大数据时代下的临床科研数据挖掘

郭有　主编

江西科学技术出版社

图书在版编目（CIP）数据

大数据时代下的临床科研数据挖掘 / 郭有主编. --
南昌：江西科学技术出版社，2022.10
ISBN 978-7-5390-8289-9

Ⅰ. ①大… Ⅱ. ①郭… Ⅲ. ①临床医学—科学研究—
数据采掘—研究 Ⅳ. ① R4-39

中国版本图书馆 CIP 数据核字（2022）第 148097 号

国际互联网（Internet）地址：
http://www.jxkjcbs.com
选题序号：ZK2022177

大数据时代下的临床科研数据挖掘 郭有 主编

出版发行	江西科学技术出版社
社址	江西省南昌市蓼洲街 2 号附 1 号 邮编：330009 电话：（0791）86623491 86639342（传真）
印刷	南昌市红星印刷有限公司
经销	全国各地新华书店
开本	720mm × 1000mm 1/16
字数	183 千字
印张	15.25
版次	2022 年 10 月第 1 版
印次	2022 年 10 月第 1 次印刷
印数	1—3040 册
书号	ISBN 978-7-5390-8289-9
定价	68.00 元

赣版权登字 -03-2022-215

主编简介

郭有　医学博士。副教授。现任赣南医学院第一附属医院医药大数据中心、赣南医学院创新团队、赣州市医药大数据重点实验室负责人，赣南医学院公共卫生、江西理工大学计算机科学与技术专业硕士生导师。中国生物工程学会计算生物与生物信息学专业委员会委员，中国研究型医院学会临床数据与样本资源专业委员会委员，江西省研究型医院学会生物样本库分会副主任委员，江西省医学会大数据与人工智能分会委员。主持国家自然科学基金、江西省重点研发计划、江西省自然科学基金等项目。

《大数据时代下的临床科研数据挖掘》编委会

◎序

张小康

医疗大数据时代的到来，特别是国家卫健委以电子病历和互联互通标准化测评为推手，使得国内医院信息化建设和临床科研数据中心建设速度大大提升，为临床科研数据的治理、集成、挖掘和利用提供了全新而广阔的舞台。同时，高通量生物芯片、二代测序、质谱等生物组学技术的快速成熟和使用成本的迅速下降，极大提升了临床生物样本高通量检测的普及速度、生物组学数据积累分享速度。随之，这也将数据挖掘推到了临床科研领域的中心地带，形成了医学研究模式的变革性动力。

临床数据挖掘流程复杂、多学科融合度最深，其核心任务是让数据产生临床价值。从这个角度看，临床科研数据挖掘包括三个必要环节：将临床科学问题定义为数据分析问题、数据处理与建模、模型评估与临床应用，这三个环节缺一不可。但是，现有相关图书主要关注的是第二个环节，阐释详细的数据处理与建模方法，而对第一个、第三个环节基本没有涉及。原因也很简单，第一个和第三个环节必须来自临床科研数据挖掘实践。如果不能把临床科学问题转换为可执行的数据分析问题，那么临床数据挖掘就成了无源之水、无本之木。只有把临床科学问题准确定义为数据分析问题，数据处理和建模才会有用武之地。而且，数据处理和模型必须经过临

床应用评估，否则无法证明数据挖掘的临床价值。

赣南医学院第一附属医院在国内率先成立了医药大数据中心。中心的同仁们长期专注于临床科研数据挖掘的研究与教学工作，是国内临床数据挖掘领域中的最早实践者、代表者和推广者，积累了宝贵的分析案例和实践经验。因此，从临床数据挖掘流程角度而言，临床科研数据挖掘的第一个、第三个环节更为重要，这也是《大数据时代下的临床科研数据挖掘》作者想和大家分享的核心理念。

站在分享角度，我很赞赏本书以临床数据挖掘实践为主线，以最贴近临床场景的数据挖掘案例为核心，围绕临床实践需求和高频数据挖掘方法，着重突出完整案例分析的多学科系统整合。同时，该书在系统回顾吸收国内外最佳临床数据挖掘实践案例的基础上，首次从六个方面系统地、深入地论述了临床数据挖掘中应该坚持的基本思想及其实践方法，凝聚了多年实践心血与宝贵经验。这是一本基于持续性改进的临床数据挖掘实践的参考书，也可以作为临床研究生培养的教科书，有助于提高对临床科研数据认知思维、处理与利用能力。

