基于大数据分析的卵巢癌生物标志物筛选与预后评估模型的构建与验证

杨 璐 乔 娟(通讯作者)

北京大学肿瘤医院内蒙古医院 内蒙古 010000

【摘 要】：卵巢癌作为妇女健康的重大威胁之一，其早期诊断和预后评估具有极其重要的临床意义。本研究旨在利用大数据分析技术筛选卵巢癌的生物标志物，并构建相应的预后评估模型。研究通过收集和分析来自多个数据库的大规模卵巢癌患者基因表达数据，运用机器学习和统计分析方法来识别与卵巢癌发病和预后相关的关键生物标志物。基于筛选结果，我们构建了一个预后评估模型，并在独立患者队列中进行了验证。验证结果显示，该模型具有较高的准确性和稳定性，能够有效预测卵巢癌患者的生存率和疗效反应。此研究不仅为卵巢癌的早期诊断和个体化治疗提供了新的生物标志物，还为临床预后评估提供了可靠的工具，具有重要的临床应用价值。研究意义：该研究提供了一种基于大数据的方法学框架，为卵巢癌的早期诊断和个体化治疗提供强有力的科学依据，推动了精准医疗在妇科肿瘤领域的发展。

【关键词】：卵巢癌；生物标志物；大数据分析；预后评估模型；精准医疗

DOI:10.12417/2982-3676.25.03.016

引言

卵巢癌是全球范围内女性健康面临的主要威胁之一，其诊断和治疗的难度在于疾病通常在晚期阶段才被发现。根据世界卫生组织的数据，卵巢癌的早期发现和治疗成功率高于晚期。因此，早期诊断和有效的预后评估对于提高患者的生存率具有至关重要的意义。近年来，随着大数据和机器学习技术的飞速发展，这些技术在医疗领域尤其是在肿瘤生物标志物的发现和疾病预测模型的构建中展现出巨大的潜力。本研究通过集成分析多个数据库中的大规模卵巢癌患者基因表达数据，运用先进的数据分析技术，旨在筛选出卵巢癌的关键生物标志物，并构建一个准确性和稳定性都较高的预后评估模型。通过对模型在实际临床样本中的验证，本研究力求为卵巢癌患者提供更为精准的早期诊断与个体化治疗方案，推动精准医疗在妇科肿瘤领域的进一步发展。

1 概述卵巢癌的临床挑战与大数据的应用

1.1 卵巢癌的概况与临床影响

卵巢癌是妇科领域内最为严重的恶性肿瘤之一，其发病率和死亡率在女性生殖系统肿瘤中均居高不下[1]。由于早期症状不明显，卵巢癌常在晚期才被确诊，预后较差，患者五年生存率较低，严重威胁女性健康与生命安全。这种疾病的复杂性与异质性使得临床诊断与治疗充满挑战，传统诊断方法难以提供足够的准确性与敏感性。卵巢癌的病理机制尚未得到全面理解，新型生物标志物的发现以及个体化治疗方案的制定变得尤为重要。在此背景下，大数据的应用为卵巢癌的诊断及预后评估提供了新的思路，通过整合多维度数据，可以更全面地捕捉疾病生物学特征，以提高早期诊断的准确性和治疗效果。

1.2 大数据在医学研究中的作用

随着信息技术和数据采集能力的飞速发展，大数据已成为现代医学研究的核心驱动力之一。在医学领域，大数据能够整合多种来源的海量信息，如基因组数据、蛋白质组数据、电子病历和影像数据，为疾病的研究和临床应用提供深度支持。基于大数据的分析方法能够揭示疾病的复杂生物学特征，挖掘潜在的诊断和治疗靶点，并显著提升对疾病发展的预测能力。特别是在癌症研究中，大数据技术为患者个体化诊治提供了可能，其精准性和广泛适用性推动了医学领域的深刻变革。这种数据驱动型研究方式对卵巢癌等具有高异质性和复杂病理特征的疾病具有重要意义，可以有效提升相关研究的深度与广度。

1.3 生物标志物的重要性

生物标志物在肿瘤诊断、治疗和预后评估中具有重要意义。通过生物标志物可以实现疾病的早期检测，提高诊断准确性，并为制定个性化治疗方案提供参考依据[2]。在卵巢癌研究中，生物标志物有助于识别高危患者群体，监测治疗效果以及预测疾病进展，为精准医疗提供了关键支持，显著提升了临床干预的效果和效率。

