

Advances in Experimental Medicine and Biology 1082

Bertram K. C. Chan

Biostatistics for Human Genetic Epidemiology

 Springer

Advances in Experimental Medicine and Biology

Volume 1082

Editorial Board

IRUN R. COHEN, *The Weizmann Institute of Science, Rehovot, Israel*

ABEL LAJTHA, *N.S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*

JOHN D. LAMBRIS, *University of Pennsylvania, Philadelphia, PA, USA*

RODOLFO PAOLETTI, *University of Milan, Milan, Italy*

NIMA REZAEI, *Tehran University of Medical Sciences, Children's Medical Center Hospital, Tehran, Iran*

More information about this series at <http://www.springer.com/series/5584>

Bertram K. C. Chan

Biostatistics for Human Genetic Epidemiology

 Springer

Bertram K. C. Chan
Epidemiology and Biostatistics
Loma Linda University School of Medicine
and Public Health
Sunnyvale, CA, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-319-93790-8 ISBN 978-3-319-93791-5 (eBook)
<https://doi.org/10.1007/978-3-319-93791-5>

Library of Congress Control Number: 2018953701

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to the glory of God and to my better half Marie Nashed Yacoub Chan

Preface

Some genetic epidemiologic experiences of, and concomitant challenges for, this writer are as follows -

Experience and Challenge #1: Type-1 Diabetes

CASE SUBJECT: A child with Type-1 Diabetes – the case subject was a 14-year-old child who was clinically diagnosed, some 2 years previously, as suffering from Type-1 (juvenile) diabetes (Chan 2015). Now, at the currently accepted level of understanding, only about 5% of people with diabetes have this form of the disease. In a case subject with Type-1 diabetes, the body does *not* produce insulin. Normally, the human body breaks down the starches and sugars, that one eats, first into a simple sugar (called glucose) which it then used for energy. In this process, insulin is a hormone that the body *needs* to get glucose from the bloodstream into the cells of the body (American Diabetes Association 2017). And with the help of insulin therapy and other treatments, young children may learn to manage their conditions and live long, healthy, and productive lives. In almost all cases of Type-1 diabetes, the medical and health communities focus on the medical engineering aspect of how best to effectively “pump” insulin into the patient’s system. The disease is generally considered as permanently irreversible, viz., “incurable”!

One would certainly like to learn more about the genetic basis of case subjects diagnosed with Type-1 diabetes!

Moreover, this particular 14-year-old case subject was enrolled in a test in which the child orally took a prescribed medication for a period of about 3 months. Interestingly, this special medication was a Traditional Chinese Medicine (TCM) formulation of herbal origin. During this period of special medication, A1C blood tests were taken to monitor the progress of the case subject. The progressive A1C test results were as follows:

$$9+ \rightarrow 8.4 \rightarrow 7.8 \rightarrow 7.45 \rightarrow 6.7(\%)$$

The low/last reading was below the 6.9% level – which may be considered as the A1C reading for a normal and non-diabetic person!

How does this particular test result affect the accepted medical position that Type-1 diabetes is permanently irreversible? Can epidemiologic research help? Clearly much epidemiologic investigation is called for in this situation. Actually, there had been a clinical trial in which the same TCM treatment was given to more than 10,000 case subjects resulted in a positive response (viz., improved stability of blood glucose control *without* insulin) in about 30% of the test population. Such results should be

considered as strong justification for further epidemiologic studies (including genetic-epidemiologic investigations) in this particular area!

Experience and Challenge #2: Autism Spectrum Disorders (ASD) – A Costly Condition!

Recently, this writer experienced a cultural shock: a certain rental property placed on the open market received the highest bid from an organization which provides daily care to autistic children. This serendipitous result came when it was discovered the State Health Department considers it appropriate to heavily support such an organization which provides daily health and educational care to all qualified children with ASD! (Newschaffer et al. 2007) Can one expect some relief for the financial cost for supporting such a societal program? And what about the concomitant social and personal costs for supporting such a program?

