Lecture Notes in Bioengineering

Mowafa Househ
Andre W. Kushniruk
Elizabeth M. Borycki *Editors*

Big Data, Big Challenges: A Healthcare Perspective

Background, Issues, Solutions and Research Directions

Springer

Lecture Notes in Bioengineering

More information about this series at http://www.springer.com/series/11564

Mowafa Househ · Andre W. Kushniruk ·
Elizabeth M. Borycki
Editors

Big Data, Big Challenges: A Healthcare Perspective

Background, Issues, Solutions and Research Directions

Springer

Editors
Mowafa Househ
Division of Information and Computing Technology, College of Science and Engineering
Hamad Bin Khalifa University, Qatar Foundation
Doha, Qatar

Elizabeth M. Borycki
School of Health Information Sciences
University of Victoria
Victoria, BC, Canada

Andre W. Kushniruk
School of Health Information Sciences
University of Victoria
Victoria, BC, Canada

ISSN 2195-271X ISSN 2195-2728 (electronic)
Lecture Notes in Bioengineering
ISBN 978-3-030-06108-1 ISBN 978-3-030-06109-8 (eBook)
https://doi.org/10.1007/978-3-030-06109-8

Library of Congress Control Number: 2018964922

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Much has been written about the utilization of big data analytics methods, tools and technologies to collect, process, visualize and make use of high volume structured and unstructured data in a number of fields such as finance, insurance, sports, agriculture, and health. With the fast and ever-increasing growth of user-generated data from the Internet, such as social media content, data from wireless medical devices and mobile apps, big data analytical methods, tools and technologies have become recognized as the only plausible "go-to" solutions that are able to make sense of such voluminous, disorganized, fluid and free-flowing data. Within health care, there is a growing knowledge base of big data related studies and implementations in public health, clinical decision making, disease prevention, and healthcare cost reduction. As with any new field, much of the research and discussions center upon the added value and opportunities that new technologies, such as big data analytics methods, tools and technologies can provide. However, as the domain area begins to mature through increased implementation, evaluation studies, and user experiences, the problems, and challenges relating to the methods, tools, and technologies used for big data analytics begin to emerge. For the past five years, much of the literature on big data analytics has focused on the benefits of big data in improving all areas of health care. A new wave of research is beginning to emerge challenging some of the assumptions made by the positive assertions for big data analytics in health care. That is the motivation behind this book, which is not only about sharing success stories or opportunities for big data in health care, but also to address the arising challenges that many researchers have overlooked.

What makes this book unique is that it examines both the opportunities and focuses more on the challenges in applying big data analytics methods, tools and technologies within health care from a number of perspectives. The book is divided into three parts and eleven chapters. The first part of the book examines the healthcare professional perspective on the challenges and opportunities of big data analytics from a nursing, medical, public health, and health administrator perspective. Most of the chapters are included in the first part of the book. The second part of the book focuses on human factors and ethical challenges and opportunities related to big data analytics in health care. There are three chapters in part two

of the book that address topics related to patient safety, user-centered design, and ethical issues. Part three of the book includes two chapters that examine the technical challenges in the utilization of big data analytics in health care. The first chapter examines the challenges and opportunities of big data analytics from a data scientist's perspective. The second chapter examines the integrative exposum/expotype perspective related to big data analytics in health care.

The book provides health data scientists, health care professionals, and health-care managers and policymakers the first comprehensive insight into the challenges and opportunities of big data analytics in health care. The book will challenge some of the pre-held conceptions and notions students and professionals of big data analytics in health care currently possess and challenge them to derive new solutions and ideas to the proposed challenges suggested within the book.

Doha, Qatar — Mowafa Househ
Victoria, Canada — Andre W. Kushniruk
Victoria, Canada — Elizabeth M. Borycki

Contents

Part I
Health Professional Perspective

Big Data Challenges from a Nursing Perspective

Suzanne Bakken and Theresa A. Koleck

1 Introduction

The International Council of Nurses provides a global definition of nursing as "Nursing encompasses autonomous and collaborative care of individuals of all ages, families, groups and communities, sick or well and in all settings. Nursing includes the promotion of health, prevention of illness, and the care of ill, disabled and dying people. Advocacy, promotion of a safe environment, research, participation in shaping health policy and in patient and health systems management, and education are also key nursing roles" [1]. In contrast to physicians who focus on cure, nurses focus on individual, family, and group "responses to actual or potential health problems" [2]. Importantly, nurses consider the individual within the context of their family, sociocultural, and physical environments. Nursing's holistic perspective as well as the focus on responses to actual or potential health problems has major implications for the benefits, promise, and challenges of big data streams and data science methods for nursing.

