6.S897/HST.956 Machine Learning for Healthcare

Lecture 4: Risk stratification Using EHRs and Insurance Claims

Instructors: David Sontag, Peter Szolovits

1 Risk Stratification

1.1 What Is It?

At a high level, risk stratification is way of separating a patient population into one of 2+ categories (e.g. separating into patients with high-risk, low-risk or in-between). The reason for risk stratification is to act on these predictions and couple those predictions with known interventions. For patients in the high risk pool, we would attempt to do something for them to prevent whatever outcome is of interest from occurring.

Risk stratification is **quite different** from diagnosis. Diagnosis has a highly stringent criteria on performance. A misdiagnosis could lead to severe consequences like the patients being treated for conditions that they don't have, or patients dying because they were not diagnosed in time. The performance characteristics of risk stratification are different, instead looking at quantities such as positive predictive value (PPV). In today's economic environment, the goal of risk stratification is reducing cost in the healthcare setting and improving patient outcomes.

Definition 1 (Positive predictive value or PPV)**.** *Fraction of patients that were predicted to be high risk and are actually high risk.*

Data used for risk stratification is often different from diagnosis and very diverse. Things you might use include multiple views of the patient or auxiliary data such the patient's demographics or socioeconomic information that would highly affect their risk profile but unused in an unbiased diagnosis of the patient.

1.2 Examples

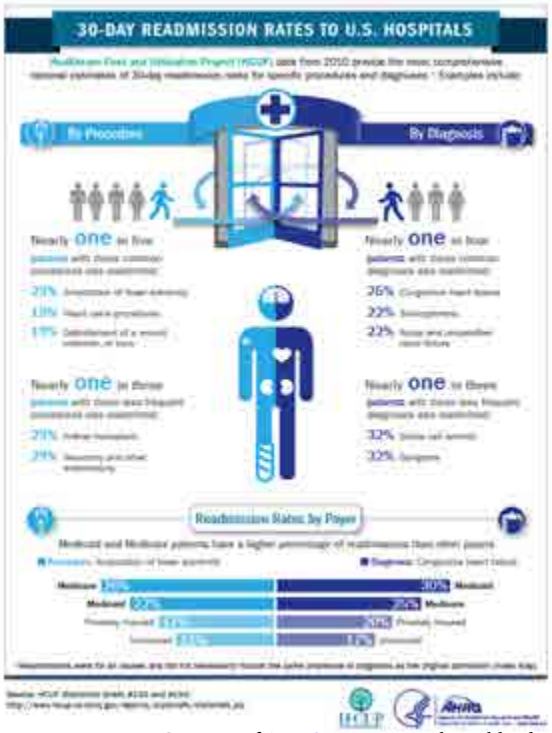
1.2.1 Predicting preterm infant's risk of severe morbidity

The outcomes of premature babies have dramatically improved over the last century. Of the many different interventions that led to this improvement, one of them was having a very good understanding of a particular infant's risk level. A very common score that is used to try to characterize risk for premature infant is the Apgar score [Apg66], but this metric is not as accurate as it could be. Saria et. al uses a machine learning approach to really improve our ability to predict morbidity in infants [SRG+10].

1.2.2 Predicting if patient needs to be admitted to coronary-care unit (CCU)

For patients who coming into the ER with a heart-related condition, the question is: should they be admitted to the CCU, or is it safe for them to be discharged and managed by their physician or cardiologist outside the hospital? A study was performed in 1984 using over 2000 patients, nontrivial amount of variables and logistic regression to predict such cases [PDS+84]. The goal was cost-oriented, as identifying patients who are not high-risk and don't have to be admitted to the CCU leads to reduction in the costs associated with CCU admissions.

1.3 Predicting likelihood of US hospital readmission



Courtesy of AHRQ. Image is in the public domain.

Figure 1: Infographic on 30-day readmission rates to US hospitals.

