Protecting Patients from Unnecessary Emergency Room Visits and Hospitalizations: Harnessing Big Data to Directly Improve Clinical Care at UCSF

Team: Alvin Rajkomar, Sara Murray, Joanne Yim

Background

Many believe that the era of Big Data will enable precision medicine through the integration of high-dimensional molecular data with digital clinical data. Increasingly, molecular markers will serve as digital fingerprints, which when analyzed with modern computational methods can shed light into personalized predictors of response to targeted therapies. The premise is that large swaths of data are dormant, waiting for the launch sequence of asking the right clinical question, gathering appropriate data and applying suitable computational techniques. Instead of waiting for a million genomes to be collected over the coming decade, we believe that there is a multitude of unharnessed data currently awaiting us in our electronic health record (EHR) that we can use to improve care for patients at UCSF over the coming year.

For example, imagine a young pregnant woman who emails her obstetrician in the morning that she is experiencing a fever and contractions. If the obstetrician does not check her messages until after clinic closes in the evening, that message would sit unread for hours. In a Big Data world, an algorithm should automatically recognize that the woman is pregnant and is experiencing high-risk symptoms that require immediate attention and mobilization of resources to prevent adverse events – like pre-term labor - as soon as possible.

A major challenge of developing such a system is pre-defining all the rules which would indicate a high risk of adverse events. Given the combinatorial explosion of possible diseases and descriptions of symptoms, it would be impossible to manually elucidate the conditions meriting immediate attention. Fortunately, machine learning allows us to accomplish this daunting task by using algorithms on previously collected Big Data to learn which features predict a possibility of a pre-term labor or hospitalization – a process referred to as predictive analytics.

Study Overview

We will employ machine learning on millions of records collected in the EHR at UCSF to predict which patients will require emergency services or hospitalizations before they happen. As many as 70% of emergency room visits may be preventable, with a proportion of these resulting in hospitalizations that may have also been avoided. These preventable escalations of care, which we define as decompensations, cause personal and financial stress to patients and use of costly services by health systems which are increasingly focused on high-value care. There is a pressing need to be able to identify patients at greatest risk for imminent decompensation, with the intent on intervening prior to the need emergency or hospital care.

Our hypothesis is that there is a set of escalations of care that can be predicted, and that a subset of those can be avoided with early intervention by the health system. For example, a nurse might call the patient to give further instructions of how to dress a wound whose dressing has fallen off.

Much of the assessment of clinical status and risk of decompensation is contained within the clinical notes as unstructured free text (e.g. “The patient is calling to report worsening fever and chest pain.”). Therefore, algorithms must not only able to quickly gather information about patients but also draw upon modern machine learning techniques to extract meaning from unstructured data to assess patient status. **Here we propose developing a novel algorithm that leverages the EHR to perform real-time prediction of an individual patient’s risk of emergency room visit or hospitalization within the next 7 days.** Ultimately, our plan is for this algorithm to be implemented in the healthcare system and improve outcomes for our highest risk patients across UCSF Health.

Project Plan

Our plan has three distinct components: building a machine learning model on structured data, deploying the model, and testing the efficacy of adding features from basic natural language processing to the algorithm. Given our team's extensive experience working with clarity data and building predictive models, we believe our project is feasible within this time period.

1. Predict emergency department visits and emergent hospitalizations using structured data in the electronic health record

Here we will assemble a dataset of structured data about patients, including health care encounters, medications, and laboratory values and building and testing a variety of machine learning models to predict decompensation. We will apply support vector machines, linear models, regression trees, and deep neural-networks to assess predictive accuracy.

Our entire research team is Clarity certified and has direct access to the electronic health record database, and AR and YM have already built a machine learning model on a subset of the structured data of health care encounters and medications for the accountable care organization. Therefore, the addition of further structured data from the electronic health record, including laboratory values, is a mere extension of our current work and does not require outside report writers. The entire research team has experience with building machine learning algorithms with the techniques we propose to use. <u>We will test the success by using standard machine learning methodology</u>: we will develop the model on historical data from 2012-2014, using cross validation to help determine hyper-parameters required by the learning algorithm. Once a model is proposed, we will see how accurately the predictions perform on data from 2015, defined by discrimination of area-under-the-curve (AUC) statistics and calibration – the actual rates of emergency room utilization and hospitalizations compared to the predicted rates in different strata of risk.

2. Deploy our novel machine learning algorithms on a UCSF server that integrates with real-time clinical data

Here we will create a pipeline that feeds live clinical data into a computational server that runs this model in order to continuously deploy our models in real-time. We will build on the approach we have already developed for our current algorithm with the accountable care organization and modify our custom software to accommodate the algorithm that we will create in Aim 1.

