Nurs Admin Q
Vol. 39, No. 4, pp. 311–318

The Veterans Affairs's Corporate Data Warehouse

Uses and Implications for Nursing Research and Practice

Lauren E. Price, BSN; Kimberly Shea, PhD; Sheila Gephart, PhD

The Department of Veterans Affairs Veterans Healthcare Administration (VHA) is supported by one of the largest integrated health care information systems in the United States. The VHA's Corporate Data Warehouse (CDW) was developed in 2006 to accommodate the massive amounts of data being generated from more than 20 years of use and to streamline the process of knowledge discovery to application. This article describes the developments in research associated with the VHA's transition into the world of Big Data analytics through CDW utilization. The majority of studies utilizing the CDW also use at least one other data source. The most commonly occurring topics are pharmacy/medications, systems issues, and weight management/obesity. Despite the potential benefit of data mining techniques to improve patient care and services, the CDW and alternative analytical approaches are underutilized by researchers and clinicians. **Key words:** *Big Data, Corporate Data Warehouse, Department of Veterans Affairs, data mining, nursing*

AS OF 2012, 2.5 exabytes of data were being generated globally each day, with a projected annual growth rate of 40%.[1-3] To put this into perspective, 5 exabytes of data are equivalent to 37 000 libraries the size of the current Library of Congress.[4] The term used to describe this trend in the rapid evolution and expansion of available data is "Big Data." Big Data refers to the high-velocity, large-volume, and enormous variety of data being generated as the result of the advent of new technologies.[1] Big Data requires extensive technology infrastructure and analytical tools to generate knowledge for informing research and practice.[5] In 2014, the American Academy of Nursing beseeched health care organizations to support and implement a sustainable system for nursing utilization of Big Data in research and practice.[6]

In health care, Big Data emerged mainly as the result of the implementation of electronic health records (EHRs), which were conceptualized as early as the 1960s. In the 1970s, the Department of Veterans Affairs Veterans Healthcare Administration (VHA) became one of the first adopters of the EHR.[7] In the 1980s, the VHA instituted the information technology infrastructure that it still uses today, known as the Veterans Health

Author Affiliation: *The University of Arizona College of Nursing, Tucson.*

Ms Price acknowledges the Southern Arizona VA Healthcare System librarians, Lynn Flance and Karen Douglas, who aided with the supplemental MEDLINE searches for this project. Dr Gephart acknowledges funding from the Robert Wood Johnson Foundation Nurse Faculty Scholars Program and the Agency for Healthcare Research and Quality (grant #K08HS022908). The content is solely the responsibility of the authors and does not represent the official views of the Agency for Healthcare Research and Quality or the Department of Veterans Affairs, Veterans Healthcare Administration.
Ms Price is a VA employee. No conflicts of interest for Dr Gephart or Shea are identified.

Correspondence: *Lauren E. Price, BSN, The University of Arizona College of Nursing, 1305 N. Martin, PO Box 210203, Tucson, AZ 85721 (LPRICE1@EMAIL.ARIZONA.EDU).*

DOI: 10.1097/NAQ.0000000000000118

Information Systems Technology Architecture (VistA).[5]

The VHA is one of the largest integrated health care systems in the United States with data from more than 20 years of sustained EHR use.[8] When large amounts of clinical data accumulate over time, there are issues in data quality, storage, retrieval, and computation for use.[9] To combat these types of issues, the VHA created a complex network of both regional and national databases, which are controlled by the Department of Veterans Affairs, Office of Information and Technology.[10] The VHA then began to integrate the content of these databases into a more usable and accessible form, known now as the Corporate Data Warehouse (CDW) (Figure 1).

The greatest challenges faced by researchers and administrators seeking to generate new knowledge from EHR data originate from a lack of accessibility and standardization in data collection and storage.[11] The goal of the VHA's CDW is to "incorporate data from multiple data sets throughout the VHA into one standard database structure to facilitate reporting and data analysis at the enterprise level."[12(p3)] The goal of the VA Informatics and Computing Infrastructure (VINCI) is to "provide data and tools to support management decision making, performance measurement, and research objectives."[10,12(p3)]

Given the newness of Big Data to the nursing profession, it is important to share information about the CDW and other clinical data warehouses. If the data are not prepared and analyzed correctly, it is easy to misinterpret the significance of the results. Clinical data represent many observations across time that have been filtered through the lens of the health care system. Only as Big Data science has evolved have researchers begun to emphasize the importance of validating results against clinically significant outcomes to ensure that they are meaningful.[13,14] Clinical significance and the timeliness of analysis are required to produce evidence that informs changes in the health care system, improves patient outcomes, and supports practitioner decision making.

