工业和信息化普通高等教育“十三五”规划教材

医药信息处理与分析

Medical Information Processing and Analysis

晏峻峰 占艳 主编

- 阐述医药科研活动中信息科学的基本方法、技术和工具
- 医药信息采集与预处理—分析方法—MATLAB工具—案例
- 培养学生的医药信息获取、分析与利用能力

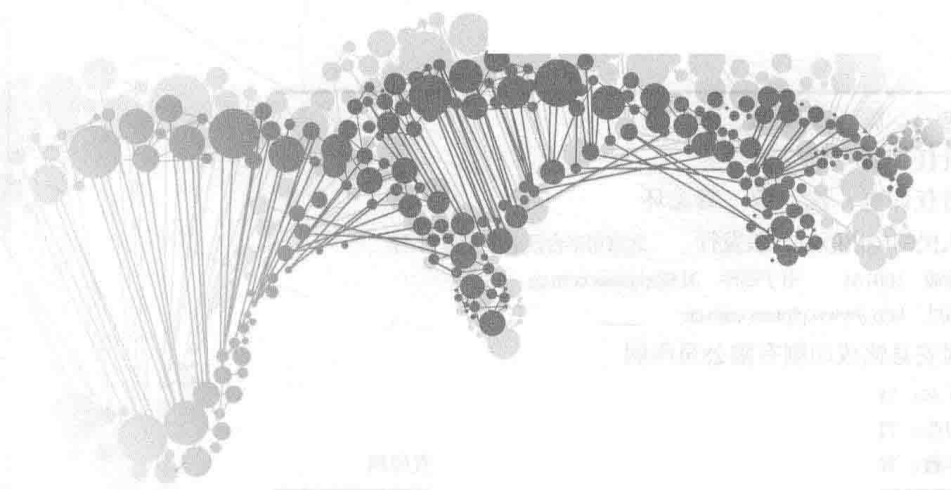
中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

工业和信息化普通高等教育“十三五”规划教材

医药信息处理与分析

Medical Information Processing and Analysis

晏峻峰 占艳 主编



高校系列

人民邮电出版社
北京

图书在版编目（CIP）数据

医药信息处理与分析 / 晏峻峰，占艳主编. -- 北京：人民邮电出版社，2018.3
ISBN 978-7-115-47821-4

Ⅰ. ①医… Ⅱ. ①晏… ②占… Ⅲ. ①医药学－情报检索－高等学校－教材 Ⅳ. ①G252.7

中国版本图书馆CIP数据核字(2018)第016309号

内 容 提 要

本书是面向高等医药院校开展信息处理与分析教学活动设计的教材。本书以培养学生的信息获取能力、分析与利用信息的能力为目标，全面阐述了医药科研活动中运用到的信息科学基本方法、技术和工具。全书共分为 5 章，主要包括信息处理与分析的内涵及相关技术、医药信息处理与分析的基本流程、医药信息分析质量保证的几个关键因素、医药信息采集与预处理、医药信息分析方法和工具、医药信息处理与分析案例和医药信息标准。

本书可作为高等医药院校本科、研究生的信息处理与分析基础教材，也可作为医药卫生领域科技人员开展医药信息处理与分析相关工作的参考书。

◆ 主　　编　晏峻峰　占　艳
责任编辑　邹文波
责任印制　沈　蓉　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：12.25　　2018 年 3 月第 1 版
字数：301 千字　　2018 年 3 月河北第 1 次印刷

定价：45.00 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

本书编委会

主　编：晏峻峰　占　艳

副主编：韦昌法　刘青萍　辛国江　涂　珊

编　委：穆　珺　吴世雯　晏峻峰　占　艳

韦昌法　刘青萍　辛国江　涂　珊

刘东波　王林峰　张　蕾　陈陵芳

前言 PREFACE

“医药信息处理与分析”是一门提升医药类学生科研能力的课程，该课程立足培养学生在医药科研活动中运用信息科学基本方法、技术与工具的能力，其目的是加深学生对信息技术相关工具的了解与掌握，使学生对信息、数据的采集、加工变得更加熟练；促进他们对分析方法的学习与研究，使学生能对已处理的信息、数据进行全面解读，以获取更深层次的知识，开展创新研究、辅助决策等工作。

全书共5章。第1章是绪论，包括信息的概念及相关知识、信息处理与分析的内涵及相关技术、医药信息与医药信息学、医药信息处理与分析的基本流程、医药信息分析质量保证的几个关键因素；第2章是医药信息采集与预处理，包括医药信息资源的获取与处理、数据预处理的基本方法、医药信息管理及相关技术；第3章是医药信息分析方法和工具，包括医药信息分析方法、常用信息分析工具、MATLAB工具；第4章是医药信息处理与分析案例，包括多元线性回归分析、主成分分析、聚类分析、判别分析、决策树分析、支持向量机算法、贝叶斯分类算法、时间序列分析；第5章是医药信息标准，包括常用医药信息标准介绍、医学信息分类与编码、医药信息处理与分析中应用标准的原则。

