·专题报道——公共卫生信息化建设与人工智能·

基于大数据的肿瘤危险因素监测数据管理平台设计与实现

王崴，葛辉，曹煜，万明，苏雪梅

中国疾病预防控制中心，北京 102206
通信作者：苏雪梅，E-mail：suxm@chinacdc.cn

【摘　要】 为提升全链条肿瘤危险因素数据整合与共享利用效能，本研究利用大数据相关技术开展多源异构肿瘤危险因素监测数据整合与分析，在肿瘤危险因素监测业务的标准数据集指导下，构建形成包含数据生产层、数据交换层、数据计算层、数据应用层、用户访问层和数据监管层等 6 层架构的多部门资源共享肿瘤危险因素监测数据管理平台，同时提供肿瘤危险因素监测数据共享、综合查询、可视化分析等功能，从而可以较好地掌握居民肿瘤发病、患病、死亡及其危险因素的流行现状和变化趋势，使肿瘤危险因素监测业务跨部门协作成为可能，为国家制定肿瘤防控策略及相关卫生政策评估提供科学依据和有力数据支撑。

【关键词】 肿瘤；危险因素；大数据；可视化；多源异构数据；综合监测平台

Design and implementation of a tumor risk factor surveillance data management platform based on big data

WANG Wei, GE Hui, CAO Yu, WAN Ming, SU Xuemei (*Chinese Center for Disease Control and Prevention*, *Beijing* 102206, *China*)
Corresponding author: SU Xuemei, E-mail: suxm@chinacdc.cn

【Abstract】 To enhance the integration, sharing, and utilization of full-chain tumor risk factor data, this study employed big data technologies to integrate and analyze multi-source heterogeneous tumor risk factor surveillance data. Guided by the standardized datasets for tumor risk factor surveillance, a six-tiered multi- departmental resource-sharing tumor risk factor surveillance data management platform was constructed, encompassing data production, data exchange, data computation, data application, user access, and data supervision layers. This platform provides functionalities for data sharing, comprehensive querying, and visualized analysis of tumor risk factor surveillance data. Consequently, it enables a better understanding of the prevalence and trends of tumor incidence, morbidity, mortality, and associated risk factors among residents. This facilitates cross-departmental collaboration in tumor risk factor surveillance, providing scientific evidence and robust data support for national tumor prevention and control strategy formulation and related health policy evaluation.

【 Keywords】 tumor; risk factors; big data; visualization; multi-source heterogeneous data; integrated monitoring platform

恶性肿瘤是我国居民健康重要影响因素，2022 年我国恶性肿瘤新发病例 482.47 万，总死亡人数 257.42 万，全癌种标化发病率平均每年增加 1.4%[1]。如不采取有效防控，疾病负担将更加沉重，恶性肿瘤属于慢性病，同样遵循三级防控规律。肿瘤诊疗费用高昂且越到晚期治疗效果越差，国内外已逐渐形成共识预防是控制肿瘤疾病负担最具有成本效益的策略，因此亟须整合不同部门肿瘤危险因素监测数据，为其精准防控奠定数据基础[2]。然而，我国虽已建立起比较完善的慢性病危险因素监测系统、不同病种（包括肿瘤）的发病登记系统，以及全死因监测系统[3]，但从肿瘤危险因素监测角度来看，现有监测系统存在监测数据覆盖面不足、监测数据标准不完全一致、数据质量不高、同系统间数据关联分析与共享利用能力较为欠缺等问题，尚未形成全链条恶性肿瘤危险因素数据整合与共享利用的全新模式[4]。因此，本研究探索利用大数据分布式存储建立肿瘤危险因素数据管理系统，实现对不同维度、跨时空离散的肿瘤危险因素数据统一存储和管理；开展多源异构危险因素监测数据的分析，构建多部门资源共享的肿瘤危险因素监测数据综合管理平台，可以连续、系统地收集恶性肿瘤的相关信息，从而掌握居民恶性肿瘤发病、患病、死亡及

开放获取：CC BY-NC-ND 4.0　DOI：10.11847/zgggws1145238
基金项目：国家科技基础条件平台国家人口与健康科学数据共享服务平台公共卫生科学数据中心（NCMI-ZB01N-201905）
第一作者：王崴（1987—），助理研究员，硕士，研究方向：医学信息资源管理与利用。
收稿日期：2024-06-19　修回日期：2024-12-06　录用日期：2025-03-14
利益冲突：不存在　伦理审查：免除审查　出版授权：全体作者已与编辑部签署作者贡献声明及版权转让协议

