Advances in Clinical Medicine 临床医学进展, 2024, 14(3), 894-899
Published Online March 2024 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2024.143788

Hans汉斯

机器学习在胃癌的分子特征筛选及预后模型建立应用的回顾

宝　仰[1]，和红阳[2]，管云飞[2]，张　阳[2]，方俊伟[2]

[1]大理大学临床医学院，云南 大理
[2]大理大学第一附属医院普外一科，云南 大理

收稿日期：2024年2月19日；录用日期：2024年3月12日；发布日期：2024年3月20日

摘　要

胃癌在我国恶性肿瘤的死亡率中位列第三，严重危害我们的生命和健康，目前治疗方式多以手术治疗为主，但许多患者发现时已经是晚期，基本无手术机会。微卫星不稳定是一种在胃癌中的遗传变异类型，与基因组不稳定性和肿瘤进展相关。机器学习作为一种近年来常用于肿瘤的数据分析工具，可以对大规模的基因表达数据进行分析，筛选出与胃癌微卫星不稳定相关的特征基因，建立预后模型，有助于指导胃癌的预后评估和治疗决策。本文综合归纳了机器学习在肿瘤中分子特征筛选及预后模型建立的应用，分析其在胃癌诊断、治疗及预后判断中的应用价值，并对今后的研究方向进行展望。

关键词

胃癌，机器学习，MSI

A Review of the Application of Machine Learning in Molecular Feature Screening and Prognostic Model Establishment of Gastric Cancer

Yang Bao[1], Hongyang He[2], Yunfei Guan[2], Yang Zhang[2], Junwei Fang[2]

[1]School of Clinical Medicine, Dali University, Dali Yunnan
[2]Department 1 of General Surgery, The First Affiliated Hospital of Dali University, Dali Yunnan

Received: Feb. 19th, 2024; accepted: Mar. 12th, 2024; published: Mar. 20th, 2024

文章引用: 宝仰, 和红阳, 管云飞, 张阳, 方俊伟. 机器学习在胃癌的分子特征筛选及预后模型建立应用的回顾[J]. 临床医学进展, 2024, 14(3): 894-899. DOI: 10.12677/acm.2024.143788

Abstract

Gastric cancer ranks third in the mortality rate of malignant tumors in China, which seriously endangers our lives and health. Microsatellite instability is a type of genetic variant in gastric cancer that is associated with genomic instability and tumor progression. As a data analysis tool commonly used in tumors in recent years, machine learning can analyze large-scale gene expression data, screen out characteristic genes associated with microsatellite instability of gastric cancer, and establish prognostic models, which will help guide the prognosis assessment and treatment decisions of gastric cancer. This article comprehensively summarizes the application of machine learning in the screening of molecular features and the establishment of prognostic models in tumors, analyzes its application value in the diagnosis, treatment and prognosis of gastric cancer, and looks forward to future research directions.

Keywords

Gastric Cancer, Machine Learning, MSI

1. 引言

胃癌是我国常见的恶性肿瘤严重危害我们的生命和健康。WHO 国际癌症研究机构(IARC)于 2020 年发布的癌症数据报告[1]显示，我国胃癌的发病率占恶性肿瘤的第 2 位，死亡率占恶性肿瘤的第 3 位，胃癌患者 5 年总体生存率仅为 35.1% [2]。因胃癌早期临床症状不典型，许多患者治疗时已是进展期，且胃癌目前治疗方式单一，多采用手术治疗为主[3]。近年随着免疫领域的发展，胃癌治疗方式也日益增多，但晚期胃癌患者 5 年的生存率依旧较低[4]。

