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

IARC SCIENTIFIC PUBLICATIONS No. 82

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER
LYON 1987

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

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

CONTENTS

1. THE ROLE OF COHORT STUDIES IN CANCER EPIDEMIOLOGY

1.1 Historical role

1.2 Present significance and specific strengths of cohort studies

1.3 Limitations of cohort studies

1.4 Implementation

1.5 Interpretation

1.6 Proportional mortality studies

CHAPTER 1

THE ROLE OF COHORT STUDIES IN CANCER EPIDEMIOLOGY

Longitudinal studies are of fundamental importance in human biology. In the study of physical growth, of mental and hormonal development, and in the process of ageing, the longitudinal approach has played a central role. The essential feature of such investigation is that changes over time are followed at the individual level. Most chronic diseases are the result of a process extending over decades, and many of the events occurring in this period play a substantial role. The longitudinal surveillance and recording of these events is therefore a natural model of study to obtain a complete picture of disease causation. Fortunately, for the study of a large number of chronic diseases, most of the relevant information on exposure can be summarized in a few relatively simple measures, so that continuous monitoring is not required. But the regular assessment of exposure variables may well be necessary, and in the epidemiology of cardiovascular disease, with its emphasis on physiological and biochemical explanatory measures, this approach has been the one of choice.

The essence of longitudinal studies in epidemiology is the identification of a group of individuals about whom certain exposure information is collected; the group is then followed forward in time to ascertain the occurrence of the diseases of interest, so that for each individual prior exposure information can be related to subsequent disease experience. Since the first requirement of such studies is the identification of the individuals forming the study group – or cohort – longitudinal studies in cancer epidemiology are usually referred to as *cohort* studies. (This use of the word 'cohort' first appeared in the literature in a demographic setting in 1944, according to the Oxford English Dictionary. It had apparently been introduced informally in 1935, as described by Wall & William, 1970.)

There are two ways in which the follow-up over time may be conducted. First, one may assemble the cohort in the present, and follow the individuals prospectively into the future. This type of study is often referred to as a *prospective cohort* study. It has the advantage that one may collect exactly the information thought to be required, and the disadvantage that many years may elapse before sufficient cases of disease have developed for analysis.

Second, one may identify a group with certain exposure characteristics, by means of historical records, at a certain defined time in the past, and then reconstruct the disease experience of the group between the defined time in the past and the present. This type of study has been called a *historical cohort* study. The advantage is that results are

potentially available immediately; the disadvantage is that the information available on the cohort may not be completely satisfactory, since it would almost certainly have been collected for other purposes. Much may be missing, and it may not correspond closely to the question of interest. The term 'retrospective cohort study' is also commonly used, but is slightly misleading, since the essential viewpoint in most such studies is forward in time, although starting in the past. The term 'historical cohort study' is preferable logically. In both types of study, the individuals comprising the cohort are identified, and information on their exposure obtained, before their disease experience is ascertained.

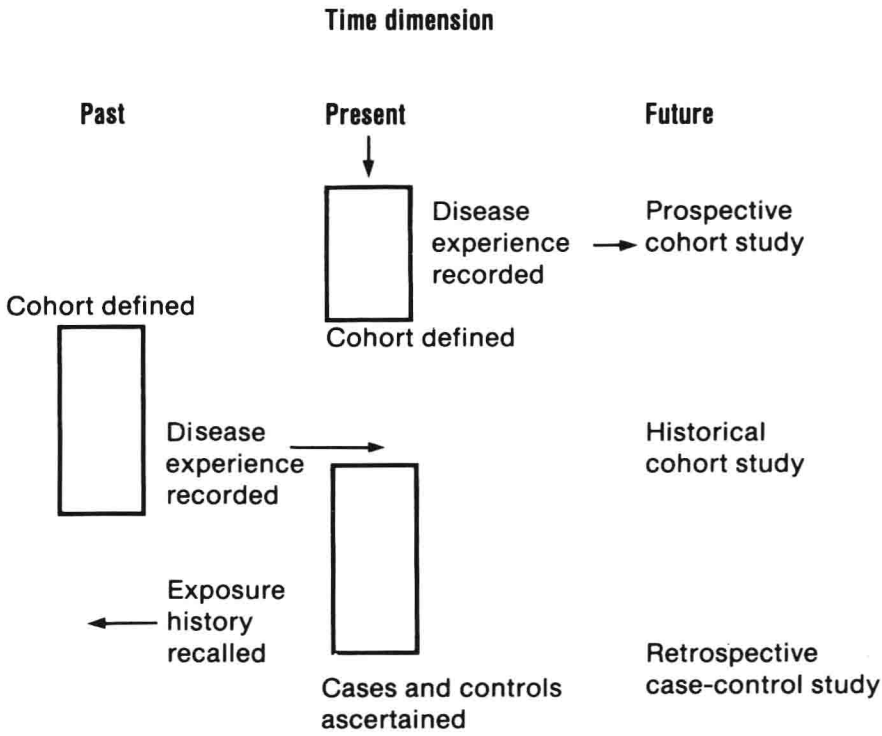
Cohort studies, by recording disease occurrence in a defined group, provide measures of incidence, or mortality rates, and it is these rates that provide the basic measures of disease risk. By allowing one to measure the basic risk associated with different levels and types of exposure, cohort studies provide the foundation of cancer epidemiology. It so happens, however, that a frequently convenient way of expressing the excess risk in one group compared to another is in terms of the ratio of the rates in the two groups, and to estimate the ratio of the rates one can use just a sample of the overall cohort. Since it is often easier and cheaper to obtain information on a sample rather than on the entire cohort, the case-control study has become widely adopted in cancer epidemiology as an alternative to the cohort study.

