◁信息管理▷

肿瘤大数据平台实践与思考

王海伟[1],石晶[2]
(1. 明智医疗科技(上海)有限公司,北京市 10111; 2. 北京市朝阳区卫生信息中心,北京市 100027)

【摘要】 当前,医疗服务模式正在从传统的经验医学向循证诊疗、个体差异化诊疗、连续诊疗不断迈进,实现疾病的发病风险评估、早期预防、早期诊断、个体化诊疗、参与式管理成为现代医疗服务体系化发展的重要方向。同时医疗行业诊治数据也呈现了爆炸式的增长,医疗领域迎来了自己的"大数据时代"。医院积累了大量不同类型的数据,当前三甲医院,每天接待上万例的患者就诊,患者的诊疗信息是一个庞大的数据集。如何采用大数据技术对海量的医疗数据进行整理、分析、挖掘,提高疾病诊断与治疗的效率,促进疑难病症诊治研究的进展、新药的开发、远程监控以及改善疾病防控研究等,迫切需要大数据平台来解决。

【关键词】 医疗信息;大数据;数据挖掘
【中图分类号】 R197　【文献标识码】 B　【文章编号】 1672-4232(2024)01-0115-03
【DOI编码】 10.3969/j.issn.1672-4232.2024.01.031

1　项目背景及意义

随着移动互联网、智能传感器、云计算、大数据等技术的快速发展,医疗过程的各个环节的数据都可以被完整、准确地记录下来,大数据挖掘也得到了国内外卫生主管部门越来越多的重视。我国医院信息化发展已20余年,80%以上的医院部署了信息系统[1],但在大数据应用层面国内的诊疗信息系统大多还仅处于数据保存这一层级,并且不同的数据零散地存储在不同的业务系统中,无法被有效整合及挖掘。虽然部分领先的医疗机构已经开始建立基于集成平台的临床数据中心(CDR)[2],以解决数据孤岛及后续数据挖掘的问题,但受限于传统的技术手段,数据的处理效率还比较低下。另外,医疗数据中有很大一部分(如电子病历、各种检查报告、影像图片等)是非结构或半结构化数据,传统的技术也无法从中提取有价值的信息[3]。总而言之,医疗大数据相关工作在我国虽已开展多年,但尚处于行业发展初期,各大医院的信息资源数据并未真正应用起来。作为典型的实践科学,医学中有很多知识来源于经验积累,而目前经验积累的最直接、客观的体现就是"诊疗数据"。

肿瘤大数据平台基于大数据技术、机器学习、自然语言处理等技术设立,解决患者诊疗数据分散、重复、孤立等问题,实现对非结构化的数据的标准化、结构化处理[4],深度挖掘出数据的价值,打通院内外患者的诊疗和健康数据,通过随访平台与患者建立更全面的沟通渠道,采集更多患者院外的康复数据,服务于临床诊疗、科研及患者管理等方面。旨在建立患者连续诊疗档案数据库[5],以患者为核心、以诊疗周期为主线,涵盖患者历次诊疗信息、追溯患者在其他机构的就诊记录、搜集患者院外康复信息和体征数据的全方位档案数据,形成一个动态、完整的档案。建立临床决策支持系统,利用大数据挖掘技术使临床决策支持系统更智能,如可以使用图像分析和识别技术,识别医疗影像数据,或者挖掘医疗文献数据建立医疗专家数据库,从而给医生提出诊疗建议。此外,临床决策支持系统还可以使医疗流程中大部分的工作流向护理人员和助理医生,使医生从耗时过长的简单咨询工作中解脱出来,从而提高诊疗效率和质量。因此,利用医疗过程中产生的海量数据,开发其潜在价值,使其助力医疗健康事业的发展,成为医疗行业、技术研发领域等相关有识之士共同努力的目标。

2　平台设计及实现

肿瘤大数据平台根据数据结构和特点分为业务源数据层、源数据采集层、数据标准化整理层、数据集市层、数据应用层。源数据对应着医院业务系统生产的诊疗数据,包括医院信息系统(HIS)、电子病历(EMR)、实验室信息系统(LIS)、医学影像存储与传输系统(PACS)、随访数据等。源数据采集层是将诊疗数据抽取、导入大数据平台,用到的技术有Flume、java等ETL工具。数据标准化整理层主要打破信息孤岛融合诊疗数据,通过自然语言处理技术实现文件数据结构化。数据集市层实现诊疗数据关系型存储,构建疾病的知识图谱。数据应用层可提供病历调阅、科研管理、患者管理、辅助决策、数字疗法等服务。

2.1　多源异构数据的集成

大数据平台需要处理来源于不同业务系统中的数据[6],数据来源的多样化导致数据结构上的不同;通过对数据库内容的机器智能学习[7],实现源数据库的数

据列以及列之间关系的自动判断,构建以患者为中心、以诊疗时间为轴线的诊疗档案。

2.2　数据标准化

在实际操作过程中,数据的标准化又分为两个层次,即数据存储标准化和数据内容标准化。数据存储的标准化主要保证不同数据库之间数据存储格式的统一,数据内容的标准化主要保证不同系统之间,对统一内容采用同样的描述问题。大数据平台在系统设计时参照国内外通用标准,细化了医疗各类数据字段,以确保医疗数据规划的标准始终处于国内外一流水平。

