

面向科研的临床结构化数据预处理方法研究^{*}

车贺宾¹ 何昆仑¹ 吴 欢¹ 陈媛媛¹ 王万玲¹ 王 飞² 刘立永² 尚亚飞²

(¹ 中国人民解放军总医院医学创新研究部 北京 100853

² 北京嘉和海森健康科技有限公司 北京 100085)

〔摘要〕 目的/意义 分析临床回顾性研究中结构化数据的主要问题和预处理技术, 为临床数据预处理研究提供新思路。方法/过程 基于中国人民解放军总医院多个医学中心 10 余年历史数据治理实践, 梳理一套以数据融合、简约、清洗为主的预处理方法, 并总结其效果。结果/结论 数据预处理解决了临床科研中历史数据的常见问题, 减轻了数据二次加工工作量。基于此建立的临床科研检索系统, 有效提高了数据的可用性和价值。

〔关键词〕 临床科研; 数据预处理; 数据清洗

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.03.010

Study on the Preprocessing Method of Clinical Structured Data for Scientific Research

CHE Hebin¹, HE Kunlun¹, WU Huan¹, CHEN Yuanyuan¹, WANG Wanling¹, WANG Fei², LIU Liyong², SHANG Yafei²

¹ Medical Innovation Research Department of PLA General Hospital, Beijing 100853, China; ² Beijing Goodwill Hessian Health Technology Co. Ltd., Beijing 100085, China

〔Abstract〕 **Purpose/Significance** To analyze the main problems and preprocessing technical methods of clinical structured data in clinical retrospective studies, so as to provide new perspectives for clinical data preprocessing research. **Method/Process** On the basis of more than 10 years historical data governance in several medical centers of the PLA General Hospital, a set of preprocessing methods focusing on data fusion, simplification and cleaning are sorted out, and the effects are summarized. **Result/Conclusion** Data preprocessing solves the common problems of historical data in clinical research and reduces the workload of data secondary processing. The clinical research retrieval system is established on this basis to effectively improve the availability and value of data.

〔Keywords〕 clinical research; data preprocessing; data cleaning

1 引言

临床结构化数据包含患者在医院诊疗过程中的

人口学基本信息、就诊情况、体征、检验结果、检查报告、病历文书等, 普遍存储于关系型数据库。此类数据对医学科研至关重要, 通过数据分析可深入了解疾病发病机制、诊断及治疗方法, 进而提高医疗质量和患者生存率。高质量数据是数据分析的基础, “脏数据”会导致不可靠输出, 其后果可能难以弥补^[1]。数据预处理有助于提升大数据的准确性、完整性、规范性、时序性和唯一性^[2], 因此, 数据预处理方法研究意义重大。

针对临床原始结构化数据存在的数据分散、结

〔修回日期〕 2024-10-23

〔作者简介〕 车贺宾, 工程师, 发表论文 17 篇; 通信作者: 何昆仑, 主任医师, 教授, 博士生导师。

〔基金项目〕 新一代人工智能国家科技重大专项 (项目编号: 2021ZD0140406)。

构复杂、内容冗余、缺失、逻辑矛盾及值域混乱等问题，本研究提出一套以数据融合、简约、清洗为主的预处理方法，并用于真实世界数据治理，取得了良好效果。

2 临床数据的主要问题

2.1 数据分散和结构复杂

医院内部信息系统众多，因上线时间、厂商不同，数据存储和结构多样化，阻碍数据共享和交流，融合难度大^[3]。医院信息系统服务于医院运营，对于科研而言，临床原始数据存在范围大、字段多、报告文书难拆分等问题。

2.2 内容冗余和逻辑矛盾

临床数据中存在同类数据来源于不同系统的情况，内容相似但不一致，易导致检索结果缺失。真实世界中，医院临床数据存在逻辑矛盾，如业务流程重复（同一人次入院记录却有多条），需对比生成时间和内容完整度删除重复；业务流程缺失（检查、检验记录缺失患者门诊就诊或住院记录），需关联最近就诊或新生成就诊；业务流程时间差逆向（会诊时间小于入院时间），需对比该次住院临床数据生成最早时间，追溯问题源头并更正。

2.3 值域混乱

临床数据的准确性直接影响疾病诊断、治疗和预防。然而，数据元值域标准不一致或缺乏行业标准，导致数据元值域混乱，严重影响数据质量，如手术名称、辅助检查项目名称、检验项目名称等。实验室检验项目和生命体征结果中的错误值和异常值区分困难，缺乏明确标准和自动化识别工具。诊断记录中的发病部位和疾病严重等级备注也干扰数据准确性。

