医疗健康大数据的研究进展与趋势

CCF 大数据专家委员会

彭绍亮[1,2]　杨亚宁[1]　张彦春[3,4]　胡　斌[5]　阮　彤[6]　邢春晓[7]

[1]湖南大学，长沙

[2]国家超级计算长沙中心，长沙

[3]澳大利亚维多利亚大学，墨尔本

[4]复旦大学，上海

[5]兰州大学，兰州

[6]华东理工大学，上海

[7]清华大学，北京

摘　要

近年来，强大的计算平台以及数据存储技术的发展使得医疗数据正在向电子数字化转化，并呈现爆发式增长。挖掘医疗数据，可以辅助医生进行临床科研和临床诊疗。本文对医疗健康大数据的概念进行了剖析，并对其进行了分类，总结出医疗健康大数据所面临的挑战。然后，针对这些挑战，较系统给出了医疗大数据的采集、分析、保护和应用的国内外发展现状，着重讨论了医疗大数据在医学影像、辅助诊断、药物研发、健康管理和基因测序等五个方面的应用。最后，展望了未来医疗健康大数据行业的发展趋势。

关键词：医疗大数据，临床科研，临床诊疗，数据采集，数据分析

Abstract

In recent years, the development of powerful computing platforms and data storage technologies has enabled medical data to be digitally transformed, and the data growth has also exploded. Excavating medical data can assist doctors in clinical research and clinical diagnosis and treatment. This paper analyzes the concept of medical health big data and classifies it to summarize the challenges faced by medical health big data. Then, in response to these challenges, we provide a detail survey in medical big data collection, analysis, protection and application. The application of medical big data in medical imaging, auxiliary diagnosis, drug development, health management and gene sequencing is discussed. Finally, the future development trends of the health care big data are discussed.

Keywords: medical big data, clinical research, clinical diagnosis and treatment, data collection, data analysis

1　引言

近年来，随着移动互联网、云计算、物联网等信息技术的快速发展与广泛应用，使

得全球范围内的数据正在以前所未有的速度增长。大数据是指难以被传统的数据管理系统有效且低成本地存储、管理以及处理的复杂数据集[1]。与传统的数据相比，大数据包含 5V 特性：数据规模巨大（Volume）、数据的产生速度非常快（Velocity）、数据类型繁多（Variety）、分析结果取决于数据准确性（Veracity）、大数据一般具有重要的价值（Value）[2]。大数据不仅在于数据规模的巨大，更代表着处理大数据所需要的方法和手段，以及分析和应用大数据所带来的新发明、新服务和新机遇[3-5]。现如今，多个行业已经不得不利用大数据来提高技术质量，提升自身的价值。医疗健康数据是其中一个最具代表性的应用。

1.1　医疗健康大数据

医疗健康大数据研究对辅助医生给病人选择更好的治疗方案，进而提升医疗服务质量，降低医疗成本有积极的作用，得到了各国政府的大力支持。从 2013 年起，美国、英国在医疗大数据应用方面投入了大量资金[6]。2015 年 3 月，我国在国家卫生计划生育委员会网络安全和信息化工作组全体会议上提出“推进健康医疗大数据应用，制定促进健康医疗大数据应用的相关方案，推动健康医疗大数据有序发展”的意见。2016 年 6 月，国务院办公厅颁发了《关于促进和规范健康医疗大数据应用发展的指导意见》，明确指出健康医疗大数据是国家重要的基础性战略资源，要通过其应用，激发深化医药卫生体制改革的动力和活力，提升健康医疗服务效率和质量。

人口的增长和老龄化，发展中国家医疗市场的扩张、医学技术的进步和人力成本的不断上涨将推动支出增长。2017-2021 年全球医疗支出预计将以每年 4.1% 的速度增长，而 2012-2016 年的增速仅为 1.3%。慢性病发病率提升，变化的饮食习惯以及日益增加的肥胖度加剧了慢性病的上升趋势[7-8]，特别是癌症、心脏病和糖尿病，目前中国糖尿病患者约有 1.14 亿，而全球患者人数预计将从目前的 4.15 亿增加至 2040 年的 6.42 亿。传统研发（R&D）成本上升，产品上市速度慢，2004 年至 2014 年药物开发成本增加了 145%。劳动力不足，在人口结构的变化和技术的迅速发展下，熟练和半熟练医疗保健工作者将大幅减少。

2015 年，国家卫计委提出分级诊疗制度将在 2020 年全面确立，包括基层首诊、双向转诊、急慢分诊、上下联动分诊诊疗等。新模式的搭建过程中，主要存在以下 3 个问题：

1）信息不流通，各医疗机构间多为信息孤岛，患者信息无法进行快速共享流通；

2）资源不流通，优质医生多集中在各大省会的顶级医院，且三甲医院医生精力有限，每年可支援的基层医疗更是有限；

3）利益不互通，医院之间缺乏有效的利益捆绑机制，以促进患者在院间的流通。

医疗大数据技术的应用，将从体系搭建、机构运作、临床研发、诊断治疗、生活方式五个方面带来变革性的改善[9-12]。例如，腾讯公司最近发布了其最新医疗 AI 引擎——腾讯睿知，从诊前环节切入，推出智能导诊，其目的利用大数据与人工智能解决医疗资源错配的问题。腾讯睿知的主要数据类型分为三个方面：一是来自于互联网用户的数据，

包括一系列远程医疗服务平台；二是真实的患者语料库，腾讯睿知学习了百万级的脱敏数据；三是静态数据。包括权威的医学先验知识以及各种症状体征、检验检查指标、用药治疗的疾病知识库。技术上，它以自然语言处理技术为核心，结合医学图像 OCR 能力（Optical Character Recognition，图像文字识别）、深度学习等 AI 算法模型构建引擎，包括疾病判断引擎和沟通对话引擎，进而实现对疾病及病程的预判，可以说这是医疗 AI 应用的关键一环。广州妇女儿童医疗中心已经试运行其智能导诊功能 3 个多月，疾病判断准确率 94%，医生推荐准确率 96% 以上。此外在全科领域，腾讯睿知仍在不断拓展，覆盖疾病范围已扩大至 23 个学科，但这些病种主要是集中在全科领域。从产品的角度，腾讯医疗更关注患者诊疗过程中的哪些环节可以用技术替代人工。理想状态下，所有的技术能力都可以在诊前、诊中、诊后的过程中得到应用。由于我国医疗体系的强监管性，大数据若要在行业内实现其价值，需由国家建立一套自上而下的战略方针，从而引导医院、药企、民办资本、保险等机构企业构建项目，相互合作，最终实现从“治疗”到“预防”的就医习惯的改变，降低从个人到国家的医疗费用。

目前以高通量测序技术[13]为基础的生物大数据和序列分析技术正在推动着医疗健康领域的技术革命，相比于 2000 年，2010 年的基因组数据产量增大了 8 个数量级。基因组研究产生的海量数据正以每 12 ~ 18 个月 10 倍以上的速度增长，已远远超过摩尔定律。二代测序同时衍生出了 RNA-seq[14]、exome-seq[15]、ChIP-seq[16]、DNA 甲基化测序[17]、染色质交互分析[18]等针对特定生物分子或靶标的测序方法。另外，蛋白质质谱分析法[19-20]被用于蛋白质组学研究，医学影像[21]（如 CT[22]、核磁共振 MRI[23]）已成为医学研究及诊断的常用技术。在新型生物技术的协助下，大型生物/医学项目（如 1000 基因组[24]、ENCODE[25]、modEN- CODE[26]、Cancer Genome Atlas[27]、Human microbiome[28]等）得以完成或正在稳步推进。然而，在利用相关技术进行研究的同时也产生了规模庞大的数据，其累积速度已超过了摩尔定律（Moore’s Law）所标量的计算机发展速度，形成了“医疗健康大数据”。当前，医疗健康大数据处理的相关问题已成为生命科学和健康医疗研究发展的重要挑战，医疗健康大数据的分析和计算需要在保持高精准度的前提下，尽可能地提高时效性。这些数据只有经过精确的分析、挖掘和计算才能发挥巨大的作用，并应用到人类疾病与健康产业中，例如以医院电子病历为核心的临床数据记录了病人的疾病、诊断和治疗信息，挖掘此类数据，可以辅助医生进行临床科研与临床诊疗。

健康医疗大数据领域涉及的相关技术范围非常广，如底层数据采集中包括信息化、物联网、5G 技术，处理分析中包括深度学习、认知计算、区块链、生物信息学及医院信息化建设等。全球大健康数据正以每年 48% 的速度增长，在 2020 年数据量将超过 2300Exabytes。预估 2020 年，全球健康物联网设备出货量将达到 161 万台。院内数据方面，2016 年医院管理信息系统整体已实施比例在 70% ~80%，且集中于三级医疗机构，大量健康医疗数据的积累为算法搭建提供了基础。在处理分析方面，人工智能、生物信息学需要与实际应用场景相结合，以便搭建有效模型。

医疗健康大数据增长快、应用范围广、贡献价值大，近几年引起人们的特别关注。

医疗健康大数据在辅助医生进行临床科研、临床诊疗、药物研发等诸多领域的应用前景十分广阔，能够创造巨大的价值和收益[29-30]。

1.2 医疗健康大数据分类

健康医疗大数据分为三大类，院外数据、院内数据以及基因数据。院外数据包括健康档案、智能硬件体征及环境监测/检测，院内数据包括就医行为、临床诊疗等，基因数据包括外显子、全基因等。在具体场景应用方面，多为不同种类的数据相互交叉结合应用，如预防预警，需要结合智能硬件监测、诊疗用药历史等数据才能为用户提供及时的预警监测。

1.3 医疗健康大数据特性

医疗健康大数据具有以下特性：

1）体积大。医疗健康数据体积巨大，例如，一张 CT 图像包含的数据量大约为 100MB，一个标准病历图接近 5GB。

2）多态性。数据来源多样性，涵盖的数据种类丰富，包括文本型、数字型、医疗图像等。多类型的数据对数据处理能力提出了更高的要求。多态性是医疗数据区别于其他领域最显著和最根本的特性，这种特性也在一定程度上加大了医疗数据分析和处理的难度。

3）不完整性。医疗数据的搜集和处理过程经常相互脱节，这使得医疗数据库难以对任何疾病信息全面反映。大量的医疗数据来源于人工记录，这导致数据记录通常存在偏差和残缺，同时许多数据的表达、记录本身也具有不确定性。另外，随着医疗技术和手段的不断创新与发展，可能会产生新的医疗数据类型，数据的维度是不断地在增长。

4）冗余性。每天都会产生大量的医学数据，同一人在不同医疗机构就可能产生相同的信息；整个医疗数据库会包含大量重复和无关紧要的信息，如常见疾病的相关描述信息，与病理特征无关的信息。

5）时效性。数据的创建速度快，更新频率高，许多数据的采样周期已从周、天升级到分、秒，甚至是连续性记录。这对响应速度及处理速度提出更高要求。另外，就诊、疾病进程等并非是在某一时间点上发生的瞬时事件，在前、中、晚期可能会呈现出不同的特点。此外，疾病也可能具备季节性的特征。

