国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

“十三五”国家重点图书出版规划项目

Precision Medicine

精准医学出版工程

精准医学基础系列

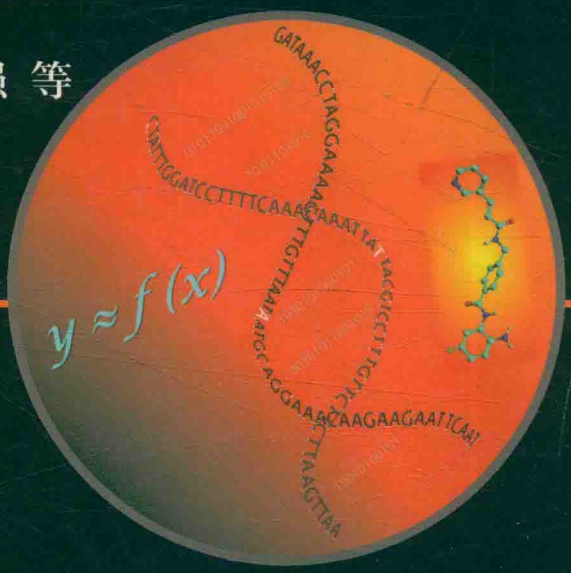
总主编 詹启敏

大数据与精准医学

Big Data and Precision Medicine

石乐明 郑媛婷 苏振强 等

编著



上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

Precision Medicine

精准医学出版工程

精准医学基础系列

总主编 詹启敏

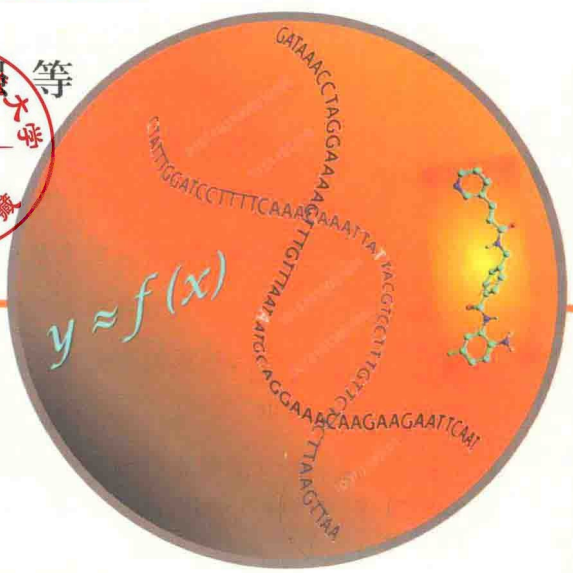
“十三五”国家重点图书出版规划项目

大数据与精准医学

Big Data and Precision Medicine

石乐明 郑媛婷 蒋振强 等

编著



上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

精准医学是以生物医学特别是多组学数据为基础，根据患者个体在基因型、表型、环境和生活方式等各方面的特异性，制订个性化的精准预防、精准诊断和精准治疗方案，是大数据最具应用前景的领域之一，同时也体现了临床实践发展的新方向。

本书围绕“生物大数据到精准医学实现”的全过程，介绍了生物大数据的基本概念、大数据研究中的共性方法论、健康人群队列研究、临床大数据及其标准化、组学大数据及其标准化、大数据的挖掘与融合分析、精准医学知识库及临床决策支持系统等关键性技术方法。同时还介绍了生物大数据在精准医学研究中的成功应用实例，包括遗传病与精准医学、药物基因组学与精准用药、基于组学大数据的肿瘤精准医学、HLA 基因多态性与药物不良反应、基于大数据的新药研发、精准医学与美国 FDA 监管作用等，为从事大数据与精准医学研究的读者提供较为全面的参考。

图书在版编目(CIP)数据

大数据与精准医学/石乐明等编著. —上海：上海交通大学出版社，2017

精准医学出版工程

ISBN 978-7-313-18401-6

Ⅰ. ①大… Ⅱ. ①石… Ⅲ. ①医学—数据处理 Ⅳ. ①R319

中国版本图书馆 CIP 数据核字(2017)第 278708 号

大数据与精准医学

编　　著：石乐明　郑媛婷　苏振强等
出版发行：上海交通大学出版社
邮政编码：200030
出 版 人：谈　毅
印　　制：苏州市越洋印刷有限公司
开　　本：787mm×1092mm　1/16
字　　数：413 千字
版　　次：2017 年 12 月第 1 版
书　　号：ISBN 978-7-313-18401-6/R
定　　价：248.00 元

地　　址：上海市番禺路 951 号
电　　话：021-64071208
经　　销：全国新华书店
印　　张：24.5
印　　次：2017 年 12 月第 1 次印刷