The US government has imposed penalties on hospitals that have a large amount of patients who had been released from the hospital but within the next are re-admitted in the next 30 days. This is part of the transition to value-based care and receiving a lot of attention. The premise is that there are many patients who are hospitalized, but not managed appropriately on or after discharge. There might be poor communication between the hospital staff and the patient about what to do after discharge, leading to poor outcomes. Predicting which patients are likely to be readmitted before even being discharged could lead to better discharge practices and reduce the likelihood of readmission. For example, the hospital could send a nurse to go slowly through discharge instructions or follow up at the patient's home over the next few weeks.

1.4 Old vs. New

1.4.1 Traditional Approaches

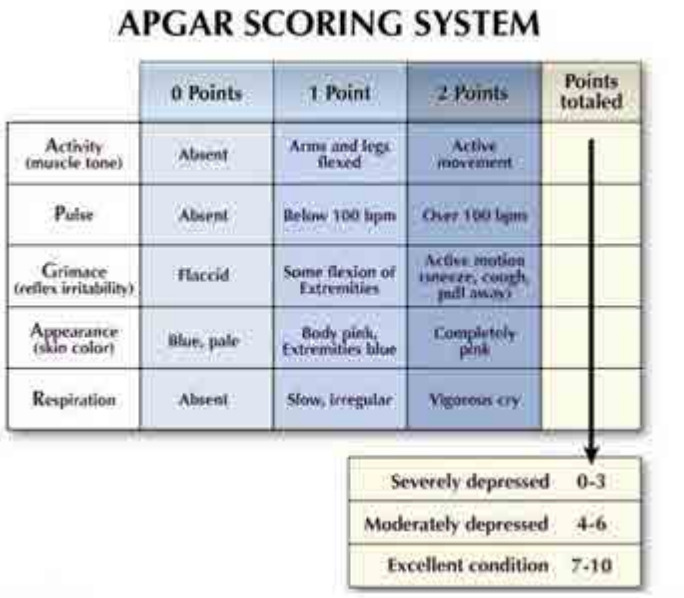


Figure 2: Chart of the Apgar scoring system used to predict infant morbidity.

Traditional approaches to risk stratification are based on scoring systems. The Apgar score, shown in Figure 2, is one such system [Apg66]. It is based on different criteria, with each answer having a specific point value. After answering, one adds up the points to obtain the risk-level score. There are hundreds of such scoring rules that are carefully derived through studies and are widely used in today's healthcare system.

1.4.2 Machine Learning-based Approaches

Now, most of industry is moving towards machine learning based methods that can work with a much higher dimensional set of features and solve a number of key challenges of these early approaches. Machine learning based approaches can:

- Fit more easily into clinical workflows. Scores from traditional approaches are often done manually. One has to figure out the corresponding inputs, so it is often not used as frequently.
- Be much quicker to derive. Traditional scoring systems have a very long research and development process that led to their adoption. With machine learning based approaches, given enough data or access to data, one can predict narrow outcomes or conditions that may occur infrequently.
- Lead to higher accuracy.

However, these new ML approaches also introduce new dangers. This will be discussed more in future lectures.

1.4.3 Example of Commercial Product By Optum

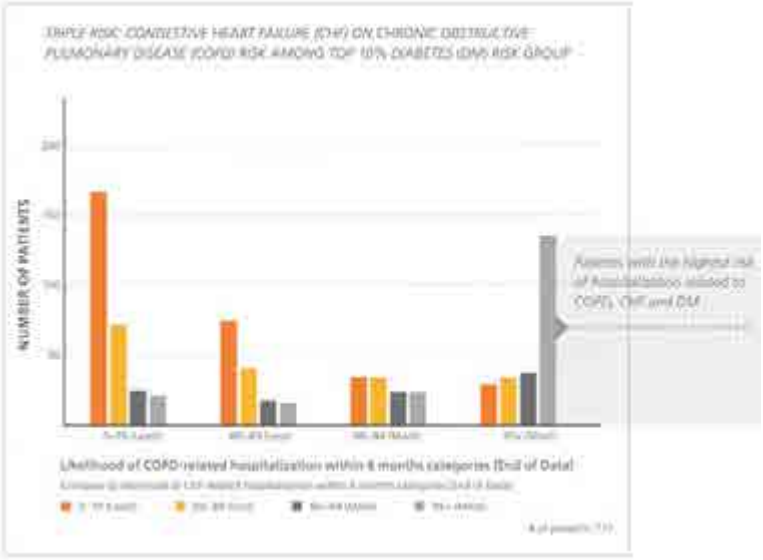


Figure 3: Predictive model by Optum showing likelihood of COPD-related hospitalizations.

