Secondary Data Analysis, Big Data Science and Emerging Academic/Corporate Partnerships

Thomas R. Clancy, PhD, MBA, RN, FAAN
Clinical Professor and Associate Dean, Faculty Practices, Partnerships and Professional Development
School of Nursing, The University of Minnesota
Minneapolis, MN
Objectives

• Define big data and factors that are driving its growth
• Discuss the current impact of big data on nursing research and practice.
• Describe common methods used to analyze and create predictive models using big data.
• Discuss differences in education between nurse informaticists and data scientists.
• Review various team models used to conduct big data projects
• Provide an exemplar of a big data study in nursing.
Big Data Drivers

- HITECH Act & Meaningful Use (9.4% - 96.4% EHR since 2008)
- Accountable Care Organizations (744 ACO’s/23.5 Million covered lives)
- CPU Speed/Cost of Memory (25% increase /30% decrease/year)
- Database Architecture (Hadoop/MapReduce – streaming, real time data)
- Cloud Computing & Wireless
- Mobile Health
  - Sensor technology
  - Smart phones
- Intranet (Information doubling every 2 years)
Sources of Data

- Electronic Health Record
- Health Insurance Claims
- Sensor Data (2.9 billion)
- Geo-spatial Data (GPS mapping)
- Intranet of Things (IoT)
- Social Media (1.8 billion subscribers – top 5)
- Patient Reported Outcomes (quantified self movement)
- Human Genome (6 billion/pair)
- Financial Systems (credit cards, bank accounts)
- Environmental and Weather Data
**The Data Trilogy**

**Data Analytics** (DA) is the science of reporting and analyzing raw data with the purpose of drawing conclusions about that information.

**Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms. It is a continuation of data analysis fields such as statistics, machine learning, data mining, and predictive analytics.

**Big Data:** Electronic data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware.

http://www.hissjournal.com/content/2/1/3
Big Data Characteristics

• **Volume** (Terabyte – 1 trillion Gigabytes, laptop is 4G’s)
• **Velocity** (Speed data is stored & accessed – streaming/real time)
• **Variability** (Multiple representation for the same term)
• **Variety** (Structured, unstructured, audio, video, XML, streaming)
• **Veracity** (Accuracy and completeness of data)
• **Visualization** (Ability to visualize patterns/signals in the data)
• **Value** (Capacity of the data to provide value)
Is This Big Data?

• The digital information available on one person's life, encompasses more than the entire Library of Congress.

Benefits of Big Data

• **Explanation**
  • To understand patterns hidden in the data

• **Knowledge Discovery**
  • To extract new knowledge from large, complex data sets

• **Prediction**
  • Generalize patterns and new knowledge to predict outcomes
History of Nursing and Big Data Studies

Secondary Data Analysis
The use of existing data to test new hypotheses or answer new research questions\(^1\).

Nursing Studies\(^2\):
- 1997 to 2003 = 82
- 2003 to 2008 = 99
- 21% increase


Public and Private Databases

Public and Private

- Medicare Claims Use Files (ResDac)
- National Center for Health Statistics (CDC)
- Agency for Healthcare Quality (AHRQ)
  - Medical Expenditure Panel Survey
  - Healthcare Cost and Utilization Data (HCUP)
- Patient Centered Outcomes Research Institute (PCORI)
- Clinical Transformation Science Institute (CTSI)
- Healthcare Cost Institute (HCI)
- OptumLabs Data Warehouse (OLDW)
- Optum Insight
- National Science Foundation Big Data Hubs
- National Institute of Health Clinical Registries
Big Data Research Today

Criteria

• Author: Nurse
• Published in a peer reviewed journal
• Focus: Nursing practice
• Multivariate analysis or contemporary methods
• At least 1 of 3 big data V’s
• Period: 2009 - 2015

Results

• 650 manuscripts (17 nursing informatics journals)
• 17 studies in 18 articles
  • 6 studies: knowledge discover
  • 5 studies: predict process & outcomes
  • 6 studies: impact of technical or nursing interventions on outcomes

Westra et al. Working paper3 Big Data Science: A Literature Review of Nursing Research Exemplars
Where is big data research in Nursing being conducted?

