Predictive Analytics at University of Utah Health Care

Samir E. Abdelrahman, Ph.D.
Research Scientist
Department of Biomedical Informatics
University of Utah
samir.abdelrahman@utah.edu
Academic Career

- **B.Sc.:** operations research (mathematical and statistical modeling for solving industrial optimization problems)

- **M.Sc. and PhD: computer science**
  - Faculty of Computers and Information - Cairo University
  - Major: artificial intelligence
  - Minor: programming languages and compilers

- **Assistant professor at Cairo University**

- **Cairo University research fellowship:**
  - Minnesota University – Computer Science Department

- **Associate professor at Cairo University**

- **USA biomedical informatics research fellowships**
  - Vanderbilt University, University of Illinois, and University of Utah
Research Interests

● **Predictive analytics**

● **Data mining and multi-agent Systems**

● **Natural language processing applications**

● **Ontologies**
Outline

● Background
  o What is predictive analytics?
  o Health care predictive analytics stages
  o Examples of health care applications
  o Definitions
    ▪ Predictive model performance measures
    ▪ P-value risk factor analysis

● Predictive analytics at University of Utah Health Care (UUHC)
  o Predictive analytics pipeline
  o Clinical projects
Predictive analytics is the practice of extracting information from existing data sets to learn patterns and predict future outcomes and risks.
Health Care Predictive Analytics Stages

Data Collection

Providers

Patients

Visits

Predictive Modeling

Data Mining

Statistics

Optimization

Actions

Deployment

What happened?
How many and how often?
What exactly is the problem?
What should we do?

Why is this happening?
What if these trends continue?
What will happen next?
What are the risk factors?
What are the outcomes?

What are the policies?
What are the decisions?
What are the interventions?

Patient care improvement
Model implementation
Operational follow-up

© Samir E. AbdelRahman – All rights reserved
### Example Health Care Applications

<table>
<thead>
<tr>
<th>Outcome of Interest</th>
<th>Potential Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital readmission</td>
<td>Intensively manage high-risk patients post-discharge</td>
</tr>
<tr>
<td>Appointment no-shows and last-minute cancellations</td>
<td>Contact intensively prior to appointment</td>
</tr>
<tr>
<td>Gastro-intestinal (GI) bleeding following Left Ventricular</td>
<td>Modify management and follow-up of high-risk patients post-op</td>
</tr>
<tr>
<td>Assist Device (LVAD) implantation</td>
<td></td>
</tr>
<tr>
<td>Patient satisfaction</td>
<td>Understand and address risk factors to improve patient experience</td>
</tr>
<tr>
<td>Sepsis mortality</td>
<td>Rapid initiation of sepsis protocol</td>
</tr>
<tr>
<td>Resource utilization</td>
<td>Allocate additional care management resources</td>
</tr>
</tbody>
</table>
### Predictive Model Performance Measures (Readmission Example)

<table>
<thead>
<tr>
<th>Positive (Predicted Readmitted)</th>
<th>Not-Readmitted</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readmitted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (TP – True Readmitted)</td>
<td>B (FP – False Readmitted)</td>
<td>(T_{\text{predicted readmitted}})</td>
</tr>
<tr>
<td>Negative (Predicted Not-Readmitted)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C (FN – False Not-Readmitted)</td>
<td>D (TN – True Not-Readmitted)</td>
<td>(T_{\text{predicted not-readmitted}})</td>
</tr>
</tbody>
</table>

- **Positive predictive value (PPV):** probability that patient was **readmitted** when the model predicted patient would be **readmitted**  \(\frac{A}{A+B} \times 100\)

- **Negative predictive value (NPV):** probability that patient was **not readmitted** when the model predicted patient would **not be readmitted**  \(\frac{D}{D+C} \times 100\)
Predictive Model Performance Measures (Readmission Example)