其危险因素的流行现状和变化趋势，为国家制定肿瘤防控策略和措施提供科学依据，为评估相关卫生政策和肿瘤防控效果提供有力数据支撑。

1 总体架构（图 1）

多部门共享肿瘤危险因素监测数据管理平台主要是将与恶性肿瘤相关的病例、死因、环境、营养、人口监测等相关的数据，通过数据交换共享方式接入到平台肿瘤危险因素数据资源库中，同时能实现对接入数据监管、分析、清洗和有效性验证。在此基础上，根据数据标准集生成相关的主题资源库。通过分析工具包、建模工具等分析工具，平台可实现肿瘤危险监测数据管理与关联分析计算等功能，最终实现对肿瘤危险因素综合监测分析[5]。主要包括数据生产层、数据交换层、数据计算层、数据应用层、用户访问层和数据监管层等 6 层架构，见图 1。

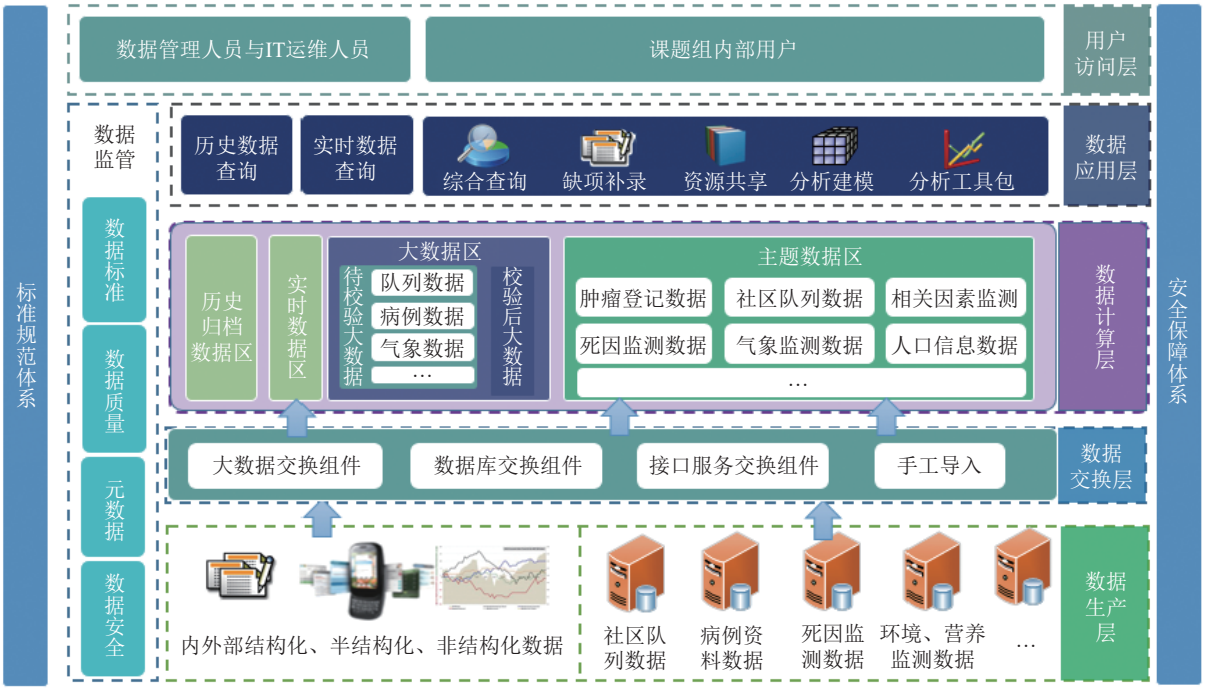


图 1 肿瘤危险因素综合监测平台总体架构图

Fig. 1 Overall framework of the integrated monitoring platform for tumor risk factors

1.1 *数据生产层* 收集与存储来源于恶性肿瘤的病例数据、死因数据、环境数据等相关业务数据。这些可利用业务数据均掌握在医院、社区、疾控中心等自行建设的业务监测系统中，包含格式化与非格式化数据。