Machine learning based models are being widely commercialized. For example, Optum has built a risk stratification tool and Figure 3 shows the output from one of their models predicting the likelihood of COPD-related hospitalizations, giving a population-level view of the results [Opt14]. Patients are scored using either the traditional or machine learning based models and placed into different categories depending on the risk level. One also has the ability to look more closely and see things like patients who are highest at risk or potential impact-able aspects of the patient's health.

High-risk diabetes patients missing tests	# of A1c tests	# of LDL tests	Last A1c	Date of last A1c	Last LDL	Date of last LDL
Patient 1	2	0	9.2	9/3/13	N/A	N/A
Patient 2	2	0	8	1/30/13	N/A	N/A
Patient 3	0	0	N/A	N/A	N/A	N/A
Patient 4	0	2	N/A	N/A	133	8/9/13
Patient 5	0	0	N/A	N/A	N/A	N/A
Patient 6	0	1	N/A	N/A	115	7/16/13
Patient 7	1	0	10.8	9/18/13	N/A	N/A
Patient 8	0	0	N/A	N/A	N/A	N/A
Patient 9	0	0	N/A	N/A	N/A	N/A
Patient 10	0	0	N/A	N/A	N/A	N/A

Figure 4: Table showing patients with high-risk diabetes.

Looking at the table in Figure 4, one could see patients with the highest risk of diabetes. Looking at Patient 3, one could see that they haven't been tracking their A1C. One could get them into the clinic to get their blood types to see whether they need a change in medication.

2 Case Study: Early Detection of Type 2 Diabetes

2.1 Background

We now consider a case study: risk stratification for Type 2 diabetes. This problem is extremely important, as an estimated 25% of individuals with diabetes in the United States are still undiagnosed, and the number is similar in other countries worldwide. If we are able to discover undiagnosed individuals who currently have diabetes, or identify people who are at high risk of developing diabetes in the future, we can provide interventions that prevent their condition from worsening, such as weight loss programs or first line diabetic treatments. In this section, we discuss the problem of identifying the population of individuals at high-risk of diabetes using machine learning algorithms.

Traditional approaches to this problem include point-based metrics similar to the APGAR score. The following image shows a sample questionnaire for evaluating diabetes risk in Finland, which produces a single score quantifying an individual's likelihood of developing diabetes.

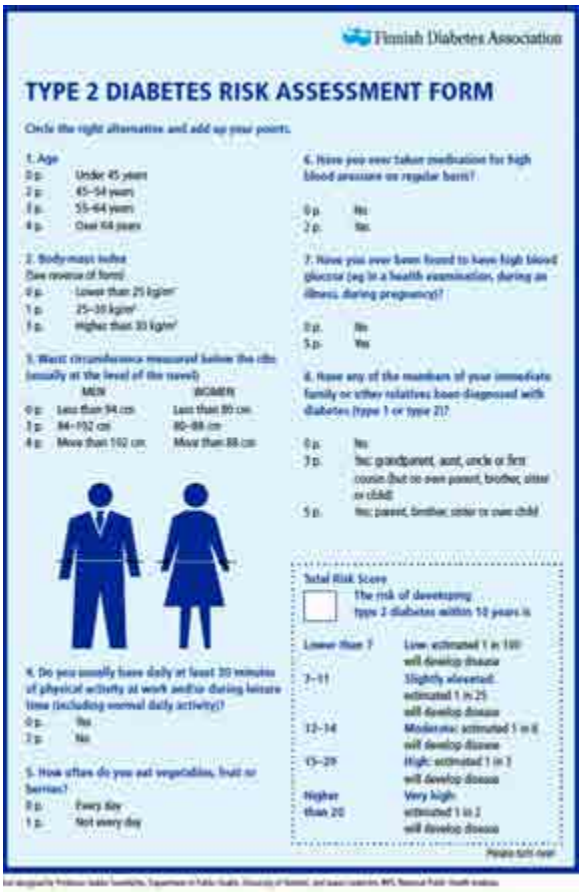


Figure 5: Questionnaire for Diabetes Risk Assessment

Unfortunately, these simpler methods have not had much impact and have not been widely used. Automation of the risk stratification process would allow us to avoid these types of manual questionnaires and lead to wider adoption. Instead of evaluating risk for every individual separately, an alternative option is to use machine learning models - trained on data from a health insurance company or other sources - that automatically identify the subpopulation at high risk of developing diabetes out of millions of individuals.