培养临床科研能力，首先体现在科研思维上。科研思维跟不上，则科研永远跟不上。学科交叉融合的广度和深度，使得临床数据挖掘能力培养实属不易，我深刻感受到赣南医学院第一附属医院探索性工作的意义和价值。期望更多的人能够读到这本《大数据时代下的临床科研数据挖掘》，为自己开启一个激情澎湃的临床研究新世界。

2022 年 8 月

（张小康系赣南医学院党委副书记、院长，赣南医学院第一附属医院党委书记）

目录

第一章 》总 论

第一节 医疗大数据的概念、来源与特点……002
一、医疗大数据的概念 ……002
二、医疗大数据的来源 ……002
三、临床表型数据的特点 ……004
四、生物组学数据的特点 ……004
五、代表性数据库简介 ……005
六、赣南医学院第一附属医院临床数据与样本资源库 ……006
第二节 临床数据挖掘的概念、特点……006
一、数据挖掘的概念 ……006
二、数据挖掘的目标 ……007
三、数据挖掘对临床研究的变革 ……007
四、临床数据挖掘的特点 ……009
第三节 临床数据挖掘的意义……010
一、揭示人群疾病负担分布和发展趋势 ……010

二、揭示发病机制和药物作用的分子互作网络 ……011
三、制订医学参考值范围 ……011
四、提供疾病精准诊断和预后预测模型 ……012
五、药效结构预测与新药发现 ……013
六、发现联合用药方案 ……014
七、医院运行管理 ……014
第四节 临床数据挖掘的基本流程……015
一、定义临床问题 ……015
二、数据准备与数据提取 ……016
三、数据预处理 ……017
四、特征提取 ……018
五、建立预测建模与模型评估 ……018
六、分析报告与结果发表 ……019
第五节 临床数据挖掘的基本思想……020
一、临床应用场景为导向 ……020
二、临床问题背后的多学科问题 ……022
三、控制多重检验的假阳性发现率 ……023
四、亚组分析 ……023
五、分层分析 ……024
六、对照选择 ……026
第六节 数据挖掘工具R软件 ……028
一、R简介……028
二、R软件安装……029
第七节 临床数据挖掘研究的展望……034
一、临床专病数据库建设与利用 ……034
二、临床数据挖掘面临的挑战 ……034

三、临床数据挖掘的研究方向 ……036

第二章 》临床数据挖掘中常用的临床表型数据库

第一节 常用临床表型数据库……040
一、SEER数据库 ……040
二、MIMIC数据库 ……041
三、Biobank数据库 ……042
四、BioLINCC数据库 ……043
五、TARGET数据库 ……044
第二节 SEER数据库 ……045
一、获取SEER数据的方式 ……045
二、利用SEER · Stat软件获取数据 ……045
第三节 SEER数据挖掘研究 ……053

第三章 》临床数据挖掘中常用生物组学数据库

第一节 常用多组学数据库……057
一、GEO数据库 ……057
二、TCGA数据库……062
三、其他常见分子数据库 ……064
第二节 差异表达基因……065
一、GEO2R分析高通量芯片数据 ……065
二、edgeR分析RNA测序数据 ……068
三、TCGA数据线上分析工具……073
第三节 通路富集分析……074

一、KEGG数据库……074
二、GO数据库 ……075
三、软件包clusterProfiler的富集分析功能 ……076

第四章 》临床科研数据挖掘中的相关分析

第一节 两个特征变量间的相关分析……085
一、连续性变量间的相关分析 ……085
二、无序分类变量间的相关分析 ……092
三、有序分类变量间的相关分析 ……098
第二节 两组临床特征变量的相关分析……101
一、典型相关分析的概念 ……101
二、R软件操作……102
第三节 多个特征变量间的相关分析……107
一、概念 ……107
二、相关系数矩阵 ……107
三、相关系数矩阵图 ……109
第四节 单个特征变量与一组特征变量间的复相关……114
一、复相关概念 ……114
二、建立预测模型 ……115
三、评估临床特征对预测模型贡献的独立性 ……120
四、评估分子特征对预测模型贡献的独立性 ……122