1.2 *数据交换层* 平台的业务应用数据均来源于其他业务系统，并且接入的数据类别与业务系统的架构不尽相同，接入方式也多种多样。数据交换层将提供大数据交换组件、数据库交换组件、接口服务交换组件、手动导入等多样的数据采集组件与方式。

1.3 *数据计算层* 将接入的数据进行原始数据归档形成历史归档数据区，并根据数据标准与规则分拣数据，并初步筛选、清洗数据及数据完整性校验，形成包含肿瘤登记数据、社区队列数据、气象监测数据、人口信息数据等的主题数据区。

1.4 *数据应用层* 提供归档历史数据检索，以及实时交换的数据检索，并能够对相关主题数据资源库的数据进行综合关联查询。同时，提供大数据分析工具包，可通过多种方式选择分析所需的数据，实现数据图表展示分析，为危险因素监测与分析提供直观的数据图表分析结果。

1.5 *用户访问层* 为数据管理员与 IT 运维人员提供数据管理与监控的功能，同时业务应用人员可以自由灵活地分析与利用研究数据。

1.6 *数据监管层* 实现对数据标准对标、数据质量控制、数据安全保障，确保数据的有效性和安全性。

2 功能设计

2.1 *肿瘤危险因素数据交换与共享（图 2）* 平台通过开放相关的接口服务，实现接口注册、审核、发布监控，同时对数据的有效性校验、查重、监控等方式进行存储。平台提供标准的接口交换文档，按照标准建设数据，接口服务设计基于 XML 和 SOAP 的 Web Service 技术，能够实现 XML 到数据库、数据库到 XML 和数据库到数据库之间的数据交换和共享，完全实现了跨平台操

作、支持各种异构系统的动态数据交换功能。交换与共享服务由管理控制台、消息传输中间件、消息处理器、数据处理模块组成，实现了数据加工、数据传输、数据监控等数据交换的功能。

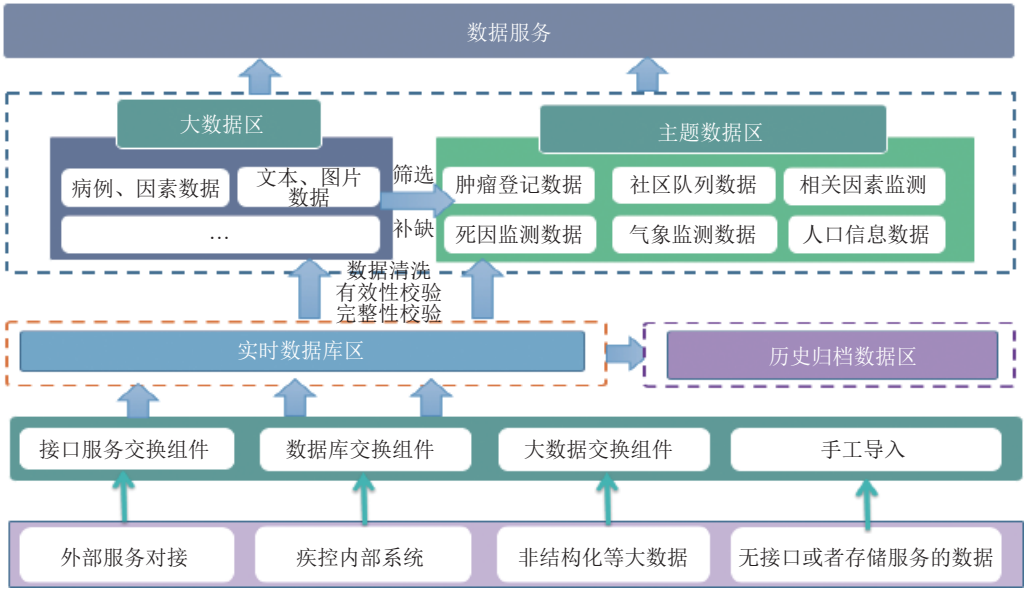


图 2　肿瘤危险因素数据交换与共享业务流程图

Fig. 2　Business process diagram of data exchange and sharing for tumor risk factors

2.1.1　管理控制台　是数据交换组件的配置、管理、监控的平台，实现了数据传输的队列配置、通过配置、路由表配置、适配器配置等功能；通过系统监控、恢复消息管理实现数据交换平台节点和消息传递的监控。

