临床数据处理流程规范构建*

车贺宾 徐洪丽

（中国人民解放军总医院医学大数据研究中心 北京 100853）

〔摘要〕 基于医院真实世界大数据应用实践，提出构建临床数据处理流程规范，详细阐述该规范制定流程及重点，包括临床问题确定、患者分组条件客观化处理、临床数据规范化校验、临床数据规则化提取等。

〔关键词〕 医学大数据；应用实践；临床数据处理；流程规范

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.03.010

Construction of the Clinical Data Processing Flow Specification *CHE Hebin, XU Hongli, Medical Big Data Research Center of PLA General Hospital, Beijing 100853, China*

〔**Abstract**〕 Based on the application practice of real world medical big data in hospitals, the paper proposes the construction of the clinical data processing flow specification, and elaborates the procedures and emphases of the specification, including the determination of clinical problems, objectified treatment of patients' grouping conditions, standardized validation of clinical data, regular extraction of clinical data, etc.

〔**Keywords**〕 medical big data; application practice; clinical data processing; process specification

1 引言

医疗卫生行业数据来源丰富且类型多样[1]。医疗信息平台等数据端汇聚庞大数据资源，充分挖掘医学数据价值有利于辅助临床诊断、拓展科研思路、提高科研效率、强化医院数据治理能力[2]。医学大数据挖掘利用以医学问题为先导，医疗行业业务场景较多[3]，所面临问题不同，对医学数据资源的加工处理需求也不同，导致临床研究方案设计的个性化需求较强。医学工程师应理解临床研究需要解决的问题，具备定位问题、发现问题、拆解问题和解决问题的能力[4]。具体来说需要长时间反复和临床研究者沟通以便充分理解研究过程，在此基础上分析和筛选，进行适当抽象和简化，将临床问题转化为数学问题，进而定义规则处理临床数据[5]。

2 基本概念与数据处理流程

2.1 回顾性队列研究

首先明确研究目的，确立结局指标和研究因素，提出 PICO（P：研究对象，I：干预或暴露因素，C：对照组，O：结局指标）问题[6]。临床数据处理主要包括确定患者入选标准（如性别、年龄、诊断等）、

〔修回日期〕 2021－09－13

〔作者简介〕 车贺宾，硕士，工程师，发表论文 8 篇；通讯作者：徐洪丽，工程师。

〔基金项目〕 国家重点研发计划项目“医学多源异构数据规范研究及典型标准数据集构建”（项目编号：2019YFB1404801）；国家老年疾病临床医学研究中心开放课题“基于医疗大数据的老年共病临床科研平台建设及应用”（项目编号：NCRCG－PLAGH－201905）；中国人民解放军总医院项目“医疗数据质量评价与控制系统的研发”（项目编号：2019MBD－058）。

剔除标准（如既往史不符、关键指标缺失等）、临床研究因变量（如体征、检验、检查、用药等）和结局变量（如生存状态、预后评分等）。

2.2 临床数据处理流程规范

临床数据处理需要严格规范，才能保证完整性和准确性。第一，医学大数据应用实践需经临床研究者、数据工程师、统计分析师协同合作完成，共同制定、逐步完善并严格贯彻临床研究方案。数据工程师和统计分析师的介入使得方案更加明晰，临床研究者更方便掌握研究进展和调整人力、财力和资源配置；第二，流程规范能及时发现错误，数据工程师可以及时解决疑问数据；第三，流程所涉及程序脚本可复用，大幅降低研究难度[7]。总之规范流程可以显著提高临床研究执行效率，是获得具有科学性和标准性研究结论的前提。