本书主要体现了以下特点。

（1）注重领域特色、组织结构合理

本书以医药信息为主线来组织素材，将信息处理技术与信息分析方法有机融合。全书采用总分式结构进行阐述：绪论中呈现了信息处理与分析的基本流程，使学生从宏观上了解信息处理与分析的每一步骤，帮助学生掌握信息处理与分析的每一阶段的关键技术与方法；紧紧围绕绪论，其他章节从内容上将信息处理与分析的每一环节有序贯通，使学生在学习时对各章内容一目了然，做到有的放矢，达到事半功倍的学习效果。

（2）选材经典、注重理论与实践的结合

本书的素材来自多年的教学积淀，选取了经典、有代表性的信息处理与分析的技术、工具与方法，列举了大量医药信息处理案例的分析，同时每章还配有难度适中的练习，使得理论知识与实践操作有机地结合在一起，做到真正意义上的学以致用。

（3）引导思维、促进学科交叉

本书通过介绍医药信息处理技术以及信息分析方法和工具，使学生能够从信息科学的角度对医药科学研究工作中的问题进行思考与分析，以解决跨学科交流、合作带来的诸多问题，从而促进多学科交叉合作的协同研究。

本书内容体系经晏峻峰与占艳多次组织教学团队研讨后制定。全书编写分工如下，第1章由晏峻峰编写，第2章由占艳编写，第3章及第4章4.8节由辛国江编写，第4章的4.1、4.2、4.3、4.4节由韦昌法编写（4.3.3小节由涂珊编写），第4章的4.5节由涂珊编写，第4章的4.6、4.7节由穆珺编写，第5章由刘青萍编写，吴世雯、刘东波、张蕾、王林峰老师以及研究生陈陵芳参与了整书的审阅与统稿工作。另外，在编写的过程，编写人员参阅了本书所列的参考文献，在此，编者对文献作者表示衷心感谢。由于编者水平有限，书中难免存在不足之处，敬请读者批评指正。

编　者

2018年1月

目录 CONTENTS

01 第1章 绪 论

信息同物质和能源一样，是人们赖以生存和发展的重要资源。人类通过信息认识各种事物，借助信息的交流建立人与人之间的联系，相互协作，从而推动社会进步。通常意义下，医药信息处理偏重于信息技术相关工具的应用，目的是使信息的采集、加工、利用更方便、快捷。而医药信息分析则更着重于对已处理信息的综合与解读，对分析方法学、领域知识综合能力有更高要求，以获得更多对所关注信息的认识与理解、更全面精准地认识信息的内涵为主体目标。医药信息处理与分析是医药科技工作者开展创新研究的重要手段。加强相关思维方法、技能工具的学习、培养多学科交叉合作的协同研究能力已为业界共识。而理解信息的一般概念、本质及关联知识，是其学习的首要任务。

1.1 信息的概念及相关知识

1.1.1 信息的概念

信息是自然界和人类社会的一个重要范畴，也是客观存在的一种基本现象。一般意义下信息的概念是指消息、情报，是互相交流中要传递的某种内容，与知识、见闻、通知、情报、事实、数据等概念在某些场合中常会交叉互用。通信意义上的信息是指在通信的任何可逆的重新编码或翻译中那些保持不变的东西。在通信技术领域，1948年，香农在《贝尔系统技术杂志》上发表的“通信的数学理论”一文中提出“信息是使（通信系统中）不确定程度减小的量”。“不确定程度”指通信系统中未收到有关信源信号的状态，在收到信号后，系统的有序程度增加，不确定程度减少。情报科学领域常用数据（Data）、信息（Information）和知识（Knowledge）这一组概念来表达情报领域中传输的内容。美国的情报学家麦克唐纳认为：“信息是特定情况下评价未经评价的数据的东西，知识是在一般使用过程中评价这种数据的东西”。而现代哲学认为，信息是与物质和能量并存的一种自然现象，是物质和意识存在方式的表现，同时也是物质和意识的桥梁。宇宙间一切事物都处于相互联系、相互作用之中，信息就是对事物之间相互联系、相互作用的状态的描述；从微观世界到宏观世界，从无机世界到有机世界，从植物到动物，从机器到人，都能产生信息，也能接收信息。正因为如此，信息成为了许多学科的研究对象，不同领域的学者都从不同的角度来研究信息。

1.1.2 信息的基本特征

信息是不同于物质和能量的一种特殊的资源，它具有可存储性、可传递性、可加工性、共享性、时效性和可替代性六大基本特征。

（1）可存储性。信息借助载体可在一定条件下存储起来。信息的可存储性为信息的积累、加工和不同场合下的应用提供了可能。

（2）可传递性。传递是信息存在的基本状态之一，信息传递的基本要求是速度快。传递具有动态性和方向性特征。信息的传递依赖于物质媒介。信息的传递必然伴随着物质或能量的传递，并且须消耗一定的能量。传递的基本方式有物质的传递和能量的传递两大类。物质传递较为显见，如运输、交通等。而能量的传递则不易察觉，如阳光照耀，多米诺骨牌等。信息传递的方式是多种多样的。按照流向的不同，可以有单向传递、反馈传递和双向传递三种方式。按信息传递时信息量的集中程序不同，有集中和连续两种方式。按信息传递范围或与环境关系的不同，可有内部传递和外部传递两种方式。

