流行病学与计算机应用

LIUXINGBINGXUE YU JISUANJI YINGYONG

俞顺章　姜庆五　主编

复旦大学出版社

流行病学与计算机应用

LIUXINGBINGXUE YU JISUANJI YINGYONG

流行病学与计算机应用

主编 俞顺章 姜庆五

编委 俞顺章 董传辉 郜艳晖 沈洪兵 夏 毅
周宝森 蔡 琳 陶旭光 孙晓武 赵守军
赵根明 张 涛 柏建岭 周晓明 赵 宁
任秋芳 张作风 黄德生 孟 炜 吕 桦
沈福民 陈启光 霍 翔

审定 俞国培 徐德忠

复旦大学出版社

图书在版编目(CIP)数据

流行病学与计算机应用 / 俞顺章,姜庆五主编. —上海:
复旦大学出版社,2011.4
ISBN 978-7-309-06523-7

Ⅰ.流… Ⅱ.①俞…②姜… Ⅲ.计算机应用—流行病学
Ⅳ.R18-39

中国版本图书馆 CIP 数据核字(2009)第 027074 号

流行病学与计算机应用
俞顺章　姜庆五　主编
责任编辑/魏　岚

复旦大学出版社有限公司出版发行
上海市国权路 579 号　邮编:200433
网址:fupnet@fudanpress.com　http://www.fudanpress.com
门市零售:86-21-65642857　　团体订购:86-21-65118853
外埠邮购:86-21-65109143
江苏省句容市排印厂

开本 787×1092　1/16　印张 25.75　字数 626 千
2011 年 4 月第 1 版第 1 次印刷

ISBN 978-7-309-06523-7/R·1074
定价:70.00 元

内 容 提 要

《流行病学与计算机应用》一书系统地介绍了国内外流行病学与计算机应用领域的发展动态，尤其是在疾病控制中的应用进展。书中包括3大篇：第一篇总论中除概述计算机在流行病学上的应用外，还包括疾病数据管理、统计分析、图形处理、数学模型等内容；第二篇实践中介绍了流行病学调查中经常遇到的样本大小、生存分析、多变量分析、决策分析等；第三篇程序中介绍了流行病学和疾病控制中常用软件等。本书重点介绍了流行病学工作者应用较多的软件 Epi Info 和 Epi Data；还着重列举了流行病学常用分析程序 CP Epi，SAS，logistic 回归分析，生存分析，分子遗传流行病学等程序。书后光盘附有常用的流行病学软件程序，以便使用。

本书的特点：内容新颖，方法全面，应用方便，实践和理论并重，近况与远景共存。大部分资料来自复旦大学公共卫生学院流行病学教研组在科研和教学中常年的积累、总结和实践；一部分来自国际合作和应用；还有一部分由我院在约翰·霍普金斯大学、费城大学等的博士和兄弟单位的教授共同参与编写。

本书可以用作研究生教材，亦可供专业人士参考。

序

我国提出的“以信息化带动工业化，以信息化推动现代化”方针，使信息化和现代化的步伐大大加快。流行病学的信息化、现代化是离不开现场工作的，而现场工作更离不开数据处理和统计分析。过去调查所得的数据只能靠算盘、计算尺、计算器来运算。1980 年当俞顺章教授到加拿大多伦多大学进修时，国外应用计算机蓬勃发展，按照 Moore 定律，集成电路上可以容纳的零件数量，每过一年半便会增长一倍，性能亦提升一倍。他除了学习流行病学外，还紧跟计算机的发展，努力钻研计算机的应用。从一个机盲开始，一鼓作气听了 10 门课，其中“数据处理”，每周要交一份计算机程序的作业报告：从基本的作图程序，到用逻辑语言编写国际象棋下法等应有尽有。由于基础较差，他除了认真听课外，还经常请教周围高手。他们手把手地教他写程序，制图形，白天排队打孔、输入，晚上计算，隔天才能拿到结果。后来俞教授所在的加拿大肿瘤研究所流行病学室有了自己的微机系统，虽然容量小、速度慢，但每人桌上一台终端机，计算起来非常方便。为了学好计算机，在那里他用计算机完成了所有与流行病学、统计学有关的计算作业。他还进行了肝癌趋势面分析、营养流行病学研究，用记录联动分析了加拿大铁路工种与恶性肿瘤的关系等。

