健康医疗大数据与罕见病的精准用药

武志慧[1],王飞[1],姜召芸[1],闵浩巍[1],王心慰[1],弓孟春[2],史文钊[1]

1. 神州数码医疗科技股份有限公司,北京 100080
2. 中国医学科学院北京协和医院中心实验室,北京 100730

摘要 药物治疗是罕见病的主要治疗方式,然而目前仅有1%的罕见病能够得到有效的药物治疗。不同来源的基因组、转录组等组学数据与临床表型数据融合起来形成的"健康医疗大数据",可以通过集中小规模罕见病临床数据的方式有效弥补罕见病样本量少的不足。大数据信息可用于研究罕见病新的药物靶点、探索成熟药物在罕见病领域新用法、分析药物不良反应实现个体化用药,并可进一步通过大数据技术建立本地化罕见病知识库,从而实现罕见病的精准诊断和治疗。随着大数据研究的不断深入,仍然需要突破多组学融合及分析技术、基于真实世界的知识提取技术、基于组学的临床决策支持等技术壁垒才能使大数据在罕见病的诊疗中得到最大应用。

关键词 健康医疗大数据;罕见病;精准用药

罕见病,又称为"孤儿病",是指发病率低、患病人数相对较少的疾病。目前尚无对罕见病的确切定义,不同国家或组织对其定义不同。世界卫生组织(WHO)将罕见病定义为患病人数占总人口0.065%~0.1%的疾病或者病变。欧盟健康与消费者保护总司(DG SANCO)认为患病人数占总人口0.05%的疾病或者病变即为罕见病。2002年,美国《罕见病法案》(《Rare Disease Act of 2002》)将罕见病定义为每年患病人数少于20万的疾病。尽管单个病种罕见病发病率低,但由于全世界罕见病大约有7000多种,因此全世界罕见病患者人数可达3.5亿~4亿。所以说,罕见病是一类严重危害公众健康的疾病。

健康医疗大数据包含基因组、暴露组、电子病历等临床数据以及个人生活习惯和地理位置等信息,可帮助临床医生理解疾病的发生发展过程,使得数据驱动的用药决策成为可能。对于罕见病而言,健康医疗大数据能够有效地弥补罕见病样本量较少的缺点,有效地整合各种信息,从中挖掘出各种相关关系,进而对罕见病患者个体进行用药指导。

1 健康医疗大数据指导罕见病用药

药物治疗是罕见病治疗的主要方式,据统计,目前仅1%的罕见病能够得到有效的药物治疗[1]。用于罕见病的药物(又称"孤儿药")在开发和上市的过程中面临着受众小、科研投入可能无法收回、基础研究落后等重重困难,这些使得孤儿药研发进度缓慢。不同来源的基因组、转录组等组学数据与临床表型数据融合起来形成的"健康医疗大数据"使得获取信息更加容易,对于罕见病而言,更利于将小规模的临床试验等数据集中起来进行分析,进而指导用药。

1.1 利用健康医疗大数据寻找潜在药物靶点

化学信息学工具能够整合真实世界的多维度数据,通过化学结构相似性检索[2]、数据挖掘、机器学习[3]、生物活性谱[4]等算法模拟药物设计过程,寻找潜在的药物靶点。此外,也可联合药物-蛋白互作网络和蛋白-疾病互作网络,通过研究基因组、转录组、蛋白组、代谢组、微生物组和药物基因组之间的关系来寻找药物靶点。Zhao等[5]基于药物和基因组开发的drugCIPHER就是一个预测药物靶点的计算工具,利用该模型在全基因组范围内构建了726种药物的生物指纹图谱,其中,在501例中发现新的药物互作关系。通过多组学分析发现的与嗜铬细胞瘤和副神经节瘤的发生发展密切相关的13个位点可作为药物研发的靶点[6-8]。2003年,Abifadel等[9]对一法国高胆固醇血症家族进行测序,发现该家族的高胆固醇血症患者发生了*PCSK9*基因突变,进而导致该家族胆固醇含量是正常人胆固醇含量的5~10倍。2003年,该基因首次被报道与高胆固醇血症相关,成为继*LDLR*和*APOB*之外第3个与常染色体显性家族性高胆固醇血症有关的基因。这是因