（3）可加工性。信息可以通过一定的手段进行加工，如扩充、压缩、分解、综合、抽取、排序等。加工的方法和目的反映信息接收者获取和利用信息的特定需求。加工后的信息是反映信息源和接收者之间相互联系、相互作用的更为重要和更加规律化的因素。应当注意的是，信息的内容是语法、语义和语用三者的统一体。信息在加工过程中要注意保证上述三者的统一，以免造成信息的失真，即原始信息（加工前的信息）的有些内容丢失或被歪曲。

（4）共享性。一个信息源的信息可以为多个信息接收者享用。一般情况下增加享用者不会使原有享用者失去部分或全部信息。一些特殊的信息和特殊形式的信息在共享上存在明显的障碍，但并不影响信息共享性这一本质属性。有的信息涉及商业的、政治的、军事的秘密，扩大对这类信息的享有者范围，可能影响某些享用者对这类信息的利用，但不会改变信息本身的内容。这是信息不同于物质和能量的一个本质特征。共享性指接收者在获得全部的信息的同时而不会减少信息的信息量（指记忆信源，如文献等）。并且，数个接收者可以获得同一信源发出的同样的信息。

（5）时效性。信息的时效性表现是多种多样的。例如，信息的滞后性是表明客观事物总是在前，认识总是在后，人类获得信息总是滞后的。例如，各种星体信息，即使是以光速传播，在人类接收到时，也是滞后的，也正是这种滞后性，使人们可以了解到不同历史时期的星体特征。超前性是指人类在把握各种规律的前提下，能够对发展中的事物进行预测。此外，信息在一定时间内相对说来会变成过时的信息，尤其是经济信息的有效期非常短。例如，国际金融市场信息的时滞一般不应超过 6 小时，否则，过了时限的信息，再详尽也只是昨日的黄花，非但无用，反而会使人做出错误的决定。大多数经济活动过程都很短暂，如国际股票交易市场的变化往往发生在几秒钟内。但某些信息的时效性却表现在越古老的信息越有价值，如考古研究等。任何信息从信息源传播到接收者都要经过一定的时间。信息接收者所得到的与自己有关的信息源的状况的信息都是反映信息源已经出现的状况。时滞的大小与载体运动特性和通道的性质有关。（技术性强）信息的传输、加工与利用都必须考虑这种时滞效应，特别对于需要实时或及时处理与利用的信息，必须通过合理选用载体与通道来把这种时滞控制在允许的范围内。

（6）可替代性。人的任何行为，都可以概括为一个不断从外界获取信息，对信息进行处理，并在这个基础上，通过一定的物质和能量，对事物进行调整、控制和组织行动的过程。因此，信息具有替代性。它可以替代资本、劳动力或其他有形的物质。最简单的事实是把信息编成程序，输入计算机，

就可以在工厂、矿山、交通运输、商业、医疗乃至家庭等各个领域代替人的劳动。信息可以在不同的层次上，在不同的状态之间和不同的信号系统之间进行转换。如自然语言和机器语言的转化属于不同的层次之间的转换；不同状态：光电信号转换，电声转换；不同的信号系统：不同的语种，方言等。可替代性的作用是使得信息能以不同的方式存储和传递，使信息的处理有可能得以实现，也使得各种交流方式得以存在。

1.1.3 信息、数据与知识的关系

数据是事实或观察的结果，它是对客观事件的记录和可以鉴别的符号。数据不仅指狭义上的数字，还可以是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频、音频等，是客观事物的属性、数量、位置及其相互关系的抽象表示，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。例如，“0、1、2…”“心、肝、脾、肺、肾”“病人的病历记录、CT扫描片子”等都是数据。数据的表现形式还不能完全表达其内容，需要经过解释，数据和关于数据的解释是不可分的。例如，“75”是一个数据，可以是某人的成绩，也可以是其体重，还可以是汽车的速度。数据的解释是指对数据含义的说明，数据的含义称为数据的语义，数据与其语义是不可分的。数据都有其属性和客观值，如“年龄25岁”，其“年龄”是数据的属性名称，“25岁”是数据的客观值。因此说数据是对客观事物的属性、数量、位置及其相互关系的一种抽象的描述，数据是事物原始性状的记载，没有经过任何加工处理，数据是杂乱的，但它是真实的、可靠的，并且具有累积的价值。数据是信息的表达载体，信息是数据的内涵，是形与质的关系。知识是经人为组织的可理解的系统信息，与信息不属同一范畴。知识是相对的，因人而异，因时而异，因地而异，属于意识范畴，具有主观的特征。而信息是绝对的，是客观的。信息是构成知识的基本要素，没有信息就没有知识可言。信息可以通过信息仪器设备获取；而知识则必须通过学习获得。

