

DOI: 10.3969/j.issn.1005-8982.2020.09.025

文章编号: 1005-8982 (2020) 09-0124-05

学术报告

基于信息技术的临床科研数据平台建设

段鹤李, 陈廷寅

[中南大学湘雅医院 网络信息中心 (中南大学医院管理研究所), 湖南 长沙 410008]

摘要: **目的** 分析信息技术在临床科研中心的作用与功能架构方法, 为建立一套高效的临床科研中心工作体系提供依据。**方法** 通过分析临床科研中心的工作现状及患者信息数据的处理方法, 寻找现阶段信息技术在临床科研中心功能架构中可能存在的问题, 并提供可行的解决方案。**结果** 建立临床数据分析方法, 并据此形成符合理论和实践经验的临床诊疗决策路径和规则。**结论** 可通过应用诸如机器学习等信息技术手段, 推进临床科研中心的工作效率。

关键词: 信息技术; 临床科研数据平台; 数据分析; 数据库

中图分类号: R197.323

文献标识码: C

近年来, 数据分析与挖掘技术快速发展, 其在医疗健康领域的影响与作用也越来越受到重视^[1]。全国卫生与健康大会、《健康中国 2030 规划纲要》、《“十三五”卫生与健康规划》和《全民健康保障工程建设规划》都对医疗健康信息化与大数据应用发展作了规划部署。国务院办公厅于 2016 年下发的《关于促进和规范健康医疗大数据应用发展的指导意见》, 对发展健康医疗大数据, 推动健康医疗数据互联互通、开放共享具有推动和指导意义^[2]。

医学要发展和突破就离不开科研, 在当下还有很多疾病的发病机制、发病原因没有明确, 治疗方法、诊断方法等都还有待进一步改进, 但随着医疗信息化程度的提升, 通过对医疗过程和结果数据的分析、监控, 可以有效地发现医疗过程中出现的问题, 并分析发生的原因, 有针对性地进行改善, 提升医疗质量。这就需要医务人员, 特别是临床一线医务人员从日常诊疗过程中积累经验, 见微知著, 就未解问题积极进行研究。然而医疗数据量大且关系复杂, 要对这种大量数据的关系进行发现和分析, 依靠传统的计算框架已很难满足要求。

1 医院科研数据应用现状分析

医院临床业务每天会产生大量的诊断、医嘱、生

命体征、检验、检查等方面的数据, 但这些数据在信息化过程中未以科研为目标进行采集和组织, 从而导致无法直接为科研所用, 需要再次将这些数据根据科研的需求进行组织。不仅需要整合本院的患者数据, 而且还需要整合患者在其他医疗机构和保健机构产生的医疗、健康、生活习惯、居住环境、职业等方面的全生命过程数据, 为科研建立起完整的、以患者/居民为中心的科研大数据中心。但这些数据大多是基于患者治疗产生的数据, 更多是以完成诊疗工作为主, 而没有从更全面、更深入、更智能的角度上对数据进行很好地分析和利用, 从而将数据提炼成信息, 凝聚成知识, 为医院的临床和管理服务。

随着医学科学的深入发展, 临床分科精细化更有利于医生学术有专攻和对疾病的有效治疗, 而目前的电子病历针对不同专科和疾病的特色进行细分方面还需要进一步完善, 以支撑医护人员快速精准定位、精准分析, 并适应医疗专科化、专病化的发展。患者病历、检查、检验和随访资料是宝贵的科研资源, 医生可以对病例数据进行筛选和分析, 为临床科研提供坚实的基础^[3]。目前, 医院虽然实现院内信息系统的互联互通, 汇聚不同业务系统的数据, 但由于数据质量不高、数据异构性矛盾、缺乏专业的数据分析平台等问题, 离实际科研应用尚有一段距离, 还需要根据专科需要

收稿日期: 2019-12-05

[通信作者] 陈廷寅, E-mail: 366395856@qq.com

实现科研数据的合理、高效使用。临床数据与科研数据尚未紧密融合,科研与临床的良性互动还不够。

2 建立临床数据分析的方法

2.1 数据预处理

数据处理是以提高临床科研效率、促进业务协同服务、辅助临床决策支持、夯实大数据技术支撑为目标,按标准构建院级科研大数据中心,并向上层应用提供数据服务。数据范围主要包括基于医院临床数据中心的医院临床业务数据、基于专科联盟及医联体的医疗卫生机构服务数据、基于医疗设备(含穿戴设备)的健康体征监测数据及基于专科疾病的科研量表数据等。主要方法包括数据清洗、数据变换和数据规约^[4]。

