

经方医案数据采集、处理方法与工具构建^{*}

文志华^{1,3} 张龙信¹ 刘青萍^{2,3} 肖 莉³ 夏帅帅³ 晏峻峰^{2,3}

(¹ 湖南工业大学计算机学院 株洲 412007 ² 湖南中医药大学信息科学与工程学院 长沙 410208
³ 湖南省智慧中医工程技术研究中心 长沙 410208)

〔摘要〕 目的/意义 采集和处理经方医案数据，为中医智能诊疗模型提供高质量训练数据。方法/过程 构建中医经方医案数据采集与处理框架，提出针对网页、纸质文献、电子病历、已有数据库的经方医案数据采集、清洗、相似度检测方法。结果/结论 共获得有效经方医案数据 12 499 条，为下游中医智能诊疗模型训练奠定基础。该经方医案数据采集和处理方法，可为中医药领域大数据处理提供参考。

〔关键词〕 医案；数据集；人工智能；中医药

〔中图分类号〕 R - 058 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 - 6036. 2025. 04. 006

Construction of Methods and Tools for Collection and Processing of Classical Formula Medical Case Data

WEN Zhihua^{1,3}, ZHANG Longxin¹, LIU Qingping^{2,3}, XIAO Li³, XIA Shuaishuai³, YAN Junfeng^{2,3}

¹ School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China; ² School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China; ³ AI TCM Lab Hunan, Changsha 410208, China

〔Abstract〕 **Purpose/Significance** To collect and process classical formula medical case data and to provide high - quality training data for intelligent diagnosis and theatment models in the field of traditional Chinese medicine (TCM). **Method/Process** A framework is developed for the collection and processing of classical formula medical case data in TCM. Methods are proposed for data acquisition, cleaning, and similarity detection from multiple sources, including web resources, printed literatures, electronic medical records, and existing databases. **Result/Conclusion** A total of 12 499 valid classical formula medical case records are obtained, offering a foundation to support the training of downstream TCM intelligent diagnosis treatment models. The proposed data collection and processing method can provide a reference for big data handling in the field of TCM research.

〔Keywords〕 medical case; dataset; artificial intelligence; traditional Chinese medicine

1 引言

近年来，人工智能技术发展迅速，在中医药领

域广泛应用，如中医四诊、辨证诊断、处方推荐、中药识别等方面研究^[1-5]，以及中医智能诊疗系统与设备研制^[6-8]。中医诊疗大数据是中医智能诊疗研究的重要基础，然而其建设工作没有得到足够重视，已有数据集数据量较小，缺乏高质量、开放共享的数据集，制约了人工智能技术在中医智能诊疗领域的应用与发展。中医诊疗领域拥有数千年临床医案和海量医院电子病历数据，但这些数据来源多样，且呈现非结构化、半结构化特征，只有经过采集、清洗、融合等处理后，才能满足人工智能模型

〔修回日期〕 2025 - 01 - 03
〔作者简介〕 文志华，博士，讲师，发表论文 12 篇。
〔基金项目〕 湖南省自然科学基金联合项目（项目编号：2024JJ7143）；湖南省教育厅一般科学研究项目（项目编号：23C0181）。

训练的需求。本研究针对中医智能诊疗研究所需经方医案,提出针对网页、纸质文献、电子病历及已有医案库的经方医案数据采集、数据清洗、医案分割、相似度检测方法,开发经方医案校对工具,为中医药数据集构建奠定基础并提供参考。

2 相关工作

早期中医医案数据采集与整理研究主要集中在医案数据库建设,建立了“新安医案数据库”“中医历代医案数据库”等^[9]。早期医案数据获取方式以手工整理为主,医案数据量偏小,且大部分医案数据库没有共享发布。随着人工智能技术的发展,对于数据量有了更高的要求,医案数据获取方式转向自动化采集。朱彬等^[10]针对传统医案采集效率低、形式单一、整理和分析困难等问题,提出一种多媒体医案采集移动平台。针对中医药网站信息爬取中数据结构化程度低且信息维度高的问题,章亚东等^[11]提出基于网络爬虫的名老中医医案数据采集与清理算法。在系统化采集方面,韦昌法等^[12]提出中医数字辨证配套医案智能采集与分析系统,实现医案采集、医案展示、医案检索、数据统计分析及可视化等功能;马亚楠等^[13]设计开发用于中医智能诊疗的信息采集系统,支持患者四诊数据采集并存储为医案,同时获得可用于智能诊疗算法训练与验证的数据集。针对医案数据管理与校对,中国中医科学院中医药信息研究所^[14]开发了包含名医信息管

