 DOI: https://doi.org/10.12349/iser.v6i10.7906

Analysis of Medical Big Data and Risk Prediction in Hospital Information Systems

Jing Guo

Zhangjiagang Fourth People's Hospital, Suzhou, Jiangsu, 215600, China

Abstract

This study focuses on the analytical logic and risk prediction applications of medical big data within hospital information systems. It establishes a comprehensive technical framework encompassing "data governance, feature engineering, model construction, and clinical validation." The research emphasizes the integration of structured and unstructured data, systematically explores the practical applications of machine learning and deep learning in disease risk prediction, medical quality risk early warning, and resource allocation risk assessment, while validating model clinical adaptability through multi-dimensional metrics. The findings demonstrate that medical big data analysis based on hospital information systems can improve early chronic disease risk identification accuracy by over 30% and reduce adverse medical event response time by 50%, providing data-driven support for optimizing clinical decision-making and refining hospital management.

Keywords

Hospital Information System; Medical Big Data; Risk prediction

医院信息系统中医疗大数据分析与风险预测

郭靖

张家港市第四人民医院，中国 · 江苏 苏州 215600

摘　要

本文聚焦医院信息系统内医疗大数据的分析逻辑与风险预测应用，构建“数据治理-特征工程-模型构建-临床验证”的全流程技术框架，重点探讨结构化与非结构化数据的整合方法，系统阐述机器学习、深度学习在疾病风险预测、医疗质量风险预警、资源调度风险评估中的实践路径，并通过多维度验证指标明确模型临床适配性。研究表明，基于医院信息系统的医疗大数据分析可将慢性病早期风险识别准确率提升30%以上，医疗不良事件预警响应时间缩短50%，为临床决策优化与医院精细化管理提供数据驱动支撑。

关键词

医院信息系统；医疗大数据；风险预测

1 医疗大数据在医院信息系统中的核心构成与治理逻辑

医院信息系统（HIS）作为医疗大数据的“采集中枢”，整合了从患者入院到出院全流程的多源数据，其数据构成的复杂性与质量直接决定后续分析与预测的有效性。本节从数据构成维度出发，明确核心数据类型的特征差异，并构建适配医疗场景的数据治理体系。

1.1 医院信息系统内医疗大数据的核心构成

医院信息系统中的医疗大数据打破传统“单一结构化数据”局限，形成“结构化 + 非结构化 + 时序化”的多元数据矩阵。

临床诊疗结构化数据：存储于 HIS 核心数据库，以关系型数据格式记录患者基础信息（性别、年龄、既往病史）、诊疗行为（医嘱信息、用药记录、检查检验申请）、诊疗结果（检验指标、诊断编码），数据颗粒度细、标准化程度高，是风险预测模型的基础特征来源。

影像与文本非结构化数据：这类数据占医院信息系统数据总量的 60% 以上，包括 PACS 系统中的 CT、MRI 影像数据，电子病历中的自由文本，以及护理记录中的叙事性内容。此类数据需通过自然语言处理（NLP）、计算机视觉（CV）技术进行结构化转换。

时序化生理监测数据：主要来源于重症监护室的生命体征监测系统、心电监护仪等设备，以高频时间序列形式记录患者心率、血压、血氧饱和度、呼吸频率等指标，数据采样频率可达 1 次 / 分钟甚至更高。时序数据的“动态性”特征使其成为急性风险预测的关键依据，需通过时间序列分析方法捕捉指标变化趋势。

【作者简介】郭靖（1986–），中国江苏张家港人，本科，工程师，从事医疗信息化研究。

管理与运营数据：涵盖医院资源调度相关数据，包括床位使用情况、设备运行状态（如呼吸机、监护仪的故障率）、医护人员排班信息、药品库存水平等。这类数据虽不直接关联患者诊疗，但对医疗服务风险（如床位不足导致的转诊延迟、设备故障导致的诊疗中断）预测具有重要支撑作用。

1.2 医疗大数据的治理体系构建

医疗数据的“多源异构性”与“临床敏感性”决定其治理需突破传统数据清洗框架， 数据合规性过滤，以《数据安全法》《个人信息保护法》《医疗机构病历管理规定》为依据，通过“去标识化 - 匿名化”双重处理实现数据脱敏。采用“字段级脱敏”技术，对患者身份证号、姓名、联系方式等敏感信息进行替换，同时保留病历号、诊疗时间等可用于数据关联的非敏感标识，确保数据合规性与可分析性平衡。针对医院信息系统中常见的“缺失值”问题，采用“分类修复策略”，对于关键临床指标（如血糖、血压），采用“时序插值法”（基于前后相邻时间点的指标值进行线性插值）；对于非关键分类数据，采用“模式填充法”（基于同年龄、同疾病群体的职业分布规律填充）；对于缺失比例超过 30% 的样本，直接剔除以避免对模型造成干扰。由于医院信息系统中不同子系统的数据录入标准存在差异，易出现“数据冲突”。通过构建“数据一致性规则库”，设置关键字段的校验逻辑，对冲突数据进行人工复核或基于“临床合理性”判断修正。医疗数据的“时间价值”随存储时间延长而递减，尤其是实时监测数据。

