Statistical Methods
in Cancer Research
Volume IV
Descriptive Epidemiology
International Agency for Research on Cancer

The International Agency for Research on Cancer (IARC) was established in 1965 by the World Health Assembly, as an independently financed organization within the framework of the World Health Organization. The headquarters of the Agency are at Lyon, France.

The Agency conducts a programme of research concentrating particularly on the epidemiology of cancer and the study of potential carcinogens in the human environment. Its field studies are supplemented by biological and chemical research carried out in the Agency's laboratories in Lyon and, through collaborative research agreements, in national research institutions in many countries. The Agency also conducts a programme for the education and training of personnel for cancer research.

The publications of the Agency are intended to contribute to the dissemination of authoritative information on different aspects of cancer research. A complete list is printed at the back of this book.

Cover illustration: Victor Vasarely, HAT-1V, 1971
Copyright: © ADAGP, Paris, 1994

Authors

Jacques Esteve
Adviser on Biostatistics, International Agency for Research on Cancer

Ellen Benhamou
Epidemiologist, Gustave Roussy Institute, Paris, France

Luc Raymond
Epidemiologist, Geneva Cancer Registry and Department of Social and Preventive Medicine, University of Geneva, Switzerland

Translator

Mary Sinclair
Scientific Editor, Alpha Biomedical Communications, Sydney, Australia
Contents

Preface ............................................................ IX
Foreword .......................................................... XI
Acknowledgements ................................................ XIII

1. Fundamental concepts ........................................... 1

Introduction ........................................................... 1
Basic concepts of descriptive epidemiology ........................................ 3
Time and the concept of incidence ............................................. 3
Group characteristics and place .............................................. 6
Statistical concepts for the analysis of incidence data .......................... 11
Formal definition of the incidence rate ........................................ 11
Estimation of the instantaneous incidence rate ................................ 14
An approximation useful in descriptive epidemiology ........................ 14
When individual observations are available and complete .................... 15
When individual observations are available but possibly censored .......... 18
Statistical concepts in survival analysis ...................................... 21
Follow-up studies ...................................................... 21
Survival probability ................................................... 23
The life table .......................................................... 26
Classical models for survival distribution ...................................... 28
Interactive risks ......................................................... 34
Competing risks ......................................................... 34
Relationship between incidence, mortality, survival and prevalence ....... 37

Bibliographical notes .................................................. 44
References .......................................................... 45

2. Techniques for the analysis of cancer risk .......................... 49

Measurement of the risk of cancer .......................................... 49
Age- and sex-specific rates ............................................... 49
The calculation of a specific rate .......................................... 49
The accuracy of the estimate of a rate ..................................... 51
The incidence curve ..................................................... 54
Standardized rates ....................................................... 56
Direct standardization ..................................................... 56
Cumulative rates ......................................................... 60
Indirect standardization .................................................... 62
Probability of developing a specific form of cancer ........................ 66
The number of years of life lost ........................................... 69
Years of life lost with respect to a fixed age limit .......................... 70
Years of life lost with respect to life expectancy .......................... 71
Methods for comparison ............................................. 73
Comparison of incidence of a disease in two groups .......... 74
   The approximate method ...................................... 74
Mantel-Haenszel test ............................................ 77
Overall measure of incidence ratio ............................ 79
Test of a multiplicative model .................................. 81
Trend test ......................................................... 82
Example ............................................................ 83
Comparison of incidence among several populations ....... 85
   Comparison with an overall expected value ............... 85
Homogeneity test for incidence ................................ 87
Use of the log-linear model ..................................... 90
Extension and limitations of the present methodology ...... 95
   Risk analyses in the absence of denominators ............ 95
   Standardized indices of relative frequency ............... 96
   Modelling incidence data in the absence of the denominator .... 96
Choosing between various risk measures ....................... 98
Bibliographical notes ............................................ 103
References .......................................................... 104