数据清洗:即进行异常数据清除,错误纠正及重复数据的清除。清洗数据时,主要解决空缺值、错误数据、孤立点、噪声等问题。清洗后的数据达到格式标准化。

数据变换:将数据转换成适合挖掘的形式称为数据变换,采用线性或非线性的数学变换方法,将多维数据压缩成较少维数的数据,以消除其在时间、空间、属性、精度等特征上的差异。

数据规约:即去噪处理后,还要对数据的属性进行相应处理。在减少数据存储空间的同时尽可能保证数据的完整性,获得比原始数据小得多的数据,并以合乎要求的方式对数据进行表达^[5-6]。

2.2 自然语言处理

自然语言处理是将非结构化数据转化为统一的结构化数据的重要方法,医学信息抽取一直是医学自然语言处理的研究热点。患者信息在医疗病历中常以文本形式存储,而文本描述容易出现非标准化描述,因此对自然语言必须进行技术处理,包括医学名实体识别、医学名实体自动编码、医学名实体修饰词识别、时间信息抽取等^[7]。

医学自然语言处理可以分为两类:①分析内容特征,主要包括中文分词、命名实体识别、文档语义分析、机器学习、同义词替换等;②分析对象粒度,主要包括句法结构分析、依存关系分析、文本聚类、预制模板、知识推理等。通过处理,研究者能够从已转化为真实于原有医学信息的结构化数据中方便地抽取有用的医学信息,从而减少运行成本^[8-9]。

2.3 知识库构建

从诊疗指南、医学教材、医学文献等资料中根据

一定的模型整理出疾病的静态知识库,再经由信息化手段形成诊疗决策路径,基于诊疗决策路径自动提取出疾病的关键指标形成疾病分析模型,提取路径中的决策规则形成决策规则库。通过对医院历史临床业务数据的整合后形成数据中心,根据语料库、关系和规则库、基础知识库、疾病模型,对历史的数据进行深度的挖掘和分析,提取数据中的医学术语丰富语料库、提取关系和规则、提取诊疗经验。

将人工整理的知识库和系统自动提取或学习的知识进行相关验证,并修正疾病诊疗决策路径和规则,最终形成符合理论和实践经验的临床诊疗决策路径和规则;并基于这些路径和规则形成知识图谱,形成的知识图谱、诊疗决策路径和规则、疾病模型可以应用于医院端的知识推荐、风险评价、用药推荐、医保控费、医疗质量控制、相似病例推荐等;应用于居民端的合理医疗评估、个人疾病评估、健康风险评估、疾病知识学习等。见图 1。

2.4 数据分析挖掘

医疗数据需要借助于大数据技术的发展,结合传统的数据挖掘分析理论与技术,从海量数据中高效地提取有价值的信息。数据分析挖掘以数据为中心,以数据预处理、特征工程、统计分析技术、机器学习与深度学习算法模型为基础,存储及计算能力以元数据、数据标准、数据质量为治理核心,以信息可视化技术作为呈现手段,构建统一的数据分析挖掘平台。

把数据属性转化为数据特征,减少受到噪声的干扰,以满足数据挖掘和分析,是数据进行挖掘和分析之前对数据进行特征工程处理的必要过程,但是对于特征工程中引用的新特征,需要验证其是否提高预测的准确度而排除其是否为无用的特征,以免增加算法运算的复杂度^[10]。然后从数据统计分析方法、机器学习算法、深度学习算法等方法中选择合适的算法模型或组合基于特征数据进行模型的设计与挖掘分析。

2.5 数据质控

数据质量是数据的正确性、完整性、一致性及最小性这 4 个指标在信息系统中的满足程度^[11-12]。首先要进行质控规则的管理,根据业务领域,提供数据质控指标的定义,编写数据质量校验程序,确保数据唯一性,避免从多处提供数据来源,减少数据不一致性的麻烦,并提供完整的数据质控日志,可通过日志查看数据质控情况;其次要提供数据校验功能,具体包括数据规范性评价指标、完整性评价指标、及时性

