Designing a System to Predict Inflammatory Bowel Disease Flares Using Machine Learning

Emergent Research Forum (ERF) Paper

Mohamed Abouzahra
College of Business, California State University, Monterey Bay
mabouzahra@csumb.edu

Abstract

This exploratory study aims to design a system that allows physicians to take advantage of the available data and data sources to manage inflammatory bowel disease (IBD). This study will: (1) Explore the sources of data relevant to IBD management including EMR, wearable devices, and mobile personal health records. (2) Design a system that interfaces with these data sources to capture and store data in an IBD data warehouse. (3) Develop the algorithms necessary for data cleansing and for applying descriptive and predictive analytics to provide physicians with relevant data to predict future IBD flares. (4) Design a system interface that is easy to use and that can be integrated into physicians’ workflows. This study will inform the development of a system that enhances patients’ quality of life and reduce cost associated with patients’ hospitalization and medication.

Keywords

IBD, Machine learning, predictive analytics, healthcare analytics

Introduction

Recent years have witnessed a remarkable growth in the digital transformation of healthcare. This growth stemmed from the call for new ways to enhance healthcare quality and reduce costs (Black et al. 2011). A useful side effect of the use of information technology is the growth of healthcare data in terms of velocity and volume and the availability of these data to decision makers. Despite the availability of data from sources such as electronic medical records (EMR), lab results, wearable devices, and personal health records (PHR). There is a lack of interoperability between different systems making them silos of useful but inaccessible information.

For example, one of the main healthcare information systems widely adopted by healthcare organizations is EMR. EMR provide the means to acquire and store clinical data that support clinical care whilst facilitating clinical research, practice audit and quality improvement (Hillestad et al. 2005). However, currently there is a multitude of different EMRs which vary with respect to their structure, functionality and data entry requirements; as a result of this and different design decisions, the data acquired and stored by these EMRs will differ very significantly (Black et al. 2011). These problems constitute an important barrier to standardization of clinical care and the application of evidence-based principles to clinical practice at individual centres; these problems are compounded when one considers other sources of healthcare data such as wearable devices. Attempting to standardize all these technologies would be very difficult and expensive to implement. Furthermore, it would be very restrictive and would hamper the developments needed to adapt to changes in our understanding of diseases and their management.

An alternative solution, to be explored in this study, is to develop a system that allows the extraction of comparable data from different healthcare systems and the organization of these data in a standardized, structured format for subsequent, aggregate analysis. Based on the HL7 v3 specifications (Eggebraaten et al. 2007), it is possible to develop a disease-specific system that will collate clinical and demographic EMR data as well as other sources of data such that these data can be evaluated in the context of published. In this study, we aim at designing such a system for inflammatory bowel disease (IBD).
System context

IBD is a chronic disease that results in inflammatory conditions that affect the gastrointestinal system. These conditions often occur unexpectedly in the form of “flares” leading to significant reduction in patients’ quality of life (Podolsky 1991). IBD has also been associated with colon cancer (Kaser et al. 2009). There is currently no permanent cure for IBD patients (Podolsky 1991). In the United States, more than three million patients are affected by IBD (Dahlhamer 2016).

IBD was chosen as a context for this study because of its significant impact on both quality of life and healthcare cost in the United States as well as its complexity. Therefore, IBD represents an interesting area of applying analytics.

Several studies have attempted to use data to predict patients’ flares and to alleviate the effects of these flares. The most prominent study in this area includes the use of EMR longitudinal data to predict IBD flares in veteran patients using machine learning (Waljee et al. 2017). While this study was successful in predicting flares with almost 80% accuracy, it focused mostly on EMR quantitative data. Using other sources and forms of data may improve this accuracy considerably given the expected role of external environment in IBD flares. The other prominent study involved integrating EMR data with patients and care givers input to improve IBD care for children (Ramsey et al. 2015). While this study utilized data from multiple sources, it did not take a quantitative approach to predict IBD flares, rather, it focused on improving patient care. However, this study shows the potential of integrating data from multiple sources to improve IBD care.

