

**STATISTICAL METHODS
IN
CANCER RESEARCH**

**Volume II—THE DESIGN AND
ANALYSIS OF COHORT
STUDIES**

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

WORLD HEALTH ORGANIZATION



INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

STATISTICAL METHODS
IN
CANCER RESEARCH

VOLUME II – THE DESIGN AND ANALYSIS
OF COHORT STUDIES

BY
N.E. BRESLOW & N.E. DAY

TECHNICAL EDITOR FOR IARC
E. HESELTINE

IARC Scientific Publications No. 82

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER
LYON

1987

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.

Distributed for the International Agency for Research on Cancer
by Oxford University Press, Walton Street, Oxford OX2 6DP, UK

London New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland

Oxford is a trade mark of Oxford University Press

Distributed in the United States
by Oxford University Press, New York

ISBN 92 832 1182 0
ISSN 0300-5085

© International Agency for Research on Cancer 1987
150 cours Albert Thomas, 69372 Lyon Cedex 08, France

The authors alone are responsible for the views expressed in this publication. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of the International Agency for Research on Cancer
Printed in the UK

CONTENTS

Foreword	v
Preface	vii
Acknowledgements	ix
List of Participants at IARC Workshop 25–27 May 1983	xi
Chapter 1. The Role of Cohort Studies in Cancer Epidemiology	2
Chapter 2. Rates and Rate Standardization	48
Chapter 3. Comparisons among Exposure Groups	82
Chapter 4. Fitting Models to Grouped Data	120
Chapter 5. Fitting Models to Continuous Data	178
Chapter 6. Modelling the Relationship between Risk, Dose and Time	232
Chapter 7. Design Considerations	272
References	316
Appendices	
I. Design and conduct of studies cited in the text	
IA. The British doctors study	336
IB. The atomic bomb survivors – the life-span study	340
IC. Hepatitis B and liver cancer	345
ID. Cancer in nickel workers – the South Wales cohort	347
IE. The Montana study of smelter workers	349
IF. Asbestos exposure and cigarette smoking	352
II. Correspondence between different revisions of the International Classification of Diseases (ICD)	355
III. U.S. national death rates: white males (deaths/person-year × 1000)	358
IV. Algorithm for exact calculation of person-years	362
V. Grouped data from the Montana smelter workers study used in Chapters 2–4.	363
VI. Nasal sinus cancer mortality in Welsh nickel refinery workers: summary data for three-way classification	367
VII. Lung and nasal sinus cancer mortality in Welsh nickel refinery workers: summary data for four-way classification	369

VIII. Continuous data (original records) for 679 Welsh nickel refinery workers	374
IX. England and Wales: age- and year-specific death rates from nasal sinus and lung cancer and from all causes	391
Combined Index to Volumes 1 and 2 of <i>Statistical Methods in Cancer Research</i>	396

FOREWORD

Epidemiological studies provide the only definitive information on the degree of cancer risk to man. Since malignant diseases are clearly of multifactorial origin, their investigation in man has become increasingly complex, and epidemiological and statistical studies on cancer require a correspondingly complex and rigorous methodology.

The past 15 years have seen rapid developments of the analytic tools available to epidemiologists. These advances now permit a more flexible and quantitative approach to the use of epidemiological data, and thus greatly enhance the utility of such data for the primary purpose of disease prevention. For society now expects that if preventive measures are to be introduced, then quantitative assessments of the expected benefit should be available. The first volume in this series focused on case-control studies, reflecting the concentration on this approach in the 1970s for the identification of cancer hazards. Attention has recently turned to the more basic line of attack provided by cohort studies, and the more general modelling of risk that can ensue. This second volume gives an authoritative account of the methods now available for the interpretation of the results from this type of study.

The two volumes together give a comprehensive development of the principles and concepts underlying the design and analysis of both types of study currently used in analytic cancer epidemiology, and a detailed treatment of the quantitative methods now available. The IARC hopes that this text will be of value to the epidemiological and statistical community for many years to come.