Multiple authors have highlighted the relevance of data science to nursing [3–5]. Bakken and Brennan further argue that nursing policy statements inform a principled and ethical approach to big data and data science [3]. Nurses' use of data science methods is on the rise. A recent systematic review of application of data science in nursing evaluated 17 studies conducted in 2009–2015 [5]. The focus was on nursing practice and systems that affect nurses. Although most studies were in acute care settings, community, home health, and public health settings were also represented reflecting the variety of settings in which nursing occurs. In terms of

S. Bakken (✉)
School of Nursing, Department of Biomedical Informatics, and Data Science Institute, Columbia University, 630 W. 168th Street, New York, NY 10032, USA
e-mail: sbh22@cumc.columbia.edu

T. A. Koleck
School of Nursing, Columbia University, New York, NY, USA

M. Househ et al. (eds.), *Big Data, Big Challenges: A Healthcare Perspective*, Lecture Notes in Bioengineering, https://doi.org/10.1007/978-3-030-06109-8_1

characterizing the data used according to the criteria for big data [6, 7], all studies met the criterion of volume, most met the criterion of variety, and a minority met the criterion of velocity. Veracity and value were not explicitly analyzed. Electronic health records (EHRs) were the primary data source for 14 studies although several studies integrated EHR data with other data sources. The study purposes were categorized as knowledge discovery, prediction, and evaluation. Since the time of this review, additional nursing studies have been conducted that reflect data sources beyond EHRs and structured data sources including omics [8], social media [9], and sensors [10]. Moreover, health policy considerations for data science have been delineated from a nursing science perspective [11].

The purpose of this chapter is to summarize the benefits and key challenges related to big data streams and data science from the perspective of nursing. The benefits and challenges are considered from the perspective of data governance as well as data science infrastructure and pipeline and illustrated through six case examples. In addition, two cross-cutting issues (ethical conduct of research and data science competencies) are addressed.

2 Data Governance and Data Science Infrastructure and Pipeline

A number of authors have published data science pipelines. However, from the perspective of nursing, data science starts with a question and because of the nature of the data, which is often protected health information from a U.S. Health Insurance Portability and Accountability Act (HIPAA) standpoint, requires careful consideration of data governance (Fig. 1). In addition, the infrastructure required for data science is often significantly different from the data management and analytic pipelines typically available to nurse scientists and clinicians due to the volume of data and the processing power needed to ingest, wrangle (i.e., pre-process using semi-automated tools), compute and analyze, model and validate, and interpret (visualize and report) the data. Moreover, data science requires platforms beyond SAS, STATA, and R such as Apache Hadoop Map-Reduce, Apache Mahout (machine learning algorithms), Sparks Machine Learning Library, and R-Hadoop to support reduction and analysis of multi-dimensional data through methods such as K-means, random forest classifier, neural network backpropagation, support vector machines, and Gaussian discriminative analysis. A data science infrastructure must also support visualization of the data for analysis, interpretation, and reporting through general tools such as Tableau and tools for special purposes (e.g., Sentiment Viz for visualization of Tweet contents, ORA for visualization of network structures).

Table 1 displays a summary of challenges related to aspects of data governance, data science infrastructure, and data science pipeline in a set of case examples that are described in more detail in the following section.

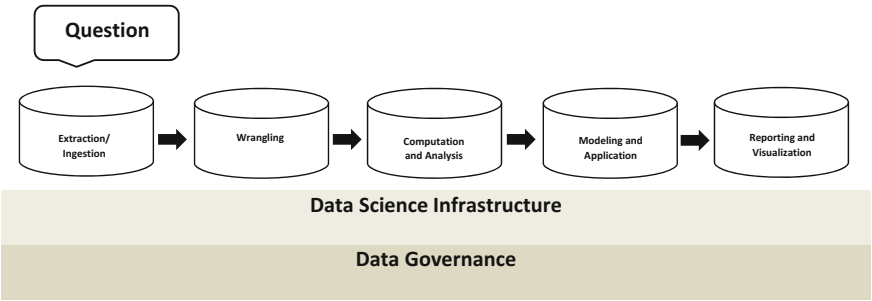

Fig. 1 Data governance, data science infrastructure, and data science pipeline. Adapted from Tesla Institute [12]

3 Case Examples

Three case examples from the authors' experience, which are focused on knowledge discovery from electronic health records (EHRs), omics, and social media and reflect multiple challenges, are described first. This is followed by briefer descriptions of three more case examples from the literature that highlight a specific challenge.