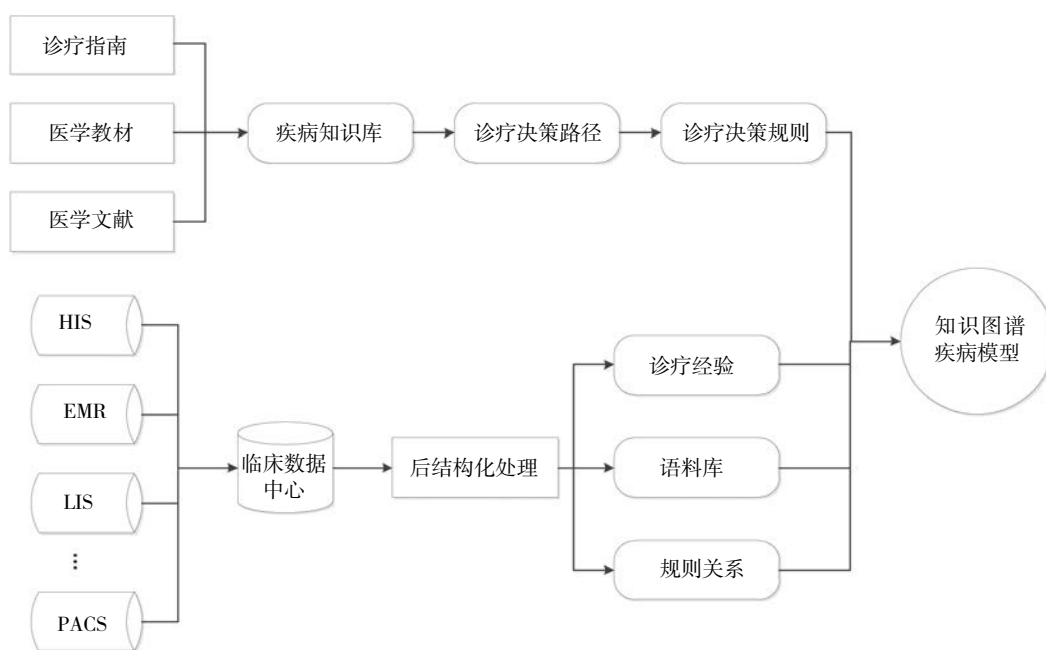


图 1 知识库构建流程

评价指标、逻辑性评价指标、一致性评价指标和异常值探测指标^[13]；最后要进行数据质量监管，在每一个 ETL 阶段中设置检查点，如在数据源层、业务存储层、数据仓库明细数据等，在每个检查点上对用户要求的关键指标进行计算，检查各阶段计算出来的指标是否一致，以及误差是否在用户的许可范围内，并提供数据质量监控审计功能，利用业务和校验规则，实时或定时监控数据质量，当数据质量低于设定的阈值时发出警告。

3 应用实践

临床科研中心的总体设计是在信息标准体系、安全保障体系的支撑下，通过对院内各种不同数据标准、技术接入方式的系统的数据集成，形成院内临床数据中心；专科联盟各医院的数据通过数据报送或上传的方式将其院内的数据进行汇聚；对于各种医疗设备和个人穿戴设备通过提供标准化接口的模式汇聚后形成医疗设备数据接入中心，最后通过数据交换与共享引擎完成 3 个来源数据的采集、清洗、标化、质控及共享形成科研数据库；再根据不同科研主题的需求建设相应的科研专题数据库；通过数据服务体系中的主索引服务、数据服务、数据资源服务、数据安全、数据挖掘分析、搜索引擎等服务为上层 10 大类数据应用提供服务支撑，同时基于数据服务体系来支撑其他专科数据应用的扩展。

3.1 数据汇聚

院内各业务系统包括院级建设和科室级建设的系统，通过以患者为中心的临床数据整合，将 HIS、EMR、LIS、PACS 等系统的数据抽取到临床数据中心 CDR，针对数据的来源可以直接基于其逻辑结构，按科研数据中心的结构进行转换。

对院外各联盟机构的数据，采用 CRF 表单填报、SOAP 协议传输服务、消息传输等数据汇集与集成方式，一种方式是由临床科研中心提供统一数据上报系统，对接联盟机构的数据资源管理平台或前置数据库，主动抓取经过标化清洗的数据；另一种方式是由各个联盟医院成员以协定的数据格式，应用 CRF 表单进行数据填报，经审核后向临床科研中心上传数据。