The CDW is continuously updated to incorporate data from additional data sources within the VHA and the Department of Defense. Currently, the CDW houses data from the following databases: VistA, VHA Regional

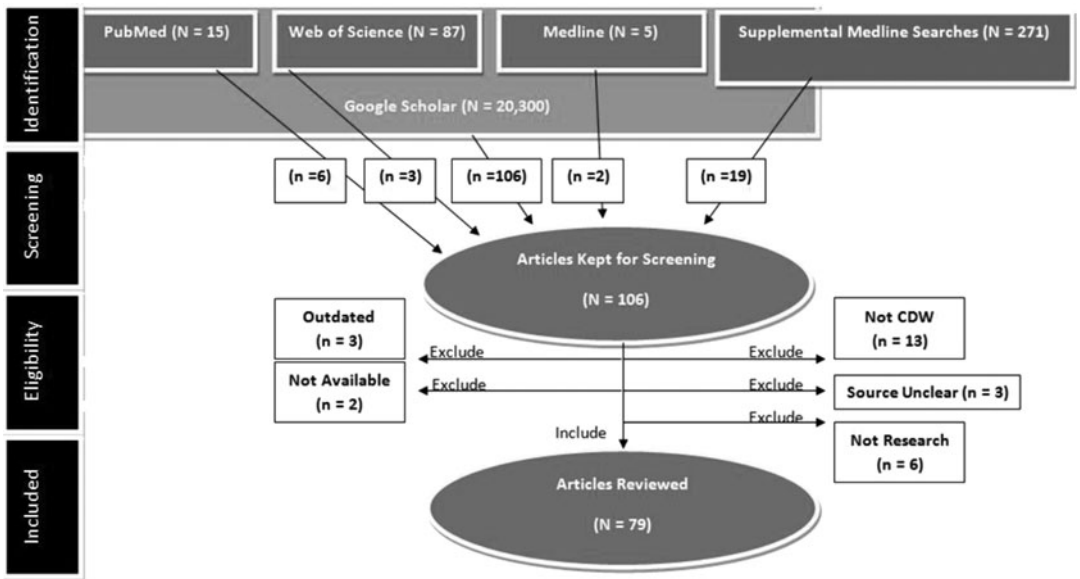

Figure 1. Inclusion/exclusion criteria. CDW indicates Corporate Data Warehouse.

Data Warehouses, VHA Decision Support System, VHA National Patient Care Database, VA Compensation and Pension Exams, and MyHealtheVet.va.gov. Because of varying CDW data domains before 2010, the authors chose to explore recent CDW uses in literature published between 2010 and 2014.[15]

Our investigation was designed to answer the following questions:

1. What are the scope and number of published research projects that have been conducted since 2010 using the VHA's CDW?
2. What are the implications of and special considerations for using CDW and Big Data sources?
3. Are nurses embracing and utilizing Big Data to improve practice and the science of the nursing profession?

METHODS

A systematic search of the literature was conducted between 2010 and 2014 in MEDLINE, Web of Science, Google Scholar, and PubMed using the key terms "Corporate Data Warehouse," "Corporate Data Warehouse and Veteran," and "VHA Corporate Data Warehouse." Figure 2 depicts the breakdown of search results and articles included for review. Articles included in the review needed to (1) clearly articulate the data source, (2) use the CDW as a data source, (3) be original research, and (4) be published between 2010 and 2014 in a scholarly journal. Studies that used only the CDW to identify a cohort were excluded.

Two additional searches in MEDLINE were conducted to ensure that eligible articles were not being overlooked on the basis of inadequate key term selection. The Google Scholar searches captured all articles identified in the other databases. Unlike the other databases, Google Scholar searches are based on string frequency (throughout the entire article), and relevance, resulting in a greater yield than the others. All articles were prescreened on the basis of their titles and abstracts before undergoing the formal screening process for inclusion. (Please contact L.E.P. for MEDLINE search paths for additional supplementation; see the Table.)

