Analytical Challenges in the Era of Big Data

Alvin D. Jeffery, PhD, RN

Post-Doctoral Fellow

United States Department of Veterans Affairs

Nashville, Tennessee, USA

Sigma Theta Tau International's

28th International Nursing Research Congress

Dublin, Ireland – July 2017

This material is based upon work supported by the Office of Academic Affiliations, Department of Veterans Affairs, VA National Quality Scholars Program and with resources and the use of facilities at VA Tennessee Valley Healthcare System, Nashville, TN. Support for travel was provided by the Iota Chapter of Sigma Theta Tau International.

Objectives

- List at least 2 analytical challenges encountered within large datasets
- Describe at least 2 solutions to analytical challenges encountered within large datasets

Background

- “Big Data” is increasingly popular
- Predictive Analytics holds great promise for nursing care delivery
- *However,* large datasets are not the panacea many would taut
- Solutions to challenges are not always evident to clinician subject matter experts

Challenges

Data Acquisition & Management

- Ethics approval
- Ensuring individual patient privacy
- Preventing undesired user access
- Collecting & storing “big data”

Missing Data

- All large datasets contain some degree of missing data
- Must find cause of missingness
- Each imputation approach has advantages & disadvantages

Challenges (cont'd)

Statistical Model Assumptions

- *Many* statistical models & machine learning techniques exist!

Model Evaluation

- Evaluating expected future performance
- Bias/Variance Trade-Off
- Clinicians might not understand the model

"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."