- University of Pennsylvania
- University of Minnesota
- Columbia University
- University of Michigan
- Duke
- Michigan State University
- Cincinnati Children's Medical Center
- University of Alabama
- University of California
Top 10 Non-Profit Health Systems  
(by number of hospitals)

1. Ascension Health—73
2. CHE - Trinity Health - 45
3. Adventist Health System—36
4. Dignity Health -- 34
5. Catholic Health Initiatives — 32
6. Sutter Health (Sacramento) — 26
6. Providence Health and Services — 26
7. CHRISTUS Health — 22
8. UPMC — 20
9. Catholic Healthcare Partners — 17
9. Intermountain Health Care — 17
9. New York-Presbyterian — 17
9. SSM Health Care — 17
10. Banner Health — 16

Themes: Value Based Care

- JC Core Measures (VTE, ECU, Other)
- Nurse Sensitive Indicators (CMS, NDNQI, Magnet)
- CMS Never Events
- Chronic Disease
- Hospital EHR is the primary data source

https://www.google.com/search?q=ndnqi+measures&espv=2&biw=1517&bih=735&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjsu9i477bPAhXI6CYKHSeqB6IQ_AUICCgD&d
National Institute of Nursing Research

- Precision Medicine Initiative ($200 million).
- 2016 National Institute of Nursing Research Strategic Plan (funding areas)

- Symptom Science: Promoting Personalized Health Strategies
- Wellness: Promoting Health and Preventing Illness
- Self-Management: Improving Quality of Life for Individuals with Chronic Conditions
- End-of-Life and Palliative Care: The Science of Compassion
Personalized Health Strategies

- Determine key interceding points in symptom management that can alter (improve or adversely affect) the trajectory of chronic conditions.

Exemplar:
- **Population:** 13 years of EHR data (43,509 diabetic patients)
- **Period:** Tracked patients from baseline to development of T2DM
- **Methods:** Multivariate logistics regression model
- **Results:** Most likely sequence of comorbidities (30%), or “typical” trajectory, was HLD, HTN, IFG and DM
- **Benefit:** Predict phenotypes characteristic of typical and atypical trajectories of T2DM. Clinical trial would have been 13 years..
The investigation of key biological, behavioral, and social factors that promote long-term health and healthy behaviors and prevent the development of disease across health conditions, settings, and the lifespan.


- Population: 500,000 patients; 25,000 nurses; 300 hospitals
- Methods: Survey
- Results: For every 10-percent increase in nurses on staff with bachelor’s degrees, the likelihood of patient death decreased by 7 percent.
Benefits of Big Data: Experimental vs Observational Studies

Clinical Trials

- Long duration
- Expensive
- Small sample size
- Highly dependent on veracity of data.
- Generalizability
Inferential Statistics

Inferential statistical methods infer results from a sample of data to the general population under investigation.

- t-test,
- Analysis of Variance (ANOVA),
- Analysis of Covariance (ANCOVA),
- regression analysis,
- multivariate methods (factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis and others)

http://www.macalester.edu/~kaplan/ISM/testing/15.3.html?access=not-defined&docname=15.3
Observational Studies Using Big Data: Opportunities for AI

Secondary analysis of data
• Short duration
• Less expensive
• Large sample size
• Less dependent on veracity of data
• Need many examples

AI - Machine Learning
• Decision Trees
• Neural Networks
• Bayesian Methods
• Evolutionary Computation
Machine Learning

• The science and technology of systems that learn from data.
• Used to solve complex problems and describe the structure of the data generating processes.