收稿日期:2017-06-20;修回日期:2017-08-02
基金项目:国家高技术研究发展计划(863计划)项目(2015AA020106);国家重点研发计划项目(2016YFC0901500);上海市出生缺陷防治重点实验室开放课题(16DZKF1007);国家卫生计生委2016年信息化与统计项目
作者简介:武志慧,硕士,研究方向为大数据,电子信箱:wuzhj@dchealth.com;王飞(共同第一作者),博士,研究方向为大数据,电子信箱:wangfeiab@dchealth.com;史文钊(通信作者),博士,研究方向为健康医疗大数据,电子信箱:shiwza@dchealth.com
引用格式:武志慧, 王飞, 姜召芸, 等. 健康医疗大数据与罕见病的精准用药[J]. 科技导报, 2017, 35(16): 20-25; doi: 10.3981/j.issn.1000-7857.2017.16.002

为PCSK9蛋白与肝脏表面的低密度脂蛋白受体(LDLR)结合后,促进LDLR的降解,导致血液中低密度脂蛋白(LDL)水平的升高。将该基因作为药物靶点,安进研发的Evolocumab(抗PCSK9单抗)和赛诺菲研发的Alirocumab(抗PCSK9单抗)可显著抑制体内PCSK9的功能,降低"坏胆固醇",减少心血管疾病发生率和死亡率。

1.2　利用大数据分析药物不良反应实现精准化治疗

罕见病的治疗手段有限,并且部分药物的常规方案不良反应显著而无法用于临床,导致可供罕见病患者选择的治疗方案少之又少。这种对药物不良反应"一刀切"的方式无疑会给需要这种治疗方式的罕见病患者带来不便。融合了基因组、环境、生活习惯等多个维度信息的健康医疗大数据能够系统地获得大量的个案报告,通过对大量数据的分析可以得到特定罕见病治疗药物的有效及安全数据,从而提供适合罕见病患者的精准治疗方式。治疗癫痫的丙戊酸钠药物的使用就是这样一个例子。丙戊酸钠是一种有效的抗癫痫药物,尤其是对青少年肌阵挛癫痫非常有效,但因其不良反应而仅作为治疗癫痫的三线或者四线药物。2015年一项前瞻性队列研究[10]发现当患有癫痫的孕妇服用高剂量丙戊酸钠(每日剂量>800 mg)时,所产婴儿的认知能力较服用拉莫三嗪和卡马西平的癫痫孕妇所产婴儿弱,但是当患有癫痫的孕妇服用低剂量丙戊酸钠(每日剂量<800 mg)时,对婴儿认知能力的影响则不显著。将这个研究结果与先前的认知结合起来,对多个研究结果组成的大数据进行分析,提示对青少年肌阵挛癫痫患者可使用低剂量的丙戊乙酸(<800 mg)进行治疗。这一发现大大改善了青少年肌阵挛癫痫患者的预后和生活质量。利用健康医疗大数据,可实现对治疗方案进行精准亚分类,更加精准地进行治疗。

1.3　利用大数据发现老药在罕见病领域新用法

老药新用是一个为现有药物发现新适应症的过程,也是药物研发的一个重要方面[11]。对于罕见病而言,老药新用能够有效解决罕见病无药可医的问题,为罕见病患者带来福音。甲磺酸伊马替尼原是针对Bcr/Abl位点研发的治疗费城染色体阳性慢性髓性白血病(Ph+CML)的一种孤儿药,后来发现其对C-Kit位点也有作用,还可以用来治疗胃肠道间质瘤(GIST)、成人复发的或难治的费城染色体阳性的急性淋巴细胞白血病(Ph+ALL)、嗜酸细胞过多综合症(HES)/慢性嗜酸粒细胞白血病(CEL)伴有FIPlLl-PDGFRα融合激酶等多种罕见病适应症[12]。

