UNIVERSITY OF CALIFORNIA

Los Angeles

Discovering Data-Driven Actionable Intelligence for Clinical Decision Support

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Ahmed M. Alaa H. H. Ibrahim

2019

ABSTRACT OF THE DISSERTATION

Discovering Data-Driven Actionable Intelligence
for Clinical Decision Support

by

Ahmed M. Alaa H. H. Ibrahim
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Mihaela van der Schaar, Chair

The rapid digitization of healthcare has led to a proliferation of clinical data, manifesting through electronic health records, biorepositories, and disease registries. This dissertation addresses the question of how machine learning (ML) techniques can capitalize on these data resources to assist clinicians in predicting, preventing and treating illness. To this end, we develop a set of ML-based, data-driven models of patient outcomes that we envision to be embedded within systems of decision support deployed at different stages of patient care.

We focus on two broad setups for analyzing clinical data: (1) the *cross-sectional* setup wherein data is collected by observing many patients at a particular point of time, and (2) the *longitudinal* setup in which repeated observations of the same patient are collected over time. In both setups, we develop models that are: (a) capable of answering *counter-factual* questions, i.e., can predict outcomes under alternative treatment scenarios, (b) *interpretable* in the sense that clinicians can understand how the model predictions for individual patients are issued, and (c) *automated* in the sense that they adaptively tune their modeling choices for the dataset at hand, with little or no need for expert intervention. Models satisfying these *three* requirements would enable the realization of actionable, transparent and automated decision support systems that operate symbiotically within existing clinical workflows.

Our technical contributions are multi-faceted. In the cross-sectional data setup, we develop ML models that fulfill the aforementioned requirements (a)-(c) as follows. We start by

developing a comprehensive theoretical framework for causal inference, whereby we quantify the limits to how well ML models can recover the causal effects of counter-factual treatment decisions on individual patients using observational (retrospective) data, and we build ML models — based on *Gaussian processes* — that achieve these limits. Next, we develop a novel *symbolic meta-modeling* approach for interpreting the predictions of any ML-based prognostic model by converting the "black-box" model into an understandable symbolic equation that relates patients' features to their predicted outcomes. Finally, we develop a model selection approach based on *Bayesian optimization* that enables the automation of predictive and causal modeling. In the longitudinal data setup, we develop a novel deep probabilistic model for sequential clinical data that satisfies requirements (a)-(c) by capitalizing on the strengths of both state-space models and deep recurrent neural networks.

To demonstrate the utility of our models, we evaluate their performance on various real-world datasets for cohorts of breast cancer, cardiovascular disease and cystic fibrosis patients. We show that, compared to existing clinical scorers, our ML-based models can improve the accuracy of predicting individual-level prognoses, guide treatment decisions for individual patients, and provide insights into underlying disease mechanisms.

The dissertation of Ahmed M. Alaa H. H. Ibrahim is approved.

Greg Pottie

Yahya Rahmat-Samii

Patricia Ganz

Douglas Bell

Mihaela van der Schaar, Committee Chair

University of California, Los Angeles

2019

To my parents and my brothers.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would not have been able to complete this dissertation without the help and support of my family, my adviser, my colleagues and friends. I am deeply grateful.

First, I wish to thank my adviser, Professor Mihaela van der Schaar, for her kindness and unyielding support. This dissertation would not have been possible without her research vision. Her new ideas, insights, opportunities and willingness to intrepidly explore multidisciplinary research terrains have had a profound impact on my development as a researcher. I am deeply grateful for her invaluable support and immense help. Thank you for everything, Mihaela. I also thank the other members of my dissertation committee, Professor Greg Pottie, Professor Yahya Rahmat-Samii, Professor Patricia Ganz, and Dr. Douglas Bell for their valuable perspectives and thoughtful feedback.

I would also like to thank all of my co-authors and collaborators at UCLA; Jinsung Yoon, Scott Hu, William Zame, Changhee Lee, Kartik Ahuja, William Hsu, Kyeong Moon and Martin Cadeiras. It has been a great privilege to be able to work with such talented people. I have also been fortunate to enjoy the collegiality of my labmates; Onur Atan, Trent Kyono, William Whoiles, Yangbo Song, Simpson Zhang, Jie Xu and Yuanzhang Xiao, with whom I had many insightful research discussions. I would also like to thank my friends at UCLA for their companionship and support. I especially want to thank Ahmed Hareedy, Mohammed Karmoose, Yahya Ezzeldin, Omar Hussien, Moustafa Alzantot, and Safa Cicek.

