• 医学信息研究 •

面向开放应用的生物医学大数据分类研究*

张胜发 罗 葳 马玉环 张晓宇 赵远志 周 伟

（中国医学科学院国家人口健康科学数据中心 北京 100730）

〔摘要〕 **目的/意义** 探索适用于开放应用的生物医学大数据分类方法，以提升数据管理效率并加强数据安全。**方法/过程** 采用文献研究法和焦点小组讨论法，总结归纳面向开放应用的数据分类目的、原则和维度，采用线面结合的方法构建生物医学大数据分类体系，并提出分类实施的关键步骤。**结果/结论** 明确生物医学大数据分类目标，基于数据特征、安全管理需求和共享应用 3 个维度，构建包含学科归属、重要程度、可共享性等因素的综合分类体系，提出生物医学大数据分类实施具体步骤，助力数据管理效率和安全管理水平提升，促进数据开放共享与再利用。

〔关键词〕 生物医学大数据；数据分类；开放应用；数据安全

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2025.05.006

Study on the Classification of Biomedical Big Data for Open Applications

ZHANG Shengfa, LUO Wei, MA Yuhuan, ZHANG Xiaoyu, ZHAO Yuanzhi, ZHOU Wei

National Population Health Data Center, Chinese Academy of Medical Sciences, Beijing 100730, China

〔**Abstract**〕 **Purpose/Significance** To explore a biomedical big data classification method for open applications, and to improve data management efficiency and strengthen data security. **Method/Process** Literature research and focus group discussion are used to summarize the purpose, principles and classification dimensions of data classification for open applications. A combination of linear and planar classification method is adopted to build a biomedical big data classification system, and key steps for classification implementation are proposed. **Result/Conclusion** The study proposes objectives of the classification of biomedical big data, and constructs a comprehensive classification system that includes factors such as subject affiliation, importance and shareability based on 3 dimensions: data characteristics, security management needs and sharing applications. The specific steps for the implementation of biomedical big data classification are proposed to help improve the efficiency and security of data management, and promote the open sharing and reuse of data.

〔**Keywords**〕 biomedical big data; data classification; open applications; data security

1 引言

随着全球数据资源流动和配置趋势的增强，许多发达国家重点加强了生物医学大数据的汇聚融合、开放共享和开发利用[1]。生物医学大数据广泛涉及人类生命健康相关的各个领域，类型复杂多样，不仅包括基础医学、临床诊疗、公共卫生、药

〔修回日期〕 2025－02－02

〔作者简介〕 张胜发，副研究员，发表论文 41 篇；通信作者：周伟，高级工程师。

〔基金项目〕 国家科技重大专项（项目编号：2023ZD0509702）；国家重点研发计划（项目编号：2023YFC2508801）；国家科技基础条件平台中心委托课题（项目编号：2023WT31）。

物研发、环境卫生等医疗卫生与人口健康数据，还包括基因组学、微生物学等生物学数据[2-3]。鉴于生物医学大数据种类繁多、应用场景复杂、安全管理要求严格等特点，科学合理地分类成为生物医学数据管理、开放共享及有效应用的基础支撑[4]。然而，现有的分类体系主要聚焦于业务管理，在数据共享和安全管理方面支持不足，迫切需要构建一个面向共享应用的生物医学大数据分类体系，以实现数据的安全管理、开放共享和创新应用[5-6]。

2 生物医学大数据分类研究综述

数据分类是组织利用持久标签对其数据资产进行特征描述的过程，以使这些数据资产能够得到恰当管理[7]。常见分类方法包括线分类法、面分类法和混合分类法[8]。

2.1 基于数据特征的数据分类

根据 2024 年 3 月发布的《数据安全技术 数据分类分级规则》，数据应按照“先行业领域、再业务属性”思路进行分类，其中行业领域划分多参照 2017 年发布的《国民经济行业分类》[9-10]。在行业分类的基础上，再根据业务属性、主题等进一步分类，常见分类维度包括描述对象、数据主题、数据用途等[11-16]。如《基础电信企业数据分类分级方法》采用线分类法，按业务属性将基础电信企业数据分为若干大类，再分为若干层级，每个层级分为若干子类，最终形成分类目录树[17]。《政务数据安全分类分级指南》从数据存储方式、数据所在地理位置等维度进行分类[18]。《信息技术 大数据 数据分类指南》从技术选型、业务应用和安全隐私保护等视角出发，按来源、应用场景、安全要求等维度进行数据分类[11]。

