Exploring Clinical Care Processes Using Visual and Data Analytics: Challenges and Opportunities

Vikas Kumar[1,3], Hyunwoo Park[1,4], Rahul C. Basole[1,2], Mark Braunstein[1,2], Minsuk Kahng[3], Duen Horng Chau[3], Acar Tamersoy[3], Daniel A. Hirsh[5,6], Nicoleta Serban[4], James Bost[3], Burton Lesnick[5], Beth Schissel[5,6], Michael Thompson[5]
[1]Tennenbaum Institute, Georgia Tech
[2]School of Interactive Computing, Georgia Tech
[3]School of Computational Science & Engineering, Georgia Tech
[4]School of Industrial & Systems Engineering, Georgia Tech
[5]Children's Healthcare of Atlanta
[6]Pediatric Emergency Medicine Associates, LLC, Atlanta, Georgia

ABSTRACT

Healthcare big data is being widely touted as a potential resource for curbing costs and improving outcomes. However, numerous challenges remain for leveraging this data to its full potential. In this position paper we identify the difficulties that characterize clinical data, based on our experiences working with pediatric asthma data from Children's Healthcare of Atlanta. The specific dataset we explored includes administrative items, medications, lab results, clinical respiratory scores (outcome), timestamps, and demographic information from 5,785 emergency department (ED) visits for asthma exacerbations. We argue that new data and visual analytic techniques are needed that are specifically tailored for solving challenges in healthcare, and we propose characteristics that these techniques should have and give our design rationale. To demonstrate how a tool that embodies these desirable features may be designed, we introduce AsthmaFlow, a prototype interactive visual analytics tool that helps clinicians explore and understand the processes involved in pediatric asthma emergency department care.

Categories and Subject Descriptors

Human-centered computing [**Visualization**]; Information systems [**Data mining**]; Applied computing [**Health care information systems**]

Keywords

Visual analytics, process mining, asthma, emergency care, pediatric hospital

1. INTRODUCTION

Achieving cost effective healthcare is one of the most pressing problems facing society today. Many industry experts point to the incoming tsunami of electronic health record (EHR) data as a possible resource to help solve this problem [5, 13]. However, many obstacles remain for leveraging this data to its full potential. A central question we are exploring is how we can overcome the challenges posed by healthcare data to discover care patterns that produce optimal outcomes at the lowest cost [3].

To illustrate the magnitude of this problem, consider the many types of information that are recorded for a typical patient entering a hospital: demographic information including race, gender, age, payment type, and numerous other variables; a clinical description of the problem, including a chief complaint, history of the present illness, past medical history, review of systems, and physical examination results, each of which may include dozens of structured and unstructured data elements; laboratory tests including bloodwork, other diagnostic tests, and imaging studies; and prescribed medications with specific dosages and instructions. The timestamps of all of these events are recorded in the EHR but this may be done after the fact. These data are also sparse and may even be missing – one patient may have many imaging studes while another has none. The challenge is to analyze and visualize this complex data set to yield insights about clinical care and to inform further investigations for doing so.

We describe some of the issues that we have faced when attempting to understand the underlying clinical processes based on our experience working with pediatric asthma data from the Children's Healthcare of Atlanta (Children's) emergency department. To solve these challenges we require a tool that combines data mining and analytics with effective visualization.

Our main contributions include:

1. Identifying the challenges facing healthcare analytics, specifically when understanding clinical care processes.

2. Proposing data and visual analytic solutions for these challenges.

3. Demonstrating AsthmaFlow, our prototype tool, in the context of possible solutions.

The structure of this paper is as follows: in Section 2 we discuss the factors that make healthcare data difficult to work with. In Section 3 we identify specific features that will be required of new tools for analyzing and visualizing care processes effectively. We also introduce AsthmaFlow. In Section 4 we review previous attempts at using data and visual analytics to understand clinical care processes. Finally, in Section 5 we conclude with future directions and implications for research.