6）隐私性。数据隐私性是医疗大数据的重要特点。个体的患病情况、诊断结果、基因数据等医疗健康数据的泄露会对个人产生负面影响，且涉及侵犯公民权。

1.4 医疗健康大数据面临的挑战

当前，大数据技术的涌现，在给医疗领域的发展与进步带来机遇的同时，也给其带

来了巨大的挑战。首先，医疗大数据并不是单一的学科，它涉及医疗、人工智能、以及数据挖掘等多个领域的知识。另外，由于医疗数据本身的复杂性、多态性、不完整性、冗余性、时效性以及隐私性等特点，使得医疗大数据的研究在数据采集存储、数据保护、数据的分析处理以及数据应用等方面面临着巨大的挑战。当前，医疗数据的发展及应用增长迅猛，医疗领域在动态数据安全监控和隐私保护方面存在不足，带来了数据安全和个人隐私保护方面的困扰。同时，从数据角度出发，医疗领域的数据采集，以及数据的分析处理和应用方面也面临着很大的压力。总体来说，医疗健康大数据目前存在着数据的采集、分析处理、保护和应用四个方面的挑战。

1）数据采集的挑战。医疗大数据的来源是多元的，质量是不受控制的，有些数据是拿来也不可用的，比如：不可及的碎片化数据，可及但又是错误的数据，可及、正确但是残缺、无法修补的数据；

2）数据分析处理的挑战。医疗健康大数据增长非常迅速，半结构化和非结构化，以及来源于多个位点等这些现状，使得使用传统的数据管理工具变得非常困难。这些传统的数据管理系统首先存储能力不够，且一般都是独立的，不能分享与合并数据，而任何集中式的数据库系统仍然要面对如单一的控制点、瓶颈问题等。

3）数据保护的挑战。挖掘医疗数据时，不可避免地会涉及患者的隐私信息，这些隐私信息的泄露会对患者的生活产生不良的影响。大数据分析中，个人隐私的保护需要注意两个方面，一是患者身份、地址和疾病等相关的敏感信息应该被隐藏，二是能够通过数据挖掘或者其他方式得到的敏感信息也应该被除去。

4）数据应用的挑战。医疗健康大数据本身来源广泛，设计到医学影像、辅助诊断、药物发现、健康管理以及基于测序等众多领域的数据信息，使得医疗大数据非常复杂。另一方面，医疗大数据涉及医疗、人工智能、数据挖掘等多个领域的交叉学科研究，因此对其展开研究存在很大的困难。

2 国内外研究现状

下面主要对医疗大数据研究进展进行介绍，包括医疗大数据的采集、分析、保护和应用等四个方面，其中，对于医疗大数据的应用，我们将分别从医学影像、辅助诊断、药物研发、健康管理和基因测序等多个方面进行介绍。

2.1 数据采集

随着医院的信息化程度日趋成熟及物联网和互联网的发展，医、药相关的行为数据量大大提升。我国医疗数据主要来自于医院信息系统（HIS）、电子病历系统（EMR）/电子健康档案（EHR）、影像采集与传输系统（PACS）、实验室信息系统（LIS）、病理系统（PS）、医疗器械等信息化系统和设备所记录下来的疾病、体征数据。还包括医院

物资管理、医院运营系统所产生的数据。通过对医疗大数据的分析和加工，可以挖掘出和疾病诊断、治疗、公共卫生防治等方面的重要价值。中商产业研究院整理了关于我国医疗数据的采集途径、难度以及数据价值的信息，见表 1。医疗大数据采集的三个关键环节是：多源异构数据融合、数据清洗转换、数据脱敏。

1）多源异构数据融合：消除多源信息之间可能存在的冗余和矛盾，加以互补，改善信息提取的及时性和可靠性，提高数据的使用效率；

2）数据清洗转换：数据清洗的任务是“洗掉”不符合要求的“脏数据”。该过程需严格遵守清洗规则，补全不完整数据、挑出并修正错误数据、对重复数据进行去重操作；

3）数据脱敏：是指以特定的脱敏规则对某些敏感信息进行变形，实现敏感隐私数据的保护，让其可以正常使用而不被非法利用的一项技术。

表 1　我国医疗数据采集信息表

序号	来源	数据采集难度	数据价值
1	电子病历数据	★★★☆☆	★★★★★
2	检验数据	★★☆☆☆	★★★★☆
3	影像数据	★★☆☆☆	★★★★★
4	费用数据	★★★★☆	★★★★☆
5	基因测序数据	★★★★☆	★★★★★
6	医药研发数据	★★★★☆	★★★★☆
7	药品流通数据	★★★★☆	★★★☆☆
8	智能穿戴数据	★★★☆☆	★★☆☆☆
9	移动问诊数据	★★★☆☆	★★☆☆☆
10	体检数据	★★★☆☆	★★☆☆☆

2.1.1　医院信息系统

医院信息系统（HIS）就是利用计算机软硬件技术、网络通信技术等现代化技术手段，对医院及其所属各部门对人流、物流、财流进行综合管理，对在医疗活动各阶段中产生的数据进行采集、存储、处理、提取、传输、汇总、加工生成各种信息，从而为医院的整体运行提供全面的、自动化的管理及各种服务的信息系统。HIS 的主要功能按照数据流量、流向及处理过程分为临床诊疗、药品管理、经济管理、综合管理与统计分析、外部接口五部分。

上海交通大学医学院附属仁济医院每年的诊疗人次在 400 万以上，出院人次在 10 万左右。如此大的业务量，如果要考虑能满足更高等级数据一致性的话，仁济医院觉得要采取一种“单体数据库 + 数据集成平台”的模式，最终选择了 IBM 开发的 LinuxONE 这样一个系统。一是 LinuxONE 能满足医院信息系统、电子病历、实验室检查信息系统、移动医疗共享单体数据库的要求；二是在数据库整合方面具有水平扩展能力，其弹性架构具备垂直扩展能力；三是其内部全冗余系统的设计和应用，能保证医院 7 × 24 小时稳定可靠不宕机的要求；四是具备强大的 IO 数据处理能力，满足了医院未来对医疗数据的处

理需求；五是能实现医院的基础信息平台升级与“云架构”建设，其私有云平台稳定可靠、数据有最高安全保障；六是能将新业务系统快速部署上线，能实现服务器和存储资源云化共享、资源按需动态调配。

行心 HIS 系统（标准版）是广州市行心信息科技有限公司累积十多年医疗信息化行业经验，精心打造的新一代医院信息管理系统。按照国家相关规定，高标准、高要求的开发条件，符合国家卫计委《医院信息系统基本功能规范》和卫计委与财务部共同发布的新《医院财务管理制度》的规范与要求。除了全面和成熟以外，她最大的亮点和优势是带有临床辅助决策支持系统，这个系统是基于世界权威的 BMJ 临床数据库，经过我们全结构化的开发，完全嵌入到医生诊疗行为里面，帮助医生快速提高其诊疗能力，减少医疗事故和医疗纠纷。行心 HIS 系统（标准版）是一套面向中小医疗机构使用的“一体化”信息管理软件。门诊、住院、电子病历、检验、影像、临床知识库、临床决策、领导决策、药房药库、物资后勤、财务绩效、人事 OA 等系统一体化建设，一步到位。行心 HIS（标准版）搭配远程医疗，以及免费的全科信息系统，可以帮助医院联合周边区域内的基层医疗机构，快速搭建自己的医联体。

医院信息系统提高了医院的现代管理水平、提高了工作效率、优化了医疗流程、促进了医教研质量的提高、增加了医院的经济效益、有利于医改的落实、在突发公共卫生事件中发挥了特有的作用，同时提升了医院文化。

2.1.2 电子健康档案

电子病历（EMR）是病人在诊断和治疗过程中产生的数字医疗信息文档，是“以医疗为中心”的数字化健康档案。电子健康档案（EHR）是以医院的电子病历为主体，以信息共享为核心的数字化健康档案。EHR 将跨越不同的机构和系统，在不同的信息提供者和使用者之间实现医疗信息的互换和共享。

MIMIC-III（Medical Information Mart for Intensive Care III）[31]是一个基于重症监护室病人监测情况的医学开源数据集。该数据集的公布旨在促进相关医学及数据研究，提升 ICU 重症监护室在治疗、急救时的执行效率水平。MIMIC-III 包含了 2001 年至 2012 年期间入住重症监护病房的成年患者（16 岁或以上）的 53423 例不同住院患者的数据。此外，它还包含了 2001 年至 2008 年期间接纳的 7870 例新生儿的数据。数据涵盖 38597 名成年患者和 49785 名住院患者。成年患者的平均年龄为 65.8 岁，其中 55.9% 为男性，院内死亡率为 11.5%。数据库内的患者在 ICU 的住院时间中位数为 6.9。

与 MIMIC-III 数据库不同，eICU[32]数据库的来源为 Philips Healthcare，不同的来源导致这两个数据库的结构与记录方式完全不同。Philips Healthcare 是基于 Philips 各项医疗系统与设备产生的数据的医疗科技平台。eICU 数据库囊括了多家医院的重症监护数据，这与 MIMIC-II 围绕以色列贝斯女执事医疗中心的数据来源非常不同，eICU 数据库的病患样本数量更多，而 MIMIC-II 数据库记录的病患样本信息则更具体。eICU 协作研究数据库由美国大陆多个重症监护病房的数据组成。合作数据库中的数据涵盖了 2014 年和 2015 年入住重症监护室的患者。

2.1.3　影像采集与传输系统

医学影像系统（Picture Archiving and Communication Systems，PACS）是影像归档和通信系统，系统中存储了病人的医学影像（包括核磁，CT，超声，各种 X 光机，各种红外仪、显微仪等设备产生的图像）数据，并提供一些辅助诊断管理功能[33-34]。PACS 是以高性能服务器、网络及存储设备构成硬件支持平台，以大型关系型数据库作为数据和图像的存储管理工具，以医疗影像的采集、传输、存储和诊断为核心，是集影像采集传输与存储管理、影像诊断查询与报告管理、综合信息管理等综合应用于一体的综合应用系统[35-37]。

目前几乎所有欧美先进 PACS 厂家都用正式 DICOM3.0 文件格式来储存图像。新一代的 PACS 大多采用 DICOM 支持的标准压缩算法，如 JPEG、JPEGLossless、JPEG2000、JPEG-LS 和 Deflate 等。厂家用自定义算法来压缩图像的现象越来越少。三级储存模式（在线、近线和离线）已经转变成两级（在线和备份）：目前欧美先进 PACS 厂家都在推行在线和备份两级储存。备份只是为了防意外，如火灾、地震等。在线用的是硬盘，用 RAID（冗余存储磁盘阵列）加 NAS（Network Attached Storage）或 SAN（Storage Area Network）。而前几年 PACS 界最常见的是用三级图像储存模式：在线（online）、近线（near-line）和离线（off-line）。新的图像在线存在硬盘上、老一点的图像近线存在网路服务机里、再老一点的图像离线存在 MOD 或磁带里。

由于我国开发和引进 PACS 系统较晚，目前已经建立并有效运行的 PACS 系统并不多见（特别是内陆省市）。究其原因主要是标准化程度低、兼容性差，一般为封闭式的专用系统，既不经济、价格也昂贵，配置的硬件不够合理，对工作量大的医院缺乏强大的存储子系统，无法支持数据量巨大的常规放射影像，因此不能真正实现“无片化”管理。多数 PACS 系统也没有其有效的工作流程和自动化管理功能，也不能向临床诊断提供所需的全部，表现在在线信息少，响应速度慢。对网络安全、保密和符合法律要求方面还不可靠。现有的 PACS 系统设计大多数没有考虑技术发展和扩展需要的可能，难于与现有的 HIS/RIS 整合为一个系统。