2.2 基于数据安全管理需求的数据分类

随着《数据安全法》等法规的实施，数据安全分类分级成为关注重点。2021 年发布的《网络安全标准实践指南——网络数据分类分级指引》从国家、行业、组织等视角进行分类，维度包括公民个人、公共管理等，并将数据划分为一般、重要、核心 3 个级别[12]。2025 年发布的《科学数据安全分类分级指南》提出，科学数据安全分类应由多个维度的多级分类标签构成，包括安全属性、数据形态、共享方式等，并细分安全属性为国家安全、公共利益等主题。2024 年发布的《数据安全技术 数据分类分级规则》提出，数据分类应基于业务特点和数据属性，如个人信息、商业秘密等，维度包括业务领域、责任部门等[10]。

2.3 生物医学大数据分类研究与实践

生物医学大数据涵盖广泛，涉及生物科学、医学、健康等多个学科，不同学者和机构根据自身研究视角和业务需求提出不同分类方法[13]。杨朝晖等[14]根据健康活动来源将医疗健康大数据分为临床大数据、健康大数据等。俞国培等[15]按照数据来源将医疗健康大数据分为医院医疗大数据、疾病监测大数据等。《信息安全技术 健康医疗数据安全指南》将健康医疗数据分为个人属性数据、健康状况数据等类型[16]。邬金鸣等[19]将人口健康领域个人敏感信息细化为身份标识信息、生物特征信息等类别。国家人口健康科学数据中心依据人口健康科学数据分类表和教育部学科分类表进行学科领域分类，然后基于业务管理需求按数据类型、数据来源等进行分类。

3 面向开放共享的生物医学大数据分类

3.1 研究思路

一是采用文献研究法，以“生物医学数据”“数据分类”“数据安全”“数据共享”等为关键词，在中国知网、PubMed 等数据库进行检索，并通过相关网站、电子文献进行补充检索，筛选出与生物医学大数据分类、安全管理与共享相关的法律法规、标准、政策等文献 129 篇，通过主题分析法归纳和整理其中的数据分类维度。二是采用焦点小组讨论法，在研究小组内部初步讨论并建立数据分类体系架构的基础上，邀请生物医学、数据科学、信息安全等领域的专家学者和数据管理人员进行多

轮讨论，基于对当前问题的认识和数据共享利用问题的分析，以面向开放共享的生物医学大数据分类需求为导向，提出生物医学大数据的分类目的、原则和体系。

3.2 面向开放共享的生物医学大数据分类目的

一是便于数据管理和用户检索。通过将复杂的生物医学数据按照类型、来源和用途等维度进行系统化组织，提高数据的可发现性和管理效率，使用户在数据检索过程中能够更直接和便捷地找到所需信息。二是便于开展数据安全保护。作为数据分级的基础，数据分类可以更准确地评估安全风险，为制定相应的数据保护措施提供依据。三是促进数据共享与再利用，最大化地利用数据资源，加速科研和医疗领域的创新步伐。

3.3 面向开放共享的生物医学大数据分类原则

在系统性地梳理和归纳国内外既有研究中数据分类原则的基础上，结合生物医学大数据分类目的，提出以下面向开放共享的分类原则。一是科学性原则。分类应基于生物医学大数据的多维特性及其内在逻辑联系，进行合理、系统划分。二是目的性原则。分类设计应紧密围绕数据管理、安全保护及共享应用的需求，便于用户根据具体需求快速定位和访问数据。三是唯一性原则。应确立明确无误的分类界限和定义，确保数据能够被精确归类和检索，在单一分类维度下不重叠或重复。四是安全性原则。应确保数据的安全性和隐私保护，支持数据的安全存储、访问控制和数据泄漏防护。五是扩展性原则。分类系统应具有良好的适应性和灵活性，能够容纳新出现的数据类型和业务需求，分类应能够适时调整，以适应数据管理、安全保护以及共享应用需求的变化。

