基于电子健康档案的老年人多病共存风险预测模型构建研究

白成德
宁东能源化工基地疾病预防控制中心　宁夏　银川　750411

摘　要：本文旨在系统探讨基于EHRs构建老年人多病共存风险预测模型的理论框架、关键技术路径、面临的挑战及未来发展方向。文章首先梳理了多病共存的概念内涵、流行病学特征及其对医疗体系的冲击；其次，深入剖析了EHRs在多病共存研究中的数据价值与局限性；进而，详细阐述了从数据预处理、特征工程到模型选择与验证的完整建模流程，并对比分析了逻辑回归、机器学习及深度学习等不同算法的适用场景与优劣；最后，针对数据质量、模型可解释性、伦理隐私及临床转化等关键瓶颈，提出了相应的对策建议。研究表明，融合先进人工智能技术与高质量EHRs数据的风险预测模型，有望实现对老年人多病共存风险的早期识别与动态预警，为优化资源配置、实施精准预防和推动主动健康管理提供科学决策支持。

关键词：电子健康档案；多病共存；风险预测模型；老年人；机器学习；健康管理

引言

21世纪以来，全球人口老龄化加速演进，中国已步入中度老龄化社会，60岁及以上人口达2.97亿。伴随老龄化而来的是慢性非传染性疾病高发，多病共存（即同时罹患两种及以上慢性病）在老年人群中已成为常态，不仅显著降低生活质量、增加失能失智与死亡风险，更导致医疗方案复杂化、药物相互作用增多、住院率上升，并暴露出传统单病诊疗模式的碎片化与低效性。在此背景下，推动医疗服务从“疾病治疗”向“健康管理”转型，亟需构建能精准预测多病共存风险的模型，以实现早期识别与主动干预。近年来广泛普及的电子健康档案（EHRs）系统性记录了个体全生命周期的多维健康数据，为大规模、高精度风险预测提供了理想平台[1]。然而，EHRs数据存在噪声、缺失与异构性，加之多病共存定义复杂、特征高维稀疏、模型可解释性与伦理合规等挑战，制约了其临床应用。本文聚焦该议题，旨在系统梳理理论基础、探讨关键技术路径、分析现实瓶颈并提出对策，为推动老年健康服务智能化与精准化提供参考。

1 文献综述

1.1 多病共存的概念界定与研究现状

多病共存指个体同时有两种及以上慢性疾病，但“慢性疾病”范畴和“同时存在”时间窗口无统一标准，导致研究可比性差。流行病学研究证实，其在老年人群中普遍，患病率达50%-80%，疾病组合呈聚类模式，反映共同病理生理机制或社会因素。研究集中于描述流行病学特征、探究影响因素、评估健康结局影响，前瞻性预测多病共存风险的研究较少，且多依赖传统统计模型，预测效能待提升。

1.2 电子健康档案在健康预测研究中的应用

EHRs超越单纯临床记录功能，利用其进行疾病风险预测成研究热点。早期集中于单一疾病预测，后探索复杂健康状态，在多病共存领域崭露头角。但初步尝试未充分利用时序信息和细粒度数据。

1.3 风险预测模型的发展：从传统统计到人工智能

风险预测模型构建方法从传统统计学演进到人工智能。传统模型优势明显，但难捕捉复杂非线性交互效应，易过拟合。机器学习模型预测精度高，但“黑箱”特性限制临床应用。深度学习模型为处理EHRs开辟新路径，但数据量和计算资源需求高，可解释性问题突出，平衡性能与可解释性是前沿课题。

2 基于 EHRs 的多病共存风险预测模型构建框架

构建一个有效的风险预测模型是一个系统工程，需要严谨的方法论指导。本文提出的框架主要包括以下五个核心环节。

2.1 研究设计与数据准备

首先，研究应基于一个大型、具有代表性的老年人群队列，理想情况下，该队列应来源于覆盖广泛地域和不同级别医疗机构的EHRs数据库，以确保研究结论的外部效度。研究起点（基线）通常设定为个体首次进入数据库并满足年龄条件（如 ≥ 60岁）的时间点，而随访期则应足够长（如3-5年），以便充分观察多病共存这一相对缓慢的健康事件的发生过程。其次，结局变量的定义是整个研究的基石。鉴于多病共存概念的复杂性，本文建议采用一种动态、累积的定义方式，例如将“多病共

