Chapter 1

Machine learning for healthcare technologies – an introduction

David A. Clifton[1]

1.1 The changing needs of healthcare

Much has been written concerning the manner in which healthcare is changing, with a particular emphasis on how very large quantities of data are now being routinely collected during the routine care of patients. The use of machine learning methods to turn these ever-growing quantities of data into interventions that can improve patient outcomes seems as if it should be an obvious path to take. However, the field of machine learning in healthcare is still in its infancy. This book, kindly supported by the Institution of Engineering and Technology, aims to provide a "snapshot" of the state of current research at the interface between machine learning and healthcare.

Necessarily, this is a partial and biased sampling of the state of current research, and yet we have aimed to provide a wide-ranging introduction to the depth and scale of work that is being undertaken worldwide. In selecting material for this edited volume, we have placed special emphasis on machine learning projects that are (or are close to) achieving improvement in patient outcomes. For many reasons, uncovered variously in some of the chapters that follow, it is a truism that "healthcare is hard"; there are unique constraints that exist, and considerations that must be taken, when working with healthcare data. However, for all its difficulties, working with healthcare data is exceptionally rewarding, both in terms of the computational challenges that exist and in terms of the outputs of research being able to affect the way in which healthcare is delivered. There are few application areas of machine learning that have such promise to benefit society as does that of healthcare.

1.2 Online resources

The remainder of this chapter seeks to survey the various research programmes described in this book, and draws the readers' attention to the fact that many of

[1] Balliol College, Oxford

the projects described were presented by the authors in person, at a workshop held at Balliol College, Oxford, during the summer of 2015. The Institution of Engineering and Technology (IET) video-recorded the event, and has made the resulting recordings available for free via its IET.tv online resource. There is seldom a better introduction to a research programme than to hear it described by its originators, and so we hope that the reader finds the online resources to be a useful and accessible complement to the more comprehensive descriptions provided in this volume.

Balliol College celebrated its 750th anniversary a little while before the "Machine Learning in Healthcare" workshop took place, and it has a good claim to being the oldest college, in what is the oldest university of the English-speaking world. Since coming to Balliol from a series of other colleges in the university, I have been greatly impressed by the open spirit of enquiry that exists in the place, and the substantive and very real manner in which it supports its fellows to do the same. The workshop combined the extensive efforts of both Balliol and the IET, in which the old buildings were repurposed for professional video-filming and audio recording, and where attendees came from both academia and (healthcare) industry to discuss the work presented in this volume. As a "snapshot" of some of the best work at the interface between machine learning and healthcare, it struck me as being fitting that this should be recorded against the backdrop of an institution that remains both ancient and modern, in the best of ways. It is our hope that some of this spirit is apparent in both the videos maintained on the IET.tv website, and the volume before us.[2]

1.3 Survey of contents

Chapter 2: The team led by Prof. Chris Williams at the University of Edinburgh, UK, has long been at the forefront of various aspects of machine learning, and an important theme of their work is the application of time-series analysis methods to healthcare applications – most notably, those pertaining to the intensive care unit (ICU) in the hospital. The ICU is a data-rich environment, in which patients are typically monitored continuously for the duration of their stay, and where the nurse-to-patient ratio is typically 1:1 in many healthcare systems. Entering an ICU is to be deluged by data in all its forms: various machines, which may or may not be interoperable, report measurements to the clinician almost constantly; there is typically a great deal of "alarm noise" from the various devices, and one receives the impression that there is far more data being generated than can be meaningfully interpreted by a human – even a highly trained expert as is typically the case with ICU clinicians. On seeing such an environment, one almost immediately concludes that machine learning has a key role to play in aiding the clinician, by guiding their attention to those components of the data that are most pertinent. Chapter 2 describes one such approach, in which a factorial switched linear dynamical system (FSLDS) is used to make sense of the data,

[2](https://tv.theiet.org/?event=3534)

with the goal of understanding what within a signal can be described as artefact, and what is clinically important information. It seems fitting to encounter this material first in the book, given the themes of signal understanding and subsequent modelling that Prof. Williams and colleagues describe.