3.4 面向开放应用的生物医学大数据分类体系

生物医学大数据分类应综合考虑数据特征、安全管理需求和共享应用 3 个维度，采用线面结合的方法进行分类，见图 1。

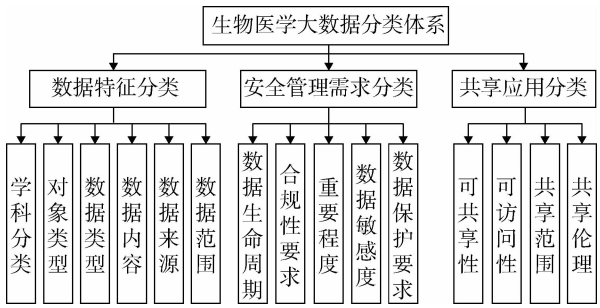

图 1 面向开放应用的生物医学大数据分类体系参考

3.4.1 基于数据特征的分类 数据特征分类是理解和组织生物医学大数据的首要步骤。(1) 学科分类。学科分类是将数据归类到相应的学科领域，以反映数据的专业特征和应用范围，如临床医学、基础医学等。参考教育部学科分类方法，先按照生物医学涉及的生物学和医学两个大类划分，随后采用线分类法对每个大类进行子类划分。对于涉及生物学和医学交叉领域的学科数据，需同时参考两个大类进行分类，并最终将数据标记为融合两个学科分类结果的综合分类。分类参考框架，见表 1。(2) 对象类型分类。数据可按照所涉及的对象特征进行有效分类，有助于确定数据的敏感性、隐私保护需求，以及数据共享和使用的限制条件。采用线分类法，将数据划分为非生物、非人类和人类 3 个类型。分类参考框架，见表 2。(3) 基于数据类型、内容、来源与范围分类。采用面分类法，基于数据的类型进行分类能够适应不同的数据获取方式、处理方法和共享应用场景。内容分类需要分析数据包含的具体信息，如是否包含个人标识信息、生物标志物等。数据来源分类旨在识别数据的原始出处，以评估数据的可靠性、可用性等。数据范围分类需要界定数据的覆盖范围，如特定人群、特定地区等，以评估数据的代表性和适用性。分类参考框架，见表 3。

表 1 基于学科分类的生物医学大数据分类参考

学科类别	一级学科	二级学科
生物学	生物学	植物学、动物学、生理学等
医学	基础医学	人体解剖与组织胚胎学、免疫学、病原生物学等
	临床医学	内科学、儿科学、老年医学等
	公共卫生与预防医学	流行病与卫生统计学、劳动卫生与环境卫生学、营养与食品卫生学等
	……	……

表 2 面向对象的生物医学大数据分类参考

一级分类	二级分类
非人类数据	实验动物数据、植物数据、微生物数据等
人类数据	身份标识信息、生物标识信息、健康医疗信息等
非生物数据	环境数据、物理数据、化学数据等

注：来源于国家人口健康科学数据中心等机构的实践总结。

表 3 基于数据类型、内容、来源与范围的生物医学大数据分类参考

分类维度	分类依据参考	分类名称
数据类型	国家人口健康科学数据中心等机构实践总结	文本数据、图像数据、视频数据、音频数据、表格数据等
数据内容	《数据安全技术 数据分类分级规则》等	个人标识信息、医疗健康信息、基因组学数据等
数据来源	《健康医疗大数据的管理与应用》等	实验室研究数据、调查普查数据、临床研究数据、临床诊疗数据、生物样本库数据等
数据范围	《科学数据安全分类分级指南》等	特定人群数据、特定地区数据、特定时间段数据等