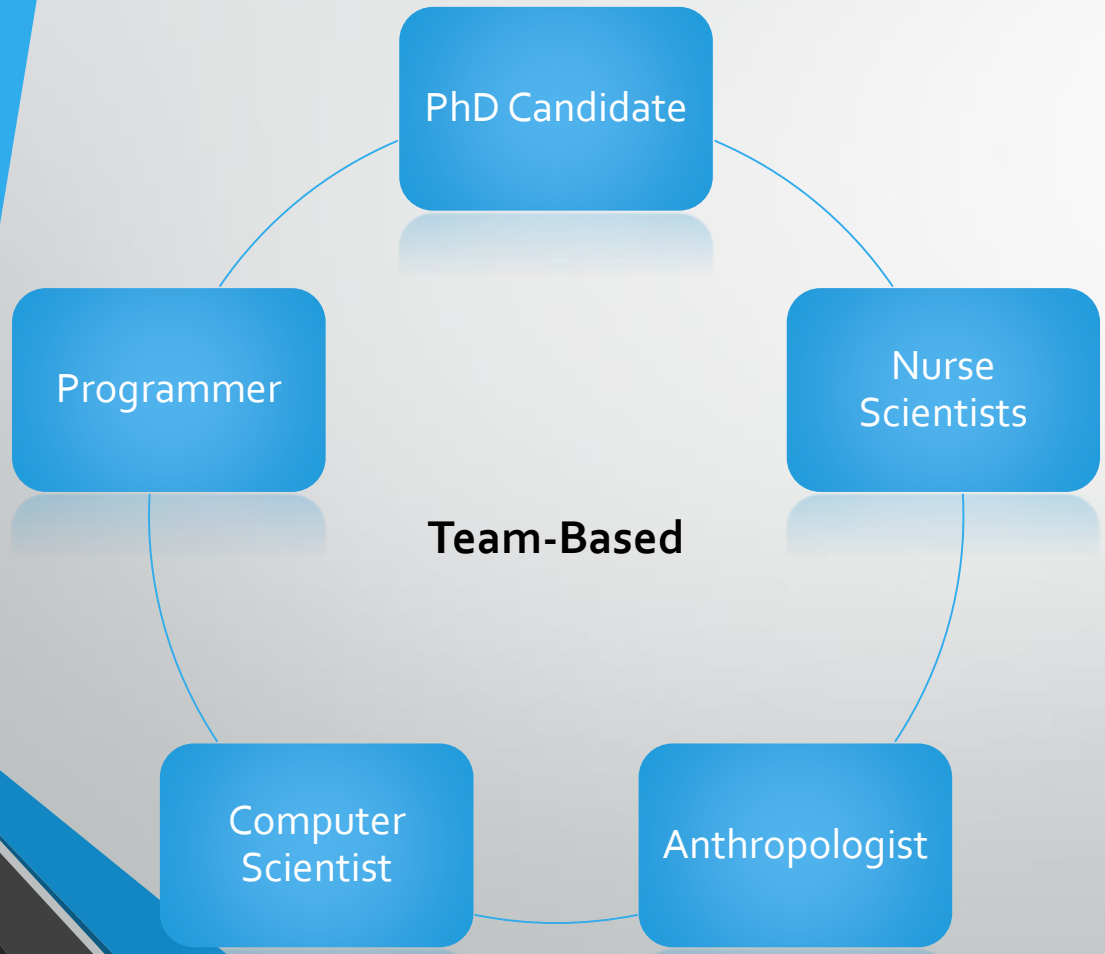
Atul Butte, MD, PhD

Associate Professor of Pediatrics, Stanford

Exemplar

- Create a prediction model for in-hospital cardiopulmonary arrest
- ~170,000 adult patients
- Predictors: demographics, vital signs, laboratory values, billing codes
- De-identified electronic health record

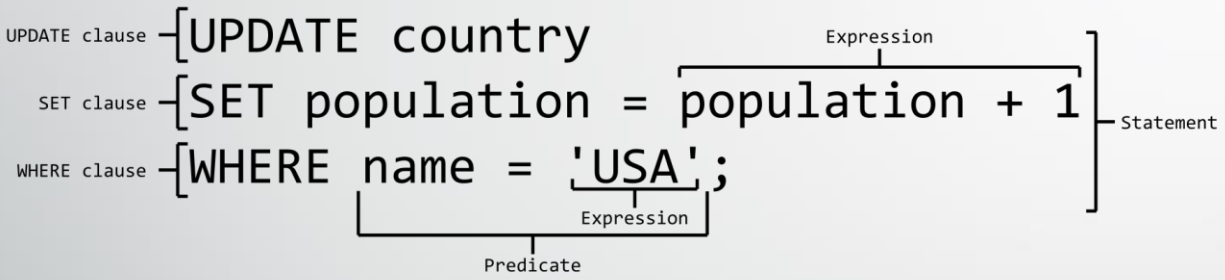
Data Acquisition & Management: Exemplar



Local Server

Data Acquisition & Management: Other Solutions

Structure Query Language (SQL)



Secure Servers

amazon web services

Google Cloud Platform

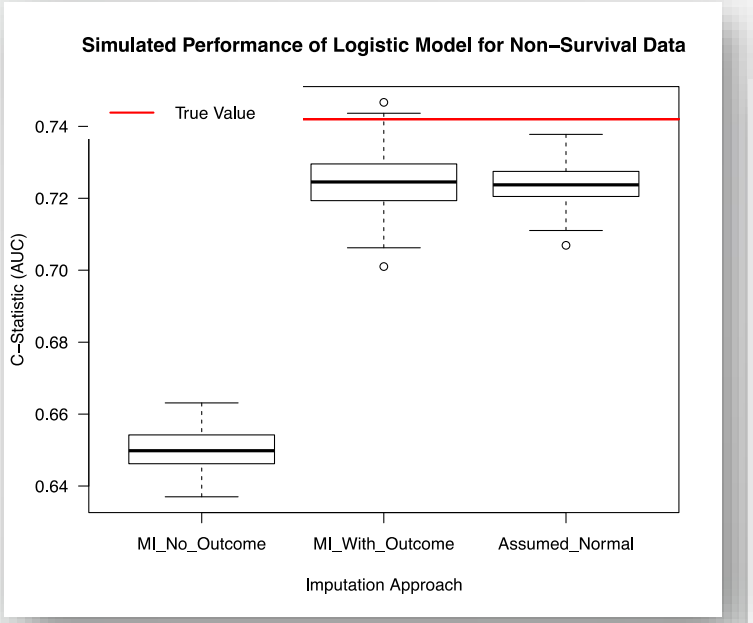
Many free online courses (MOOCs) available, for example…