理、临床医案导入、数据标准化、医案数据管理、医案数据分析、数据字典管理模块的中医名医医案分析系统,并迭代升级为“古今医案云平台”,在中医行业具有较大影响力。基于该平台,于琦等^[15]利用大数据和机器学习方法构建中医医案语音识别模型,支持中医临床信息实时采集。中医医案数据获取逐步发展到应用网络爬虫、光学字符识别(optical character recognition, OCR)等技术的自动采集模式,出现了一批可用于医案展示、分析、管理的系统平台。但在数据集方面,中医诊疗领域缺乏高质量、公开共享的开放数据集,也缺乏从数据采集到融合处理完整过程的方法共享。

3 数据采集与处理

3.1 经方医案数据采集与处理框架

经方医案数据采集与处理框架由数据源选定、数据采集、数据处理、系统工具 4 部分组成,见图 1。本研究将经方限定为张仲景《伤寒论》《金匮要略》中出现的方剂,经方医家限定为伤寒流派医家,古代如许叔微、陈修园、徐灵胎,近代如陆渊雷、曹颖甫,当代如胡希恕、刘渡舟等。中医历史发展特征决定了高疗效医案主要保存于纸质书籍中,也有部分医案经过整理发布于互联网,本研究数据采集对象以纸质医案数据为主,以网页和已有医案库中数据为辅。因电子病历类医案难以确定真实疗效,仅提供采集思路,未予正式采集。

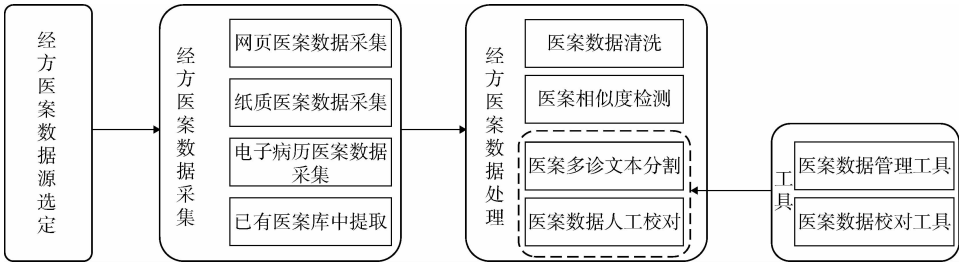


图 1 经方医案数据采集与处理框架

3.2 经方医案数据采集

3.2.1 网页经方医案数据采集 网页经方医案数据采集包含发起超文本传输协议(hypertext transfer

protocol, HTTP) 请求、获取地址列表、解析网页内容、提取医案数据、保存医案数据 5 个步骤,采用 Python 语言编写程序。在前两步利用 Requests 进行 HTTP 请求以解析并获取待访问医案的统一资源

定位符 (uniform resource locator, URL)。在第 3 步对 URL 列表循环遍历, 采用 Selenium 驱动浏览器访问网页医案内容。当目标网站具有较强反爬机制时, 多次分批采集, 将已经爬取过、访问异常的 URL 记录入库, 在后续爬取中忽略被记录的 URL。在第 4 步根据网页标签内容解析并提取所需结构化医案数据内容。在第 5 步将结构化数据保存入数据库。以“360 个人图书馆”中“经方医案”栏目为爬取目标, 共爬取 URL 记录 1 799 条, 其中爬取异常记录 23 条、爬取失败记录 150 条, 最终采集经方医案数据 1 626 条。