2.3 数据处理方案（图 1）

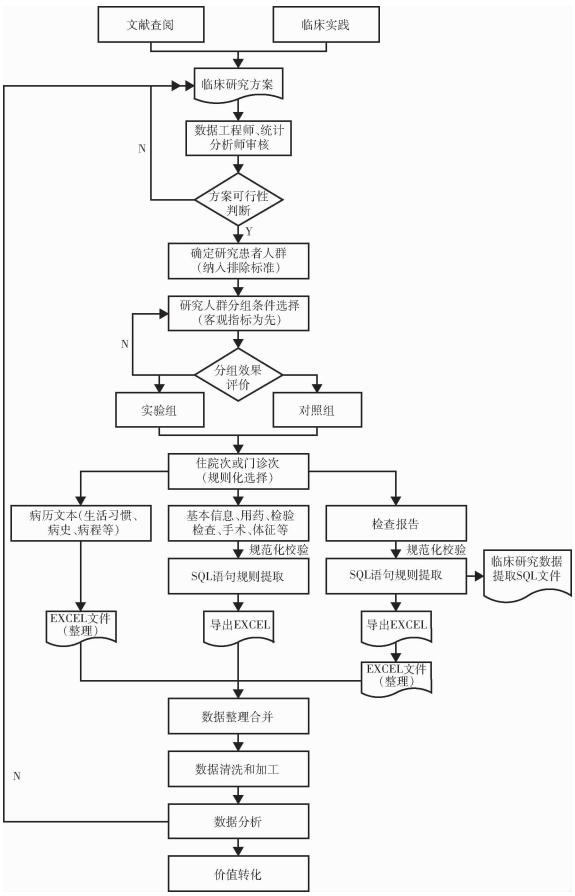

图 1 临床数据处理流程

具体包括患者纳入与排除标准、分组条件、就诊次选择以及对应具体诊疗数据筛选，包括非结构化数据（病历文本）、半结构化数据（检查报告、病理报告等）和结构化数据（病案首页、检验结果等）。临床数据一般由数据工程师利用结构化查询语言（Structured Query Language，SQL）脚本定义规则批量处理，其中非结构化和半结构数据需要自然语言处理技术配合人工整理提取具体数据项[8]，整理后的数据由统计分析师合并、清洗、统计分析并校正混杂因素等[9]，最终实现临床数据价值转化。大多数临床研究方案制定是一个长期过程，经常会因后期数据分析结果不理想被重新修正[10-12]。

3 临床问题确定与处理

3.1 临床问题的提出

临床问题的提出是医学大数据分析应用的核心，一个好的可回答的问题是保障临床研究质量的关键，有助于制定证据收集策略，提高解决临床问题的针对性[13]。要找准临床问题应具备对患者的责任心、丰富的基础和临床医学知识、扎实的临床基本技能、一定的人文科学及社会心理知识、综合分析和判断能力。统计分析师和数据工程师从方法学和工程学角度，基于大数据思想和统计方法，结合医院数据实际情况审核临床问题，为临床研究者提出建议。

3.2 慢病及危急症患者临床数据特点

一般来说，慢病相关研究涉及患者数量多、治疗周期长，一家医院包含患者临床数据的完整程度不高，许多重要指标需要随访跟踪，完成困难[14]。例如课题“恶性肿瘤患者服用化疗药导致高血压预测分析”中，影响研究结果因素较多。首先，恶性肿瘤患者离院后，很难掌握其是否遵医嘱服药、是否存在中间停服或者换药等情况；其次，患者化疗周期长、医院人流量大，持续在本院复查患者占比较低，缺少疾病发展过程中的临床数据；最后，部分患者高血压患病时间点难以判断，难以确定高血压与服用化疗药之间的关系。危急症患者治疗周期短，见效

快，患者临床数据完整度高，完成相对容易。如课题“住院急性胰腺炎患者经口进食不耐受风险因素分析”中，住院急性胰腺炎患者相对较少、治疗周期短、临床数据完整度高，并且经治疗后进食是否耐受在医生医嘱或病程记录中有所体现。