2. CHALLENGES FOR UNDERSTANDING CLINICAL CARE PROCESSES

In order to understand the healthcare analytics techniques that are needed, we first discuss the challenges presented by healthcare data. We acknowledge the substantial body of prior work in machine learning / process mining analytics in healthcare; a comprehensive review is beyond the scope of this paper and interested readers are referred to key work [4, 9, 11, 17]. Our contribution is that we focus on the synthesis of visual and data analytics to help clinicians understand entire clinical processes. This is an important, but sparsely explored, area of research that is relevant to helping clinicians, who will be the ultimate beneficiaries of our work. Below, we highlight the major challenge areas our research aims to address (Figure 1).

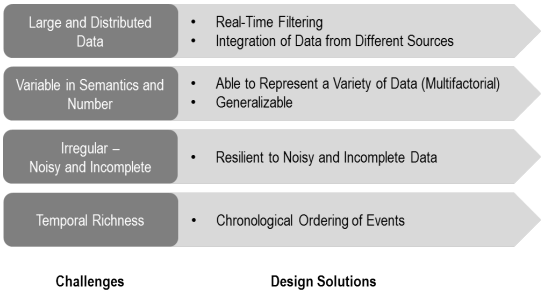

Figure 1: Our proposed challenges and corresponding solutions for visualizing and analyzing clinical care processes.

2.1 Large data

Clinical databases are often quite large, requiring storage space in the petabyte range. They are large not only in the number of patients: the 2011 Mayo Clinic data repository contained over 1.1 million patient registrations [18]—but also in terms of the data elements stored for them, as discussed in the introduction. With this large number of attributes comes the "curse of dimensionality": when dividing patient populations using multiple filters, very small sample sets result that are not amenable to statistical testing. Furthermore, visual analytic tools have a finite array of features (e.g. x-position, y-position, color) with which to represent this multiplicity of variables. Finally, the data may be distributed throughout the electronic records of various departments, including billing, administration, clinical care, and even medical devices. Assembling and organizing these large and distributed datasets can be difficult and time consuming.

2.2 Variable semantics and number

Clinical data is incredibly variable, both *semantically* and *numerically*. Semantically, EHR data contains different types of variables and events. While a particular demographic trait (such as age) can be represented as one variable per patient, events are more complex. An outcome event may be an associated measure (e.g. HgbA1c) and a timestamp. A medication treatment event might require the medication name, dose, form and route of administration along with multiple timestamps representing each time it is given. Administrative events might simply be a timestamp indication, for example, when the patient was admitted. Also there are many different data types: some variables are categorical, some ordinal, some continuous and some in date and time format [4].

Finally, the variables are not fixed in number: there might be a few or dozens of medications administered to a patient during a single admission. This leads to a relational database structure requiring multiple tables, instead of the traditional one-table structure consisting of observations and attributes.

2.3 Irregularity

A third challenge is irregularity, by which we mean both *noisiness* and *incompleteness*. Noisiness refers to the high frequency of errors in the data. This may occur because of incorrect logging of data into the EHR by care professionals, either at wrong timestamps or with incorrect values [11]. Noise is also present in the values themselves; for example a lab test that is repeated multiple times in the same patient will likely return different values.

Additionally, clinical data is rarely complete; the data is recorded at irregular intervals and sometimes it is not recorded at all. The irregular intervals pose a problem for standard time series analysis. For example, consider a diabetic patient who has been hospitalized for high blood glucose. During the visit, the sequence of blood glucose measurements and that of medication administrations will be irregular with respect to time and each other, making it difficult to determine how the medication affects the glucose level over time in that patient. Also, other medications may have been given during the treatment, further confounding the results. In clinical trials this problem is largely circumvented by conducting randomized controlled studies, in which the medications and intervals between variables of interest are strictly regulated while determining the effect that the medication has on outcome. If we are to leverage the data collected in the clinical care environment we require alternative solutions to this problem.