3.2.2 纸质经方医案数据采集 纸质书籍中的医案经过医家筛选首肯、编辑校对, 具有治疗效果突出、数据质量可靠的特点, 是经方采集的重点对象。采用百度智能 OCR 应用程序编程接口 (application programming interface, API) 进行纸质医案数据采集。采集过程分为经方医案书籍扫描、图片 PDF 书籍生成、OCR 识别、文本电子书籍生成、结构化提取、存储数据入库 6 个步骤。前两个步骤需手工处理, 后 4 个步骤通过程序算法自动化采集, 参考既有研究^[15]使用经方医案书籍体例模板优化相关算法。选取《经方实验录》《刘渡舟验案精选》《胡希恕医论医案集粹》《伤寒方历代治案》《伤寒名医验案精选》《伤寒论方医案集》《伤寒论方医案选编》《金匱方百家医案评议》《金匱名医验案精选》《国医大师经方验案精选》共 10 本经方医案纸质书籍, 采用 CZUR ET16 设备快速扫描书籍并保存为 PDF 格式。每本医案书籍体例格式有所不同, 预先提取其体例模板特征, 通过算法提取结构化数据。根据目标书籍共性特点及后续研究需要, 提取“方剂”“证型”“医案”“医案和按语”“方药”“提取的方药”“用法”“原文”“解说”“来源”“医家”共 11 种属性, 存储于 CSV 文件中。从 10 本纸质书籍中共提取经方医案数据 3 092 条。

3.2.3 电子病历经方医案数据采集 电子病历是结构化数据, 采集方式相对简单, 可通过电子病历系统开放式 API 或数据库接口直接采集, 只需解决数据库字段内容转换与合并问题。考虑到电子病历中数据质量问题以及疾病临床疗效评价的复杂性,

本研究谨慎使用电子病历数据。

3.2.4 已有医案中经方医案数据采集 以本团队历年积累的医案资源库作为补充, 扩展可用经方医案数据量。提取思路为: 建立经方方剂名称索引列表, 依据索引对库中数据进行二重遍历, 提取文本中含有经方方剂名称的医案并进行内容解析, 提取所需结构化数据, 并保存为 CSV 文件, 待后期融合处理。该医案中大部分数据为时方医案, 需从中提取经方医案。医家在治疗过程中有时会使用经方与时方相结合的合方, 因为下游智能诊疗模型目标标签中不包含时方, 所以只保留经方与合方医案数据。利用正则表达式循环遍历经方方剂名称索引列表, 采用最长匹配模式, 避免出现错误结果。在二重循环遍历全部医案数据后, 经方方名共出现 17 218 次, 得到经方医案数据 10 894 条, 剔除空值后获得经方医案数据 9 357 条。

3.3 经方医案数据处理及工具

医疗领域数据质量尤为重要, 本研究通过数据清洗保证数据的完整性、一致性; 通过相似度检测去除重复医案, 保证数据的唯一性; 通过医案分割、人工校对确保医案数据的准确性。

3.3.1 数据清洗 数据清洗是对数据进行重新审查和校验的过程, 主要包括重复值删除、缺失值处理、异常数据处理等。重复值删除即删除出现重复内容的医案。由于采集源不同, 同一则医案在出版、发布过程中可能有少许改动, 需结合相似度检测算法判定重复度, 再删除重复数据。缺失值处理是判断医案数据中是否存在空值, 空值为非法记录, 应删除。异常数据处理主要是对医案内容的合法性进行判断, 如医案内容过短, 则属于属性采集错误, 应删除。

3.3.2 相似度检测 重复数据会影响下游模型训练效果, 应剔除经方医案中的重复数据。采用在医学文本相似度分析中表现优秀的余弦相似度来计算医案文本的相似性。其主要思想是, 将医案文本分割为 n 个连续字符的片段 (n -gram 片段), 然后比较这些序列的重合度, 计算方式如下。其中 X 、 Y 表示两个医案文本语言片段集合, x_i, y_i 分别表示 X 、

Y 对应的语言片段， $\sum_{i=1}^k \cos(x_i, y_i)$ 是两个医案文本语言片段集合的交集，表示两个文本的相似性。 $x_{\text{dif}_i}, y_{\text{dif}_i}$ 分别表示 X、Y 中没有进行匹配的语言片段，m 表示没有匹配语言片段集合的大小。

$$\text{Sim}(X, Y) = \frac{\sum_{i=1}^k \cos(x_i, y_i)}{\sum_{i=1}^k \cos(x_i, y_i) + \sum_{i=1}^m (1 - \cos(x_{\text{dif}_i}, y_{\text{dif}_i}))} \quad (1)$$