3 临床数据预处理方法

3.1 临床科研通用数据模型和标准术语集

临床科研通用数据模型以患者为中心，梳理

医疗事件，抽象门诊、急诊、体检、住院就诊类型，删除就诊过程中的预约、排队、审核等记录，只保留结局数据，并进行标准化转换^[4]。按此标准，筛选、合并中国人民解放军总医院（以下简称解放军总医院）多个医学中心信息系统千余张数据表，形成由14部分组成的临床科研通用数据模型，包含人口学信息、病史、风险因素、就诊、诊断、用药医嘱、处置医嘱、手术、费用、体征、护理评估、检查报告、检验报告、病历文书和临床结局等。业务主数据包括机构数据标准（科室、人员、设备、耗材编码等），以及区域和行业标准（国际疾病分类、医学系统临床标准术语、实验室逻辑测试结果和体检发现、开放医疗与健康联盟标准术语集等）^[5]。除现有编码数据标准外，还需制定临床科研医学检验、检查、症状等标准术语集。

3.2 患者主索引

医院门诊、急诊、住院、体检以及其他专科检查、检验系统的患者标识不同，导致患者数据分散。为整合不同系统患者信息，采用患者主索引（enterprise master patient index, EMPI）技术生成患者唯一标识。根据患者基本信息建立数据模型，包含姓名、性别、出生日期、证件号码、住址、联系电话等。EMPI匹配引擎需先配置规则，基于不同规则计算最终得分，进行最优化匹配。规则配置包括“证件号码规则”“基本信息规则”和“联系方式规则”，见表1。按顺序匹配，未通过则进入下一规则，见图1。

表 1 EMPI 匹配规则

规则名称	合并分	匹配字段	匹配分值	匹配类型
证件号码规则	≥100	身份证号	100	精确匹配
基本信息规则	≥80	姓名	15	精确匹配
		性别	5	精确匹配
		身份证号	80	相似匹配
联系方式规则	≥95	姓名	15	精确匹配
		联系电话	80	相似匹配
		联系地址	5	相似匹配

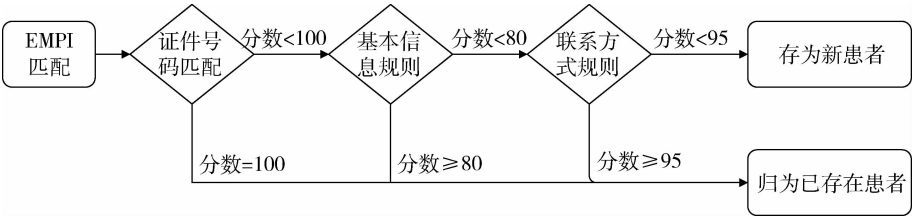


图1 EMPI 匹配流程

3.3 数据补全

不同于特定科研课题中采用的缺失值处理方法，如删除法、插补法等，全院级临床数据预处理中补全的数据应明确且可追溯。采用目标数据的多来源合并及非直接业务系统的数据补充方式。一是目标数据的多来源合并，如诊断信息存在于医师诊断记录、病案科归档诊断记录、入院记录文本中初步诊断、出院记录文本中出院诊断等多个数据源^[6]。病案首页诊断记录模板有条目限制，共病过多则无法完全记录，检索结果可能存在偏差，因此将包含病案首页的各来源诊断信息融合为一张表，并标明诊断来源。二是目标数据非直接业务系统的数据补充，因专科系统上线时间不一或记录不全，需依据其他系统侧面补充数据。如患者输血量计算，输血系统上线晚、维护差，导致数据缺失；研究手术耗材 S-ROM 假体与患者预后关系时，手术耗材存在于病历文书章节-手术记录中，但并非所有记录均明确耗材类型。鉴于费用明细表上线早、运行稳定、内容全面，上述情况可采用费用明细表中缴费项目补全。原记录条数与费用明细补全后条数对比，见表 2。

表 2 费用明细补充输血量 and 手术耗材检索条目前后对比			
字段名称	原记录数 (条)	费用明细补全数 (条)	补全前 缺失率(%)
输血量	17 310	18 755	9.23
手术耗材 S-ROM 假体	965	1 217	24.07