Confronting the question: “Is there a relief, or a cure, in sight for the autistic state of human conditions?” A recent report (<http://edition.cnn.com/2017/04/05/health/autism-cord-blood-stem-cells-duke-study/index.html>) points to a study on the safety and effectiveness of infusion of umbilical cord blood into children with autism did yield some promising results! Perhaps one may raise the question: Is genetics still a critical factor involved in such an extraordinary and heroic approach to medicine and health care?

Experience and Challenge #3: Childhood Brain Tumors

The challenge of understanding the genetic epidemiology of fatal childhood brain cancer tumors was experienced by this writer – he was acquainted with a married couple (both of whom came from very close and similar ethnic backgrounds). Later, this couple was blessed with the birth of a child, who, at the age of 12 months, developed fatal brain cancer tumors. The infant spent the next few months in the hospital before passing away! It was, indeed, a very sad occasion at the memorial service of that precious child. Sometime later, the parents decided to forego the conceiving and birthing another offspring, and chose to adopt an infant – across ethnic and racial lines!

This instance seems to call for a much-motivated understanding of the “Genetic Epidemiology of Childhood Brain Cancer Tumors” (Bondy 1990).

Genetic Epidemiology holds a critically important and effective role in understanding the critical factors in the aforementioned diseases and health issues, especially with respect to hereditary and environmental factors (the “Nature vs. Nurture” issues). Starting from population-based methods, the magnitude of genetic effects on health and diseases may be assessed (Austin 2013). To these approaches one may added quantitative methods, including biostatistical analysis. The latter methodology may be efficiently enhanced with the now-popular R programming software (Chan 2015).

To understand, and ultimately to apply the knowledge of, the etiology of a disease, it seems imminently helpful to unravel the relationships (if any) that govern the genetic bases of the disease. While genetics is complicated, it is to be hoped that the use of available biostatistical power, supported by the efficiencies of the computer program (developed largely for biostatistical applications), will go some positive way toward resolving some of the complex relations within genetic epidemiology.

Experience and Challenge #4: CMT (Charcot-Marie-Tooth) Disease^[w]

From a personal friend of the author was learnt that a rather common hereditary disease of genetic origins can cause severe weakness of the limbs that required supporting metallic braces to aid simple daily walking. This disease, known as CMT, was recently diagnosed in a personal acquaintance!

CMT is one of the most common inherited neurological disorders, affecting approximately 1 in 2,500 people in the United States of America. The disease is named after the three physicians who first identified it in 1886 – Jean-Martin **Charcot** and Pierre **Marie** in Paris, France, and Howard Henry **Tooth** in Cambridge, England. CMT, also known as **Hereditary Motor and Sensory Neuropathy (HMSN)** or **Peroneal Muscular Atrophy (PMA)**, comprises a group of disorders that affect peripheral nerves. The peripheral nerves lie outside the brain and spinal cord and supply the muscles and sensory organs in the limbs. Disorders that affect the peripheral nerves are called peripheral neuropathies.

Although there is no known cure for CMT, physical therapy, occupational therapy, braces and other orthopedic devices, and even orthopedic surgery may help individuals deal with the disabling symptoms of the disease. In addition, pain-killing drugs can be prescribed for individuals who have severe pain.

Physical and occupational therapy, the preferred treatment for CMT, involves muscle strength training, muscle and ligament stretching, stamina training, and moderate aerobic exercise. Most therapists recommend a specialized treatment program designed with the approval of the person's physician to fit individual abilities and needs. Therapists also suggest entering into a treatment program **early** as muscle strengthening may delay or reduce muscular atrophy, so strength training is most useful if it begins before nerve degeneration and muscle weakness progresses to the point of disability!