各类医疗设备的医疗数据也是临床科研数据的重要组成部分，医疗设备数据汇集常用 SOAP 协议传输、前置库采集或消息传输模式，通过建立统一的设备数据接入中心，对各类医疗设备进行数据采集。见图 2。

3.2 搭建全量科研数据库

全量科研数据是整合患者和疾病全生周期的数据，通过从数据汇聚库经过标化处理后，根据相关标准建立的数据模型而存储的数据，主要的数据域包括门诊和急诊患者全量医疗数据、住院患者全量医疗数据、生物样本库数据、医疗设备和仪器数据、联盟医院全量数据、穿戴设备全量数据等，并将这些数据按

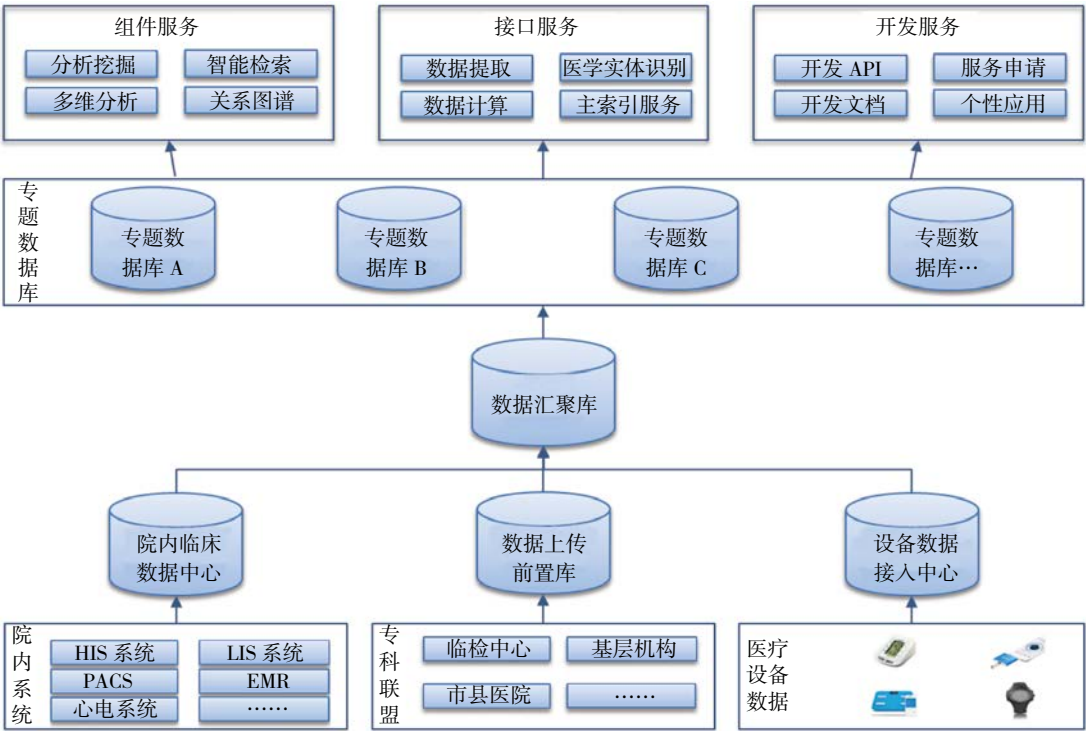


图 2 临床科研中心架构图解

患者、医疗实体对象管理、医疗事件等闭环业务数据链进行组织和逻辑存储。

对非结构化数据，进入数据中心，即存储结构化处理后的数据，又存储原始的非结构化数据，以方便数据的利用。对医疗影像数据可以通过与 PACS 系统建立数据交互的模式，数据中心只存储影像的索引信息而达到影像数据的共享，以减少存储的投入成本。

3.3 专病数据库

专病数据库的建立是为专病研究的需要，以专病的数据模型为核心，从全量数据中心抽取专病的数据而形成，并建立起数据自动增量更新的机制，保证专病科研数据库数据的完整性与及时性。专病科研数据库是交由科研的管理人员，在授权的前提下，以诊疗决策路径为基础建立起数据模型后，利用数据平台提供的数据提取功能自行从全量数据中心提取，并根据科研的需求和科研人员安排，建立起专病科研数据应用的管理体系，包括数据授权、数据访问日志、数据质量审查等。

