Real-Time Data, Predictive Analytics Can Reduce Infections

John W. Cromwell, MD, FACS, FASCRS; The University of Iowa Hospitals & Clinics

Thomas Hill, Ph.D., Dell Statistica Software

DISCLAIMER: The views and opinions expressed in this presentation are those of the author and do not necessarily represent official policy or position of HIMSS.
Conflict of Interest

John W. Cromwell, MD., FACS, FASCRS

Dr. Cromwell serves as Associate Chief Medical Officer, Director of Surgical Quality & Safety, Director of the Division of Gastrointestinal, Surgery, and a faculty member in the Interdisciplinary Graduate Program in Informatics at University of Iowa.

Dr. Cromwell serves as a consultant to Dell, Inc. and receives travel expenses for such.
Conflict of Interest

Thomas Hill, Ph.D.

Dr. Hill is Executive Director of Analytics at the Dell Software Group. He is working closely with Dell Services, and in particular with Dell Health Care and Life Sciences.

Dell Software provides analytics solutions to health care and life sciences customers world-wide.

The Dell Health Care and Life Sciences group is an established provider of services and solutions to health care providers and payers world wide.
Agenda

• Learning objectives
• How benefits were realized for the value of Health IT
• Summary and overview
• Problem: Controlling surgical site infections (SSI) and readmission
  – Building institutional support
  – Available data sources and data
• A brief primer on analytics (Thomas Hill, Ph.D.)
  – Statistical analysis vs. pattern recognition
  – Big data and big data technologies
  – Architecture overview
• Implementation details (Dr. John Cromwell)
  – Data, predictor variables
  – Pre-coding/binning of data; Weight-of-Evidence (WoE) coding
  – Modeling and results
  – Implementation
• Practical considerations and summary
• Questions and answers
Learning Objectives

• Identify opportunities for using predictive analytics and big data sets to improve patient care decisions
• Analyze the practical steps needed to take advantage of the opportunities
• Identify key challenges in using predictive analytics to inform critical treatment decisions in real-time
• Calculate the cost, resources, expected outcomes and return on investment (ROI) for analytics projects
• Demonstrate a practical example of using real-time predictive analytics to reduce complications, improve patient outcomes and decrease the cost of care
An Introduction of How Benefits Were Realized for the Value of Health IT

**TREATMENT**: Guided by the use of predictive analytics tools, we were able to reduce surgical site infections by about 58%.

**SATISFACTION**: The system has increased satisfaction for involved providers and increasingly engaged them in quality improvement efforts.

**SAVINGS**: The decrease in surgical site infections is estimated to save us $740,000 for every 1,000 patients undergoing surgery. We perform greater than 19,000 surgeries annually,
Summary

• Surgical site infections (SSIs) are a major cause of morbidity and hospital readmissions in surgical patients.

• Real-time prediction of risk is needed prior to and during the time of an operation so that preventative strategies can be applied.
  – Naïve Bayes (NB) and support vector machines (SVMs) can effectively predict patients at risk for any SSI, or superficial SSIs

• A number of challenges have to be addressed to stand up a robust, “production-quality” system that reliably delivers valid risk profiles to the operating room
  – Data preparation and coding
  – Effective model management and monitoring for *validated analytics*
  – Real time delivery of actionable predictions
  – Building trust and support among surgical staff

• The technologies and methods we used to achieve significant improvements and ROI are readily available and highly cost-effective
Problem

• Surgical site infections (SSIs) are a major cause of morbidity and hospital readmission, affecting 3-11% of general surgery patients in the United States

• The additional healthcare costs attributed to SSIs are estimated to be over $20,000 per infection (Shepard 2013)

• SSIs are categorized into 3 distinct types by the Centers for Disease Control (CDC 2016); the 3 types are
  – Superficial (most common type)
  – Deep
  – Organ space (most severe and costly type)

• The goal was to determine SSI risk at the time of completion of the surgical procedure, including features of the operation itself, so that alternative wound management strategies and care protocols can be evaluated
Building Organizational Support

- It is critical to build support among
  - Leadership
  - All affected stakeholders
- Critical activities to build support at U of Iowa:
  - Educated *leadership* and other *stakeholders* on the motivation to move from reporting and dashboard analytics to proactive advanced analytics
  - Demonstrated consideration for avoiding huge changes to IT infrastructure – started in sandbox environment
  - Developed an important but limited-scope use-case with a small group of highly-engaged clinicians
  - Planned who and how analytics output would be USED to IMPROVE clinical care
A Brief Primer on Analytics

Thomas Hill, Ph.D.
Dell Software
What is pattern recognition?