Figure 1: Diagram of a typical learning problem.
# Nine Factor Binary Matrix

One Person = 1,024 Combinations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Ht.</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Wt.</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Pul.</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Inc.</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Edu</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Occ</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Lab</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Xray</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>
The Ill Posed Problem

State space of all possible combinations

\[
\text{Var.} \times \text{Quest.} = \text{Comb.}
\]

\[
2 \times 2 = 2^2
\]

\[
3 \times 2 = 2^3
\]

\[
30 \times 2 = 2^{30}
\]

As accurate as flipping a coin

As accurate as flipping a coin

The more examples (patients) the more accurate the results

Domain experts prime the pump with knowledge and reduce the state space
Machine Learning: Decision Trees

1. Select dataset concepts

2. Remove irrelevant factors
   - Data Scientist
   - Informatician
   - Statistician
   - Domain Experts
   (nurse scientist)

3. Search for conjunctive conjectures

4. Remove negative conjunctive conjectures

5. Build algorithmic rules around conjunctive conjectures

6. Test data on training and hold out group
1. Select concepts

2. Remove irrelevant factors

3. Search for positive conjunctive conjectures

4. Remove negative conjunctive conjectures

5. Build Algorithmic rules around positive conjunctive conjectures

Acute pain + back surgery
Acute pain + abdominal surgery
Acute pain + knee surgery
Acute pain + back surgery + age
Acute pain + abdominal surgery + age
Acute pain + knee surgery + age
Acute pain + back surgery + age + smoking
Acute pain + abdominal surgery + age + smoking
Acute pain + knee surgery + age + smoking
Acute pain + back surgery + age + smoking + cancer
Neural Networks

Cell assembly for a patient in bed with acute pain
Synaptic Neurotransmitters and Action Potentials

Action Potential

http://hyperphysics.phy-astr.gsu.edu/hbase/biology/neurtran.html

Neural Networks

Input Layer

Acute Pain

Back surgery

Age

Smoking

Cancer dx

Hidden Layer

Output Layer

Action Potential

Opioid Use

Cancer dx
Bayes Theorem

• A simple rule for updating your belief in a hypotenuse after you receive new evidence.

• If the new evidence is consistent with your hypotenuse, then the probability of it goes up.
Bayes Theorem

• Estimates the likelihood of an outcome based upon a series of conditional probabilities.
• Excellent formula for classifying diagnosis based upon a patients symptoms.

• What is the probability of flu given fever, cough, aching joints, loss of appetite and nausea?
## Machine Learning Applications

<table>
<thead>
<tr>
<th>Adaptive websites</th>
<th>Game playing</th>
<th>Robot locomotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affective computing</td>
<td>Information retrieval</td>
<td>Search engines</td>
</tr>
<tr>
<td><strong>Bioinformatics</strong></td>
<td>Internet fraud detection</td>
<td>Sentiment analysis (or opinion mining)</td>
</tr>
<tr>
<td><strong>Brain-machine interfaces</strong></td>
<td>Machine perception</td>
<td>Sequence mining</td>
</tr>
<tr>
<td>Cheminformatics</td>
<td>Medical diagnosis</td>
<td>Software engineering</td>
</tr>
<tr>
<td><strong>Classifying DNA sequences</strong></td>
<td>Natural language processing</td>
<td>Speech and handwriting recognition</td>
</tr>
<tr>
<td>Computational finance</td>
<td>Optimization and metaheuristic</td>
<td>Stock market analysis</td>
</tr>
<tr>
<td><strong>Computer vision, including object recognition</strong></td>
<td>Online advertising</td>
<td>Structural health monitoring</td>
</tr>
<tr>
<td>Detecting credit card fraud</td>
<td>Recommender systems</td>
<td>Syntactic pattern recognition</td>
</tr>
</tbody>
</table>

Machine Learning accessed on November 1 at: https://en.wikipedia.org/wiki/Machine_learning
Examples in Nursing

• Classifying data into dashboards
• Classification of data into diagnosis (medical, nursing)
• Optimizing best practices (clinical pathways)
• Comparative effectiveness (drugs, technology, practice)
• Prediction (risk profiling: diabetes, stroke, MI, readmission, pressure ulcers, falls)
• Personalized medicine (genomic, claims, EHR, social media, GPS, wearable technology…)}
What Do You Need to Conduct Big Data Studies?