3.4.2 基于安全管理需求的分类 该分类维度主要聚焦于数据生命周期中的安全管理要求、安全风险与安全保护。采用面分类法，从数据生命周期、合规性要求、重要程度、数据敏感度和数据保护要求等维度进行分类。数据生命周期维度，涵盖数据从创建到销毁的整个生命周期过程，包括创建、存储、处理、传输、共享和销毁等阶段。合规性要求维度，强调数据管理需严格遵守相关法律法规，如《个人信息保护法》等，以确保个人隐私和数据安全。重要程度维度，依据《数据安全法》等要求评估数据的重要程度并进行分类。数据敏感度维度，根据数据所涉及的隐私程度和保密级别等因素进行分类，以识别出需要特别关注的高敏感数据。数据保护要求维度，基于数据的敏感性和重要性等因素，设定不同的保护级别和相应措施，以确保数据得到适当的安全防护。分类维度参考框架，见表 4。

表 4 基于安全管理需求的生物医学大数据分类参考

分类维度	分类依据参考	分类名称	数据特征
数据生命周期	《数据安全法》《信息安全技术 大数据安全管理指南》	采集	指数据的初始获取阶段
		存储	涉及数据保存的方法和位置
		处理	涵盖数据的分析、清洗、转换等操作过程
		传输	指数据在不同系统或地理位置之间的移动
		共享	包括数据在不同用户或组织间的共享
		销毁	指数据的最终删除或使数据无法恢复的处置过程，以保护隐私和安全
合规性要求	《个人信息保护法》《人类遗传资源管理条例实施细则》《涉及人的生命科学和医学研究伦理审查办法》等	遵循《个人信息保护法》	数据是否涉及个人信息，如涉及，应符合《个人信息保护法》等法律法规要求
		遵循《人类遗传资源管理条例》	如数据涉及人类遗传资源信息，应符合《人类遗传资源管理条例》等法律法规要求
		符合伦理审查管理规定	如数据涉及伦理审查，应符合《涉及人的生命科学和医学研究伦理审查办法》等规定
数据重要程度	《数据安全法》《信息安全技术 重要数据识别指南（征求意见稿）》等	一般数据	指在日常业务活动中产生的，对组织运行和个人权益影响较小的数据
		重要数据	指对组织运行、个人权益或公共利益具有较大影响的数据
		核心数据	指对国家安全、经济运行、社会稳定、公共健康和安全具有重大影响的数据
数据敏感度	《信息安全技术 大数据安全管理指南（G/BT 37973－2019）》等	不敏感	不包含任何个人信息、商业秘密或其他需要保护的内容，泄漏这类数据不会对个人、组织或国家安全造成损害
		低敏感	包含少量个人信息或轻微敏感的数据，泄漏这类数据可能导致轻微的不利影响，但通常不会造成严重后果
		中敏感	包含较多个人信息或具有一定敏感度的数据，泄漏这类数据可能导致个人或组织遭受一定程度的不利影响

续表 4

分类维度	分类依据参考	分类名称	数据特征
		高敏感	包含大量个人信息、商业秘密或具有高度敏感性的数据，泄漏这类数据可能导致个人或组织遭受严重损害
		涉密	包含国家机密、商业秘密或具有极高敏感度的数据，泄漏这类数据可能导致国家安全、重大经济利益或个人安全受到严重威胁
数据保护要求	《数据安全法》《信息安全技术 网络安全等级保护 大数据基本要求》等	基础	适用于不太敏感或价值较低的数据
		增强	适用于包含个人隐私信息、商业敏感信息或对日常运营至关重要的数据
		高级	适用于对国家安全、商业秘密或个人安全至关重要的数据