2.3　自然语言处理

自然语言处理用于将医疗机构中的电子病历抽取整理为病人知识库,结合以疾病为核心的知识图谱,联系疾病、症状、治疗等诊疗要素之间的联系,形成完整的病历知识体系[8]。

2.4　数据分析评介模型

基于深度学习技术构建数据分析评介模型,动态反馈诊疗数据存在的问题,实质性提高数据的可用性[9]。深度学习是一种试图使用包含复杂结构或多重非线性变换构成的多个处理层对数据进行高层抽象的算法。目前深度学习使用的神经网络结构包括卷积神经网络与递归神经网络等多种不同的结构的网络。通过对临床诊疗病案、文献指南进行深度学习构建的数据分析评介模型,可以实时分析、反馈、优化诊疗数据。

肿瘤大数据平台主要包括:患者管理、随访管理、临床决策等主要功能模块。其中患者管理功能模块主要实现患者结构化病历的录入(或导入)工作,包括患者基本资料、电子病历信息[10]、检查检验结果、治疗情况、疗效评价、不良反应等数据的采集、修改和查看。随访管理功能模块,可以实现针对患者疾病情况和治疗方案不同制定个性化随访计划,并按时采集患者院外康复数据和患者在其他医疗机构复查复诊信息,包括随访概况、复诊复查、不良事件、生存状态、患者教育等信息的采集、修改和查看。临床决策功能模块,可以实现针对患者疾病进展阶段和检查检验情况个性化推荐治疗方案和复查计划,并按时进行随访和复诊提醒。

3　平台效果及总结

肿瘤大数据平台处理的诊疗数据范围包括:患者基本资料、体格检查、主诉、既往史、个人史、肿瘤家族史、血常规、血生化、乙肝五项、HBV-DNA、丙肝、凝血功能检查、CT/增强CT/PET-CT、超声、DSA、肿瘤标志物、基因检测、病理检查、诊疗、治疗前评估、肿瘤治疗、伴随用药、疗效评价、不良事件、院外随访等,贯穿患者整个治疗生命周期。诊疗数据结构化程度高、颗粒度细,一份病历多达2 000个结构化、标准化数据元。通过大数据平台的建设,即可以实现医院积累的大量数据进行深度的分析、挖掘,建立专项的科研课题进行回顾性和前瞻性的科研分析,还可以试探找寻体征、诊断、用药、治疗方式等指标的内在相关性,分析医生的诊疗路径,优化指南,形成更加科学的诊疗知识库,机器辅助诊疗的基础;诊疗中医生、护士可及时对患者自身情况制定个性化诊疗方案并优化整个诊疗过程。其中肝癌大数据平台涵盖了肝癌的临床、流行病学的各个方面,是目前肝癌研究项目中数据项最全的一项研究,建立了规范化的标准数据格式,为后续的相关疾病的诊疗平台提供了统一的可比较的数据,为医疗机构间的业务沟通、数据共享打下了坚实基础。

通过大数据处理技术、机器学习技术和人工智能结合分析海量庞杂数据并挖掘疾病诊断、治疗、转归等领域中的因果关系或相关关系建立模型以提高疾病诊疗的整体效能,避免了传统统计学方法先假设后验证思路的挂一漏万的缺点。可有效分析疾病各种复杂因素之间的影响。规范化的疾病领域词语、词条可对临床大量的非结构化的数据进行清洗、归一、结构化处理,帮助临床医生更加轻松和准确地利用这些数据,实现病历的多维度检索。

肿瘤大数据平台首先较好地实现了以疾病为模板的结构化数据采集系统。将既往仅电子化的病历导入结构化病历系统,形成结构化、标准化数据。结构化电子病历是指从医学信息学的角度将以自然语言方式录入的医疗文书按照医学术语的要求进行结构化、标准化分析,并将这些语义结构最终以关系型(面向对象)结构的方式保存到数据库中。从而将所有的医疗数据元素化、简单化、模块化,为医疗科研及管理创建大容量的知识库,为今后数据挖掘奠定了基础,同时也利于医院数字化信息化建设。其次,建立了数据清洗治理平台,保证医疗数据由业务系统向结构化病历系统导入过程中的数据准确性,垃圾数据、重复数据的自动匹配删除,为数据挖掘应用打下了坚实基础。再次,建立肿瘤辅助诊疗AI学习系统及诊疗机器人。以数据驱动的临床决策系统,基于平台中的癌患者数据并利用大数据分析技术使得诊断治疗更加智能,提高医疗工作者工作效率和医疗服务的质量。最后,可配置的结构化病历输入系统,使新增病人入院后即形成结构化病历。基于已形成的结构化病历系统、结构化数据及分词系统,建立结构化病历输入系统,与医院信息系统对接,即系统生成结构化病历。