With the rapidly developing techniques for data collection, storage and computation, the era for big data has come. For clinical studies, in particular, there is an increasing need in handling big data issues, as various forms of high dimensional data from clinical trials are rapidly accumulated with massive sample size. These include the ever-expanding electronic medical record (EMR) system, metabolomics and proteomics data from lab tests, and imaging data from MRI. These data provide us great opportunities for improvements in clinical research. However, they also bring big challenges for data analysis, for example, the famous "curse of dimensionality".

There has been substantial researches done in the area of high dimensional statistics and machine learning to handle big data issues, but these research areas are still quite open and actively studied. Furthermore, big data in clinical studies has their own features, providing more challenges, for example, how to handle missing data in high dimensional settings, how to perform high dimensional survival data analysis, how to study structured data and how to extend causal inference analysis to the high dimensional settings. Theoretical efforts have been made in the last several decades in all these areas and there are many useful tools developed that can be potentially applied to clinical studies.

This book addresses the above challenges encountered in clinical studies and review a broad range of methods from modern statistics and machine learning for big clinical data analysis. This book also provides instructions for many useful R packages with detailed example codes so that reader could apply these new statistical methods to their own studies. The organization of the book is as follow.

There are six chapters in this book. Chapter 1 provides an overview of the big data clinical research, including the perspective, the general accessing workflow, a brief review of machine learning methods and data acquisition and management. Chapter 2 discusses about exploratory data analysis and data management. It focuses on the missing data problem that is frequently encountered in clinical studies by introducing a number of methods and their applications. First it discusses about missing data exploration and data reshaping and aggregating. Then it introduces several imputation methods including single imputation, multiple imputation, and multivariate imputation. Chapter 3 discusses methods for variable selection for both parametric and non-parametric models that are commonly used in clinical studies. It also discusses about methods for diagnostic and introduced a useful R package to draw Nomograms. Chapter 4 discusses about the analysis of survival data. In this chapter both the application of parametric and semi-parametric models are illustrated, as well as the competing risk model. Chapter 5 discusses several commonly used unsupervised and supervised machine learning methods including the k nearest neighbor, naïve Bayes classification, decision tree and neural network. Chapter 6 addresses a number of other important statistical areas that has applications in clinical studies, for example, the hierarchical cluster analysis and its visualization with R, causal mediation analysis, structural equation modeling, and case-crossover design.

With this book, we hope to provide reader a comprehensive introduction to big data clinical studies and easy to follow data analysis applications with R examples that will potentially encourage big data analysis in clinical studies. I would like to sincerely thank Dr. Zhongheng Zhang for the opportunity to write the preface for this book.

Cheng Zheng, PhD
Joseph. J. Zilber School of Public Health,
University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA