MIT CRITICAL DATA

Secondary Analysis of Electronic Health Records

EXTRAS ONLINE

Springer Open

Secondary Analysis of Electronic Health Records

MIT Critical Data

Secondary Analysis of Electronic Health Records

Springer Open

MIT Critical Data
Massachusetts Institute of Technology
Cambridge, MA
USA

Additional material to this book can be downloaded from http://link.springer.com/978-3-319-43740-8.

ISBN 978-3-319-43740-8 ISBN 978-3-319-43742-2 (eBook)
DOI 10.1007/978-3-319-43742-2

Library of Congress Control Number: 2016947212

Printed on acid-free paper

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

Preface

Diagnostic and therapeutic technologies continue to evolve rapidly, and both individual practitioners and clinical teams face increasingly complex decisions. Unfortunately, the current state of medical knowledge does not provide the guidance to make the majority of clinical decisions on the basis of evidence. According to the 2012 Institute of Medicine Committee Report, only 10–20 % of clinical decisions are evidence based. The problem even extends to the creation of clinical practice guidelines (CPGs). Nearly 50 % of recommendations made in specialty society guidelines rely on expert opinion rather than experimental data. Furthermore, the creation process of CPGs is "marred by weak methods and financial conflicts of interest," rendering current CPGs potentially less trustworthy.

The present research infrastructure is inefficient and frequently produces unreliable results that cannot be replicated. Even randomized controlled trials (RCTs), the traditional gold standards of the research reliability hierarchy, are not without limitations. They can be costly, labor-intensive, slow, and can return results that are seldom generalizable to every patient population. It is impossible for a tightly controlled RCT to capture the full, interactive, and contextual details of the clinical issues that arise in real clinics and inpatient units. Furthermore, many pertinent but unresolved clinical and medical systems issues do not seem to have attracted the interest of the research enterprise, which has come to focus instead on cellular and molecular investigations and single-agent (e.g., a drug or device) effects. For clinicians, the end result is a "data desert" when it comes to making decisions.

Electronic health record (EHR) data are frequently digitally archived and can subsequently be extracted and analyzed. Between 2011 and 2019, the prevalence of EHRs is expected to grow from 34 to 90 % among office-based practices, and the majority of hospitals have replaced or are in the process of replacing paper systems with comprehensive, enterprise EHRs. The power of scale intrinsic to this digital transformation opens the door to a massive amount of currently untapped information. The data, if properly analyzed and meaningfully interpreted, could vastly improve our conception and development of best practices. The possibilities for quality improvement, increased safety, process optimization, and personalization of clinical decisions range from impressive to revolutionary. The National Institutes of

Health (NIH) and other major grant organizations have begun to recognize the power of big data in knowledge creation and are offering grants to support investigators in this area.

This book, written with support from the National Institute for Biomedical Imaging and Bioengineering through grant R01 EB017205-01A1, is meant to serve as an illustrative guide for scientists, engineers, and clinicians that are interested in performing retrospective research using data from EHRs. It is divided into three major parts.

The first part of the book paints the current landscape and describes the body of knowledge that dictates clinical practice guidelines, including the limitations and the challenges. This sets the stage for presenting the motivation behind the secondary analysis of EHR data. The part also describes the data landscape, who the key players are, and which types of databases are useful for which kinds of questions. Finally, the part outlines the political, regulatory and technical challenges faced by clinical informaticians, and provides suggestions on how to navigate through these challenges.

In the second part, the process of parsing a clinical question into a study design and methodology is broken down into five steps. The first step explains how to formulate the right research question, and bring together the appropriate team. The second step outlines strategies for identifying, extracting, Oxford, and preprocessing EHR data to comprehend and address the research question of interest. The third step presents techniques in exploratory analysis and data visualization. In the fourth step, a detailed guide on how to choose the type of analysis that best answers the research question is provided. Finally, the fifth and final step illustrates how to validate results, using cross validation, sensitivity analyses, testing of falsification hypotheses, and other common techniques in the field.

The third, and final part of the book, provides a comprehensive collection of case studies. These case studies highlight various aspects of the research pipeline presented in the second part of the book, and help ground the reader in real world data analyses.