存”定义为在随访期内，累计被诊断出属于预设慢性疾病清单（如包含20-30种常见老年慢性病）中的至少两种疾病，并要求疾病诊断基于可靠的ICD编码，且设定合理的诊断频次或确认规则（如至少两次独立就诊记录）以减少误诊噪音[2]。最后，暴露变量（即特征）的提取应充分利用EHRs的多维性，涵盖人口学特征、既往病史、用药史、实验室与检查指标、生命体征与健康行为以及医疗服务利用等多个层面，力求构建一个全面反映个体健康状况的特征集。

2.2 数据预处理与特征工程

原始EHRs数据质量参差不齐，必须经过严格的清洗和转换才能用于建模。数据清洗是首要步骤，需要处理普遍存在的缺失值、异常值、重复记录和逻辑错误。对于缺失值，可根据其缺失机制选择多重插补、KNN填充或直接剔除等策略；对于异常值，则需结合临床知识和统计方法进行识别与修正。在数据清洗的基础上，特征工程是提升模型性能的关键。对于类别变量（如诊断代码），常用One-Hot编码将其转化为数值形式，或采用嵌入技术将其映射到低维稠密向量空间以保留语义信息。对于时序变量（如多次检验结果），不能简单地取均值，而应将其视为时间序列，计算均值、方差、最大/最小值、变化率等统计特征，或者直接保留其原始序列形态以供时序模型使用。用药史则可通过计算各类药物的使用天数比例（PDC）或简单计数来量化。面对成千上万的初始特征，必须进行降维以避免“维度灾难”和过拟合。特征选择可采用过滤法、包装法或嵌入法，筛选出最具预测力的特征子集，这不仅能提高模型效率和泛化能力，也有助于后续的模型解释。

2.3 模型选择与训练

模型的选择应基于研究的具体目标、数据特性和对可解释性的要求。为了建立一个性能基准，通常会先构建一个多变量逻辑回归模型，该模型可纳入经过临床专家筛选的核心变量，其结果易于理解和沟通。在此基础上，可以引入更强大的机器学习模型，如XGBoost或LightGBM。这些集成学习模型能自动进行特征组合和非线性拟合，通常能取得远优于传统模型的预测精度。对于希望深度挖掘EHRs时序信息的研究，可以构建基于LSTM或Transformer的深度学习模型[3]。这类模型将按时间排序的就诊事件序列（每个事件包含诊断、用药、检验等信息的嵌入向量）作为输入，通过其内部的记忆单元或注意力机制，学习个体独特的健康轨迹演化规律，并在序列末尾输出未来发生多病共存的概率。无论选择何种模型，训练过程都必须严谨，应采用交叉验证（如k折交叉验证）来评估模型的稳定性和防止过拟合，确保模型性能评估的客观性。

2.4 模型评估与验证

对模型的评估绝不能仅看单一指标，而需进行多维度、全方位的审视。区分度是衡量模型区分高风险与低风险个体能力的核心指标，常用AUC（曲线下面积）、敏感性、特异性等来量化。然而，一个仅有高区分度的模型可能是“虚胖”的，因此校准度同样至关重要，它衡量的是模型预测概率与实际发生概率的一致性，可通过校准曲线和Hosmer-Lemeshow检验进行直观和定量的评估。一个理想的模型应在区分度和校准度上均表现良好。此外，从临床决策的角度出发，还需评估模型的实用性。决策曲线分析（DCA）通过量化模型在不同风险阈值下带来的净收益，能够判断该模型是否真正具有临床应用价值，而非仅仅停留在统计学意义层面。最后，也是最关键的一步，是在另一个独立的EHRs数据集上进行外部验证。这是检验模型泛化能力的金标准，只有通过了外部验证的模型，才有可能在未来的真实世界场景中可靠地发挥作用。

3 挑战、对策与未来展望

尽管前景广阔，但基于EHRs构建多病共存风险预测模型仍面临严峻挑战。

3.1 数据层面的挑战

EHRs数据固有的质量问题，如缺失、录入错误和编码不规范，严重威胁模型可靠性。应对之策需双管齐下：在管理层面强化源头数据治理，建立统一的录入与质控规范；在技术层面开发鲁棒的缺失值填补与异常检测算法。更深层挑战在于系统异构性——不同机构采用的EHRs平台、数据标准和术语体系差异巨大，阻碍数据整合与共享。根本解决路径是推动全国乃至全球范围内的标准化建设，如采纳FHIR等互操作性规范。此外，EHRs主要覆盖就医人群，存在选择偏倚，难以代表健康老人或居家失能等脆弱群体。为此，未来研究应融合社区健康档案、智能手机及可穿戴设备等多源实时数据，构建更全面、无偏的老年人健康画像，提升预测模型的代表性与公平性。