3.4.3 面向共享应用的分类 该分类维度旨在针对开放共享应用场景和共享管理要点进行分类。采用面分类法，从可共享性、可访问性、共享范围和共享伦理等维度进行分类。一是可共享性，评估数据开放共享的潜力与可能性。二是可访问性，评价数据的可用性以确保其能够被用户利用。三是共享范围，考虑法律和政策影响，不同范围数据共享可能从开放到严格限制不等，要确定数据共享的边界。四是共享伦理，数据共享过程中需考虑伦理问题，以确保数据使用不侵犯个人权利或违反社会伦理标准。分类参考框架，见表 5。

表 5 面向共享应用的生物医学大数据分类参考

分类维度	分类依据参考	分类名称	数据特征
可共享性	《科学数据管理办法》等	公开共享	数据可以无限制地被任何人获取和使用
		有条件共享	数据可以在满足特定条件或协议的情况下被获取和使用
		不共享	数据由于隐私、安全或法律原因不能被共享
可访问性	《信息安全技术 健康医疗数据安全指南》等	公开访问	数据可被任何人访问，通常无需身份验证或授权
		受限访问	数据访问受到限制，可能需要身份验证或其他授权步骤
		私有访问	数据只能被特定的个人或团体访问
共享范围	《数据出境安全评估办法》等	境外	共享范围不局限于一个国家或地区，允许数据跨越国界被其他国家或地区的个人或组织获取和使用
		境内行业/机构	共享范围限定在特定国家或地区内部的特定行业或机构之间，不向公众或其他行业开放
		境内个人	共享范围限定在我国境内个人之间
共享伦理	《涉及人的生命科学和医学研究伦理审查办法》等	伦理审查通过	数据的使用已通过伦理审查委员会的审查和批准
		伦理限制	数据的使用受伦理审查的某些限制
		伦理禁止	由于伦理原因，数据的使用被明确禁止

4 面向开放应用的生物医学大数据分类实施

面向开放共享的生物医学大数据分类实施是系统性过程，从需求分析与规划开始，最终形成分类管理策略，关键实施步骤如下。第 1 步：需求分析与规划。先明确分类目标，再结合数据开放应用实际需求，规划如何通过分类策略满足上述目标，以提高数据检索和管理效率，保障数据安全，并促进数据的共享和再利用。第 2 步：梳理数据资源。全面评估现有数据资源，识别和记录所有数据资产，包括数据集、数据库、数据类型、来源等，明确数据资源的特征、内容、来源等元数据内容及实体数据内容。第 3 步：制定分类框架。从数据特征、安全管理需求和共享应用 3 个方面，结合实际需求，确定各方面的分类维度（如学科归属、数据类型、数据来源、数据风险、共享范围等），据此制定分类框架。第 4 步：制定数据分类标准。在分类框架

下，定义数据分类标准，包括分类的命名、定义和分类边界的识别规则等内容。第 5 步：实施分类标签。根据分类框架和分类标准，对数据实施分类标记，确保每个数据资源都有明确的分类标识。第 6 步：分类审核与维护。审核分类结果的科学性、准确性和相关性，并根据实际需求进行动态调整。第 7 步：形成数据分类目录。设计和建立综合性的数据分类目录，确保数据分类的逻辑性和系统性。第 8 步：制定分类管理策略。基于分类结果，制定针对性的访问控制、加密、分级管理、共享应用等措施，确保数据安全合规管理，为数据分级奠定基础，推动基于分类分级的数据共享与应用。

5 结语

本研究提出一种面向开放应用的综合性生物医学大数据分类体系，采用线面结合方法进行综合分类，主要包括数据特征、安全管理需求和共享应用 3 个方面，涵盖学科归属、数据类型、共享范围等维度。最后，本研究提出数据分类的实施步骤，为数据管理、安全保护和共享应用提供了系统的解决方案和参考。该分类体系可以提高数据检索效率、加强数据安全管理，并促进生物医学大数据的共享与再利用，为生物医学研究和临床应用提供支持。

作者贡献：张胜发负责论文撰写；罗葳、马玉环、张晓宇、赵远志负责文献分析、数据分类与统计；周伟负责研究设计、论文撰写。

利益声明：所有作者均声明无利益冲突。

参考文献

1 YANG X, HUANG K, YANG D, et al. Biomedical big data technologies, applications, and challenges for precision medicine: a review [J]. Global challenges, 2024, 8 (1): 2300163.