作者简介：杨璐，出生年：1992 年，性别：女，民族：汉，籍贯：内蒙古巴彦淖尔市，单位：北京大学肿瘤医院内蒙古医院，职称：主治医师，学位：硕士研究生，主要研究方向：妇科恶性肿瘤。

通讯作者：乔娟。

2 数据收集与预处理

2.1 数据源的选择与收集策略

卵巢癌患者基因表达数据的收集对于后续生物标志物筛选和预后评估模型的构建至关重要。以开放性医疗数据库和癌症基因组数据库为主要数据来源，重点关注具有高质量标注和大样本量的公开数据集。选择数据源时，严格遵循以下原则：患者信息和基因组数据的完整性、数据集的代表性以及数据的更新频率。特定数据筛选标准包括患者的年龄、肿瘤分期和治疗方案等相关临床信息，以确保数据的异质性和统计学意义。在数据收集过程中，多中心数据库中的基因表达数据得以整合，结合既往相关文献，筛选出符合研究目标的优质数据。所有数据在收集阶段确保来源合法性并经过标准化预处理，为后续分析提供可靠基础。

2.2 数据清洗和预处理方法

数据清洗和预处理是保证分析质量的关键环节。为提高数据的准确性和完整性，采用去噪处理以剔除不必要的背景信号，并通过缺失值填补算法解决数据缺失问题。异常值通过统计方法和机器学习算法进行识别和处理，以减少对分析结果的干扰。数据标准化技术被用来消除不同批次间的数据偏差，使数据能够在统一尺度上进行比较。为进一步提升分析准确性，采用基因表达水平归一化方法，对不同样本间的表达量进行校正。所有处理环节均严格按照既定标准操作，以确保数据质量达到分析要求[3]。

2.3 确保数据质量与完整性

为了确保数据质量与完整性，采用严格的数据质量控制标准，包括丢失数据填补、异常值检测和重复数据去除等措施。使用多种统计方法评估数据的偏倚和一致性，以减少因技术误差或样本异质性带来的影响[4]。对所有基因表达数据进行归一化处理，以确保数据间的可比性。通过多次数据校验和交叉验证确认数据的完整性与可靠性，从而为后续分析提供坚实的数据基础。

3 生物标志物的筛选与模型构建

3.1 生物标志物的筛选方法

生物标志物的筛选是构建卵巢癌预后评估模型的关键环节。通过整合多种筛选技术，全面分析卵巢癌患者的基因表达数据。利用差异表达分析方法识别在正常组织与癌症组织间显著变化的基因。随后，采用加权基因共表达网络分析（WGCNA）构建基因模块，与临床特征进行相关性分析，以确定潜在生物标志物。通过机器学习技术如 lasso 回归、支持向量机（SVM）、随机森林等进一步筛选关键基因，以增强对重要标志物的识别能力。筛选过程中辅以功能注释与通路富集分析，验证标志物的生物学意义及其在卵巢癌中的潜在机制。经上述方法筛选出的生物标志物为后续预后评估模型的构建提供了坚实的基础。

3.2 预后评估模型的构建技术

在预后评估模型的构建过程中，采用了多种机器学习算法和统计技术，以确保模型预测的高效性与准确性。利用线性回归、Cox 比例风险模型等经典方法初步评估生物标志物与患者预后之间的关系[5]。随后引入随机森林、支持向量机和深度学习等先进算法，通过特征筛选与权重赋值优化预测性能。在模型训练中采用交叉验证技术，以降低过拟合风险并提高模型的泛化能力。通过贝叶斯优化及梯度提升方法进一步优化参数，实现模型对生存率和疗效反应的精准预测。

3.3 模型优化与特性选择

模型优化与特性选择是提升预后评估模型性能的重要环节。通过采用正则化方法，如 LASSO 回归，可有效去除冗余变量，确保模型的简洁性与稳定性。特性选择过程中，利用主成分分析（PCA）降低数据维度，从而保留与预后相关的关键特征。结合交叉验证评估模型性能，以确保选择的特性具有较高的预测能力。为进一步优化模型，应用网格搜索调整超参数，在不同配置下测试模型的准确性与鲁棒性。通过上述技术，最终获得了一个具有良好泛化能力和预测性能的模型，为临床应用提供了可靠的支撑。