信息与数据、知识密切相关，是在不同层次上对事物的认识。要从数据中获得认知，需要对掺有大量杂质的数据进行清洗，以形成干净的数据。干净的数据也意味着数据的质量高，干扰少，从数据中获取的信息价值高，所形成的认知更具应用意义。由图 1-1 数据、信息与知识的关系中可见，数据与信息、信息与知识均互为蕴含关系，能够有效开展信息分析。在开展医药相关科研活动中，经常会做一些诸如：数据统计、数据分析、数据挖掘等工作，本质上就是对数据、信息、知识之间关系的探索，均属于信息处理与分析，因此，本书中，没有特别声明时，对信息处理与数据处理、信息分析与知识发现概念不做严格区分。

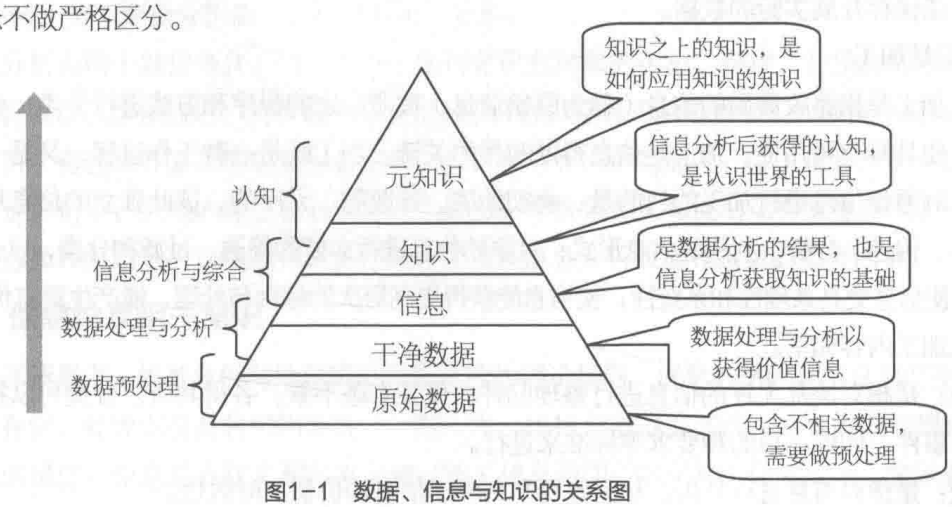


图1-1 数据、信息与知识的关系图

1.2 信息处理与分析的内涵及相关技术

信息的处理和分析是信息发挥作用的关键环节之一。在信息处理和分析中，“信息内核”非常关键。信息内核也称“特征信息”。在信息处理的过程中如何鉴别信息特征和保存信息特征是关键问题。即便是信息压缩与扩展都是要在保持“信息内核”的基础上进行的；而信息分析和鉴别正是通过诸多的表象来分析“信息内核”的过程。如漫画家可以“几笔”把一个人画出来，不管怎么美化或丑化，不管怎么极度夸张，就是画得很像，神似得很。为什么那么像？因为那“几笔”不是别的什么，就是拓扑学中的“特征不变量”，就是事物最本质的东西；画得“神似”，“神似”就是“特征不变量”这一信息内核的体现。再如，大家最常用的手机手写输入法，同一款手机，不同的人手写的字体千差万别，为何都能高效输入呢？这也是因为所安装的文字识别软件对每一个手写体文字的特征了如指掌。要做到这一点，手写体识别研究人员，就经历了从大量手写体开始进行分析（数据），找出每一个字的核心特征（信息），并依此特征对文字进行识别（知识）的研发过程。

1.2.1 信息处理的内涵

信息处理是运用科学合理的手段与方法对原始数据进行整理，或按照事先设计的信息提取标准来采集、存储与加工信息等技术活动。

1. 信息采集

信息采集是指从各类信息载体中通过一定的方式与方法获得信息的过程，包含手工信息采集与自动信息采集。一般意义下，从文献中采集信息的过程叫文献检索，从各类仪器设备中获得一维、二维、三维信号的过程叫信号采集，从计算机的数据库中采集信息的过程称数据检索，这些均属于信息采集。本书第 2 章将会重点介绍网络文献检索的信息采集方式，其他章用例所涉及的数据与信息采集方法将不会述及。

2. 信息存储

因为信息往往具有可重复利用性和历史参考价值，所以必须安全、准确、长期地保存信息，确保信息存储的连续性和安全性。信息存储技术、设备、容量、速度和安全管理相关软硬件的使用等，都是信息处理与分析必备的知识。本书默认读者已掌握文件、数据库、数据库管理、Office 办公软件等部分知识，能保存开展实验的数据。

3. 信息加工

信息加工是指将收集到的信息（称为原始信息）按照一定的程序和方法进行分类、分析、整理、编制等，使其具有可用性。加工是信息得以利用的关键。加工既是一种工作过程，又是一种创造性思维活动。对原始信息进行加工的目的是：将初始的、零散的、无序的、彼此独立的信息形式，变换成便于观察、传递、分析、利用的信息形式；对原始信息进行必要的筛选、过滤和分类，以去粗取精、去伪存真，使信息更具条理性和系统性；使信息能获得更高层次的综合与处理，能产生更有价值的认识。