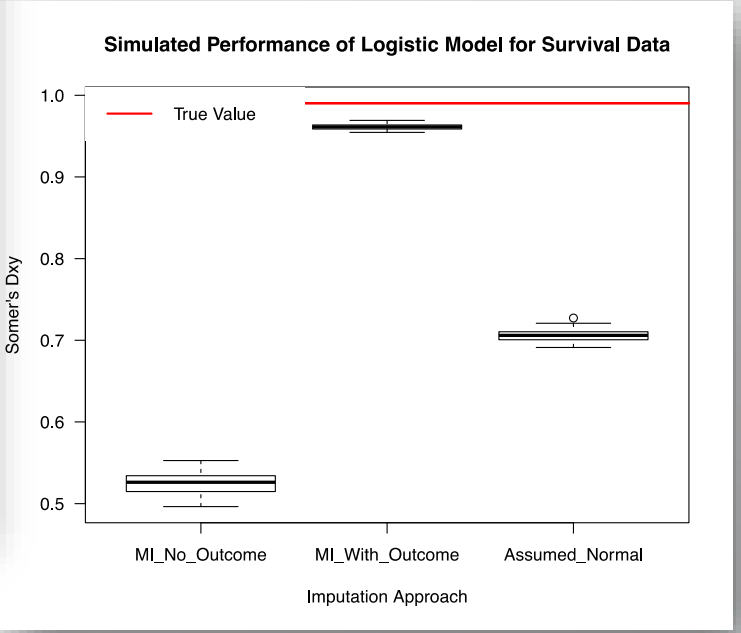
coursera

w3schools.com

Missing Data: Exemplar

- Lab Values & Vital Signs missing in 40-60% of cases
- Conducted simulation studies (10 million patients) to identify preferred imputation approach





Leveraging Statistical Simulation Studies to Gain Insights from Data:

A New Type of Simulation for Nurses

ALVIN D. JEFFERY, PHD, RN | U.S. DEPARTMENT OF VETERANS AFFAIRS | TENNESSEE VALLEY HEALTHCARE SYSTEM, NASHVILLE, TN, USA

Missing Data: Other Solutions

	Advantages	Disadvantages
Complete Case	Simple	*Highly* biased (unless MCAR) Loss of power
Median Imputation	Simple Easy to implement prospectively	Likely biased Unable to represent uncertainty
Multiple Imputation	Minimal bias Accounts for uncertainty	Computationally intense Challenging to implement prospectively