3.2 模型层面的挑战

模型的“黑箱”特性是阻碍其临床采纳的核心障碍。临床医生和患者难以信任一个无法解释其决策逻辑的模型，尤其是在涉及重大健康决策时。因此，发展可解释人工智能（XAI）技术已成为当务之急。通过SHAP、LIME等方法，可以为模型的每一个预测结果提供直观、局部的解释，阐明是哪些关键特征及其组合导

致了高风险预测，从而增强模型的透明度和可信度。另一个常被忽视但至关重要的问题是，大多数预测模型只能发现变量间的相关性，而非因果关系[4]。这意味着模型可能识别出与多病共存相关的标记物，但这些标记物本身未必是可干预的靶点。将因果推断方法（如倾向得分匹配、工具变量法）融入预测框架，有助于从纷繁的相关性中剥离出真正的因果驱动因素，从而为精准干预提供更有价值的洞见。

3.3 伦理与临床转化挑战

EHRs包含高度敏感的个人健康信息，其使用必须置于严格的伦理和法律框架之下。隐私保护是不可逾越的红线。必须在模型开发和部署的全过程中贯彻“隐私设计”原则，积极采用联邦学习、差分隐私等前沿技术，在不共享原始数据的前提下实现多方协同建模，从根本上保障数据安全。同时，算法公平性问题不容忽视。如果训练数据本身存在对特定人群（如低收入、少数族裔）的系统性偏差，模型很可能会继承甚至放大这种偏见，导致不公平的预测结果。因此，需在模型评估中加入公平性指标，并采取针对性的去偏技术，确保模型对所有人群都公正有效。最终，模型的价值必须在临床实践中得到检验。这要求我们将预测工具无缝嵌入到医生的日常工作流中，例如作为EHRs系统的一个智能插件，在医生接诊时自动弹出风险预警。更重要的是，需要在真实世界中开展严格的实效性研究，系统评估该工具对临床决策过程、患者依从性、健康结局以及医疗成本的实际影响，这是实现从科研成果到临床价值转化的最后一公里。

3.4 未来展望

未来的多病共存风险预测模型将朝着更智能、更整合、更个性化的方向发展。一方面，模型将不再局限于EHRs数据，而是融合多模态数据，包括基因组学、蛋白质组学、环境暴露数据和来自物联网设备的实时生理行为数据，构建一个动态更新的、逼近真实的“数字孪生”健康画像。另一方面，预测模型将不再是孤立的预警工具，而是与个性化干预推荐系统、远程监测平台和患者自我管理APP紧密结合，形成一个“预测-预警-干预-评估”的闭环健康管理生态系统。在这个生态中，技术不再是冰冷的算法，而是赋能医患双方、共同守护健康的智慧伙伴，真正实现从被动治疗到主动健康的伟大跨越。

4 结语

多病共存是老龄化社会的核心健康挑战，精准预测其风险对实现主动健康管理与优化资源配置至关重要。电子健康档案（EHRs）凭借全面、动态、海量的数据优势，为构建预测模型提供了坚实基础。本文系统探讨了从数据准备、特征工程到模型选择与评估的完整建模路径。然而，模型构建仅是起点，数据质量、可解释性、隐私伦理及临床转化等深层次问题才是决定其能否落地应用的关键。未来需推动临床医学、流行病学、人工智能与伦理学等多学科协同，通过完善数据治理、发展可解释AI、深化临床整合，使该技术真正赋能老年健康，助力“健康中国2030”战略实施。

参考文献

[1]刘洋,于洪臣,李金津.智慧医院视域下电子健康档案模式构建对老年慢性病管理应用研究[J].山东档案,2024,(05):53-55.

[2]李秀华,黄伟彬.基于居民电子健康档案数据的分析与应用[J].中国农村卫生,2024,16(12):33-35.

[3]林巾粲,郝晓宁.健康档案建立在老年流动人口社会支持与健康相关生命质量间中介效应[J].中国公共卫生,2023,39(08):958-964.

[4]孟祥仪.基于电子健康档案的社区卫生信息服务探索[D].上海师范大学,2021.