信息加工内容如下。

分类：是指对凌乱无序的信息进行整理归并，使其有条不紊，各得其所。分类可以按时间、空间（地理）、事件、问题、目的和要求等标准来进行。

比较：是指对信息进行分析，从而鉴别和判断出信息的价值、时效性。

综合：是按一定的要求和程序对零散的数据资料进行综合性的处理。

表达：对加工过的信息整理成易于理解，易于阅读的形式，如文字、图表、音视频等。信息表达是理解信息的基本条件。如，在临床过程中收集了一批高血压病人的血压数据，这些数据是连续几个月每日 4 次测量血压得到的，通过这些数据来判断一种抗高血压药物是否有效。在对这些数据进行分析之前，通常会要求对病人治疗前及治疗后一个月及每两个月的数据进行计算，如给出血压平均值。这些计算后的数据若能大量地应用图表表达，并且以不同的方式从不同的角度制作图表，将大大提高对这些信息的认知。本书第 3 章介绍的 MATLAB 软件工具对此将有述及。

另外，为了使信息被加工后能被更大范围的人和机器识别与处理，往往还会对信息进行编码。信息编码实际上是赋予信息元素以代码的过程，用不同的代码与各种信息中的基本单位建立一一对应的关系。任务和目的不同，信息编码的方式也会不同，信息编码会涉及规范化、标准化等问题，本书第 5 章将会进行介绍。

1.2.2 信息分析的内涵

信息分析是指以用户的特定需求为依托，以定性和定量研究方法为手段，通过对信息的收集、整理、鉴别、评价、分析、综合等系列化加工过程，形成新的有价值的信息产品或认知的过程。主体任务是：从混沌的信息中萃取有用的信息；从表层信息中发现相关的隐蔽信息；从过去和现在的信息中推演出未来的信息；从部分信息中推知总体的信息，揭示相关信息的结构和变化规律。信息分析需要系统地采集与之相关的各种原生信息，进行定向的筛选和整序，通过逻辑思维过程对其内容进行去伪存真的鉴定、由表及里或由此及彼的推理，运用科学的理论和方法对原生信息进行分析处理和提炼，以得出有助于解决实际问题的知识，揭示研究对象的内在变化规律及其与之相关联对象的联系，满足研究需求。

信息分析方法是一个庞大的体系，对一个具体的信息分析课题而言，可采用的方法往往并非唯一，而是有多种现实的方案可供选择或组合，它与研究的具体情况相关。具体而言，信息分析方法包括定性和定量两种。定性分析方法包括对比与类比、分析推理和综合抽象。定量分析方法包括因果关系类（回归分析法、时间序列分析法）、趋势外推类（回归分析法、时间序列分析法）、变量变化类（主成分分析法、因子分析法、典型相关分析法）、定性—定量转化类（德尔菲法、层次分析法、交叉影响法）以及定量-定性转化类（聚类分析法、判别分析法）五种。

做好信息分析有两个前提条件：（1）充分了解特定研究对象的历史、现状，并预测其未来的发展趋势，经过分析鉴别、综合归纳、判断推理的研究加工过程，结合实际需要和工作深度，提出有依据、有分析、有评价、有预测性意见的信息分析结果，为决策等相关活动服务；（2）信息分析方法的选择与应用。在信息分析中，大量的原生信息被深加工成对科学决策相关智能活动有支撑作用的新信息，对方法的合理选择和应用是决定信息分析水平和效率以及信息分析质量和效益的重要因素。

1.2.3 信息论及相关常识

在现代科学背景下，信息分析与处理离不开信息论的理论指导。信息论是研究信息的产生、获取、变换、传输、存储、处理识别及利用的学科。一般认为，1948 年香农发表的《通信的数学理论》一文标志着信息论的诞生。信息论有狭义和广义之分。狭义信息论即香农早期的研究成果，它以编码理论

为中心，主要研究信息系统模型、信息的度量、信息容量、编码理论及噪声理论等。广义信息论又称信息科学，主要研究以计算机处理为中心的信息处理的基本理论，包括评议、文字的处理、图像识别、学习理论及其各种应用。信息论的研究与很多学科密切相关，例如，数学、物理学、控制论、计算机科学、逻辑学、心理学、语言学、生物学、仿生学、管理科学等。信息论在各个方面得到了广泛的应用。信息科学是在信息论的基础上发展起来的，包括系统论、控制论、信息论、耗散结构论、协同论、突变论、超循环论等学科。随着现代科学技术的发展，信息科学也在不断向纵深方向深化和发展。现代信息科学实际上是以信息作为研究核心的一系列主导学科与边缘学科群。信息科学是以信息作为主要研究对象，以信息的运动规律作为主要研究内容，以现代科学方法论作为主要研究方法，以扩展人的信息功能作为主要研究目标的一门科学。信息科学包括对信息的描述和测度、信息传递理论、信息再生理论、信息调节理论、信息组织理论、信息认识理论等内容。它研究信息提供、信息识别、信息变换、信息传递、信息存储、信息检索、信息处理、信息施效等一系列问题和过程。在信息处理与分析中，通常需要了解并考量以下知识。