1982 年俞教授回校后带一批毕业班同学去启东实习。学生用我们自编的《Casio180 程序计算器在流行病学上的应用》在现场进行了统计分析。现场回来后再用苹果机做了肝癌标化死亡率比(*SMR*)等的分析。随后我们建立了教学机房，编写了《计算机在流行病学上的应用》讲义，开设了有关课程。1985 年俞教授再次到加拿大肿瘤研究所进修时，盲人教授 Dr. Howe 不但帮助他完成了乳腺癌的研究课题，还无私地赠送他自己用广义线性模型编写的源程序。回国后俞教授带领研究生进一步发展了 Meta 分析、logistic 程序、交互作用评估等模型。

1990 年第 5 版 Epi Info 疾病数据管理程序出版。征得著者同意后，俞国培等共同努力翻译了这本书。在加拿大国际开发署支持下，与新加坡大学合作，举办了计算机在流行病学上应用学习班。进入 21 世纪信息时代后，网络和计算机的应用更加广泛。姜庆五院长于 2001 年赴美国加州 UC Berkley 学习了地理信息系统、遥感分析等。回国后开展了科研和教学工作，用于血吸虫病监测和太湖水面藻类污染的评估。我院流行病学教研组各位教授、教师都根据自己所长，撰写了有关章节，尤其是赵琦和张涛同志自始至终参加编写，在这里对教研组同志的辛勤劳动表示衷心的感谢。最近又征得 CPEPI 作者 Gablinger 同意，才得以免除将书译成中文出版所需费用。在这里表示感谢。本书中生存分析、图形处理、SAS 程序应用等章节由海外同道们协助完成。有些章节还得到南京医科大学、东南大学医学院、福建医科大学等同道的协助，在此一并表示感谢。

计算机在流行病学上的应用，主要是数据处理、统计分析、文字处理、图形处理、信息交流以及数学模拟和数理模型，再加上近年来发展的分子流行病学等，是每个流行病学工作者必学

的技能。我们力求把书编得全面一些，实用一些，既适合于教师、研究生、医疗预防工作者和学生阅读参考，又可供疾病控制预防工作者和医院管理工作者等使用。但是，由于我们并非计算机专业出身，错误和遗漏一定不少，希望大家指正和赐教。

本书是在上海医科大学与复旦大学合并时开始编写的，最后在复旦大学公共卫生学院和研究生院支持下定稿。在此过程中，受到校领导的关心和帮助，在这里谨向他们表示感谢。我们希望通过总结过去，展望未来，把流行病学与计算机应用这门课程办得更好，应用得更加广泛，紧紧跟上时代前进的步伐。

复旦大学公共卫生学院

俞顺章　姜庆五

2011 年 3 月 1 日

目　录

第二篇　实　　践

第三篇　程　　序

第一篇 总 论

第一章　计算机在流行病学中的应用

流行病学是研究疾病在人群中分布、生态和对策的一门科学。一方面，流行病学要关注传染病的新发和再发，进行传染病防治；另一方面，又要开展非传染病病因研究和疾病预防。流行病学除传统的传染病流行病学外，分支学科已愈来愈多，如血清流行病学、分子流行病学、遗传流行病学等；有时根据对象不同再分为职业流行病学、营养流行病学、社会流行病学等。流行病学从宏观发展到微观，从研究具体疾病的流行规律到参与政策的制定，防灾反恐等，这一切都离不开计算机的应用。流行病学发展到今天已经比较成熟，它有明确的定义、完整的体系，受到各界广泛应用，流行病学如果结合计算机的应用就相得益彰了。计算机的应用，经历了以下几个历史阶段(图 1.1)。

- 20 世纪 50 年代前：算盘、手摇和电动计数机、计算器。
- 20 世纪 60—70 年代：程序计算器，大型计算机如 IBM360 等，用打孔机输入程序并进行统计分析。
- 20 世纪 80 年代：个人计算机发展。从台式到笔记本式、掌式等。
- 在计算机语言上有 BASIC，FORTRAN，PASCAL，C+。
- 20 世纪 70 年代以自编程序为主。
- 20 世纪 80 年代已编制许多商用计算机程序包(package)。用先进的统计方法处理数据。
- 20 世纪 90 年代数据输入、统计处理、图形构建、信息交流等全部由计算机协助完成。