精神分裂症是一种致病机理复杂的罕见病,目前尚无有效的治疗方法,再加上制药公司对治疗精神分裂症的药物研发兴致不高,因此迫切需要新的方法来开发精神分裂症治疗药物。为此,Xu等[13]开发了一个新系统——PhenoPredict,该系统可通过知识库推断表型相似疾病的治疗药物对精神分裂症的治疗效果 。这一药物预测新方法的关键是要构建一个广泛全面、包含疾病与治疗药物的知识库。该模型较之于药物适应症预测模型(PREdicting Drug IndiCaTions,PREDICT ,当前常用的老药新用系统之一),查全率(PR)曲线下面积(AUC)显著提高98.8%。全表型组关联分析(PheWAS)是与全基因组关联分析(GWAS)类似的一项分析方法,二者区别在于GWAS探究变异与疾病的关系,而PheWAS探究疾病与变异之间的关系。PheWAS通过病人队列的电子表型数据能够将遗传变异与众多疾病关联起来,从而对疾病的病因进行深入解释。2013年,Denny等[14]对3144个已进行GWAS分析的单核苷酸多态性(SNP)位点进行PheWAS分析,发现PheWAS不仅能够鉴别出之前GWAS分析出的SNP位点,还可以发现新的关联。利用PheWAS技术分析药物和疾病的关系时,需要使用MetaMap工具对文本中的字符串提取疾病名称[15]。Rastegarmojarad等[16]利用PheWAS鉴定出大约14800对药物-疾病对,3800多种具老药新用潜力的药物。由此可见,健康医疗大数据在挖掘治疗罕见病的老药方面具有巨大的潜力。

1.4　使用大数据手段建立本地化罕见病知识库

罕见病诊断非常困难,这是因为临床表型和基因型之间存在一定差异:临床表型相同,基因型不一定相同;同样的基因型,不同器官、部位,表型特征不尽相同。全部的罕见病的遗传表型若要单纯靠医生进行记忆,显然不切实际。此外,环境、人种等因素的差异导致各个地区的罕见病类型不同,致病基因不同。现有的知识库多来源于外文文献[17]~[22],很难将国外研究的结果直接临床应用到中国患者身上。因此建立本地化的知识库,指导罕见病的诊疗非常有必要。国内罕见病的数据多按照数据类型和疾病类型进行分类存放,个人在融合多数据方面的力量非常有限,因此,若要建立一个服务于罕见病的知识库,就必须将临床表型、基因组、生物样本以及各种研究、实验数据整合起来,利用大数据的手段建立知识库,将个体和队列的水平都相互关联,方便研究人员全方位了解疾病,或者帮助患者获得感兴趣的团体的数据。2016年,国内已开始开展罕见病全国疾病谱调查,建立罕见病登记系统,并在此基础上建立和健全罕见病多中心临床资料库和生物样本库2个国家数据库,开展超过50种5万例罕见疾病的注册登记研究,获得国际罕见病研究最大的患者人群[23];同时开展各种罕见病家系调查,如郭奕斌等[24]按照国际遗传性骨病的最新分类标准对遗传学骨病家系进行调查并分类。此外,还应加快对中国罕见病相关的文献进行整理,使其成为罕见病知识库中的一部分。

2　技术壁垒与进展

2.1　多组学融合及分析技术

罕见病低频的性质导致与罕见病相关的研究充满挑战,依靠关联分析和图位克隆鉴定致病基因需要消耗大量的人力和时间[25]。随着测序技术、生物信息学和计算科学的迅速发展,多组学数据的融合必然快速推动精准医疗发展,但是现阶段在数据的收集、传输、存储和分析几个方面还存在着技术壁垒。

2.1.1 多层次生命组学数据整合的思路

生命组学数据包括基因组、蛋白组、转录组和表型组等多维度的数据,利用系统生物学方法,对多组学数据进行系统整合和深度挖掘,为研究疾病发病机制及治疗靶点、分子生物学特征、治疗相关预测因子、疾病发病及发展预测因子等提供指导,实现疾病的精准分类和诊断,并制定个性化的疾病预防和诊疗方案,实现精准治疗、精准用药[26]。多组学数据整合分析的基本思路是:不同来源的数据进行标准化处理,建立不同组学数据之间的关联性和差异性,对候选因子进行筛选过滤,建立模型,预测和验证候选因子的作用[27]。

