Use R!

Jan Beyersmann
Martin Schumacher
Arthur Allignol

Competing Risks and Multistate Models with R

Springer

Use R!

For further volumes:
http://www.springer.com/series/6991

Jan Beyersmann • Arthur Allignol
Martin Schumacher

Competing Risks and Multistate Models with R

Springer

Jan Beyersmann
Institute of Medical Biometry
and Medical Informatics University
Medical Center Freiburg
Freiburg Center for Data Analysis
and Modelling University of Freiburg
D-79104 Freiburg, Germany

Arthur Allignol
Institute of Medical Biometry
and Medical Informatics University
Medical Center Freiburg
Freiburg Center for Data Analysis
and Modelling University of Freiburg
D-79104 Freiburg, Germany

Martin Schumacher
Institute of Medical Biometry
and Medical Informatics University
Medical Center Freiburg
D-79104 Freiburg, Germany

ISBN 978-1-4614-2034-7 e-ISBN 978-1-4614-2035-4
DOI 10.1007/978-1-4614-2035-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011941794

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is about applied statistical analysis of competing risks and multistate data.

Competing risks generalize standard survival analysis of a single, often composite or combined endpoint to investigating multiple first event types. A standard example from clinical oncology is progression-free survival, which is the time until death or disease progression, whatever occurs first. A usual survival analysis studies the length of progression-free survival only. A competing risks analysis would disentangle the composite endpoint by investigating the *time* of progression-free survival *and the event type*, either progression or death without prior progression. Competing risks are the simplest *multistate model*, where events are envisaged as transitions between states. For competing risks, there is one common initial state and as many target states as there are competing event types. Only transitions between the initial state and the competing risks states are considered.

A multistate model that is more complex than competing risks is the illness-death model. In the example of progression-free survival, this multistate model would also investigate death after progression. In principle, a multistate model consists of any finite number of states, and any transition between any pair of states can be considered.

This book explains the analysis of such data with R. In Part I, we first present the practical data examples. They come from studies conducted by medical colleagues where at least one of us has been personally involved in planning, analysis, or both. Secondly, we give a concise introduction to the basic concepts of hazard-based statistical models which is a unique feature of all modelling approaches considered. Part II gives a step-by-step description of a competing risks analysis. The single ingredients of such an analysis serve as key tools in Part III on more complex multistate models. Thus, our approach is in between applied texts, which treat competing risks or multistate models as 'further topics', and more theoretical accounts, which include competing risks as a simple multistate example. Our choice is motivated, firstly, by the outstanding practical importance of competing risks. Secondly, starting with competing risks allows for a technically less involved account, while at the same time providing many techniques that are useful for general multistate models.

The statistical concepts are turned into concrete R code. One reason for using R is that it provides for the richest practical toolbox to analyse both competing risks and multistate models. However, the practical implementation is explained in such a way that readers will be to able to, e.g., run Cox analyses of multistate data using other software, provided that the software allows for fitting a standard Cox model. Nonparametric estimation and model-based prediction of probabilities, however, are, to the best of our knowledge and at the time of writing, an exclusive asset of R.

The typical reader of the book is a person who wishes to analyse time-to-event data that are adequately described via competing risks or a multistate model. Such data are frequently encountered in fields such as epidemiology, clinical medicine, biology, demography, sociology, actuarial science, reliability, and econometrics. Most readers will have some experience with analysing survival data, although an account on investigating the time until a single, composite endpoint is included in the first two parts of the book. We do not assume that the reader is necessarily a trained statistician or a mathematician, and we have kept formal presentation to a minimum.

Likewise, we have refrained from giving mathematical proofs for the underlying theory. Instead, we encourage readers to use simulation in order to convince themselves within the R environment that the methodology at hand works. This *algorithmic perspective* is also used as an intuitive tool for understanding how competing risks and multistate data occur over the course of time.

Although refraining from a mathematically rigorous account, the presentation does have a *stochastic process flavor*. There are two reasons for this: firstly, it is the most natural way to describe multiple event types that happen over the course of time. Secondly, we hope that this is helpful for readers who wish to study more thoroughly the underlying theory as described in the books by Andersen et al. (1993) and Aalen et al. (2008).

How to read this book: The most obvious way is to start at the beginning. Chapter 1 presents the practical data examples used throughout the book. In Chapter 2, we recall why the analysis of standard survival data is based on *hazards*, and we then explain why the concept of a hazard is amenable to analysing more complex competing risks and multistate data. A further consequence is that the data may be subject to both the common *right-censoring*, where only a lower bound of an individual's event time may be observed, and *left-truncation*, where individuals enter the study after time origin. Such a delayed study entry happens, e.g., in studies where age is the time scale of interest, but individuals enter the study only after birth. The practical implications of Chapter 2 for competing risks are considered in Part II. Part III is on multistate models and frequently makes use of the competing risks toolbox.