3.4 医学术语归一处理

医学术语归一处理包括通用基本规则、个性基本规则、获取知识库标准、分词、分词结果处理等流程。通用基本规则主要包括英文大小写转换、标点符号全半角转换等常规字符统一规则。个性基本规则基于真实数据，参考知识图谱中不同本体的标准术语，针对不同本体设计一些单独的规则，例如否定项去除、字母简写判断、拆分、倒置处理等。获取知识库标准指通过调用医学术语数据，获取本体中同义词和层级关系，从而找到数据资源体系中真实数据所对应的标准术语名称。分词基于医学术语中的数据，利用自然语言处理技术识别实体和关系，并进行结果处理，摒弃干扰内容，生成标准名。医学数据归一处理流程，见图 2。

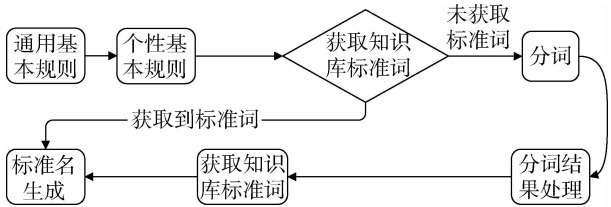


图2 医学数据归一处理流程

针对手术和检查本体在临床数据中存在较多复合词的特点，分词结果需要细粒度细分处理，如对手术复合词中的入路、部位、操作等，检查复合词中的部位、检查类型等信息的再提取，从而对应到知识库中的多个标准词。手术和检查本体复合词拆分归一处理流程及结果，见表 3，某数据集医学术语归一处理效果示例，见表 4。

表 3 手术和检查本体复合词拆分归一处理

序号	原始词	实体识别	语义提取结果	标准词
1	足血管、神经、肌腱探查术、坏死组织清创术	足血管（1）、神经（1）、肌腱（1）、坏死组织（1）、探查术（0）、清创术（0）	足血管探查术、足神经探查术、足肌腱探查术、足坏死组织清创术	足血管探查术、足神经探查术、足肌腱探查术、足坏死组织清创术
2	右下肢吸脂、病变组织切除术	右下肢（2）、病变组织（1）、吸脂术（0）、切除术（0）	右下肢吸脂术、右下肢病变组织切除术	右下肢吸脂术、右下肢病变组织切除术

表 4 医学术语归一处理效果示例

类型	源数量（个）	归一数量（个）	源样例	归一结果样例
检验	1 419	1 146	C 反应蛋白、C 反应蛋白（CRP）、C 反应蛋白测定	C 反应蛋白测定
检查	954	593	腕关节正侧位单、腕关节正侧位（单）、腕关节正位、腕关节正位（双）	腕关节 X 线检查
诊断	31 234	18 923	肺恶性肿瘤、肺上叶恶性肿瘤、肺癌、肺 CA	肺恶性肿瘤
手术	2 977	2 506	Carto 指导下心房扑动射频消融术、CARTO 心房颤动射频消融术、Carto 指导下房颤射频消融术	CARTO & 房颤射频消融术

3.5 数值规范清洗

检验、体征等项目存在大量数值类型结果值，在实际业务系统中经常出现>、<、+、-、阳性、阴性等附加说明字符，以及“（已复查）”等临床科研无意义字符。通过去除数值数据中的特殊字段规范定性和定量结果，将定性结果、定量结果作为不同字段存储，见表 5。检验、体征等项目中的部分极值很难作错误值和异常值判断，根据医学知识和临床经验，为常见项目设定正常范围的上下阈值。当数据超出这些阈值，标记为异常数据待处理，见表 6。

表 5 数值规范清洗示例

项目名称	原始数据	定性结果	定量结果
潜血试验	阴性（已复查）	阴性	—
C 反应蛋白测定	≤138.0	—	138
血浆 D-二聚体测定	>20（已复查）	—	20
¹³ C 尿素呼气试验检查	DOB: 7.4（阳性）	阳性	7.4
幽门螺杆菌			

表 6 体征错误值清洗示例

项目类型	错误值样例
呼吸	113, 020, 1 020, 2
体温	120, 00, 6.4, 36 *5, 14: 00
心率	65 536, 183, 0, 23