Related work

EMR and PHR

EMR systems are healthcare information systems that contain a comprehensive set of patient data including for example, medical history, allergies, demographic data, and medicine prescriptions (Hoerbst et al. 2010). The goal of EMR systems is to have patients’ data readily available to guide patient treatment and support physicians’ decision making. Despite the validity of EMR benefits, these benefits are not always realized because of the lack of interoperability between different systems. This lack of interoperability can lead to silos of patient information and prevent physicians in one hospital from using patients’ medical history stored in a different hospital (Blumenthal 2010). A second reason behind the ineffectiveness of EMR systems is the enormous amount of patient information available to physicians. This information is not always relevant to current patient condition and hence the physician may have to spend more time going through this information to make the right decision.

PHR systems are systems developed to enhance patients’ involvement in their care plans. They allow patients to record their activities, symptoms, and vital signs and allow physicians to access these systems, upon patients’ authorization. Integrated EMR and PHR systems can improve communications between patients and providers and create a centralized repository of health data (Archer et al. 2011). However, PHR information is entered by patients, and therefore, this information may be of lower quality than EMR data limiting the ability to use it in decision making (Archer et al. 2011). The growth in smartphone uses has encouraged the creation of mobile PHR (mPHR) which allow patients to enter information more frequently, and provide them with the interface to enter information in a more reliable way (Archer et al. 2011).

While several studies have used EMR data to predict patients’ outcome (Waljee et al. 2017), few studies have used the integrated data from EMR and PHR together. Using both types of data is valuable because together, they would provide a rich dataset that encompasses providers’ view of the patients’ condition as well as the patients’ day-to-day data.

Wearable devices in healthcare

Wearable devices are devices that can be worn by individuals to monitor their activities continuously without limiting their ability to interact with their environment (Gao et al. 2016). Wearable devices use in healthcare has been growing significantly in the past few years. Modern wearable devices extend beyond
measuring activities to measuring vital signs such as heart rate, skin temperature and blood pressure (Haghi et al. 2017). The main issues with the use of wearable devices are that sometimes, their data are not very accurate (Haghi et al. 2017) and they are difficult to integrate with other healthcare sources because they do not share the same integration standards.

While wearable devices and internet-of-things represent an active research area, and while they have been integrated with EMR system before, there is scarce literature on the efficacy of this integration and on the application of analytics on the data resulting from the integration.

**Data analytics in healthcare**

Data analytics aims at extracting useful knowledge from data. In healthcare, data analytics has attracted a lot of attention and has been used in different areas such as detecting epidemic trends (Ramsey et al. 2017), associating specific patterns with diseases, and predicting patients’ future condition. The data source for most of these applications is EMR systems. For example, mined knowledge from EMR systems has been used to predict high risk patients and to predict treatment outcome for HIV patients (Tang et al. 2009). The data mining techniques in healthcare adopted a wide range of techniques including clustering, grouping patients with similar symptoms; association, for example connecting certain parts of patient’s history with her current condition; and predictive models that aim at predicting possible future conditions or treatment outcomes based on models rooted in patients’ data. For inflammatory bowel disease, data analytics has not been used frequently. An important study, data analytics was used to predict flare with 80% precision using EMR data. This study implemented several analytics algorithms including logistic regression and longitudinal random forests (Waljee et al. 2017).

In this study, analytics is extended beyond EMR data to include other forms of data (Miotto et al. 2017) including wearable devices and PHR. One of the goals of this study is to develop novel algorithms that will use the integrated data to predict patients’ outcomes.

**System Design**

Figure 1 below shows the conceptual system design. The Data collected by healthcare providers in EMR, by patients in mPHR and through wearable devices are cleansed and anonymized using an anonymization algorithm. The data are then stored in an IBD database. Machine learning algorithms are then applied to IBD data and the resulting knowledge will be stored in a knowledge warehouse. Finally, the physicians will obtain the system outcome through a clinical decision support system designed to accommodate physicians’ workflows.