We have written the book so that a reader at different levels may easily start at different parts. For the novice researcher, the book should be read from start to finish. For individuals who are already acquainted with the challenges of clinical informatics, but would like guidance on how to most effectively perform the analysis, the book should be read from the second part onward. Finally, the part on case studies provides project-specific practical considerations on study design and methodology and is recommended for all readers.

The time has come to leverage the data we generate during routine patient care to formulate a more complete lexicon of evidence-based recommendations and support shared decision making with patients. This book will train the next generation of scientists, representing different disciplines, but collaborating to expand the knowledge base that will guide medical practice in the future.

We would like to take this opportunity to thank Professor Roger Mark, whose vision to create a high resolution clinical database that is open to investigators around the world, inspired us to write this textbook.

Cambridge, USA MIT Critical Data

MIT Critical Data

MIT Critical Data consists of data scientists and clinicians from around the globe brought together by a vision to engender a data-driven healthcare system supported by *clinical informatics without walls*. In this ecosystem, the creation of evidence and clinical decision support tools is initiated, updated, honed, Oxford, and enhanced by scaling the access to and meaningful use of clinical data.

Leo Anthony Celi has practiced medicine in three continents, giving him broad perspectives in healthcare delivery. His research is on secondary analysis of electronic health records and global health informatics. He founded and co-directs Sana at the Institute for Medical Engineering and Science at the Massachusetts Institute of Technology. He also holds a faculty position at Harvard Medical School as an intensivist at the Beth Israel Deaconess Medical Center and is the clinical research director for the Laboratory of Computational Physiology at MIT. Finally, he is one of the course directors for HST.936 at MIT—innovations in global health informatics and HST.953—secondary analysis of electronic health records.

Peter Charlton gained the degree of M.Eng. in Engineering Science in 2010 from the University of Oxford. Since then he held a research position, working jointly with Guy's and St Thomas' NHS Foundation Trust, and King's College London. Peter's research focuses on physiological monitoring of hospital patients, divided into three areas. The first area concerns the development of signal processing techniques to estimate clinical parameters from physiological signals. He has focused on unobtrusive estimation of respiratory rate for use in ambulatory settings, invasive estimation of cardiac output for use in critical care, and novel techniques for analysis of the pulse oximetry (photoplethysmogram) signal. Secondly, he is investigating the effectiveness of technologies for the acquisition of continuous and intermittent physiological measurements in ambulatory and intensive care settings. Thirdly, he is developing techniques to transform continuous monitoring data into measurements that are appropriate for real-time alerting of patient deteriorations.

Mohammad Mahdi Ghassemi is a doctoral candidate at the Massachusetts Institute of Technology. As an undergraduate, he studied Electrical Engineering and graduated as both a Goldwater scholar and the University's "Outstanding

Engineer". In 2011, Mohammad received an MPhil in Information Engineering from the University of Cambridge where he was also a recipient of the Gates-Cambridge Scholarship. Since arriving at MIT, he has pursued research at the interface of machine learning and medical informatics. Mohammad's doctoral focus is on signal processing and machine learning techniques in the context of multi-modal, multiscale datasets. He has helped put together the largest collection of post-anoxic coma EEGs in the world. In addition to his thesis work, Mohammad has worked with the Samsung Corporation, and several entities across campus building "smart devices" including: a multi-sensor wearable that passively monitors the physiological, audio and video activity of a user to estimate a latent emotional state.

Alistair Johnson received his B.Eng. in Biomedical and Electrical Engineering at McMaster University, Canada, and subsequently read for a DPhil in Healthcare Innovation at the University of Oxford. His thesis was titled "Mortality and acuity assessment in critical care", and its focus included using machine learning techniques to predict mortality and develop new severity of illness scores for patients admitted to intensive care units. Alistair also spent a year as a research assistant at the John Radcliffe hospital in Oxford, where he worked on building early alerting models for patients post-ICU discharge. Alistair's research interests revolve around the use of data collected during routine clinical practice to improve patient care.