2.1.2　消息处理器　负责消息路由、消息加工、消息缓冲和消息分拣等工作，基于消息传输中间件实现消息的传递功能，并对消息进行监控。

2.1.3　消息传输中间件　负责消息的传输，消息传输中间件的配置完成后，消息处理器通过传输的调用进行数据的传输。

2.1.4　数据处理　数据交换工具对预传输的数据进行数据抽取、转换、装载的组件，实现数据的抽取、装载、转换，并统一管理数据源、监控数据处理过程。

2.2　肿瘤危险因素数据综合查询　平台可以管理各类主题数据区。通过发布数据资源地址索引，建立共享的数据资源目录，其中包含 6 大肿瘤信息库、气象信息库、水质信息库、死因、出生、三大队列信息资源等信息资源目录库。通过建立各业务数据集标准，将各业务系统中原有的数据规范化验证、清洗、整合，对不规范数据自动标记、提醒，并能够在线修订、删除数据等。各业务系统也可以通过标准数据中心申请数据借阅，最终导出标准数据。在此基础上，平台管理人员可以通过肿瘤危险因素数据管理功能，对研究分析过程中数据、文档等内容进行管理，便于后续的资料归集与查询。综合查询能够通过相关工具全库检索接入的格式化、半格式化、非格式化的数据，便于用户检索到对应的资源数据，从而导出数据资源实现线下分析与数据利用。

2.3　肿瘤危险因素数据可视化分析　平台集成基于开源的 Echarts 可视化引擎，抽取、转换、分析和模型化处理数据库中的大量格式化、半格式化、非格式化业务数据，可提取辅助决策的关键性信息。同时，将大数据技术与 *R* 统计软件集成可视化分析模块中，完成数据挖掘、分析、计算、展示等功能。平台通过对肿瘤危险因素数据进行大数据挖掘分析，产出 4 类图形，分别是动态聚集分布图、动态空间分布图、流动分析图、动态预警图等。

2.3.1　动态聚集分布图　以特殊高亮的形式显示恶性肿瘤所在的地理区域的图示，通过颜色由浅变深可以显示不同区域疾病严重情况[6]。

2.3.2　动态空间分布图　频数代表某区县内报告病例数，流量代表某个线条表示的流动方向上的病例数，某个点的值在整个区域内所处的百分位数，在整个区域内按地区所有位置上统计出的一个值。把这些值从大到小排序，某个位置的百分位数就是这个位置的值在排序里的大小，值越大地图中显示的颜色越重。它可以直观地体现出某种病情在某地区的聚集暴发性。

2.3.3　流动分析图　是一种特定类型的流程图，图中延伸的分支的宽度对应数据流量的大小的可视化分析。

2.3.4　动态预警图　可以展示出疾病预警值与当前值数量之间关系，而且还能够清楚地表明数量增减变化的情况，直观地反映疾病预警值与当前值在某段时间之间的差别。

3 系统主要功能特点与可能存在的问题

3.1 探索形成一套适用于肿瘤危险因素监测业务的标准数据集 本研究所涉及的肿瘤危险因素相关的监测数据来源广泛，数据类型多样且分散，包括病案数据、前瞻性队列研究数据、与环境相关的水质、土壤等监测数据。为应对数据融合与集成的挑战，依据《国家卫生健康委关于印发卫生健康系统解决形式主义冲突问题为基层减负措施的通知》，结合国家数据集标准以及业务实际需求，按照一数一源、一源多用、整合共享、统一口径的原则，编制了一套肿瘤危险因素监测业务的标准数据集，用于指导数据抽取、加工、整合及统一存储、使其各种来源数据标准统一与资源共享。数据集覆盖了个人基本信息数据集、体检信息数据集、肿瘤死亡信息数据集、肿瘤危险因素数据集、行为危险因素数据集等五大类数据集与值域代码，基本已经覆盖恶性肿瘤相关危险因素的全部数据集。标准数据集的编制为后续信息化数据交换、数据整合汇总、关联性分析等奠定了有力基础。