编 委 会

《大数据与精准医学》
编委会

主　编

石乐明(复旦大学生命科学学院、复旦大学附属肿瘤医院教授)
郑媛婷(复旦大学生命科学学院副教授)
苏振强(美国汤森路透集团研究员)

副主编

郭　力(中国科学院过程工程研究所研究员,中国科学院大学教授)
李亦学(中国科学院上海生命科学研究院研究员)

编　委

陈兴栋(复旦大学泰州健康科学研究院研究员)
洪汇孝(美国食品药品监督管理局资深科学家)
刘　雷(复旦大学生物医学研究院教授)
楼敬伟(上海宝藤生物医药科技股份有限公司董事长)
鲁先平(深圳微芯生物科技有限责任公司 CEO)
罗　衡(美国 IBM 公司 Thomas J. Watson 研究中心助理研究员)
索　晨(复旦大学生命科学学院助理研究员)
童伟达(美国食品药品监督管理局研究员)
吴　杰(复旦大学计算机学院教授)
夏诏杰(中国科学院过程工程研究所副研究员)
徐　萍(中国科学院上海生命科学信息中心研究员)
郁　颖(复旦大学生命科学学院副研究员)
张国庆(中国科学院上海生命科学研究院研究员)

主编简介

石乐明，1964年出生。中国科学院过程工程研究所计算化学专业博士，现为复旦大学生命科学学院教授、复旦大学附属肿瘤医院教授。主要研究方向为药物基因组学、精准医学、医学大数据、生物信息学及化学信息学等。1985年毕业于湖南大学化学化工系，获理学学士；1988年毕业于中国科学技术大学近代化学系，获理学硕士；1991年毕业于中国科学院过程工程研究所，获计算化学专业博士，后留所任助理研究员（1991年）和副研究员（1993年）。1994年赴美国留学，先后在凯斯西储大学（研究助理）、美国国立卫生研究院肿瘤研究所（访问学者）、美国食品药品监督管理局（FDA）、美国家用（惠氏）及巴斯夫等机构和公司任（资深）研究科学家职务；2001年作为原创人之一加入深圳微芯生物科技有限责任公司，任信息学部主任，负责计算机辅助药物分子设计并创建了基于化学基因组学的创新药物研发和筛选平台；2003年作为资深研究员再次加入美国FDA，并被阿肯色医科大学聘为兼职研究教授。先后发起并领导了大型国际基因芯片和下一代测序质量控制（MAQC/SEQC）研究计划，相关研究成果由 *Nature* 出版集团于2006年、2010年和2014年以专辑发表，并建立了专门的网站予以重点推荐，美国FDA依此制定了相应的《药物基因组学指南》。国家特聘专家，入选国家第三批“千人计划”。发表学术论文200多篇（其中11篇发表于 *Nature Biotechnology*），SCI引用7 000多次，单篇最高引用1 700多次，应邀参与10多本英文

专著有关章节的撰写，在国际学术会议上做大会报告或特邀报告 100 余次。获 4 个创新药物化合物美国专利授权，2 个化合物已进入中国Ⅲ期临床试验，1 个进入美国和日本临床试验。参与研发的一个原创 1.1 类新药西达本胺（爱谱沙®）于 2014 年被中国国家食品药品监督管理总局（CFDA）批准上市，用于治疗复发及难治性外周 T 细胞淋巴瘤，1.1 类原创抗糖尿病新药西格列他钠的Ⅲ期临床试验即将完成，已于 2017 年申报上市。

精准医学出版工程

总序

“精准”是医学发展的客观追求和最终目标，也是公众对健康的必然需求。“精准医学”是生物技术、信息技术和多种前沿技术在医学临床实践的交汇融合应用，是医学科技发展的前沿方向，实施精准医学已经成为推动全民健康的国家发展战略。因此，发展精准医学，系统加强精准医学研究布局，对于我国重大疾病防控和促进全民健康，对于我国占据未来医学制高点及相关产业发展主导权，对于推动我国生命健康产业发展具有重要意义。

2015 年初，我国开始制定“精准医学”发展战略规划，并安排中央财政经费给予专项支持，这为我国加入全球医学发展浪潮、增强我国在医学前沿领域的研究实力、提升国家竞争力提供了巨大的驱动力。国家科技部在国家“十三五”规划期间启动了“精准医学研究”重点研发专项，以我国常见高发、危害重大的疾病及若干流行率相对较高的罕见病为切入点，将建立多层次精准医学知识库体系和生物医学大数据共享平台，形成重大疾病的风险评估、预测预警、早期筛查、分型分类、个体化治疗、疗效和安全性预测及监控等精准防诊治方案和临床决策系统，建设中国人群典型疾病精准医学临床方案的示范、应用和推广体系等。目前，精准医学已呈现快速和健康发展态势，极大地推动了我国卫生健康事业的发展。