Chapter 3: We noted previously the old adage that "healthcare is hard," and a contributing factor to this is that biomedical devices typically operate independently, without knowledge of other aspects of the patient's physiology other than that which it is measuring. Prof. Gari Clifford of Emory University, USA, is a long-standing contributor to the field of computational approaches to cardiology, and in performing analysis in the presence of the substantial noise that typically exists when patients are monitored while ambulatory. The latter is an important factor in the limited impact that "mobile health" (or m-health) has had in clinical practice, due to the fact that most ambulatory monitoring systems are typically insufficiently robust due to an inability to cope with such data uncertainty. Chapter 3 combines these two themes, by developing methods that permit the identification of atrial fibrillation from the electrocardiogram (ECG) under the most testing of circumstances. Again, these themes appear early in the book, due to the commonality that it shares with most chapters subsequently described.

Chapter 4: Continuing the topic of handling noisy biomedical waveform data, Dr. Julien Oster of the University of Oxford describes advances in the development of Bayesian filters for detecting important features of clinical relevance in the ECG. Dr. Oster's research career has focussed on the interface between cardiology and machine learning, where the goal is to improve upon human annotations and analysis of the ECG – as is appropriate when one is, for example, faced with screening very large quantities of cardiac data. This chapter provides a helpful tutorial in how the framework of Bayesian filtering may be applied to cardiology, and how one can use a generative model to understand the ECG waveform.

Chapter 5: Readers will probably be familiar with the traditional decompositions often used to summarise high-dimensional data by using a lower-dimensional version that can be used for parsimonious inference. For example, the likes of principal components analysis and its derivatives are becoming popular in many branches of genomic analysis. However, such representations of data are typically simplistic, and fail to capture much of the structure that may be present – Prof. Maarten de Vos and his team at the University of Oxford present a tutorial for using methods based on tensor decompositions for better understanding the structure in EEG (and other biomedical) time-series. Chapter 5 provides a helpful introduction to how and why tensor decomposition can be used in such situations, with illustrations of the method in the application area of detecting epileptic seizures.

Chapter 6: With earlier chapters concerning themselves with the understanding and modelling of biomedical waveform data, this chapter looks at methods by which data may be compared across patients. Based on some of the work from the Computational Health Informatics (CHI) Lab at Oxford, Dr. Marco Pimentel describes the development of principled, probabilistic methods for performing inference across entire time-series of patient data. We undertook a study of over 300 post-operative patients at Oxford University Hospitals, and present the results of how such methods might be used to provide risk stratification – the ultimate goal is to understand, as early as possible, which patients are at the highest risk of deterioration – a problem that is critical, because the mortality rate in this patient group approaches 1 in 6.

Chapter 7: In the second chapter from the CHI Lab at Oxford, Dr. Tingting Zhu presents the means by which the outputs of multiple algorithms may be fused to improve accuracy of classification for biomedical tasks. She focuses her attention on a cardiology application, similar to that addressed in Chapters 3 and 4, but where a panel of algorithms exist. She describes a fully Bayesian methodology for assuming that the outputs of each of these algorithms (which may be automated computational algorithms or, in the case of cardiology, human experts) are "noisy" with respect to the correct output; the noise distribution for each algorithm is then learned in an *unsupervised* manner. Dr. Zhu shows that the resulting estimates, across all algorithms, typically outperform even the single-best algorithm. This is especially appealing for the use of machine learning systems, whereby we may have multiple algorithms that have been created for a single task, the results of which may be fused to produce robust outputs.

Chapter 8: Prof. Suchi Saria and her team at Johns Hopkins University, USA, take a novel look at the case of competing models: the dollar-value associated with acquiring individual data-points for a patient in a healthcare setting is incorporated into a regulariser. This recognises the fact that acquiring different data may be associated with more costly measurement procedures; for example, ordering blood tests for an ICU patient may be more expensive (in terms of dollar-value) that acquiring another heart-rate estimate from a bedside monitoring. By taking this information into account within the regularisation framework, an estimate is provided of how the predictive accuracy of risk assessment (here, for risk of developing septic shock) varies as available dollar-value increases. While the "true" costs of estimating various data types is notoriously difficult to quantify (especially in centralised healthcare systems, such as the UK National Health Service), Prof. Saria's approach helps us make informed choices concerning which risk assessment system should be used, for example – where such decisions are typically made using predictive accuracy alone, without any information of the costs of data acquisition.