2.1.2 多层次生命组学数据整合的研究现状

数据整合的前提是收集和共享数据,目前很多国际组织致力于罕见病各组学数据的收集,比如国际罕见疾病研究联盟(IRDiRC)拥有40个成员,收集其资助项目的原始/源数据,并在保证安全的基础上提供互操作性,预计在2020年开发诊断罕见病的手段,并提供200种新的罕见病治疗手段[25]。国家罕见病注册登记平台选择20家全国领先的罕见病研究单位,开展超过50种5万例罕见疾病的注册登记研究。通过这一基础工程,获得国际罕见病研究最大的患者人群数据,加快中国罕见病资源的收集和整合,加速罕见病研究和创新。

通过整合多组学数据预测疾病的靶点。例如Zhang等[28]对基因表达、DNA甲基化、miRNA表达以及拷贝数的变化进行了整合分析,通过Cox比例风险回归模型、贝叶斯信息准则以及无监督的超级K聚类方法找到了卵巢癌的7个亚型,其中基因组和转录组的数据在卵巢癌分型中的作用更大。在罕见病的研究中,可以借鉴癌症靶点预测的方法和思路,通过整合多组学数据预测罕见病的靶点。

2.1.3 多层次生命组学数据整合的难点

罕见病的患病率低,数据量相对其他疾病不够丰富,其多组学数据的收集和共享对罕见病研究来说尤其重要,需要多机构的共同努力。另一方面,生命大数据的储存和维护通常超出了单个研究小组的能力:管理患者的高通量组学数据需要在计算基础设施方面大量投资,开发和维护生物信息学数据分析流程需要花费的时间和精力也会越来越多[25]。

生命组学数据库中集成了不同来源、不同维度、不同质量的临床数据和组学数据,多组学数据的融合不仅是医学研究的难题,也为数学和计算机科学带来挑战。现阶段多组学数据的整合分析研究还不成熟,亟需开发通用的数据分析和整合方法,这已经成为生物信息学研究领域的一个瓶颈。同时,生物各组学数据量大,呈指数级增长,数据分析的运算量巨大,对计算机的性能要求非常高[29]。大量数据的解读能力成为一个重要的限制因素。

2.2 基于真实世界的知识提取技术

按照美国食品药品监督管理局(FDA)的定义,真实世界数据(real world data)的来源包括大规模简单临床试验、实际医疗中的临床试验、前瞻型观察性研究或注册型研究、回顾性数据库分析、病例报告、健康管理报告、电子健康档案等。这些数据包含了大量的患者相关的医学信息,然而这些文本形式存在的文件,缺乏对医学概念的标准化描述[30],需要对生物医学文本进行挖掘和利用。根据中国国情,当前最有价值的数据还是电子病历,电子病历对患者疾病发生、发展和转归,以及检查、诊断和治疗等活动过程进行了详细记录,这对于指导临床、医药研发和医学研究有很重要的意义。

2.2.1 生物医学文本挖掘技术

生物医学文本挖掘涵盖了自然语言处理、生物信息学 、机器学习等多个领域,研究重点主要是信息抽取和数据挖掘,包括生物医学命名实体识别、缩写词和同义词的识别、命名实体关系抽取、利用推理生成抽取的关系假设、文本分类以及上述工作的集成框架等[31]。

2.2.2 生物医学文本挖掘技术的进展

当前生物医学文本挖掘的主要研究热点集中在文本挖掘的基本技术研究、文本挖掘在生物信息学领域里的应用、文本挖掘在药物相关事实抽取中的应用3个方面[32]。在生物医学文本挖掘方面,中国学者做了大量研究,也取得了极大的进展。比如Huang等[33]利用动态规划算法的模式匹配方法,抽取蛋白质交互作用关系,取得了80%的召回率和精确率;进一步采用最小描述长度原理进行模式优化,提高了抽取精度[34]。通过结合模式匹配与浅层句法分析,再次将原模式匹配方法的精确率和F测度提高了7%[35]。龚凡等[36]基于症状构成模式的非监督学习方法,实现了中文电子病历文本中症状实体的自动识别。李昀泽等[30]利用潜在语义分析方法,构建了一套医学病历数据服务系统,提供病历检索、病历总结、病历语料库的语义总结服务。

