WORLD HEALTH ORGANIZATION

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER

STATISTICAL METHODS IN CANCER RESEARCH

VOLUME 1 - The analysis of case-control studies

BY

N.E. BRESLOW & N.E. DAY

TECHNICAL EDITOR FOR IARC
W. DAVIS

IARC Scientific Publications No. 32

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER
LYON
1980

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.

First reimpression, 1982
Second reimpression, 1983
Third reimpression, 1989
Fourth reimpression, 1990
Fifth reimpression, 1992
Sixth reimpression, 1994
Seventh reimpression, 1998
Eighth reimpression, 2000

ISBN 92 832 0132 9

REPRINTED IN THE UNITED KINGDOM

N.E. Breslow & N.E. Day (1980) *Statistical Methods in Cancer Research. Volume I – The Analysis of Case-Control Studies,* Lyon, International Agency for Research on Cancer (*IARC Scientific Publications* No. 32)

ERRATA

Page 47, line 8, should read "... in each age group in 1970,..."

Page 60, line 21, should read "... lines are larger than would be expected..."

Page 61, line 3, should read "... $\mathbf{x}$ = UK (Birmingham);..."

Page 75, line 33, should read "$(1\times0.2+6\times0.1)/(1\times0.2+3\times0.3+6\times0.1+18\times0.4)=0.0899$"

Page 76, line 2, should read "$62.5\times0.0899+74.1\times0.9101=73.0\%$"
line 5, should read "73.0% of cancers by eliminating smoking, 60.6% by..."

Page 94, line 25, should read "... with both E and disease, then we should be..."

Page 141, line 4, delete the sentence beginning "Its only major drawback..."

Page 167, last line, should read "... = 0.98 corresponding to..."

Page 174, line 9, should read

$$= \frac{\left(|110-13|-\frac{5}{2}\right)^2}{302} = 29.57.$$

Page 180, last line, fourth box should read "$\frac{3}{M+1}$"

Page 200, line 22, should read "... variables $\mathbf{x} = (\mathbf{x}, \ldots, \mathbf{x}_K)$"
last line, should read "$\mathbf{x}^*$ and $\mathbf{x}$ of risk variables is"

Page 201, line 1, should read "... with a standard ($x = \mathbf{0}$)"

Page 203, line 12, should read

$$= \frac{\text{pr}(z=1\,|\,y=1,\mathbf{x})\text{pr}(y=1\,|\,\mathbf{x})}{\text{pr}(z=1\,|\,y=0,\mathbf{x})\text{pr}(y=0\,|\,\mathbf{x})+\text{pr}(z=1\,|\,y=1,\mathbf{x})\text{pr}(y=1\,|\,\mathbf{x})}$$

Page 204, line 21, should read "$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1\,\boldsymbol{\mu}_0)$."
and throughout, $\boldsymbol{\beta}$ should read $\boldsymbol{\beta}$
line 17, should read "(Truett, Cornfield & Kannel, 1967)"
line 24, should read "... in place of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$,"
last line, should read "... likelihood..."

Page 205, line 1, should read "... and covariances for $\hat{\boldsymbol{\beta}}$ generated..."
line 18, should read "... $\boldsymbol{\beta}$ parameters of interest."

Page 206, line 28, should read "... The α and the $\boldsymbol{\beta}$'s are the..."
line 32, should read "... which are often denoted $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$,"

Page 207, line 1, should read "... $\mathbf{S} = \mathbf{S}(\alpha,\boldsymbol{\beta})$, while..."
line 2, should read "... denoted $\mathbf{I} = \mathbf{I}(\alpha,\boldsymbol{\beta})$."
line 5, should read "Covariance matrix for $(\alpha,\boldsymbol{\beta}) = \mathbf{I}^{-1}(\alpha,\boldsymbol{\beta})$."
line 6, should read "... as the value $\alpha,\boldsymbol{\beta}$ for which..."

Page 212, last line, last column, should read "-0.125 ± 0.189"

Page 218, line 16, should read "... see Table 4.2),"

Page 225, line 8, should read "... of α parameters in (6.12)"

Page 229, line 7, should read "(G_3–G_6 = 0.4, p = 0.5; G_3–G_7 = 2.1, p = 0.15)"

Page 244, line 24, should read "*J. Am. stat. Assoc., 73,*"

Page 245, last two lines, should read "$\boldsymbol{\mu}_1$" and "$\boldsymbol{\mu}_0$"
insert before line beginning $\boldsymbol{\beta}_k$: "$\boldsymbol{\beta}$ vector of log relative risks associated with a vector $\mathbf{x}$ of risk factors"

Page 249, line 33, should read "... to all cells in the $2 \times 2 \times I$ dimensional..."