Readers who urgently need to analyse competing risks data may proceed to the competing risks part of the book right away. They should at least skim over the description of competing risks as a multistate model in Chapter 3. The common nonparametric estimation techniques are in Chapter 4, and Cox-type regression modelling of the *cause-specific hazards* is explained in Section 5.2. These readers are, however, encouraged to read Chapter 2 later in order to understand why the techniques at hand work. In our experience, a practical competing risks analysis often raises questions such as whether the competing risks are independent or whether and when a competing risk can be treated as a censoring. Some of these issues are collected in Section 7.2. The theory outlined in Chapter 2 is necessary to clarify these issues.

Readers who wish to analyse multistate data in practice should have a clear understanding of competing risks from a multistate model point of view and as explained in detail in Part II. As stated above, this is so, because Part III frequently draws on competing risks methodology. The connection is that we are going to consider multistate models that are realized as a *nested sequence of competing risks experiments*; see Chapter 8.

This book is also suitable for *graduate courses* in biostatistics, statistics, and epidemiological methods. We have taught graduate courses in biostatistics using the present material.

The *R packages* and the *data* used in this book can be downloaded from the Comprehensive R Archive Network

http://cran.r-project.org/

The book is also accompanied by web pages, which can be found at

www.imbi.uni-freiburg.de/comprisksmultistate

The web pages provide the complete R code used to produce the analyses of this book as well as solutions to the Exercises. Sweave (Leisch, 2002) has been used to generate the LaTeX files of this book and to extract its R code. We also hope that readers will visit the web pages and leave us a message if they find any mistakes or inconsistencies.

We thank our medical colleagues who have granted us permission to use the data of their studies and to publish the data as part of R packages. This book has profited from collaborative work and/or comments from Adrian Barnett, Ronald Geskus, Nadine Grambauer, Stefanie Hieke, Aurélien Latouche, Reinhard Meister, Hein Putter and Christine Porzelius. We thank them all. Parts of this book have been written while the authors were supported by grant FOR 534 'Statistical modeling and data analysis in clinical epidemiology' from the Deutsche Forschungsgemeinschaft. This is gratefully acknowledged.

Freiburg,

Jan Beyersmann
Arthur Allignol
Martin Schumacher

Contents

Part III Multistate models

Part I

Data examples and some mathematical background

1

Data examples

In this book, we use both real and simulated data. One idea behind using simulated data is to illustrate that competing risks and multistate data can be conveniently approached from an algorithmic perspective. The data simulations are explained in their respective places in the book. In this section, we briefly introduce the real data examples. All of them are publicly available as part of the R packages used in this book.

Pneumonia on admission to intensive care unit, data set `sir.adm`

The data set is part of the `mvna` package. It contains a random subsample of 747 patients from the SIR 3 (*S*pread of nosocomial *I*nfections and *R*esistant pathogens) cohort study at the Charité university hospital in Berlin, Germany, with prospective assessment of data to examine the effect of hospital-acquired infections in intensive care (Wolkewitz et al., 2008). The data set contains information on pneumonia status on admission, time of intensive care unit stay and 'intensive care unit outcome', either hospital death or alive discharge. Pneumonia is a severe infection, suspected to both require additional care (i.e., prolonged intensive care unit stay) and to increase mortality.

The entry `sir.adm$pneu` is 1 for patients with pneumonia present on admission, and 0 for no pneumonia. A patient's status at the end of the observation period is contained in `sir.adm$status`, 1 for discharge (alive) and 2 for death. `sir.adm$status` is 0 for patients still in the unit when the data base was closed. These patients are called (right-) censored. A patient's length of stay is in `sir.adm$time`.

There were 97 patients with pneumonia on admission. Overall, 657 patients were discharged alive, 76 patients died, and 14 patients were still in the unit at the end of the study. 21 of the patients who died had pneumonia on admission.

The data set `sir.adm` is a competing risks example; that is, we investigate the time until end of stay *and* the discharge status, either alive discharge or hospital death. A challenge in the analysis of this data set is that pneumonia is found to increase the probability of dying in hospital, but appears to have no

J. Beyersmann et al., *Competing Risks and Multistate Models with R*, Use R,
DOI 10.1007/978-1-4614-2035-4_1, © Springer Science+Business Media, LLC 2012