微卫星不稳定(Microsatellite Instability, MSI)，是由于错配修复系统功能缺陷，在 DNA 复制时出现插入或缺失突变进而引起微卫星序列长度改变的现象[5]。目前相关研究表明 MSI 对胃癌免疫治疗有一定的预测作用，尤其是 MSI-H 胃癌患者在免疫治疗中明显获益[6] [7]，如 PD-1、PD-L1 等免疫检查点抑制剂的广泛应用。NCCN 胃癌指南(2022.V2)也明确推荐，所有新诊断胃癌患者都应进行 MSI 的 PCR 检测或 MMR 的免疫组化检测[8]。但目前的临床研究中 MSI-H 胃癌患者数量较少，所产生的研究结果多为回顾性分析，缺乏实际的临床数据与结论。

随着人工智能的发展，机器学习步入我们的视野，机器学习的应用可帮助临床医师准确判断胃癌患者的预后及相关免疫治疗，同时机器学习也可应用在胃癌的诊断、疗效监测、预后判断相关生物标志物之中[9]，机器学习技术的发展为 MSI 型胃癌的发生、发展及个体化免疫治疗提供了简单有效的方法。

2. MSI 的相关概述及检测方法

2.1. MSI 的相关概述

微卫星不稳定(Microsatellite Instability, MSI)，是由于错配修复系统功能缺陷，在 DNA 复制时出现插入或缺失突变进而引起微卫星序列长度改变的现象。2014 年，癌症基因图谱计划(The Cancer Genome

Atlas, TCGA)将胃癌分为 4 个分子亚型：EB 病毒(EBV)阳性型、微卫星不稳定(MSI)型、基因组稳定(GS)型和染色体不稳定(CIN)型[10]。临床研究中研究者发现 MSI 型胃癌患者的预后明显优于微卫星稳定(MSS)型胃癌患者，且不同分子亚型产生不同的预后情况[11]。同时相关研究已经证实 MSI-H 可以作为结直肠癌预后的预测因素，但是不同癌种之间形成 MSI 的原因不全相同，MSI 与胃癌的发病机制、临床特征、治疗反应和预后方面的关系尚无统一定论。

2.2. MSI 的检测方法

目前我们常用的检测方法主要有二种：一、免疫组化(IHC)法，通过检测 MMR 蛋白中的“MLH1、MSH2、MSH6 和 PMS2”4 种蛋白。如肿瘤样本中 4 种蛋白均存在时为错配修复功能完整(Proflcient Mismatch Repair, pMMR)，即为 MSI-H 型；任何一种蛋白缺失即为错配修复缺陷(Deflcient Mismatch Repair, dMMR)，即为 MSI-L 型或 MSS 型[12]。但不同地方的病理报告结果存在一定的差异，因此免疫组化(IHC)法所产生的判读结果也不全相同，且外在影响因素较大。二、PCR 法，与免疫组化法相比，该法具有更高的灵敏度与特异度。美国国立癌症协会(NCI)推荐的对“BAT-25、BAT-26、D5S346、D2S123、D17S250”5 个位点进行检测为主，检测微卫星位点均未见基因位点表达异常，定义为 MSS；有其中 1 个标志物或<40%的标志物表达显示异常，则被定为 MSI-L；当≥40%的标志物出现表达异常时，则被定为 MSI-H [13]。但也有学者在研究中发现组成的样本不同，标记点和参考标准及研究方法不同所得出的结论中 MSI 型胃癌发生率也产生很大的差异[14]。为了进一步探明 MSI 型胃癌的特征，构建一个可靠的 MSI 检测标准成为研究的重中之重。