RESULTS

Of the 106 articles reviewed, 27 were excluded for not meeting 1 or more of the inclusion criteria. More than half of the remaining studies (n = 44) combined CDW data with data from another source. Additional sources included data sets from outside organizations (such as Medicare), data from other VHA databases, and prospective data. Sample sizes across studies ranged from 22 to 11 041 855.

Forty-five topics were identified on the basis of key words that could be used to succinctly describe each article's purpose. The topic labels were determined on the basis of a consensus between the authors of this study. Some articles focused on more than 1 topic. The most frequent topics included systems issues and health services utilization (n = 21), pharmacy and medications (n = 10), weight/weight loss/obesity (n = 9), and mental health (n = 9). The remaining topics were varied and included lipid management (n = 1), chronic obstructive pulmonary disease (n = 1), Barrett's esophagus (n = 1), endoscopy (n = 1), dry eye (n = 1), infusion pump testing (n = 1), phenotyping (n = 1), transplantation (n = 1), bone mineral density (n = 1), musculoskeletal disorders (n = 1), chiropractic services (n = 1), telehealth (n = 1), smoking (n = 1), nursing issues (n = 1), adherence (n = 1), myocardial infarction (n = 1), *Clostridium difficile* infection (n = 1), spinal cord injury (n = 1), influenza (n = 1), heart failure (n = 2), gender issues (n = 2), ulcerative colitis (n = 2), pain (n = 2), sexual dysfunction (n = 2), homelessness (n = 2), rural health (n = 2), human immunodeficiency virus (HIV) infection (n = 2), travel (n = 3), blood pressure management (n = 3), brain injury (n = 3), hepatitis C (n = 3), surgical outcomes (n = 4), disparities (n = 5), race (n = 5), diabetes (n = 6), kidney disease (n = 7), cancer (n = 7), and mortality rates (n = 8).

The majority of the 79 studies gave descriptive statistics regarding the sample and