Experience and Challenge #5: Alzheimer Disease

To this author, this experience has a rather personal and emotional background: the beloved pastor of the home church retired, but soon his wife died suddenly owing to the rupture of an abdominal aneurism – and a year later, the pastor himself rapidly lapsed into severe symptoms of Alzheimer Disease (AzD) – unable even to remember the first name of the author who had been his personal friends for years! Thus, the beloved pastor had become a “stranger,” then passed away within a year or so! And, more than that:

On a national, if not worldwide, scale, it has been reported that (<http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>):

Could Alzheimer's really bankrupt Medicare and Medicaid? By **Lindsay Carlton** Published March 01, 2017



The disease that could collapse Medicare, Medicaid

The most expensive medical condition in America threatens to bankrupt Medicare, Medicaid and the life savings of millions of Americans. But the perpetrator isn't cancer or heart disease — it's Alzheimer's.

Fox News' Dr. Manny Alvarez sat down with Dr. Rudolph Tanzi, a professor of neurology at Harvard Medical School who participated in PBS' "Alzheimer's: Every Minute Counts" documentary, which takes a closer look at the critical financial problem Americans are facing with the disease, to discuss the issue.

"Because we're living so long, our health span, especially our brain health span, is not keeping up with our life span," Tanzi told Fox News. "All of modern medicine has us living on average till 80 years old, and by 85 years old you have a 40 to 50% chance of having Alzheimer's."

In 2016, total payments for health care, long-term care and hospice were estimated to be \$236 billion for people with Alzheimer's and other dementias, according to the Alzheimer's Association.

Tanzi explained that right now, \$1 of every \$5 (20.0%) in Medicare and Medicaid funding goes toward Alzheimer's patients' care. Given how many more Alzheimer's patients are expected to be diagnosed within the next decade, that number is predicted to increase to every \$1 in \$3 (33.3%). In that case, the program's funding may collapse, which would leave insufficient funds to prevent other age-related disease, he said.

"It hits every sector from the burden on the family: the caregiver taking care of their loved one who they're losing in front of their eyes, and then the government costs, assisted living," Tanzi said.

Much have been achieved in the study of population-based genetics, biostatistical genetics, epidemiology, and hopefully and finally make a significantly useful impact on genetic epidemiology. To that end, the author is prepared to introduce the title

"Biostatistics for Genetic Epidemiology: An Introduction Using R"

in terms of the following chapters:

1. **Introduction to Genetic Epidemiology**
2. **Data Analysis Using R Programming**
3. **Human Genetics and Genetic Epidemiology**
4. **Statistical Human Genetics Using R**
5. **Genetic Epidemiology Using R**

Sunnyvale, CA, USA

Bertram K. C. Chan

References

- American Diabetes Association (2017) <http://diabetes.org>
- Austin MA (2013) Genetic epidemiology: methods and applications (Modular Text Series). CABI Publishing, Wellingford
- Bondy ML (1990) Genetic epidemiology of childhood brain tumors, Texas Medical Center Dissertations (via ProQuest). AA119109972. <http://digitalcommons.library.tmc.edu/dissertations/AA119109972>
- Chan BKC (2015) Biostatistics for epidemiology and public health disorders. *Ann Rev Pub Health* 28:235–258. 10.1146/annurev.pubhealth.28.021406.144007
- <http://edition.cnn.com/2017/04/05/health/autism-cord-blood-stem-cells-duke-study/index.html>
- <http://www.foxnews.com/health/2017/03/01/could-alzheimers-really-bankrupt-medicare-and-medicaid.html>
- Newschaffer CJ et al (2007) The epidemiology of autism spectrum using R, Springer Publishing Company, New York