Chapter 9: Dr. James Hensman of the University of Lancaster and Prof. Theodore Kypraios of the University of Nottingham have an ongoing collaboration in which they have developed Bayesian non-parametric models for understanding the outbreak and spread of infectious disease. This chapter explores the log Gaussian Cox process, which is an interesting extension of the much-used Cox process, with its relationship to the traditional *Susceptible-Infective-Removed* epidemic model. This chapter provides, among other contributions, a helpful tutorial on the use of variational inference methods for estimating the values of the hyperparameters of a Gaussian process, used within the log Gaussian Cox process. There are many applications in which "arrival times" or rates are of interest, to which the methods described in this chapter are directly applicable.

Chapter 10: Perhaps one of the most troubling recent developments in healthcare is that of increasing antibiotic resistance, whereby bacteria are developing (via accelerated natural selection) resistance to the various classes of antibiotics with which we treat infection. As resistance increases, our ability to combat infectious disease becomes more limited, and we must turn to treatments that are potentially harmful for the patient. This problem is compounded by the fact that assessing resistance to antibiotics involves taking a biological sample from a patient, isolating the bacteria that are causing infection, and then growing those bacteria in a microbiological lab such that various antibiotics can be tested on those bacteria. For some strains, such as *Mycobacterium tuberculosis*, this process can take over one month. In Chapter 10, Dr. Yang Yang describes work from the CHI Lab at Oxford concerning the use of near-same-day genetic sequencing, in which the bacterium itself is sequenced. Machine learning methods are then applied to the results to estimate antibiotic resistance – in a fraction of the time taken by conventional methods, thereby allowing us to treat infectious disease in a timely manner, which timely treatment of the patient is especially important.

Chapter 11: Chronic disease is one of the greatest burdens on most healthcare systems, and, in this chapter, Katherine Niehaus introduces work from the Oxford CHI Lab on improving our understanding of various classes of immune disease. This work involves close collaboration with medical colleagues from Oxford University Hospitals, in which we have acquired genomic, time-series and other data for a large cohort of patients suffering from these types of disease. The challenge for machine learning is to determine how best one should link these very different classes of data; this chapter includes a description of methods from extreme value theory that are being used to assess "beyond normal" data – as are often acquired from patients with immune disease.

Chapter 12: Representing contributions from MIT and Harvard, Prof. Shamim Nemati describes "big data" approaches to a number of exemplar applications within healthcare, including the estimation of the dose of medication that should be provided to a patient. Conventional clinical methods of performing this estimation typically derive from simplistic factors, such as using initial measurements of the weight of the patient followed by a laboratory test performed some hours later. With the wide range of data available via the electronic medical record, this chapter describes how data-driven approaches can be used to improve upon standard clinical practice. Shamim's work includes analysis of the MIMIC-2 open-source dataset, created and curated by the Laboratory for Computational Physiology at MIT of which he is a member. Readers may be familiar with this resource as being a great asset to global critical-care research; it is no exaggeration to report that the editor alone knows at least 25 young data scientists from across the world who obtained their doctoral degrees thanks to the availability of MIMIC.

Chapter 13: Few application areas for healthcare are more testing than that of monitoring patients in their own homes. Such is the focus of the research of Prof. Bart Vanrumste and his team at KU Leuven, Belgium, in which sensors are embedded throughout a subject's home and where systems based on machine learning seek to identify patterns in the activities of daily living. A *novelty detection* approach can be taken, whereby deviation from a previously established model of normality can be used to highlight significant changes in mental- or physical-health status for a patient. Chapter 13 describes a number of approaches to this problem, including one based on extreme value theory using point processes, which is a branch of statistics typically employed to identify extremal observations – often from finance, meteorology or climate data.

Acknowledgement

DAC gratefully acknowledges the support of the Royal Academy of Engineering, the UK Engineering and Physical Sciences Research Council, the Wellcome Trust, the UK National Institute of Health Research, UNICEF, the UK Natural Environment Research Council, the UK Department for International Development, the Bill & Melinda Gates Foundation and Balliol College.