2.1.4　实验室信息系统

实验室信息系统（LIS）是医院信息系统的一个重要组成部分。LIS 的主要目的是将各种免疫、生化、临检、放免、细菌及实验室用的分析仪器，用微机完全联网，管理和传输在实验分析过程中产生的全部数据[38-39]。它不仅可以满足检验科的生化、免疫、临检以及血液等常规检验专业的要求，还可以全面支持微生物、同位素、基因检测等特殊检验专业。从整体来看，LIS 系统发展现状基本可以从三个方面阐述：临床实验室自动化、自动审核系统、无纸化进程。

LIS 系统是以检验科的生产活动为主要内容，所以 LIS 系统的发展是与检验科的发展密不可分的。近年来，随着检验医学的快速发展和检验标本的急剧增长，实验室自动化成为目前检验科的发展新趋势。实验室自动化是指将临床实验室自动化分析仪通过传送

系统连接起来进行流水线作业检测实现样品运输、分类、前处理、检测、结果报告、后处理等全检验过程自动化[40-41]。自2000年开始，由浙江大学附属第一医院采购了国内首条实验室自动化系统，拉开了中国实验室自动化的序幕。随着检验事业的飞速发展，从2008年开始，自动化流水线的装机数量每年以65% ~70%的速度增长，到2014年年底，全国安装样本前处理或流水线系统可达到400台。雅培、西门子、罗氏、贝克曼等检验仪器厂家均有各自的实验室自动化流水线系统以及配套的中间体软件。LIS系统与临床实验室自动化流水线中间体软件的对接，对减少错误率、提高检验质量、缩短实验室标本转运时间（Turn Around Time，TAT）等起到举足轻重的作用。

计算机自动审核（Auto verification）是指临床实验室指挥计算机基于系统设定的一套规则对检验结果进行自动审核的过程。自动审核通常由实验室信息系统（LIS）单独或由LIS和中间件系统共同实现[42]。自动审核系统的建立，既要考虑结果自身的准确性，又要考虑与历史核对的符合率，还要考虑到数字型结果的可报告范围，对LIS系统判断执行提出了更高的要求，除此以外，质量控制、仪器报警、危急值、组合项目完整性、多个项目之间的逻辑关系等亦是影响自动审核的独立影响因素。自动审核系统快速而准确的识别分析结果、执行逻辑关系判断，仅筛选出不符合自动审核系统规则的结果进行人工审核，可对提高检验工作效率，减少人工审核工作量，缩短TAT时间做出重要贡献。

医院的无纸化进程，是将所有关于诊疗相关的内容，全部实现电子化处理，患者只持一张诊疗卡，便可进行挂号、就诊、开具申请单、抽血检验、功能检验、结果查询与打印等一系列诊疗活动[43]。对于医学实验室来说，无纸化是指通过信息系统来实现实验室管理流程的数字化，是将整个检验科的生态系统整合进LIS系统中，包含但不限于检验申请、排队抽血、报告打印等基本功能。

2.2 数据分析处理

在过去，医疗机构为了集中一切资源，不得不购买和维护所有必需的硬件和软件，并招募大量医护人员，却不考虑这些资源是否全部使用，并且安全性通常较差。而现如今医疗大数据增长快速，且来源于多个位点，因此数据会呈现出半结构化、非结构化等特点，使得通过传统的数据库管理工具来管理数据变得非常困难。分布式计算模式的引入成功地缓解了医疗大数据增长快速的挑战。如云计算、MapReduce、Hadoop等分布式系统在一些医疗健康研究单位对于存储和计算大量数据的使用已经变得非常普遍。

医疗机构通过使用云计算技术来处理和交付数据，并将数据分析成有意义的信息，这可以缓解数据呈现半结构化和非结构化的挑战。通过使用云计算技术，医疗机构只需为使用的资料和服务支付费用，例如存储、应用程序和基础设施服务[44]，并且医疗健康组织只需要更少的技术来管理和处理产生的医疗数据[45]。云计算技术已经获得医疗健康组织持续的关注来克服许多互联网医疗障碍[46]。Peddi等人[47]我们为移动电子健康多媒体应用提出了一种基于云的智能数据处理代理机制，并通过将其与我们的特定电子健康

移动应用程序 Eat Health Stay Health（EHSH）集成，展示了其可行性和性能。EHSH 包括食品图像处理，食品识别和分类的深度学习以及卡路里估算。这些数据处理功能是计算密集型的，并且需要大多数当前移动设备无法提供的资源。为了克服计算密集型资源问题，将部分或应用程序完全卸载到云将使负载远离移动设备并改善移动设备的性能和资源消耗[48-49]。IBM 提供了一种用于健康文书工作内容管理的工具。该工具可帮助医疗保健提供者记录患者的健康数据，并提供数据分析和可视化工具。

Hadoop、MapReduce 等其他分布式系统可以分享和共用多个位点和资源的数据。Hadoop 能帮助人们整理病历，与医生、患者和组织沟通，处理实验室结果、临床数据、影像学报告等医疗设施输出的文档[50]。Yao 等人[52]研究了基于 Hadoop 的应用程序的实际案例，以智能地处理以来欧大数据并揭示 HIS 用户行为的一些特征。他们基于由日常工作中使用的 HIS 生成的结构化，半结构化和非结构化数据来研究用户行为。另外他们还构建了一个五节点 Hadoop 集群来执行分布式 MapReduce 算法。与单节点算法相比，此分布式算法有望促进医疗服务研究和临床研究中的有效医疗大数据处理。Sebaa 等人[53]基于当前对大数据建模和工具的研究，开发了基于 Hadoop 的架构和医学大数据仓库的概念数据模型，并已经证明 Apache Hive 中的主键和外键问题可以使用嵌套分区来解决，通过设计和实施数据仓库平台以确保公平分配卫生资源，将拟议的解决方案应用于所提出的案例研究。Bao 等人[54]提出了一个基于 Hadoop 和 HBase 的新颖的界面应用程序接口（API），在异构集群中提供更好的数据分配，以便通过离线负载均衡器加快处理速度，并描述找到大数据集的拆分块大小的最优标准，另外使用 HBase 表方案启用快速数据查询并启动 MapReduce 性能。Antony 等人[55]选择糖尿病作为 MongoDB 分析的对象，把各种资源实时产生的数据收集起来并输入到 MongoDB 数据库中，MongoDB 能够实现快速存储和查询，结果由 MapReduce 自动产生。Antony 首先通过 MapReduce 得到病人的详细信息，如患者 ID、胆固醇水平等，这些信息存储在 MongoDB 数据库中，然后在利用朴素贝叶斯分类器筛选出信息特征，再对这些特征信息进行分类，最后得出各种特征的风险评估概率，得出预测结果。Wang 等人[56]介绍了 Hadoop- GIS（Geographic Information System），一种可扩展的高性能空间数据仓库系统，用于在 Hadoop 上运行大规模空间查询。Hadoop-GIS 通过倾斜感知空间分区，按需索引，可自定义空间查询引擎 RESQUE，MapReduce 上隐式并行空间查询执行以及通过处理边界对象修改查询结果的有效方法，在 MapReduce 上支持多种类型的空间查询。为了加速计算密集型几何运算，基于 GPU 的几何计算算法被集成到 MapReduce 管道中。实验证明，Hadoop-GIS 具有高效性和可扩展性，并且在计算密集型空间查询方面优于并行空间 DBMS。为了加速计算密集型几何运算，基于 GPU 的几何计算算法被集成到 MapReduce 管道中。实验证明，Hadoop-GIS 具有高效性和可扩展性，并且在计算密集型空间查询方面优于并行空间 DBMS。为了加速计算密集型几何运算，基于 GPU 的几何计算算法被集成到 MapReduce 管道中。实验证明，Hadoop-GIS 具有高效性和可扩展性，并且在计算密集型空间查询方面优于并行空间 DBMS。

大数据的分析在医疗领域的应用包含很多的方向，比如临床操作的比较效果研究、

临床决策支持系统、医疗数据透明度、远程病人监控、对病人档案的先进分析；临床试验数据分析、个性化治疗、疾病模式的分析等；还有患者临床记录和医疗保险数据集等。大数据的分析和挖掘技术的运用可以在一定程度上帮助医疗行业提高生产力，改进护理水平，增强竞争力。比如有大数据参与的比较效果研究可以提高医务人员的效率，降低病人的看病成本和身体损害；另外，利用大数据对远程病人的监控也可以减少病人的住院时间，实现医疗资源的最优化配置，在使用远程监护系统实现疾病预防的过程中，不仅能够降低病人出现意外的风险，同时也可以节约医疗资源，创造社会和经济价值。

2.3 数据保护

在医疗健康大数据快速增长的同时，其数据安全问题也日益凸显。在数据采集、存储、应用等各个环节均存在不安全因素。而随着移动医疗、AI 医疗影像、电子病历等数字化程序的普及，如果不加强对其数据的保护，医疗数据很有可能被泄露。例如，2017 年，亚马逊数据库存储的 47GB 医疗数据意外泄露，其中包含 315363 份 PDF 文件，预计至少有 15 万名患者受此影响。这些文件的内容不仅包括患者的验血结果、姓名、家庭住址等个人信息，另外还有医生和他们的病例管理笔记等。2018 年，新加坡遭受了史上最严重的网络攻击，近 150 万人的医保资料遭到泄露，这些信息包括病患姓名、国籍、地址、性别、种族和出生日期等。各类危害数据安全的事件层出不穷，医疗健康大数据的安全保护刻不容缓。

许多研究者开始利用各种信息安全技术来保障医疗健康大数据的安全性，如，Sriti 等人[57]使用变换域技术为远程健康应用提供了一种强大而安全的水印方法。患者报告/身份嵌入到宿主医学图像中以用于认证，注释和识别。为了获得更好的机密性，以不太复杂的方式在水印图像上应用基于混沌的加密算法。实验结果清楚地表明，所提出的技术对于各种形式的攻击具有高度鲁棒性和足够的安全性，而水印和覆盖图像之间没有任何明显的失真。Kai 等人[58]提出了一种基于区块链的高效隐私保护和共享方案，通过结合访问控制协议和加密技术，可以保证数据中包含的用户隐私。这种不向半信任的第三方上传数据的方法确保其他代理商不能访问患者的原始医疗数据。在分类账上使用面包屑的设计可以快速检索加密信息的位置并提高系统的效率。Gong 等人[59]讨论了当前智能医疗系统存在的主要问题。然后他们设计并完成了一个基于轻量级私有同态算法和 DES 改进的加密算法的原型系统。最后，基于上述工作，他们设计并完成了基于软件和硬件的原型系统。Hu 等人[60]提出了一种基于云计算的物联网传感器方案，该方案涉及数字包络，数字认证，签名，时间戳机制和非对称加密技术，以监测老年人的生物数据和其他个人信息。拟议的方案可以提供更灵活和准确的医疗服务，同时减少医疗资源的浪费。Li 等人[61]提出了一种新的以患者为中心的框架和一套机制，用于对存储在半局域服务器中的个人健康记录进行数据访问控制。为了实现对个人健康记录的细粒度和可扩展的数据访问控制，他们利用基于属性的加密（ABE）技术来加密每个患者的个人健康记录文件，并通过利用多权 ABE 来保证高度的患者隐私。Miao 等人[62]设计了一个安全