图 1.1　计算机的发展和应用历史

在非传染病的病因方面流行病学方法的应用更加广泛，并已取得了许多突出的成就。由于非传染病潜隐期长、发病率或死亡率低，因此观察的对象数量要大、时间要长。又由于影响非传染病的病因非常复杂，因素之间相互制约、互相影响，因而许多精确的、多变量分析的方法应运而生。这些工作更离不开计算机这个工具。

近年来还由于计算机硬件和软件的广泛应用，CPU 运行的速度愈来愈快，硬件和软件已愈来愈多，贮存量愈来愈大，这些都大大促进了计算机在流行病学上的应用。计算机在流行病学中主要用于：数据处理(data processing)、文字处理(word processing)、图形处理(graphical processing)、统计处理(statistical processing)、数学模型(mathematical model)、分子流行病学(molecular epidemiology)、信息交流(communication)等方面。

第一节　数 据 处 理

由于流行病学的数据众多(表 1.1)，因此要输入、贮存、插入、删除、修改、查询、检索以及进行简单统计等。在救灾、反恐、疾病暴发中，计算机协助进行信息直报、贮存、调查表设计、数据录入、图形显示，甚至各种组织和措施实施步骤规范等。当今世界各种数据和文件的输入、输出也都离不开计算机，其流程见图 1.2。

表 1.1 计算机处理流行病学各种变量举例及统计方法

变量类型	计算机操作	举 例	统计方法
名词性	分类描述	民族,宗教,地区	棒图,饼图,频数,构成比等
顺序性	最小,最大,分等级	社会经济阶层,病人状态	频数,构成比,平均得分等
间歇性	按不同水平分组	温度,日期	连续直方图,均数,标准差等
比例性	按不同比例分组	体质指数,身高,体重	均数,几何均数,变异系数

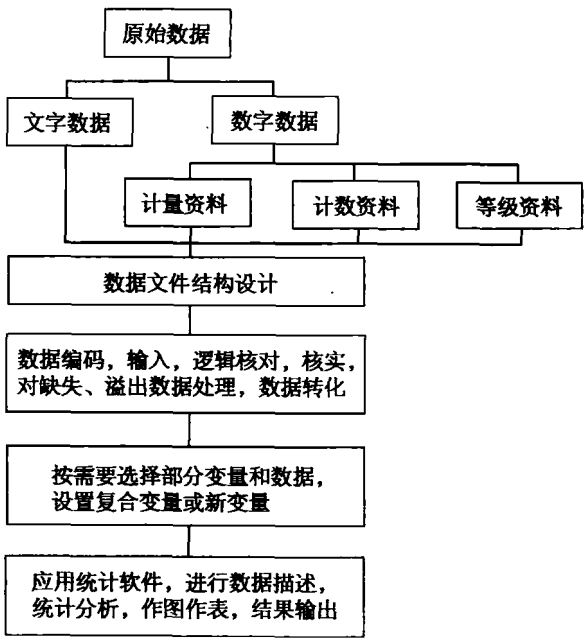

图 1.2 数据处理流程

1990 年由美国疾病预防控制中心和世界卫生组织 AIDS 全球计划部合作编写的Epi Info软件即疾病数据处理软件,可以将流行病学调查表或问卷数据直接输入,并对输入的数据进行逻辑检查、统计分析、图表处理、数据交换、地图显示、信息通讯和文字总结等;还可进行logistic 回归、Kaplan－Meier 生存分析、同份文件两人输入数据比较等;并且有各种接口或转变软件可与 Microsoft Access, Visual Basic, SPSS 和 SAS 等软件对接。在 Epi Info 基础上发展的 Epi Data 数据处理软件,它的中文输入等功能还优于 Epi Info。