Knowledge Discovery vs. Statistical Analysis

• Statistical Analysis
  – Focuses on “hypothesis testing” and “parameter estimation”
  – Fits “parsimonious statistical models” with the goal to “explain” complex relationships with fewer parameters
  – Examples: Logistic Regression, nonparametric statistics, factor analysis,

• Pattern Recognition (Data Mining)
  – The data are your model!
  – Algorithms find reliable repeated patterns in historical data:
    – Trees, boosted trees, voted trees (forests), SVM, neural nets, deep learning, numerous clustering methods, Kohonen networks, ...
  – Association and sequence rules, ...

• These methods are available in many commercial packages, or open-source R
• Are almost fully automated; straightforward to use
Big Data and Big Data Technologies

Pre-computer technologies:
- Printing press
- Dewey decimal system
- Punched cards

Magnetic tape
“flat” (sequential) files

Magnetic Disk

IDMS

ADABAS

System R

Oracle V2

Access

Postgres

MySQL

HBase

Dynamo

MongoDB

Redis

VoltDB

Neo4J

HBase

Aerospike

Hana

Riak

Cassandra

Vertica

Hadoop

1940-50
1950-60
1960-70
1970-80
1980-90
1990-2000
2000-2010

Relational Model defined
IMS
Network Model
Hierarchical model
Indexed-Sequential Access Mechanism (ISAM)

Sybase
Informix
Ingres

DB2
dBase

SQL Server

Adapted from: Guy Harrison, Dell Software, 2014 internal communication
Big Data, No-SQL Data, and ETL

- Big data generally means that (some) processing has to happen in/at the database, because the data are too large to be moved.
- Big data also often means “unstructured data”
  - Physician/nurses’ notes, Tweets, blogs, etc.
  - Hadoop and other inexpensive storage technologies allow these to be stored in raw form for future analyses.
  - To make use of unstructured data, they need to be pre-processed and turned into numbers via Extract-Transform-Load (ETL) processes.
Architecture Overview - UIHC

Data Modelers, Analysts, Data Scientists

Model Storage, ETL Processes

Application and Computation Server

Model Management, Version Control

Real-Time Data Collection
Real-Time Risk Predictions
Risk Profiles, Best Next Action

Model Monitoring, Sentinel

Real Time Livescore Server

Operating Room, During Surgery

Operating Room, During Surgery
How it Works

• Model builders/data-scientists/”business users” and “Citizen Data Scientists” build models
• Prediction models are “checked-into” model repository
  – Version control
  – Approval processes
  – Audit logs
• Some real-time data are collected during surgery
  – Validated data entry to Livescore™ server
• The real-time (Livescore) scoring server delivers
  – Predicted SSI risk
  – Risk profiles, causes, best-next-action ("prescriptive analytics")
Implementation Details

John W. Cromwell, MD
The University of Iowa Hospitals & Clinics
Practical Staffing Challenges

• Different skill sets are required to build out a modeling and real-time prediction solution

• Medical expertise, expertise about specific surgery
  – What data are available, can be utilized
  – What are relevant data and measurements that can be retrieved in real time during surgery

• Data architecture skills
  – How to retrieve and integrate data
  – Architecture for real-time scoring

• Data science, model building, statistics
  – Building models
  – Measuring and documenting accuracy
  – Measuring and documenting KPI’s and ROI
Practical Staffing Challenges

• Enterprise deployment skills
  – Automation
  – Meeting regulatory requirements

• Staffing challenges can easily be overcome in larger and/or university affiliated hospitals
  – There is significant interest in these technologies
  – Different stakeholders are excited to participate
Details: Data

• The modeling data included (real-time data in **bold**):
  – Patient Age
  – Patient Gender
  – Patient Home ZIP code
  – Surgeon Identifier
  – Body mass index (BMI)
  – **EBL** – Estimated Blood Loss
  – Blood transfusion volume
  – Lowest patient core temperature during surgery
  – Surgical Apgar Score – score interval
  – Hemoglobin level within prior 60 days
  – **ASA** - physical status classification system
  – **CDC Wound Class**– wound contamination class
  – Train/Test – train and test samples identifier
Important Predictors

- Feature Selection showed the following variables as the most important in predicting Hospital Readmission:

  - Surgeon ID
  - Apgar Score
  - EBL
  - HMB_within_60_days
  - ASA
  - Age
  - Gender

![Importance plot](https://example.com/plot.png)

- Dependent variable: Unplanned Readmission
- Importance (Chi-square)
Pre-Coding/Binning of Predictors