精准医学几乎覆盖了所有医学门类，是一个复杂和综合的科技创新系统。为了迎接新形势下医学理论、技术和临床等方面的需求和挑战，迫切需要及时总结精准医学前沿研究成果，编著一套以“精准医学”为主题的丛书，从而助力我国精准医学的进程，带动医学科学整体发展，并能加快相关学科紧缺人才的培养和健康大产业的发展。

2015 年 6 月，上海交通大学出版社以此为契机，启动了“精准医学出版工程”系列图

书项目。这套丛书紧扣国家健康事业发展战略，配合精准医学快速发展的态势，拟出版一系列精准医学前沿领域的学术专著，这是一项非常适合国家精准医学发展时宜的事业。我本人作为精准医学国家规划制定的参与者，见证了我国精准医学的规划和发展，欣然接受上海交通大学出版社的邀请担任该丛书的总主编，希望为我国的精准医学发展及医学发展出一份力。出版社同时也邀请了刘彤华院士、贺福初院士、刘昌效院士、周宏灏院士、赵国屏院士、王红阳院士、曹雪涛院士、陈志南院士、陈润生院士、陈香美院士、金力院士、周琪院士、徐国良院士、董家鸿院士、卞修武院士、陆林院士、乔杰院士、黄荷凤院士等医学领域专家撰写专著、承担审校等工作，邀请的编委和撰写专家均为活跃在精准医学研究最前沿的、在各自领域有突出贡献的科学家、临床专家、生物信息学家，以确保这套“精准医学出版工程”丛书具有高品质和重大的社会价值，为我国的精准医学发展提供参考和智力支持。

编著这套丛书，一是总结整理国内外精准医学的重要成果及宝贵经验；二是更新医学知识体系，为精准医学科研与临床人员培养提供一套系统、全面的参考书，满足人才培养对教材的迫切需求；三是为精准医学实施提供有力的理论和技术支撑；四是将许多专家、教授、学者广博的学识见解和丰富的实践经验总结传承下来，旨在从系统性、完整性和实用性角度出发，把丰富的实践经验和实验室研究进一步理论化、科学化，形成具有我国特色的精准医学理论与实践相结合的知识体系。

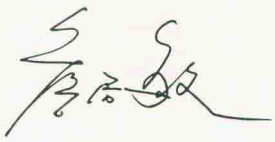
“精准医学出版工程”丛书是国内外第一套系统总结精准医学前沿性研究成果的系列专著，内容包括“精准医学基础”“精准预防”“精准诊断”“精准治疗”“精准医学药物研发”以及“精准医学的疾病诊疗共识、标准与指南”等多个系列，旨在服务于全生命周期、全人群、健康全过程的国家大健康战略。

预计这套丛书的总规模会达到 60 种以上。随着学科的发展，数量还会有所增加。这套丛书首先包括“精准医学基础系列”的 11 种图书，其中 1 种为总论。从精准医学覆盖的医学全过程链条考虑，这套丛书还将包括和预防医学、临床诊断（如分子诊断、分子影像、分子病理等）及治疗相关（如细胞治疗、生物治疗、靶向治疗、机器人、手术导航、内镜等）的内容，以及一些通过精准医学现代手段对传统治疗优化后的精准治疗。此外，这套丛书还包括药物研发，临床诊疗路径、标准、规范、指南等内容。“精准医学出版工程”将紧密结合国家“十三五”重大战略规划，聚焦“精准医学”目标，贯穿“十三五”始终，力求打造一个总体量超过 60 本的学术著作群，从而形成一个医学学术出版的高峰。

本套丛书得到国家出版基金资助，并入选了“十三五”国家重点图书出版规划项目，体现了国家对“精准医学”项目以及“精准医学出版工程”这套丛书的高度重视。这套丛书承担着记载与弘扬科技成就、积累和传播科技知识的使命，凝结了国内外精准医学领域专业人士的智慧和成果，具有较强的系统性、完整性、实用性和前瞻性，既可作为实际工作的指导用书，也可作为相关专业人员的学习参考用书。期望这套丛书能够有益于精准医学领域人才的培养，有益于精准医学的发展，有益于医学的发展。