- Many Examples
- Database Architecture
- Usable Data
- Standardized Terms
- Team Science
Machine Learning Requires Many Examples: Database of 150 Million Records

- Evidence of pharmacy-based initiation of antihypertensive medication: 595,731
- No dialysis prior to treatment initiation: 520,254
- Continuous enrollment for at least 1 year prior to treatment initiation: 266,932
- Treatment initiated within the study period: 192,234

Exclusion Criteria

Cohort Flow Chart
Database Architecture Must be Able to Handle Big Data

Current healthcare database platforms cannot handle the 7 V’s

http://www.slideshare.net/LarryCover/big-data-solutions-for-healthcare
Data Must be Usable
Preprocessing (80% of time)

Preprocessing
• File Transfer
• Data extraction
• Dimensionality reduction
• Data Cleansing
## Data Needs to be Standardized

<table>
<thead>
<tr>
<th>abrasion</th>
<th>excoriation</th>
<th>pressure ulcer</th>
</tr>
</thead>
<tbody>
<tr>
<td>blister</td>
<td>fragile</td>
<td>pressure ulcer(s)</td>
</tr>
<tr>
<td>body piercing</td>
<td>inci</td>
<td>rash</td>
</tr>
<tr>
<td>burn(s)</td>
<td>incision</td>
<td>rash(s)</td>
</tr>
<tr>
<td>cracked</td>
<td>incision(s)</td>
<td>rash(s)</td>
</tr>
<tr>
<td>cut(s)</td>
<td>intact</td>
<td>scab</td>
</tr>
<tr>
<td>cyst</td>
<td>itchy</td>
<td>scar</td>
</tr>
<tr>
<td>drain/device</td>
<td>mass</td>
<td>skin tear</td>
</tr>
<tr>
<td>ecchymosis</td>
<td>other (see comments)</td>
<td>subcutaneous emphysema (specify)</td>
</tr>
<tr>
<td>erosion</td>
<td>petechiae</td>
<td>tattoo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wound</td>
</tr>
</tbody>
</table>
NDNQI - Skin/Pressure Ulcer Information Model

Patient Information
- MRN
- Admission & Discharge Date/Time
- Age
- Gender
- Race/Ethnicity
- Payor
- Temperature
- Color
- Moisture
- Turgor
- Integrity

Skin

Skin Assessment

Risk Assessment
- Date/Time
- Method
- Score
- Risk Status
- Identifier
- Location
- Qualifier
- Category/Stage
- Device Related
- Start/Stop Date/Time

Pressure Ulcer
- Pressure Redistribution Surface
- Routine Repositioning
- Moisture Management
- Nutritional Support

Skin Interventions
- Planned Care
- Moisture Management
- Nutritional Support
- [Actual Care]
Team Science

- Nurse Scientist
  - Deep Domain Knowledge
  - Data Visualization
  - Data Exploration
  - Hypothesis Testing
  - Pattern Discovery
  - Correlations
  - Serendipitous Discovery

- Data Analyst
- Project Manager
- Statistician
- Data Scientist
- Data Engineer
- Geneticist
- Nurse Scientist
- Predictive Modeler
Big Data: Top 20 Skills for a Data Scientist