3.6 自然语言处理

临床文本数据中包含大量非结构化信息，需借助自然语言处理技术，根据科研需求进行数据后结构化。调研和理解电子病历及其他来源的临床数据，探索层次型实体关系抽取模式^[7]。该模式将文本数据表达为实体和嵌套关系，递归抽取，解决传统方法准确率随实体和关系类型增多而下降的问题。新实体类别提取时，通过增量训练快速添加，无需重训整个模型，减少训练时间和资源消耗，提高适应性和灵活性。通过对电子病历文书类别、段落、结构等属性信息的分析，本研究对电子病历进行后结构化处理主要经过以下 3 个处理阶段。一是病历文书预处理阶段，主要对文书内容进行预处理操作。以 html 格式入院记录文书为例，首先对 html 内容进行解析，提取病历文书中所有正文信息，表格信息按规则解析合并；其次对解析后文本数据进行规范化处理，包括全半角转化、非法字符去除、标点补全等；最后是完整性校验，根据电子病历规范要求，如入院记录中必须包括“一诉五史”等信息，制定关键词匹配规则确保病历完整性。对于发现的异常，及时修正处理程序，保证数据的准确性及完整性。二是段落级治理阶段，结合文书段落特点及规范，采用规则匹配和相似度计算完成。根据

章节名称与内容分隔规则分割段落，对章节名称进行归一化处理，如“既往史”和“过往史”归一化。实现病历文本关键章节精细拆分，为后续语义级信息抽取、提取奠定基础。三是语义级治理阶段，主要分为两个步骤，见图 3。先采用规则模型、中文分词工具、BiLSTM + CRF 模型抽取实体，对文本内容进行命名实体识别；再通过规则模型结合 BERT + BiLSTM 模型，对命名实体结果进行实体关系构建，完成段落内容的语义级信息抽取。该抽取模式提升了文本后结构化的处理能力和复杂医疗信息处理的效率和准确性，为临床决策提供更丰富和精准的数据支持。

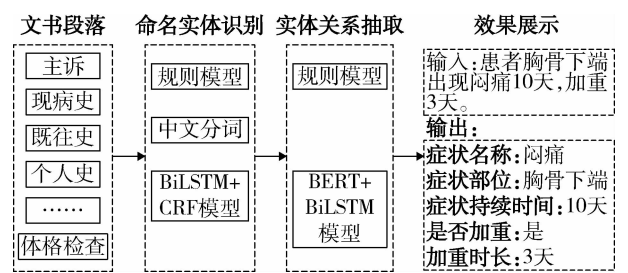


图3 文本数据处理流程

3.7 元数据数据血缘

数据从原始记录到目标字段的转换历经复杂映射与加工，增加了溯源难度。数据血缘描述数据流转关系，旨在解决数据源追溯、一致性、异常分析、质量控制、资源优化及合规审核等问题。基于元数据的数据血缘分析，凭借全生命周期管理优势，成为数据标准管理的关键。构建数据血缘能精确追踪数据路径，确保数据标准一致实施，优化资源配置，满足合规要求^[8]。数据预处理阶段，元数据根据处理方式划分为直接映射和非直接映射。直接映射指元数据直接来自原始数据源，未经额外处理。非直接映射包括简单加工、归一化、自然语言处理等，经过清洗、格式转换、标准化、文本后结构化等数据治理操作。非直接映射元数据经多次映射及复杂加工后，溯源原始记录困难，溯源机制的建立尤为关键。一是映射类型划分，治理后字段分为直接映射和非直接映射。二是治理任务溯源，非直接映射字段根据数据治理结果时间戳，溯源至相

应治理任务，再溯源至数据任务版本。数据任务版本包括表关联、字段映射、加工函数等信息。三是加工函数解析，解析数据任务版本中的加工函数，追踪字段原始来源。解析过程包括明确函数输入、输出及加工过程涉及字段，分析字段清洗、转换、标准化等处理算法和循环、条件判断等控制结构。基于上述溯源过程，直至治理后字段可直接映射到原始字段，循环结束，见图 4。完整数据血缘链条支持可视化展示，直观展现数据流动和转换过程，帮助用户快速理解和追踪数据源头。