Pages 284–289, first column heading should read "YEAR OF BIRTH"

Page 314, line 15, should read "16 IF(CC(NDIAG).EQ.000)GOTO 11"

CONTENTS

Foreword 5

Preface 7

Acknowledgements 9

Lists of Symbols 12

1. Introduction 14

2. Fundamental Measures of Disease Occurrence and Association 42

3. General Considerations for the Analysis of Case-Control Studies 84

4. Classical Methods of Analysis of Grouped Data 122

5. Classical Methods of Analysis of Matched Data 162

6. Unconditional Logistic Regression for Large Strata 192

7. Conditional Logistic Regression for Matched Sets 248

Appendices 281

FOREWORD

Epidemiological and biostatistical studies on cancer and other chronic diseases have expanded markedly since the 1950s. Moreover, as recognition of the role of environmental factors in human cancer has increased, there has been a need to develop more sophisticated approaches to identify potential etiological factors in populations living in a wide variety of environments and under very different socioeconomic conditions.

Developments in many countries have required that appropriate governmental agencies establish regulations to control environmental cancer hazards. Such regulations may, however, have considerable social and economic impacts, which require that they be based on careful risk-benefit analyses. Epidemiological studies provide the only definitive information as to the degree of risk in man. Since malignant diseases are clearly of multifactorial origin, their investigation in man has become increasingly complex, and epidemiological and biostatistical studies on cancer require a correspondingly complex and rigorous methodology. Studies such as these are essential to the development of programmes of cancer control and prevention.

Dr N.E. Breslow and Dr N.E. Day and their colleagues are to be commended on this volume, which should prove of value not only to established workers but also to all who wish to become acquainted with the general principles of case-control studies, which are the basis of modern cancer epidemiology.

John Higginson, M.D.
Director,
International Agency for Research
on Cancer, Lyon, France

PREFACE

Twenty years have elapsed since Mantel and Haenszel published their seminal article on statistical aspects of the analysis of data from case-control studies. Their methodology has been used by thousands of epidemiologists and statisticians investigating the causes and cures of cancer and other diseases. Their article is one of the most frequently cited in the epidemiological literature, and there is no indication that its influence is on the wane; on the contrary, with the increasing recognition of the value of the case-control approach in etiological research, the related statistical concepts seem certain to gain even wider acceptance and use.

The last two decades have also witnessed important developments in biostatistical theory. Especially notable are the log-linear and logistic models created to analyse categorical data, and the related proportional hazards model for survival time studies. These developments complement the work done in the 1920s and 1930s which provided a unified approach to continuous data *via* the analysis of variance and multiple regression. Much of this progress in methodology has been stimulated by advances in computer technology and availability. Since it is now possible to perform multivariate analyses of large data files with relative ease, the investigator is encouraged to conduct a range of exploratory analyses which would have been unthinkable a few years ago.

The purpose of this monograph is to place these new tools in the hands of the practising statistician or epidemiologist, illustrating them by application to *bona fide* sets of epidemiological data. Although our examples are drawn almost exclusively from the field of cancer epidemiology, in fact the discussion applies to all types of case-control studies, as well as to other investigations involving matched, stratified or unstructured sets of data with binary responses. The theme is, above all, one of unity. While much of the recent literature has focused on the contrast between the cohort and case-control approaches to epidemiological research, we emphasize that they in fact share a common conceptual foundation, so that, in consequence, the statistical methodology appropriate to one can be carried over to the other with little or no change. To be sure, the case-control differs from the cohort study as regards size, duration and, most importantly, the problems of bias arising from case selection and from the ascertainment of exposure histories, whether by interview or other retrospective means. Nevertheless, the statistical models used to characterize incidence rates and their association with exposure to various environmental or genetic risk factors are identical for the two approaches, and this common feature largely extends to methods of analysis.

Another feature of our pursuit of unity is to bring together various methods for analysis of case-control data which have appeared in widely scattered locations in the epidemiological and statistical literature. Since publication of the Mantel-Haenszel procedures, numerous specializations and extensions have been worked out for particular types of data collected from various study designs, including: 1-1 matching with binary and polytomous risk factors; 1:M matching with binary risk factors; regression models for series of 2×2 tables; and multivariate analyses based on the logistic function. All these proposed methods of analysis, including the original approach based on stratification of the data, are described here in a common conceptual framework.