Top 20 Backgrounds: Data Scientists

<table>
<thead>
<tr>
<th>PhD Nursing</th>
<th>Masters in Data Science</th>
<th>Doctor of Nursing Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>School of Nursing</td>
<td>Department of Computer Science and Engineering, Department of Electrical and Computer Engineering, School of Statistics and Division of Biostatistics.</td>
<td>Specialty: Nursing Informatics University of Minnesota School of Nursing</td>
</tr>
<tr>
<td>Core Courses</td>
<td>Core Courses</td>
<td>Core Courses</td>
</tr>
<tr>
<td>Principles of Database Systems</td>
<td>HINF 5510 Applied Health Care Databases: Database Principles and Data Evaluation</td>
<td>Statistics**</td>
</tr>
<tr>
<td>Graduate Statistics Course I</td>
<td>Applied Regression Analysis</td>
<td></td>
</tr>
<tr>
<td>Graduate Statistics Course II</td>
<td>STAT 5401 - Applied Multivariate Methods</td>
<td></td>
</tr>
<tr>
<td>Nursing and Nursing Theory Core</td>
<td>CSCI 5523 - Introduction to Data Mining</td>
<td>Nurs 6105 Systems Analysis and Design</td>
</tr>
<tr>
<td>NURS 8134: Interventions and Outcomes</td>
<td>EE 5239 - Introduction to Nonlinear Optimization</td>
<td>Nurs 7400 Health Policy Leadership</td>
</tr>
<tr>
<td>NURS 8172: Theory and Theory Development for Research</td>
<td>Elective</td>
<td>Nurs 5116 Consumer Health Informatics</td>
</tr>
<tr>
<td>Research Methodology Core</td>
<td>Elective</td>
<td>Nurs 6200 Science of Nursing Intervention</td>
</tr>
<tr>
<td>NURS 8171: Qualitative Research Design and Methods</td>
<td>Capstone Project (first half)</td>
<td>Nurs 7600 Nursing Research and Evidence Based Practice</td>
</tr>
<tr>
<td>NURS 8173: Principles and Methods of Implementing Research</td>
<td>Capstone Project (second half)</td>
<td>Nurs 7113 Clinical Decision Support: Theory</td>
</tr>
<tr>
<td>NURS 8175: Quantitative Research Design and Methods</td>
<td></td>
<td>Nurs 7105 Knowledge Representation and Interoperability</td>
</tr>
<tr>
<td>NURS 8177: Research Practicum</td>
<td></td>
<td>HINF 8406 User Interface Design and Usability in Healthcare</td>
</tr>
<tr>
<td>NURS 8180: Doctoral Pro-Seminar: Scholarly Development</td>
<td></td>
<td>Nurs 7200 Economics of Health Care</td>
</tr>
<tr>
<td>NURS 8152: Scholarship in Healthcare Ethics</td>
<td></td>
<td>Nurs 7112 DNP Project Direction III: Evaluation</td>
</tr>
<tr>
<td>NURS 8190: Critical University of Minnesota Review of Health Research</td>
<td></td>
<td>Nurs 7108 Population Health Informatics</td>
</tr>
<tr>
<td>Nursing Electives (1 or 2 courses; see below for examples)</td>
<td></td>
<td>NURS 6110 Epidemiology in Nursing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nurs 7610 Health Innovation and Leadership</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nurs 7202 Moral and Ethical Positions and Actions in Nursing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nurs 5115 Interdisciplinary Healthcare Informatics</td>
</tr>
</tbody>
</table>
Difference In Nurse and Data Scientist Training

- Mathematics for modeling.
- Big data framework
- Software programming
- Data munging/ingestion
- Machine learning
- Data visualization tools
- Evaluation methods

http://nirvacana.com/thoughts/becoming-a-data-scientist/
Filling the Short Term Gaps

- Hire faculty with expertise into your department and/or consult with faculty in other departments.
- Internal consultation service (CTSI)
- External research collaborative
- Contract an external consulting service

Data Science
- Computer Science
- Engineering
- Epidemiology
- Statistics
- Physics
- Mathematics
- Information Technology
Academic/Industry Partner

- **Exploratory Sandbox**
  - Statistical Tools are available
  - May add additional software applications (SAS, MatLab, R)
  - Data is statistically de-identified and cannot leave the sandbox
  - Multiple partners may work on the same project simultaneously