的加密原语，称为基于属性的多关键字搜索，通过多主机设置加密的个人健康记录，通过基于密文策略属性的加密（CP-ABE）支持细粒度访问控制和多关键字搜索。另外对现实世界数据集进行了实证研究，实验结果显示其在广泛实际情景中的可行性和实用性。Chandrasekhar 等人提出了基于云的健康信息交换（HIE）的授权协议，填补了加密和非加密方法之间的空白。该系统由三个主要组成部分组成：健康信息交换云，医疗保健组织（HCO）和患者。他们开发了一种新的基于代理签名的协议，基于一种新颖的基于离散日志的陷门散列方案，通过基于云的 HIE 实现经过身份验证和授权的患者健康信息的选择性共享。根据他们详细的安全性和性能分析，所提出的协议使用其基于陷门哈希的代理签名方案，在可证明安全的同时实现了最佳的全面性能。

综上，由于医疗健康行业特性、医疗健康大数据自身特征、信息技术的自身缺陷等因素，医疗健康大数据的不安全状态将长期存在，很难彻底根除医疗数据被盗用、滥用等问题。总体来说，医疗健康大数据的安全保护应该注意以下六个方面：

1）防窃取。避免不拥有数据所有权的地方将数据据为己有；

2）防滥用。避免危害国家安全、违反社会伦理、侵犯个人隐私权益的应用；

3）保隐私。避免个人隐私或商业秘密被泄露；

4）保价值。保证数据资产的价值不折损；

5）防超限使用。外部合作方不得将数据超出授权界限使用；

6）保物理安全。避免数据非正常丢失、损伤或不能读取。

2.4　数据应用

医疗大数据本身来源广泛，涉及临床诊断、临床治疗、制药、患者病历、健康管理等众多领域的数据信息。在日常生活中，我们每个人都不可避免地与这些数据息息相关。因此，通过对这些医疗健康数据进行有效的应用，可以实现其在医疗领域的价值。将医疗大数据与机器学习、深度学习等技术与循证医学、影像、组学等学科相结合，可以为健康管理、辅助诊断等场景提供解决方案。打通底层数据，构建互联互通的数据平台，可以优化诊疗疗程、提升医疗行为的效率。因此，我们就医疗大数据在医学影像、辅助诊断、药物研发、健康管理和基因测序等五个重点领域的应用展开探讨。

2.4.1　医学影像

影像组学这一概念起源于肿瘤学领域，之后其外延扩大到整个医学影像领域，即从 CT、MRI、PET 或 SPECT 等影像中高通量地提取大量影像信息，实现感兴趣区（通常指病灶）图像分割、特征提取与模型建立，凭借对海量影像数据信息进行更深层次的挖掘、预测和分析来定量描述影像中的空间时间异质性，揭示出肉眼无法识别的图像特征。影像组学可直观地理解为将视觉影像信息转化为深层次的特征来进行量化研究。理解医学图像、提取其中具有诊断和治疗决策价值的关键信息是诊疗过程中非常重要的环节。以往，医学影像前处理与诊断需要 4 ~ 5 名医生参与。而基于影像组学与大数据技术，训练

计算机对医学影像进行分析，只需 1 名医生参与质控及确认环节，这对提高医疗行为效率有很大帮助。影像组学解读“数据语言”、AI 辅助阅片将作用于疾病早期筛查及诊断，是医学影像的发展方向。

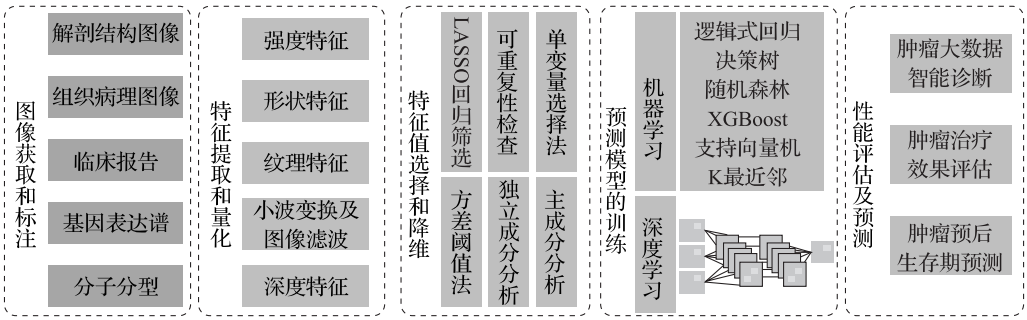


图 1　基于影像组学与医疗大数据的医学影像分析流程

将医学大数据应用到影像医学中的处理过程见图 1，首先我们获得 CT 等医疗影像并对其进行标注，在进行特征提取和量化，然后选取特征值并对数据降维，再然后我们利用机器学习、深度学习等方法进行预测模型的训练，最后对模型的性能进行评估。如，Zhang 等人[63]使用深度学习来描述脑肿瘤分割（Brain tumor segmentation），构建了一个 6 层密集卷积神经网络，它以前馈方式将每个层连接到每个后续层。这种特定的连接架构可确保网络中各层之间的最大信息流，并加强层与层之间的特征传播。用 2017 年多模式脑肿瘤图像分割挑战（BRATS）提供的成像数据[64]对此方法进行了训练和评估。脑肿瘤的分割迄今为止有多种半自动或自动脑瘤分割方法。多数分割方法在训练卷积神经网络（CNN）时使用图像补丁（Image patches）作为输入，他们首先将图像块分类为不同的肿瘤类别，例如坏死，水肿和健康组织，然后根据这些预测类别标记图像块的中心体素，一个缺点是这些方法没有考虑肿瘤的外观和空间一致性。此模型考虑脑肿瘤的空间一致性和其外观信息，使用全尺寸 MRI 图像作为输入来训练 CNN，神经网络的特殊连接模式极大地改善了网络中各层之间的信息流，使网络易于训练。Godinho 等人[65]提出了一种创建模拟医学成像库（Medical Imaging Repositories）的方法，基于模型数据集的索引（Indexing of Model Datasets），模式的提取（Extraction of Patterns）和研究生产的建模（Modelling of Study Production）。系统根据实际存储库的代表时间窗口创建模型，并根据正在进行的研究需求进行扩展。此外，该解决方案提供了减少生成的数据集大小的不同方法。图片存档和通信系统（PACS）是管理医学成像数据和相关工作流程的信息系统的通用名称[66]，典型的 PACS 环境包括三大类应用程序：图像存储库，采集设备和查看器应用程序。该系统能够仅使用从真实环境中提取的代表性数据部分生成大数据集，不会对生产过程产生任何影响。生成的存储库是支持 PACS 技术的研究，开发和验证的元素。所提出的方法非常适合需要模拟 DICOM 元数据[67]的任何场景。此方法的结果强调了结果数据集的质量使得它们可以用于在负载环境中测试多个 PACS 服务，例如存储，查询和检索。拟议系统的一个主要方面是它的直接效用。Dicoogle[68]一直在其存储库的

DICOM 图像中包含的所有元数据上提供搜索和检索服务。其索引引擎的开发受到连续验证过程的影响，以评估其对增加负载的响应，以及各种补丁如何影响系统行为。此过程需要强大的数据集来验证解决方案。

深度学习方法一般用于完成医学成像中的大量任务，包括图像生成步骤，例如图像重建。因此，这些技术和应用不仅会对图像分析产生潜在影响，而且会影响整个医学影像和医疗保健[70-71]。该领域临床应用的主要挑战是检测组织或器官异常。算法必须能够准确，精确地检测病变，以提供一致的临床表现。将需要各种临床试验和医学成像分析研究来验证这些学习算法方法[72]。

机器学习已越来越多地用于放射学，因为在医学图像中呈现的典型成像对象（例如病变和器官）在大多数情况下太复杂而不能通过某个简单的方程或手工制作的模型可靠地表示。这些简单的模型以及从它们计算出的简单特征通常不能提供辨别能力，以便可靠地检测和分类具有可变指示的个体患者图像中的感兴趣对象[73]。新的放射成像技术，重建和后处理技术提供了新的和大多数非线性图像输出[74]。与传统的滤波反投影重建方法相比，这种发展的一个例子是 CT 中的迭代重建。这些新方法还要求对临床相关图像质量进行更全面的描述，最好通过客观量化来实现一致的图像分析方法[75]。

传统的大医学图像数据分类方法主要是基于图像灰度特征的变化，提取边缘和轮廓特征信息，或者在医学图像坐标集之间进行转换。然而，算法复杂，实时性能差，分类速度慢，准确性低。Zhu 等人[76]结合深度学习算法，提出了基于偏微分方程的大医学图像数据的分类研究，并利用大医学图像处理中的偏微分方程提取医学图像的纹理特征。此外，根据医学图像对比度调制的纹理特征，该方法滤除了图像噪声干扰。基于深度学习算法，图像距离分层，目标对象大小，通过拟合等信息，实现对大医学图像数据的准确分类。该方法传统方法和方法对大医学图像数据分类过程中的误差率进行统计计算。该方法可以将分类误差控制在较低水平，优于传统方法。仿真和对比实验结果表明，与传统方法相比，该方法在医学图像大数据的分类速度，KAPPA 系数（用于一致性检验，也可以用于衡量分类精度）和分类误差性能方面具有一定的优势[77-79]。在不同的噪声干扰水平下，研究了该方法和传统方法的大医学图像数据分类精度。同时，还研究了所设计的滤波器干扰性能。在不同的噪声下，大医学图像数据的分类精度始终保持在 95% ~ 98%。分类精度结果基本上没有波动，表明所提出的方法不受噪声强度的影响。证明了该方法能够有效地滤除噪声。传统算法的分类精度随着噪声的增加而降低，表明传统方法不考虑噪声对大图像数据分类精度的影响，使其受到较大的噪声波动，性能不稳定，不利于实际应用。

医学影像现已成为人工智能在医疗领域最热门的方向，但在实际应用过程中还是存在一定挑战，例如，数据获取及数据标注问题、缺乏行业标准、注册审批缺乏指导原则、技术创新问题等等。但随着人工智能相关技术的不断发展，国家相关政策的不断完善，相信人工智能与医学影像的结合能够推动医学的不断发展。

2.4.2　辅助诊断

人类目前临床疾病约有 3 万种，常见的疾病约 3000 多种，且以每年 20 ~ 30 种疾病

数量递增。国内当前医卫工作的主要问题是医疗资源分布不平衡，基层医疗机构需要的全科医生缺乏，培养周期长且难留住人才；病人满意度低，误诊率高，医疗事故死亡率呈不断增长态势。医疗行业已经普遍认为解决基层医疗机构需要的全科医生，贯彻实施分级诊疗制度；提高医疗质量、控制人为医疗差错、提高病人安全为优先和急迫的任务。疾病多样性、体征多样性、药物多样性的发展趋势，急需计算机来辅助人脑做知识存储和快速检索，大部分的人为因素所致医疗差错也可以通过计算机辅助系统避免。通过引入大数据临床诊断辅助决策人工智能平台可以实现以大数据技术作为使能手段，提高医生的诊断能力和诊断效率、提升病人就诊满意度、拉动医院知识能力的传承和积累！