L. Tomatis, MD
Director
International Agency
for Research on Cancer

PREFACE

Long-term follow-up (cohort) studies of human populations, particularly of industrial workers, of patients treated with radiation and cytotoxic chemotherapy, and of victims of nuclear and other disasters, have provided the most convincing evidence of the link between exposure to specific environmental agents and cancer occurrence. Of the chemicals and industrial processes for which working groups convened by the IARC have decided that there is 'sufficient evidence' of human carcinogenicity, cohort studies provided the definitive evidence in the great majority of cases. In the studies carried out in the 1950s and 1960s, high risks were associated with specific exposures. Relatively simple statistical methods were sufficient to demonstrate the effect, and the finer quantitative features of the relationship were not emphasized. It was not uncommon for reports of occupational hazards to be based primarily on the computation of standardized death rates or mortality ratios (SMRs) for a few causes of death, with virtually no attention paid to internal comparisons among differentially exposed workers. Since then, the picture has changed. More attention is now paid to the quantification of risk and the use of more refined dose-response models. Interest has also turned to a wider range of exposures and the interplay between physiological measures of nutritional status, dietary factors and other variables of modes of life. Multivariate methods are then necessary, often making use of serial measurements on the same individuals.

Increasingly, modern concepts of statistical inference and modelling are being used to maximize the information obtainable from these major endeavours and to provide the most precise estimates possible of quantitative risk. Indeed, some cohort studies have stimulated the development of new statistical methods of particular relevance to this field.

The primary purpose of this monograph is to bring together in one place the statistical developments that have taken place during the past few years that are of relevance to the design and analysis of cohort studies, and to illustrate their application to several sets of data of importance in the field of cancer epidemiology. We hope to present these new statistical methods in such a way that epidemiologists and other research workers without extensive statistical training can appreciate the possibilities they offer and, in many cases, can apply them to their own work. In addition, by providing a thorough introduction to the design and execution of cohort studies, including a detailed description of six landmark investigations of this type, we hope to interest students of statistical science in this field so that they may turn their attention

both to the proper application of current methods and to the further development of those methods.

In the preface to the first volume in this series we stressed the essential similarity of statistical methods applicable to the case-control and cohort approaches to epidemiological research, the flexibility of new methods for handling a variety of data configurations and the wide range of problems that could be approached from a common conceptual foundation. This pursuit of unity and flexibility continues to be our goal. We show how elementary methods that have long been used for analysis of cohort data relate to explicit statistical models, and how they may be extended so as to achieve greater understanding of the collected data. The SMR, for example, has been used virtually without change for over 200 years to make age-adjusted comparisons of regional and occupational mortality. We show how this statistic may be derived as a maximum likelihood estimate in a well-defined statistical model, and how an extension of that model leads to a regression analysis of the SMR as a function of one or more risk factors. This approach shows us that the well-known 'lack of comparability' of SMRs is due to the problem of statistical confounding and may be alleviated by a proper analysis. Further extensions of the basic model permit variations in the SMR to be estimated as a nonparametric function of time for purposes of exploratory analyses of data.

Experience with the first volume taught us that one of its most important features, made possible through the generosity of our collaborators, was the provision of appendices containing several condensed, but nonetheless bona-fide, sets of data. These were used in worked examples that readers could follow to test their understanding of the material (and, occasionally, to find our mistakes). The present volume contains appendices that give grouped data from a study of respiratory cancer among smelter workers in Montana, USA, and both grouped and individual data records on 679 Welsh nickel refiners who had high rates of lung and nasal sinus cancer. Summary data from several other studies that appear in tables scattered throughout the monograph may also be useful for this purpose.

A major source of dissatisfaction with the first volume was its lack of a subject index. We have attempted to remedy the situation by including a combined index to both volumes.