近年来,生物医学领域命名实体识别的研究也在不断扩展和深入,形成了医学术语体系,例如观测指标标识符逻辑命名与编码系统(logical observation identifiers names and codes,loinc)、医学系统命名法——临床术语(systematized nomenclature of medicine-clinical terms,SNOMED-CT)、中文人类表型标准用语联盟(the human phenotype ontology of China,CHPO)等。目前主要是基于机器学习的方法识别生物医学命名实体,如贝叶斯模型、隐马尔可夫模型(HMM)、支持向量机(SVM)、条件随机场(CRFs)、最大熵(ME)等[37]。

2.2.3 生物医学文本挖掘技术的难点

电子病历的语法结构松散,通常由很多短句拼凑在一起,简略语多,医学概念的表述更是因人而异。如何通过生物医学文本挖掘,提取其中有用的诊疗信息,形成知识本体或者知识网络,为后续的文本挖掘任务提供标准和便利[38],并应用到临床实践中,是精准医学信息学需要解决的重要问题之一。

生物医学文本挖掘的技术难点主要是由生物医学文本本身特点导致的。以生物医学命名实体识别为例,新的命名实体不断出现,所以简单的文本匹配算法不能直接使用;另外识别命名实体的边界非常困难,例如很多命名实体是多词短语,有些命名实体名称很长,以及同义词和缩写词。尽管

目前多采用机器学习的方法识别命名实体,但是仍然存在识别不准确等缺点。希望随着医学术语体系的发展和应用,能够解决这一难题。

2.3　基于组学的临床决策支持

基于组学的临床决策支持是一种复杂的健康信息技术,它能够翻译和整合基因组知识与电子病历及其他临床系统的病人信息,自动处理患者数据并给出智能的医疗和护理建议,为临床医生的决策提供干预措施、诊断和治疗建议。

2.3.1　组学数据从融合到临床应用

利用临床决策支持系统(CDS)将组学数据融合到临床应用涉及多个方面:包括各层次生命组学和医疗数据的收集、基因型-表型-药物等多重数据对应关系的构建、基因变异和表型相关数据库和知识库的建设,开发应用于真实场景的临床决策工具等[39]。

将临床决策支持系统部署在电子病历层面,将来自不同数据源的患者遗传变异数据与知识库整合后,通过电子病历系统进行结构化的显示,并反馈给临床医生,从而支持临床决策。将临床决策支持系统与各生命组学数据库相连接可以提供基于基因变异的风险预测、预后评估、在特定临床环节的药物剂量制定等建议[40]。

2.3.2　基于组学的临床决策支持系统应用现状

临床决策支持系统与患者个体数据相结合,可以在诊疗过程的各个环节提供可干预的信息[41]。目前,国际上已经开发了一些临床决策支持工具,例如分子图谱与可行治疗方法整合系统(the integrating molecular profiles with actionable therapeutics,IMPACT),在临床上利用全外显子测序数据,预测可干预药物[42];IBM公司的Watson通过获取的基因序列数据及与药物匹配的医学文献信息,包括病人独特的基因突变,确定最有可能的驱动突变及作用的药物靶标[43]。

2.3.3　基于组学的临床决策支持系统开发难点

临床决策支持系统基于病人和疾病的特点,包括基因组突变检测和所有患者特异性数据的临床应用,然而临床医生缺乏利用基因组信息的知识,再加上每个人的基因组都包含大量的变异信息,只依靠医生的经验和知识,不能将基因组数据和表型数据整合。