<table>
<thead>
<tr>
<th></th>
<th>Readmitted</th>
<th>Not-Readmitted</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td>A (TP – True Readmitted)</td>
<td>B (FP – False Readmitted)</td>
<td>$T_{predicted\ readmitted}$</td>
</tr>
<tr>
<td>(Predicted Readmitted)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td>C (FN – False Not-Readmitted)</td>
<td>D (TN – True Not-Readmitted)</td>
<td>$T_{predicted\ not-\ readmitted}$</td>
</tr>
<tr>
<td>(Predicted Not-Readmitted)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_{true\ readmitted}$</td>
<td></td>
<td>$T_{true\ not-\ readmitted}$</td>
<td>$T_{total}$</td>
</tr>
</tbody>
</table>

- **Positive predictive value (PPV):** probability that patient was readmitted when the model predicted patient should be readmitted  
  \[
  \frac{A}{A+B} \times 100
  
  \text{Maximize PPV and NPV} ...
  
- **Negative predictive value (NPV):** probability that patient was not readmitted when the model predicted patient would not be readmitted  
  \[
  \frac{D}{D+C} \times 100
  
  \text{Maximize PPV and NPV} ...
## Predictive Model Performance Measures (Readmission Example)

<table>
<thead>
<tr>
<th>Positive (Predicted Readmitted)</th>
<th>Readmitted</th>
<th>Not-Readmitted</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (TP – True Readmitted)</td>
<td>A</td>
<td>B (FP – False Readmitted)</td>
<td>T_{predicted readmitted}</td>
</tr>
<tr>
<td>C (FN – False Not-Readmitted)</td>
<td></td>
<td>D (TN – True Not-Readmitted)</td>
<td>T_{predicted not-readmitted}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>T_{True Readmitted}</th>
<th>T_{True Not-Readmitted}</th>
<th>Total</th>
</tr>
</thead>
</table>

### Area Under Curve (AUC/a.k.a. c-statistic):

- $\geq 50$ and $\leq 60 \rightarrow$ By chance prediction
- $> 60$ and $\leq 70 \rightarrow$ Limited prediction
- $> 70$ and $\leq 80 \rightarrow$ Good prediction
- $> 80$ and $\leq 100 \rightarrow$ Very good prediction

### Sensitivity:

$$\frac{A}{(A+C)} \times 100$$

### Specificity:

$$\frac{D}{(D+B)} \times 100$$
P-value for Risk Factor Analysis

- Is risk factor $X$ associated with outcome of interest $Y$?
  - If there is a difference among the distributions of the $X$ values between the two populations of $Y$, e.g. readmitted and not-readmitted, then $X$ is considered as a risk factor.

- Readmission example:
  - Is Age a risk factor?
    - $H_0 : \mu_{Age(\text{readmitted})} = \mu_{Age(\text{not readmitted})}$
    - $H_1 : \mu_{Age(\text{readmitted})} \neq \mu_{Age(\text{not readmitted})}$
    - For example:
      - if p-value $\leq$ threshold (e.g. 0.005), then reject $H_0$
        - Age has a significant impact on readmission risk
      - else we fail to reject $H_0$
Predictive Analytics at University of Utah Health Care (UUHC)
Health Care Predictive Analytics Stages

Data Collection

- Providers
- Patients
- Visits

Data Mining

- Why is this happening?
- What if these trends continue?
- What will happen next?
- What are the risk factors?
- What are the outcomes?

Predictive Modeling

- What happened?
- How many and how often?
- What exactly is the problem?
- What should we do?

Actions

- What are the policies?
- What are the decisions?
- What are the interventions?

