Methods in Molecular Biology 1793

Springer Protocols

Evangelos Evangelou *Editor*

Genetic Epidemiology

Methods and Protocols

Humana Press

Methods in Molecular Biology

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Genetic Epidemiology

Methods and Protocols

Edited by

Evangelos Evangelou

Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece
Department of Epidemiology and Biostatistics, Imperial College London, London, UK

Humana Press

Editor
Evangelos Evangelou
Department of Hygiene and Epidemiology
University of Ioannina Medical School
Ioannina, Greece

Department of Epidemiology and Biostatistics
Imperial College London
London, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7867-0 ISBN 978-1-4939-7868-7 (eBook)
https://doi.org/10.1007/978-1-4939-7868-7

Library of Congress Control Number: 2018942871

Printed on acid-free paper

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Genetic epidemiology is a relatively new field of epidemiology that studies the role of genetic factors in health and diseases and has witnessed some exciting findings in our quest to understand the nature of genetic inheritance. It is an amalgam of methods and approaches applied in traditional epidemiology, statistics, genetics, and bioinformatics and it brings together several scientific disciplines. In the last few years, scientists have been able to map thousands of genetic variants contributing to complex diseases helping to unravel the genetic causes of diseases on a population scale.

This book is a broad overview written at a level that should be accessible to a wide range of interested scientists including epidemiologists, genetic statisticians, human geneticists, clinicians, and bioinformaticians. I hope that this book will be also helpful for graduate students pursuing research in related fields. Some chapters of the book assume a basic level of competence with regard to statistic and probabilistic reasoning; however it was written and edited having in mind that a noncompetent reader will be able to follow, if not all, most of the text. For many scientists, genetic epidemiology is too convoluted to understand; however I hope to persuade the reader that this view is not correct. My goal was to provide a unifying overview of a fast-moving research while providing a description in some depth of the techniques and data that are helping us to understand our genome and how it is related to mainly complex diseases.

Chapter 1 provides an introduction to basic terms of epidemiology whereas Chapter 2 introduces the reader to the key principles of genetic epidemiology including genetic models of inheritance and associations. The next three chapters describe the process of quality control (Chapter 3), the analysis and the detection of common (Chapter 4) and rare variation (Chapter 5) whereas Chapter 6 outlines state-of-the-art meta-analyses approaches for the synthesis of such data. Chapter 7 outlines methods for detecting both gene-gene and gene-environment interactions as well as approaches for increasing statistical power.

The next seven chapters cover novel, state-of-the-art methods that go beyond the conventional approaches for the detection of common variation including analysis in the HLA region (Chapter 8), novel family-based approaches (Chapter 9), approaches for polygenic traits (Chapter 10), multivariate methods for meta-analysis of genetic associations and meta-analysis of gene expression data (Chapters 11 and 12). Chapter 13 covers the rapidly evolving method of Mendelian Randomization that is used for the estimation of causal effects of an exposure on an outcome, whereas computational methods for the analysis of Copy Number Variation are presented in Chapter 14. We conclude in the last two chapters by assessing the functional role of the identified variants (Chapter 15) and the challenges we are facing to use human genetics to identify and validate novel drug targets (Chapter 16).

I thank sincerely all those who have helped to bring this book together and I am grateful to the coauthors who accepted my invitation and contributed to this book, devoting valuable time and effort.

Ioannina, Greece *Evangelos Evangelou*

Contents

Contributors

EMIL VINCENT ROSENBAUM APPEL • *Section for Metabolic Genetics, Faculty of Health Sciences, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark*

PANTELIS G. BAGOS • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

LAZAROS BELBASIS • *Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece*

VANESA BELLOU • *Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece*

ANTONIO J. BERLANGA-TAYLOR • *Department of Epidemiology and Biostatistics, Faculty of Medicine, School of Public Health, MRC-PHE Centre for Environment and Health, Imperial College London, London, UK*

GEORGIA G. BRALIOU • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

ABBAS DEHGHAN • *Department of Epidemiology and Biostatistics, Imperial College London, London, UK*

ANDREW T. DEWAN • *Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA*

NIKI L. DIMOU • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece; Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece*

KAROL ESTRADA • *Translational Genome Sciences, Biogen, Cambridge, MA, USA*

JAVIER GUTIERREZ-ACHURY • *The Wellcome Trust Sanger Institute, Cambridgeshire, UK*