• It is important to pre-bin predictors such as *Age* into optimal (for prediction) intervals
  – This also helps with interpretation of models (e.g., “if 25<=Age<=30 then …”)
• Weight-of-evidence (WoE) coding is an established method used in risk scoring for binary outcomes
• The Weight-of Evidence (WoE) statistic is calculated as follows (Siddiqi 2006) *(Read_Rela_unplan = Unplanned Admission, 0/1)*:

\[
WoE = \left[ \ln \left( \frac{\text{Distr} \ "\text{Read_Rela_unplan} = 0" }{\text{Distr} \ "\text{Read_Rela_unplan} = 1" } \right) \right] \cdot 100
\]

• Smaller WoE values for a binned group means higher re-admissions risk
• The goal of WoE coding in risk modeling is to identify binning (interval boundaries) based on historical data, so that the observations in each interval are maximally uniform with respect to (re-admission) risk, while the differences in (readmission) risk across the intervals (groups) are maximized
Pre-Coding/Binning of Predictors

- Here are examples of pre-coded (optimally binned) predictors:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>(1) 18-35.4 (2) 35.5-65.4 (3) 65.5-66.4 (4) 66.5-100</td>
</tr>
<tr>
<td>Preop Hgb</td>
<td>(1) 6-11.54 (2) 11.55-15.94 (3) 15.95-16.04 (4) 16.05-18.04</td>
</tr>
<tr>
<td>EBL</td>
<td>(1) 0-31.4 (2) 31.5-33.9 (3) 34.0-112.4 (4) 112.5-177.4 (5) 177.5-189.9 (6) 190.0-587.4 (7) 587.5-7100</td>
</tr>
<tr>
<td>Transfusion</td>
<td>(1) 0-312.4 (2) 312.5-1712.4 (3) 1712.5-3900</td>
</tr>
<tr>
<td>BMI</td>
<td>(1) 13.0-13.9 (2) 14.0-39.4 (3) 39.5-41.4 (4) 41.5-84</td>
</tr>
<tr>
<td>Min Temp</td>
<td>(1) 87.8-87.94 (2) 87.95-88.04 (3) 88.05-90.74 (4) 90.75-90.84 (5) 90.85-94.14 (6) 94.15-94.24 (7) 94.25-98.24 (8) 98.25-101.3</td>
</tr>
<tr>
<td>Duration</td>
<td>(1) 6-41.4 ... (36) 567.5-675 ... Total of 36 intervals</td>
</tr>
</tbody>
</table>
Data for Validation Study

- The surgical dataset was extracted from common data sources
- EHR (Epic, Verona, WI) extracted to SQL datamart
- Data were then combined with outcomes data obtained from the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP), a validated, institution-based surgical database of patient risk factors and 30-day postoperative outcomes.

- Samples:

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>NUMBER OF CASES</th>
<th>SSI RATE</th>
<th>SSI RATE (SUPERFICIAL SSIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULL DATASET</td>
<td>2211</td>
<td>0.084</td>
<td>0.049</td>
</tr>
<tr>
<td>TRAINING SET</td>
<td>2085</td>
<td>0.084</td>
<td>0.047</td>
</tr>
<tr>
<td>VALIDATION SET</td>
<td>126</td>
<td>0.087</td>
<td>0.071</td>
</tr>
</tbody>
</table>
Predictor Variables
(Disease Factors)

• American Society of Anesthesiology score (ASA score)*
  – class I = healthy
  – Class II = mild or controlled systemic illness
  – Class III = severe or uncontrolled systemic illness
  – class V = expected death within 24 hours

• Preoperative hemoglobin

• Wound class* (clean, clean-contaminated, contaminated, or dirty/infected wound)
Predictor Variables
(Procedure Factors)

- Heart rate*
- Estimated blood loss*
- Blood transfusion*
- Minimum intraoperative temperature*
- Operating room*
- Duration of operation*
- Procedure type *
  (categorized by organ)
- Laparoscopic *or robotic vs open procedure
Predictor Variables
(Other Factors)

• Community factors:
  - Patient Home Zip code

• Provider factor:
  - Surgeon
Real-Time Data Interface

<table>
<thead>
<tr>
<th>Data Entry Setup:</th>
<th>OR SSI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surgery</strong></td>
<td></td>
</tr>
<tr>
<td>Place</td>
<td>1</td>
</tr>
<tr>
<td>Reasonable Val</td>
<td>1000000/999999</td>
</tr>
<tr>
<td><strong>Measures</strong></td>
<td></td>
</tr>
<tr>
<td>Place</td>
<td>1</td>
</tr>
<tr>
<td>Reasonable Val</td>
<td>88/105</td>
</tr>
<tr>
<td>Sample Comments</td>
<td></td>
</tr>
</tbody>
</table>