1. 信息量与熵

信息量是信息论中量度信息多少的一个物理量。它从量上反映具有确定概率的事件发生时所传递的信息。信息的量度与它所代表的事件的随机性或意外事件发生的概率有关，当事件发生的概率大，事先容易判断，有关此事件的消息排队事件发生的不确定程度小，则包含的信息量就小；反之则大。从这一点出发，信息论利用统计热力学中熵的概念，建立了对信息的度量方法。在统计热力学中，熵是系统的无序状态的度量，即系统的不确定性的度量。

熵和信息量是信息学中的一组重要概念，是描述信息处理和信息传递的重要指标。

信息熵是信息论中的一个基本量。例如，在试验甲和乙中，两种结果 A 和 B 出现的概率如表 1-1 所示。

表1-1　A和B出现的概率

	出现 A 的概率	出现 B 的概率
试验甲	0.50	0.50
试验乙	0.99	0.01

那么，在试验之前，就试验甲而言，很难断定 A 和 B 中哪个将出现；但就试验乙而言，就很有把握地断定 A 将出现。由此可见，在不同的试验中，其不确定性是有大有小的，试验甲的不确定性就比试验乙的大。熵就是描写不确定性大小的量，熵越大不确定性就越大。一般来说，设在试验中有 N 个可能出现的结果，$A(1)$、$A(2)$、$\cdots A(N)$，假如它们出现的概率分别是 $P(1)$、$P(2)$、$\cdots P(\mathrm{N})$，通常规定这个试验的熵为：

$$H=P(1)\lg P(1)-P(2)\lg P(2)\cdots-P(N)\lg P(N)$$

2. 信息科学研究方法

信息科学的研究方法：信息科学研究有其独特的方法，这些方法包括信息分析综合法、行为功能模拟法和系统整体优化法。

（1）信息分析综合法。复杂系统、高级过程一般都具有极其复杂的成分、复杂的结构、复杂的联

系和复杂的行为。从信息的观点出发，抓住事物的信息特征，分析事物间的相互联系，提示其本质规律，从而实现决策目标的完成。

（2）行为功能模拟法。是从行为的观点出发，以行为的相似性为基础，从功能上来模拟事物或系统对环境影响的反应方式，是信息分析综合法的一个重要发展和实用化。这一方法常常又称作“黑箱方法”。所谓“黑箱”，就是指那些既不能打开，又不能从外部直接观察其内部状态的系统，比如人们的大脑只能通过信息的输入/输出来确定其结构和参数。“黑箱方法”从综合的角度为人们提供了一条认识事物的重要途径，尤其对某些内部结构比较复杂的系统，对迄今为止人们的力量尚不能分解的系统，黑箱理论提供的研究方法是非常有效的。“黑箱”研究方法的出发点在于：自然界中没有孤立的事物，任何事物间都是相互联系，相互作用的，所以，即使我们不清楚“黑箱”的内部结构，仅注意到它对信息刺激作出的反应，注意到它的输入/输出关系，就可对它做出研究。如果我们能设计出一个系统，在同样的输入作用下，它的输出和所模拟对象的输出相同或相似，就可以确认实现了模拟的目标。在此，信息的输入，就是一个事物对黑箱施加影响；信息的输出，就是黑箱对其他事物的反作用。事实上人们在对信息进行分析和综合时，很少追求结构上的相似性，而总是把握信息的观点、行为功能的观点。

（3）系统整体优化法，即是从系统的观点出发，着重从整体与部分之间、整体与外部环境之间的相互联系中，综合地考察对象，从而得到全面地、最佳地解决问题的方法。实践证明，物质具有系统属性，科学研究的对象，都可以把它看成是一个由基本要素组成的动态系统。在这个系统内外，不仅存在着信息传递、交换，还有对信息的处理和控制。同行为功能模拟法一样，系统整体优化法也是信息分析综合法的一个重要的发展和实用化。

3. 信息技术

计算机技术与现代通信技术一起构成了信息技术的核心内容。信息技术能够延长或扩展人的信息功能（包括传感技术，通信技术，计算机技术和缩微技术、多媒体技术等）。传感技术的任务是延长人的感觉器官收集信息的功能，通信技术的任务是延长人的神经系统传递信息的功能，计算机技术则是延长人的思维器官处理信息和决策的功能，缩微技术是延长人的记忆器官存储信息的功能。当然，这种划分只是相对的、大致的，没有截然的界限。如传感系统里也有信息的处理和收集，而计算机系统里既有信息传递，也有信息收集的问题。目前，传感技术已经发展了一大批敏感元件，除了普通的照相机能够收集可见光波的信息、微音器能够收集声波信息之外，现在已经有了红外、紫外等光波波段的敏感元件，帮助人们提取那些人眼所见不到的重要信息。还有超声和次声传感器，可以帮助人们获得那些人耳听不到的信息。不仅如此，人们还制造了各种嗅敏、味敏、光敏、热敏、磁敏、湿敏以及一些综合敏感元件。这样，还可以把那些人类感觉器官收集不到的各种有用信息提取出来，从而延长和扩展人类收集信息的功能。