4 预后评估模型的验证与应用

4.1 验证模型的统计方法

验证模型的统计方法是确保其可靠性和稳定性的重要环节。研究采用了多种统计分析方法，对预后评估模型的预测能力和适用性进行了全面评估。利用受试者工作特征曲线（ROC 曲线）计算模型的曲线下面积（AUC），以量化模型对卵巢癌患者生存率预测的准确性。为了进一步验证模型的稳健性，进行了交叉验证分析，通过不同的数据分割方式，评估模型在多个数据集上的表现一致性。通过卡方检验和 Cox 回归分析方法，评估模型所筛选出的生物标志物与临床特征之间的关系，以确保标志物的显著性与预测能力。还应用决策曲线分析评估模型在治疗决策中的临床获益，为模型的实际应用提供理论依据。各项验证分析结果表明，该模型具有良好的预测性能和临床适用性。

4.2 独立患者数据集的验证结果

在独立患者数据集验证环节，选取多个公开数据库中的卵巢癌患者数据对预后评估模型进行性能检测。通过将模型预测的生存率与实际临床结局进行对比，计算其准确性、灵敏度、特异性及 AUC 值等指标，并分析模型在不同临床亚组中的适用性。验证结果显示，该模型能够准确预测患者的生存时间及治疗反应，且在独立队列中稳定性较高，表现出对不同亚型卵巢癌患者具有广泛适用性。模型的高预测性能证明了所选生物标志物的可靠性，为临床预后评估提供了重要支持，有助于个体化治疗策略的制定。

4.3 预后评估模型在临床实践中的应用

预后评估模型在临床实践中的应用主要体现在辅助临床决策、优化治疗方案和指导患者管理等方面。通过将模型嵌入医疗系统，医生可以利用模型预测患者生存率和治疗反应，从而制定更加精准的治疗计划。模型的高准确性和稳定性为个体化医疗提供了科学依据，显著提升了卵巢癌患者的治疗效果和预后管理水平，对于改善患者预后具有重要意义，有助于推动精准医疗在妇科肿瘤领域的进一步发展。

5 结语

此项研究通过对卵巢癌患者基因表达数据的综合分析，成功筛选出关键生物标志物，并基于此构建了一个具有高准确性和稳定性的预后评估模型。该模型已在独立患者队列中进行验证，验证结果显著，有效提升了对卵巢癌患者生存率和疗效反应的预测能力。不仅如此，本研究提出的基于大数据的方法学框架为卵巢癌的早期诊断和个体化治疗提供了新角度和科学依据，极大推动了精准医疗在妇科肿瘤领域的进展。尽管研究成果令人鼓舞，但在未来研究中仍需解决一些局限性和挑战。例如，模型的泛化能力和在不同人群中的适用性还需进一步验证。同时，生物标志物的动态变化与临床路径的关联性也需要更深入的研究。进一步的研究应聚焦于扩大样本量，丰富多样性，并探索更多潜在的生物标志物，以增强模型的实用性和准确性。总之，本研究不仅在理论上拓展了卵巢癌生物标志物的研究领域，而且在实际应用上为患者的个体化治疗方案提供了科学依据，期待在未来的临床实践中发挥更大的作用。

参考文献：

[1] 刘真,及霄杨,王秀,吴晶晶,王秀丽.筛选重要生物标志物用于卵巢癌早期诊断和预后评估[J].南昌大学学报:医学版,2021,61(06):51-54.

[2] 万喻婷,王治,洪莉.基于 WGCNA 探索与卵巢癌预后和免疫浸润相关的生物标志物[J].中国计划生育和妇产科,2023,15(07):52-56.

[3] 龚姗,白波,李苗,付静静,姜丽,金海红.浆液性卵巢癌预后生物标志物的筛选及免疫细胞浸润的分析[J].中国性科学,2022,31(08):68-73.

[4] 李耀威李力.外泌体生物标志物在卵巢癌诊断及预后评估方面的价值--系统评价[J].国际妇产科学杂志,2020,47(04):462-468.

[5] 谢美强,张丹,颜劲,巫源博.卵巢癌化疗相关的耐药标志物分析[J].北方药学,2020,17(06):190-191.