由于医生在临床诊断过程只需要组学数据的解读报告,并不需要了解组学相关的知识和生物信息分析流程,因此组学数据的分析和变异的解读报告需要整合到临床决策支持系统中。但是现在生物信息的分析软件种类非常多,没有统一的行业标准;另外,变异位点的知识和解读依赖于庞大的基因型-表型关联数据库和知识库,在此基础上结合机器学习等人工智能手段,才能够嵌入临床决策支持系统。这对人才和技术的要求较高,需要医学、生物信息学、机器学习等多种技术人员的配合和复合型人才的共同努力。

3　结论

大数据及大数据技术的出现,使得各行各业面临着新的变革,健康医疗大数据的发展使重大疾病及罕见病的诊疗逐渐颠覆传统的方式。通过积累的药物信息、治疗方案、病例信息等数据,逐步实现一个完全个性化的诊断结果以及理想的治疗方案。罕见病的研究也将推动精准医学的发展,疾病谱的细分也对新治疗方法的探索有帮助。目前罕见病的主要问题还是能否进行明确诊断,大部分疾病尚缺乏针对性治疗,部分疾病的治疗也主要是长期服药、特殊饮食及生活干预,因此门诊依然是罕见病的主要医疗方式。而健康医疗大数据的发展对罕见病的用药指导也有革命性的推动作用。多组学融合及分析、基于真实世界的知识提取等技术的突破和进步,必将推进健康医疗大数据在实际的应用中。在健康医疗大数据的指导下,罕见病的诊断、治疗和用药等都将获得更精准的方案。

参考文献(References)

[1] 谷景亮, 鲁艳芹, 钟彩霞, 等. 国外罕见病药物政策发展现状对比分析[J]. 卫生软科学, 2013, 27(7): 393-396.
Gu Jingliang, Lu Yanqin, Zhong Caixia, et al. Comparative analysis to rare disease pharmaceutical policy development status in foreign countries[J]. Soft Science of Health, 2013, 27(7): 393-396.

[2] Keiser M J, Setola V, Irwin J J, et al. Predicting new molecular targets for known drugs[J]. Nature, 2009, 462(7270): 175-181.

[3] Nidhi N, Glick M, Davies J W, et al. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases[J]. Journal of Chemical Information & Modeling, 2006, 46(3): 1124-1133.

[4] Cheng T, Li Q, Wang Y, et al. Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining[J]. Journal of Chemical Information and Modeling, 2011, 51(9): 2440-2448.

[5] Zhao S, Li S. Network-based relating pharmacological and genomic spaces for drug target identification[J]. PloS One, 2010, 5(7): e11764.

[6] Nölting S, Grossman A B. Signaling pathways in pheochromocytomas and paragangliomas: Prospects for future therapies[J]. Endocrine Pathology, 2012, 23(1): 21-33.

[7] Björklund P, Pacak K, Crona J. Precision medicine in pheochromocytoma and paraganglioma: Current and future concepts[J]. Journal of Internal Medicine, 2016, 280(6): 559-573.

[8] Castro-Vega L J, Letouzé E, Burnichon N, et al. Multi-omics analysis defines core genomic alterations in pheochromocytomas and paragangliomas[J]. Nature Communications, 2015, 6: 6044.

[9] Abifadel M, Varret M, Rabès J P, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia[J]. Nature Genetics, 2003, 34(2): 154-156.

[10] Baker G A, Bromley R L, Briggs M, et al. IQ at 6 years after in utero exposure to antiepileptic drugs A controlled cohort study[J]. Neurology, 2015, 84(4): 382-390.

[11] Novac N. Challenges and opportunities of drug repositioning[J]. Trends in Pharmacological Sciences, 2013, 34(5): 267-272.

[12] 田苗, 田红, 解学星, 等. 罕见病用药现状分析[J]. 现代药物与临床, 2014, 29(7): 701-707.
Tian Miao, Tian Hong, Xie Xuexing, et al. The development of orphan drugs[J]. Drugs & Clinic, 2014, 29(7): 701-707.

[13] Xu R, Wang Q Q. PhenoPredict: A disease phenome-wide drug reposi-

tioning approach towards schizophrenia drug discovery[J]. Journal of Biomedical Informatics, 2015, 56: 348-355.