I have been so lucky to spend some time on the other side of the Atlantic as a visiting student at Oxford University. There, I was fortunate to work with my colleagues at Oxford University, Ioana Bica, James Jordon, Bryan Lim, and Michael Weisz, and my colleagues at Cambridge University, Yao Zhang, Zhaozhi Qian, Alexis Bellot, and Daniel Jarrett.

During my visit in the UK, I was blessed to be able to collaborate with many clinicians who provided me with the clinical data, guidance and feedback that made the clinical application of my work come to life. I would like to especially thank Dr. Jem Rashbass (National Director for Disease Registration at Public Health England) for granting me access to the UK breast cancer registry data, and Dr Janet Allen (Director of Strategic Innovation at the UK Cystic Fibrosis Trust) for enabling my access to the UK Cystic Fibrosis registry. I would also like to thank my clinical collaborators at Cambridge University, Andres Floto, Fiona Gilbert, Emanuele Di Angelantonio, James Rudd, and my collaborator at Queen Mary University of London, Deepti Gurdasani.

Finally, and most importantly, I wish to thank my parents, Hadeel and Alaa, and my borthers, Ali and Sherif for their love and support. No words can express what your encouragement and support have meant to me.

VITA

2011	Bachelor in Communications and Computer Engineering, Cairo University, Cairo, Egypt
2011–2014	Teaching Assistant, Communications Engineering Department, Cairo University, Cairo, Egypt
2014	Master of Science in Electronics and Communications Engineering, Cairo University, Cairo, Egypt
2014	Graduate Division Fellowship, Electrical Engineering, University of California, Los Angeles.
2014–2019	Research Assistant, University of California, Los Angeles.
2017–2018	Recognized PhD Student, University of Oxford, United Kingdom.

PUBLICATIONS

A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, M. van der Schaar, "Cardiovascular Disease Risk Prediction using Automated Machine Learning: A Prospective Study of 423,604 UK Biobank Participants," *PloS One*, 2019.

A. M. Alaa, M. van der Schaar, "Demystifying Black-box Models with Symbolic Metamodels," *Neural Information Processing Systems* (NeurIPS), 2019.

A. M. Alaa, M. van der Schaar, "Attentive State-Space Modeling of Disease Progression," *Neural Information Processing Systems* (NeurIPS), 2019.

A. M. Alaa, M. van der Schaar, "Validating Causal Inference Models via Influence Functions," *International Conference on Machine Learning* (ICML), 2019.

I. Bica, **A. M. Alaa**, M. van der Schaar, "Estimating Counterfactual Treatment Outcomes over Time through Adversarially Balanced Representations," *NeurIPS Machine Learning for Health Workshop*, 2019.

I. Bica, **A. M. Alaa**, M. van der Schaar, "Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders," *NeurIPS Machine Learning for Health Workshop*, 2019.

C. Lee, W. R. Zame, **A. M. Alaa**, M. van der Schaar, "Temporal Quilting for Survival Analysis," *International Conference on Artificial Intelligence and Statistics* (AISTATS), 2019.

A. M. Alaa, M. van der Schaar, "Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning," *Scientific Reports*, 2018.

A. M. Alaa, M. van der Schaar, "Bayesian Nonparametric Causal Inference: Information Rates & Learning Algorithms," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

J. Yoon, W. R. Zame, A. Banerjee, M. Cadeiras, **A. M. Alaa**, M. van der Schaar, "Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation," *PloS One*, 2018.

B. Lim, **A. M. Alaa**, M. van der Schaar, "Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks," *Neural Information Processing Systems* (NeurIPS), 2018.

A. M. Alaa, M. van der Schaar, "AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning," *International Conference on Machine Learning* (ICML), 2018.

A. M. Alaa, M. van der Schaar, "Limits of Estimating Heterogeneous Treatment Effects:

Guidelines for Practical Algorithm Design," *International Conference on Machine Learning* (ICML), 2018.