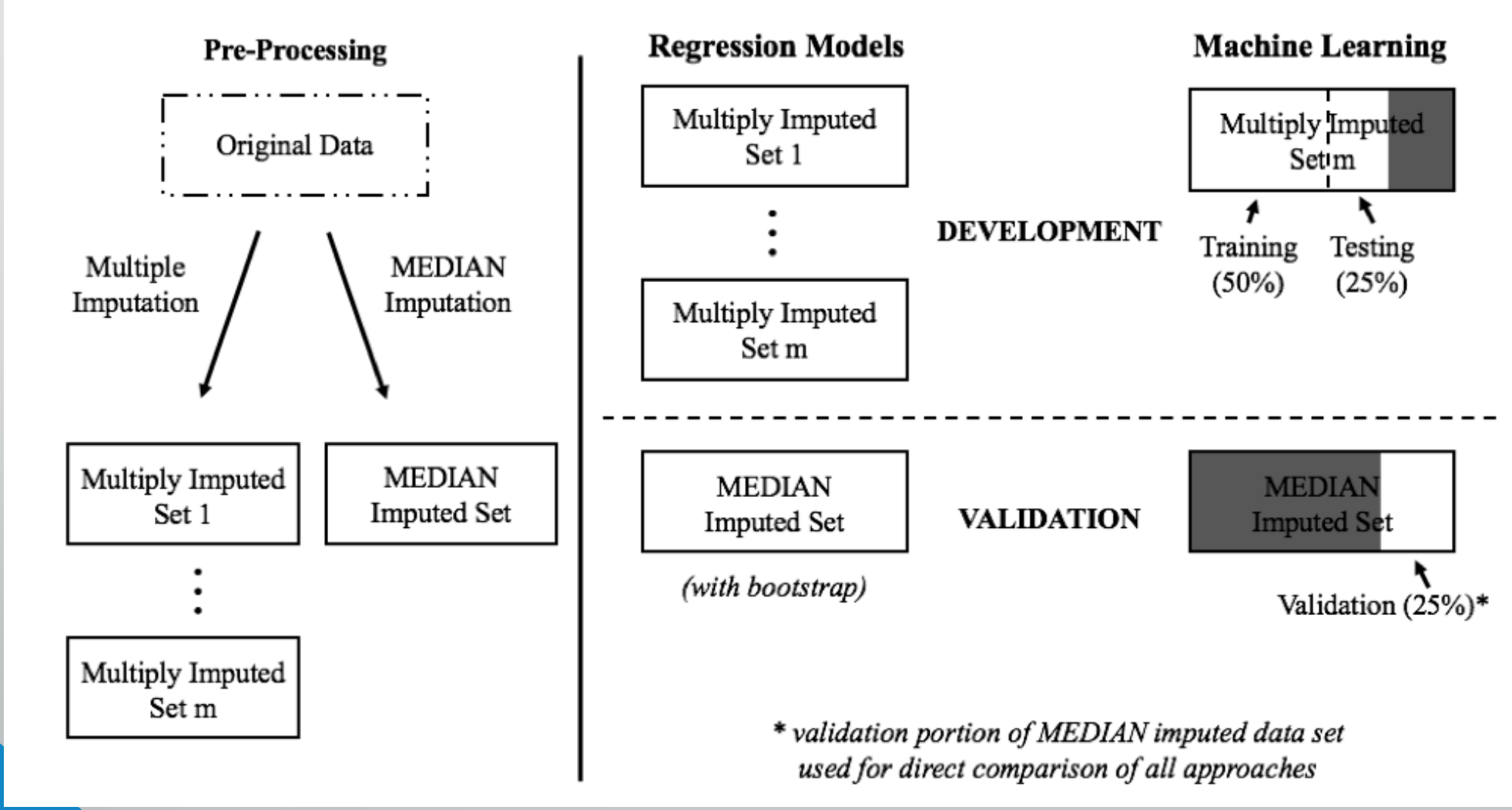
Statistical Model Assumptions: Exemplar

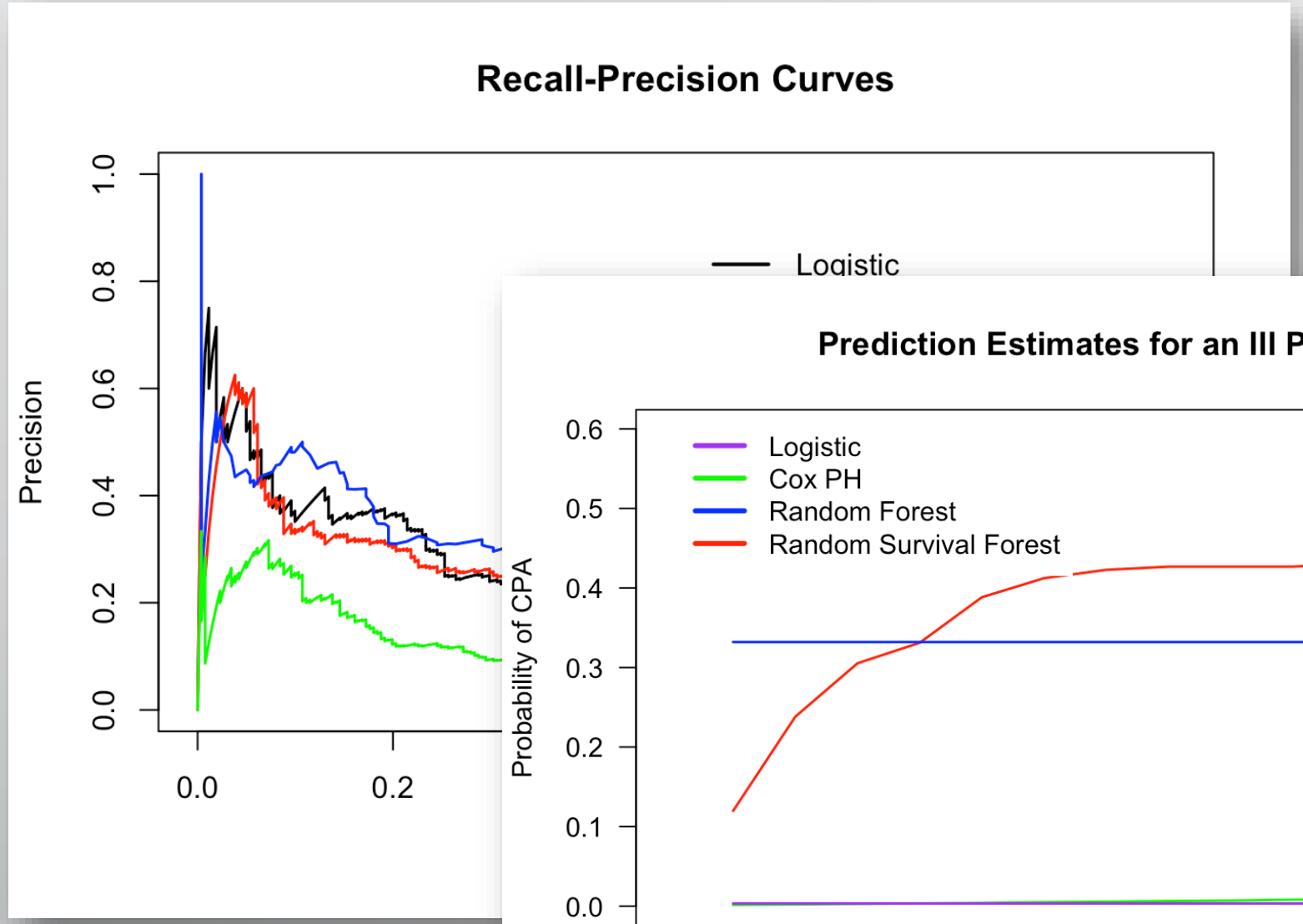
	Approach	
Purpose	*Statistical*	*Machine Learning*
Classification Predicts whether an event will occur	Logistic Regression	Random Forest
Survival/Time-to-Event Predicts how likely an event is at each time point	Cox Proportional Hazards Regression	Random Survival Forest

Statistical Model Assumptions: Other Solutions

	Examples	Advantages	Disadvantages
Traditional Statistics	Linear Models (e.g., OLS) Generalized Linear (e.g., logistic) Time-to-event (e.g., Cox)	Available in many statistical packages	Many assumptions to meet Can be slow
Machine Learning	k-nearest neighbors Naïve Bayes Decision Trees (e.g., random forests) Neural Networks	Handle large data quickly Fewer assumptions to meet	Less interpretable Less familiarity within healthcare community

Model Evaluation: Exemplar




Recall-Precision Curves
Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
Logistic

Prediction Estimates for an Ill Patient
Probability of CPA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Logistic
Cox PH
Random Forest
Random Survival Forest
0
2
4
6
8
10
12
14
Hospital Day

Model Evaluation: Other Solutions

Validation

- External:
  - Split-Sample
  - New Setting
  - Chronological
- Internal:
  - k-fold cross validation
  - Bootstrap

Interpretation

- Predictor Importance
- Partial Dependence Plots
- Graphs vs. Numbers

Conclusion

- Clinically meaningful big data insights require multifaceted expertise & teamwork
- Nurses & other clinicians are equipped to identify problems "big data" can help solve
- **As nurses become more knowledgeable of big data research challenges & solutions, they position themselves to be leaders in research teams & advocates for implementation of novel findings**

Thank you!

alvinjeffery@gmail.com

quantitativenurse.com

Twitter: @Nurse_Alvin

References

- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123-1131. doi: 10.1377/hlthaff.2014.0041
- Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience, 5*, 12. doi: 10.1186/s13742-016-0117-6
- Harrell, F. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York, NY: Springer.
- *The routledge international handbook of advanced quantitative methods in nursing research*. (2016). (S. J. Henly Ed.). New York, NY: Routledge, Taylor & Francis Group.
- Steyerberg, E. W. (2009). *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer.
- van der Heijden, G. J., Donders, A. R., Stijnen, T., & Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology, 59*(10), 1102-1109. doi: 10.1016/j.jclinepi.2006.01.015

Recommended Resources