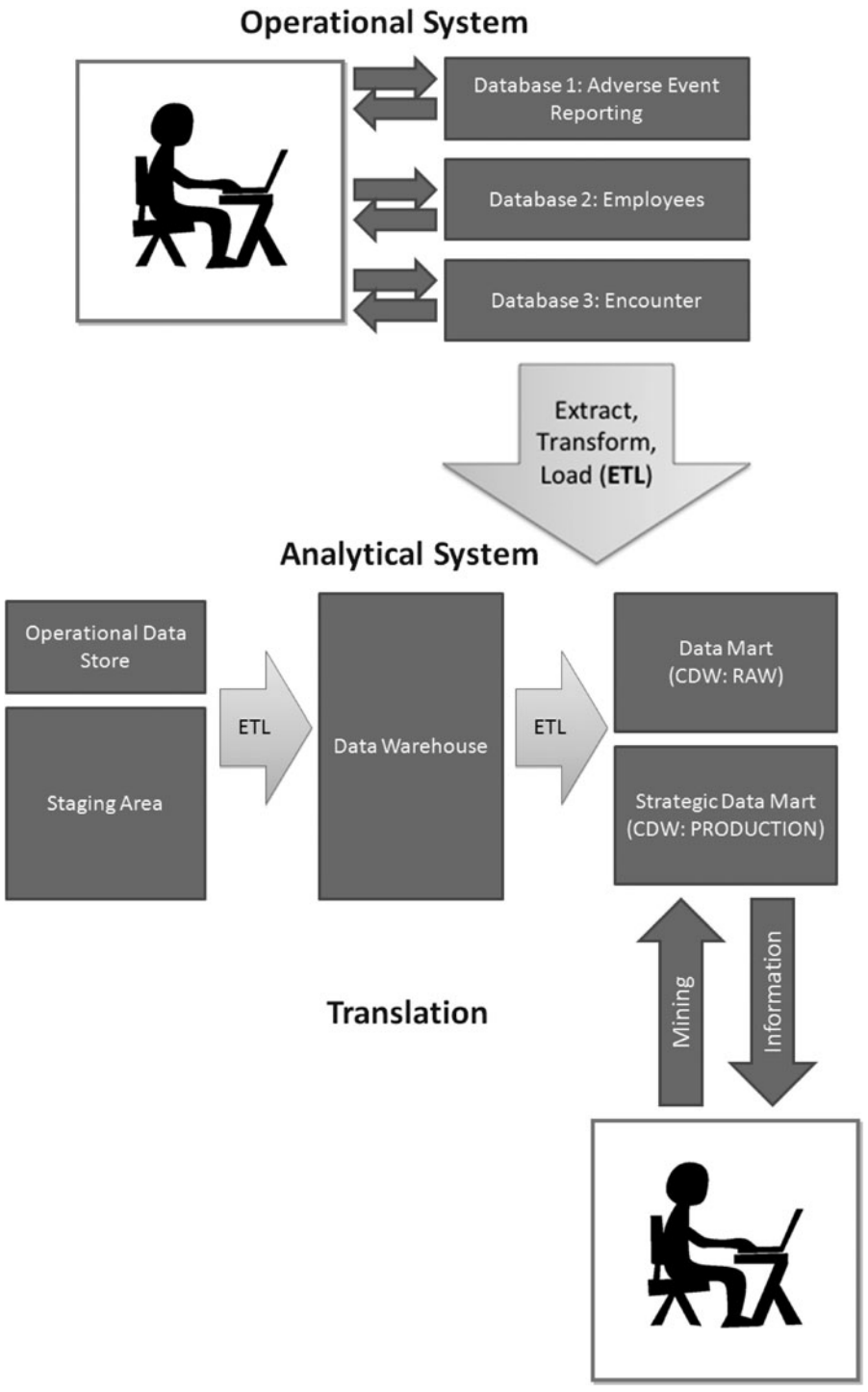

Figure 2. Graphical depiction of a data management system and the movement of data. CDW indicates Corporate Data Warehouse.

Table. MEDLINE Search Paths for Additional Supplementation

Search 1	Search 2
1. veterans.in. (62 231)	1. department of veterans affairs.in. (9 566)
2. "United States Department of Veterans Affairs"/ (4 757)	2. "United States Department of Veterans Affairs"/ (4 757)
3. 1 or 2 (65 836)	3. 1 or 2 (13 849)
4. retrospective studies.mp. or Retrospective Studies/ (484 251)	4. CDW.mp. (235)
5. 3 and 4 (3 549)	5. clinical data warehouse.mp. (64)
6. national.ab. (224 986)	6. (clinical adj data adj warehouse).mp. (64)
7. 5 and 6 (251)	7. corporate data warehouse.mp. (5)
	8. 4 or 5 or 6 or 7 (295)
	9. 3 and 8 (2)
	10. informatics consortium.mp. (5)
	11. 3 and 10 (0)
	12. exp Medical Records Systems, Computerized/ (24 657)
	13. 3 and 12 (284)
	14. exp Medical Records Systems, Computerized/sn [Statistics & Numerical Data] (920)
	15. 3 and 14 (19)
	16. Retrospective studies/ or cohort studies/ or retrospective cohort study.mp. (609 117)
	17. 3 and 16 (987)
	18. Retrospective studies/ (480 183)
	19. 3 (13 849)
	20. 18 and 19 (710)
	21. 14 and 20 (6)
	22. Electronic health records/ (5 706)
	23. 3 and 22 (92)
	24. 9 or 15 (20)

conducted some form of regression analysis. Data mining, natural language processing, or qualitative approaches were less frequently employed. The most commonly used analytical approach was a logistic regression (38%), Cox proportional hazard modeling was used the second most frequently (30%), and profiling or natural language processing techniques were infrequent (11%). Studies using data mining, natural language processing, or qualitative approaches tended to be the only studies that elaborated on data quality, missingness, and/or use of cross-validation procedures. These elements of analysis are essential to ensure that clinically meaningful results are obtained when working with the large samples noted in these studies.[16]

DISCUSSION

Four databases were systematically searched and 79 original research studies were identified as utilizing data from the VHA's CDW. The 2014 estimate of ongoing uses of this data source, provided by the VA Information Research Center, has grown substantially to include 1500 research and 200 quality management projects in progress. The published studies we reviewed covered a wide range of topics but most frequently

focused on addressing systems issues (including assessment of health care quality), pharmacy/medications, and weight/weight loss/obesity. Quality management is one of the primary reasons the CDW was initially developed.[12] Therefore, the frequent use of the CDW to evaluate the contribution of systems issues to patient outcomes was an expected finding.

