**45th Biennial Convention (16-20 November 2019)**

**Learning to Do Data Science Healthcare Research**

**Lana Pasek, MSN**
*School of Nursing, SUNY@Buffalo School of Nursing, Buffalo, NY, USA*

**Purpose:** Data science describes the process of asking questions, analyzing, and manipulating large data sets in the search for patterns and knowledge by using diverse mathematical models and methods (Sullivan, 2018). With the great advances of technology and computer science, scientists now have quick access to large amounts of data from a variety of sources. These sources are digitized and can be monetary as in credit card/banking information or even the most personal from Electronic Medical Records. Concurrently there has been increased capacity to store and retrieve these huge amounts of data electronically. The speed of delivering all this data by electronic transmission is something never experienced before. This speed coupled with organization and labeling of the data within software programs makes it easy to use the data for different analyses.

Health care research focusing on improved managing and delivery of care is evolving quickly with the accessibility of these large data bases. But this requires interdisciplinary teamwork of computer software engineers, mathematics and statistics, and those with substantive area expertise, like nurses (Grus, 2015). Nurses are experts in clinical knowledge and wisdom. Clinical wisdom comes from the general truths and principles that emerge from the understanding and intimate knowledge of clinical practice (Benner, Hooper Kyriakidis, & Stannard, 2011). Therefore they are ideal to contribute to data science health research and to learn how to perform data science processes.

**Methods:** Data science research starts with a research question. With this research question is an understanding of its significance in solving a problem and the background of the variables involved interacting as predictors of the outcome, the literature would have data on the problem's incidence and prevalence in healthcare. The next step is to define the data the researcher needs to analyze. These definitions include being explicit as what data warehouse or government or private database from which your data was obtained and the dates explored. Defined data elements are located within a data registry which can be from the clinical codes such as the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10). The data registry also includes the data to be excluded such as certain age ranges or conditions like pregnancy. The strategy is to request broadly to get as much data as think you need initially and through data cleaning the duplicates and data not related to the research question can be eliminated. The data is then normalized for the research question with certain data flagged such as deaths. Next in the process is the analytic database to be established from which the data is described with all the independent and dependent variables and the data science approaches are applied. There are numerous data science approaches and the most important of healthcare data analytics is clinical prediction. These approaches try to uncover the underlying relationship between attributes and a dependent variable and include regression models and survival analyses, your mathematics and statistics (Reddy & Yi, 2015). Data science approaches, therefore, help data scientists in identifying patterns in these massive data sets to answer questions such as, "What is odd about my data?" (anomaly detection) or "What will happen next?" (predictive analytics) (Sullivan, 2018).

**Results:** The results from the different approaches can be compared for hypothesis testing and to start to understand what is going on in relation to the research question. Nurses use data science to uncover knowledge in large datasets (usually Electronic Health Record data). They determine how large data sets can define new meaning for patient care by discovering patterns, associations, or factors related to patient outcomes. Predictive models were developed by nurses to identify risks for adverse health outcomes such as infection or mortality. Moreover, nurses use data science to develop, assess, and evaluate patient outcomes by way of clinical decision support tools, health portals, and care coordination activities (Westra, et al., 2017; Delany & Weaver, 2018). Lessons learned from performing data science research are that some data may be missing and the need to re-think the data science approach (but there are many to choose from), difficulty in filtering a population, or discovery of an incomplete data registry.

---

**Title:**
Learning to Do Data Science Healthcare Research

**References:**
Benner, P., Hooper Kyriakidis, P., & Stannard, D. (2011). *Clinical Wisdom and Interventions in Acute and Critical Care A Thinking-in-Action Approach (2nd Ed.).* New York: Springer Publishing.
Delaney, C.W., & Weaver, C. (2018). 2018 Nursing knowledge big data science initiative. *CIN: Computer, Informatics, Nursing*, 36(10), 473-474.
Grus, J. (2015). *Data Science from Scratch.* Sebastopol, CA: O'Reilly Media Inc.
Reddy, C.K. & Li, Y. (2015). A review of clinical prediction models. In C. K. Reddy & C.C. Aggarwal (Eds.), *Healthcare Data Analytics* (pp. 344-371). New York: CRC Press.
Sullivan, S. (2018). Strength in numbers: Using data science to advance nursing knowledge. Retrieved from University of New York at Buffalo website: http://nursing.buffalo.edu/news-events/nurses-report.host.html/content/shared/nursing/articles/nurses-report/posts/big-data-science-analytics.detail.html
Westra, B.L., Sylvia, M., Weinfurter, E.F., Pruinelli, L., InPark, J., Dodd, D., Keenan, G.M., Senk, P., Richesson, R.L., Baukner, V., Cruz, C., Gao, G., Whittenburg, L., Delaney, C.W. (2017). *Big data science: A literature review of nursing research exemplars*. Nursing Outlook, 65(5), 549-561.

**Abstract Summary:**
Data science research is a growing discipline. It involves engaging core multidisciplinary teams of computer science, math and statistical techniques and substantive knowledge in order to solve healthcare related problems. Nurses have a core role because of distinct expertise in healthcare and can learn the processes of performing this research.

**Content Outline:**
I. Introduction

A. Increasing prominence of data science
1. Access/handling
2. Storage/retrieval
3. Delivery/analyzing
B. Healthcare research evolving
1. Manage care
2. Delivery of care
C. Interdisciplinary teams
1. Computer science
2. Math/statistics
3. Substantive area expertise
II. Body
A. Research Question
1. Background/Significance
a. incidence
b. prevalence
B. Data Definitions
1. Sources and dates of data
2. Data Registry
C. Cleaning/Extracting data
D. Analytic Database
E. Analysis
1. Descriptives
2. Data science approaches
III. Results
A. Compare results of different data science approaches
B. Lessons learned

First Primary Presenting Author
*Primary Presenting Author*
Lana Pasek, MSN
SUNY@Buffalo School of Nursing
Adjunct faculty
Buffalo NY
USA

**Author Summary:** Adult Nurse Practitioner with over 16 years of experience in acute, specialty and sub-acute care. Masters in Science Education and Nursing with leadership in coordinating a nursing research committee, and presentations of original research internationally, nationally and locally. Certified by the American Nurses Credentialing Center, American Association of Critical Care Nurses and CITI program for Biomedical Research Investigators.