3. Space-time variations and group correlations .............. 107

Geographical analysis ............................................ 107
   The objectives of cartography ................................ 107
Methods .......................................................... 111
   Geographical division ........................................ 111
   Choice of a risk indicator ................................... 114
   Definition of risk classes ................................... 114
Tools to interpret geographical data ........................... 119
   Autocorrelations ............................................. 119
   Identifying risk clusters .................................... 122
   Time-space clustering ........................................ 131
   Smoothing and the empirical Bayes method ............... 134
   Concluding remark .......................................... 140
Ecological studies ................................................. 141
   Aim and methodological principles ......................... 141
   Strengths and limitations of a measure of group exposure ... 148
      Group versus individual exposure ......................... 148
      Risk estimation in the context of an ecological study .... 149
      The ecological fallacy .................................... 150
Specific techniques and examples .............................. 154
   Definition of groups ........................................ 154
   Multivariate analysis ....................................... 158
   Migrant studies ............................................ 166
Time trends ........................................................ 170
   Objectives ..................................................... 170
   Methods ......................................................... 172
Contents

Components of temporal evolution ............................................................... 172
Effect of data quality ....................................................................................... 176
Role of modelling ............................................................................................. 178
Description of trend by period .......................................................................... 179
Description of trend by cohort .......................................................................... 189
Age-period-cohort models ................................................................................ 194
Specific techniques and examples .................................................................... 198
   Epidemiological evaluation of a secondary prevention campaign ................. 198
   Trends in cancer of the uterine cervix .......................................................... 202
Bibliographical notes ......................................................................................... 204
References ........................................................................................................ 206

4. Techniques for survival analysis ..................................................................... 214

Survival analysis in descriptive epidemiology .................................................. 214
Estimation of survival distribution ..................................................................... 216
   Estimation of crude survival ......................................................................... 216
      Actuarial method ....................................................................................... 216
      Kaplan-Meier method .............................................................................. 219
   Confidence interval for a survival rate .......................................................... 222
   Median survival time ..................................................................................... 225
   Collecting data for survival analysis ............................................................. 227
Estimation of net survival ................................................................................... 229
   Cause-specific survival ................................................................................ 230
   Relative survival .......................................................................................... 231
   Constructing a life table .............................................................................. 236
   Alternative methods ..................................................................................... 237
Methods of comparison ....................................................................................... 245
   Introduction .................................................................................................. 245
   Comparison of crude survival probabilities .................................................. 246
      Test to compare two survival rates ............................................................. 246
      Rank tests to compare survival curves ....................................................... 247
   Stratified comparison of crude survival ......................................................... 255
   The Cox model ............................................................................................. 260
      The proportional hazard model ................................................................. 260
      Principle of the Cox model estimation ...................................................... 263
      The likelihood ratio test ........................................................................... 266
      Example of an adjustment using Cox's model ........................................ 268
   Comparison of net survival ......................................................................... 272
Bibliographical notes ......................................................................................... 276
References ......................................................................................................... 277

Appendix 1. Life table for Switzerland; 1978-1983 .......................................... 281
Appendix 2. Using GLIM in descriptive epidemiology ..................................... 284
As the authors of this book remind us in their introduction, classical descriptive epidemiology was long regarded as simply a first, rather crude, step in the exploration of an epidemiological problem. Based essentially on comparisons between populations, it could do no more than stimulate ideas and hypotheses. Instead, it was up to analytical epidemiology, a more precise science since it involves measurements at the individual level, to produce firm evidence on risk factors, and it is mainly on the methodology of this area that various books have focused over the last thirty years.

Nevertheless, I can recall a number of major successes of descriptive epidemiology. For example, simply mapping the distribution of mortality rates for oesophageal cancer and of alcoholic cirrhosis enabled us, with Daniel Schwartz and Odile Lasserre, to demonstrate a relationship between alcohol consumption and oesophageal cancer. Likewise, comparison between the rising curve of lung cancer mortality and that of cigarette consumption certainly played a decisive role in focusing the attention of Richard Doll on the link between tobacco and cancer.

In recent years, the establishment of numerous cancer registries has encouraged many researchers to attempt to draw the maximum advantage from the data collected on cancer incidence and mortality. This has led to many original ideas and raised many important questions. Some biostatisticians entered this arena and they have gradually been able to lay the necessary statistical and mathematical foundations that were lacking. Problems such as those posed by the study of risks when the denominator is unknown, competing risks, and autocorrelation have led to the development of solid methodological concepts.