此次集束出版的“精准医学基础系列”系统总结了我国精准医学基础研究各领域取得的前沿成果和突破，内容涵盖精准医学总论、生物样本库、基因组学、转录组学、蛋白质组学、表观遗传学、微生物组学、代谢组学、生物大数据、新技术等新兴领域和新兴学科，旨在为我国精准医学的发展和实施提供理论和科学依据，为培养和建设我国高水平的具有精准医学专业知识和先进理念的基础和临床人才队伍提供理论支撑。

希望这套丛书能在国家医学发展史上留下浓重的一笔！



北京大学副校长
北京大学医学部主任
中国工程院院士

2017 年 11 月 16 日

序

精准医学源自2011年美国国家科学研究委员会的报告《迈向精准医学：构建生物医学研究的知识网络和新的疾病分类法》，此后该理念受到重视，相关技术不断发展成熟。精准医学是指在大样本研究获得疾病分子机制的知识体系基础上，以生物医学，特别是组学数据为依据，根据“患者个体”在基因型、表型、环境和生活方式等各方面的特异性，应用现代遗传学、分子影像学、生物信息学和临床医学等方法与手段，制订个性化精准预防、精准诊断和精准治疗方案。

与传统医学不同，精准医学可以精准地优化诊疗效果，减少无效、有害和过度医疗，避免医疗资源浪费，降低医疗成本，优化医疗资源配置。而且，在精准医学研究证据的指导下，通过识别高危人群，有的放矢地进行针对性防控，将推动预防为主的健康医学发展，可以极大节约医疗费用。目前，精准医学的成功应用包括采用全基因组测序方法寻找罕见疾病的病因和治疗方案以及靶向药物在肿瘤临床治疗中的应用。

2015年1月20日，美国总统奥巴马在作国情咨文报告时提出精准医学计划（Precision Medicine Initiative，PMI），短期目标是癌症的研究与应用，长期目标是把精准医学推广到更多的疾病类型。我国于2016年3月8日正式启动了“精准医学研究”重点专项，并发布了首批项目指南。我国的精准医学研究计划充分调研和考虑了社会重大需求，确定了以我国常见高发、危害重大的疾病及若干发病率相对较高的罕见病为切入点，基于中国人群独特的遗传背景和环境多样性，实施针对我国人群的精准医学研究计划的战略。这将有利于建立中国自己的精准医学体系，避免前沿技术成果对外依赖和相关产业受制于人的局面。

当前，我国精准医学发展的挑战主要体现在缺少国家级关键共性技术平台。为解决这一瓶颈，“精准医学研究”专项将构建总量超过百万人级自然人群国家大型健康队列和特定疾病队列，建立多层次精准医学知识库体系和安全、稳定、可操作的生物医学大数据分析关键技术，建立创新性的大规模研发疾病预警、诊断、治疗与疗效评价的生物标志物、靶标、制剂的实验和分析技术体系。

精准医学更为重要的目标是解决现实需求的问题。通过“精准医学研究”专项的引导，形成重大疾病的风险评估、预测预警、早期筛查、分型分类、个体化医疗、疗效和安全性预测及监控等精准防诊治方案和临床决策系统，建成可用于精准医学应用全过程的生物医学大数据参考咨询、分析判断、快速计算和精准决策的系列分类应用技术平台，最终实现提升人口健康水平、减少无效和过度医疗、遏制医疗费用支出快速增长等目标。

本书的出版，及时而全面地介绍了精准医学实现过程中生物大数据的关键技术，包括大数据研究的共性方法论、大型健康队列数据的利用、生物大数据的标准与质量控制、生物大数据挖掘与融合分析、精准医学知识库的构建以及实现临床决策支持系统的关键技术和方法等，可以为从事精准医学研究的临床医师及科研人员提供生物大数据的方法学基础。其中，将生物大数据应用于精准医学的临床实践，数据的可靠性及其分析流程的标准化是必备的前提条件。特别值得一提的是，本书主编石乐明教授领导的国际组学质量控制联盟（MAQC），在组学数据的质量控制与标准化研究方面做出了具有国际影响力的研究成果。

此外，本书从迫切且可行的临床应用方面选取典型研究进行了介绍，如精准医学在遗传病和肿瘤精准诊疗中的研究和应用，为临床医师提供了研究范例和应用参考。同时，本书结合主编石乐明教授在国际制药工业界和美国食品药品监督管理局（FDA）任职十几年的研究工作经验，重点介绍了采用生物大数据方法进行创新药物研发与药物精准应用研究的实例与前景，以及美国 FDA 在药物基因组学及生物标志物相关研究中的监管作用，为精准医学时代的新药研发提供了前沿方向与实例参考。