2.2 Data

This case study used administrative data from a health insurance company containing information about a patient population in Philadelphia. The types of data found in this dataset are summarized in the following diagram:

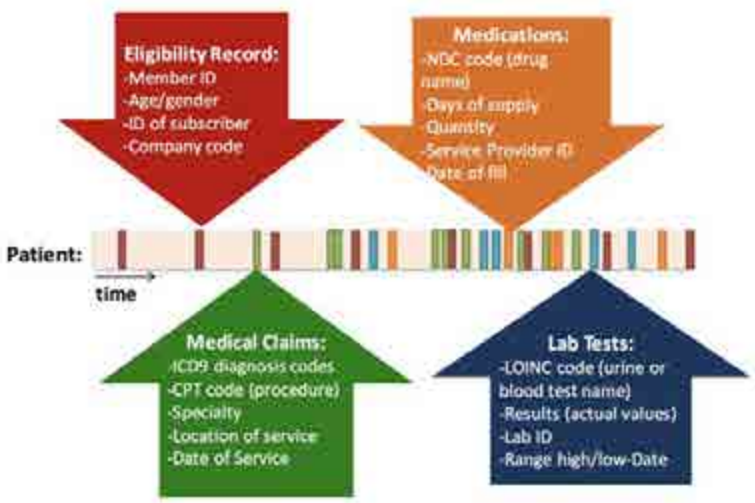


Figure 6: Overview of Data Types

The top diagnoses and administered lab tests in this dataset are shown in the tables below. Many of the most frequent lab tests come from the CBC panel, a common set of tests carried out during annual physicals and checkups. Note that the end of the lab test table contains the hemoglobin A1C test, used to measure blood glucose levels and track the status of diabetes patients. Type 2 Diabetes is also one of the most common diagnoses among the patient cohort.

Lab test	
2160-0 Creatinine	1284737
3094-0 Urea nitrogen	1282344
2823-3 Potassium	1280813
2345-7 Glucose	1299897
1742-6 Alanine aminotransferase	1187809
1920-8 Aspartate aminotransferase	1187965
2885-2 Protein	1277338
1751-7 Albumin	1274166
2093-3 Cholesterol	1268269
2571-8 Triglyceride	1257751
13457-7 Cholesterol.in LDL	1241208
17861-6 Calcium	1165376
2951-2 Sodium	1167675

Lab test	
2085-9 Cholesterol.in HDL	1155666
718-7 Hemoglobin	1152726
4544-3 Hematocrit	1147893
9830-1 Cholesterol.total/Cholesterol.in HDL	1037730
33914-3 Glomerular filtration rate/1.73 sq M.predicted	561309
785-6 Erythrocyte mean corpuscular hemoglobin	1070832
6690-2 Leukocytes	1062980
789-8 Erythrocytes	1062445
787-2 Erythrocyte mean corpuscular volume	1063665

Lab test	
770-8 Neutrophils/100 leukocytes	952089
731-0 Lymphocytes	943918
704-7 Basophils	863448
711-2 Eosinophils	935710
5905-5 Monocytes/100 leukocytes	943764
706-2 Basophils/100 leukocytes	863435
751-8 Neutrophils	943232
742-7 Monocytes	942978
713-8 Eosinophils/100 leukocytes	933929
3016-3 Thyrotropin	891807
4548-4 Hemoglobin A1c/Hemoglobin.total	527062