2.4 Temporal richness

Another obstacle that characterizes healthcare data is its rich temporal domain which prevents the breakdown of data into simple rows and features onto which traditional clustering, classification, and prediction tasks can be applied. Instead, a patient may have hundreds of tests performed and medications administered in interleaving sequences. Moreover, the sequences may be of different lengths, precluding a simple item-by-item comparison. Events may be associated with multiple timestamps: for example, a medication event has a timestamp for when it was ordered and one for each time it was administered. In this case, should events be treated as intervals, or as separate "sub-events" (a similar question has been asked in [12])? The data includes not only the relative sequence of events, but also the actual time of each event, something that many current sequence mining algorithms don't consider. Finally, an additional problem is concurrent events [15]. In the clinical setting the care team may submit multiple lab tests and medication orders at once, in various combinations. This makes it hard to form true sequences. Are the resulting "super-events" treated as one big event, or as separate events?

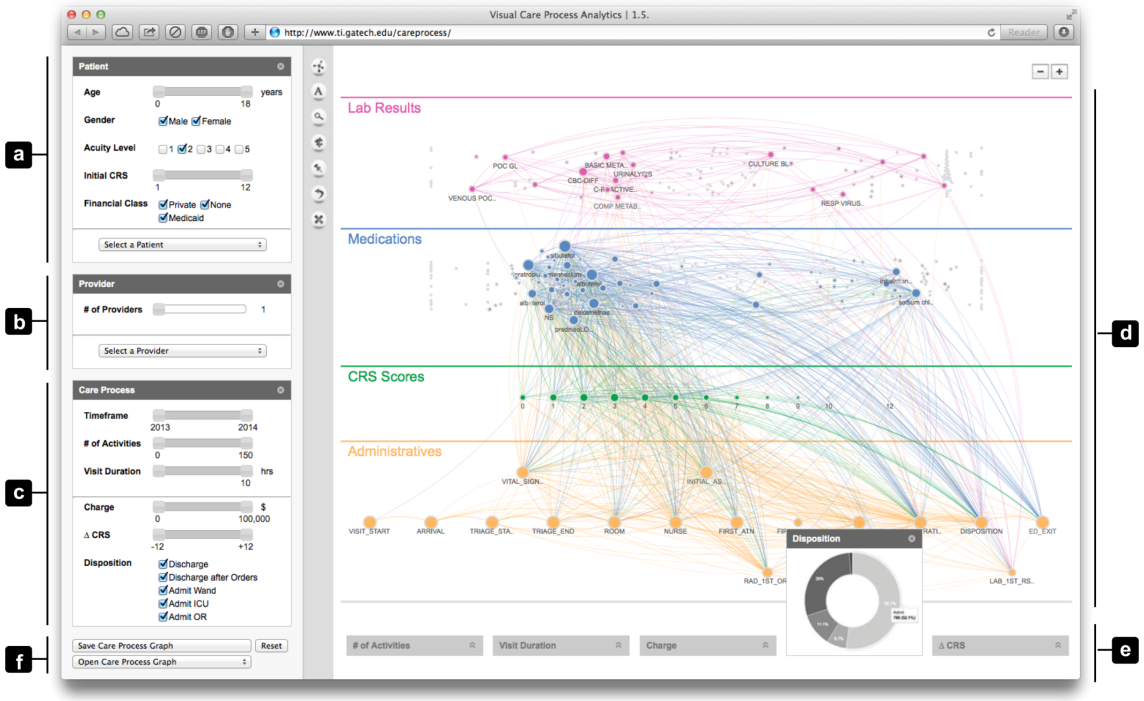

Figure 2: A screenshot of AsthmaFlow. (a), (b), (c) are the filter criteria panels for patient (e.g. age, sex, race), provider, and care process characteristics (e.g. visit duration, charge, disposition), respectively. (d) is the main network visualization area. The clinical events are each represented as nodes, and edges represent the connections between consecutive events in a case. (e) is a collection of analytical tools including histograms and pie charts. (f) allows users to save and load filtered subsets of populations.