The pharmacy and medications studies were also anticipated. In no other area of clinical research has the implementation of EHR data storage systems been of greater benefit than in the area of pharmacovigilance.[17] The VHA has been praised on multiple occasions for the impact its information technology systems have had on improving pharmacovigilance.[8] Adverse drug events are often accompanied by serious cost and health implications. The trend of using the CDW in studies of pharmacovigilance is anticipated to continue in the years to come.[17]

Weight loss trends and the management of obesity, the third most frequently researched topics, make intuitive sense as overweight and obese veterans compromise the majority of the population cared for by the VHA. As a result, 10 years of research has been focused on the impact of weight on comorbid diseases such as diabetes, cardiovascular disease, and mental health disorders.[18] The VHA is invested in developing and evaluating program outcomes targeted at weight loss.[19] The CDW is currently one of the most accurate and accessible resources for obtaining longitudinal height and weight data.[18] Consequently, the CDW will likely continue this research focus.

Only one of the 79 studies was conducted by nurses: research to address the frequency of blood pressure monitoring in the emergency department.[20] The lack of nurses utilizing the CDW is of concern to us. It was recently presented at the National Advisory Council on Nursing Research that nurses have worked with large samples of retrospective data for many years.[21] Yet, just as this examination of the CDW confirms, published nursing studies are underrepresented in the context of using this data source. It is imperative that nurses publish their findings for both quality improvement and research endeavors. Nurses must learn about and utilize Big Data sources so that we may continue to grow and recognize the significance of Big Data analytics within our discipline.

LIMITATIONS

A limitation of this study is that we may have failed to capture many of the studies that made use of CDW data. Not specifying a data source is a serious problem among retrospective records review studies, as manifested by the 3 cases that were excluded from this study. Not only is it difficult to get an accurate estimate of CDW utilization, the scientific rigor and ethics of these studies must be questioned when data sources, manipulation techniques, and prepreprocessing procedures are not specified.

Because of the increasing overlap between data mining and traditional statistical approaches, distinguishing between the 2 can be difficult.[22] For this reason, we chose not to make this distinction when conducting this review but do want to point out the significance of methodology when working with Big Data.

According to statistical learning theory, traditional statistics emerged from deductive logic and the primary assumptions of the parametric model. However, observations in nature do not always follow the rules of the parametric model and variations emerge.[16] The multiple statistical approaches found in this review provide evidence of the variations that can result depending on the size, dimensions, and natural distribution of the data. Data mining has emerged from inductive, Bayesian and Platonic logic and the need to improve computational efficiency when working with large, organic data sources.[16] For this reason, it is the ideal analytical mate when deriving new knowledge from large clinical data sources and data warehouses.[23] Data mining approaches also yield results that

are often directly translatable to real-world problem solving and decision making and are not limited to research endeavors. The appropriateness of data mining for research purposes versus quality improvement depends on the quality of data and the scope of the question.[24]

The American Medical Informatics Association advocates for the use of data mining when working with large clinical data sources.[13] An exploratory type of data mining, known as electronic phenotyping, has specifically been developed to ensure that clinical data are being adequately cleaned and transformed in a standardized way before conducting research. The intent is to minimize error and ensure that findings retain clinical significance and relevance.[13] One important (but often forgotten) feature of clinical data is that they reflect many observations that have been filtered through the lens of the health care system and its data collection methodologies.[13] Therefore, new knowledge generated from large clinical data sources should *only* be used to augment nursing findings through direct assessment of the patient, not replace it.

CONCLUSION

The VHA's CDW is a well-developed resource for enhancing scientific discovery. This data source is a valuable tool for both knowledge discovery and improving clinical practice, particularly when data mining techniques are applied. On the basis of the current state of science, the average length of time it takes original research to affect clinical practice is 17 years.[25] The application of advanced analytical techniques to CDW data has the potential to halt this trend and expedite the translation of evidence into practice. Nurses should learn about and utilize resources for Big Data to improve the science of our profession to build the foundation for clinical practice. Future studies are needed that use the CDW to produce directly translatable results, both in nursing and across the health care continuum.