Contents

1	Introduction to Human Genetic Epidemiology	1
1.1	Medicine, Preventive Medicine, Public Health, and Epidemiology	1
1.1.1	An Overseas Vacation Tour and Worldwide Infectious Diseases	1
1.1.2	Genetics and Infectious Diseases [<i>infectious diseases and genetics => Genetics of infectious diseases => academic.oup.com</i>]	5
1.2	Human Genetic Epidemiology (HGE)	5
1.2.1	The Human Genome Project (HGP) ^[W]	6
1.2.2	Human Genes, Genetics, and Health	12
1.2.3	A Glossary of Common Terms in Human Genetics	18
1.2.4	Human Genetics in Medicine ^[W]	18
1.2.5	Human Genetic Epidemiology	44
1.2.6	Applied Statistical Human Genetics	46
	References	46
2	Data Analysis Using R Programming	47
2.1	Data and Data Processing	48
2.2	Beginning R	53
2.2.1	A First Session Using R	56
2.2.2	The R Environment	68
2.3	R As a Calculator	70
2.3.1	Mathematical Operations Using R	70
2.3.2	Assignment of Values in R, and Computations Using Vectors and Matrices	72
2.3.3	Computations in Vectors and Simple Graphics	72
2.3.4	Use of Factors in R Programming	73
2.3.5	Simple Graphics	74
2.3.6	x As Vectors and Matrices in Statistics	78
2.3.7	Some Special Functions That Create Vectors	79
2.3.8	Arrays and Matrices	80
2.3.9	Use of the Dimension Function dim() in R	81
2.3.10	Use of the Matrix Function matrix() in R	81
2.3.11	Some Useful Functions Operating on Matrices in R	81
2.3.12	NA ‘Not Available’ for Missing Values in Datasets	82
2.3.13	Special Functions That Create Vectors	83

2.4	Using R in Data Analysis in Human Genetic Epidemiology	87
2.4.1	Entering Data at the R Command Prompt	87
2.4.2	The Function list() and the Construction of data.frame() in R	98
2.5	Univariate, Bivariate, and Multivariate Data Analysis	100
2.5.1	Univariate Data Analysis	101
2.5.2	Bivariate and Multivariate Data Analysis	103
	Appendix 1	121
	Documentation for the plot function	121
	Special References	122
3	Applied Statistics for Human Genetics Using R	123
3.1	Some Fundamental Concepts in the Theory of Probability and Applied Statistics in Epidemiology	123
3.2	Biostatistical Concepts and Measures in Genetic Association	124
3.2.1	Familial Aggregation Studies	125
3.2.2	Segregation Studies	127
3.2.3	Linkage Studies	128
3.2.4	Association Studies	130
3.3	Genome-wide Association Studies (GWAS)	136
3.3.1	A Worked Example of SNPs-based Whole Genome Association Study	137
3.4	Big Data and Human Genomics	142
3.4.1	What Is Big Data? [W]	142
3.4.2	What Is Genetic Big Data? And Where Is It Taking Genetics?	142
3.4.3	Analysis of Human Genomics	143
	References	144
4	Applied Human Genetic Epidemiology	145
4.1	The Study of Human Genetic Epidemiology	146
4.1.1	Family Studies in Genetic Epidemiology	148
4.2	Human Genetic Influences on Diseases	165
4.2.1	Genetic Relationships in a Familial Aggregation ^[*]	165
4.2.2	Familial Risk of Diseases	167
4.2.3	Heritability Analysis ^[G]	174
4.2.4	Molecular Variation Study Methods	181
4.3	Genomics for Human Genetic Epidemiology	183
4.3.1	Complex Traits and Mendelian Inheritance	184
4.4	Factors in Human Genetic Epidemiology	198
4.4.1	Linkage Analysis	199
4.4.2	Family Association Studies	200
4.5	Human Genetic Association	204
4.6	Genetic Epidemiology Owing to Population Stratification	205
4.7	Environmental Effects on Genetic Epidemiology ^[Google]	206
4.7.1	Environmental Factors on Genetic Epidemiology ^[Google]	207
4.8	Genetic Epidemiology and Public Health ^[Google]	209
	Special References	216