第五章 》临床数据挖掘中的聚类分析

第一节 聚类分析中的距离与相似度……127

第二节 层次聚类……128
第三节 划分聚类……130
第四节 聚类分析步骤……131
一、特征变量选择 ……131
二、特征变量处理 ……131
三、聚类方法选择及距离计算 ……132
四、聚类结果解读和验证 ……132
第五节 聚类分析的R实践 ……132
一、层次聚类 ……133
二、K-means聚类……138
三、层次聚类与K-means聚类结果的对比……141
第六节 聚类分析在结直肠癌分子分型中的应用……143

第六章 》 倾向评分法控制临床数据挖掘中的混杂因素

第一节 概念与原理……149
一、倾向评分法的概念 ……149
二、倾向评分匹配的基本原理 ……149
三、倾向评分匹配法的特点 ……150
四、倾向评分法的其他类型 ……151
第二节 倾向评分匹配过程……152
一、常用的匹配方法 ……152
二、倾向评分匹配步骤 ……153
三、倾向评分匹配适用情况 ……154
第三节 PSM的R实践 ……155
一、安装加载MatchIt包 ……155

二、倾向评分匹配函数“matchit（）” ……155
三、倾向性评分逆概率加权法 ……161
四、PSM注意事项 ……162
第四节 倾向评分法在临床数据挖掘中的应用……162

第七章 》临床数据挖掘中的生存分析

第一节 基本概念与原理……166
一、生存分析中的基本概念 ……167
二、生存数据的统计学分析方法 ……169
三、Cox比例风险回归模型……170
四、比例风险假定的检验 ……171
第二节 生存分析的R实践 ……172
一、安装加载survival包 ……172
二、函数“survfit（）”估计总体生存率 ……174
三、函数“survdiff（）”比较总体生存率 ……176
四、利用函数“coxph（）”进行单因素分析 ……178
五、利用函数“coxph（）”进行多因素分析 ……180
六、“cox.zph（）”评估模型比例风险假定 ……181
第三节 肿瘤研究中生存分析实例……182

第八章 》诺莫图在构建预测模型中的应用

第一节 诺莫图基本原理……187
一、基本原理 ……187
二、诺模图的类型 ……187

三、诺模图的要素 ……187
第二节 构建诺莫图预测模型流程……188
一、定义患者群体 ……188
二、定义事件结局 ……189
三、识别临床结局相关的特征变量 ……189
四、构建诺莫图 ……189
第三节 诺莫图的R实践 ……197
一、安装加载rms软件包 ……197
二、数据准备 ……197
三、数据转换 ……198
四、构建模型 ……199
第四节 临床应用诺莫图的注意事项……200
一、重视风险预测值的置信区间 ……200
二、重视患者临床特征变量的差异 ……201
第五节 诺莫图在数据挖掘中的应用……202
一、在SEER数据库中的研究应用 ……202
二、TCGA数据库中的研究应用……203
三、门静脉血栓预测模型研究 ……204

第九章 》 ROC曲线与临床决策曲线

第一节 ROC曲线原理 ……208
一、基本原理 ……208
二、ROC曲线比较方法和使用条件 ……208
第二节 ROC曲线下面积 ……209
一、曲线下面积（AUC） ……209

二、AUC的分布范围 ……210
三、约登指数 ……210
第三节 R绘制ROC曲线 ……211
一、安装和加载pROC软件包 ……211
二、数据准备 ……211
三、拟合ROC曲线 ……212
四、绘制ROC曲线 ……213
五、绘制多条ROC曲线 ……214
第四节 ROC在临床数据挖掘研究中的应用 ……216
一、在公共数据库中的研究应用 ……216
二、在临床数据中的研究应用 ……218
第五节 临床决策曲线 ……219
一、基本原理 ……219
二、R绘制临床决策曲线 ……220
三、临床决策曲线应用 ……223
第六节 ROC曲线与临床决策曲线的比较 ……226

第一章

总　论

科研能力是医学特别是临床医学领域从业人员职业特质的核心要素。跨越了仅依赖经验、理论、假设和价值观去探索医学世界边缘的“无数据时代”，进入随机抽样靶向收集数据验证科学假说的“抽样数据时代”后，临床从业人员及研究生科研能力的培养，是在抽样验证思维的指导下，培养掌握观察性研究、随机对照研究以及队列研究等为方法核心的科研能力。但抽样验证研究具有一定局限性，不仅难以平衡样本代表性与可行性之间的矛盾，而且对人力、物力、时间有特定要求。