PANAGIOTA I. KONTOU • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

KAROLINE KUCHENBAECKER • *Wellcome Trust Sanger Institute, Cambridge, UK; University College London, London, UK*

ZOLTÁN KUTALIK • *Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland*

AURÉLIEN MACÉ • *Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland; Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland*

KYRIAKI MICHAILIDOU • *Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus*

LOUKAS MOUTSIANAS • *The Wellcome Trust Sanger Institute, Cambridgeshire, UK*

KALLIOPE PANOUTSOPOULOU • *Wellcome Sanger Institute, Wellcome Genome Campus, Cambridgeshire, UK*

KATERINA G. PANTAVOU • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

ATHANASIA PAVLOPOULOU • *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece; International Biomedicine and Genome Institute (iBG-Izmir), Dokuz Eylul University, Konak, Turkey*

RAHA PAZOKI • *Department of Epidemiology and Biostatistics (inc MRC-HPA Centre), School of Public Health, Imperial College London, London, UK*

KONSTANTINOS K. TSILIDIS • *Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK*

ARMAND VALSESIA • *Nestlé Institute of Health Sciences, Lausanne, Switzerland*

KLAUDIA WALTER • *Wellcome Sanger Institute, Wellcome Genome Campus, Cambridgeshire, UK*

ELEANOR WHEELER • *Wellcome Sanger Institute, Wellcome Genome Campus, Cambridgeshire, UK*

QI YAN • *Division of Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA, USA*

Chapter 1

Introduction to Epidemiological Studies

Lazaros Belbasis and Vanesa Bellou

Abstract

The basic epidemiological study designs are cross-sectional, case-control, and cohort studies. Cross-sectional studies provide a snapshot of a population by determining both exposures and outcomes at one time point. Cohort studies identify the study groups based on the exposure and, then, the researchers follow up study participants to measure outcomes. Case-control studies identify the study groups based on the outcome, and the researchers retrospectively collect the exposure of interest. The present chapter discusses the basic concepts, the advantages, and disadvantages of epidemiological study designs and their systematic biases, including selection bias, information bias, and confounding.

Key words Bias, Case-control study, Cohort study, Confounding, Information bias, Observational studies, Selection bias, Study design

1 Definition of Epidemiology

Epidemiology is defined as "the study of the occurrence and distribution of health-related events, states, and processes in specified populations, including the study of the determinants influencing such processes, and the application of this knowledge to control relevant health problems" [1]. It is apparent that the scope of Epidemiology is very wide and mainly includes the study of incidence and prevalence of health conditions and traits, the study of their determinants (i.e., risk and protective factors), and the design of potential strategies for disease prevention.

Many subfields of Epidemiology have been developed, including environmental epidemiology, genetic epidemiology, and nutritional epidemiology. An early definition of *Genetic Epidemiology* defined it as "the field that addresses the etiology, distribution, and control of disease in groups of related individuals and the inherited causes of diseases in population [2, 3]. Later, this definition was broadened to include the role of interaction between the environment and the genetic factors in the occurrence of diseases [3]. Also, the term *Human Genome Epidemiology* was coined to

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793, https://doi.org/10.1007/978-1-4939-7868-7_1, © Springer Science+Business Media, LLC, part of Springer Nature 2018

describe "the field that uses systematic applications of epidemiologic methods and approaches to the human genome to assess the impact of human genetic variation on health and disease" [4]. The present chapter constitutes a brief introduction to epidemiologic study designs for Genetic Epidemiology.

2 Cross-Sectional Studies

The defining characteristic of cross-sectional studies is that both exposure and outcome are ascertained at the same time. The temporal sequence is often impossible to work out, because exposure and outcome are identified at one time point. However, cross-sectional studies are useful in Genetic Epidemiology, because genetic exposures cannot change over time and unquestionably preceded the outcome [5, 6].

3 Cohort Studies

A cohort study is "an observational epidemiological study in which subsets of a defined population can be identified who are, have been, or in the future may be exposed or not exposed to a factor or factors hypothesized to influence the occurrence of a given outcome" [1].

A cohort study tracks two or more groups forward from exposure to outcome. This type of study can be done by going ahead in time from the present (prospective cohort study) or by going back in time to identify the cohorts and following them up to the present (retrospective cohort study) [7]. In both prospective and retrospective designs, a cohort study moves in the same direction, although data gathering might not. The exposure of interest is measured at the beginning of the study, and the two groups are defined based on the exposure or the level of exposure to a specific factor.