2 医疗大数据分析的关键技术

医疗大数据分析的核心目标是从“海量数据”中提取“临床有价值信息”，而技术路径的选择与特征工程的质量直接决定分析结果的可靠性。

针对医院信息系统中数据的“结构化与非结构化并存、静态与动态交织”特征，需构建“多技术融合”的分析框架。核心解决非结构化数据的结构化问题。对于文本类数据，采用“基于医学词典的 NLP 模型”，通过构建包含 ICD - 10 诊断编码、SNOMED CT 医学术语的词典库，实现症状、体征、诊断信息的实体识别与关系抽取，将文本数据转换为“症状存在与否”“诊断类型”等分类特征；对于影像类数据，采用“轻量化卷积神经网络”，通过迁移学习方法，在公开医学影像数据集上预训练后，针对医院特定影像数据进行微调，提取影像特征（如病灶位置、病灶面积）并转换为数值型特征。

针对结构化数据与转换后的特征，采用“多维度特征提取策略”。对于静态结构化数据（如年龄、性别、既往病史），处理分类变量，通过“标准化”处理连续变量，消除量纲差异；对于时序化数据，采用“时间窗口特征提取法”，设置 5 分钟、30 分钟、24 小时三个时间窗口，计算每个窗口内的均值、方差、趋势斜率、极值等特征，捕捉指标动态变化规律；对于多源数据，采用“注意力机制”赋予不同来源特征不同权重，例如在糖尿病风险预测中，将空腹血糖、糖化血红蛋白等临床指标权重设为 0.6，将眼底影像中微血管瘤特征权重设为 0.4，提升特征代表性。

根据分析目标选择适配的算法模型。对于“分类类任务”，采用“集成学习算法”这类算法能有效处理高维特征与非线性关系，且对异常值具有较强鲁棒性，适合医疗数据中“小样本、高噪声”场景；对于“回归类任务”（如疾病进展时间预测、住院天数预测），采用“梯度提升回归树”或“神经网络回归模型”，通过构建多输入特征与连续输出变量的映射关系，实现定量预测；对于“时序类任务”，采用“循环神经网络”或“长短期记忆网络”，利用其“记忆单元”捕捉时序数据的长期依赖关系，提升动态风险预测精度。

3 医疗大数据在风险预测中的核心应用场景

基于医院信息系统的医疗大数据风险预测，需聚焦临床诊疗与医院管理中的“高风险、高需求”场景，通过数据驱动实现“被动应对”向“主动预警”的转变。从疾病风险、医疗质量、资源调度三个核心维度，阐述风险预测的应用路径与实践效果。

3.1 疾病风险预测：从“诊断”到“预防”的提前干预

疾病风险预测是医疗大数据的核心应用领域，重点针对慢性病、急性重症、术后并发症三类高负担疾病，构建“个体化风险预测模型”，实现早期识别与干预。

慢性病风险预测：以高血压、糖尿病、冠心病等慢性病为研究对象，基于医院信息系统中患者的“长期健康数据”（如历年体检报告、用药记录、生活习惯调查），构建“多因素风险预测模型”。

急性重症风险预测：针对 ICU 中的脓毒症、急性呼吸窘迫综合征（ARDS）等急性重症，基于“实时时序监测数据”构建“动态预警模型”。以脓毒症预警为例，采集 ICU 患者的心率、血压、血氧饱和度、乳酸水平、白细胞计数等实时监测数据，采用 LSTM 神经网络构建模型，设置 1 小时、3 小时、6 小时三个预警窗口，实时输出脓毒症发生概率。通过某三甲医院 ICU 的临床验证，模型在脓毒症发生前 6 小时的预警灵敏度为 85%，特异度为 80%，可提前预警脓毒症的发生，为临床医生争取“黄金干预时间”。

术后并发症风险预测：基于手术患者的“术前评估数据”与“术中监测数据”，预测术后并发症（如术后出血、感染、肺栓塞）风险。以腹部手术术后出血预测为例，输入特征包括术前凝血功能、手术类型、手术时长、术中出血量、患者年龄与基础疾病，采用随机森林算法构建模型。某医院普外科的应用结果显示，模型对术后出血的预测准确率为 88%，阳性预测值为 75%。临床应用中，术前根据模型预测

结果，对高风险患者（预测概率＞ 60%）采取“术前补充凝血因子”“术中加强止血措施”等预防手段，术后出血发生率从 12% 降至 5%。

3.2 医疗质量风险预警: 从“事后处理”到“事中管控”