3.3 患者分组条件客观化处理

临床问题涉及指标应当尽量来源于客观数据，避免人为主观干预造成数据分析偏差[15-16]。以不良结局为例讨论如何客观化处理患者分组条件。许多临床研究方案中结局变量为死亡，但筛选结果往往不符合临床实践认识，而且本院数据无法满足临床研究对数据量的需求[17]。原因在于一方面先进的医疗技术延长了危重症患者生命；另一方面濒死患者存在转院或者自行出院返家的情况。建议采用临床不良事件发生代替死亡事件，并将医嘱处置作为参考条件，即医嘱中包含死亡、尸体、电除颤、心外按压或盐酸肾上腺素注射液 3 次以上的患者为不良结局组。

4 临床数据规范化校验

4.1 概述

不同医院业务系统和流程存在差别，导致诊疗信息数据结构和存储内容不同[18]。数据工程师需深入医院业务流程充分了解数据源头、临床数据存储结构，才能根据临床科研需求有的放矢。

4.2 相似的临床检验项目在不同医院检验倾向不同

如肌钙蛋白，有的医院倾向查全血肌钙蛋白，有的倾向查肌钙蛋白 T，有的倾向查肌钙蛋白 I。同一临床检验项目可能对应不同名称，包括简称、别名等。如谷草转氨酶（英文缩写为 AST 或 GOT），又称天门冬氨酸氨基转氨酶、门冬氨酸氨基转氨酶、天冬氨酸氨基转移酶。同一临床检验项目可能对应不同结果单位。如血气肌酐结果值的单位可能为 umol/L 或 mg/dl。

4.3 临床研究方案涉及数据项可能有多个数据源

临床研究方案涉及数据项可能有多个数据源，存储在不同数据表中。以“恶性肿瘤患者服用化疗药导致高血压预测分析”为例，判断高血压可以通过体征中血压记录，至少不同日 3 次测量的血压大于 140/90mmHg，也可以通过医嘱中用药记录，至少包含一定剂量降压药，也可直接以医生下达的诊断为依据。这需要根据实际数据情况选择不同方案。

5 临床数据规则化提取

5.1 确定研究人群具体诊疗信息

临床研究方案首先要确定患者人群，然后确定其具体诊疗信息。诊疗信息可能需要入组患者历次门诊、住院次信息或者其满足一定条件的门诊、住院次信息，根据方案实际需求而定。确定患者就诊次后，一次住院可能会有多次检验、检查，具体选择可能是在院期间第一次或最后一次，也有可能是服用某药或接受某项手术前后最近的一次[19]。

5.2 研究案例

以“恶性肿瘤患者服用化疗药导致高血压预测分析”研究为例，要求首先确定结局事件标准，然后提取患者结局事件前最近一次指标结果。恶性肿瘤患者通常一年有多次住院诊疗记录，不同住院次检验项目不同。提取患者结局事件前最近一次指标，可能会导致同一患者结局事件前最近一次血常规检验与最近一次肿瘤标志物检验时间间隔过大，影响分析结果。为避免此类情况发生，采用时间间隔范围限定。如预测 180 天内结局变量发生，检验指标 90 天内有效期。脚本提取结局变量发生前 180 天内患者所有在院相关检验指标，然后以 90 天作为窗口在 180 天范围内滑动，选定囊括最多不同类别检验项的窗口作为目标值。服用化疗药导致高血压，需要一定服药期限和剂量，需同时统计恶性肿瘤患者住院医嘱用药和门诊取药记录。具体做法为：确定研究的化疗药在数据库中具体药品名；患者一年或两年住院化疗次数和门诊化疗取药次数大于阈值；患者用药剂量大于阈值。