Prospective cohort studies constitute the most reliable type of observational studies, and they present many advantages. The temporal sequence between putative cause and outcome is usually clear, given that the exposed and unexposed can often be seen to be free of the outcome at the beginning of the study [7]. Also, cohort studies are useful in the investigation of multiple outcomes that might arise after a single exposure [7]. For example, a cohort study can be designed to assess the association between birth weight and multiple health outcomes or traits. However, in this case, publication bias and reporting bias are often observed when the researchers publish or report only the statistically significant findings [8]. - Furthermore, testing multiple associations may lead to many false-positive findings due to chance. To avoid reporting and publication bias, several approaches have been proposed. Study registration and pre-specification of the project design and the analysis plan are important initiatives to avoid post-hoc secondary analyses seeking additional statistically significant findings [8]. To reduce the rate of false-positive findings in the field of Genetic Epidemiology, several statistical approaches for multiple testing have been suggested, including a more stringent P-value, q-values, or false discovery rate [9].

The cohort studies are also useful in the study of rare exposures [7]. For example, they are the appropriate study design to examine the health effects of occupational exposures, such as ionising radiation and chemicals. Moreover, cohort studies reduce the risk of survivor bias, i.e., diseases that are rapidly fatal are difficult to study because of this factor [7]. Finally, cohort studies allow the calculation of incidence rates, risk ratios, and other outcome measures, such as survival curves and hazard ratios [7].

However, cohort studies also have important shortcomings. Ideally, both cases and controls should be the same in all important aspects, except for the exposure of interest [7]. This seldom occurs, and the absence of comparability between groups being studied results in *selection bias* [10]. Also, cohort study is not the optimum study design for rare diseases or diseases that take a long time to develop, such as cancer [7]. Moreover, loss to follow-up can be an important issue in this study design, especially for longitudinal studies that have a long follow-up period. In this case, differential losses to follow-up between exposed and unexposed can bias results [7]. Another drawback is the potential change of the exposure status of study participants during the follow-up period [7].

One of the most frequent variations of cohort studies is the *nested case-control study*, which is “a type of case-control study in which cases and controls are drawn from the population in a cohort study” [1]. The rational for designing a case-control study nested in a cohort study is that some exposure variables are too expensive to determine on the full cohort [7]. Nested case-control studies can be designed to examine genetic associations for a specific health-related outcome. The variable of interest is measured in the cases group, and then the investigator chooses a random sample of all participants who did not develop the outcome. This approach minimizes the cost of measuring the variable of interest and ensures that the exposure was present before the development of the outcome. During the study design, a matching process is used and the controls are matched to cases by important characteristics, such as age and sex [7].

4 Case-Control Studies

A case-control study is an observational epidemiological study of persons with the disease of interest and a suitable control group of persons without the disease [1]. In this study design, study groups are defined by outcome, and the study population is divided into two groups, cases and controls, based on whether the outcome of interest has occurred [11, 12]. Case-control studies cannot yield incidence rates, but they provide an odds ratio, derived from the proportion of individuals exposed in each of the case and control groups [12]. When the incidence rate of a particular outcome in the population of interest is low (*rare disease assumption*), the odds ratio from a case-control study is considered a good approximation of risk ratio [12].

Case-control studies are considered an efficient study design in terms of time, money, and effort. Specifically, this study design is appropriate to investigate diseases with a low incident rate and diseases that have a long latency period, such as cancer [12]. However, case-control studies have also some disadvantages and, in these cases, cohort studies are considered a more efficient design. If the frequency of exposure is low, case-control studies quickly become inefficient, because researchers would have to examine many cases and controls to find one who had been exposed [12]. A simplified rule has been proposed indicating that cohort studies are more efficient in settings in which the incidence of outcome is higher than the prevalence of exposure [12]. Also, selection of a control group and obtaining exposure history [12] are two main methodological issues affecting the validity of the results of case-control studies and are discussed in more detail.

The term "*selection bias*" is used to describe "the bias in the estimated association or effect of an exposure on an outcome that arises from the procedures used to select individuals into the study or the analysis" [1]. Investigators can reduce selection bias by minimizing judgement in the selection process, and the selection process should be defined and described in detail for both case and control group [12]. Often only a sample of cases from a population is included as participants in a case-control study. During the selection process, investigators should focus on incident cases rather than prevalent cases, since diagnostic patterns change over time and this can affect the consistency of diagnosis between incident and prevalent cases [12]. Controls should be free of the disease being studied, but they should also be representative of those individuals who would have been selected as cases had they developed the disease. For example, if the case group included all affected individuals in a specified region, then the control group could be chosen at random from the general population of the same area [12].