CDSS 系统将人工智能技术应用于医疗行业，核心算法融合一系列人工智能 VQ、RBF、BP 等算法，并融合循证医学和经验医学两大模型，显著提高临床疾病的诊断能力。

围绕医疗大数据、算力和算法这三大关键挑战问题，2017 年 8 月国家超级计算长沙中心、全国高等医学院校诊断学联盟和湖南智超医疗科技有限公司签订了关于建设“医疗大数据中心与人工智能辅助诊疗系统”三方合作协议。合作目标是建立医疗大数据中心并联合探索和研发人工智能机器人医生，其中基于国家超级计算长沙中心，建设面向全国的医疗大数据中心，并建设辐射全国的医疗大数据产学研基地及孵化中心；全国高等医学院校诊断学联盟负责协调联盟内各个医院和专家，提供医院案例和医疗数据等临床信息，并参与指导和组建医疗大数据中心；湖南智超医疗科技有限公司负责人工智能机器人医生的研发和系统测试，项目筹建和运营管理，并筹建医学大数据企业孵化器及“双创”基地。

该合作是产学研跨界合作的具体体现，有效地解决了数据、算力和算法这三大挑战。为湖南智超医疗科技有限公司研发 CDSS 系统提供了全面的支撑。智超公司核心研发人员结合循证医学和经验医学两大模型，利用医疗大数据、超级运算和人工智能等技术，并联合研发具有自然语言的处理系统、基于循症的自动问答系统、基于用户反馈的机器学习和系统优化体系三大系统，从“诊前”、“诊中”、“诊后”三个阶段，全力支持和辅助医院和医生进行全方位、全科的智能诊疗。加强挂号、诊疗决策和处方精准度，提高医疗服务效率和质量，连接分导诊、临床检查与临床知识，提高临床决策精准度，提升临床诊疗效率，减少医院和患者的时间和经费等开销，为国家卫生系统、分级诊疗体系的做出有益补充。

Guo 等人[80]开发了基于透明学习的开放和可扩展的医疗辅助诊断的 MADP 平台。该平台基于医学图像进行辅助诊断，利用深度学习对大量数据进行测试和分析，并且建立了基于 TensorFlow 的疾病辨别模型，实现了诊断的稳定性和精确性。MADP 作为一个开放的平台，提供了云服务的功能，任何医院和研究机构都可以使用该平台来提升他们的医疗经验，同时进行辅助诊断。甲状腺疾病的发病率在全世界都呈现显著增加的趋势，而且很多甲状腺疾病是阴性的不易被发现。通过辅助诊断的措施，可以帮助患者快速发现疾病，包括节点的良性和恶性的预测，疾病分类，推荐检查指标等。Chen 等人[81]提出了一种基于张量分解的多任务模型，通过对患者自身较高纬度的数据进行处理和分析，

从来提高预测的准确性，并且加入了主动学习的方式来有效使用样本，从未降低人工成本。从结果来看，研究提出的模型提高了甲状腺疾病预测的效率和准确性，起到了辅助诊断的作用。前列腺癌（PCa）已经成为世界上第二大常见的恶性肿瘤，也是造成男性发病甚至死亡的主要原因之一。大多数发展中国家存在大量的患病人群，但是医生诊断面临着复杂和重复的工作，误诊率很高。通过对中国三甲医院中 1 933 535 项医疗住院数据信息的研究，一方面，通过对各种诊断间隔之间的生理指标进行对比分析，可以检测 PCa 的进展，另一方面，结合统计分析和医疗数据决策为医生提供快速准确的医疗选择。通过实验表明，基于大数据的 PCa 数据分析系统，可以为 PCa 提供辅助诊断，提高医师的工作效率，同时尽可能减少误诊率[82]。

总体来说，辅助诊断能够提高医院诊疗准确度，提升医院接诊率、医疗服务效率和质量，提升群众医疗服务满意度，全面减少医生工作量、减少病人等待时间，缓解医患矛盾、平衡医疗资源分布，解决我国医疗资源缺乏问题，降低社会疾病就诊平均费用。

2.4.3　药物研发

目前医药研发主要存在研发周期长、研发成本高、研发失败率高等问题，大数据可以应用到医药研发多个阶段以缓解痛点。在临床前研究阶段，可通过大量的文献挖掘和生物信息分析，较快确认药物作用靶点、提升化合物筛选效率。在临床试验阶段，通过大数据优化临床试验设计，提高药物试验有效应答率，筛选受试对象，降低临床试验成本，缩短研发时间。在上市后再评价阶段，可较快实现不同数据库不良事件的识别、计算不良事件发生率，收集大量用药反馈并做出分析，指导后续研发设计。另外，真实世界研究日益成为医药研发的一大趋势，建设合规的真实世界数据查询平台可以很大程度地提升真实世界研究效率和准确率。然而，大数据在医药研发场景下的应用却受到两个因素的极大制约：第一，大数据助力医药研发领域的方式更多的是提供基因组数据作为底层支持，而基因组数据具有极高的隐私特性及敏感性；第二，我国在原研药方面力量较薄弱，传统医药研发尚未扎实根基，此时谈大数据在该领域的应用为时尚早、步伐过快。通过利用大数据开发个性化药物，针对不同患者的实际情况提供个性化的治疗方案，不仅可以减少药物副作用，而且可以降低误诊率。

分子的多重药理的理解变得越来越重要。由于副作用和毒性等问题，药物有时会退出市场。这些问题通常是由于分子与体内多种蛋白质靶标相互作用的能力而产生的，称为多药理学。此外，由于与多种生理途径相关的疾病的复杂性质，正在设计更多的药物以与这些途径中的多个靶相互作用，以便对单个分子具有增加的作用[83]。即使在考虑使用经批准的药物重新利用时，例如，如果该药物最初设计用于外周适应症，现在它的安全性可能是不合适的，现在它正在针对 CNS 中的替代适应症进行调整，其中神经毒性评估是现在更相关的。虽然化合物的基本多药理学保持不变，但作用部位可能会改变哪些靶标成为导致某些毒性的必要脱靶。常用于确证生物活性结构的方式是相似性原理。如果化学结构保持很高的结构相似性，则认为其具有更高相似生物活性的可能性。据报道，基于路径指纹[84]的 Tanimoto 相似度为 0.85 足以保持相似生物活性的高概率。即便如此，

已经表明30%的近邻将维可以保持相似的生物活性[85]。然而，高度相似的分子结构可能完全失去效力，这种现象被称为活性断崖[86]。通过传统的相似性度量可以认为两种化合物是相同的，但它们的活性变化很大。最重要的是要记住，没有单一的分子相似性度量，相似性的值，高或其他，将完全取决于所使用的描述符和相似系数，最常见的是Tanimoto系数。此外，成功的可能性将取决于所需的终点和溶液空间的形态。无论是探索还是开发，分子设计最有效的方法之一都是从头分子设计的概念[87]。从头设计的一种流行方法是迭代进化算法，例如遗传算法。在这种类型的实现中，对候选解决方案的群体（在这种情况下为分子结构）进行迭代评分以确定其适合性，并且这些得分及其结构用于通过重组和突变的计算类似物生成新的群体。一旦满足适当的成功条件，例如建模的生物学终点，该过程终止并返回最佳解决方案。合成可及性的评估是合成有机化学家在化合物优化过程中始终存在的事情，通常可以使他们选择使用他们熟悉的经过试验和测试的反应的分子[88]。此外，药物化学中使用新反应的速度随着时间的推移而降低，在过去的20年中仅发现了两种常用的反应[89]。当选择候选药物的分子时，这种偏差导致制药公司内可获得的化学空间缩小。相比之下，从头设计遭受相反的挑战，尽管方法可以产生数百万个理论分子，但实验室化学家可能只有一小部分可以合成获得，而不考虑合成可行性。其余部分需要如此多种不同的起始材料和反应路径，以使其在时间和资源上过于昂贵。这一难题导致了许多评估分子合成可及性的方法的产生。Ertl和Schuffenhauer开发了一种方法，该方法将分子复杂性得分与有益分子的分数相结合，所述分子在已知的合成可及分子内有益于分子[90]。这些方法可以将大型从头数据集分类为更小，更易于管理的数据集，以便合成化学家进行更严格的审查，但需要进行大量的进一步研究。

2.4.4 健康管理

在健康医疗大数据的驱动下，健康管理学也遇到了空前的发展机遇。健康管理（Health management，HM）是对个体（包括健康个体、亚健康个体与患者）或群体（包括健康人群、亚健康人群、患病人群）的健康危险因素进行全面监测、分析、评估、预测，旨在提供健康咨询与指导以及对健康危险因素进行干预的全过程。它通过系统检测疾病危险因素，评估发病或预后风险，实施有针对性的预防性干预，以阻断、延缓甚至逆转疾病的发生和发展，实现维护健康之目的。完整的健康管理应包括采集个人健康信息（如健康体检）、进行健康及疾病风险评估、实施健康干预3个基本环节。其核心是以调动多方面的积极因素为宗旨，对危害健康或导致疾病的危险因素进行全面管理。

Timothy等人[91]描述了利用大数据分析技术为400万拥有AARP Medigap计划的65岁及以上的被保险人执行全面的人口健康计划，包括健康计划、整体护理协调计划、两条电话咨询热线、礼宾部对保险和医疗保健需求的支持，以及旨在帮助减少不必要的急诊室就诊的计划等。结果表明，人口健康管理计划有助于帮助老年人更健康地生活。其中，在2009-2011年期间，一些计划促进了医疗保健的一些改进；每1000名加入急诊决策的被保险人比没有参加该项计划的被保险人减少了178次急诊就诊；加入抑郁症管理

计划减少了59%的被保险人的抑郁症症状。在实施计划过程中产生的数据将被分析，用作后续计划实施和质量的改进。Jannet等人[92]使用他们的电子病历的数据来建立健康管理计划。在退伍军人综合服务网络（VISN）21中，药剂数据分析师使用SQL语言编写复杂程序，开发了用于人口健康管理的数据量丰富的分层仪表板，前线临床医生团队使用之后提供持续性的反馈，以改善退伍军人的健康状况。医疗的质量、安全和价值是VISN 21衡量人口健康计划的标准，衡量指标、基准、特别工作组、目标群众和团队的明确是应用这些工具的关键。Karen[93]以健康管理中的“护理计划”（也称为“协作实践组”、“自我管理组”）为例，介绍了护理人群、护理功能、护理范围、角色分布、以及护理的步骤，表明健康管理旨在改善患者的治疗效果。并列举了目前的健康管理机构，如PACE、CHS、Carr、Atrius Health等，详细介绍了它们的服务人群、角色分布、服务范围等。另外还介绍了夏威夷的一份将无家可归作为新的医学问题的创新提案，表明人口健康越来越受到重视，最终得出在人类医疗保健史上人口健康管理将会更加重要、开展人口健康计划十分必要的结论。Lv等人[94]创建了一种移动健康护理模式，为中国医疗保健提供者的局限性问题和医疗保健资源分布不均问题提供了解决方案。该平台迎合移动互联网通信快速发展和医疗服务需求不断增加的趋势，面向中国医生和慢性疾病患者，提供的服务包括医疗咨询、电子病历、慢性病管理等，能更好地应对慢性病和人口老龄化带来的日益沉重的负担。移动医疗方案的实施受到中国卫生政策的大力推动。