---

**Team**
- Project Mgt
- Domain Ex
- Machine Learn
- Data Dic. Analyst

**Data**
- Explore
- Research Views
  - Unified (claims/EHR)
  - Death Index
  - SES (social/economic)

**AHC**
- Medicine
- Pharmacy
- Public Health

**Optum Labs**
- Project Mgr
- Data Engineer
- Data Dic. Analyst

**UM Nursing**

**OLDW 150M Lives**

---

**Optum Labs**
- Project Mgr
- Data Engineer
- Data Dic. Analyst
<table>
<thead>
<tr>
<th>Partners</th>
<th>Sample of Academic/Industry Partnerships</th>
<th>Funding Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Otolaryngology</td>
<td>Prediction model: causal factors in patients presenting with dizziness</td>
<td>NIH</td>
</tr>
<tr>
<td>Nursing</td>
<td>Prediction model: Patients experiencing adverse effects of statin therapy</td>
<td>UM Internal</td>
</tr>
<tr>
<td></td>
<td>Prediction model: Cardiovascular disease risk prediction using EHR/claims data</td>
<td>UM Internal AHC Seed</td>
</tr>
<tr>
<td></td>
<td>Symptom management of liver transplant patients</td>
<td>RO1</td>
</tr>
<tr>
<td></td>
<td>Prevention of urinary tract infections in young women</td>
<td>RO1</td>
</tr>
<tr>
<td>Public Health</td>
<td>Prediction model: Diffusion of knowledge from clinical trials to practice.</td>
<td>NIH</td>
</tr>
<tr>
<td></td>
<td>Comparative effectiveness of extended oral anticoagulant use</td>
<td>PCORI</td>
</tr>
<tr>
<td></td>
<td>Contemporary Venous Thromboembolism Treatment - NIH</td>
<td>NIH</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>Comparative effectiveness between surgical and nonsurgical intervention of low back pain.</td>
<td>NIH</td>
</tr>
</tbody>
</table>
Summary
Before You Embark on a Big Data Study

• Determine study aims and questions first.
• What data elements do you need?
• Where are the best source(s) of data?
• How clean is the data?
• Are the data elements coded (ICD10, CPT, LOINC, SNOMED)
• What is the study design: (explanatory, knowledge discovery, predictive)
• What are the best study methods (machine learning, statistics)
• Are experts available (who is on the team)?
• Is funding available?
Purpose and Specific Aims of Study

- **Purpose**
  - To develop a predictive model for hospital-acquired CAUTIs using multiple data sources

- **Specific Aims**
  - **Aim 1**: Create a quality, de-identified dataset combining multiple data sources for machine learning tasks
  - **Aim 2**: Develop and evaluate predictive models to find the best predictive model for hospital-acquired CAUTI

Research Design

Steps in the Analysis

1. Data Selection
2. Data Preprocessing
3. Data Transformation
4. Data Mining
5. Interpretation / Evaluation
6. Knowledge Discovery

Aim 1
Create the dataset

Aim 2
Develop & evaluate the predictive model

Data Selection

Total number of unique ICU admissions for final analysis – 11,226

Total number of patients in three ICUs – 8,496

1,292 patients with multiple hospitalizations

Total unique hospitalizations – 10,420

Several ICU admissions during one hospitalization
Data Quality:
Data Preprocessing and Transformation

Missing data
- Race → “unknown”
- Immunosuppressive agents → “No immunosuppressive agents”
- Out of range lab results → “No”
- Pre-existing urinary catheter → imputed using the \( k \)-nearest neighbors
- Rationale for continued use of catheter → “Not having rationale”

Data transformation
- Age: Continuous → Categorical
- Use of immunosuppressive agents: Nominal → Binary
- Charlson index score: continuous → Categorical
- Lab result - glucose: Continuous → Binary
- Rationale for continued use of catheter: Nominal → Binary

Attributes with no variance removed
- Race, WBC lab results, surgical procedure history,
- and total RN hours per patient day
Machine Learning: The Model Results from Decision Trees