1.3 医药信息与医药信息学

1.3.1 医药信息

医药信息是医药信息学研究的对象。医药信息的特点在于：医药信息面广量大，更新速度快。仅以病人为例，其信息不仅牵涉面广、数量庞大，而且十分复杂细致，再加上病人流动频繁，以及每天的病情变化，造成信息更新快，形成极为复杂的海量信息。医药信息种类繁多，包括数值、文字、图

像、声音、气味等，各种类的信息表达形式不一、所包含数据的标准不一、单位不一，难以标准化。医药信息量化困难，它不同于工程信息，往往概念不精确，难以量化，各变量的相互关系及变化规律难以用数学语言表达。例如，头痛的性质和程度会因患者的文化素质、痛阈高低、意志力不同而表达不一。另外，医疗病历中的病史、病程记录、病情讨论分析多采用自然语言，常因医师的学术水平、文化素质、性格习惯不同而迥然不一，自然语言标准化是全球共同的难题。

医药信息大致可分为三大类：医药公用信息、医药临床信息和医药管理信息。医药公用信息一般是指不涉及病人隐私权的医药信息和不属于内部管理机构的医药管理信息，包括医药学情报、书籍、期刊、医疗卫生档案，以及医学决策支持信息、医药卫生年鉴、政府公布医药卫生统计信息等各种文字、图像、音频、视频。医药临床信息是以病人临床数据为核心的系列临床诊疗信息，包括临床信息（病人信息、医嘱信息、护理信息等）、临床检验信息、临床检查信息、医学影像信息以及与治疗有关的信息（如药品剂量、药物性质、各种治疗方法有关的数据、营养饮食配餐信息、药物监测信息、临床监测、监护，信息等）。医药管理信息包括医药卫生机构组织信息、医药技术管理信息、物资与设备管理信息（包括医学设备与仪器管理信息、消耗性物资、卫生材料管理信息、药品管理信息等）、医药卫生机构经济管理信息、教学科研管理信息、人事人才管理信息、后勤保障服务管理信息等。

医药信息涉及的学科包括基础医药学、临床医学、预防医学、临床专科与辅助学科、生物医学等。

（1）基础学科：解剖学、组织胚胎学、生物化学、遗传学、细胞及分子生物学、免疫学、微生物学、病理生理学、药理学、药剂学、寄生虫学和神经生物学等。

（2）临床学科

内科：肾脏病学、心血管病学、感染与传染病学、老年病学、呼吸病学、内分泌病学、免疫与风湿病学、血液病学、神经病学、消化病学和儿科学等。

外科：普通外科学、整形外科学、烧伤外科学、胃肠外科学、胸外科学、心脏外科学、创伤及骨科学、麻醉学。

其他：妇产科学、眼科学、耳鼻喉科学和移植学等。

（3）预防医学包括营养食品卫生学、毒理学和劳动与环境保护学等。

（4）临床专科与辅助学科包括放射诊断学、超声诊断学、急诊医学、肿瘤学、口腔医学、护理学、中医学、皮肤病学等学科。

（5）生物医学包括遗传学、发育生物学、细胞生物学等。

医药信息是医药相关学科开展科学研究、临床服务、药品开发生产与利用过程中，各种系列数据的具体表达与内涵解读，医药信息的分析是获取更多医药知识，并加以充分利用的必由之路。医药信息的系统研究是一个大的学科体系。由于医药信息表达与传输处理都离不开计算机技术、数字化技术、网络技术和通信技术，因此，医药信息的源头所属学科，都有与信息科学交叉研究的内容，并正在逐步形成新型交叉学科，如医学信息学、药学信息学、临床信息学、护理信息学等。

医药信息的特点：医药信息的数据量大，复杂性高；医药信息源是以人为本的信息收集对象，因此涉及的数据源是海量的，数据的类型、属性、表达方式也是错综复杂的；医药信息标准化工作的水平较低，表现在信息分类、编码工作存在着不统一，造成交流与共享困难；医药信息的处理难度大；医药信息种类繁多，流程复杂；医药信息的私密性强，连续性、时效性显著。

1.3.2 医药信息学

医药信息学是以医药信息为主要研究对象，以医药信息的运行规律、应用方法为主要研究内容，以现代计算机为主要工具，以解决医药工作人员在处理医药信息过程中的各种问题为主要研究目标的新兴学科，是一门介于医药与信息学之间的交叉学科。随着以计算机技术为代表的信息技术在医疗工作包括数据通信、医疗质量评估、辅助决策过程、管理规划和科学研究中越来越广泛的应用，医药信息学受到了世界各国的普遍重视，获得了快速的发展，已经深入渗透到医疗领域的方方面面，如电子病历、生物信号分析、医药图像处理、临床支持系统、医药决策系统、医院信息管理系统、卫生信息资源等，为提高医疗效果、效率、效力并降低医疗支出，合理配置医疗资源做出了杰出的贡献。

