

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
ScienceLab, 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>

© The Editor(s) (if applicable) and The Author(s) 2020. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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
 Boston, USA
 San Juan, USA
 Seattle, USA
 Cambridge, USA
 London, UK

Leo Anthony Celi
 Maimuna S. Majumder
 Patricia Ordóñez
 Juan Sebastian Osorio
 Kenneth E. Paik
 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. *BMJi*, 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

Part I Building a Data Science Ecosystem for Healthcare		
1	Health Information Technology as Premise for Data Science in Global Health: A Discussion of Opportunities and Challenges	3
	Louis Agha-Mir-Salim and Raymond Francis Sarmiento	
2	An Introduction to Design Thinking and an Application to the Challenges of Frail, Older Adults	17
	Tony Gallanis	
3	Developing Local Innovation Capacity to Drive Global Health Improvements	35
	Christopher Moses	
4	Building Electronic Health Record Databases for Research	55
	Lucas Bulgarelli, Antonio Núñez-Reiz, and Rodrigo Octavio Deliberato	
5	Funding Global Health Projects	65
	Katharine Morley, Michael Morley, and Andrea Beratarrechea	
6	From Causal Loop Diagrams to System Dynamics Models in a Data-Rich Ecosystem	77
	Gary Lin, Michele Palopoli, and Viva Dadwal	
7	Workshop on Blockchain Use Cases in Digital Health	99
	Philip Christian C. Zuniga, Rose Ann C. Zuniga, Marie Jo-anne Mendoza, Ada Angeli Cariaga, Raymond Francis Sarmiento, and Alvin B. Marcelo	

Part II Health Data Science Workshops

8	Applied Statistical Learning in Python	111
	Calvin J. Chiew	
9	Machine Learning for Patient Stratification and Classification	
	Part 1: Data Preparation and Analysis	129
	Cátia M. Salgado and Susana M. Vieira	
10	Machine Learning for Patient Stratification and Classification	
	Part 2: Unsupervised Learning with Clustering	151
	Cátia M. Salgado and Susana M. Vieira	
11	Machine Learning for Patient Stratification and Classification	
	Part 3: Supervised Learning	169
	Cátia M. Salgado and Susana M. Vieira	
12	Machine Learning for Clinical Predictive	
	Analytics	199
	Wei-Hung Weng	
13	Robust Predictive Models in Clinical Data—Random Forest	
	and Support Vector Machines	219
	Siqi Liu, Hao Du, and Mengling Feng	
14	Introduction to Clinical Natural Language Processing with	
	Python	229
	Leo Anthony Celi, Christina Chen, Daniel Gruhl, Chaitanya Shivade, and Joy Tzung-Yu Wu	
15	Introduction to Digital Phenotyping for Global Health	251
	Olivia Mae Waring and Maiamuna S. Majumder	
16	Medical Image Recognition: An Explanation and Hands-On	
	Example of Convolutional Networks	263
	Dianwen Ng and Mengling Feng	
17	Biomedical Signal Processing: An ECG	
	Application	285
	Chen Xie	

Part III Data for Global Health Projects

18	A Practical Approach to Digital Transformation: A Guide	
	to Health Institutions in Developing Countries	307
	Alvin B. Marcelo	

19	Establishing a Regional Digital Health Interoperability Lab in the Asia-Pacific Region: Experiences and Recommendations	315
	Philip Christian C. Zuniga, Susann Roth, and Alvin B. Marcelo	
20	Mbarara University of Science and Technology (MUST)	329
	Richard Kimera, Fred Kaggwa, Rogers Mwavu, Robert Mugonza, Wilson Tumuhimbise, Gloria Munguci, and Francis Kamuganga	
21	Data Integration for Urban Health	351
	Yuan Lai and David J. Stone	
22	Ethics in Health Data Science	365
	Yvonne MacPherson and Kathy Pham	
23	Data Science in Global Health—Highlighting the Burdens of Human Papillomavirus and Cervical Cancer in the MENA Region Using Open Source Data and Spatial Analysis	373
	Melek Somai, Sylvia Levy, and Zied Mhirsi	
 Part IV Case Studies		
24	A Digital Tool to Improve Patient Recruitment and Retention in Clinical Trials in Rural Colombia—A Preliminary Investigation for Cutaneous Leishmaniasis Research at Programa de Estudio y Control de Enfermedades Tropicales (PECET)	385
	Dr. James Alexander Little, Elizabeth Harwood, Roma Pradhan, and Suki Omere	
25	A Data-Driven Approach for Addressing Sexual and Reproductive Health Needs Among Youth Migrants	397
	Pragati Jaiswal, Amber Nigam, Teertha Arora, Uma Girkar, Leo Anthony Celi, and Kenneth E. Paik	
26	Yellow Fever in Brazil: Using Novel Data Sources to Produce Localized Policy Recommendations	417
	Shalen De Silva, Ramya Pinnamaneni, Kavya Ravichandran, Alaa Fadaq, Yun Mei, and Vincent Sin	
27	Sana.PCHR: Patient-Controlled Electronic Health Records for Refugees	429
	Patrick McSharry, Andre Prawira Putra, Rachel Shin, Olivia Mae Waring, Maiamuna S. Majumder, Ned McCague, Alon Dagan, Kenneth E. Paik, and Leo Anthony Celi	

28 Using Non-traditional Data Sources for Near Real-Time Estimation of Transmission Dynamics in the Hepatitis-E Outbreak in Namibia, 2017–2018 443
Michael Morley, Maiamuna S. Majumder, Tony Gallanis, and Joseph Wilson

29 Building a Data Science Program Through Hackathons and Informal Training in Puerto Rico 453
Patricia Ordóñez Franco, María Eglée Pérez Hernández, Humberto Ortiz-Zuazaga, and José García Arrarás

Epilogue: MIT Critical Data Ideathon: Safeguarding the Integrity of Health Data Science 469