数据处理软件首先用于数据输入,输入最普通的方法是由键盘输入,亦可用光电笔或扫描输入,或用固定表格或涂黑标记后自动输入。其次,可进行数据编辑、逻辑校对,例如,男性不应该生育和不应该发生女性特有的肿瘤如宫颈癌、卵巢癌,儿童很少发生冠心病等。对缺失值应该在分析时剔除或用内插程序插入。再次,要检查输入数据是否正态分布,如遇非正态分布数据可用对数或平方根转化成正态后分析,必要时要对年龄重新分组(如每 5 岁一组)等。最后,为了统计分析,数据以统计软件包所要求格式进行输出。

一个好的数据输入,首先要有一个设计好的调查表。一般调查表应包括的基本内容:①标志符,识别符(姓名、登记号、标本号、地址、电话等);②社会人口学信息(年龄、性别、民族等);③结局(疾病或伤害);④暴露因素;⑤可能发生的混杂因素等。关于调查表的格式现以对吸烟史调查为例表述如下(表 1.2)。

表 1.2　吸烟史调查表示例

吸烟史调查

① 您曾吸过烟吗？吸过_____。如果是的话，现在戒烟了吗？_____	□
② 最常吸的纸烟牌号：中华_____牡丹_____红双喜_____其他_____	□
③ 一天吸多少支烟？_____支	□□
④ 吸入深度：深_____，一般_____，不吸入_____；烟吸入到口腔_____咽_____肺_____	□
⑤ 您初次吸烟时年龄：_____岁	□□
⑥ 如果您现在不吸烟，有几年不吸了？_____年	□□
⑦ 回忆过去吸烟时有几年偶然吸？_____年	□□
经常吸烟的量和年数？<5 支/d_____年	□□
5—20 支/d_____年	□□
>20 支/d_____年	□□
⑧ 共同生活中有谁吸烟？夫妻_____子女_____其他_____无_____	□
⑨ 工作环境中，周围有无吸烟者？有_____无_____	□

近年来国内外都在开发疫情报告、慢性病监测（恶性肿瘤报告、糖尿病监测、职业病监测等）以及传染病诊断、细菌学鉴定等软件。例如，以色列 Berger（2001）发表医学和急诊疾病现场信息诊断模拟软件（Global Infectious Diseases and Epidemiology Network，GIDEON）对 300 多种传染病，250 多种药物和疫苗，1 500 种细菌、病毒、寄生虫和真菌进行监测。内容分 3 个部分：第一部分对疾病症状、征候、实验室检查、原发地区和潜伏期进行 Bayes 鉴别诊断；该程序曾用于 495 例病人，敏感度达 94.7%，特异度为 75%。第二部分进行流行病学分析，阐明疾病对人群的影响，过去和现在暴发的情况以及与该疾病有关的病原、媒介、宿主、时间和地区信息。第三部分是有关药物和标准检测方法等。

第二节　统 计 处 理

统计方法选择要看反应变量是单变量、双变量还是多变量；要看是计量资料、计数资料还是等级资料（有时多种情况并存）；要判断资料的类型，看适合于单因素分析还是多因素分析；看样本是单样本、双样本或是多样本；看是否是配对设计或成组设计（病例对照研究）；看是否已经满足各不同检验方法所需的前提条件（如正态分布、方差齐性等）。

如是计量资料可用均数±标准差，分年龄、性别、地区和时间等；如是计数资料，可进行率和比例的计算，并按三间分布特征进行描述。特别应该注意：样本的代表性（注意抽样方法和应答率），率还要作标准化处理（注意不能以比代替率）。在进行统计推断和假设检验时，分组比较要特别注意各组间的可比性。计量资料的均数比较可采用 t 检验（两样本比较）、方差分析（多样本比较），必要时可采用协方差分析或线性回归模型处理。对一些比的数据如相对危险度，应用卡方检验、趋势检验和比数比统计，进行 logistic 回归分析或采用广义线性模型方法。对一些率的数据，应用卡方检验、趋势检验和分层率的总结，进行 Poisson 回归分析或采用生存模型方法。对时间序列数据，可采用生存分析和时序检验，进行比数风险模型分析；对配对数据进行 McNeman 卡方检验和条件 logistic 回归（详见 http://gisserver.yale.edu/holford/example）。统计处理的步骤较多，一般可分：数据描述、联系确定、分层分析、多因素分析及深入分析等步骤（图 1.3）。联系确定后，不同的调查方法对病因的确定，其贡献大小是不同的，其顺序可见图 1.4。