- 5 Human Genetic Epidemiology Using R** 217
 - 5.1 Biostatistical Human Genetics 219
 - 5.1.1 Some Preliminary Remarks on the *T*-Test in Statistics 222
 - 5.2 Human Genetic Data Concepts 231
 - 5.2.1 The Study of Human Genetic Variation 231
 - 5.2.2 Manhattan Plots^[Wikipedia] 252
 - 5.3 Procedures for Multiple Comparison 263
 - 5.3.1 Worked Examples of Statistical Tests and Utilities for Genetic Association 265
 - 5.3.2 Worked Examples of Statistical Tests and Utilities for Genetic Association 281
 - 5.4 Regression Decision Trees and Classifications^[Google] 308
 - 5.5 Multi-dimensional Analysis in Genetic Epidemiology 311
 - 5.5.1 Biomedical Background Challenges to Genetic Epidemiology 312
 - 5.5.2 Worked Examples in Epidemiology 324
 - References 341

- References** 343

- Index** 353

About the Author

Bertram K. C. Chan PhD, PE, Life Member-IEEE, completed his secondary education in Sydney, Australia, having passed the New South Wales State Leaving Certificate (viz., university matriculation examination) with excellent results in mathematics and in honours physics and in honours chemistry.

He then completed both a Bachelor of Science degree in Chemical Engineering, with First Class Honours (summa cum laude), and a Master of Engineering Science degree in Nuclear Engineering at the University of New South Wales, and a PhD degree in Engineering at the University of Sydney.

This was followed by 2 years of work as a Research Engineering Scientist at the Australian Atomic Energy Commission Research Establishment, and 2 years of a Canadian Atomic Energy Commission postdoctoral fellowship at the University of Waterloo, Canada.

He had undertaken additional graduate studies at the University of New South Wales, at the American University of Beirut, and at Stanford University, in mathematical statistics, computer science, and pure and applied mathematics (abstract algebra, automata theory, numerical analysis, etc.), and in electronics and electromagnetic engineering.

His professional career includes over 10 years of full-time, and 10 years of part-time, university-level teaching and research experience in several institutions, including an appointment as a research associate in biomedical and statistical analysis, Perinatal Biology Section, ObGyn Department, University of Southern California Medical School, teaching at Loma Linda University, Middle East University, and research engineering staff positions at Lockheed Missile & Space (10 years), Apple (7 years), Hewlett-Packard (3 years), and at a start-up company (Foundry Networks) in the manufacture of Internet hardware and software: gigahertz switches and routers (7 years).

In recent years:

- He supported the biostatistical work of the Adventist Health Studies II research program at the Loma Linda University Health (LLUH) School of Medicine, California, and consulted as a forum Lecturer for several years in the LLUH School of Public Health (biostatistics, epidemiology, and population medicine). The LLUH lectures formed part of this book. In these lectures, Dr. Chan introduced the use of the programming language R and designed these lectures for the biostatistical elements for courses in the MPH, MsPH, DrPH, and PhD programs, with special reference to epidemiology in particular and public health and population medicine in general.
- Dr. Chan has three US patents in electromagnetic engineering, has published over 30 engineering research papers, and authored a 16-book set in educational mathematics (Chan 1978), as well as two

monographs entitled: *Biostatistics for Epidemiology and Public Health Using R* (Chan 2016) and *Applied Probabilistic Calculus for Financial Engineering: An Introduction Using R* (Chan 2017).

- He is a registered Professional Engineer (**PE**) in the State of California, as well as a life member of the Institute of Electrical and Electronic Engineers (**IEEE**).

References

- Chan BKC (1978) A new school mathematics for Hong Kong, 10 Volumes: 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5B, 6 Workbooks: 1A, 1B, 2A, 2B, 3A, 3B. Ling Kee Publishing Co., Hong Kong
- Chan BKC (2016) *Biostatistics for epidemiology and public health using R*. Springer, New York (with additional materials on the Publisher's website)
- Chan BKC (2017) *Applied probability calculus for financial engineering: an introduction using R*. Wiley, Hoboken