Research on Machine Learning for Biomedical Research

ZHAO, Kai

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
School of Biomedical Sciences

Supervised by

Prof. So Hon-cheong

The Chinese University of Hong Kong
July 2020

Thesis Assessment Committee

Professor Cheng Sze Lok Alfred (Chair)

Professor So Hon-cheong (Thesis Supervisor)

Professor Chen Yangchao (Committee Member)

Professor Sham Pak Chung, HKU (External Examiner)

Abstract of thesis entitled:

Research on Machine Learning for Biomedical Research

Submitted by ZHAO, Kai

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in July 2020

This is the abstract in no more than 350 words.

Acknowledgement

I would like to thank my supervisor Prof. So Hon-cheong for offering me the opportunity to work with him. He teaches me how to conduct research works and gave me generous help in my daily life. He is humorous and highly empathic. Truth to be told, it is a wonderful journey to study and work with him, and I am lucky to have the experience.

I also would like to thank my wife. It's nearly ten years since the first meet, and she has become the most important person in my life. Thanks for supporting my decision to pursue a higher degree and bearing all burden from family to free me from distractions. This is not easy to her. You are a brave girl and wonderful mother for the kids. My any achievement is impossible without you.

Thanks my father and mother for never saying NO to the decision of further my education and for taking care of my kids in the past years. I know the hardness for you to bear the burden. I highly appreciate the support.

Thanks my kids for tolerating my absence of parental responsibility for the past year. You let me be your father and share tremendous joy with me. I promise my love to you will alway be the same.

Thanks my lab mates for having some awesome years with you. You are a part of my daily life in those years. The times we spent together will be a precious memory.

Thanks everybody who helped me for their kindness!

Dedicated to those who risked their life to fight against COVID-19.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

□ **End of chapter.**

Chapter 2

Background Study

□ **End of chapter.**

Chapter 3

Drug Repurposing

□ End of chapter.

Chapter 4

Drug Target Discovery

□ **End of chapter.**

Chapter 5

Evaluating ITE of Genetic RFs on Survival

5.1 Motivation

Traditional biomedical or clinical studies in the area of estimating treatment effect mainly focus on the average effect of risk factors (RFs) or treatment (tx) in population level. However, in the clinical environment we can easily find that the same risk factor may affect patients differently. Thus, patients pay more attention to how a risk factor will affect them in an individual level rather than in a population level, given their clinical backgrounds and genetic characteristics. The main objective of this study is to

resolve this concern by estimating the ITEs for each patients, with consideration of their unique genetic and clinical information. Here we consider the two term ”risk factor” and ”treatment” conceptually equivalent, since a risk factor can be treated as a ”treatment” with side effects. This approach allow us to offer tailored health management to individual patients. This enables us to deliver more cost-effective prevention or treatment strategies to bring the most benefits to them. This idea is also in accordant with the aim of ”personalized medicine”, which has been advocated in recent years.

In spite of an increasing number of studies in this area, current studies in ITE are rather limited. Some critical limitations include lack of well-established validation method for treatment effect estimations and for key features contributed to ITE estimation and failure of handling censored data. Even though genetic factors may determine heterogenous response to tx/RFs, especially to cancer treatments, current studies on ITE have not included genomic features. Here we propose methodologies to overcome the above limitations and apply the ITE framework to genomic data. In our approach genomic features will be consider as risk factors or co-variates that explain to heterogeneity of treatment effect.

5.2 Background

It has been long recognized that the same risk factor (RF) or treatment (tx) can impact differently on different individuals. For example, while being overweight or obese is a risk factor for cardiometabolic (CM) diseases, not all overweight subjects will develop such complications2. The type and severity of CM complications can also vary among subjects2. While stressful life events are an RF for depression, only a subset of people will be affected11. The same is applicable to other RFs or treatments (an RF can be considered as a 'treatment' with adverse effect; the two entities are conceptually equivalent). Such heterogeneity of responses to RFs/tx may be attributed to different genetic and environmental backgrounds of subjects. In addition to clinical factors, one may consider genetic variants or mutations as RFs. Notably, the same variant/mutation can have varying effects on different subjects3 4,5 (see 'objectives' section).

The past decade has witnessed remarkable developments in omics technology and massive growth in biomedical data. However, epidemiology research and genetic studies in cancer and other complex diseases are still largely limited to one clinical/genetic RF at a time, ignoring interactions

with the individual's genetic/clinical characteristics. For the patient, the most pertinent question is: how would a risk factor or treatment affect people with a similar (genetic and clinical) background like me? Nevertheless, current studies have largely focused on the average effect of RFs in the population instead of individualized effects.

Here we aim to develop and apply an analytic framework for unraveling the individualized effects of RFs (or tx), such that we can predict their impact for each person, given his/her unique genetic/clinical background. We also develop methods to reveal the key features contributing to such tx effect heterogeneity. We will apply the method to large-scale genome-wide association studies (GWAS) to uncover and predict individualized effects of RFs (e.g. obesity, hyperlipidemia or expression changes of risk genes) on disease risks/outcomes. The framework will also be applied to cancer data to discover the individualized effects of genetic changes (e.g. mutations, CNVs etc.) and other RFs on survival.

5.3 Overview of Related Work

5.4 ITE Framework

5.5 Experiment Results

5.6 Conclusion

Bibliography data is put in database.bib.

□ End of chapter.

Chapter 6

Conclusion

Appendix A

Proof of Propositions

□ **End of chapter.**

Appendix B

Publication List

□ **End of chapter.**

Bibliography