2.4.5　基因测序

随着人类基因组测序技术的飞速提升、生物医学分析技术的快速发展和大数据分析工具的日益完善，我们正进入全新的医疗健康时代——精准医疗。精准医疗是一种基于“个人”的定制医疗模式，其以个体的组学信息和遗传信息为基础，以环境、生活方式、既往病史及诊疗方式等为跟踪对象，搜集全方位、可量化、有前瞻性和时效性的个体数据，通过数据的综合分析、挖掘形成有价值的医学信息，最终设计出针对个体的最优解决方案。而基因测序技术凭借灵敏度高、精度和通量高等优势，成为基因检测技术中获取人体基因组数据的主流技术，通过将基因组数据与无线生物传感器获取的生命体征信息（如血压、心跳、脑电波、体温等），成像设备中的个体信息（如CT、MRI、超声等）以及传统医学数据相结合，精准医疗为个体提供全新的定制医疗。精准医疗的基础在基因测序，基因测序是建立“组学”大数据库和分析的基础，推动精准医疗实现“同病异治”和“异病同治”。基因测序相关技术的发展有两个要素：一是构建“组学”大数据样本库，如基因组学、转录组学、蛋白组学等；二是探究基因型与样本表型的关联。

自Pauling等人[95]确定镰刀型细胞贫血症（sickle cell anemia）的分子遗传机制以来，目前已被确定的遗传病超过5000种，主要包括单基因遗传病、多基因遗传病、染色体异常遗传病等三大类。高通量测序和生物大数据分析已成功用于多基因遗传病检测、无创产前筛查（NIPT）[96]和胚胎植入前遗传学检测（PGD）[97-100]等临床实践，取得了良好的社会效益和经济效益。此外，高通量测序在检测外周循环血液中的肿瘤细胞或肿瘤DNA/RNA，用于早期肿瘤筛查、检测肿瘤复发、观察临床疗效等方面也具有其独特的

优势[101-102]。

我国科学家已经在疾病队列人群的全基因组关联分析（genome-wide association study，GWAS）等多组学研究中积累了丰富的工作经验，为阐明复杂疾病发生的分子机制提供了重要的理论依据。1998 年，中南大学夏家辉院士等人[103]成功地克隆人类遗传性神经性耳聋的致病基因 GJB3。交通大学贺林院士的团队率先完成第一例孟德尔常染色体遗传病 A-1 型短指（趾）症致病基因的克隆与突变检测[104-105]；通过对患病家系的遗传连锁分析，定位了第一例以中国人姓氏命名的罕见恒齿缺失的孟德尔常染色体显性遗传病“贺-赵缺陷症”的致病基因[106]。安徽医科大学张学军教授等人[107-110]在银屑病、系统性红斑狼疮、麻风、白癜风等复杂疾病的 GWAS 研究中发现一系列疾病易感基因。中山大学肿瘤医院曾益新院士等人[111]开展的鼻咽癌 GWAS 除证实人类白细胞抗原与鼻咽癌的关联性外，发现多个新的易感基因。军事医学科学院贺福初院士和周钢桥教授开展的肝脏蛋白质组和肝癌的 GWAS 研究，发现乙型肝炎病毒相关肝癌的易感基因[112]。中国医学科学院基础医学研究所张学教授[113]对于遗传性脱发相关药物靶点以及他与沈岩院士合作的反常性痤疮家族基因的研究[114]。中医学科学院肿瘤研究所詹启敏院士团队在广东潮汕地区开展食管鳞状细胞癌研究[115]；林东昕院士课题组开展的肺癌、食管癌流行病学研究[116]；林东昕院士与郑州大学王立东教授等[117-118]对河南安阳地区的食管癌队列研究等。

一些复杂疾病往往是由遗传、环境等多重因素导致的，仅依赖于临床上的影像诊断和病理分析等难以对疾病做出准确的诊断和分类。综合分析多种组学数据和临床数据，能够更加准确地确定各疾病的亚型。在乳腺癌中，不同的分子亚型在临床症状、治疗反应和效果方面有明显差异[119]。由转录组数据确定的胰腺癌三种亚型，患者治疗后的反应具有差异性[120]。而不同分子亚型的结直肠癌患者的存活时间显著差异[121]。除癌症外，这种综合分析也被用于其他复杂疾病的诊疗，如自闭症谱系障碍。综合分析外显子组数据、基因表达谱、蛋白质表达谱以及临床上的心理测试和影像诊断，研究人员提出了新的自闭症亚型，这一成果不仅加强了自闭症诊断，也为后期选择有效的治疗方案提供了依据[122]。

3 发展趋势与展望

医疗健康大数据是一门融合了多种学科的科学。本文是从医疗大数据所面临的挑战这一角度，主要介绍了医疗大数据在数据采集、分析、保护和应用四个方面的最新进展。

合规性是医疗大数据领域的重要问题。医疗大数据采集及管理、分析的任一环节都存在合规性问题，相关主体需要根据从事的业务领域关注相应的合规义务。需要规范数据质量、数据来源合法性，数据采集合规性，个人信息授权和脱敏化处理的保证责任。获得优质数据是企业在挖掘医疗大数据价值方面制胜的关键，把握优质医院资源将使企

业在该领域拥有先发优势。医疗大数据的真正落地需要政府、医院和企业三方共同合作实现，政府负责制定相应的法律法则、标准制度、管理要求、监督规范，同时要消除信息不对称、资源不均衡。医院提供医学专业知识并进行合规采集、存储、传输相关医疗数据。企业则负责前沿技术研发并承担一部分数据采集、存储、传输、追踪的任务，提升市场化竞争实力，为挖掘医疗大数据的价值提供支持。医疗健康大数据将搭上云计算、人工智能等技术的“高速列车”，海量的医疗大数据需要强大的计算能力、存储能力与前沿的分析技术。云计算能够提供算力、存储能力支持，人工智能的实现离不开底层数据作为“原材料”。在医疗中运用大数据，重点可以辅助各种临床诊断及临床治疗，同时对辅助临床路径也有一定的优化作用。临床诊断治疗及临床路径分析的基础是电子病历，大数据可以对医疗中的各种临床病历及各种辅助临床路径进行分析，其中分析内容包括具体的治疗效果及具体的费用，从而得到最优的临床诊断治疗方法以及临床路径。但是仅依靠大数据技术对这些数据进行分析是不够的，还需要在结合人的各种经验的基础上得出最佳的临床治疗方案，从而使实际临床效果得到很大的提升。

总体来说，医疗健康大数据将更快发展和更广泛应用，其电子数据将会向着精细化、智能化和便捷化的方向发展，互联网环境下更有助于实现个性化与社会化的健康管理制度，医疗健康大数据将更加注重开放共享和隐私保护，人工智能的发展与医疗健康大数据的结合将驱动临床决策支持和精准医学研究。

4 结束语

医疗健康大数据在医学影像、辅助诊断、药物研发、健康管理、基因测序等医疗各方面都有着非常积极的作用，尽管现阶段大数据技术在医疗信息化中的应用还面临着一些关键问题，但只要我们坚持以大数据为基础，对各类医疗数据进行不断的整合分析及利用，就可以促进整个医疗行业的进一步发展。

本报告的整理得到科技部重点研发计划“精准医学大数据的有效挖掘与关键信息技术研发-课题5 高通量生物医学数据高效算法与并行计算”（编号：2018YFC090002）、国家自然科学基金“面向大规模异构体系结构的生物医药大数据并行算法及优化关键技术研究”（编号：61772543）的支持。王小奇、王建民、高明玉、袁玉洁、刘孝炎、叶琪等同志参加了整理工作，在此一并致谢。

参考文献

[1] Snijders C, Matzat U, Reips U D. Big Data: big gaps of knowledge in the field of internet science[J]. International Journal of Internet Science, 2012, 7(1): 1-5.

[2] Sharma S, Mangat V. Technology and trends to handle big data: Survey[C]. 2015 Fifth International

Conference on Advanced Computing & Communication Technologies. IEEE, 2015: 266-271.

[3] Issa N T, Byers, Stephen W, Dakshanamurthy, Sivanesan. Big data: the next frontier for innovation in therapeutics and healthcare[J]. Expert Rev Clin Pharmacol, 2014, 7(3): 293-298.

[4] Ali A, Qadir J, Rasool R U, et al. Big data for development: applications and techniques[J]. Big Data Analytics, 2016, 1(1): 2.

[5] Ozkuran M A . A Study on Trends In Information Technologies using Big Data Analytics[J]. 2017.

[6] Xi W . British State Strategy of Developing Big Data[J]. Global Science Technology & Economy Outlook, 2013.

[7] Mozaffarian D, Benjamin E J, Go A S, et al. Executive summary: heart disease and stroke statistics-2015 update: a report from the American Heart Association[J]. Circulation, 2015, 131(4): 434-441.

[8] Mozaffarian D, Benjamin E J, Go A S, et al. Executive summary: heart disease and stroke statistics-2016 update: a report from the American Heart Association[J]. Circulation, 2016, 133(4): 447-454.

[9] Kuo M H, Sahama T, Kushniruk A W, et al. Health big data analytics: current perspectives, challenges and potential solutions[J]. International Journal of Big Data Intelligence, 2014, 1(1-2): 114-126.

[10] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential[J]. Health information science and systems, 2014, 2(1): 3.

[11] Farahani B, Firouzi F, Chang V, et al. Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare[J]. Future Generation Computer Systems, 2018, 78: 659-676.

[12] Wang Y, Kung L A, Wang W Y C, et al. An integrated big data analytics-enabled transformation model: Application to health care[J]. Information & Management, 2018, 55(1): 64-79.

[13] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads[J]. EMBnet. journal, 2011, 17(1): 10-12.

[14] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nature reviews genetics, 2009, 10(1): 57.

[15] Yadav M, Jhunjhunwala S, Phung Q T, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing[J]. Nature, 2014, 515(7528): 572.

[16] Park P J. ChIP-seq: advantages and challenges of a maturing technology[J]. Nature reviews genetics, 2009, 10(10): 669.

[17] Capper D, Jones D T W, Sill M, et al. DNA methylation-based classification of central nervous system tumours[J]. Nature, 2018, 555(7697): 469.

[18] Buisine N, Ruan X, Ruan Y, et al. Chromatin interaction analysis using paired-end-tag (ChIA-PET) sequencing in tadpole tissues[J]. Cold Spring Harbor Protocols, 2018, 2018(8): pdb. prot104620.

[19] Patrick, John W, Laganowsky, Arthur. Generation of Charge-Reduced Ions of Membrane Protein Complexes for Native Ion Mobility Mass Spectrometry Studies[J]. Journal of the American Society for Mass Spectrometry, 2019: 1-7.

[20] Molly C Goodier, Lun-Yi Zang, Paul D Siegel, et al. Isothiazolinone Content of US Consumer Adhesives: Ultrahigh-Performance Liquid Chromatographic Mass Spectrometry Analysis[J]. Dermatitis, 2019, 30.

[21] Bankman I N. Handbook of medical imaging[M]. Handbook of Medical Imaging. 2000.

[22] Guo, Zhe, Li, Xiang, Huang, Heng, et al. Deep Learning-based Image Segmentation on Multi-modal Medical Imaging[J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2019, 3(2): 1-1.

[23] Byun J H. Imaging: MRI with MRCP[J]. 2019.