Deployment

- Treatment Plans
- Operational Follow-up
- New Patient Services
- Patient care improvement Model implementation Operational follow-up

© Samir E. AbdelRahman – All rights reserved
Predictive Analytics Pipeline

Problem Analysis -> Data Analysis -> Systematic Model Development -> Risk Factor Analysis -> Operational Application

Data Collection -> Predictive Modeling -> Actions -> Deployment

What should we do?
What are the risk factors?
What are the outcomes?
Operational follow-up

© Samir E. AbdelRahman – All rights reserved
Predictive Analytics Pipeline

Problem Analysis → Data Analysis → Systematic Model Development → Risk Factor Analysis → Operational Application

Data Collection → Predictive Modeling → Actions → Deployment

What should we do? → What are the risk factors? → What are the interventions? → What are the outcomes? → Operational follow-up

1. Problem Analysis
2. Data Analysis
3. Systematic Model Development
4. Risk Factor Analysis
5. Operational Application

© Samir E. AbdelRahman – All rights reserved
Predictive Analytics Pipeline

Data Collection > Predictive Modeling > Actions > Deployment

Problem Analysis > Data Analysis > Systematic Model Development > Risk Factor Analysis > Operational Application

Data Mining
Statistics
Optimization

Providers
Patients
Visits

© Samir E. AbdelRahman – All rights reserved
Pipeline: Problem Analysis

- Collaborate with experts and partners
- Determine the aim and the objectives
- Identify the explanatory and response (yes/no) variables— in some cases, the response variable is numeric (regression)
  - Expert and literature
- Collect an operational dataset
  - Data retrieval (e.g., from data warehouse)
Discover potential challenges

- Variable(s) distribution may change over time
- Dataset selection to represent the population
- Missing values/variables
  - For example:
    - Consider lab tests of 4 patients with values 60, 70, N/A, 65
    - Then we have 1 missing value
- Imbalanced class values
  - For example:
    - If we have 10 patients with 8 readmitted and 2 not readmitted
    - Then we have an imbalanced situation towards the readmitted class
Pipeline: Data Analysis

• Analyze the explanatory variables using descriptive statistical methods

• Divide the data into partitions of derivation and validation datasets
  o All variable exclusion and risk analysis should use only derivation datasets

• Exclude (rule out) irrelevant variables
  o Correlation analysis among explanatory variables
    ▪ Identify independent/dependent variables
  o Statistical variable exclusion methods
  o Statistical tests or p-value analysis between explanatory variables and response variable
Investigate all possible supervised learning techniques with variable selection methods to maximize positive/negative predictive value

- Feature selection/rankers methods and/or p-value statistical test analysis
To maximize positive/negative predictive value, we may also handle the data using some methods such as:

- Discretization methods
  - Methods for transferring a numeric variable into discrete counterparts
  - For example: presenting age as intervals of 0-10, 11-20, ...

- Imputation methods
  - Methods for filling missing variable values
  - For example: given 4 patient lab tests: 60, N/A, 70, 65, then the values may become 60, 65, 70, 65

- Imbalance methods
  - Methods for balancing samples for each data class
  - For example: under-sampling methods to deal with the highly represented cases and others to over-sample the low ones
We use statistical test and p-value analysis maximizing positive/negative predictive values to assess the significance of:

- Risk factors (e.g. age)
- Risk factors values (e.g. age ≥ 65 has high readmission ratio)
Pipeline: Operational Application

- Implement the computational core
  - Model implementation
  - Backend implementation

- Meet with stakeholders to design interventions leveraging prediction model outcomes

- Design and implement the user interface application
  - Web interface
  - Client/server interface
Predictive Analytics at UUHC

- Congestive heart failure readmission*
- No-shows and last-minute cancellation appointments
- Non-surgical bleeding for patients receiving continuous flow Left Ventricular Assist Devices (LVAD)
- Ongoing projects

Predictive Analytics at UUHC

- Congestive heart failure readmission*

  The goals are (1) to propose and modify methods to contribute to the clinical predictive analytics approaches, and (2) to predict potential risk factors in our UUHC settings using these methods.