Disease	count
4011 Benign hypertension	447017
2724 Hyperlipidemia NEC/NOS	382030
4019 Hypertension NOS	372477
25000 DMII wo cmp nt st uncntr	339522
2720 Pure hypercholesterolem	232671
2722 Mixed hyperlipidemia	180015
V7231 Routine gyn examination	178709
2449 Hypothyroidism NOS	169829
78079 Malaise and fatigue NEC	149797
V0481 Vaccin for influenza	147858
7242 Lumbago	137345
V7612 Screen mammogram NEC	129445
V700 Routine medical exam	127848

Figure 7: Most Frequent Diagnoses and Administered Lab Tests in Patient Cohort

2.3 Machine Learning Formulation

This problem is treated as a binary classification problem: predicting whether or not a patient will develop diabetes. Data is collected for each patient before January 1, 2009, and the classification models predict

a patient's likelihood of developing diabetes at some time period in the "future" (after 1/1/2009). Three different prediction tasks were considered, using different gaps between the data collection and prediction windows, shown in Figure 8:

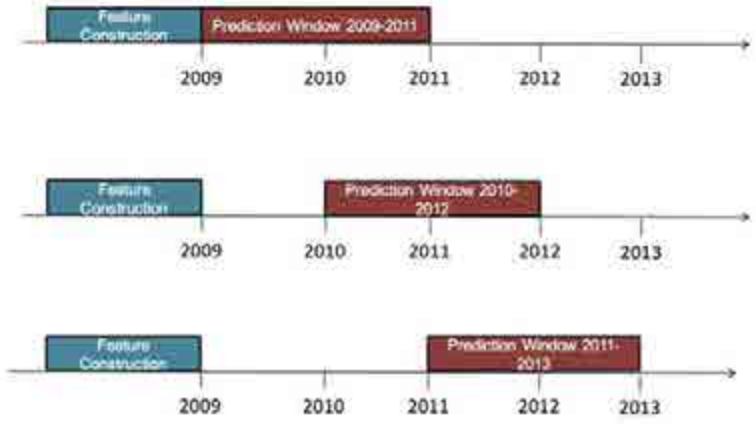


Figure 8: Prediction Tasks

In each case, we exclude patients who develop diabetes before the start of the prediction window. For instance, in the task with the 1-year gap, we exclude any patients diagnosed with diabetes prior to 1/1/2010.

One reason for including this gap is **label leakage**. In certain situations, it is possible that a doctor is very certain that a patient has diabetes, even though this has not been explicitly coded in a way that our algorithms can detect. The doctor may already be doing interventions based on this "pre-diagnosis". The models will pick up on these signals and predict that this patient is very likely to develop diabetes.

However, such a prediction is not very interesting, as the doctor has already identified the patient as being at high-risk for diabetes and is carrying out appropriate interventions. Instead, our models should be able to identify patients at high risk that the doctor may not expect.

Another issue is **data censoring**. For example, a patient may have only enrolled with an insurer in 2012, so they will have no data prior to 2009, and our models will not be able to construct any features for these individuals. There are two types of censoring that are handled:

- **Left Censoring**: Patient data absent prior to some point in time
- **Right Censoring**: Patient data absent after some point in time

For patients with left-censoring, the models attempt to construct as many features as possible; patients with less data simple have sparser feature vectors. Right-censored patients are dropped from the dataset if data in the full relevant prediction window is unavailable.

This simple exclusion criteria can be problematic in some cases. For instance, a patient may have switched insurers as a result of their diabetes diagnosis at some point in the prediction window, leaving them with no data after that time. Thus, we may actually be excluding patients who would benefit from our model's predictions and biasing the model's results. In the next lecture, we will discuss alternative approaches for handling right-censoring. For the rest of this section, we focus on the prediction task corresponding to the 1-year gap.

2.4 Model Overview

Logistic regression with L1 regularization is used for this task. L1 regularization is useful because it encourages sparsity in the feature weights of the trained model, which has multiple benefits:

1. Prevents overfitting in settings with a good risk model containing a small number of features.
2. Improves interpretability. Can potentially enumerate all non-zero features to better understand how the model makes predictions.
3. Improves translatability. If a model has only a small number of features, it is more likely that data needed for the models can be found at many hospitals, allowing the same model to be used more widely.