[24] Duan Q, Flynn C, Niepel M, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures[J]. Nucleic Acids Research, 2014, 42(Web Server issue): W449.

[25] Gerstein M B, Bruce, C, Rozowsky, J. S, et al. What is a gene, post-ENCODE? History and updated definition[J]. Genome Research, 2007, 17(6): 669-681.

[26] Martin P C N, Zabet N R. Dissecting the binding mechanisms of transcription factors to DNA using a statistical thermodynamics framework[J]. BioRxiv, 2019: 666446.

[27] Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic signaling pathways in the cancer genome atlas [J]. Cell, 2018, 173(2): 321-337. e10.

[28] Turnbaugh P J, Ley R E, Hamady M, et al. The human microbiome project[J]. Nature, 2007, 449 (7164): 804.

[29] Zillner S, Neururer S . Technology Roadmap Development for Big Data Healthcare Applications[J]. KI - Künstliche Intelligenz, 2015, 29(2): 131-141.

[30] Zillner S, Oberkampf H, Bretschneider C, et al. Towards a technology roadmap for big data applications in the healthcare domain[C]. IEEE International Conference on Information Reuse & Integration. 2015.

[31] Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific Data, 2016, 3: 160035.

[32] Pollard T J, Johnson A E W, Raffa J D, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research[J]. Scientific data, 2018, 5.

[33] Euclid Seeram. Picture Archiving and Communication Systems: Physical Principles and Quality Control [M]. 2019.

[34] Arenson R L . Picture archiving and communication systems. [J]. Western Journal of Medicine, 1992, 156(3): 16-22.

[35] Naveen M, Gonzalez Glebys, Rodgers Richard, et al. Gestures for Picture Archiving and Communication Systems (PACS) operation in the operating room: Is there any standard? [J]. Plos One, 2018, 13(6): e0198092.

[36] Kovacs M D, Cho M Y, Burchett P, et al. Benefits of Integrated RIS / PACS / Reporting Due To Automatic Population of Templated Reports [J]. Current Problems in Diagnostic Radiology, 2018: S0363018817302050.

[37] Jin J, Zhang J, Chen X, et al. HIPAA-compliant automatic monitoring system for RIS-integrated PACS operation[J]. Proceedings of SPIE - The International Society for Optical Engineering, 2006: 61451B-61451B-9.

[38] McAdam, Alexander J, Burnham, Carey-Ann D. Total Laboratory Automation in Clinical Microbiology: a Micro-Comic Strip[J]. Journal of Clinical Microbiology, 2018, 56(4): e00176-18.

[39] Burckhardt, Irene. Laboratory Automation in Clinical Microbiology[J]. Bioengineering, 2018, 5(4).

[40] Naugler, Christopher, Church, Deirdre L. Automation and artificial intelligence in the clinical laboratory [J]. Critical Reviews in Clinical Laboratory Sciences, 2019, 56(5): 1-13.

[41] Archetti C, Montanelli, Alessandro, Finazzi, Dario, et al. Clinical laboratory automation: a case study [J]. Journal of Public Health Research, 2017, 6(1).

[42] Sediq M E, Hala G A. Designing an autoverification system in Zagazig University Hospitals Laboratories: Preliminary evaluation on thyroid function profile[J]. Annals of Saudi Medicine, 2014, 34(5): 427.

[43] Tsoromokos D. Design of an Innovative Information System for the Intensive Care Unit in a Public Hospital σ[J]. Studies in health technology and informatics, 2016, 226: 157-160.

[44] Gu D, Li J, Li X, et al. Visualizing the knowledge structure and evolution of big data research in healthcare informatics[J]. International Journal of Medical Informatics, 2017, 98: 22-32.

[45] Sultan, Nabil. Making use of cloud computing for healthcare provision: Opportunities and challenges[J]. International Journal of Information Management, 2014, 34(2): 177-184.

[46] Kuo M H. Opportunities and Challenges of Cloud Computing to Improve Health Care Services[J]. Journal of Medical Internet Research, 2011, 13(3): e67.

[47] Peddi S V B, Kuhad P, Yassine A, et al. An intelligent cloud-based data processing broker for mobile e-health multimedia applications. [J]. Future Generation Computer Systems, 2017, 66: 71-86.

[48] Peddi S V B, Yassine A, Shirmohammadi S. Cloud Based Virtualization for A Calorie Measurement E-Health Mobile Application[C]. IEEE International Conference on Multimedia and Expo Workshops. IEEE, 2015.

[49] A virtualization mechanism for real-time multimedia-assisted mobile food recognition application in cloud computing[J]. Cluster Computing, 2015, 18(3): 1099-1110.

[50] Toh S, Reichman M E, Houstoun M, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data[J]. Pharmacoepidemiology and Drug Safety, 2013, 22(11): 1171-1177.

[51] Batarseh F A, Latif E A. Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare[J]. Big Data Research, 2016, 4: 13-24.

[52] Yao Q, Tian Y, Li P F, et al. Design and Development of a Medical Big Data Processing System Based on Hadoop[J]. Journal of Medical Systems, 2015, 39(3): 23.

[53] Sebaa A, Chikh F, Nouicer A, et al. Medical Big Data Warehouse: Architecture and System Design, a Case Study: Improving Healthcare Resources Distribution[J]. Journal of Medical Systems, 2018, 42(4): 59.

[54] Shunxing Bao * a, Yuankai Huob, Prasanna Parvathanenib, et al. A Data Colocation Grid Framework for Big Data Medical Image Processing - Backend Design[J]. 2018.

[55] Antony Basco J, Senthilkumar N C. Real-time analysis of healthcare using big data analytics[J]. IOP Conference Series: Materials Science and Engineering, 2017, 263: 042056.

[56] Wang F, Aji A, Vo H. High performance spatial queries for spatial big data: from medical imaging to GIS[M]. ACM, 2015.

[57] Sriti T, Kumar S A, Prakash G S, et al. Multi-layer security of medical data through watermarking and chaotic encryption for tele-health applications[J]. Multimedia Tools and Applications, 2018.

[58] Kai F, Shangyang W, Yanhui R, et al. MedBlock: Efficient and Secure Medical Data Sharing Via Blockchain[J]. Journal of Medical Systems, 2018.

[59] Gong T, Huang H, Li P, et al. A Medical Healthcare System for Privacy Protection Based on IoT[C]. Seventh International Symposium on Parallel Architectures. IEEE, 2016.

[60] J-X Hu, C-L Chen, C-L Fan, K-H Wang. An intelligent and secure health monitoring scheme using IoT sensor based on cloud computing[J]. Journal of Sensors, 2017.

[61] Li M, Yu, Shucheng, Zheng, Yao, et al. Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption[J]. IEEE Transactions on Parallel & Distributed

Systems, 2013, 24(1): 131-143.

[62] Miao Y, Ma J, Liu X, et al. m2-ABKS: Attribute-Based Multi-Keyword Search over Encrypted Personal Health Records in Multi-Owner Setting[J]. Journal of Medical Systems, 2016, 40(11).

[63] Zhang Z, Odaibo D, Skidmore F M, et al. A Big Data Analytics Approach in Medical Imaging Segmentation Using Deep Convolutional Neural Networks[M]. Big Data and Visual Analytics. 2017.

[64] MICCAI: BraTS 2017 dataset[OL]. http://braintumorsegmentation.org/. 2017.

[65] Godinho T M, Costa C, José Luís Oliveira. Intelligent Generator of Big Data Medical Imaging Repositories [J]. IET Software, 2017, 11(3).

[66] Huang H K. PACS and Imaging Informatics: Basic Principles and Applications[M]. Basic Principles and Applications. Springer Berlin Heidelberg, 1984.

[67] Peck D. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide[M]. Digital Imaging and Communications in Medicine: A Practical Introduction and Survival Guide. 2012.

[68] Costa C, Carlos Ferreira, et al. Dicoogle-an Open Source Peer-to-Peer PACS[J]. Journal of Digital Imaging, 2011, 24(5): 848-856.

[69] Caruana C J, Christofides, S, Hartmann, G H. European Federation of Organisations for Medical Physics (EFOMP) Policy Statement 12.1: Recommendations on Medical Physics Education and Training in Europe 2014[J]. Phys Med, 2014, 30(6): 598-603.

[70] Feain I J, Court L, Palta J R, et al. Innovations in radiotherapy technology[J]. Clinical Oncology, 2017, 29(2): 120-128.

[71] Kisling K, McCarroll R, Zhang L, et al. Radiation planning assistant-a streamlined, fully automated radiotherapy treatment planning system [J]. JoVE (Journal of Visualized Experiments), 2018 (134): e57411.

[72] Amuasi J H, Kyere A K, Schandorf C, et al. Medical physics practice and training in Ghana[J]. Physica medica, 2016, 32(6): 826-830.

[73] Mahdavi S R, Rasuli B, Niroomand-Rad A. Education and training of medical physics in Iran: The past, the present and the future[J]. Physica Medica, 2017, 36: 66-72.

[74] Tsapaki V, Tabakov S, Rehani M M. Medical physics workforce: A global perspective[J]. Physica Medica, 2018, 55: 33-39.

[75] Kron T, Healy B, Ng K H. Surveying trends in radiation oncology medical physics in the Asia Pacific Region[J]. Physica Medica, 2016, 32(7): 883-888.

[76] Zhu W, Xian L, Wang E, et al. Learning classification of big medical imaging data based on partial differential equation[J]. Journal of Ambient Intelligence and Humanized Computing, 2019.

[77] Yoshida H, Shimazu T, Kiyuna T, et al. Automated histological classification of whole-slide images of gastric biopsy specimens[J]. Gastric Cancer, 2018, 21(2): 249-257.

[78] Kang X, Xiang X, Li S, et al. PCA-based edge-preserving features for hyperspectral image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(12): 7140-7151.

[79] Rokni K, Ahmad A, Solaimani K, et al. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques[J]. International Journal of Applied Earth Observation and Geoinformation, 2015, 34: 226-234.

[80] Guo K, Liu D, Li T, et al. MADP: An Open and Scalable Medical Auxiliary Diagnosis Platform[J].

Computing in Science & Engineering, 2018: 1-1.

[81] Dehua Chen, Jinxuan Niu, Qiao Pan. Auxiliary treatment of thyroid disease tensor combined with active learning method for multiple tasks[C]. 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB) 15-17 Nov. 2018.

[82] Kanghuai Liu, Zhigang Chen, Jia Wu. Big Medical Data Decision-Making Intelligent System Exploiting Fuzzy Inference Logic for Prostate Cancer in Developing Countries[J]. Publisher: IEEE Volume: 7 Page (s): 2348-2363.

[83] Reddy AS, Zhang S. Polypharmacology: drug discovery for the future[R]. Expert Rev.

[84] Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors[C]. J Med Chem, 1996, 39: 3049-59.

[85] Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity [C]. J Med Chem, 2002, 45: 4350-8.

[86] Maggiora GM. On outliers and activity cliffs-why QSAR often disappoints[C]. J Chem Inf Model, 2006, 46: 1535.

[87] Schneider G. De novo molecular design[M]. John Wiley & Sons, 2013.

[88] Roughley SD, Jordan AM. The medicinal chemist' s toolbox: an analysis of reactions used in the pursuit of drug candidates[C]. J Med Chem, 2011, 54: 3451-79.

[89] Brown DG, BostrCom J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone[C]. J Med Chem, 2016, 59: 4443-58.