3.1 Electronic Health Records and Symptom Science

Symptoms (e.g., pain, fatigue, sleep disturbance, anxiety, depression, nausea) related to a disease process and/or clinical intervention are complex—they are subjective, vary over time, and lack clear biological mechanisms [20, 21]. Despite challenges, EHRs and clinical data repositories are two related big data resources that can be used to facilitate symptom research [22, 23]. Koleck and colleagues [13] investigated demographic and clinical predictors of one of the most common and distressing postoperative symptoms, nausea, and its frequently accompanying sign, vomiting, in women undergoing gynecologic surgical procedures. The team took advantage of EHRs, which capture real-life symptom data over time, available within their institution's clinical data repository containing records for over 5.5 million patients. The first challenge addressed by the team was related to governance, specifically obtaining access to the clinical data necessary to answer their research questions. The institution at which this study was completed has formal data stewards who maintain datasets for a variety of clinical applications. The process for obtaining data involves submitting an electronic form to a central committee that reviews, approves, prioritizes, and fulfills requests. Time to completion varies depending on a number of factors including the complexity and priority of the request as well as the request queue. Procuring proper infrastructure for storing data was a second challenge but is of upmost importance to both ensure

Table 1 Summary of key data science challenges for case studies

Case example	Governance	Infrastructure	Extraction/ ingestion	Wrangling	Computation and analysis	Modeling and application	Reporting and visualization
EHRs and symptom science [13]	x	x	x	x	x		
Omics [14]			x	x			x
Twitter and dementia caregiving [15, 16]		x	x				
Prediction and sepsis campaign guideline [17]				x			
Intelligent sensors and aging in place [18]		x					
Dashboards and nurse numeracy and graph literacy [19]							x

that patient confidentiality is maintained and for ease of data manipulation. Requested data was stored in MySQL relational database tables on a secure HIPAA-compliant server. Relational database tables also allowed the team to alleviate the key challenge of data integration (a component of wrangling) by enabling direct connection of data from multiple clinical applications via a primary key (e.g., the patient's master identification number). The next challenge was related to extracting the relevant symptom information for the population of interest. Women undergoing gynecological procedures were identified using structured ICD-9 and ICD-10 procedure codes related to operations on the ovary, fallopian tube, or uterus. In contrast, while billing codes for symptoms do exists (e.g., ICD-10 R11 nausea and vomiting), symptom data were not well represented by these codes. The team overcame this challenge by using postoperative medication administration records. A comprehensive list of antiemetic medications was compiled from the literature. Administration of an antiemetic medication was treated as a surrogate for postoperative nausea and/or vomiting. The team was able to limit instances of nausea/vomiting to the postoperative period (within the first 48 h after surgery) by subtracting the antiemetic medication administration time from the surgery anesthesia finish time. Unstructured clinical narratives can be used to overcome this challenge for symptoms without structured surrogates available. Common strategies to extract information from clinical narratives include text mining and natural language processing, but these strategies introduce additional data pre-processing challenges [24, 25]. Finally, one of the most significant challenges for analysis of EHR data is assessment of data quality. In order to mitigate data quality concerns, the team addressed the five dimensions of data quality for EHR data reuse research—completeness, correctness, concordance, plausibility, and currency [26].

3.2 Omics

Large scale omic (e.g., genomic, epigenomic, transcriptomic, proteomic, metabolomic, microbiomic) studies aim to enhance our understanding of the molecular basis of disease, disease risk, and patient outcomes [27, 28]. Arockiaraj and colleagues [14] conducted an epigenome-wide association study to explore how changes in DNA methylomic profiles following acute subarachnoid hemorrhage impact and potentially explain observed variability in prognosis and recovery from hemorrhage. Epigenetic changes are DNA modifications that affect gene expression without changing the DNA sequence [29, 30]. These changes, including methylation (i.e., the process of attaching methyl groups to DNA), can change overtime and differ substantially between tissue and cell types. Consequently, data collection considerations, related to selection of tissue type, timing of serial sampling, plate design, and phenotype assessment, were of critical importance when designing this study [31]. Two to five cerebral spinal fluid (CSF) samples were collected from patient ventricular drains (placed as part of standard care) over the first 14 days