[14] Denny J C, Bastarache L, Ritchie M D, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data[J]. Nature Biotechnology, 2013, 31(12): 1102-1111.

[15] Aronson A R, Lang F M. An overview of MetaMap: historical perspective and recent advances[J]. Journal of the American Medical Informatics Association, 2010, 17(3): 229-236.

[16] Rastegarmojarad M, Ye Z, Kolesar J M, et al. Opportunities for drug repositioning from phenome-wide association studies[J]. Nature Biotechnology, 2015, 33(4): 342-345.

[17] MacArthur D G, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes[J]. Science, 2012, 335(6070): 823-828.

[18] 1000 Genomes Project Consortium. A global reference for human genetic variation[J]. Nature, 2015, 526(7571): 68-74.

[19] Maaroufi M, Choquet R, Landais P, et al. Towards data integration automation for the French rare disease registry[C]//AMIA Annual Symposium Proceedings. Bethesda: American Medical Informatics Association, 2015, 2015: 880.

[20] Roy A J, Van den Bergh P, Van Damme P, et al. Early stages of building a rare disease registry, methods and 2010 data from the Belgian Neuromuscular Disease Registry (BNMDR)[J]. Acta Neurologica Belgica, 2015, 115(2): 97-104.

[21] Tattersfield A E, Glassberg M K. Lymphangioleiomyomatosis: A national registry for a rare disease.[J]. American Journal of Respiratory & Critical Care Medicine, 2006, 173(1): 2-4.

[22] Nagel G, Ünal H, Rosenbohm A, et al. Implementation of a population-based epidemiological rare disease registry: Study protocol of the amyotrophic lateral sclerosis (ALS)-registry Swabia[J]. BMC Neurology, 2013, 13(1): 22.

[23] 冯时, 弓孟春, 张抒扬. 中国国家罕见病注册系统及其队列研究:愿景与实施路线[J]. 中华内分泌代谢杂志, 2016, 32(12): 977-982.
Feng Shi, Gong Mengchun, Zhang Shuyang. The national rare diseases registry system of China and the related cohort studies: Vision and roadmap[J]. Chinese Journal of Endocrinology and Metabolism, 2016, 32(12): 977-982.

[24] 郭奕斌, 李荣. 罕见遗传性骨病大家系调查[J]. 中华骨科杂志, 2014 (8): 880-882.
Guo Yibin, Li Rong. Rare hereditary bone disease family survey[J]. Chinese Journal of Orthopaedics, 2014(8): 880-882.

[25] Johnston L, Thompson R, Turner C, et al. The impact of integrated omics technologies for patients with rare diseases[J]. Expert Opinion on Orphan Drugs, 2014, 2(11): 1211-1219.

[26] 谢兵兵, 杨亚东, 丁楠,等. 整合分析多组学数据筛选疾病靶点的精准医学策略[J]. 遗传, 2015, 37(7): 655-663.
Xie Bingbing, Yang Yadong, Ding Nan, et al. Identification of disease targets for precision medicine by integrative analysis of multi-omics data[J]. Hereditas, 2015, 37(7): 655-663.

[27] Yoon S H, Han M J, Jeong H, et al. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12[J]. Genome biology, 2012, 13(5): R37.

[28] Zhang W, Liu Y, Sun N, et al. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer[J]. Cell Reports, 2013, 4(3): 542-553.

[29] 王锋. 基于稀疏偏最小二乘算法的生物组学数据融合算法研究[D]. 长春: 吉林大学, 2012.
Wang Feng. Research into biological omics datasets integration based on sparse partial least-square algorithm[D]. Changchun: Jinlin University, 2012.

[30] 李昀泽. 基于潜在语义分析的病历文本挖掘应用研究[D]. 杭州: 浙江大学, 2015.
Li Yunze. Research and apply on patient record text mining based on latent semantic analysis[D]. Hangzhou: Zhejiang University, 2015.

[31] Ananiadou S, Kell D B, Tsujii J. Text mining and its potential applications in systems biology[J]. Trends in Biotechnology, 2006, 24(12): 571-579.