The cost function for such a model is of the form:

$$L(w) = \sum_{i=1}^{n} l(x_i, y_i; w) + \lambda \|w\|$$

where l is any loss function, w are the weights of the model, and $\|w\|$ is the L1-norm of the weight vector.

2.5 Features

Features were designed to account for the large amount of missing data in the records for most patients. Instead of choosing features in a way that would potentially require the imputation of many values, the models use several binary features indicating whether a particular observation was ever made for an individual in their records.

For instance, there are features for each type of specialist a patient could have visited. The corresponding feature value is "1" if the patient ever visited a particular type of specialist and "0" otherwise. Similar features are constructed for the most common medications ("1" if a person has ever taken, "0" otherwise).

A slightly different approach is used for featurizing lab test results. In addition to features indicating whether a patient ever took a particular lab test, there are features for each of the following:

- Is lab test result high/low/normal?
- Is result increasing/decreasing?
- Is result fluctuating?

If a patient had never been given a particular lab test, all the corresponding feature values would be 0. As constructed here, all these features are very simple. One could potentially use recurrent neural networks or other models to automatically learn features about the time series data present in lab test results. We will discuss more complex feature construction techniques in future lectures.

Each of these features are then computed for different time windows: the last 6 months, the last 24 months, and all of a patient's past history. In the end, a patient's feature vector consists of approximately 42,000 elements.

2.6 Model Evaluation

We first examine some of the features that were determined to be most predictive in the trained model. The top feature is found to be a diagnosis of "Impaired Fasting Glucose". While one may think that such a diagnosis would indicate that a patient has already been diagnosed as diabetic, these could also correspond to pre-diabetic patients in the dataset who are not guaranteed to develop diabetes. Other top features are shown in Figure 9.

Top History of Disease	Odds Ratio
Impaired Fasting Glucose (Code 790.21)	4.17 (3.87 4.49)
Abnormal Glucose NEC (790.29)	4.07 (3.76 4.41)
Hypertension (401)	3.28 (3.17 3.39)
Obstructive Sleep Apnea (327.23)	2.98 (2.78 3.20)
Obesity (278)	2.88 (2.75 3.02)
Abnormal Blood Chemistry (790.6)	2.49 (2.36 2.62)
Hyperlipidemia (272.4)	2.45 (2.37 2.53)
Shortness Of Breath (786.05)	2.09 (1.99 2.19)
Esophageal Reflux (530.81)	1.85 (1.78 1.93)

Figure 9: Most Predictive Disease Features

The criteria used to evaluate risk stratification models are slightly different from standard diagnosis criteria. One common metric is the model's **positive predictive value (PPV)**. In this case study, this metric corresponds to the proportion of predicted high-risk patients who actually went on to develop diabetes in the prediction window. The results for these particular models are shown in Figure 10, compared with a simpler model that did not perform as well.

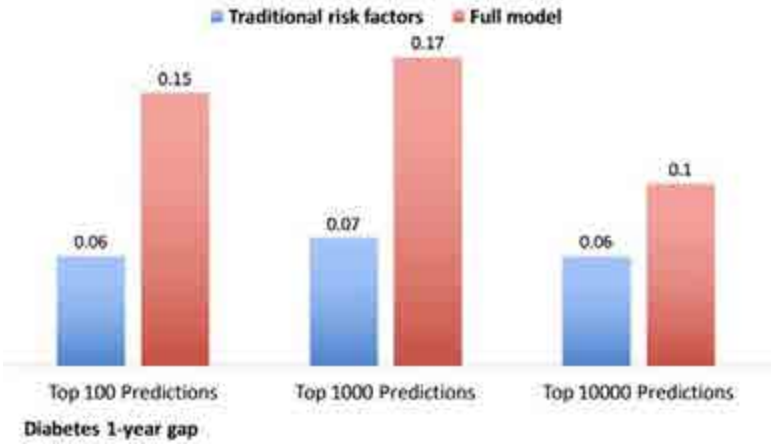


Figure 10: Positive Predictive Value (PPV) of Models for Different Groups

We evaluate this metric at different levels: for the top 100, top 1000, and top 10000 most risky patients. We can observe that 15% of the top 100 and 10% of the 10000 riskiest patients go on to develop diabetes. By performing these separate analyses, one could target different interventions for patients at different risk levels. Cheap interventions (i.e, an eye checkup for diabetic retinopathy) could be recommended for the top 10,000 riskiest patients. On the other hand, a more expensive intervention could not be implemented at such a large scale, and it may only be recommended for the 100 riskiest patients.

3 Interview with Leonard D'Avolio

3.1 Introduction

Leonard D'Avolio is an Assistant Professor at Harvard Medical School and the CEO and founder of Cyft. He spent the last 15 years "trying to help healthcare learn from its data in new ways" for governments, academia, researchers, publishing papers and nonprofits. Things D'Avolio worked on included

- Working with the Department of Veterans Affairs to build out their genomic science infrastructure and recruiting and enrolling millions of veterans to donate blood.
- Working at Ariadne Labs in improving neonatal care in India.

3.2 Interview

Q: What is risk stratification to you?

Risk stratification depends entirely on the problem. Risk could be:

- Running out of medical supply in an operating room
- The Apgar score
- A patient going from pre-diabetic to diabetic
- An older person falling down in their home

Risk startification is a set of wonderful tools with which skilled craftsmen can go ahead to solve specific problems.

Q: What are some of the areas that Cyft is applying today?

What we do is essentially performance improvement; more specifically, the performance in keeping people out of the hospital. The most logical application for these technologies is to help do preventative things, but only between 8-12% of healthcare is **financially incentivized** for that. As a company, you focus on where there's a financial incentive. I wanted to build a company where the financial incentives aligned with keeping people healthy.

We focus on older populations where it is important to understand who care managers should approach because their risk levels are rising. The traditional approach of risk stratification identifies people that are already at their most acute. We try to help care management organizations find people that are rising risks and bring a more granular approach to healthcare. *The power of these technologies is to move away from one-size-fits-all.* Examples of what Cyft does includes:

- Tackling rising risk of an inpatient psychiatric admission
- Predicting which older people are likely to fall down
- Finding which children with Type 1 diabetes should be scheduled an appointment now rather than every 3 months

The theme of the above examples is helping organizations move away from rather generic decisions towards things that are more actionable.

Q: What areas have you worked on the longest?

Cyft works with a large behavioral healthcare organization who is contracted by health plans to treat people that have mental health challenges. The traditional way of identifying the most acute people is through a risk score. The steps Cyft takes are

1. Get an understanding of where is the greatest opportunity, the greatest cost and what types of things are happening the most frequently
2. Find out what types of resources the team have like personnel and interventions to find the greatest possible return on investment from a data and financial standpoint
3. Get full agreement from executive team on what exactly is the narrow problem that they can address
4. Try to apply machine learning to solve the problems

Q: What was the problem that you decided to address?

We decided to address reducing inpatient psychiatric admissions. The traditional way of doing reducing admissions always been thought of in terms of 30 days out, but for this particular condition, it takes more like 90 days to have an impact.

Q: What kind of data is the most useful?

I think the philosophy that you should all take is that your data should be your competitive advantage in solving the problem. Our approach is: whatever data you have, we will consider it.

Behavioral health is incredibly under-diagnosed. There's a stigma attached to carrying diagnosis codes that would describe you as having mental health challenges. Because of the challenges, claims data alone is not enough. An important data source is from care managers who assesses the patient and fills out forms as well as written notes from clinicians like psychologists and psychiatrists.

Q: What is the development process?

The team looks for the simple solution to the problem first, like throwing logistic regression at it, before iterating back and forth based on how the data looks and its characteristics. Then, the team works through algorithms and feature selection approaches that seem to fit for the available data. A huge education component of the process is helping people understand what they're seeing, how to interpret it, and helping them connect it to what they're going to do with it.

Q: What is the deployment process?

It is far too late to start getting the client ready when the model is built and ready. I don't completely agree with the idea that these approaches are easier to plug into a workflow. For care managers who have spent years training and learning who needs their help the most, it is hard to just start trusting a computer.

