A Statistical Approach to Modeling Risk with Data from Electronic Medical Records

Joseph E. Lucas, Ph. D.
Clinical Trials from the EMR

<table>
<thead>
<tr>
<th>RUNNING STUDY FOR</th>
<th>FRACTION OF COST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharma</td>
<td>40-50%</td>
</tr>
<tr>
<td>Decrease cost of data management, data verification, administrative staff, site monitoring, RN and physician costs, site overhead</td>
<td>This doesn’t include new costs for managing the trial and analyzing the data through the EMR. How much of that can we automate?</td>
</tr>
<tr>
<td>Comparison of standards of care</td>
<td>20%</td>
</tr>
<tr>
<td>Retain IRB, site recruitment and retention, patient recruitment</td>
<td>Learning healthcare:</td>
</tr>
<tr>
<td>Health system</td>
<td>&lt;4%</td>
</tr>
<tr>
<td>No site recruitment or retention costs</td>
<td>• Compare standards of care (or care innovations?)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Randomize at the bedside</td>
</tr>
<tr>
<td></td>
<td>• Track outcomes through dashboard</td>
</tr>
</tbody>
</table>

Randomized Trials for Pharma

Phase III trial ~ $20M\(^1\)

Add 1 month to patent life of blockbuster ~ $250M

Speed of trial is critical
  EMR: Identify trial sites, directly recruit patients

CRO market ~ $24 Billion\(^2\)

Will FDA accept EMR data? Will working with FDA on novel trial delay completion?

\(^1\) https://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development
\(^2\) http://www.pharmsource.com/market/how-big-is-the-market-for/#R&D CRO
Interests of Health Systems

May be in direct opposition to pharma

Depends on business model
  Payer contract: At risk versus risk based versus fee for service
  Providers: Private practice versus on staff

Example: Diabetes
  At risk: head off adverse outcomes at low cost
  Risk based: Contract pays more for diabetics than pre-diabetics
  Fee for service: React to adverse events as they happen

Trend toward at risk and vertical integration
Health innovations: What are they worth? For whom?

HAP Innovations, pill dispenser with PRO
Trials: Evidence for payers

OVERALL APPROACH

Choose a patient population

Use EMR to identify high risk subset
   Which of them will respond to intervention?

Intervention

Evaluation of success
   Total cost over year after intervention
   Total hospital days or number of hospital visits
   Disease specific outcomes

ADVANTAGES

Health system / payer budget

Agnostic to disease and intervention

Automate Learning Healthcare
   Once implemented, novel interventions can be continuously tested
Risk Prediction: Complicated data

Electronic Health Records data has
- Continuous data (labs, age, vitals)
- Categorical data (gender, race, family history)
- Written text (nurses’ and physicians’ notes, radiology reports)
- Images (x-ray, CT, EKG)

Important information everywhere
- Example: A diabetic might have any or all of the following
  - Synonym of “diabetes” in a note
  - High lab values (glucose, HbA1C)
  - Relevant medications
  - Billing codes related to treatment of diabetes
  - Predisposing demographics (weight, race, family history)
  - Genetic predisposition (TCF7L2, JAZF1, HHEX, etc)

Temporal gaps – People are only patients when they are sick
We want to incorporate all of this information
Don’t want to be fooled by mistakes and bias
Predictive Modeling from EMR

Spots are observations about a patient. 10 patients shown

We have a trail of data
- Very different for each patient
- 10’s of thousands of different codes, labs, meds
- Radiology/pathology images? EKG?
- What is important?
- How much history to examine?

What will happen to patient?

Typical approach:
- Aggregate the data over a window
  - Counts, means, maxima, variance, quantiles, etc
- Put the aggregated data into a predictive model
  - Penalized regression, random forest, deep learning

4 day stay is just one vertical column
Versus Traditional Approach

UKPDS

Predict heart disease in diabetics
Age, sex, race, smoking, HbA1C, Systolic BP, total/HDL cholesterol
Averages .05 lower C-statistic compared to machine learning
Hierarchical Model – Multiple Outcomes

\[ N \] patients with \( K \) outcomes

\[ y_j \] is outcome \( j \) of dimension \( N \)

\( X \) is \( N \times D \) matrix of independent variables

\[ y_j = X\beta_j + \epsilon_j \]

Ridge regression: \( \beta_1 \sim N(0, \phi^{-1}I_D) \)

We want to borrow strength across outcomes: \( \beta_1 \sim N(\beta_i, t_{i,1} \Phi^{-1}) \)
Model fit: Diabetics

Most predictive of future cost
Improvement in Test Accuracy

Accuracy improvement over either MLE or ridge regression

Automatically links related outcomes

Generates predictors of composite outcomes as well
Future costs of at risk patients
Power: Sample size = 100

- Intervene on 100 highest risk patients
- Suppose intervention lowers them to the next risk quantile
- 90% confidence bounds for control and experimental group separate at around 15 months
“We know our high risk patients!”
(Stratification by hospital days)
Stratification: Only patients with zero hospital days
Acknowledgments

Electrical and Computer Engineering
Ricardo Henao

Center for Predictive Medicine
Michael Pencina

Center for Predictive Medicine
Tom Wall
Jenna Willingham

Cone Health
Robert Slaughter

UNC Greensboro
Carol Lucas
Daniel Bibeau
Regina McCoy

http://EHRanalytics.blogspot.com

joe@stat.duke.edu