医案文本相似度检测流程为：提取医案内容存入列表，循环抽取列表中文本的所有 n - grams，使用 Hash 函数映射到 0—d 之间的整数（d 为医案内容向量的维度），并使医案内容向量归一化，设 n 取值为 5（医案文本中重复出现的词汇较多，n 值不宜过小，此处取常见词汇的两倍长），文档总长度取值为 5 000（取医案文本长度最大值整数）；循环计算各分割语言片段之间的余弦相似度，将结果保存为 CSV 文件；遍历 CSV 中相似度数据，相似度大于等于 0.95、0.90、0.85、0.80、0.75、0.70、0.65、0.60 的成对医案数量统计结果，见表 1。对不同段的成对医案进行人工抽样比对，发现相似度大于等于 0.90 的医案属于不同来源的重复医案。因此，确定 0.90 为两个医案相似度阈值，剔除相似度大于等于 0.90 的医案 102 条。

表 1 经方医案数据相似度检测结果

相似度值	相似医案数量（对）
≥0.95	20
≥0.90	102
≥0.85	290
≥0.80	626
≥0.75	1 176
≥0.70	2 129
≥0.65	3 591
≥0.60	5 616

3.3.3 经方医案分割与人工校对 分析所采集的经方医案发现，近现代医案具有多诊次特点。一般而言，每诊医案对应一组诊断结果、处方用药等标签是比较合理的记录。如果多诊信息混合对应一组标签，将影响下游模型训练的准确性，因此要对多诊次医案

进行分割。此外，医家首诊医案信息记录较为详细，二诊、三诊信息记录仅提到症状减轻或加重，因此仅收集分割后的首诊经方医案数据作为下游模型训练数据。针对本研究数据处理和人工校对涉及的常用功能，开发包含医案筛选、医案分诊分割、症状、主诉、疾病、证型、方剂、中药等校对功能的软件平台。“新增分诊医案”功能将原医案按照诊次进行拆分，包含病史、症状、诊断、处方用药等信息；为方便校对症状，纳入《中医临床症状术语规范修订版》中 2 088 条症状记录，便于快速检索对应症状；为方便校对方剂及其对应中药，纳入《伤寒论》《金匱要略》《方剂学》中对应的方剂名称及其对应中药，输入部分字符可快速检索对应方剂，选中方剂后可在“方剂对应中药”框中自动呈现中药内容；软件工具具有新增、删除、修改和查询上一诊、下一诊分诊医案的功能。

4 结果与讨论

4.1 结果

通过对网页、纸质书籍、已有医案库进行经方医案数据采集，得到 14 075 条经方医案数据，经过上述清洗、分割、相似度检测等数据预处理，并使用经方医案校对工具进行人工校对，最终获得 12 449 条经方医案数据。预处理前，部分医案文本长度超过 3 000 字符，最长达 17 609 字符。预处理后，绝大部分医案文本长度在 2 600 字符以内，为正常医案记录长度。

12 449 条医案数据中共有经方方剂 358 方（包含合方），大部分经方方剂对应的医案数量在 50 个以下，少量在 150 个以上，医案数量最多的 20 个经方方剂及其医案数量，见图 2。在所有经方中，以理中汤为主方的医案数量最多（有 615 则），其次分别是小柴胡汤、白虎汤、桂枝汤、五苓散等。处方频次与经方在临床中的使用习惯比较吻合，一定程度上反映了医案数据采集的有效性。经方方剂对应医案数量占比分布，见图 3，分布中心点落在 30 ~ 50 个之间，说明绝大部分经方方剂有比较充分的样本。

