Preface

Prediction models are important in various fields, including medicine, physics, meteorology, and finance. Prediction models will become more relevant in the medical field with the increase in knowledge on potential predictors of outcome, e.g. from genetics. Also, the number of applications will increase, e.g. with targeted early detection of disease, and individualized approaches to diagnostic testing and treatment. The current era of evidence-based medicine asks for an individualized approach to medical decision-making. Evidence-based medicine has a central place for meta-analysis to summarize results from randomized controlled trials; similarly prediction models may summarize the effects of predictors to provide individualized predictions of a diagnostic or prognostic outcome.

Why Read This Book?

My motivation for working on this book stems primarily from the fact that the development and applications of prediction models are often suboptimal in medical publications. With this book I hope to contribute to better understanding of relevant issues and give practical advice on better modelling strategies than are nowadays widely used.

Issues include:

(a) Better predictive modelling is sometimes easily possible; e.g. a large data set with high quality data is available, but all continuous predictors are dichotomized, which is known to have several disadvantages.
(b) Small samples are used:

- Studies are underpowered, with unreliable answers to difficult questions such as "Which are the most important predictors in this prediction problem?"
- The problem of small sample size is aggreviated by doing a complete case analysis which discards information from nearly complete records. Statistical imputation methods are nowadays available to exploit all available information.

- Predictors are omitted that should reasonably have been included based on subject matter knowledge. Modelers rely too much on the limited data that they have available in their data set, instead of wisely combining information from several sources, such as medical literature and experts in the field.
- Stepwise selection methods are abundant, which are especially risky in small data sets.
- Modelling approaches are used that require higher numbers. Data-hungry techniques, such as tree modelling and neural network modelling, should not be used in small data sets.
- No attempts are made towards validation, or validation is done inefficiently. For example, a split-sample approach is followed, leading to a smaller sample for model development and a smaller sample for model validation. Better methods are nowadays available and should be used far more often.

(c) Claims are exaggerated:

- Often we see statements such as 'the predictors were identified'; in many instances such findings may not be reproducible and may largely represent noise.
- Models are not internally valid, with overoptimistic expectations of model performance in new patients.
- One modern method with a fancy name is claimed as being superior to a more traditional regression approach, while no convincing evidence exists, and a suboptimal model strategy was followed for the regression model.
- Researchers are insufficiently aware of overfitting, implying that their apparent findings are merely coincidental ("the curse of dimensionality").

(d) Poor generalizability:

- If models are not internally valid, we cannot expect them to generalize.
- Models are developed for each local situation, discarding earlier findings on effects of predictors and earlier models; a framework for continuous improvement and updating of prediction models is required.

In this book; I try to suggest many small improvements in modelling strategies. Combined, these improvements hopefully lead to better prediction models.

Intended Audience

Readers should have a basic knowledge of biostatistics, especially regression analysis, but no strong background in mathematics is required. The number of formulas is deliberately kept small. Usually a bottom-up approach is followed in teaching regression analysis techniques, starting with model assumptions, estimation methods, and basic interpretation. This book is more top-down: given that we want to predict an outcome, how can we best utilize regression techniques?

Three levels of readers are envisioned:

(a) The core intended audience is formed by epidemiologists and applied biostatisticians who want to develop or apply a prediction model. Both students and professionals should find practical guidance in this book, especially by the proposed seven steps to develop a valid model (Part II).
(b) A second group is formed by clinicians, policy makers, and health care professionals who want to judge a study that presents a prediction model. This book should aid them in a critical appraisal, providing explanations of terms and concepts that are common in publications on prediction models. They should try to read chapters of particular interest, or read the main text of the chapters. They can skip the examples and more technical sections (indicated with*).
(c) A third group includes more theoretical researchers, such as (bio)statisticians and computer scientists, who want to improve the methods that we use in prediction models. They may find inspiration for further theoretical work and simulation studies in this book. Many of the methods in prediction modelling are not fully developed yet, and common sense underlies some of the proposed approaches in this book.

Other sources

Many excellent text books exist on regression analysis techniques, but these usually do not have a focus on modelling strategies for prediction. The main exception is Frank Harrell's book "Regression Modelling Strategies".[174] He brings advanced biostatistical concepts to practical application, supported by the Design and Hmisc libraries for S+ software (nowadays: packages for R). Harrell's book may however be too advanced for clinical and epidemiological researchers. This also holds for the Hastie, Tibshirani, and Friedman's quite thorough text book "The Elements of Statistical Learning".[181] These books are very useful for a more in-depth discussion of statistical techniques and strategies. Harrell's book provided the main inspiration for the presented work here. Another good companion book is the Vittinghoff et al. book on "Regression Methods in Biostatistics".[472]

Various sources at the internet can be used that explain terms used in this book. Frank Harrell has a glossary at his web site: [http://biostat.mc.vanderbilt.edu/twiki/pub/Main/ClinStat/glossary.pdf]. Other useful sources include [http://www.aiaccess.net/e_gm.htm] and Wikipedia.

Structure

It has been found that people learn by example, by checklists, and by own discovery. Therefore I provide many examples throughout the text, including the essential computer code and output. I also suggest a checklist for prediction modelling (Part II).

Own discovery is possible with exercises per chapter, with data sets provided at the book's web site: http://www.clinicalpredictionmodels.org.

Many statistical techniques and approaches are readily possible with any modern software package. Personally, I work with SPSS for simple, straightforward analyses, but this package is insufficient for more advanced analyses which are essential in prediction modelling. The SAS computer package is more advanced, but may not be so practical for some. A package such as Stata is very suitable. It may be similar in capabilities to S-plus software, which was my preferred program for advanced prediction modelling since a stay at Duke University in 1996. The R software is very similar in nature to S-plus, and has several additional advantages: the software is for free, and innovations in biostatistical methods become readily available for R. Therefore, R is the natural choice as the software accompanying this book. R software is available at http://www.cran.r-project.org, with help files and a tutorial.

Some R commands are provided in this book; full programs can be downloaded from a web site (http://www.clinicalpredictionmodels.org). This web site also provides a number of data sets that can be downloaded for application of the described techniques. I provide data files in SPSS format that can readily be imported in R and other packages. Further, comments on the text can be submitted electronically.

http://www.springer.com/978-0-387-77243-1

Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Steyerberg, E.
2009, XXVIII, 500 p., Hardcover
ISBN: 978-0-387-77243-1