REFERENCES

1. Fan W, Bifet A. Mining Big Data: current status, and forecast to the future. *ACM SIGKDD Explorations Newslett*. 2013;14(2):1-5.
2. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big Data: the management revolution. *Harv Bus Rev*. 2012;90(10):61-67.
3. Manyika J, Chui M, Brown B, et al. *Big Data: The Next Frontier for Innovation, Competition and Productivity* [Technical report]. McKinsey Global Institute; 2012. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed July 21, 2015.
4. Regents of the University of California. How much information? 2003 [Executive summary]. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_execsum.pdf. Accessed July 21, 2015.
5. Fihn SD, Francis J, Clancy C, et al. Insights from advanced analytics at the Veterans Health Administration. *Health Aff*. 2014;33(7):1203-1211.
6. Clancy TR, Bowles KH, Gelinas L, et al. A call to action: engage in Big Data science. *Nurs Outlook*. 2014;62:64-65.
7. Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA—US Department of Veterans Affairs national-scale HIS. *Int J Med Inform*. 2003;69(2):135-156.
8. Byrne CM, Mercincavage LM, Pan EC, Vincent AG, Johnston DS, Middleton B. The value from investments in health information technology at the US Department of Veterans Affairs. *Health Aff*. 2010;29(4):629-638.
9. Roski J, Bo-Linn G, Andrews T. Creating value in healthcare through Big Data: opportunities and policy implications. *Health Aff*. 2014;33(7):1115-1122.
10. US Department of Veterans Affairs. *Corporate Data Warehouse. Health Services Research & Development*. Washington, DC: US Department of Veterans Affairs; 2014. http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm. Accessed July 21, 2015.
11. Almasalha F, Xu D, Keenan GM, et al. Data mining nursing care plans of end-of-life patients: a study to improve healthcare decision making. *Int J Nurs Knowledge*. 2013;24(1):15-24.
12. VA Information Resource Center. *VIReC Resource Guide: VA Corporate Data Warehouse*. Hines, IL: US Department of Veterans Affairs, Health Services Research & Development Service, VA Information Resource Center; 2012.
13. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2012;20(1):117-121.

14. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc*. 2013;20(e2):e226-e231.
15. Noël P, Copeland L. Using VA Corporate Data Warehouse for health services research [PowerPoint presentation]. http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/vdm-060412.pdf. Published 2012. Accessed July 21, 2015.
16. Nisbet R, Elder IV, Miner G. *Handbook of Statistical Analysis and Data Mining Applications* [Books 24/7 Electronic Version]. Academic Press; 2009. http://common.books24x7.com/toc.aspx?bookid=37310. Accessed July 21, 2015.
17. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. *Br J Clin Pharmacol*. 2004;57(2):127-134.
18. Noël PH, Copeland LA, Perrin RA, et al. VHA Corporate Data Warehouse height and weight data: opportunities and challenges for health services research. *J Rehabil Res Dev*. 2010;47(8):739-750.
19. Kahwati LC, Lance TX, Jones KR, Kinsinger LS. RE-AIM evaluation of the Veterans Health Administration's MOVE! weight management program. *Transl Behav Med*. 2011;1(4):551-560.
20. Miltner RS, Johnson KD, Deierhoi R. Exploring the frequency of blood pressure documentation in emergency departments. *J Nurs Scholarsh*. 2014;46(2):98-105.
21. Dr. Patti Brennan: Nursing, Big Data, & the NIH BD2K Initiative. *YouTube*. https://www.youtube.com/watch?v=S00DyTsdFm4. Updated April, 15, 2014. Accessed February 28, 2015.
22. Smyth P. Data mining: data analysis on a grand scale? *Stat Methods Med Res*. 2000;9(4):309-327.
23. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. *Proc AMIA Annu Fall Symp*. 1997:101-105.
24. Testik MC, Runger GC, Kirkman-Liff B, Smith EA. Data mining and knowledge discovery in healthcare organizations: a decision-tree approach. In:Wang J, ed. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. Vol V [Books 24/7 Electronic Version]. Information Science Reference; 2008. http://common.books24x7.com/toc.aspx?bookid=45370. Accessed July 21, 2015.
25. Balas EA, Boren SA. Managing clinical knowledge for health care improvement. *Yearbook Med Inform*. 2000;200065-200070.