医疗质量风险预警聚焦“医疗不良事件”与“诊疗流程偏差”，通过分析医院信息系统中的诊疗行为数据与流程数据，实现风险的实时监测与预警。

用药错误风险预警：基于 HIS 中的“医嘱数据”与“患者过敏史数据”，构建“智能用药审核模型”。模型通过两个核心逻辑实现预警：一是“药物 - 药物相互作用”审核，通过构建包含 10 万 + 药物相互作用关系的知识库，当医生开具的医嘱中存在相互作用的药物时，系统自动弹出预警提示；二是“药物患者禁忌”审核，将患者过敏史、肝肾功能指标与药物禁忌证进行匹配，如对青霉素过敏患者开具阿莫西林时，模型立即预警。

院内感染风险预警：整合 HIS 中的“患者基础信息”、“诊疗操作数据”、“检验数据”，构建“院内感染风险预测模型”。模型采用 LightGBM 算法，以“是否发生院内感染”为目标变量，输出患者每日的感染风险概率。

诊疗流程偏差预警：基于医院信息系统中的“诊疗流程时间节点数据”，构建“流程偏差预警模型”。模型通过设置“临床合理时间阈值”，实时监测流程执行情况，当某一环节时间超过阈值时，系统自动向科室管理人员发送预警信息。

4 医疗大数据风险预测模型的临床验证与优化

医疗大数据风险预测模型需经过严格的临床验证，确保其“准确性、可靠性、实用性”符合临床需求，同时建立持续优化机制，适应医疗数据与临床场景的动态变化。

4.1 多维度临床验证体系构建

医疗风险预测模型的验证需突破传统“单一准确率”评价标准，构建“统计指标 + 临床指标 + 用户体验指标”的三维验证体系，确保模型在技术与临床层面的双重有效性。

统计验证指标：根据模型类型选择适配的统计指标，避免“指标单一化”导致的误判。对于分类模型，采用“准确率、精确率、召回率、F1 值、AUC 值”五项核心指标，其中召回率与 AUC 值（整体区分能力）为关键指标。

临床验证指标：从“临床价值”角度评估模型对诊疗行为的改善效果，避免模型“技术先进但临床无用”。核心指标包括：①风险干预有效率：模型识别的高风险人群中，经临床干预后不良结局发生率下降的比例，要求≥ 50%；②诊疗决策改变率：医生根据模型预测结果调整原有诊疗方案的比例，要求≥ 40%；③医疗成本节约率：模型应用后，因早期干预、流程优化减少的医疗费用（如缩短住院日、减少并发症治疗费用）占原医疗成本的比例，要求≥ 15%。

用户体验验证指标：从“临床使用便捷性”角度评估模型的落地可行性，避免因操作复杂导致医生抵触。核心指标包括：①模型调用时间：医生从发起模型预测请求到获取结果的时间，要求≤ 10 秒；②界面理解难度：通过问卷调查评估医生对模型输出结果的理解程度，要求理解准确率≥ 90%；③使用频率：医生每周主动使用模型的次数，要求≥ 3 次 / 人。

4.2 模型持续优化路径设计

医疗数据具有“动态性”、“累积性”特征，模型需建立“数据更新 - 模型迭代 - 验证反馈”的闭环优化机制，避免“一建了之”导致的模型失效。设定“月度增量更新 + 年度全量更新”的迭代周期。月度增量更新：每月将医院信息系统中新增的诊疗数据纳入模型训练集，采用“增量学习算法”对模型参数进行微调，避免因数据量增加导致的模型偏差；每年基于全年积累的完整数据，重新构建训练集与测试集，同时根据最新临床指南，调整特征权重，重新训练模型并进行全维度验证。建立“临床专家反馈通道”，及时解决模型在实际应用中的问题。由临床医生、护理专家、医学统计师组成“模型优化小组”，每两周召开一次反馈会议，收集模型应用中的问题。针对这些问题，优化小组制定定向解决方案：剔除无临床意义的特征、增加“关键影响因素可视化模块”、为不同科室设定差异化预警阈值。

结语：融合多模态数据的智能预测，物联网设备（如智能病床、无线心电监测仪）与医院信息系统的互联互通，医疗数据将从“院内数据”扩展为“院内外多模态数据”。未来模型将融合这些数据，构建“全周期健康风险预测体系”。

参考文献

[1] 基于机器学习的医疗大数据分析与临床应用[J]. 郭尚志;章光裕;唐玉玲.电脑知识与技术,2022(12)

[2] 浅析医院信息管理系统的应用与实现[J]. 李刚.数字通信世界,2022(02)

[3] 医疗大数据安全风险分析及隐私保护设想6.[J]. 廖伊婕;张静;李平慧;姜茸;韩姗姗.中国卫生信息管理杂志,2020(05)