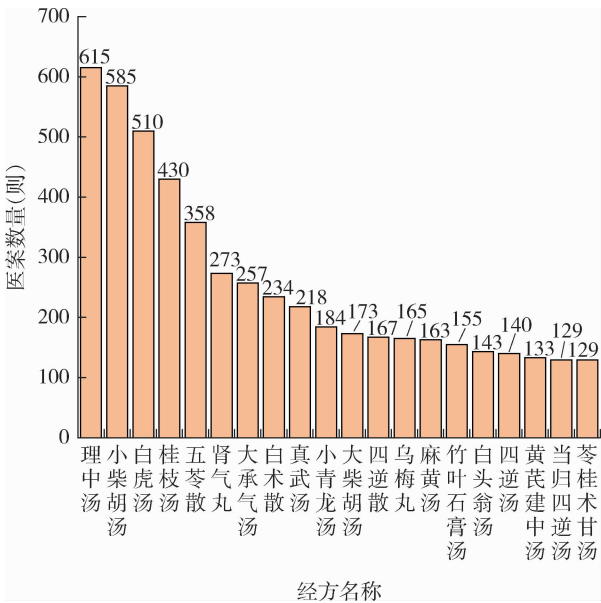


图2 对应医案数量最多的20个经方方剂

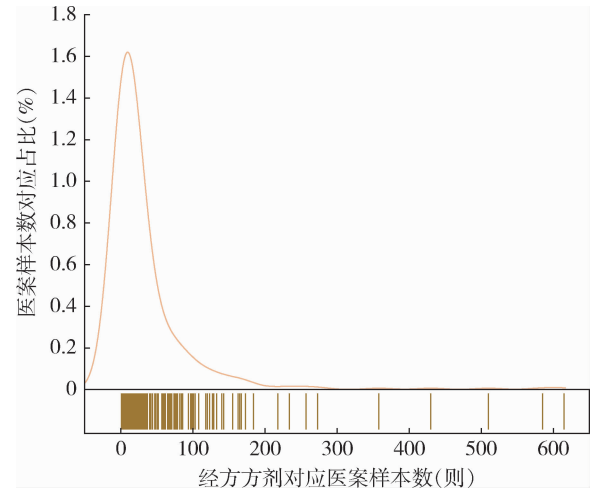


图3 经方方剂对应医案数量占比分布

4.2 讨论

4.2.1 选择高可信度数据源作为采集目标 在公开出版的中医书籍中，医案经过作者的深入整理和 后人的临床实践，具有最高的可信度，但其采集难度、成本也最高；网页公开医案数据大部分来源于 已出版书籍、文献，可信度次之，采集难度较纸质 文献低，但查询信息的时间成本较高；医院电子病 历数据多为结构化数据，采集难度最低，但因临床 疗效评价复杂，人工确认成本较高；为获得更多医 案数据，已有医案数据库也可作为数据来源，其数

据可信度要根据具体情况分别对待。

4.2.2 重视人工校对，提升数据质量 医学领域 具有特殊性，对医案准确性要求较高，低质量医案 数据将严重影响下游智能诊疗模型的训练效果。目 前医案数据质量有待提升，虽然技术手段能辅助发 现、处理部分问题，但人工校对依然是提升医案数 据质量的重要途径。

4.2.3 加强技术人员中医学专业能力 开展中医 医案数据采集，技术人员须具备一定中医学专业知 识，在数据采集与处理过程中能从容解决一些非技术 原因导致的问题，将有助于提升数据采集与处理效率。

4.2.4 仅用首诊医案数据的可行性及弊端 首诊 医案记录详细、续诊医案记录简单是近现代医案的 显著特征。首诊医案中 含有丰富的症状、体征信 息，是学习训练的良好素材。中医主要是对证治 疗，如复诊时症状减轻，可视为首诊具有有效性。 仅使用首诊医案数据而忽略后续治疗信息，无法反 映疾病在治愈过程中的时序递进关系和前后诊次的 因果关系。因此，多诊次医案的时序关系值得进一 步探索。

4.2.5 数据采集异常值分析 本研究中异常值主 要出现在纸质医案提取中，少量医案内容过短。通过 对比发现问题集中于《伤寒论方医案选编》和《金 匱方百家医案评议》。二者出版年代久远并采用铅 版印刷，部分字体存在浓淡、变形、残缺等问题， 通过模板提取容易产生错位分割，导致出现异常值。 因此，原始书籍版本选择非常重要，如果所选版本 确实存在问题，可在 OCR 识别后对文本数据模板的 提要进行初步校对，以降低异常发生概率。

5 结语

医案数据是智能诊疗模型训练的基础，本研究 提出针对纸质文献、网页、电子病历、已有医案库 的经方医案数据采集方法，共采集经方医案数据 14 075 条。为了提升数据质量，采用医案数据清 洗、相似度检测等方法，对原始数据进行预处理， 最终获得经方医案数据 12 449 条。通过医案文本长 度与医案数量、经方方剂与医案数量分析，验证本

研究预处理方法以及所获得经方医案数据的有效性, 可为下游智能诊疗模型训练提供可靠数据。

作者贡献: 文志华负责经方医案数据采集、处理与分析, 论文撰写; 张龙信负责经方医案校对工具构建; 刘青萍、肖莉、夏帅帅负责经方医案校对; 晏峻峰负责研究设计、提供指导。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

1 衣凯, 徐樱, 马云静, 等. 基于中医象思维的远程仿生复现脉诊仪的研发 [J]. 时珍国医国药, 2024, 35 (3): 747 – 750.