The term "*information bias*" is used to describe "a flaw in measuring exposure, covariate, or outcome variables that results in different quality of information between comparison groups" [1]. A type of information bias is recall bias, which is "a systematic error due to differences in accuracy or completeness of recall to memory of past events or experiences" [1]. For example, in a case-control study for risk factors of melanoma, when information for past history of sun exposure, sunburns and solarium use is retrospectively collected, melanoma cases are more prone to report an increased exposure to these factors [13]. Also, information bias could be caused by data gatherers using different techniques to elicit information based on the case or control status. Thus, data gatherers should be unaware of the case or control status of the respondents, to minimize the risk for information bias [12].

Another important issue in the design of case-control studies is the *matching ratio* of controls to cases. There is usually little marginal increase in precision from increasing the ratio of controls to cases beyond four, except when the effect of exposure is large [14]. In general, the best way to increase precision in a case-control study is to increase the number of cases by widening the base geographically or temporally rather than by increasing the number of controls, because the marginal increase in precision from an additional case is greater than from an additional control [14].

5 Confounding

Selection bias and information bias have already been discussed in the section above. Another important issue in epidemiological studies is confounding. Three criteria should be fulfilled for a variable to be a confounder [15]. First, the confounding factor must be an extraneous risk factor for the disease. Second, a confounding factor must be associated with the exposure under study in the source population. Third, a confounding factor must not be affected by the exposure or the disease, and it cannot be an intermediate step in the causal path between the exposure and the disease of interest.

Several approaches have been suggested to control for confounding [10]. These methods can be applied either during the selection of cases and controls or during the statistical analyses. The simplest approach is restriction, i.e., during recruitment period researchers exclude individuals having the exposure that is suspected to be a confounding factor. Another way is pairwise matching. In a case-control study, during the selection of controls, cases and controls can be matched by the confounding factor. However, matching can be proven challenging if it is done on several potential confounding factors. Moreover, control for confounding can be done after a study has been completed. One approach is stratification which can be considered a post hoc restriction, done during the analysis [10]. Multivariate techniques (e.g., multivariate logistic regression) have also been proposed to examine the effect of one variable while controlling for the effect of many other factors [10].

References

1. Porta M (ed) (2014) A dictionary of epidemiology. Oxford University Press, Oxford
2. Morton NE (1997) Genetic epidemiology. Ann Hum Genet 61:1–13
3. Boslaugh SE (2007) Genetic epidemiology. In: Boslaugh SE (ed) Encyclopedia of epidemiology. SAGE Publications, Thousand Oaks, pp 417–420
4. Khoury M, Little J, Burke W (2004) Human genome epidemiology: scope and strategies. In: Human genome epidemiology. Oxford University Press, New York, pp 3–16
5. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131
6. Grimes DA, Schulz KF (2002) Descriptive studies: what they can and cannot do. Lancet (London, England). 359:145–149
7. Grimes DA, Schulz KF (2002) Cohort studies: marching towards outcomes. Lancet (London, England). 359:341–345
8. Ioannidis JPA, Munafò MR, Fusar-Poli P et al (2014) Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci 18:235–241
9. Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389
10. Grimes DA, Schulz KF (2002) Bias and causal associations in observational research. Lancet (London, England) 359:248–252
11. Gordis L (2014) Case-control and other study designs. In: Epidemiology. Saunders, Philadelphia, pp 189–214
12. Schulz KF, Grimes DA (2002) Case-control studies: research in reverse. Lancet (London, England). 359:431–434
13. Parr CL, Hjartåker A, Laake P et al (2009) Recall bias in melanoma risk factors and measurement error effects: a nested case-control study within the Norwegian women and Cancer study. Am J Epidemiol 169:257–266
14. Wacholder S, Silverman DT, McLaughlin JK et al (1992) Selection of controls in case-control studies. III. Design options. Am J Epidemiol 135:1042–1050
15. Rothman K, Greenland S, Lash T (2008) Validity in epidemiologic studies. In: Modern epidemiology. Lippincott Williams & Wilkins, Philadelphia, pp 128–148