Congestive Heart Failure Readmission

- **Expert collaborators and partners**
  - Dr. Bruce E Bray, Biomedical Informatics Department, University of Utah

- **Explanatory variables**
  - **Demographic information:**
    - Age, race, religion, finance class, gender, and zip code
  - **Hospitalization data:**
    - Discharge disposition, the responsible hospital service, the length of stay, and the comorbidities
  - **48 laboratory tests**
  - **4 Vital sign Readings:**
    - Last reading prior to discharge for systolic blood pressure, reading weight, and last heart rate
    - First reading weight upon hospital admission
  - **Prior 6-month visit information**
    - Frequencies of prior emergency, outpatient visits,… with their length of stays and comorbidities
Congestive Heart Failure Readmission

● **Response variable**
  o Readmission (repeat inpatient hospitalization) for any cause within 30 days of the index CHF hospitalization

● **Dataset**
  o 2,787 CHF hospitalizations (January 2003 - June 2013) having the derivation and the first validation datasets
  o The second validation dataset of patient records
    ▪ July - October 2013
Challenges

- Vital Sign Data:
  - They were regularly saved in the EDW only starting from 2008, when a new electronic health record system was implemented at UUHC

- A new scheduling and billing system caused changes to some categorical variable values:
  - The designation of the cardiothoracic surgery service changed from the “CTI” service to the “CTS” service, we merged such values into a single CTI-CTS value

- The annual readmission rates for 2003-2012 through the first 9 months of 2013:
  - 17.06%, 14.50%, 15.38%, 24.11%, 11.81%, 10.57%, 16.78%, 23.20%, 19.94%, 15.50%, 8.50%, and, 12.8%
Challenges

- The distribution of many variables changed over time
- On average, 45% of patient records have complete values of the variables with distribution:

<table>
<thead>
<tr>
<th>Number of Records</th>
<th>Time Period</th>
<th>% of Complete Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>1122</td>
<td>2003-2007</td>
<td>5.3%</td>
</tr>
<tr>
<td>227</td>
<td>2008</td>
<td>42.7%</td>
</tr>
<tr>
<td>1250</td>
<td>2009-2012</td>
<td>47.4%</td>
</tr>
<tr>
<td>188</td>
<td>January-June 2013</td>
<td>47.3%</td>
</tr>
<tr>
<td>149</td>
<td>July-October 2013</td>
<td>45.7%</td>
</tr>
</tbody>
</table>
Excluding 21 variables
- Any having > 50% missing values

The first validation dataset
- January-June 2013

The second validation dataset
- July-October 2013

Derivation datasets
- 2003-2012
- 2008-2012
- 2009-2012
• **The selected derivation dataset:**
  o Complete records of 2008-2012 (729 records with 18.1% readmission rate)

• **The first validation dataset:**
  o The whole records (complete and missing) of January-June 2013 (188 records with 8.5% readmission rate)

• **The second validation dataset:**
  o The whole records (complete and missing) of July-October 2013 (149 records with 12.8% readmission rate)

• **Selected methods**
  o Voting classifier between voting feature interval (VFI) and logistic regression
  o Class-attribute contingency coefficient discretization method
  o The wrapper subset feature selection method
  o Three ranking strategies based on information gain, gain ratio, and symmetrical uncertainty
Congestive Heart Failure Readmission

**Results:**

<table>
<thead>
<tr>
<th>Validation Datasets</th>
<th>Our Approach</th>
<th>Others*,**</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>January-June 2013</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC:</td>
<td>86.8%</td>
<td>58.3-68.1%</td>
</tr>
<tr>
<td>PPV:</td>
<td>50.0%</td>
<td>35.2-41%</td>
</tr>
<tr>
<td>NPV:</td>
<td>96.4%</td>
<td>70.2-79.1%</td>
</tr>
<tr>
<td><strong>July-Oct 2013</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC:</td>
<td>79.0%</td>
<td>53.1-62.2%</td>
</tr>
<tr>
<td>PPV:</td>
<td>44.0%</td>
<td>25.8-38.9%</td>
</tr>
<tr>
<td>NPV</td>
<td>90.4%</td>
<td>60.2-77.6%</td>
</tr>
</tbody>
</table>