[90] Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions[C]. J Cheminform 2009: 8.

[91] Timothy S. Wells, Ronald J. Ozminkowski, Kevin Hawkins, Gandhi R. Bhattarai, Douglas G. Armstrong. Leveraging big data in population health management[J]. Big Data Analytics, 2016: 1(1).

[92] Jannet M. Carmichael, Joy Meier, Amy Robinson, Janice Taylor, Diana T. Higgins, Shardool Patel. Leveraging electronic medical record data for population health management in the Veterans Health Administration: Successes and lessons learned[J]. American Journal of Health-System Pharmacy, 2017: 74(18).

[93] Karen Zander. Population Health Management: Coming of Age[J]. Professional Case Management, 2019: 24(1).

[94] Lv Qing, Jiang Yutong, Qi Jun, Zhang Yanli, Zhang Xi, Fang Linkai, Tu Liudan, Yang Mingcan, Liao Zetao, Zhao Minjing, Guo Xinghua, Qiu Minli, Gu Jieruo, Lin Zhiming. Using Mobile Apps for Health Management: A New Health Care Mode in China[J]. JMIR mHealth and uHealth, 2019: 7(6).

[95] Pauling L, Itano HA, et al. Sickle cell anemia a molecular disease[J]. Science, 1949, 110(2865): 543-548.

[96] Wong AI, Lo YM. Noninvasive fetal genomic, methylomic, and transcriptomic analyses using maternal plasma and clinical implications[J]. Trends Mol Med, 2015, 21(2): 98-108.

[97] Guo F, Yan L, Guo H, et al. The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells[J]. Cell, 2015, 161(6): 1437-1452.

[98] Hou Y, Fan W, Yan L, et al. Genome analyses of single human oocytes[J]. Cell, 2013, 155(7): 1492-1506.

[99] Yan L, Yang M, Guo H, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and

embryonic stem cells[J]. Nat StructMolBiol, 2013, 20(9): 1131-1139.

[100] Lu S, Zong C, Fan W, et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing[J]. Science, 2012, 338(6114): 1627-1630.

[101] Dawson SJ, Tsui DW, Murtaza M, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer[J]. N Engl J Med, 2013, 368(13): 1199-1209.

[102] Schwarzenbach H, Nishida N, Calin GA, et al. Clinical relevance of circulating cell-free microRNAs in cancer[J]. Nat Rev Clin Oncol, 2014, 11(3): 145-156.

[103] Xia JH, Liu CY, Tang BS, et al. Mutations in the gene encoding gap junction protein beta-3 associated with autosomal dominant hearing impairment[J]. Nat Genet, 1998, 20(4): 370-373.

[104] Yang XP, She CW, Guo LZ, et al. A locus for brachydactyly type A-1 maps to chromosome 2q35-q36 [J]. Am J Hum Genet,2000, 66(3): 892-903.

[105] Gao B, Gu JZ, She CW, et al. Mutations in IHH, encoding Indian hedgehog, cause brachydactyly type A-1[J]. Nat Genet, 2001, 28(4): 386-388.

[106] Lui W, Wang H, Zhao S, et al. The novel gene locus for agenesis of permanent teeth (He-Zhao deficiency) maps to chromosome 10q11.2[J]. J Dent Res, 2001, 80(8): 1716-1720.

[107] Zhang XJ, Huang W, Yang S, et al. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21[J]. Nat Genet, 2009, 41(2): 205-210.

[108] Han, J W, Zheng, H F, Cui, Y, Sun, L D, Ye, D Q, Hu, Z, et al. Genome-wide association study in a chinese han population identifies nine new susceptibility loci for systemic lupus erythematosus[J]. NATURE GENETICS, 2009, 41(11): 1234-1237.

[109] Zhang FR, Huang W, Chen SM, et al. Genome wide association study of leprosy[J]. N Engl J Med, 2009, 361(27): 2609-2618.

[110] Quan C, Ren YQ, Xiang LH, et al. Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC[J]. Nat Genet, 2010, 42(7): 614-618.

[111] Bei JX, Li Y, Jia WH, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci[J]. Nat Genet, 2010, 42(7): 599-603.

[112] Zhang HX, Zhai Y, Hu Z, et al. Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers[J]. Nat Genet, 2010, 42(9): 755-758.

[113] Wen Y, Liu Y, Xu Y, et al. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause[J]. Nat Genet. 2009.

[114] Wang B, Yang W, Wen W, et al. Gamma-secretase gene mutations in familial acne inversa[J]. Science, 2010, 330(6007): 1065.

[115] Song YM, Li L, Ou Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer [J]. Nature,2014, 509(7498): 91-95.

[116] Su JG, Jiao X, Sun TG, et al. Analysis of domain movements in glutamine-binding protein with simple models[J]. Biophys J,2007, 92(4): 1326-1335.

[117] Wang LD, Zhou FY, Li XM, et al. Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54[J]. Nat Genet, 2010, 42(9): 759-763.

[118] Wu C, Wang Z, Song X, et al. Joint analysis of three genome-wide association studies of esophageal

squamous cell carcinoma in Chinese populations[J]. Nat Genet, 2014,46(9): 1001-1006.

[119] Dvorkin-Gheva, A, Hassell JA. Identification of a novel luminal molecular subtype of breast cancer[J]. PLoS One,2014, 9(7): e103514.

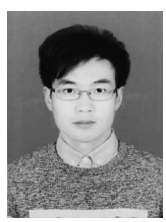
[120] Collisson EA, Sadanandam A, Olson P, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy[J]. Nat Med, 2011, 17(4): 500-503.

[121] Phipps AI, Limburg PJ, Baron JA, et al. Association between molecular subtypes of colorectal cancer and patient survival[J]. Gastroenterology, 2015, 148(1): 77-87.

[122] Higdon R, Earl RK, Stanberry L, et al. The promise of multiomics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders[J]. OMICS, 2015, 19(4): 197-208.

作者简介

彭绍亮 湖南大学计算机科学系，国家超级计算长沙中心副主任，教授，博导，研究方向：高性能计算、大数据、生物信息、人工智能、区块链等技术。CCF 杰出会员和杰出讲者，计算机应用专委和生物信息专委副主任、YOCSEF 长沙主席（2016-2017）和总部 AC 委员（2017-）、CCF 高性能计算、大数据、区块链专委委员、CCF2016 大数据技术大会程序委员会主席、CCF 大数据学术大会程序委员会副主席。



杨亚宁 湖南大学信息科学与工程学院，博士，CCF 会员。研究方向：并行计算、生物信息、机器学习等。

张彦春 复旦大学计算机科学技术学院，教授、博导；研究方向：大数据分析技术、数据库、数据挖掘、数据分析、深度学习及网络信息系统、医学健康大数据分析、健康信息学、环保数据分析、区块链技术、中医药知识发现、儿童健康数据管理等。

胡 斌 兰州大学信息科学与工程学院院长，教授，国家特聘专家，研究方向：心理生理计算、普适计算、协同工作技术和语义网等。

阮　彤　华东理工大学计算机技术研究所，教授，研究方向：自然语言处理、信息抽取、信息推荐、Web 内容挖掘、大数据挖掘。

邢春晓　清华大学信息技术研究院副院长、清华大学互联网产业研究院副院长、清华大学信息技术研究院 WEB 与软件技术研究中心主任、清华大学智慧城市大数据研究中心主任、清华大学电子政务研究中心主任。研究方向：数据库和数据仓库、软件工程、数据和知识工程、人工智能和机器人、大数据科学与技术、电子商务和电子政务关键技术等。

关键词索引

互联网 1-9, 13-23, 25-27, 29-30, 32-36, 44, 46, 55, 60, 67, 70-71, 73, 86, 96, 140, 155, 164, 166, 170, 179, 200, 204, 211, 217, 243, 301, 322, 323, 326, 330, 339, 341, 349
光通信 1, 3-5, 15-16, 21, 23
移动无线网 1, 5-6, 16, 21
IPv4 1-3, 7-8, 17-18, 24-27
IPv6 1-3, 7-8, 17-18, 20-21, 24-27
互联网路由 1, 9, 29
互联网传输 1, 30
网络管理 1, 2, 7, 9, 21, 25, 27
互联网应用 1, 3, 8, 13, 19, 25, 46, 55, 71
互联网安全 1
开源芯片 31-33, 35-36, 45, 49, 51, 53-54, 67, 69-71, 74
开放指令生态 31, 44, 67, 73-74
开放指令集 31-33, 35-38, 41, 63, 69-71
芯片敏捷开发 31, 44, 67, 73-74
大数据 337-341, 348-349
近似算法 103-105, 114, 121-122, 124, 126, 128
近似查询处理 103-106, 111-112, 121-126
系统级近似计算 103-104, 114, 120, 123, 126-127
大数据近似计算 103-104, 121, 124, 126, 128
区块链 25, 139-179, 324, 332, 348
智能合约 139-140, 143-147, 149, 151-152, 154-159, 162-163, 166-167, 171-173, 176
数字货币 139-140, 146-147, 152-154, 164, 166
安全 1, 8, 9, 13-16, 19-21, 24-25, 27, 29, 31, 40, 42-43, 62, 75, 80, 97, 104, 115, 119, 139-148, 150-163, 165-166, 168-173, 175-176, 178-179, 183, 199, 254, 323, 326, 328-330, 332-333, 336-337, 339
监管 13, 39, 139-141, 145-147, 159, 164, 172-175, 179, 324
元学习 180-183, 185-193
深度学习 45, 75, 77, 82, 172, 180-181, 186-187, 192, 199, 203, 205-206, 208, 232-234, 237-238, 241, 262, 271, 288, 302-304, 306, 310-312, 324, 331, 333-336, 348
监督学习 180-181, 188, 192-193, 235, 242, 279-280
事件知识图谱 199-201, 210-211, 213, 217-218
事件抽取 199-206, 210, 218
事件关系抽取 199-201, 204, 206, 210, 218-219
知识图谱构建 199-201, 210-213, 217-218, 222
知识推理 199, 210, 213-214, 216-217
结构化稀疏表示模型 226, 229-230, 239, 242
基于框架理论的深度网络模型 226, 229-230, 232, 240, 242
多层卷积稀疏编码模型 226, 229-230
图信号处理理论 226, 229-230, 236-238, 241, 243-244
多媒体信号处理 226-227, 229, 242
真实感绘制 255-256, 259, 262, 280, 284, 289-290
基于物理的绘制 255
基于物理的材质模型 255, 257-258, 263, 290
全局光照 255, 257, 259, 262, 268, 274, 280, 284-285, 287-290, 298
体绘制 255, 257, 260, 273-274, 282, 287, 290
蒙特卡罗降噪 255, 257, 262, 277, 279, 288, 290
并行渲染 255, 257
增强现实 5, 243, 300-302, 305-312, 321
可视计算 300-301, 309-310, 312
场景建模 300, 302-303, 310, 312
内容生成 300, 305
感知交互 300-301, 308, 312
医疗大数据 322-327, 330-331, 333-334, 336, 338, 340-341
临床科研 322, 324-325
临床诊疗 322, 324, 325, 327, 336
数据采集 236, 312, 322, 324, 326-327, 332, 340-341
数据分析 88-89, 124-125, 137-138, 173, 210, 265, 322, 325-326, 330-332, 337-339, 348