3. DESIGN OF ANALYTIC TOOLS FOR HEALTHCARE

Because of the unique and diverse set of challenges outlined earlier, it may be necessary to develop analytic tools that are specifically suited for EHR data. We propose that such new tools have certain characteristics (Figure 1). Previous work [9, 4] is not readily applicable here, since it does not consider certain aspects (e.g. temporal richness) that are present in clinical data.

To exemplify such a tool we introduce AsthmaFlow (Fig. 2), a prototype tool for analyzing and visualizing pediatric asthma care processes in the emergency department. AsthmaFlow was developed in a D3.js [6] environment while storing the data on a lightweight server. We collaborated with Children's Healthcare of Atlanta (Children's), the largest provider of pediatric health services in the nation. Our data from Children's was in a relational database format and included administrative data, medications, lab results, CRS score (outcome) timestamps, charge data, and demographic characteristics for 5,785 ED visits for which asthma was the primary problem.

3.1 Fast data filtering

Because researchers and clinicians may want to test multiple hypotheses in a short period of time or even "on-the-fly", the optimal tools would be able to filter across various patient populations and perform analyses in close to real-time. The interface for AsthmaFlow allows for real-time filtering of patients by demographic characteristics including age, sex, race, triage status, and other variables (see Figure 2a - c).

3.2 Integrative

Data relevant to health analytics is often distributed across different parts of the healthcare system such as the billing, pharmacy, and clinical care departments. Effective tools will need to draw information from these multiple sources in order to put forward all-encompassing solutions. As input, AsthmaFlow receives information from many sources with types ranging from charge data to administrative timestamps to medications administered to each patient.

3.3 Multifactorial

As we have discussed, healthcare contains different types of variables and events. Clinical care analytic tools should have built-in data types that are equipped to represent these diverse variables. The variables should also be categorized accordingly and differentiated from each other in the visual representation. AsthmaFlow divides events into four types: administrative data, CRS Scores (outcome), medications, and lab results (see Figure 2d).

3.4 Generalizable

The best tools and algorithms can also be generalized for many clinical conditions. To facilitate this, the code should not be tailored to specific datasets and data formats but should be highly flexible and adaptive. This may be difficult to achieve in the absence of a national standard for clinical

research data. AsthmaFlow is not specific to asthma: with relatively little effort it can be used for the analysis of conditions such as diabetes or heart disease. To facilitate this, in time we plan to release the AsthmaFlow pseudocode so that other users can generalize it for their own research domains.

3.5 Resilient

Because clinical data is noisy and incomplete, successful algorithms must be able to cope with data irregularity. Many techniques focus less on the relationship between data items than on missing data. Algorithms that are insensitive to data irregularity have been discussed elsewhere [10]. The graph representation used by AsthmaFlow facilitates seeing correlations and causations more directly, which may enable clinicians to more readily see the effects and outcome of noisy and incomplete data.

3.6 Chronological

Given the importance of the temporal dimension in clinical care processes, healthcare analytic solutions should have an explicit representation of time so that, for example, processes can be visualized in their chronological order. In AsthmaFlow, administrative timestamps are organized roughly in chronological order along the x-axis (e.g., patient arrival is represented on the left of the screen, while patient discharge is on the right). This shows both the relationship between individual events as well as their overall order. In the future, the edges of the AsthmaFlow graph may be directed to help visualize the sequence of events for individual cases (see Figure 2d).

4. RELATED WORK

Current methods for mining and visualizing clinical processes can be divided into two groups: those stemming from the sequence mining discipline and those originating from the related process mining field.

4.1 Association rule mining and Sequence mining

We aim to incorporate association rule and sequence mining into our work. Early sequence mining algorithms are defined by the Apriori algorithm, originally proposed for solving the market basket problem [1, 2], but which has been adapted for healthcare by substituting clinical events for purchased items and forming rules linking various medications, tests, and other clinical events to different outcomes. For example, in one study the data of more than 30,000 ICU patients was mined for associations between prolonged ICU stays and various comorbidities and medications [7]; in another study the records of 655 patients with coronary artery disease were mined for associations between various risk factors including age, smoking, and cholesterol level and the blockage of specific coronary arteries [14]; and in a third study association rule mining was combined with visual analytics to explore how diagnostic sequences are associated with the development of sepsis in lung disease patients; the authors of this study adapted the Apriori algorithm so that it could deal with the concurrency and temporal constraints commonly seen in clinical data [15].