In fact, as commonly used, the case-control approach departs more radically from a cohort study than simply by sampling. In many case-control studies, the individuals with the disease in question and some comparison group are ascertained first, and their exposure experiences for some defined period of time in the past obtained retrospectively. The results are used to derive rate ratios. A cohort study faces forwards in time, starting with the defined population and its exposure status, and observing the subsequent disease experience, whereas a retrospective case-control study faces backwards in time, starting with the disease status and reconstructing the exposure history from which it emerged. Graphically, the distinction can be expressed as shown in Figure 1.1

Notwithstanding these differences, however, the rate ratios estimated in a case-control study should refer to rates in some defined population. As argued in Volume 1 of this series, the inferences one draws from the results of a case-control study depend logically on the interpretation one can give to it as having arisen by sampling from some underlying cohort. The less clear the definition of the underlying population, the less confidence can be put in the results of the case-control study. Thus, although the case-control and cohort approaches appear clearly distinct, they share the same logical framework of inference. An increasing number of studies have components of both approaches in their design. In these hybrid designs, the cohort component would usually identify the group and ascertain the disease experiences in the follow-up period; the exposure experience would then be obtained using the case-control approach. In this way, one ensures strict definition of the study cohort, but the effort and resources devoted to obtaining accurate exposure data can be concentrated on the most informative individuals. We discuss later at some length (§1.4*i*) the interplay between the cohort and the case-control approach.

Common to both cohort and case-control studies is the extended period of

Fig. 1.1 Differences between cohort and case-control studies



observation, relating to disease experience in the former and to exposure experience in the latter, and sometimes both in either case, and the fact that the individual is the unit of observation. These two features contrast with those of studies in which populations are compared by using cross-sectional data on both exposure and disease occurrence – so-called 'population correlation' or 'ecological' studies. This type of study would normally be given little weight in assessing the basic causality of a relationship, and, in the series of *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* a prerequisite for evidence to be deemed sufficient to establish carcinogenicity in humans is that it derive from individual-based studies. Correlation studies may be useful in suggesting interesting areas of study, that is, for hypothesis generation. The distinctions, however, are not absolute. Population comparisons may be made on the basis of temporal changes or of the experience with respect to exposure and disease of different birth cohorts, rather than among populations defined geographically, and such comparisons are often given greater weight. A cohort study, on the other hand, may include little or no information on variations in exposure between individuals, it being known simply that the cohort as a whole was exposed – for example, had received Bacillus Calmette–Guerin (BCG) vaccination in the first year of life.

1.1 Historical role

In 1954, two papers were published that are landmarks in the historical development of cancer epidemiology. The first, called a 'preliminary report', described the rationale for, and the first results of, the prospective cohort study of British doctors (Doll & Hill, 1954), designed to investigate the relatonship of tobacco smoking to lung cancer. The second, a historical cohort study, reported on the risk of bladder cancer in the British chemical industry (Case *et al.*, 1954; Case & Pearson, 1954).

The prospective study of British doctors was initiated in 1951, when the results of a number of case-control studies had already been published demonstrating an association between lung cancer and cigarette smoking. (The design and execution of the study are described in detail in Appendix IA.) It is interesting to examine why, in view of the results of the case-control studies, a large scale, long-term study was felt necessary. The 1954 paper by Doll and Hill starts as follows:

'In the last five years a number of studies have been made of the smoking habits of patients with and without lung cancer. All these studies agree in showing that there are more heavy smokers and fewer nonsmokers among patients with lung cancer than among patients with other diseases. While, therefore, the various authors have all shown that there is an "association" between lung cancer and the amount of tobacco smoked, they have differed in their interpretation. Some have considered that the only reasonable explanation is that smoking is a factor in the production of the disease; others have not been prepared to deduce causation and have left the association unexplained.

'Further retrospective studies of that same kind would seem to us unlikely to advance our knowledge materially or to throw any new light upon the nature of the association. If, too, there were any undetected flaw in the evidence that such studies have produced, it would be exposed only by some entirely new approach. That approach we considered should be "prospective". It should determine the frequency with which the disease appeared, in the future, among groups of persons whose smoking habits were already known.'

In this initial report on the British doctors study, the authors stressed that the results of the prospective study were in close agreement (Table 1.1) with the results of their earlier case-control study (Doll & Hill, 1950), in terms of the ratios of the rates in the different smoking categories. The absolute level of the rates, however, appeared to be more than twice as high in the case-control study (confined to the subset of the study consisting of residents of Greater London) than in the cohort of doctors. It should be noted that the results of the case-control study were converted into absolute incidence rates for lung cancer and were not limited to a description of the effect of smoking in terms of the ratios of rates in the different smoking categories.

The results of 20 years or more of follow-up have been published in some detail (Doll & Peto, 1976, 1978; Doll *et al.*, 1980). A comparison of these results with those of the case-control study published in the early 1950s (Doll & Hill, 1950, 1952) highlights the relative merits of the two approaches. The case-control study was begun in April 1948, and the final results published in December 1952. A total of 4342 people were interviewed, of whom 1488 were lung cancer cases. Most of the analyses referred