3.2 融合多源异构数据，实现肿瘤危险因素数据交换共享 在标准数据集指导下，对多源前瞻性研究队列数据融合清洗。利用大数据采集技术实现对结构化、半结构化、非结构化的海量肿瘤危险因素数据的识别、接入、传输、转换、监控、初步处理和管理等。利用大数据预处理技术实现对数据的辨析、抽取、清洗等工作。在此基础上，平台为肿瘤危险因素数据的数据交换共享和数据管理提供了有效支撑。为实现交换共享，通过多种数据接入手段，平台实现为信息化能力不同的研究机构或单位提供多种数据收集与交换手段，共计交换共享数据 4 469 814 条。为信息化相对较弱的地区提供多源数据接入与格式化后数据导入、导出功能，针对信息化能力较强或已经有信息系统支撑的地区提供数据共享交换功能。以标准数据集为基础，通过接口、服务等形式建立数据上下行共享交换通道，方便各业务系统之间数据中心的资源共享。

3.3 研发多维数据分析和可视化展现工具包 基于开源的 *R* 语言分析引擎，提供各种基础统计功能和成熟稳定的内置函数，协助平台使用者对肿瘤危险因素监测数据的分析建模，完成对大量多源危险因素监测数据的分析。基于开源的 Echarts 可视化引擎，提供坐标系、图例、提示、工具箱等基础组件，构建折线图（区域图）、柱状图（条状图）、散点图（气泡图）、饼图（环形图）、地图、力导向布局图以及和弦图，同时支持任意维度的堆积和多图表混合展现，基于地理信息系统（GIS）空间分析及可视化应用。在可视化分析方面，平台实现了数据快速分析与相关研究工具包的使用，使研究人员能够通过系统快速地调用并利用分析肿瘤监测数据，大大缩短了数据研究与统计时间，显著提升科研效率。

3.4 可能存在的问题

3.4.1 智慧化技术应用不足 肿瘤危险因素数据分析与利用是发现和了解肿瘤等慢性病在人群中的发生、发展、变化规律的重要手段和方法。但平台未充分利用人工智能等大数据分析技术构建肿瘤危险因素的预警和预测模型，不能有效实现慢性病诊治的关口前移与精准预防。后续课题组将进一步探索生成式人工智能在监测预警方面使用的可能性。

3.4.2 肿瘤危险因素自动化采集能力不足 目前，相关数据主要采用手工录入方式，后经接口服务交换组件等汇集不同业务系统的数据。当面对大量数据涌现，手动录入工作效率制约着数据监测预警的时效性，另一方面不经校验的数据又将影响精准防控的力度。后期，课题组将尝试使用更多的智能化识别方法与自动化校验策略完善录入流程。

《“健康中国 2030”规划纲要》等政策文件均指出，需完善和提升恶性肿瘤监测体系并实施恶性肿瘤综合防控战略[7]。人工智能与大数据的快速发展，不仅可以使恶性肿瘤危险因素监测的数据更全面、监测链条更健全，而且可以更加高效、更加精准分析与利用监测数据，从而有助于恶性肿瘤防控的早期筛查、政策制定和动态评估。针对我国各业务收集的肿瘤危险因素监测数据存在来源广泛、数据类型多样、数据动态更新频繁、数据环境复杂、数据描述语法和数据元数据格式不统一等特点，需要多方共同联合，推进覆盖多部门肿瘤危险因素数据资源中心的建立，创建基于大数据的肿瘤危险因素数据管理平台，使肿瘤危险因素监测业务跨部门协作成为可能，为发现各类数据的内部联系、提升数据挖掘效力、助力精准防控提供了数据支撑。

参考文献

[1] Han B F, Zheng R S, Zeng H M, et al. Cancer incidence and mortality in China, 2022[J]. Journal of the National Cancer Center, 2024, 4(1): 47-53.

[2] 王伟进, 张晓路. 中国癌症的现状与疾病负担 [J]. 中国经济报告, 2019(4): 63-73.

[3] 吴静. 我国慢性病监测系统现状及信息技术驱动发展展望 [J]. 中国数字医学, 2023, 18(3): 1-4.

[4] 秦奎. 我国心脑血管疾病监测现状与发展 [J]. 应用预防医学, 2020, 26(3): 265-268.

[5] 任佳, 宋志鹏, 孙晓晖, 等. 基于互联网的肿瘤和心脑血管病监测系统建设研究 [J]. 通讯世界, 2019, 26(8): 219-220.

[6] 唐碧玮. 我国主要恶性肿瘤危险因素知识图谱研究 [D]. 呼和浩特: 内蒙古医科大学, 2020.

[7] 新华社. 中共中央 国务院印发《“健康中国 2030”规划纲要》[J]. 中华人民共和国国务院公报, 2016(32): 5-20.