精准医学体现了医学发展趋势，也代表了临床实践发展方向。本书的出版将推动我国生物大数据与精准医学研究和临床应用的发展。希望我国能发挥特有的举国体制优势，以及在健康人群和患病人群队列研究的规模上与生物样本多样性上的优势，同时

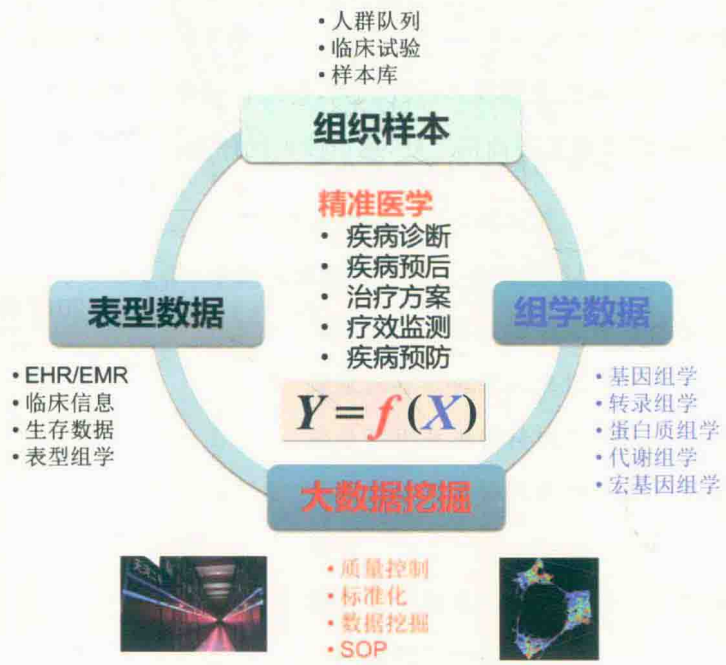
尽快弥补在核心共性技术与支撑性平台等方面的短板，在精准医学研究的国际竞争中，实现弯道超车，提升生命科学、生物医药、健康医疗等大健康产业的全产业链创新能力，驱动我国社会经济发展的转型升级，促进健康中国的建设。

复旦大学副校长
中国科学院院士

2017 年 11 月

前言

随着生物医学分析检测技术的飞速发展和相关领域研究的不断深入，健康及疾病档案，分子水平多组学指标，行为学、社会学及环境因素等多层次、多类型的生物学和医学数据呈海量式增加。如何有效地获取、融合和利用这些海量数据中隐含的信息，改进传统诊疗方式，为人类提供更好的医疗健康服务是当前生物大数据研究领域所面临的挑战。而如图所示，精准医学正是以生物医学大数据，特别是多组学大数据为基础，根据患者个体在基因型、表型、环境和生活方式等各方面的特异性，制订个性化的精准预



大数据与精准医学

防、精准诊断和精准治疗方案的全新医学模式，是生物大数据最具前景的应用领域之一，同时也体现了临床实践发展的崭新方向。

本书分为理论篇（第1～8章）和应用篇（第9～13章）。理论篇重点介绍生物大数据与精准医学研究中的共性方法和技术，包括生物大数据的基本概念与共性方法论、健康人群队列研究、临床表型数据及其标准化、组学大数据及其标准化、大数据的挖掘与融合分析、精准医学知识库的构建以及临床决策支持系统。应用篇介绍若干应用实例，包括遗传病、药物基因组学、肿瘤精准医学、基于大数据的新药研发等领域的最新研究进展，这是目前精准医学研究可以落地实施的切入点。同时本书（第14章）还介绍了美国FDA对精准医学相关的数据递交、生物标志物研究的监管，这是实现精准医学临床转化应用的重要环节。希望通过这些案例研究介绍，可以给从事大数据与精准医学研究的读者提供参考，促进在更广泛的领域进行精准医学研究和临床应用。

本书由复旦大学生命科学学院的石乐明教授团队主持编著，编写工作得到诸多科研院所、高等院校和临床医院的大力支持和帮助。衷心感谢我国“精准医学研究”重点专项指南编制专家组组长金力院士为本书作序！编写组由复旦大学、中国科学院上海生命科学研究院、中国科学院过程工程研究所、中国科学院上海生命科学信息中心、复旦大学泰州健康科学研究院、深圳微芯生物科技有限责任公司、美国FDA、美国汤森路透集团、美国IBM公司Thomas J. Watson研究中心等单位的专家组成。

本书引用了一些作者的论著及其研究成果，在此表示衷心的感谢！

书中如有疏漏、错谬或值得商榷之处，恳请读者批评指正。

编著者

2017年12月于上海

目录