3.4 数据服务

数据服务分为 3 类：①基于微服务架构的，可以直接提供给其他系统进行界面嵌入，一个较完整的业务组件服务，按照一定的业务域需求，利用数据中心底层的各种资源和能力，直接将需求封装一个个可

以直接使用的业务功能；②接口服务，将数据中心的数据或能力组织成一个个的包含特定功能的接口供上层应用调用；③开发服务，主要是将底层的数据资源、计算资源通过接口的形式提供给应用层进行开发，这一层服务开放的粒度更细，方便应用进行深层次的应用数据中心的各种资源。

3.5 数据分析应用

基于临床科研中心搭建的各个专病数据库和提供的数据服务，可以直接应用于临床应用，如临床决策支持系统是一个基于临床知识库的人机交互医疗信息技术应用系统，它把散布于各个系统的数据按不同专业关注点进行整合及区分，实现数据的抽提、加工和转换，通过数据、模型等辅助完成临床决策^[4]。如合理用药系统也是一个基于知识库的临床用药辅助系统，可以结合系统知识库与人工智能规则设定，对用药安全做到事前监测、事中控制、事后分析，全方位为用药安全提供保障。也可以根据数据挖掘分析的结果优化医疗流程，如分析历史路径数据，根据分析结果调整路径阶段设置和路径项目设置，实现临床路径的持续优化。还可以衍生出基于数据分析的应用，如基于患者的全病程数据，采用深度学习、机器学习和自然语言处理技术，对数据进行挖掘分析和患者诊疗路径关键指标结果，提取患者的特征，采用可视化技

术,形成患者画像,对患者可能患病进行精准预测。

4 结语

临床科研中心的建设,可以借助专病数据库挖掘疾病特征,支撑疾病病因探索的各种流行病学研究及临床辅助决策,可以发挥优质医疗资源优势,辐射带动区域及联盟体内各级医疗卫生机构的医疗水平提升,服务于院前、院中、院后、科研、教学及管理 6 大业务场景,改善与提高整体医疗质量,提升临床科研能力,提升医院的整体综合实力,最终为医疗及其联盟机构的相关业务人员建立起有效的业务协作机制。

参 考 文 献:

- [1] 孙仁华,刘景全,杨向红.急性肾损伤遇见大数据时代:为我们带来了什么[J].中国急救医学,2018,2: 97-102.
- [2] 国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见国办发[2016]47号[S].
- [3] 朱文婕,陈玉娥,蒋秀林.基于计算思维的医学院校大学计算机课程改革初探[J].工业和信息化教育 2014, 12: 74-77.
- [4] 杨国林,王飞,贺慧.基于数据挖掘的图书馆数据预处理方法研究[J].电子设计工程,2015,3: 26-29.
- [5] 张卫东.基于多维度属性权重优化的 FCM 聚类算法的图书管理数据聚类研究[J].农业图书情报学刊,2016,6: 50-57.
- [6] 张凯.数据挖掘技术在医疗费用数据中的应用研究[D].北京:北京邮电大学,2015.
- [7] 杨晨浩.基于深度学习的中文电子病历实体修饰与关系抽取研究及算法平台开发[D].哈尔滨:哈尔滨工业大学,2016.
- [8] 刘红霞.面向慢性病海量数据问答系统智能摘要算法的研究与实现[D].开封:河南大学,2016.
- [9] 吴宇皓,蒋琳,周奕.医疗大数据分析管理系统的设计与科研应用[J].中国数字医学,2017,6: 53-55.
- [10] 陈凤娟.面向数据流的频繁项集挖掘算法研究[J].洛阳师范学院学报,2015,2: 82-85.
- [11] 高运华.基于数据仓库的数据质量分析和评估[J].黑龙江科技信息,2014,20: 165-165.
- [12] 曹建军,刁兴春,张潇毅,等.数据质量控制研究中若干基本问题[J].微计算机信息,2010,9: 12-14.
- [13] 赵金.面向智慧医疗的区域卫生信息系统的研究和实现[D].南京:东南大学,2016.
- [14] 赵妍,王颖,邵伟.基于数据仓库的临床决策支持系统在我院的应用[J].中国医疗设备,2016,7: 95-97.

(张西倩 编辑)

本文引用格式:段鹤李,陈廷寅.基于信息技术的临床科研数据平台建设[J].中国现代医学杂志,2020,30(9): 124-128.