*Van Walraven, Carl, et al. Derivation and Validation of an index to Predict Early Death or Unplanned Readmission after Discharge from Hospital to The Community. Canadian Medical Association Journal 182.6 : 551-557, 2010

Developing risk analysis algorithm coupled with rankers, feature selection methods, and p-values on explanatory variables and selected risk factors

- Strong (High)
- Regular (Normal)
- Weak (Low)
Developing risk analysis algorithm coupled with rankers, feature selection methods, and p-values on explanatory variables and selected risk factors

- Strong (High)
- Regular (Normal)
- Weak (Low)

**If (the variable p-value ≤ the threshold)**
- If (the variable selected by Feature selection method and the 3 rankers with the threshold) then
  - The variable has a high significance
- Else
  - The variable has a normal significance

**Else**
- If (removing the variable increases the classifier performance) then
  - The variable should be removed totally.
- Else
  - The variable has a low significance
Congestive Heart Failure Readmission

42 Selected risk variables

<table>
<thead>
<tr>
<th>Significance Level</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>discharge disposition, discretized age, anemia-related labs, hospital services, the Carlson Index frequency, injury and poisoning CCS category, and the prior 6-month variables of Carlson Index frequency and ED encounters</td>
</tr>
<tr>
<td>Low</td>
<td>race, religion, and insurance/finance class</td>
</tr>
<tr>
<td>Normal</td>
<td>The remaining variables</td>
</tr>
</tbody>
</table>
Future Work

- Generalizing the predictive model to include any index cause of hospitalization
- Working on another data pattern to partition the derivation datasets
- Using time-based analysis to tackle variable distribution change over time nature
- Developing some imputation and exclusion methods (hands-on)
- Integrating with NLP techniques to get benefit from clinical notes
Predictive Analytics at UUHC

- Congestive heart failure readmission*

- No-shows and last-minute appointment cancellations

- Non-surgical bleeding for patients receiving continuous flow Left Ventricular Assist Devices (LVAD)

- Ongoing projects

No-shows and Last-minute Cancellation Appointments

- **Expert collaborators and partners**
  - Ryan M Vanderwerff, UU Orthopaedic Center, Outpatient Services Director
  - Matthew Ball Huish, Physical Medicine & Rehabilitation, Administrative Director
  - Steve Johnson, Value Engineering, UUHC

- **Explanatory variables**
  - Demographic information
  - Visit data
  - Prior 1-year complete and no-shows/cancelled visit information

- **Response variable**
  - Appointment no-shows or cancel within 1 business day
No-shows and Last-minute Cancellation Appointments

**Dataset**
- 44,406 appointments (November 2010 - September 2014)

**Challenges**
- Many records have missing values for patient information values
- Imbalanced dataset: 6% no-shows, 30% cancellations (including rescheduled), and 67% completed appointments
- Many cancellations cases were similar to no-show cases
No-shows and Last-minute Cancellation Appointments

- **Excluding 30 variables**
- **Dataset**
  - The derivation dataset: November 2010-May 2014
  - The first validation dataset: June-September 2014
  - The second validation dataset: October-November 2014
- **Bayesian network classifier and our imputation and imbalance methods**
  - Feature selection methods and rankers
- **Current results**
  - AUC: 73.3%, PPV: 51.5%, and NPV: 81.6%
- **Strong risk factors**
  - Age, gender, marital status, race, religion, zip, insurance, appointment time (AM/PM), length, and delay
  - Prior frequencies:
    - no-shows, cancellations, ED visits, hospitalizations, outpatient encounters, and some of comorbidities
- **A research paper under preparation**
No-shows and Last-minute Cancellation Appointments

- **Computational core**
  - Model development
  - Randomizing test set 50%/50% for control/treatment groups

- **Intervention:**
  - MA to call the patient prior to the visit

- **Web-based Application**
  - ITS team:
    - Vicki Wells, Steve Wood, and John DeGrey
Predictive Analytics at UUHC