## The Odds Ratios of Attributes from LR

<table>
<thead>
<tr>
<th>Attributes</th>
<th>ORs</th>
<th>Attributes</th>
<th>ORs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young Adult</td>
<td>0.29</td>
<td>Hospitalization within Previous 6 Months</td>
<td>0.65</td>
</tr>
<tr>
<td>Middle-aged Adult</td>
<td>0.69</td>
<td>Length of Stay</td>
<td>1.07</td>
</tr>
<tr>
<td>Young-old Adult</td>
<td>1.59</td>
<td>Lab Result – Glucose &gt; 200 mg/dl</td>
<td>1.13</td>
</tr>
<tr>
<td>Old-old Adult</td>
<td>1.51</td>
<td>Pre-existing Urinary Catheter</td>
<td>0.57</td>
</tr>
<tr>
<td>Male Gender</td>
<td>0.21</td>
<td>Rationale for Continued Use of Catheter</td>
<td>14.96</td>
</tr>
<tr>
<td>Immunosuppression</td>
<td>0.27</td>
<td>Total Nursing Hours per Patient Day</td>
<td>0.83</td>
</tr>
<tr>
<td>Charlson Index Score = 0</td>
<td>0.94</td>
<td>Percent of Direct Care RNs with Associate’s Degree in Nursing</td>
<td>1.04</td>
</tr>
<tr>
<td>Charlson Index Score = 1–2</td>
<td>0.96</td>
<td>Percent of Direct Care RNs with BSN, MSN, or PhD</td>
<td>1.04</td>
</tr>
<tr>
<td>Charlson Index Score ≥ 3</td>
<td>1.07</td>
<td>Percent of Direct Care RNs with Specialty Nursing Certification</td>
<td>0.99</td>
</tr>
</tbody>
</table>
# Triangulation: Results from the Predictive Models

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Number of FN</th>
<th>Accuracy (%)</th>
<th>Sensitivity TP/(TP + FN)</th>
<th>Specificity TN/(TN + FP)</th>
<th>Precision TP/(TP + FP)</th>
<th>ROC area</th>
<th>Clinical Interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DT</strong></td>
<td>200</td>
<td>TP: 54, FN: 13, FP: 2695, TN: 8434</td>
<td>75.87</td>
<td>0.81</td>
<td>0.76</td>
<td>0.02</td>
<td>0.78</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>LR</strong></td>
<td></td>
<td>TP: 50, FN: 17, FP: 2696, TN: 8463</td>
<td>75.83</td>
<td>0.75</td>
<td>0.76</td>
<td>0.02</td>
<td>0.85</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td></td>
<td>TP: 54, FN: 13, FP: 3186, TN: 7973</td>
<td>71.50</td>
<td>0.80</td>
<td>0.71</td>
<td>0.02</td>
<td>0.76</td>
<td>No</td>
</tr>
</tbody>
</table>

## Factors Associated with CAUTI

<table>
<thead>
<tr>
<th>DT model</th>
<th>LR model</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Female</td>
<td>• Age (≥56)</td>
</tr>
<tr>
<td>• Longer length of Stay</td>
<td>• Longer length of stay</td>
</tr>
<tr>
<td>• Presence of rationale for continued use of catheter</td>
<td>• Presence of rationale for continued use of catheter</td>
</tr>
<tr>
<td>• Less total nursing hours per patient day</td>
<td>• Charlson comorbidity index score ≥ 3</td>
</tr>
<tr>
<td>• Lower percent of direct care RNs with specialty nursing certification</td>
<td>• Glucose lab result &gt; 200 mg/dl</td>
</tr>
<tr>
<td>• Higher percent of direct care RNs with BSN, MSN, or PhD</td>
<td>• Higher percent of direct care RNs with associate’s degree in nursing</td>
</tr>
</tbody>
</table>

---

Resources Page

• NIH Big Data to Knowledge (BD2K) Workshops: https://datascience.nih.gov/bd2k/events/bd2kworkshops
• NINR Advancing Nursing Research through Data Science http://www.ninr.nih.gov/training/online-developing-nurse-scientists#.VtdHJvkrLIU
• American Medical Informatics Association: https://www.amia.org/
• Health Information and Management Systems Society (HIMSS): http://www.himss.org/aboutHIMSS/
• Coursera: Six courses on data science: https://www.coursera.org/
• Health Catalyst Knowledge Center: https://www.healthcatalyst.com/knowledge-center/
Questions?

Thomas R. Clancy, PhD, MBA, RN, FAAN
clanc027@umn.edu