The great merit of this volume is that it brings together and reviews in one coherent text the different techniques needed for a modern approach to descriptive epidemiology. With the help of this compilation, researchers in this field will henceforth be able to tackle the study of their data armed with a methodological arsenal giving them the optimal chance of success. Other readers such as doctors and public health specialists will be able to obtain guidance, without having to enter into all the mathematical details, on how to avoid the many pitfalls that confront those who have to interpret collections of numerical data. The authors make extensive use of examples of analysis of real data sets and show how these can be treated and interpreted, so that the reader can follow in detail the development of the methods described and better comprehend the range of their applications.

I have known Jacques Estève, Ellen Benhamou and Luc Raymond for very many years, as a productive team with complementary capabilities. I am certain that this book that they have co-authored will provide an indispensable guide for numerous researchers and for decision-makers in public health who are concerned with epidemiology. Personally, as an epidemiologist and oncologist, I have found it to be of the greatest interest.

Professor Robert Flamant

*Director of the Gustave Roussy Institute,*

*Villejuif, France*
Foreword

This monograph presents and discusses some methods used in descriptive epidemiology which are relevant to cancer research. In presenting the fundamental concepts, we have tried to keep the mathematical formulation at a level which is compatible with an elementary knowledge of statistics and probability, but which nevertheless enables the logical relationships between the concepts currently used in epidemiology to be understood.

With the above objective in mind, Chapter 1 describes briefly the epidemiological context in which the methods will be used and devotes some space to their mathematical formulation. An elementary knowledge of statistics and probability as well as some familiarity with mathematical reasoning is expected from the reader of this chapter.

Chapter 2 describes how, in practice, the analysis and comparison of incidence and mortality can be carried out. Most attention is given to the multiplicative model and to the concept of proportional hazards, which is particularly relevant to cancer research. The exposition of these notions relies on many numerical examples, but no great mathematical sophistication is needed.

Chapter 3 is devoted to geographical analysis, ecological studies and analysis of time trends. These fields are at present subjects of interesting methodological research, and we have tried to show from several examples how modern statistical tools can considerably improve the interpretation of geographical and temporal data in epidemiology.

Chapter 4 describes the methods of analysis of survival probability at an elementary mathematical level, and the emphasis is placed on the interpretation of such data when they are collected in the context of routine cancer registry operations. Much space is therefore given to the concept of relative survival and many examples are presented to show the difficulty of interpretation when the procedures for data collection may imply several types of bias.

We have tried to give the reader sufficient understanding to use the methods which are presented by giving the details of calculations whenever possible and some examples of the use of the GLIM software, which is cheap, widely available and enables many methods presented in this book to be readily implemented.

This text was first written in French and was translated by Mary Sinclair, whom we gratefully acknowledge for her careful work. With the exception of the correction of known errors and some inevitable adaptation of French to English style, no effort has been made to update the content which was essentially written before 1991; this is why some recent references which would have been relevant are not included.

The finalization of the manuscript of this monograph benefited from the careful reading of John Cheney. We gratefully acknowledge his help.
Acknowledgements

This book was originally conceived as the proceedings of two seminars on statistical methods organized by the group of cancer registries in Latin countries. However, the absence in the French scientific literature of an up-to-date text on methods in descriptive epidemiology led us to write a manual on the key concepts and statistical methods in this area.

We hope that this manual will be of value to those working in, or with cancer registries, for whom it was originally intended. We trust that the mathematical nature of some parts of the book will not be an obstacle to its use and that the inclusion of numerous examples will serve to establish the link between the theoretical development and its epidemiological interpretation.

Many people have contributed to the development of this book: most notably the participants at the seminars mentioned above, whom we gratefully acknowledge. We are especially indebted to those who have made unpublished data available to illustrate the procedures presented in this book.

We are especially grateful to Catherine Com-Nougue and Agnes Laplanche who carefully read the manuscript and assisted us with pertinent advice and suggestions; to Annick Rivoire for her invaluable contribution in the preparation and production of the book; and to Aude Jaccard for organizing the bibliography. Finally, we thank Nicholas Day for his encouragement and help when initiating this project.

St Just et Vacquières, April 1992