Q: What are the technical details?

Healthcare is pretty immature from a technical standpoint. It can be a delivered Excel spreadsheet or a real-time call to an API. What we learn as a company serving healthcare is to not create a new interface; instead, we accommodate whatever workflow and system that they already have in place and build for flexibility. As a gross generalization, clinicians hate their information technology.

Healthcare is paid for based on delivering care, and the more complex the care the more one gets paid. Currently, because of the healthcare environment, very few organizations that Cyft works with and talk to

are ready for technologies like FIHR.

Q: What do you have to give around a prediction in order for it to be acted upon effectively?

The first thing you have to do is invite the clinical team from the very beginning. As one moves through the process, triangulate any information. After the development phase, if you have done a great job, you get away from the "show me what variables matter on a per-patient basis."

Q:How do you square up the culture of 5-7 variables like the Apgar score versus an inductive approach where thousands of variables are contributing incrementally?

It's a double-edge sword. You could never show somebody several hundred variables but if you show them 3-4, their answer would be "Well, that's obvious." Striking that balance is important, so it's really a lot education.

A helpful analogy is a GPS. GPS isn't going to give you a magic highway. It's going to suggest the roads that you're familiar with, but the advantage it has is that the GPS is aware of more than you are. It's going to give you a suggestion that will save you time, but the driver still makes the decision.

3.3 Class Q&A

Q: How do you address model bias due to lack of equal representation in the dataset across different demographic groups in the population?

A: If there's a demographic group that may be lost in the shuffle that we would do something different for, we try to explicitly bring this to their attention. For instance, kids usually just don't have as much data, and we try to handle those cases separately if we need to do something differently for them.

Q: How do you interpret the risk scores produced by a model to clients? What metrics do you use to explain model performance?

A: Use tons of graphics! Never show a graphic that isn't super obvious to interpret quickly or contains any unknown statistics. The title of a slide should tell viewers exactly what they need to know.

Q: What types of clients do you have?

A: Government-sponsored health programs and others that take on financial risks to keep people healthy. Also, we're looking for organizations that are able to make meaningful and costly interventions that can improve patient care.

Q: Are you willing to trade performance for interpretability?

A: We want to get our partners to have a deep understanding of our models, and success is when they're able to trust the models without needing to understand each of the variables that go into making a prediction. We'll walk them through the patients and variables as we construct the models to build up that trust.

Q: How much time do you spend engaging with physicians before starting building models?

A: We first spend time with the CEO, CFO, and CMO of the organization. We need to have at least a 5 to 1 financial return for solving this problem in order to make it up the chain and have a chance to make an impact with our models. After figuring out the financial parts of the project, the clinicians are embedded into the team from the very start. You have to have working groups of both clinicians, engineers, users, and others to get machine learning implemented in a healthcare setting.

Q: Can we use a data-driven approach to figure out what types of data we want to acquire?

A: It's easy to bring in new data, but the hard part is deciding whether new data actually contains added value. The data usually just doesn't tell you that you should go out and get a different type of data. If model performance is low, then we'll try to go out and find data with information that we think may boost the performance.

Q: How much impact do interventions have based on the predictions made by the model?

A: No customer ever pays you for a good positive predictive value; they only care about saving or making money. We show clients how much money they would save for a particular level of improvement, then relate that to the performance of our models. We don't show clients the predictions of our models; we show them the financial impact of our models and whether it was able to make a difference.

References

[Apg66] Virginia Apgar. The newborn (apgar) scoring system. *Pediatr Clin North Am*, 13(3):645–50, 1966.

[Opt14] Optum. Predictive analytics: Poised to drive population health.

[PDS+84] Michael W Pozen, Ralph B D'Agostino, Harry P Selker, Pamela A Sytkowski, and William B Hood Jr. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease: a prospective multicenter clinical trial. *New England Journal of Medicine*, 310(20):1273–1278, 1984.

[RBS+15] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015.

[SRG+10] Suchi Saria, Anand K Rajani, Jeffrey Gould, Daphne Koller, and Anna A Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Science translational medicine*, 2(48):48ra65–48ra65, 2010.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019