(HIMSS2016)
Association of Predictors with SSIs

• Mutual Information

- A common feature selection method
- Measures the information that \( X \) and \( Y \) share
- It measures how much knowing one of these variables reduces uncertainty about the other
- \( H(X), H(Y) \) are individual entropies
- \( H(X,Y) \) is the joint entropy
- \( H(X|Y), H(Y|X) \) are the conditional entropies

- \( I(X;Y) \) is the mutual information

\[
I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X,Y) - H(X|Y) - H(Y|X)
\]
## TOP 5 FEATURES

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mutual Information</th>
<th>Chi squared statistic</th>
<th>SSI Cases (175 total cases +SSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIP</td>
<td>0.128</td>
<td>485</td>
<td>522XX (6 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>525XX (6 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>522XX (5 cases)</td>
</tr>
<tr>
<td>DURATION</td>
<td>0.073</td>
<td>260</td>
<td>Intervals 2&amp;3 (25 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intervals 19&amp;20 (50 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Interval 23 (23 cases)</td>
</tr>
<tr>
<td>EBL</td>
<td>0.048</td>
<td>173</td>
<td>Interval 3 (43 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Interval 6&amp;7 (81 cases)</td>
</tr>
<tr>
<td>SURGEON</td>
<td>0.045</td>
<td>115.0659</td>
<td>112395 (71 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>115130 (28 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>119575 (26 cases)</td>
</tr>
<tr>
<td>PROCEDURE</td>
<td>0.041</td>
<td>108.8076</td>
<td>Colon (122 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Other (31 cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gall bladder (7 cases)</td>
</tr>
</tbody>
</table>
Assessing Model Performance

- 10-fold cross validation
- AUC (area under curve of ROC)
- F score
- Threshold
- Accuracy
- Sensitivity, PPV
- Specificity, NPV
## Assessing Model Performance

<table>
<thead>
<tr>
<th></th>
<th>Training Set (Superficial SSI)</th>
<th>Training Set (All SSI)</th>
<th>Validation Set (All SSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Max F score</strong></td>
<td>0.27</td>
<td>0.39</td>
<td>---</td>
</tr>
<tr>
<td><strong>Threshold</strong></td>
<td>0.11</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td>0.76</td>
<td>0.79</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.83</td>
<td>0.85</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Assessing Model Performance

<table>
<thead>
<tr>
<th></th>
<th>Training Set (Superficial SSI)</th>
<th>Training Set (All SSI)</th>
<th>Validation Set (All SSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.42</td>
<td>0.58</td>
<td>0.64</td>
</tr>
<tr>
<td>PPV</td>
<td>0.38</td>
<td>0.34</td>
<td>0.17</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.87</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>NPV</td>
<td>0.97</td>
<td>0.96</td>
<td>0.95</td>
</tr>
</tbody>
</table>
# Model Performance

<table>
<thead>
<tr>
<th></th>
<th>Outcome No SSI</th>
<th>Outcome SSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction no SSI</td>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>Prediction SSI</td>
<td>35</td>
<td>7</td>
</tr>
</tbody>
</table>

AUC: .79  
Sen: 0.64 ; PPV 0.17  
Spec: 0.70 ; NPV 0.95
Potential ROI

• Consider negative pressure wound therapy
• 60-80% effective at reducing SSI in high-risk wounds
• Total cost approximately $1500
• We can address 64% of SSI by using therapy in ⅓ of patients
Limitations

Variables not included in model that may be helpful:

• Co-morbidities
  – Diabetes
  – Smoker
  – Malnutrition
  – Preoperative colonization with Staph aureus
  – Coexisting infection at remote body site
  – Immunodeficiency

• Medications
  – Steroid use

• Operative factors
  – Antibiotic prophylaxis
  – Duration of surgical scrub
  – Foreign material at surgical site
A Summary of How Benefits Were Realized for the Value of Health IT

**TREATMENT:** Guided by the use of predictive analytics tools, we were able to reduce surgical site infections by about 58%.

**SATISFACTION:** The system has increased satisfaction for involved providers and increasingly engaged them in quality improvement efforts.

**SAVINGS:** The decrease in surgical site infections is estimated to save us $740,000 for every 1,000 patients undergoing surgery. We perform greater than 19,000 surgeries annually,
References


Questions

• For more information, contact
  • John W. Cromwell, MD
    john-cromwell@uiowa.edu
  • @jcromwell

• Thomas Hill, Ph.D.
  Tom.Hill@Software.Dell.com
  • @DrTomHill