2 王玲, 林依凡, 李璐. 智能诊疗在舌象研究中的应用进展 [J]. 中华中医药杂志, 2021, 36 (1): 342 – 346.

3 姚帅君, 闫敬来, 杜彩凤, 等. 基于集成学习构建围绝经期综合征中医智能诊断模型 [J]. 中医杂志, 2023, 64 (6): 572 – 580.

4 DONG X, ZHENG Y, SHU Z, et al. TCMPR: TCM prescription recommendation based on subnetwork term mapping and deep learning [J]. Biomed research international, 2022 (2): 3776 – 3783.

5 JIN Y, JI W, SHI Y, et al. Meta – path guided graph attention network for explainable herb recommendation [J]. Health information science and systems, 2023, 11 (1): 5.

6 赵国桢, 郭诗琪, 庞华鑫, 等. 人工智能技术在辅助中医诊疗及诊疗标准化中的应用 [J]. 中医杂志, 2022,

63 (24): 2306 – 2310.

7 文志华, 夏帅帅, 刘东波, 等. 中医智能辨证诊断技术的演进与问题探讨 [J]. 世界科学技术 – 中医药现代化, 2021, 23 (11): 4298 – 4304.

8 孙宇衡, 王雨轩, 魏东升, 等. 基于数字孪生技术的中医智能辅助诊疗系统构建研究 [J]. 中华中医药学刊, 2024, 42 (9): 18 – 22.

9 陈玲玲, 刘诗发, 张晨. 中医医案类数据库的构建与数据处理研究 [J]. 中医文献杂志, 2014, 32 (6): 24 – 27.

10 朱彬, 朱垚, 陆明, 等. 基于 iOS 的多媒体医案采集移动平台的设计和实现 [J]. 计算机时代, 2017, 303 (9): 21 – 23.

11 章亚东, 胡孔法, 杨涛, 等. 名老中医医案数据采集与清理算法 DCCA 的研究 [J]. 计算机时代, 2018, 316 (10): 68 – 70.

12 韦昌法, 罗丽琴, 晏峻峰. 中医数字辨证配套医案智能采集与分析系统构建研究 [J]. 湖南中医药大学学报, 2020, 40 (1): 70 – 74.

13 马亚楠, 刘国华. 中医智能诊疗信息采集系统的设计与实现 [J]. 南开大学学报 (自然科学版), 2022, 55 (5): 67 – 71.

14 于琦, 王映辉, 李敬华, 等. 中医名医医案分析系统研究 [J]. 中国数字医学, 2015, 10 (3): 51 – 53.

15 于琦, 王映辉, 李宗友, 等. 基于语音识别技术的中医医案采集与应用研究 [J]. 中国数字医学, 2021, 16 (9): 80 – 83, 109.

(上接第 33 页)

2 张倩倩, 刘云, 彭远, 等. 医院中医药大数据分类分级研究与实践 [J]. 中国数字医学, 2025, 20 (2): 34 – 38.

3 路万里. 健康医疗大数据分类分级及安全策略 [J]. 网络安全技术与应用, 2024 (12): 65 – 67.

4 陈烨, 王阳, 徐亚兰, 等. 电子健康档案数据分类分级研究 [J]. 档案学研究, 2024 (3): 119 – 128.

5 许中缘, 何舒岑. 公共卫生领域大数据治理中个人信息的利用与保护 [J]. 中南大学学报 (社会科学版), 2022, 28 (3): 20 – 31.

6 毛丹, 夏天, 张诚, 等. 疾病预防控制数据治理主题设计与应用 [J]. 中国卫生信息管理杂志, 2022, 19 (2): 165 – 170, 177.

7 朱怡潇. 面向云端医疗数据流通的可验证计算研究 [D]. 西安: 西安电子科技大学, 2023.

8 李慧. 数据治理流程体系的建立与安全应用实践 [J]. 网络空间安全, 2024, 15 (4): 109 – 114.

9 林琳. 基于 DCMM 框架的医疗数据治理体系研究 [J]. 网络空间安全, 2024, 15 (4): 361 – 366.

10 祝守宇, 蔡春久. 数据治理: 工业企业数字化转型之道 [M]. 北京: 电子工业出版社, 2020.

11 唐明坤, 吴思竹, 周佳茵, 等. 医学数据共享隐私保护中基于聚类的匿名化算法关键技术研究 [J]. 医学信息学杂志, 2023, 44 (6): 65 – 71, 78.