[32] 史航, 高雯珺, 崔雷. 生物医学文本挖掘研究热点分析[J]. 中华医学图书情报杂志, 2016, 25(2): 27-33.
Shi Hang, Gao Wenjun, Cui Lei. Hotspots in text mining of biomedial field[J]. Chinese Journal of Medical Library and Information Science, 2016, 25(2): 27-33.

[33] Huang M, Zhu X, Hao Y, et al. Discovering patterns to extract protein-protein interactions from full texts[J]. Bioinformatics, 2004, 20 (18): 3604-3612.

[34] Hao Y, Zhu X, Huang M, et al. Discovering patterns to extract protein-protein interactions from the literature: Part II[J]. Bioinformatics, 2005, 21(15): 3294-3300.

[35] Huang M, Zhu X, Li M. A hybrid method for relation extraction from biomedical literature[J]. International Journal of Medical Informatics, 2006, 75(6): 443-455.

[36] 龚凡, 王梦婕, 阮彤, 等. 电子病历文本症状自动识别方法[J]. 医学信息学杂志, 2016, 37(7): 7-14.
Gong Fan, Wang Mengjie, Ruan Tong, et al. Automatic recognition methods of symptoms in texts of electronic medical records[J]. Journal of Medical Intelligence, 2016, 37(7): 7-14.

[37] Polajnar T. Survey of text mining of biomedical corpora[J]. [2009-08-20]. http://www. brc. des. gla. ac. uk/tamara/surveyoftm. pdf, 2006.

[38] 柴华, 路海明, 刘清晨. 中医自然语言处理研究方法综述[J]. 医学信息学杂志, 2015, 36(10): 58-63.
Chai Hua, Lu Haiming, Liu Qingchen. Overview of Research Methods for Natural Language Processing in Traditional Chinese Medicine[J]. Journal of Medical Informatics, 2015, 36(10): 58-63.

[39] Ohno-Machado L. Data and the clinical decision support loop[J]. Journal of the American Medical Informatics Association, 2016, 23(e1): e1.

[40] Lang R D. In search of the missing link: Data access and the next generation of CDSS[J]. Journal of Healthcare Information Management, 2002, 16(4): 2.

[41] Bietz M J, Bloss C S, Calvert S, et al. Opportunities and challenges in the use of personal health data for health research[J]. Journal of the American Medical Informatics Association, 2015, 23(e1): e42-e48.

[42] Hintzsche J, Kim J, Yadav V, et al. IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples[J]. Journal of the American Medical Informatics Association, 2016, 23(4): 721-730.

[43] 杨春华, 王天津, 黄思敏,等. 支持精准医疗的国外临床决策支持系统[J]. 中华医学图书情报杂志, 2016, 25(2): 14-19.
Yang Chunhua, Wang Tianjin, Huang Simin, et al. Precision medicine-oriented clinical decision supporting system in foreign countries [J]. Chinese Journal of Medical Library and Information Science, 2016, 25(2): 14-19.

Healthcare big data and the precise medication for rare diseases

WU Zhihui[1], WANG Fei[1], JIANG Zhaoyun[1], MIN Haowei[1], WANG Xinwei[1], GONG Mengchun[2], SHI Wenzhao[1]

1. Digital China Health Technologies Corporation, Beijing 100080, China
2. Central Laboratories, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing 100730, China

Abstract The rapid development of the ubiquitous computing and wearable devices witnesses a new challenge in the natural hand gesture recognition: to free the users from the constraints of the environment and the devices and help the users interact with the environment in a natural and effective way. And the mid-air gesture recognition is one of the effective methods, capable of dealing with the challenge. This paper describes the definition of the mid-air gesture at first, and then analyzes and summarizes the existing hand gesture recognition methods, based on the computer vision, the ultrasonic signal and the electromagnetic wave. At last, this paper discusses the applications of the mid-air gesture recognition, some open questions and the development in the future.

Keywords big data; rare disease; precision medication

（责任编辑　刘志远）