Leveraging Data Science for Global Health

Leo Anthony Celi · Maimuna S. Majumder ·
Patricia Ordóñez · Juan Sebastian Osorio ·
Kenneth E. Paik · Melek Somai
Editors

Leveraging Data Science for Global Health

Springer

Editors
Leo Anthony Celi
Massachusetts Institute of Technology
Cambridge, MA, USA

Patricia Ordóñez
University of Puerto Rico Río Piedras
San Juan, PR, USA

Kenneth E. Paik
Institute for Medical Engineering and Science
Massachusetts Institute of Technology
Cambridge, MA, USA

Maimuna S. Majumder
Boston Children's Hospital
Harvard Medical School
Boston, MA, USA

Juan Sebastian Osorio
ScienteLab, Department of Global Health
University of Washington
Seattle, USA

Melek Somai
Imperial College London
London, UK

ISBN 978-3-030-47993-0 ISBN 978-3-030-47994-7 (eBook)
https://doi.org/10.1007/978-3-030-47994-7

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

Preface

Historically, physicians have been the sole gatekeepers of medical knowledge. If a patient had a question more complicated than the choice of cold remedies or how to nurse a sprained ankle, they had to make an appointment to see a doctor. With the advent of the Internet and personal computers, patients can now quickly learn about possible diagnoses to explain the presence of blood in the urine, or even take a picture of a mole with a phone to determine if it is cancerous or not. Most of us could not have imagined that watches would be able to diagnose atrial fibrillation, or that body parts and prosthetic devices could be "printed" to produce 3D facsimiles as required. However, with this widespread access to knowledge and technology, there has been an accompanying explosion of incorrect information and data misuse to advance various agendas. Consequently, today's providers need to adapt and learn pertinent aspects of data science if they are to keep up with the information revolution.

Something else happened since the dawn of the Internet: The medical profession as a whole has become more self-critical (Graham et al. 2011; Makary and Michael 2016; Wennberg 2001). In the field of global health, the patient safety and quality improvement movement highlighted deficiencies in the traditional service provision model. The first paper looking at quality of care in developing countries was published in 2012 in the *British Medical Journal* (Wilson et al. 2012). These investigators reviewed more than 15,000 medical records randomly sampled from 26 hospitals in Egypt, Jordan, Kenya, Morocco, Tunisia, Sudan, South Africa, and Yemen. Rather than a lack of medications, laboratory services, or access to specialists, the two biggest factors that contributed the most to poor quality of care were errors in diagnosis and/or treatment. Thus, poor quality in this case ultimately centered around how medical decisions were made. In addition to this finding, a report from the World Health Organization published in 2009 (Shankar 2009) noted that more than 50% of drugs in low- and middle-income countries are prescribed, dispensed, and/or sold inappropriately, and only 1 in 3 are prescribed according to existing clinical guidelines. These two reports highlight opportunities to improve the data-driven support of clinical decision-making around the world.

Research has been traditionally viewed as a purely academic undertaking, especially in limited-resource settings. Clinical trials, the hallmark of medical research, are expensive to perform and take place primarily in countries which can afford them. Around the world, the blood pressure thresholds for hypertension, or the blood sugar targets for patients with diabetes, are established based on research performed in a handful of countries. There is an implicit assumption that the findings and validity of studies carried out in the US and other Western countries generalize to patients around the world.

MIT Critical Data is a global consortium that consists of healthcare practitioners, computer scientists, and engineers from academia, industry, and government, that seeks to place data and research at the front and center of healthcare operations. MIT Sana, an initiative to advance global health informatics, is an arm of MIT Critical Data and focuses on the design, implementation, and evaluation of health information systems. Both MIT Critical Data and MIT Sana are led by the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology. LCP develops and maintains open-access electronic health record databases to support medical research and education (Johnson et al. 2016; Pollard et al. 2018). In addition, it offers two courses at the Harvard-MIT Division of Health Science and Technology: HST.936, Global Health Informatics, and HST.953, Collaborative Data Science in Medicine. The former is now available as a massive open online course HST.936x under edX.

MIT Sana published the textbook for HST.936 of the same name under the auspices of the MIT Press (Celi et al. 2017), while MIT Critical Data members penned the textbook *Secondary Analysis of Electronic Health Records* for HST.953 with Springer (MIT Critical Data 2016). Following a strong belief in an open science model and the power of crowd-sourcing knowledge discovery and validation, both textbooks are available to download free of charge. The latter has been downloaded more than 450,000 times since its publication in 2016. A Mandarin translation is slated for release by the end of the year, and a Spanish translation is in the works.