A second major theme of this monograph is flexibility. Many investigators, once they have collected their data according to some specified design, have felt trapped by the

intransigences of the analytical methods apparently available to them. This has been a particular problem for matched studies. Previously published methods for analysis of 1:M matched data, for example, make little mention of what to do if fewer controls are found for some cases, or how to account for confounding variables not incorporated in the design. The tendency has therefore been to ignore the matching in some forms of analysis, which may result in considerable bias, or to restrict the analysis to a subset of the matched pairs or sets, thus throwing away valuable data. Such practices are no longer necessary nor defensible now that flexible analytical tools are available, in particular those based on the conditional logistic regression model for matched data.

These same investigators may have felt compelled to use a matched design in the first place in order simultaneously to control the effects of several potential confounding variables. We show here that such effects can often be handled adequately by incorporation of a few confounding variables in an appropriate regression analysis. Thus, there is now a greater range of possibilities for the control of confounding variables, either by design or analysis.

From our experiences of working with cancer epidemiologists in many different countries, on projects wholly or partly supported by the International Agency for Research on Cancer, we recognize that not all researchers will have access to the latest computer technology. Even if such equipment is available at his home institution, an investigator may well find himself out in the field wanting to conduct preliminary analyses of his data using just a pocket calculator; hence we have attempted to distinguish between analyses which require a computer and those which can be performed by hand. Indeed, discussion of the methods which require computer support is found mostly in the last two chapters.

One important aspect of the case-control study, which receives only minimal attention here, is its design. While we emphasize repeatedly the necessity of accounting for the particular design in the analysis, little formal discussion is provided on how to choose between various designs. There are at least two reasons for this restriction in scope. First, the statistical methodology for estimation of the relative risk now seems to have reached a fairly stable period in its development. Further significant advances in this field are likely to take place from a perspective which is quite different from that taken so far, for instance using cluster analysis techniques. Secondly, there are major issues in the design of such studies which have yet to be resolved completely; these include the choice of appropriate cases and controls, the extent to which individual matching should be used, and the selection of variables to be measured. While an understanding of the relevant statistical concepts is necessary for such design planning, it is not sufficient. Good knowledge of the particular subject matter is also required in order to answer such design questions as: What factors are liable to be confounders? How important are differences in recall likely to be between cases and controls? Will the exposure influence the probability of diagnosis of disease? Are other diseases liable to be related to the same exposure?

We are indebted to Professor Cole for providing an introductory chapter which reviews the role of the case-control study in cancer epidemiology and briefly discusses some of these issues.

N. E. Breslow and N. E. Day

ACKNOWLEDGEMENTS

Since the initial planning for this monograph in mid-1976, a number of individuals have made significant contributions to its development. Twenty epidemiologists and statisticians participated in an IARC-sponsored workshop on the statistical aspects of case-control studies which was held in Lyon from 12–15 December 1977 (see List of Participants). Funds for this were generously provided by the International Cancer Research Workshop (ICREW) programme, administered by the UICC. Several participants kindly provided datasets to be used for illustrative analysis and discussion during the meeting. Others sent written comments on a rough draft of the monograph which had been distributed beforehand. The discussion was very valuable in directing its subsequent development.

As various sections and chapters were drafted, they were sent for comment to individuals with expertise in the particular areas. Among those who generously gave of their time for this purpose are Professor D.R. Cox, Professor Sir Richard Doll, Dr M. Hills, Dr Kao Yu-Tong, Professor N. Mantel, Dr C. S. Muir, Professor R. Prentice, Professor D. Thompson and Dr N. Weiss. While we have incorporated many of the suggestions made by these reviewers, it has not been possible to accommodate them all; responsibility for the final product is, of course, ours alone.

An important feature of the monograph is that the statistical methods are illustrated by systematic application to data from recent case-control studies. We are indebted to Dr A. Tuyns, IARC, for contributing data from his study of oesophageal cancer in Ille-et-Vilaine, as well as for stimulating discussion. Similarly, we appreciate the generosity of Dr M. Pike and his colleagues at the University of Southern California for permission to use data from their study of endometrial cancer and for sharing with us the results of their analyses. Both these sets of data are given as appendices, as are data from the Oxford Childhood Cancer Survey which were previously published by Dr A. Stewart and Mr G. Kneale.

The data processing necessary to produce the illustrative analyses was ably performed by IARC staff, notably Mr C. Sabai and Miss B. Charnay. Mr P. Smith contributed substantial improvements to the programme for multivariate analyses of studies with 1:M matching (Appendix IV) and subsequently modified it to accommodate variable numbers of cases as well as of controls (Appendix V).