**Exploring and integrating data sources**

This phase starts with exploring different sources of IBD related data including their platforms, the data they collect, and integration requirements. For EMR, we will examine the integration protocols of these
Predicting IBD Flares Using Machine Learning

In addition to the interface requirements, we will also examine the quality of data in the EMR system. That includes common quality issues such as missing data, typos, incomplete data, different abbreviations related to IBD.

Similarly, for mPHR, the study will examine existing IBD mobile applications, their platform, the type of data they collect, their ease of use, and how they can be integrated with the designed system. While developing an IBD application is possible, there are currently several useful applications that will be suitable to use in this study. However, since most IBD applications export their data as files, we will need to develop the mechanism to automatically capture and send these files to the system. On the system end, we will develop the interface to get data from the files, clean the data, and integrate it with other data sources.

**Designing the IBD data center**

This study will develop the scheme needed to accommodate the captured data including structured and unstructured data. Because of the different data structures from different sources, designing the data scheme will not be a trivial task. It will involve selecting the technology that fits system needs, the data fields required from each data source and the relationships among these fields.

**Machine learning and predictive analytics**

Using machine learning to predict IBD flares is an important and challenging part of this study. This challenge stems from the multitude of data sources and hence data types included in this study. While several studies in the past have used machine learning in healthcare (Choi et al. 2016), most of these studies focused on using quantitative EMR data such as vital signs and lab results (Choi et al. 2016). Very few studies extended machine learning to unstructured EMR data such as physicians’ notes. This lack of unstructured healthcare data use is due to the quality issues usually associated with physicians’ notes. Unstructured data requires extensive efforts to acquire and prepare the data for processing. However, using unstructured data is likely to improve the developed model predictive power (Miner et al. 2012).

In this study, we use the following data types: (1) patients’ longitudinal quantitative EMR data including vital signs and lab results; (2) Qualitative EMR data including physicians’ notes and medication; (3) Quantitative data from wearable devices including activity, calories, sleep pattern, and heart rate; (4) Qualitative data from PHR including dietary habits, water consumption, and psychological conditions.

The machine learning process will start by exploring different data sources with a focus on data quality. Initial findings show that different data sources have different quality issues. For example, while the most common data quality issue in wearable devices is missing data, in EMR, the most salient issues are typos and abbreviations. The next step would be to apply machine learning to predict flares. Techniques such as convoluted neural networks, Bayesian analysis, and support vector machines have been used before and proved successful in predicting patients’ outcomes (Clifton 2016). This study will explore those techniques in the case of IBD, will compare their predictive power, and will develop an algorithm that effectively predicts IBD flares.

**Privacy and security**

Individuals are mostly concerned about the privacy of their medical data (Angst et al. 2009). Threats to privacy increase as more datasets are integrated together. Hence, privacy is an important design factor in healthcare analytics systems and in this study. This study addresses patient privacy on several levels: (1) Anonymization: This study will explore anonymization algorithms to maintain privacy while providing the level of details required for analytics; (2) Patient control: As a system design component, the study will explore ways to empower patients to decide on whether or not to include their data in the system (Patients’ consent); (3) Access control: The study will explore technologies and workflows to manage access level to the system. This will include using technologies such as federated identity or open authentication, and will also include rules and auditing capabilities to prevent unjustified access.
Conclusion

The goal of this paper is to provide a conceptual system design to integrate multiple healthcare data sources and apply machine learning algorithms to predict IBD flares. The study discusses the different components of the system including integrating with different data sources, machine learning, and database design as well as design considerations such as patients’ privacy and workflow considerations. The proposed system would allow healthcare providers to predict future flares in IBD patients and help them cope with these flares hence improving their quality of life and decreasing healthcare cost related to medication and hospitalization. The next step, once this study concludes, is to build the system and assess its efficacy and use by healthcare providers. This assessment will not only improve healthcare outcomes, but will provide useful insights into physicians’ perceptions of smart and artificial intelligence systems.

REFERENCES