This book, *Leveraging Data Science for Global Health*, was written and assembled by members of MIT Critical Data. In 2018, HST.936 added data science to digital health as a focus of the course. Lectures, workshops, and projects in machine learning as applied to global health data were included in the curriculum on top of HST.936x, which focuses on digital health infrastructure. *Leveraging Data Science for Global Health*: provides an introductory survey of the use of data science tools in global health and provides several hands-on workshops and exercises. All associated code, data, and notebooks can be found on the MIT Critical Data website http://criticaldata.mit.edu/book/globalhealthdata, as well as hosted in an open repository on Github http://github.com/criticaldata/globalhealthdatabook. We recommend working through and completing the exercises to understand the fundamentals of the various machine learning methods.

Parts I and II of this book are a collection of the workshops taught in the course, plus workshops organized by MIT Critical Data around the globe. The workshops in Part I focus on building an ecosystem within the healthcare system that promotes,

nurtures, and supports innovations, especially those in the field of digital health and data science. Part II dives into the applications of data science in healthcare and covers machine learning, natural language processing, computer vision, and signal processing.

Part III focuses on case studies of global health data projects. The chapters chronicle various real-world implementations in academic and public health settings and present the genesis of the projects, including the technology drivers. Other topics that are covered include the implementation process, key decisions, and lessons learned. While no implementation strategy will be universally applicable to all use cases, we hope the ones presented in this section provide useful insights to assist in successfully developing and deploying global health data projects.

For Part IV, students from the 2018 Harvard-MIT course *Global Health Informatics* have contributed findings from their course projects in the form of scientific manuscripts. Given that developing countries are uniquely prone to large-scale emerging infectious disease outbreaks due to the disruption of ecosystems, civil unrest, and poor healthcare infrastructure, the utility of digital disease surveillance serves as a unifying theme across chapters. In combination with context-informed analytics, this section showcases how non-traditional digital disease data sources—including news media, social media, Google Trends, and Google Street View—can fill critical knowledge gaps and help inform on-the-ground decision-making when formal surveillance systems are insufficient. The final chapter presents an example of how a country can incorporate data science in their curriculums to build capacity that promotes digital transformation in health care.

We believe that learning using data science tools is the best medicine for population health, and that research should be an integral part of global health operations. Every patient encounter is an opportunity that we can learn from, and every healthcare provider should be a contributor and a custodian, and not merely a passive recipient, of the medical knowledge system.

On behalf of MIT Critical Data.

Cambridge, USA — Leo Anthony Celi
Boston, USA — Maimuna S. Majumder
San Juan, USA — Patricia Ordóñez
Seattle, USA — Juan Sebastian Osorio
Cambridge, USA — Kenneth E. Paik
London, UK — Melek Somai

References

Celi, L. A., Fraser, H. S. F., Nikore, V., Osorio, J. S., Paik, K. (2017). *Global health informatics*. Cambridge: MIT Press.

Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Graham, R., Mancher, M., Miller Wolman, D., et al. (Eds.) (2011). *Clinical practice guidelines we can trust*. Washington (DC): National Academies Press (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK209539/, https://doi.org/10.17226/13058

Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci Data*, *3*, 160035.

Makary, M. A., & Michael, D. (2016). Medical error—the third leading cause of death in the US. *BMJIi*, *353*, i2139.

MIT Critical Data. (2016). *Secondary analysis of electronic health records*. New York: Springer.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., Badawi, O. (2018). The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data*, *5*, 180178.

Shankar, P. R. (2009). Medicines use in primary care in developing and transitional countries: Fact book summarizing results from studies reported between 1990 and 2006. *Bull World Health Organ*, *87*(10), 804. https://doi.org/10.2471/09-070417

Wennberg, J. (2001). Unwarranted variation in healthcare delivery: Implications for academic medical centres. *BMJ*, *325*(7370), 961–964.

Wilson, R. M., Michel, P., Olsen, S., Gibberd, R. W., Vincent, C., El-Assady, R., et al. (2012). Patient safety in developing countries: Retrospective estimation of scale and nature of harm to patients in hospital. *BMJ*, *344*, e832.

Contents