6 讨论

医学大数据预测评估实践研究需要多学科人员协同合作，在临床数据处理过程中根据过程结果修正临床研究方案，重新调整数据提取和分析策略。方案更迭容易导致多方参与人员协作失衡，因此临床数据处理有效开展离不开过程文档的支持。临床科研工作者负责课题临床背景、研究意义和数据内容详情撰写，详情中包括具体数据项的重要程度、具体名称、数据来源、取值范围、临床意义和提取备注（包括就诊次和检验结果、检查报告选择标准）等。工程师应负责数据抽取脚本、数据分析过程和结果输出等资料撰写。文档留痕使临床数据处理流程有据可依，不仅方便后期审核查验，而且有助于参与人员协同合作，及时发现漏洞并完善研究方案，提高工作效率[20]。

7 结语

本文在医学大数据研究中心日常临床数据服务工作实践基础上提出临床数据处理流程规范，紧扣医院信息系统常见数据处理工作，对其他类型数据涉及较少，如基因、微生物等。另外限于实际工作内容范围，流程规范未提及自然语言处理工程师和算法工程师，较少涉及病历文本结构化操作流程和数据建模分析流程。总之临床数据处理流程规范与临床研究成果关系密切，值得高度重视，本研究提出的流程规范还有欠缺，仍需进一步完善。

参考文献

1 陈敏，刘宁，肖树发，等. 医疗健康大数据应用关键问题及对策研究［J］. 中国数字医学，2016，11（8）：2－5.

2 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.

3 汪鹏，吴昊，罗阳，等. 医疗大数据应用需求分析与平台建设构想［J］. 中国医院管理，2015，35（6）：40－42.

4 罗万春，罗明奎. 医用数学建模中的问题解决及其对教学的启示［J］. 卫生职业教育，2005（4）：48－50.

5 魏杰，董珺. 谈如何将医学问题转化为数学问题［J］. 卫生职业教育，2010，28（3）：65－67.

6 王小钦. 如何利用临床资料进行回顾性队列研究［J］. 协和医学杂志，2019，10（1）：73－76.

7 谢高强，姚晨. 数据管理在临床研究中的地位和作用［J］. 北京大学学报（医学版），2010，42（6）：641－643.

8 杨帆，董晓平，廉恒丽，等. 从医院信息系统提取临床科研数据的研究［J］. 中国数字医学，2011，6（3）：12－14.

9 王雯，高培，吴晶，等. 构建基于既有健康医疗数据的研究型数据库技术规范［J］. 中国循证医学杂志，2019，19（7）：763－770.

10 朱立峰，左铭，万歆，等. 医院临床研究中的大数据应用需求与策略［J］. 中国数字医学，2015，10（12）：14－15，24.

11 殷亦超，高炬，何萍. 研究型医院的临床大数据管理应用与实践探索［J］. 中国数字医学，2019，14（2）：34－36.

12 谢滢滢. 医院临床研究中的大数据应用需求与策略探讨［J］. 科学咨询（科技·管理），2020（7）：51.

13 刘晓清，吴东. 临床流行病学和循证医学的学科建设［J］. 协和医学杂志，2019，10（4）：398－402.

14 王宗凡. 提高慢病保障水平，促进慢病健康管理［J］. 中国社会保障，2019（11）：84.

15 曾于珍，陈世耀. 临床研究结局指标选择与样本量估计［J］. 协和医学杂志，2018，9（1）：87－92.

16 Porter M E，Larsson S，Lee T H. Standardizing Patient Out－comes Measurement［J］. N Engl J Med，2016（374）：504－506.

17 陈旭，刘鹏鹤，孙毓忠，等. 面向不均衡医学数据集的疾病预测模型研究［J］. 计算机学报，2019，42（3）：596－609.

18 张震江，薛万国，冷金昌，等. 区域协同医疗共享平台整体设计方案［J］. 中国数字医学，2010，5（1）：12－15.

19 南岩东，李玉娟，张涛，等. 38 例新型冠状病毒肺炎死亡患者临床特征分析［J］. 临床军医杂志，2020，48（9）：1030－1033.

20 潘岳松. 临床研究的数据管理与质量控制［J］. 协和医学杂志，2018，9（5）：458－462.