Booz | Allen | Hamilton®

TEXT MINING AT THE FOOD AND DRUG ADMINISTRATION

APPLYING MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS

The U.S. Food and Drug Administration (FDA) monitors medical products such as blood transfusions for safety and effectiveness. Booz Allen worked with agency scientists in the Office of the Commissioner and the Center for Veterinary Medicine to analyze clinicians' notes in electronic health records (EHR) of 20,000 historical transfusion cases from the MIMIC-III clinical care database. Machine learning techniques and advanced natural language processing methods were used to turn unstructured notes into valuable information about actual and possible blood transfusion side effects. This text mining framework can help FDA improve clinical care and advance population health.

INTRODUCTION

About 4.5 million patients get blood transfusions each year in the United States. Transfusion side effects, including respiratory and circulatory reactions, are often hard to specifically ascribe to transfusion and can sometimes be deadly.[1] Scientists at the U.S. Food and Drug Administration (FDA) in the Office of the Commissioner and Center for Veterinary Medicine wanted to research methods that could be used to proactively monitor electronic health records (EHR) to discover new safety concerns by reviewing historical EHR data to identify previously unascribed types of reactions for further investigation.[2] Booz Allen worked with the scientists to examine more than 20,000 past transfusion cases identified in the MIMIC-III database of historical critical care admissions in a teaching hospital. Applying advanced natural language processing methods to unstructured clinicians' notes within EHRs, our team of data scientists helped them quickly glean valuable information about actual and possible transfusion reactions, paving the way to a much faster, more direct way of discovering side effects, improving clinical care, and saving lives.

THE CHALLENGE

An FDA core mission is to protect patients from harm by ensuring that medical products such as blood and blood products intended for transfusions are safe and effective. Safety monitoring traditionally involves collecting and reviewing reports—short summaries of events that healthcare providers and patients submit to FDA. While these episodic reporting systems provide the core of critical adverse event information, FDA and other scientists have investigated the additional usefulness of other data,[3] including social media posts, scientific literature, and EHRs. EHRs typically include both coded information and written clinicians' notes about patients, including both observations and conclusions. Because EHR notes generally contain much more information than the coded data, they can capture critical information about side effects that are known, suspected, or not recognized as associated with blood transfusions and other therapies. Our problem is how to find these side effects amid the large volume of raw, unstructured text data in clinical notes.

[1] Hendrickson JE, Hillyer CD. Noninfectious serious hazards of transfusion. Anesth Analg. 2009;108:759–69.

[2] For example, there were cases of lung problems described in the notes that are compatible with the definitions of transfusion-related acute lung injury (TRALI) or transfusion-associated circulatory overload (TACO), but the notes and coded diagnoses didn't mention either of these formal diagnoses. The team is looking for cases that are statistically linked to blood transfusion but not similar to any formally recognized transfusion reaction.

[3] Natsiavas P, Malousi A, Bousquet C, Jaulent MC, Koutkias V. Computational Advances in Drug Safety: Systematic and Mapping Review of Knowledge Engineering Based Approaches. Front Pharmacol. 2019 May 17;10:415. doi: 10.3389/fphar.2019.00415.

THE APPROACH

Using natural language processing, machine learning, and visualization, FDA's and Booz Allen's data scientists developed new ways to clean and organize free-text medical record data, preparing the unstructured notes for input into sophisticated algorithms. This "text mining" unlocked the information-rich portions of the EHRs, enabling scientists to gather actionable insights. Inspired by the advanced text mining methods used to verify the true authorship of literary works attributed to William Shakespeare, the use case was called the "Shakespeare EHRs" project.

Processing the enormous amount of EHRs data required a high level of computing power. Booz Allen provided a cloud-based infrastructure and analytic tools to house and manipulate the data. In addition, as part of the text mining process, the team identified blocks of exactly-duplicated patient notes, which clinicians often copied and pasted from previous entries. Our solution, named "bloatectomy," filtered out the duplicate sections, which, if left untouched, can reduce the overall quality of machine learning. The team then applied machine learning methods to find the 10,000 words or phrases (other than those that simply say the patient was transfused) that best distinguished notes about admissions that involved blood transfusion from notes about admissions that did not involve transfusion.

THE SOLUTION

With usable data in hand, our data scientists then identified terms and keywords that commonly occurred in the notes for the same admission and used them to describe topics. Those, in turn, were used to identify cases where patients experienced actual and possible transfusion reactions. In a matter of hours, our process resulted in an initial automated assessment of more than 20,000 transfusion cases, finding cases of transfusion reactions in the historical data immediately, far more than the 10 cases per year nationwide that were reported over the same years as the clinical data.

In addition, using an open source search and analytics tool, we built an interactive data visualization platform. With this digital dashboard, epidemiologists could quickly search specific terms, view our analyses, and investigate specific EHRs with intuitive graphics and point-and-click filtering right from their laptops.

Another important aspect of the solution is that it is translatable to other types of adverse events. By designing a solution that does not rely on specific interpretations of word meaning, but rather relying on statistical properties of clinical terms, we believe this approach will scale well to other applications. In testing this hypothesis, we're helping FDA scientists look at other applications from population health to individual care. For example, our data scientists are continuing to validate this advanced text mining framework by looking at historical EHRs and attempting to retrospectively uncover adverse events that were identified in the past. The team plans to publish the work in open scientific journals and the code in an open repository. Further work will need to be done to find out how adaptable the methods are for other EHRs data and formats.

In the future with further refinements, the framework could be implemented into near-real-time operations. With a faster, scalable, and more comprehensive review of EHRs text notes, FDA could be more proactive—recalling dangerous products, alerting hospital systems before outbreaks spread, and saving lives.

About Booz Allen

For more than 100 years, business, government, and military leaders have turned to Booz Allen Hamilton to solve their most complex problems. Together, we will find the answers and change the world. To learn more, visit BoozAllen.com.

Contact Us

John Larson
Senior Vice President
Larson_John@bah.com
(571) 882 3332

Lauren Neal
Distinguished Scientist
Neal_Lauren@bah.com
(240) 314-6538