The response of the IARC secretarial staff to the requests for typing of innumerable drafts and redrafts has been extremely gratifying. We would like to thank especially Mrs G. Dahanne for her work on the initial draft and Miss J. Hawkins for shepherding the manuscript through its final stages. Valuable assistance with intermediate drafts was given by Miss M. McWilliams, Mrs A. Rivoire, Mrs C. Walker, Mrs A. Zitouni and (in Seattle) Mrs M. Shumard. The figures were carefully executed by Mr J. Déchaux. We are indebted to Mrs A. Wainwright for editorial assistance and to Dr W. Davis and his staff for final assembly of the manuscript.

During the last year of preparation of this monograph, both authors were on leave of absence from their respective institutions. NEB would like to thank his colleagues at the University of Washington, particularly Drs P. Feigl and V. Farewell for continuation of work in progress during his absence, and to the IARC for financial support during the year. NED would like to thank his colleagues at the IARC, in particular Dr J. Estève, for ensuring the uninterrupted work of the Biostatistics section of the IARC, and to the National Cancer Institute of the United States for financial support during the year.

LIST OF PARTICIPANTS AT IARC WORKSHOP

12–15 December 1977

Professor E. Bjelke
Division of Epidemiology
School of Public Health
University of Minnesota
Minneapolis, Minn., USA
(now at Institute of Hygiene and Social Medicine
University of Bergen
Bergen, Norway)

Professor N. E. Breslow
Department of Biostatistics
University of Washington
Seattle, Wash., USA

Professor P. Cole
Department of Epidemiology
Harvard School of Public Health
Boston, Mass., USA

Mr W. Haenszel
Illinois Cancer Council
Chicago, Ill., USA

Dr Catherine Hill
Institut Gustave-Roussy
Villejuif, France

Dr G. Howe
National Cancer Institute of Canada
Epidemiology Unit
University of Toronto
Toronto, Ont., Canada

Dr A. B. Miller
National Cancer Institute of Canada
Epidemiology Unit
University of Toronto
Toronto, Ont., Canada

Dr B. Modan
Department of Clinical Epidemiology
The Chaim Sheba Medical Center
Tel-Hashomer, Israel

Dr M. Modan
Department of Clinical Epidemiology
The Chaim Sheba Medical Center
Tel-Hashomer, Israel

Professor M. Pike
Professor of Community and Family Medicine, Department of Pathology
University of Southern California
Los Angeles, Calif., USA

Mr P. Smith
Imperial Cancer Research Fund
Cancer Epidemiology and Clinical Trials Unit
University of Oxford
Oxford, United Kingdom

Dr V.B. Smulevich
Department of Cancer Epidemiology
Cancer Research Center
Academy of Medical Sciences of the USSR
Moscow, USSR

Dr J. Staszewski
Institute of Oncology
Gliwice, Poland

IARC Participants:

Miss Bernadette Charnay
Dr N. E. Day
Dr J. Estève
Dr R. MacLennan
Dr C. S. Muir
Miss Annie Ressicaud
Mr C. Sabai
Dr R. Saracci
Dr J. Siemiatycki

LISTS OF SYMBOLS

To aid the less mathematically inclined reader we provide detailed descriptions of the various characters and symbols used in the text and in formulae. These are listed in order of appearance at the end of each chapter. You will notice that some letters or symbols have two different meanings, but these usually occur in different chapters; it will be clear from the context which meaning is intended.

The following mathematical symbols occur in several chapters:

$\times$ denotes multiplication: $3 \times 4 = 12$

Σ summation symbol: for a singly subscripted array of I numbers $\{x_i\}$, $\sum_{i=1}^{I} x_i$ or $\Sigma_i x_i = x_1 + \ldots + x_I$. For a doubly subscripted array $\{x_{ij}\}$, $\Sigma_i x_{ij}$ denotes summation over the i subscript $x_{1j} + \ldots + x_{Ij}$, while $\Sigma\Sigma$ denotes double summation over both indices.

Π product symbol. For a singly subscripted array of I numbers $\{x_i\}$, $\prod_{i=1}^{I} x_i$ or $\Pi_i x_i = x_1 \times x_2 \times \ldots \times x_I$.

log the natural logarithm (to the base e) of the quantity which follows, which may or may not be enclosed in parentheses.

exp the exponential transform (inverse of log) of the quantity which follows, which is usually enclosed in parentheses.