AR and YM have already deployed a machine learning algorithm on UCSF Medical Center servers that use industry standard or open-source programming tools which feed results to the Office of Population Health. We will use these same tools to deploy the optimal algorithm developed in the first section. The deployment includes running SQL code to pull data daily from the Clarity server to a secure computational server, which will transform the raw data into data that can be used by the machine learning algorithm. The server will then run the algorithm and send the results back to the Clarity server, which can then be used by the health system using standard reporting tools. This will ensure that any provider at UCSF can use the results of the algorithm without requiring any special tools or equipment, and since the data lives in the electronic health record database, it can even be pushed to clinical workflows in the live EHR.

3. Automate clinical text processing tools to extract information from the free text of clinical notes that augment the accuracy of the model

Clinical notes contain detailed information about patients, but given the heterogeneous ways a single condition can be documented, it is challenging to manually extract data that can be used by machine learning algorithms. However, we believe that the information contained in the clinical notes has the potential to greatly improve the predictive accuracy of our models, such as in the preterm labor example above.

While analyzing raw clinical notes is a daunting task, there is pre-existing machine learning software that we will employ to generate preliminary analyses within the time frame of this project proposal.

SM has already developed the programming code to extract and concatenate all of the clinical notes necessary for analysis in a lupus population that can be easily adapted for this project. We will limit our analysis to the primary care population, where patients have had at least one primary care encounter within the last year. For that population, we plan to extract all clinic notes stratified by context (eg. Office visits, telephone encounter, MyChart messages). We plan to stratify by this method because we anticipate that patients may describe their symptoms differently than healthcare professionals, and this will allow improved performance of text feature extraction by our algorithm. Given that there will be over a hundred-thousand features, we will use deep-learning (a special type of machine learning) to reduce this to a smaller number of features that we will analyze in combination with the structured data-elements from section 1. We will then re-tune the algorithms to see how predictive accuracy (defined above) changes.

Future directions and implications:

Because this work is meant to improve care and reduce healthcare costs, the Office of Population Health would like to adopt the results of the algorithm in the workflow of teams responsible for high-risk populations. Once we have developed these algorithms to accurately identify high-risk patients, they will plan to support interventions to potentially reduce the risk of emergency room visit or hospitalization. Long-term, we plan to seek larger grant funding to study potential interventions derived from our machine learning algorithms and the effect on patient outcomes.

Currently, UCSF lacks institutional expertise in both machine learning and more specifically processing of unstructured data such as free text from clinical notes. In the short term, we anticipate that streamlining these machine learning tools will greatly benefit researchers across the medical center, as lessons learned (particularly from the free text processing, a highly desirable tool) can be shared with other researchers. The fundamental premise of real-time processing of clinical data at scale has multiple applications for the Department of Medicine and UCSF Health, as the same pipeline could be used to predict an infinite number of outcomes that are important to our health system and the patients we serve.

This project embodies a highly novel application of health IT in which we synthesize big data - nearly the entirety of the EHR - and use the machine learning to directly improve clinical care. Ultimately, our vision is to combine clinical expertise with big data to improve patient outcomes for UCSF Health.