A. M. Alaa, T. Daniels, R. Floto, M. van der Schaar, "Machine Learning-Based Predictions of Prognosis in Cystic Fibrosis," *Pediatric Pulmonology*, 2018. **(Abstract)**

A. M. Alaa, T. Bolton, E. Di Angelantonio, J. Rudd, M. van der Schaar, "Cardiovascular Disease Risk Prediction Using Machine Learning: A Prospective Cohort Study of 423,604 Participants," *Circulation*, 2018. **(Abstract)**

A. Banerjee, J. Yoon, W. R. Zame, M. Cadeiras, **A. M. Alaa**, M. van der Schaar, "Personalized risk prediction for wait-list and post-transplant mortality in cardiac transplantation: machine learning using predictive clusters," *European Heart Journal*, 2017. **(Abstract)**

A. M. Alaa, M. van der Schaar, "A Hidden Absorbing Semi-Markov Model for Informatively Censored Temporal Data," *Journal of Machine Learning Research*, 2017.

A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized Risk Scoring for Critical Care Prognosis using Mixtures of Gaussian Processes," *IEEE Transactions on Biomedical Engineering*, 2017.

A. M. Alaa, M. van der Schaar, "Deep Multi-task Gaussian Processes for Survival Analysis with Competing Risks," *Neural Information Processing Systems* (NeurIPS), 2017. **(Selected for a spotlight presentation)**

A. M. Alaa, M. van der Schaar, "Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes," *Neural Information Processing Systems* (NeurIPS), 2017.

A. M. Alaa, M. Weisz, M. van der Schaar, "Deep Counterfactual Networks with Propensity-Dropout," *ICML Workshop on Principled Approaches to Deep Learning*, 2017.

A. M. Alaa, M. van der Schaar, "Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis," *International Conference on Machine Learning* (ICML), 2017.

A. M. Alaa, M. van der Schaar, "Individualized Risk Prognosis for Critical Care Patients: A Multi-task Gaussian Process Model," *Big Data in Medicine: Tools, Transformation and Translation, Cambridge*, 2017.

A. Banerjee, J. Yoon, W. R. Zame, M. Cadeiras, **A. M. Alaa**, M. van der Schaar, "Personalized Risk Prediction using Predictive Pursuit Machine Learning: A Pilot Study in Cardiac Transplantation," *European Society of Cardiology Congress*, 2017. **(Selected as Best Poster in Advanced Heart Failure)**

A. M. Alaa, K. Ahuja, and M. van der Schaar, "A Micro-foundation of Social Capital in Evolving Social Networks," *IEEE Transactions on Network Science and Engineering*, 2017.

A. M. Alaa, M. van der Schaar, "Balancing Suspense and Surprise: Timely Decision Making with Endogenous Information Acquisition," *Neural Information Processing Systems* (NeurIPS), 2016.

A. M. Alaa, M. van der Schaar, "A Semi-Markov Switching Linear Gaussian Model for Censored Physiological Data," *NeurIPS workshop on Machine Learning for Health*, 2016.

A. M. Alaa, K. H. Moon, W. Hsu and M. van der Schaar, "ConfidentCare: A Clinical Decision Support System for Personalized Breast Cancer Screening," *IEEE Transactions on Multimedia — Special Issue on Multimedia-based Healthcare*, 2016.

A. M. Alaa, M. van der Schaar, "ForecastICU: A Prognostic Decision Support System for Timely Prediction of Intensive Care Unit Admission," *International Conference on Machine Learning* (ICML), 2016.

A. M. Alaa, J. Yoon, M. van der Schaar, "Personalized Risk Scoring for Critical Care

Patients using Mixtures of Gaussian Process Experts," *ICML Workshop on Computational Frameworks for Personalization*, 2016.

A. M. Alaa, K. Ahuja, M. van der Schaar, " Self-organizing Networks of Information Gathering Cognitive Agents," *IEEE Transactions on Cognitive Communications and Networking - Inaugural issue (invited paper)*, 2015.

CHAPTER 1

Introduction

Current advances in health information technology — including digital patient records and data management tools, wearable devices, efficient methods for genomic sequencing — are expected to drastically increase the amount of data collected for individual patients through electronic health records (EHR), biorepositories, and disease registries. The proliferation of health data is evident by the dramatic increase in the rate of adoption of EHR technologies in healthcare facilities all over the developed world; in 2015, 84% of hospitals in the US adopted an EHR system, which represents a 9-fold increase since 2008 [1].

