Position Paper:
The value of integrating analytics and visualizations for understanding electronic medical records: Why, when, and which?

Adam Perer

Index Terms—Electronic medical records, Analytics, Visualization

1 Introduction

There has recently been much compelling research demonstrating the effective of visualizing electronic medical records (EMR) to reach novel insights in clinical research and practice of medicine. Electronic medical records often inherently contain structure that may lend itself to visualization. However, since data is usually recorded for billing or claims purposes, and not analysis, it is often not exploited as much as it could be for advancing medical knowledge. That said, the high dimensionality yet sparse feature space of temporal electronic medical data also presents many challenges for visualizing such data. For instance, many EMR systems adopt the International Classification of Diagnosis 9th edition (ICD9) to record clinical diagnoses. In the ICD9 system, there are 14,313 different diagnosis codes, some (e.g., Overweight or Acne diagnoses) may be frequent, and others (e.g., rare genetic syndromes such as Huntington’s disease) may be sparse. Similarly, drugs may be recorded using the United States Pharmacopeia (USP) Model Guidelines, where there are 5,869 unique drug ingredients. Furthermore, when EMRs combine these diagnoses and treatments with other medical data, like procedures and lab tests, it is clear that the event dictionary for EMRs is quite large. Visualizing such a large event dictionary over large amounts of time for many patients is challenging, so many visual interfaces may simplify the range of data by filtered to specific types of diagnoses and treatments chosen by domain experts relevant to specific populations or specific diseases of interest to users. However, relying on domain experts to make these decisions may be challenging as well, given their sparseness and limited time.

In this position paper, I propose that computational analytics can be effective at providing certain types of insights on what to visualize.

2 Why Rely on Analytics?

When analyzing thousands of patients, each with clinical events from an event dictionary of tens of thousands of different event types, relying on visualization alone is often not a scalable solution. The frequency of events among patients differs greatly, as well as the frequency of different event types. Such data properties may make our favorite visualization techniques ineffective.

One solution is to leverage analytics to find statistically significant temporal patterns hiding within the complex data. As an example, we recently developed Frequency [2], a visual analytics system that integrates data mining and visualization in an interactive hierarchical information exploration system for finding frequent patterns from longitudinal event sequences. Frequency features a novel frequent sequence mining algorithm to handle multiple levels-of-detail, temporal context, concurrency, and outcome analysis. Frequency also features a visual interface (Figure 1) designed to support insights, and support exploration of patterns of the level-of-detail relevant to users. In the manuscript [2], a case study illustrates how Frequency has been used by a team of clinicians and clinical researchers to determine if there are particular patterns that lead to patients with lung disease developing sepsis, a potentially deadly medical condition.

3 When Will My Analytics Be Ready?

A potential issue when relying on analytics during exploration is that analytics may not provide their results instantaneously. This is particularly problematic for algorithms that attempt to find patterns in large temporal data, where the algorithms are typically slow. One such solution to this issue is to employ progressive visual analytics [3]. The concept of progressive visual analytics is an idea that analytic algorithms can be designed to produce semantically meaningful partial results during execution. These progressive results can then be integrated into interactive visualizations that allow users to immediately explore partial results, examine new results as soon as they are computed, and perform new rounds of exploratory analytics without waiting for previous analyses to complete.

As an example, we have built a system, Progressive Insights (Figure 2), that uses progressive visual analytics to support analysts searching for common patterns of clinical events. Using this tool, clinical researchers have been able to understand temporal patterns and correlations among patients. For instance, in the case study described in [3], clinical researchers found patterns that have led to clues on how to reduce hospital re-admissions which could be significant to hospital budgets. Without progressive visual analytics, the researchers admitted some of their findings may have never been discovered.

4 Which Analytics Do I Choose?

If one argues that analytics are useful, and their computational time requirements can be overcome, it may still be unclear which analytics to choose. There are countless algorithms and techniques for analyzing temporal events, and this number continues to grow. For this challenge, we return to another use for visualization: making sense out of all of the analytics, so we can ultimately choose the best for the users’ purposes.

Suppose clinical researchers want to use their electronic health records to predict the onset of a disease or which treatments may be the most effective. Predictive modeling techniques are increasingly being used by medical scientists to understand the probability of predicted outcomes. However, for data that is high-dimensional, a critical step in predictive modeling is determining which features should be included in the models. Feature selection algorithms are often used to remove non-informative features from models. However, there are many different classes of feature selection algorithms. Deciding which one to use is problematic as the algorithmic output is often not amenable to user interpretation. This limits the ability for users to utilize their domain expertise during the modeling process. To improve on this limitation, we developed INFUSE, a novel visual analytics system designed to help analysts understand how predictive features are being ranked across feature selection algorithms [1]. INFUSE, as shown in Figure 3, provides a way to visualize an overview of all features and how they contribute to the classification scores of their models. INFUSE has led

* Adam Perer is with IBM T.J. Watson Research Center. E-mail: adam.perer@us.ibm.com.
to important insights for clinical researchers interested in building predictive models, such as determining if a patient is at risk of developing diabetes, a chronic disease that causes serious health complications.

**ACKNOWLEDGMENTS**

The author wishes to thank his co-authors on the referenced papers whose conversations and collaborations helped develop the ideas in this paper, including Fei Wang, Chad Stolper, David Gotz, Josua Krause, and Enrico Bertini.

**REFERENCES**


Fig. 1. Frequence contains a novel algorithm for frequent sequence mining to handle real-world clinical constraints of level-of-detail, temporal context, concurrency, and outcome, and a visual analytics user interface that integrates mining and visualization to support interactive parameterization and exploration to reach insights.

Fig. 2. Progressive Insights features progressive visual analytics, and supports user-driven exploration of in-progress analytics. Partial results from the progressive analytics enhance the scatterplot, list, and tree visualizations without interfering with users cognitive workflow.

Fig. 3. INFUSE is a visual analytics tool that supports users to understand the predictive power of features in their models. Each feature is ranked by various feature selection algorithms, and the ranking information is visualized in each of the three views within the system.