014).
7. Dahl, D. B. *xtable: Export Tables to LaTeX or HTML*. (2016).
8. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer Science & Business Media, 2009).
9. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).
10. Neuwirth, E. *RColorBrewer: ColorBrewer Palettes*. (2014).
11. Wei, T. & Simko, V. *corrplot: Visualization of a Correlation Matrix*. (2016).
12. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2,** 18–22 (2002).
13. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Available at: https://rdrr.io/rforge/e1071/. (Accessed: 9th May 2017)
14. Text Mining Infrastructure in R | Feinerer | Journal of Statistical Software. Available at: https://www.jstatsoft.org/article/view/v025i05. (Accessed: 9th May 2017)
15. Snowball stemmers based on the C libstemmer UTF-8 library. Available at: https://rdrr.io/cran/SnowballC/. (Accessed: 9th May 2017)
16. Wickham, H. & RStudio. *tidyverse: Easily Install and Load 'Tidyverse' Packages*. (2017).