Matthieu Komorowski holds board certification in anesthesiology and critical care in both France and the UK. A former medical research fellow at the European Space Agency, he completed a Master of Research in Biomedical Engineering at Imperial College London focusing on machine learning. Dr Komorowski now pursues a Ph.D. at Imperial College and a research fellowship in intensive care at Charing Cross Hospital in London. In his research, he combines his expertise in machine learning and critical care to generate new clinical evidence and build the next generation of clinical tools such as decision support systems, with a particular interest in septic shock, the number one killer in intensive care and the single most expensive condition treated in hospitals.

Dominic Marshall is an Academic Foundation doctor in Oxford, UK. Dominic read Molecular and Cellular biology at the University of Bath and worked at Eli Lilly in their Alzheimer's disease drug hunting research program. He pursued his medical training at Imperial College London where he was awarded the Santander Undergraduate scholarship for academic performance and ranked first overall in his graduating class. His research interests range from molecular biology to analysis of large clinical data sets and he has received non-industry grant funding to pursue the development of novel antibiotics and chemotherapeutic agents. Alongside clinical training, he is involved in a number of research projects focusing on analysis of electronic health care records.

Tristan Naumann is a doctoral candidate in Electrical Engineering and Computer Science at MIT working with Dr. Peter Szolovits in CSAIL's Clinical Decision Making group. His research includes exploring relationships in complex,

unstructured data using data-informed unsupervised learning techniques, and the application of natural language processing techniques in healthcare data. He has been an organizer for workshops and "datathon" events, which bring together participants with diverse backgrounds in order to address biomedical and clinical questions in a manner that is reliable and reproducible.

Kenneth Paik is a clinical informatician democratizing access "to healthcare" through technology innovation, with his multidisciplinary background in medicine, artificial intelligence, business management, and technology strategy. He is a research scientist at the MIT Laboratory for Computational Physiology investigating the secondary analysis of health data and building intelligent decision support system. As the co-director of Sana, he leads programs and projects driving quality improvement and building capacity in global health. He received his MD and MBA degrees from Georgetown University and completed fellowship training in biomedical informatics at Harvard Medical School and the Massachusetts General Hospital Laboratory for Computer Science.

Tom Joseph Pollard is a postdoctoral associate at the MIT Laboratory for Computational Physiology. Most recently he has been working with colleagues to release MIMIC-III, an openly accessible critical care database. Prior to joining MIT in 2015, Tom completed his Ph.D. at University College London, UK, where he explored models of health in critical care patients in an interdisciplinary project between the Mullard Space Science Laboratory and University College Hospital. Tom has a broad interest in improving the way clinical data is managed, shared, and analyzed for the benefit of patients. He is a Fellow of the Software Sustainability Institute.

Jesse Raffa is a research scientist in the Laboratory for Computational Physiology at the Massachusetts Institute of Technology in Cambridge, USA. He received his Ph.D. in biostatistics from the University of Waterloo (Canada) in 2013. His primary methodological interests are related to the modeling of complex longitudinal data, latent variable models and reproducible research. In addition to his methodological contributions, he has collaborated and published over 20 academic articles with colleagues in a diverse set of areas including: infectious diseases, addiction and critical care, among others. Jesse was the recipient of the distinguished student paper award at the Eastern North American Region International Biometric Society conference in 2013, and the new investigator of the year for the Canadian Association of HIV/AIDS Research in 2004.

Justin Salciccioli is an Academic Foundation doctor in London, UK. Originally from Toronto, Canada, Justin completed his undergraduate and graduate studies in the United States before pursuing his medical studies at Imperial College London. His research pursuits started as an undergraduate student while completing a biochemistry degree. Subsequently, he worked on clinical trials in emergency medicine and intensive care medicine at Beth Israel Deaconess Medical Center in Boston and completed a Masters degree with his thesis on vitamin D deficiency in critically ill patients with sepsis. During this time he developed a keen interest in statistical

methods and programming particularly in SAS and R. He has co-authored more than 30 peer-reviewed manuscripts and, in addition to his current clinical training, continues with his research interests on analytical methods for observational and clinical trial data as well as education in data analytics for medical students and clinicians.