信息技术与生物技术的极速发展，以前所未有的加速度，提升了包括临床疾病表型数据与生物组学数据在内的医疗数据的产生与采集效率，革命性地推动临床医学进入了大数据时代。借助系统整体思维，培养临床科研人员和临床研究生认识、采集、整理、提取、挖掘和利用医疗大数据，可在一定程度上克服抽样研究的局限性，探索性或验证性地分析临床疾病表型与生物分子网络间的相互关系，对疾病的发生、发展和转归进行解释

和预测，从而提升临床实践效率与水平。

第一节　医疗大数据的概念、来源与特点

大数据时代预言家、奥地利著名数据科学家维克托·迈尔·舍恩伯格指出，大数据是当代社会独有的一种革命性动力，它以一种前所未有的方式采集、挖掘、利用海量数据，提供价值巨大的数据资源和技术服务。因为数据组成与规模超出了传统数据库软件工具的获取、存储、管理和分析能力，所以称之为大数据。除了被赋予规模性（Volume）和多样性（Variety）特征，大数据产生与变化速度更快（Velocity）、使用价值更高（Value），组成了广泛认可的 4V 特征。

一、医疗大数据的概念

随着信息化技术在临床实践中的快速扩散与深度融合，医疗数据电子化记录存储更加广泛而深入，特别是电子病历收集保存了大量患者的诊疗数据，构成了可供临床研究的兼有资源性与平台性的医疗大数据。

在狭义概念上，医疗大数据仅包含临床电子病历在内的诸多诊疗业务信息系统记录的医患双方活动信息数据。在广义概念上，医疗大数据还纳入了检测在院患者生物样本产生的生物组学数据，如基因组学、转录组学、蛋白质组学、代谢组学、修饰组学等多维度生物组学数据。这些组学数据与患者临床表型特征存在着强弱不等的关联关系。

二、医疗大数据的来源

医疗大数据来源于人从胚胎到死亡的各个生命周期。根据数据产生场景，医疗大数据主要来自以下几个途径。

医疗卫生机构是医疗大数据的最大的、最主要的来源。在国家政策与信息技术的多年推动下，特别是国家卫健委从数据资源、互联互通、基

础设施、应用四个方面标准化，对医院信息平台进行综合测试评估，使我国医院信息化建设水平迅速达到了前所未有的广度和深度。其中，医院信息系统（Hospital information system，HIS）成为医疗大数据的主要来源，包括电子病例系统（Electronic medical record system，EMRS）、实验室信息系统（Laboratory information system，LIS）、医学影像存档通信系统（Picture archiving & communication system，PACS）、放射治疗信息管理系统（Radiology information system，RIS）、临床决策支持信息系统（Clinical decision support system，CDSS）等。

临床患者生物样本检测产生的生物组学数据是医疗大数据的第二大来源。人类基因组碱基对约为 3 Gigabyte，考虑基因组的人群多态性，叠加生物功能的多分子参与、多水平调节、多层面修饰，使得临床生物样本检测产生的生物组学数据量非常庞大。随着生物组学数据采集技术的进步迭代和检测价格迅速下降，可以预见在未来一段时间里，生命组学检测技术将在临床实践中的使用更加广泛，因此而产生的生物组学数据将爆炸式地膨胀、积累。

医疗器械与制药企业研发活动产生的数据也达到了相当可观的程度。医疗器械与药物研发过程复杂程度高，其间需要开展大量的临床试验，中小型企业的研发数据一般可达 Trillionbyte 级，而大型企业数据可达到 Petabyte 级。

此外，愈来愈多的各型穿戴设备，可以实时记录医疗健康相关的时间序列数据，也逐步走进数据挖掘研究者的视野之中。穿戴设备实时采集数据、物联网短距离传输数据、移动互联网远距离传输数据、“云端”存储、分析、实时原路反馈的健康大数据新模式，使得健康生命体征如心电、血氧、呼吸、血压、体温、脉搏、运动等数据具备了巨大的研究价值。