N.E. Breslow and N.E. Day

ACKNOWLEDGEMENTS

Planning of this volume on cohort studies began shortly after the appearance of the first volume on case-control studies in 1980. Since then, many people have contributed to its development. Thirteen epidemiologists and statisticians participated in an IARC workshop on the statistical aspects of cohort studies that was held in Lyon on 23–27 May 1983 (see List of Participants). Initial drafts of several chapters were circulated and reviewed during that meeting, and the discussion was valuable for orientating subsequent developments. As those chapters were completed, they were sent to selected individuals for further comment. Persons who generously contributed their time in this regard include E. Bjelke, D. Clayton, T. Fletcher, E. Johnson, J. Kaldor, E. Läärä and P. Smith. We appreciate the significant efforts of these reviewers.

Data from two cohort studies are listed in the appendices and are utilized throughout the monograph in illustrative analyses that demonstrate the relationships between various statistical methods. We are indebted to Professor Sir Richard Doll and Professor J. Peto for permission to reproduce a working version of the recently updated data on Welsh nickel refiners in Appendices VI, VII and VIII. Likewise, we appreciate the generosity of Dr J. Fraumeni, Dr A. Lee-Feldstein and Dr J. Lubin in providing access to the latest follow-up data from their study of Montana smelter workers, portions of which are reproduced in Appendix V. We believe that the availability of these data sets to readers who wish to verify our results, or who wish to test their own ideas for statistical analysis on the basis of bona-fide and well-documented sets of epidemiological data, is extremely important in achieving the goals towards which the monograph is directed.

Several people assisted with the computer programming, data management and statistical analyses required for the illustrative examples, tables and figures. NEB would like to thank particularly Dr B. Langholz, who contributed to this effort over a period of several years, Mr P. Marek for computer programming and Mr J. Cologne who assisted with many of the final preparations. NED would like to acknowledge Ms D. Magnin and Dr J. Kaldor.

Primary secretarial support for this project was provided by Jean Hawkins who was responsible for the typing of innumerable drafts and the transfer of material among several word-processing systems. She also provided valuable assistance with editing, reference checking, and a myriad of necessary details. We should like to thank also Mrs A. Rivoire, Mrs E. Nasco and Mrs M. Kaad for their contributions. The figures

were carefully prepared by Mr Jacques Déchaux. We thank Mrs E. Heseltine and her staff for editing and shepherding the manuscript through the final stages of publication.

This project would not have been possible without the generous financial support of the US National Cancer Institute. During the initial years of preparation, NEB held a Preventive Oncology Academic Award, and in later years a research grant awarded by the National Cancer Institute. First drafts of several chapters were written during the 1982–1983 academic year while he was on sabbatical leave from the University of Washington at the German Cancer Research Center in Heidelberg. He would like to thank Dr H. Neurath and Dr G. Wagner, as well as the Alexander von Humboldt Foundation, for arranging this visit and his colleagues in Seattle, particularly Dr V. Farewell and Dr P. Feigl, for continuation of work in progress during his absence.

**LIST OF PARTICIPANTS AT IARC WORKSHOP
25-27 May 1983**

Professor E. Bjelke
Institute of Hygiene and Social Medicine
University of Bergen
5016 Haukeland Sykehus, Norway

Professor N.E. Breslow
Department of Biostatistics, SC-32
University of Washington
Seattle, WA 98195, USA

Dr T. Hirayama
Chief, Epidemiology Division
National Cancer Center Research Institute
Tokyo, Japan

Dr B. Langholz
German Cancer Research Center
Im Neuenheimer Feld 280
6900 Heidelberg 1, Federal Republic of Germany

Dr O. Møller Jensen
Director, Danish Cancer Registry
2100 Copenhagen Ø, Denmark

Professor J. Peto
Division of Epidemiology
Institute of Cancer Research
Sutton, Surrey, UK

Dr P.G. Smith
Department of Medical Statistics
London School of Hygiene and Tropical Medicine
London WC1E 7HT, UK

Dr D.C. Thomas
Department of Family and Preventive Medicine
University of Southern California
Los Angeles, CA 90033, USA

Dr A. Whittemore
Department of Family, Community and Preventive Medicine
Stanford University School of Medicine
Stanford, CA 94305, USA

IARC participants:

Dr N.E. Day
Dr J. Estève
Dr J. Wahrendorf
Dr A.M. Walker