The availability of large-scale data resources that keep track of patients' features and health outcomes paves the way for more *individualized* approaches to patient care, whereby examples and experiences encoded in data for previous patients are used to unravel disease phenotypic diversity. To achieve this, data by itself is not sufficient — we need *models* that learn from this data how prognoses would vary among future patients based on their individual traits. In this dissertation, we use machine learning (ML) to develop such models — we envision our models to be embedded within systems of decision support deployed at different stages of patient care to assist clinicians in predicting, preventing and treating illness.

1.1 Machine Learning for Individualized Medicine

By machine learning (ML) we mean the process by which computer systems can learn directly from data, examples and experience, rather than being taught on the basis of predetermined rules. The purpose of this chapter is to illustrate some of the progress ML has already made in healthcare and to suggest some of the progress it might make — and ought to make — in

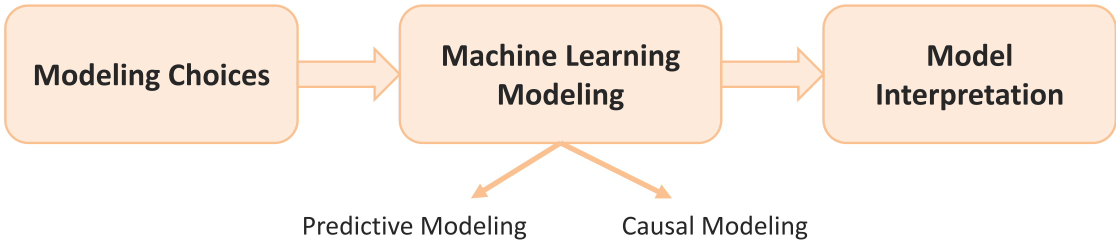

Figure 1.1: Illustration for the typical machine learning modeling pipeline.

the near future. The view presented here is deliberately optimistic; for ML to have a chance to achieve this potential — as we believe it does — there must first be a vision of what is possible. Topol has discussed at length the potential of accumulating more data; our focus here is on extracting more information from that data.

Our biggest challenge now is not whether we have enough data but whether we can combine the limitless potential of machines and the perennially limited potential of human judgement and decision making to use the oceans of data in which we are already swimming. And we should be careful to call out different dimensions of data. It covers electronic medical records, the collection and use of phenotypic and genetic data, data around performance and outcomes at an individual and population level and so on. And further, it will engage a wider range of data about the social determinants of health — in areas like housing employment, retail patterns, income and inequality data — whose impact on health intervention and outcomes will be increasingly critical.

The rapid digitization of healthcare has led to a proliferation of clinical data, manifesting through electronic health records, biorepositories, and disease registries. This dissertation addresses the question of how machine learning (ML) techniques can capitalize on these data resources to assist clinicians in predicting, preventing and treating illness. To this end, we develop a set of ML-based, data-driven models of patient outcomes that we envision to be embedded within systems of decision support deployed at different stages of patient care.

We are in the midst of a revolution in the amount of data being generated. There are many uses for this data; it can be analyzed and interpreted, often through powerful machine learning algorithms, edited, manipulated, distilled or reconstructed, and synchronized. In all cases the data must also be stored. All of these tasks must be performed in an environment of uncertainty; the underlying data operated on is never guaranteed to be reliable. The data may have been changed or edited knowingly or unknowingly, and is always subject to corruption from transmission and storage noise.

1.1.1 Models for Cross-sectional Data

1.1.2 Models for Longitudinal Data

1.2 Outline of Contributions

Chapter 2 Contributions

Chapter 3 Contributions

Chapter 4 Contributions

Chapter 5 Contributions

Chapter 6 Contributions

CHAPTER 2

Estimating Treatment Effects from Observational Data

For text, let's use the first words out of the ispell dictionary.

CHAPTER 3

Symbolic Approaches to Prognostic Model Interpretability

CHAPTER 4

Automated Prognostic Modeling

CHAPTER 5

Deep Probabilistic Modeling of Longitudinal Data

For text, let's use the first words out of the ispell dictionary.

CHAPTER 6

Clinical Application

CHAPTER 7

Conclusions

REFERENCES

[1] Karen B DeSalvo, Ayame Nagatani Dinkler, and Lee Stevens. The us office of the national coordinator for health information technology: progress and promise for the future at the 10-year mark. *Annals of emergency medicine*, 66(5):507–510, 2015.