医药信息学的研究范畴包含医药信息的获取、处理、存储、分配、分析、解释和利用的所有方面。具体来说，主要包括以下几个方面。

- 研究医药信息的概念、属性、本质和度量。
- 研究自然科学的知识综合、专门知识或临床经验及其规范化。
- 研究医药信息的产生、提取、检测、变换、传递、存储、处理和识别。
- 研究利用医药信息进行控制的原理和方法，在控制论的指导下，研制各种信息化、智能化的诊疗设备。
- 研究实现医药信息系统最佳组织的原理和方法，在系统论的指导下，运用系统工程的技术，以及硬件工程、软件工程和知识工程的方法，研制最有效的医药信息系统。

医药信息学的研究能够最大限度地延伸医药工作人员的感觉功能、思维功能和执行功能，提高其对医药信息的提取、检测、传递、存储、识别、利用等方面的能力。

1.4 医药信息处理与分析的基本流程

在医药临床与科学研究过程中，科研人员会得到大量的原始数据，其中包括大量的图片资料以及多媒体信息。医药信息处理与分析往往贯穿于医药科研工作的整个过程，最典型的工作是文献信息检索、数据统计、数据挖掘及数据的可视化表达等。

信息处理与分析的基本流程依次是：信息采集与数据化表达、数据处理与集成、数据分析和数据解释 4 个阶段。无论是什么样规模的信息处理与分析任务，其基本的工作流程包括如下几个步骤。

第一步：信息采集与数据化表达

信息的采集是指在信息资源方面做准备的工作，包括对信息的收集和处理。主要的信息来源如表 1-2 所示。

表1-2 信息来源

类　型	表现形式	特　点	主要获取方式
文献型信息源	文档。如报刊、百科书、词典及各类出版物等	量大、存储文件格式多样，如：.docx、pdf 等	文献检索
数值型信息源	数据。如临床采集的血压、心率等数据	数据测度迥异	实验室、临床及各类相关仪器记录或计算出的数据

续表

类　型	表现形式	特　点	主要获取方式
文本型信息源	文本。如临床记录的望闻问切记录的病历	自然语言、规范表达不够	人工采取或数据库检索
图形图像信息源	影像。如：x 光片、B 超图像	二维信号	设备采取或数据库检索
音频视频信息源	声音、动画等。如胎心音、动态 B 超	一维信号、多维信号	设备采取或数据库检索

第二步：数据处理与集成

数据的处理与集成主要是对已经采集到的数据进行适当的处理、清洗去噪以及进一步的集成存储。

第三步：数据分析与解释

数据分析是整个数据处理流程中最核心的部分，因为在数据分析的过程中，会发现数据的价值所在。在一个完善的数据分析流程中，数据结果的解释步骤至关重要。数据分析结果比较复杂时，除用传统的数据显示方法以外，还可用“数据可视化技术”作为解释数据的有力方式。通过可视化，可以形象地展示数据分析结果，更易于对分析结果的理解和接受。常用的分析工具一般都有基本可视化技术、基于图标的技术、基于图像的技术等。信息处理与分析流程示意图如图 1-2 所示。通过对信息的采集、处理、集成与分析等工作以达到决策、预测、知识发现等目的。

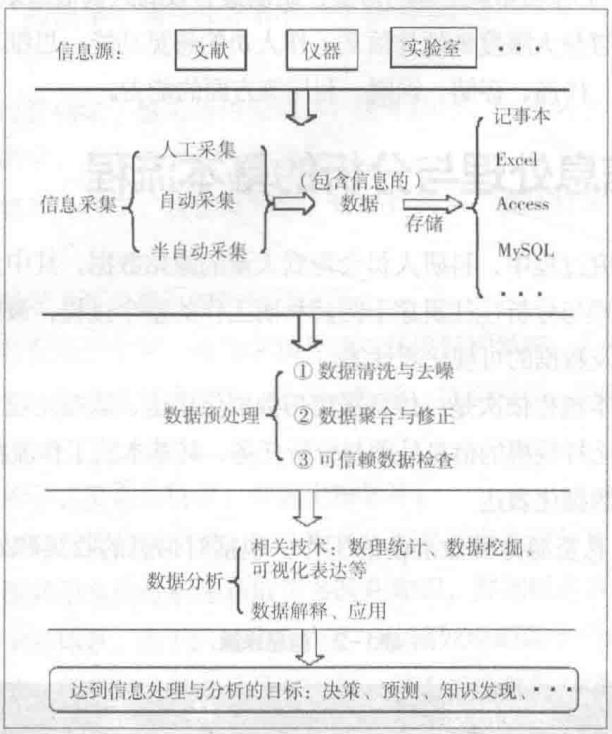


图1-2　信息处理与分析流程示意图

1.5　医药信息分析质量保证的几个关键因素

医药信息处理与分析是对各种相关信息的深加工，是深层次或高层次的医药信息处理，是一项具