4.2 Process mining

Process mining, another related area that may provide effective techniques for visualizing clinical data, is related to sequence mining [19]. Process mining was originally conceived to analyze and visualize event logs from non-healthcare related industries, and so it is not built for handling specific relationships (e.g., medication—outcome) that are seen in healthcare, but process-mining related techniques for visualization may inform our own future techniques for clinical data analysis.

Many studies combine process mining with other data mining techniques, such as clustering methodologies based on Hidden Markov Model transition matrices and k-means clustering to identify regular, infrequent, and outlier clinical cases [16], as well as spectral clustering to identify the major activity patterns in an emergency department of a Greek hospital [8]. Other studies that apply process mining to healthcare are reviewed elsewhere [11].

5. CONCLUSIONS

We have identified obstacles to analyzing EHR data, we have proposed that healthcare analytic tools of the future should have certain traits, and we have introduced our AsthmaFlow tool in the context of these traits.

In the future we will continue to make further improvements to AsthmaFlow so that it can more effectively support the goal of finding care processes that yield the best possible outcomes at the lowest possible cost.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1148903, Children's Healthcare of Atlanta, and the Tennenbaum Institute of the Georgia Institute of Technology. Acar Tamersoy was supported by the Symantec Research Labs Graduate Fellowship 2014-2015.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499, 1994.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[3] R. C. Basole, V. Kumar, H. Park, M. Braunstein, B. Kahng, D. H. Chau, D. A. Hirsh, N. Serban, J. Bost, B. Lesnick, B. Schissel, A. Tamersoy, and M. Thompson. Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. *Journal of the American Medical Informatics Association*, (forthcoming).

[4] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81 – 97, 2008.

[5] D. Blumenthal. Stimulating the adoption of health information technology. *New England Journal of Medicine*, 360(15):1477–1479, 2009. PMID: 19321856.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[7] C.-W. Cheng, N. Chanani, J. Venugopalan, K. Maher, and M. Wang. icuarm-an icu clinical decision support system using association rule mining. *Translational Engineering in Health and Medicine, IEEE Journal of*, 1:4400110–4400110, 2013.

[8] P. Delias, M. Doumpos, P. Manolitzas, E. Grigoroudis, and N. Matsatsinis. Clustering healthcare processes with a robust approach. In *26th European Conference on Operational Research*, 2013.

[9] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23:89–109, 2001.

[10] P. Liu, L. Lei, J. Yin, W. Zhang, W. Naijun, and E. El-Darzi. Healthcare data mining: Prediction inpatient length of stay. In *Intelligent Systems, 2006 3rd International IEEE Conference on*, pages 832–837, Sept 2006.

[11] R. Mans, W. van der Aalst, R. Vanwersch, and A. Moleman. Process mining in healthcare: Data challenges when answering frequently posed questions. In R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. RiaÃśo, and A. ten Teije, editors, *Process Support and Knowledge Representation in Health Care*, volume 7738 of *Lecture Notes in Computer Science*, pages 140–153. Springer Berlin Heidelberg, 2013.

[12] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2227–2236, Dec 2013.

[13] T. Murdoch and A. Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[14] C. Ordonez, E. Omiecinski, L. De Braal, C. Santana, N. Ezquerra, J. Taboada, D. Cooke, E. Krawczynska, and E. Garcia. Mining constrained association rules to predict heart disease. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 433–440, 2001.

[15] A. Perer and F. Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 153–162, New York, NY, USA, 2014. ACM.

[16] A. Rebuge and D. R. Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Inf. Syst.*, 37(2):99–116, Apr. 2012.

[17] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298, 2013.

[18] B. Techentin. Big data and graph analytics in a health care setting. In *Supercomputing*, 2012.

[19] W. M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Publishing Company, Incorporated, 1st edition, 2011.