Title:
Analytical Challenges in the Era of Big Data

Alvin D. Jeffery, PhD
Geriatric Research Education Clinical Center, TN Valley Healthcare System, U.S. Department of Veterans Affairs, Nashville, TN, USA

Session Title:
Use of Big Data to Influence Nursing Care
Slot:
H 15: Saturday, 29 July 2017: 8:30 AM-9:15 AM
Scheduled Time:
8:50 AM

Keywords:
big data, clinical decision support and quantitative methods

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

Abstract Summary:
This presentation will use an exemplar to illustrate, and propose solutions to, commonly encountered problems in large clinical datasets, such as data acquisition/management, missing data, statistical model assumptions, and model evaluation.
Learning Activity:

LEARNING OBJECTIVES	EXPANDED CONTENT OUTLINE
List at least 2 analytical challenges encountered within large datasets.	The presenter will describe 4 analytical challenges encountered in a large dataset used

	to build a prediction model for in-hospital cardiopulmonary arrest.
Describe at least 2 solutions to analytical challenges encountered within large datasets.	Leveraging an exemplar of developing a prediction model, the presenter will provide a brief overview of more than a dozen possible solutions to analytical challenges within large datasets.

Abstract Text:

Purpose: The popularity of “big data” along with an increasing capacity for real-time predictive analytics holds significant promise for nurses and other clinicians to gain new insights and develop novel decision support tools from our large clinical datasets. Unfortunately, these large datasets are not the panacea that some big data proponents would taut. For nurses with vast subject matter expertise in a clinical area who desire to leverage big data for solving practical problems, road blocks quickly surface in the form of acquisition and management of data, missing data, meeting assumptions of statistical models, and model evaluation for statistical and clinical performance. This talk will engage the audience in addressing these issues using an exemplar of the development of a prediction model for in-hospital cardiopulmonary arrest.

Methods: The following 4 topics will be addressed:

Data Acquisition and Management: From ethics approval to ensuring individual patient privacy to preventing undesired user access, collecting and storing “big data” is no simple task. The presenter will provide: (a) an overview of key concepts, (b) an exemplar for constructing a data acquisition and management team, and (c) several resources for learning more independently.

Missing Data: Almost all large datasets contain some amount of missing data. Regardless of the amount, finding the cause of missingness is of paramount importance. Approaches to determining a cause will be introduced, and disadvantages of complete case analysis will be described. Advantages and disadvantages of median imputation, multiple imputation, and machine learning imputation will be compared.

Statistical Model Assumptions: There are a variety of statistical models available, and with recent advances in machine learning methods, more approaches to retrieve information from the data are available to a wide array of users. An overview of the purpose and requirements of traditional modeling (e.g., logistic and linear regression) and machine learning approaches (e.g., random forests and cluster analyses) will be provided.

Model Evaluation: Determining how well a model performs on the current data and how well it is expected to perform on future data is essential in determining whether or not the model is helpful for clinical care. Internal (e.g., bootstrapping and cross-validation) versus external validation (e.g., split sample and chronological validation) techniques will be presented along with their respective advantages and disadvantages.

Results: Our in-hospital cardiopulmonary arrest prediction model required a team-based approach to solving the aforementioned challenges, and the audience will hear not only how we chose to solve the problems but also other approaches we considered. From the perspective of data acquisition/management, we found the best approach to be the inclusion of database and informatics specialists who used structured query language to extract the relevant data and then store it on a secure, organizational server. Following a simulation study, we discovered the missing data problem was best resolved by creating a multiple imputation model that included the outcome variable. Statistical model assumptions were best met by not assuming linearity while not permitting too many spline knots. Model

evaluation comprised internal bootstrap validation for the regression models and split-sample validation for the machine learning methods.

Conclusion: Arriving at clinically meaningful insights contained within large datasets requires multifaceted expertise and teamwork. Nurses and other clinicians are the best members of the team to identify a problem that “big data” can help solve. To ensure a clinically meaningful solution surfaces from big data efforts, nurses should be aware of common challenges in big data research. As nurses become more knowledgeable, they position themselves to be leaders in these research teams and advocates for implementation of novel findings.