2 王健伟，尹岭，刘德培，等. 加强生物医学大数据建设应用，推动健康中国战略实施 [J]. 科学通报，2024，69 (9)：1123 - 1131.

3 宁康，陈挺. 生物医学大数据的现状与展望 [J]. 科学通报，2015，60 (Z1)：534 - 546.

4 张国庆，李亦学，王泽峰，等. 生物医学大数据发展的新挑战与趋势 [J]. 中国科学院院刊，2018，33 (8)：853 - 860.

5 张胜发，马玉环，张敬晨，等. 基于数据安全的健康医疗科学数据分级指南研究 [J]. 医学信息学杂志，2023，44 (8)：19 - 24.

6 王卷乐，林海，冉盈盈，等. 面向数据共享的地球系统科学数据分类探讨 [J]. 地球科学进展，2014，29 (2)：265 - 267.

7 NEWHOUSE W, SOUPPAYA M, KENT J, et al. Data classification concepts and considerations for improving data protection [EB/OL]. [2024 - 11 - 01]. https://doi.org/10.6028/NIST.IR.8496.ipd.

8 分类与编码通用术语：GB/T 10113—2003 [EB/OL]. [2024 - 09 - 01]. https://std.samr.gov.cn/gb/search/gbDetailed?id=71F772D79989D3A7E05397BE0A0AB82A.

9 国民经济行业分类：GB/T 4754—2017 [EB/OL]. [2024 - 09 - 01]. https://xw.qianzhan.com/baike/detail/bk_990d892d.html.

10 数据安全技术 数据分类分级规则：GB/T 43697—2024 [EB/OL]. [2024 - 09 - 01]. https://std.samr.gov.cn/gb/search/gbDetailed?id=14156507D2210337E06397BE0A0AE656.

11 信息技术 大数据 数据分类指南：GB/T 38667—2020 [EB/OL]. [2024 - 09 - 01]. https://std.samr.gov.cn/gb/search/gbDetailed?id=A47A713B764014ABE05397BE0A0ABB25.

12 网络安全标准实践指南—网络数据分类分级指引：TC260 - PG - 20212A [EB/OL]. [2024 - 09 - 01]. https://www.tc260.org.cn/upload/2021 - 12 - 31/1640948142376022576.pdf.

13 教育部. 关于公布 2023 年度普通高等学校本科专业备案和审批结果的通知 [EB/OL]. [2024 - 08 - 26]. https://www.gov.cn/zhengce/zhengceku/202403/content_6940137.htm.

14 杨朝晖，王心，徐香兰. 医疗健康大数据分类及问题探讨 [J]. 卫生经济研究，2019，36 (3)：29 - 31.

15 俞国培，包小源，黄新霆，等. 医疗健康大数据的种类、性质及有关问题 [J]. 医学信息学杂志，2014，35 (6)：9 - 12.

16 信息安全技术 健康医疗数据安全指南：GBT 39725—2020 [EB/OL]. [2024 - 09 - 01]. https://www.cssn.net.cn/cssn/productDetail/f44a5eabb2fc5d767467bc4f2406f7f3.

17 基础电信企业数据分类分级方法：YDT 3813—2020 [EB/OL]. [2024 - 09 - 01]. https://bbs.biaozhuns.com/thread - 308813 - 1 - 1.html.

18 政务数据安全分类分级指南：DB2201T 17—2022 [EB/OL]. [2024 - 09 - 01]. https://std.samr.gov.cn/db/search/stdDBDetailed?id=D7F324B99ECCE20EE05397BE0A0A54CC.

19 邬金鸣，钱庆，张丽鑫，等. 人口健康科学数据中个人敏感信息分类研究 [J]. 中华医学图书情报杂志，2020，29 (11)：8 - 15.