- Congestive heart failure readmission*
- No-shows and last-minute appointment cancellations
- Non-surgical bleeding for patients receiving continuous flow Left Ventricular Assist Devices (LVAD)
- Ongoing projects

Non-surgical Bleeding For Patients Receiving Continuous Flow LVAD

● Main collaborators
  o Dr. Jose Nativi-Nicolau, Division of Cardiovascular Medicine, University of Utah
  o Omar E. Wever-Pinzon, MD, Division of Cardiovascular Medicine, University of Utah

● Dataset
  o 236 records (January 2003 – October 2013)

● Variables
  o Explanatory: demographic information and 65 lab tests and vital signs
  o Response: bleeding outcome (32.3% yes vs. 67.7% no)

● Challenge
  o Many variables and small dataset
  o Index or scoring function

● Excluding 40 variables
Non-surgical Bleeding For Patients Receiving Continuous Flow LVAD

- **Methods:**
  - Bootstrapping for validation
  - Unsupervised Discretization using quartiles – Mean
  - Regression scores and Framingham risk score function
  - Feature selection methods and rankers

- **Current Results**
  - AUC: 74.3%, PPV= 62.0%, and NPV=74.68%

- **Strong risk Factors**
  - Age, etiology of cardiomyopathy, chronic kidney disease, prior bleeding, aspartate aminotransferase, and right ventricular function

- **A research paper under preparation**
Predictive Analytics at UUHC

- Congestive heart failure readmission*
- No-shows and last-minute appointment cancellations
- Non-surgical bleeding for patients receiving continuous flow Left Ventricular Assist Devices (LVAD)
- Ongoing projects

Ongoing Projects

- **Patient satisfaction assessment and improvement**
  - Collaborators:
    - **Dr. P. Ward**, Division of Facial Plastic & Reconstructive Surgery Practice, University of Utah
    - **Steve Johnson**, Value Engineering, University of Utah Health Care
  - Most of work done and a research paper under preparation

- **Prediction of 30-day readmission for patients receiving continuous flow Left Ventricular Assist Devices (LVAD)**
  - Main Collaborators:
    - **Dr. Craig H. Selzman**, Cardiothoracic Surgery Division, University of Utah
    - **Dr. Jose Nativi-Nicolau**, Division of Cardiovascular Medicine, University of Utah
    - **Aaron Healy, MD**, Residents/General Surgery, University of Utah

- **Sepsis**
  - Collaborators:
    - **Devin Horton, MD**, General Internal Medicine
    - **John Arego**, Clinical Doc Data Manager

- **Financial predictions (UUMG leadership)**

- **Automation of predictive analytics pipeline**
Take Home Message

- Predictive analytics has high-value cases for health care

- UUHC is home to many great collaborators for driving clinical improvements leveraging predictive analytics
Acknowledgements

- Our Collaborators
  - Dr. Bruce E Bray
  - Ryan M Vanderwerff
  - Matthew Ball Huish
  - Steve Johnson
  - Vicki Wells
  - Steve Wood
  - John DeGrey
  - Dr. Jose Nativi-Nicolau
  - Omar E. Wever-Pinzon
  - Dr. P. Ward
  - Dr. Craig H. Selzman
  - Aaron Healy
  - Devin Horton
  - John Arego
Acknowledgements

- **Cairo University Professors**
  - Reem Bahgat and Aly Fahmy

- **USA University Professors**
  - William Schuler, Hua Xu, Kensaku Kawamoto, and Wendy Chapman

- **Biomedical Informatics Department Administration Staff**
  - Alisha M. Edison-Hair and Tahanie Branz

- **University of Utah Data Warehousing Team**
  - Cheri Hunter, Ming Tu, Vikrant Deshmukh, Paul Sanders, and Mingyuan Zhang

- **University of Utah ITS Team**
  - Jim Turnbull and Travis Gregory

- **KMM/Ken and Blulab/Wendy**
  - Catherine J. Staes from KMM team
  - Lee Christensen from Blulab team
Thank You