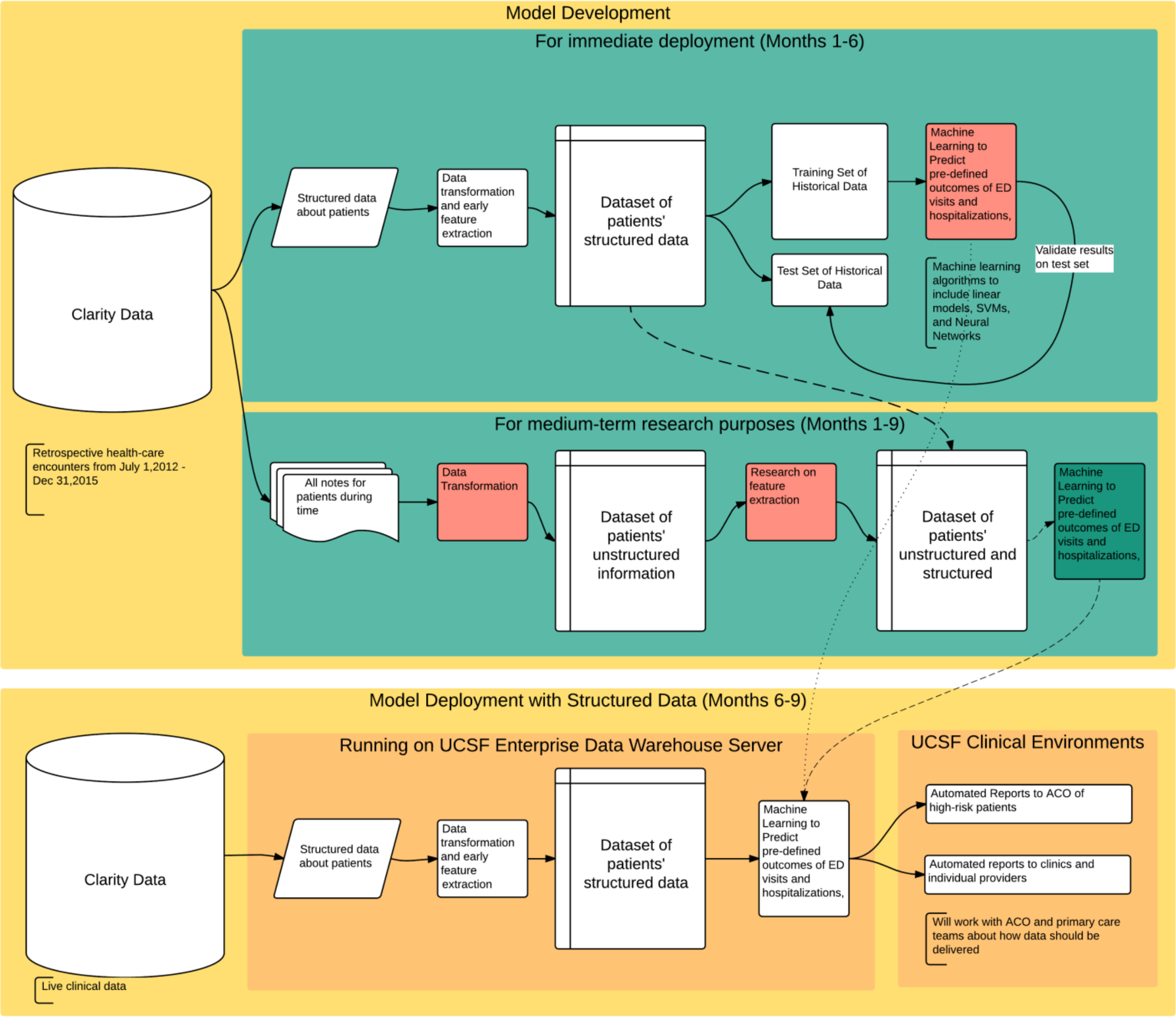
Table: Timeline for project

	April	May	June	July	Aug	Sept	Oct	Nov	Dec
Revise our currently existing data pulls to include all structured data necessary for project	X	X	X						
Employ and validated retrospective machine learning algorithms of structured data		X	X	X	X				
Integrate our machine learning algorithms with real-time clinical data on UCSF server				X	X	X			
Validate text processing tools and revise algorithms to include structured AND unstructured data				X	X	X	X	X	X

Figure: Description of work-flow

Top Panel: The model development phase will be separated into a phase with structured data from Clarity, the database of the EHR, and unstructured data. The core work funded by this proposal are high-lighted in the red-boxes.

Bottom Panel: Once the algorithms are developed they will enter a new pipeline of data processing that will live on UCSF servers that will feed back to the clinical system.


Model Development
For immediate deployment (Months 1-6)
Clarity Data
Structured data about patients
Data transformation and early feature extraction
Dataset of patients' structured data
Training Set of Historical Data
Test Set of Historical Data
Machine Learning to Predict pre-defined outcomes of ED visits and hospitalizations,
Validate results on test set
Machine learning algorithms to include linear models, SVMs, and Neural Networks
Retrospective health-care encounters from July 1,2012 - Dec 31,2015
For medium-term research purposes (Months 1-9)
All notes for patients during time
Data Transformation
Dataset of patients' unstructured information
Research on feature extraction
Dataset of patients' unstructured and structured
Machine Learning to Predict pre-defined outcomes of ED visits and hospitalizations,
Model Deployment with Structured Data (Months 6-9)
Running on UCSF Enterprise Data Warehouse Server
Clarity Data
Structured data about patients
Data transformation and early feature extraction
Dataset of patients' structured data
Machine Learning to Predict pre-defined outcomes of ED visits and hospitalizations,
UCSF Clinical Environments
Automated Reports to ACO of high-risk patients
Automated reports to clinics and individual providers
Will work with ACO and primary care teams about how data should be delivered
Live clinical data