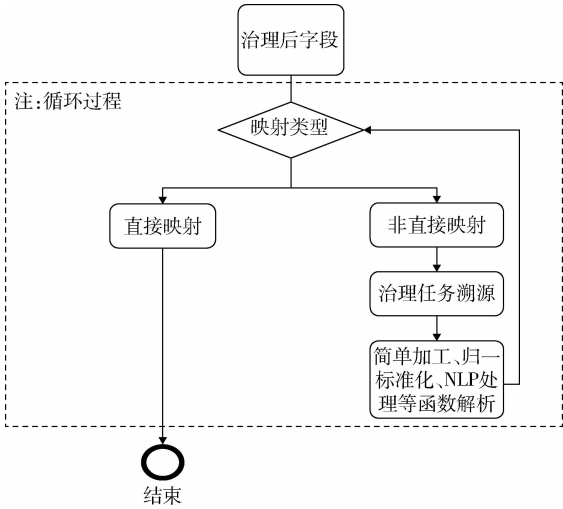


图4 数据血缘溯源流程

4 数据预处理结果和应用

数据预处理结果评价指标包括准确性、一致性、完整性、时序性、唯一性、规范性等^[9]。医院原始数据存在问题、处理方法和相关评价指标对应关系，见表 7。为验证临床数据预处理方法的有效性，采用解放军总医院 2014 年 11 月—2021 年 9 月急诊抢救间收治 5 万余例重症患者的原始急救数据集作为治理目标，取得良好效果，见表 8。急救数据集包括患者基本信息、实验室检查、用药医嘱、生命体征、手术操作、病历文书等。基于预处理的临床结构化数据可直接用于统计分析，减少重复建设，降低数据挖掘成本，有利于数据在疾病关联探索、临床结局预测、质量监测评估、专科疾病研究等临床科研场景发挥价值^[10]。

表7 数据问题、处理方法和评价指标关系

存在问题	处理方法	评价指标
数据分散和结构复杂	构建临床科研通用数据模型，ETL、EMPI 技术	准确性
内容冗余和数据缺失	目标数据多来源合并，非直接业务系统数据补充	完整性
逻辑矛盾	规则识别标记、删除重复记录、关联最近的或生成新的就诊记录	时序性、唯一性
值域混乱	标准术语集、医学术语归一、数值规范清洗、自然语言处理、数据血缘	规范性

表8 解放军总医院原始急救数据集预处理效果示例

指标	指标结果描述
准确性	32 张业务系统表、66 个字典表数据融合，总诊次数量超 5 万，数据融合后准确性达到 99.99%
完整性	入院治疗患者 2 万余人次，病案首页缺失 128 例，占比 0.47%
时序性	入院记录创建时间早于入院时间 956 例，占比 3.52%
规范性	体征结果异常 194 114 条，占比 0.73%；检验结果空值 84 890 条，占比 3.95%
唯一性	入、出院记录分别重复 945 条（3.49%）、1 310 条（4.83%）

5 结语

统一标准的临床结构化数据预处理方法对提高临床数据质量和可用性至关重要。高质量数据便于统计、分析，挖掘数据价值，加速数据向知识转化。统一标准增强数据可用性，使数据更易存储、检索和交换共享，且数据间可比性更好，便于跨机构或跨领域数据分析和综合。临床数据是以患者为中心的多源异构数据，其复杂性不言而喻。结构化数据预处理仅是数据治理的一环，为保障数据的可用性和价值，需建立有效的数据质量控制机制和完善的数据管理体系。

作者贡献：车贺宾负责论文撰写；何昆仑负责提供指导、论文审核；吴欢负责论文审核；陈媛媛、王万玲负责文献调研；王飞、刘立永负责实验数据整理；尚亚飞负责自然语言处理算法编写。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 莫祖英. 大数据处理流程中的数据质量影响分析 [J].

现代情报, 2017, 37 (3): 69–72, 115.

2 赵一凡, 卞良, 丛昕. 数据清洗方法研究综述 [J]. 软件导刊, 2017, 16 (12): 222–224.

3 王强, 易应萍. 临床医疗大数据治理和应用 [J]. 医学信息学杂志, 2018, 39 (8): 2–6.

4 王耀国, 李鹏, 刘迷迷, 等. 临床专病数据库建设现状与思考 [J]. 医学信息学杂志, 2024, 45 (3): 65–69.

5 张睿, 李蕾, 王丽, 等. 中文医学术语集在真实世界临床研究中的应用 [J]. 中国数字医学, 2024, 19 (8): 86–95.

6 车贺宾, 徐洪丽. 临床数据处理流程规范构建 [J]. 医学信息学杂志, 2022, 43 (3): 52–55.

7 王维笑, 费晓璐, 闫海荣, 等. 基于自然语言处理与结构化算法的病历信息高精度抽取方法研究 [J]. 中国数字医学, 2024, 19 (5): 40–48.

8 朱建平, 付学敬. 融合数据标准管理和数据血缘建设的数据治理方法 [J]. 上海质量, 2023 (11): 49–55.

9 冯晨阳, 刘迷迷, 刘强, 等. 大数据背景下医院数据质量评价模型及监控管理模式探索 [J]. 医学信息学杂志, 2022, 43 (7): 26–31.

10 王海星, 张靓, 杨志清, 等. 医疗大数据在临床科研中的应用探讨 [J]. 中国医院, 2020, 24 (7): 63–64.