2.3. MSI 在胃癌中的研究进展

据统计结果显示 MSI 型胃癌患者仅占胃癌患者总数的 8%~25% [10]，在人群中的发病率并不高，目前的相关研究也证实了 MSI 型胃癌的发生与患者的家族史、饮食摄入、幽门螺杆菌、PD-L1、人端粒酶逆转录酶蛋白的表达均存在着许多密不可分的联系，在研究癌前病变转变为癌症时我们也发现了 MSI 升高的现象，这可能是与靶基因突变有关，如：TGFβRII、DP2、BCL 10、APAF 1、h-MSH6、BLM、ATR 等。此外，研究也显示 MSI 与胃癌患者的临床病理特征也存在着许多关联，胃癌患者的临床病理特征直接决定了其在临床中如何选择治疗方案，目前的研究中总结起来基本围绕了老年女性患者、肿瘤的远端位置、印戒细胞型、肠型、无神经浸润、无或少见淋巴结转移、浸润较浅、早期、脉管癌栓阳性、淋巴血管侵犯、黏膜下浸润、肿瘤较大、腹膜受侵等方面进行研究[15]，研究中也证实了胃癌患者的这些病理特征对 MSI 型胃癌患者的治疗、预后及复发率均起到一定的积极作用。针对胃癌治疗的研究，MSI 的状态对胃癌的手术、化疗和免疫治疗都起到了一定的影响，复习相关文献后发现目前较为一致的结论是：MSI-H 患者进行新辅助化疗是没有益处的，反而会减少 MSI-H 患者的生存期，MSS 或 MSI-L 患者可从新辅助化疗中高度获益，对于 II 期、III 期的 MSI-H 型胃癌患者单纯行手术治疗即可，不必行新辅助化疗。但意外的是免疫治疗的相关研究中表明了 MSI 型胃癌患者在免疫检查点抑制剂的疗效上显示出来巨大优势[16]，目前临床中常用的免疫检查点抑制剂如 PD-1、PD-L1、CTLA-4 等，许多研究也论证了 MSI 与 PD-1、PD-L1 之间的相关性，MSI 对于胃癌的免疫治疗效果确实有一定的预测作用，但 MSI 与胃癌的关系相对复杂，能够完整地、正确地阐述出二者之间的关系依旧存在着许多困难。

3. 机器学习对于肿瘤分子特征筛选的应用

筛选特征基因就是从大量基因表达数据中识别出与特定生物过程或疾病相关的基因，传统的筛选方法主要是基于统计学的假设检验和差异表达分析，如 t 检验和方差分析。然而，这些方法往往受到多重

检验问题的干扰，并忽略了基因之间的相互作用和非线性关系。相比之下机器学习算法能够更准确地识别出特征基因，并进行高维数据处理，常用的机器学习方法包括支持向量机(SVM)、随机森林(Random Forest)、人工神经网络(Artificial Neural Network)和深度学习(Deep Learning)等[17]。这些方法能够准确、快捷地处理大量基因表达数据，精准筛选出与癌症相关的特征基因。对于特征基因的筛选，我们经历了从单一方法到多种方法混合的转变，Golub 等人[18]以信噪比的选择方法成功应用在了人类急性白血病的分类问题上；Wang 等人[19]运用聚类分析和基因评估排序的特征基因选择算法，选出了低冗余且分类准确率高的基因；Lin 等人[20]采用随机采样的策略，运用遗传算法及支持向量机(SVM)，使的特征基因选取简单高效。以上研究者均采用单一的研究方法，尽管准确率与便利性逐步提高，但不可避免的存在许多缺点。随着人工智能的发展，目前出现了许多混合的方法，研究者们将多种方法进行适当组合，逐步提高特征选择的性能，弥补单一方法所造成的缺陷，如：Pavithra 等人[21]运用遗传算法、决策树、十折交叉法等算法，成功得到结直肠癌的特征基因并加以验证；Radovic 等人[22]基于最小冗余最大相关的算法，结合时序信息，得出了分类准确率随选择特征的变化而变化，并在癌症数据集上加以验证。综上所述均表明，机器学习算法在挖掘癌症特征基因筛选的应用具有好的性能，但目前特征基因筛选方法多种多样，所筛选出的特征基因也不全相同，许多特征基因在临床试验中暂无法验证，未来可采用先进的多种机器学习算法进行 MSI 型胃癌特征基因的挖掘，不断改进机器学习算法，使其具有更好的特征基因分类性能并在临床试验中得以验证。

4. 基于机器学习在胃癌预后模型建立的应用

目前大多数研究肿瘤预后的影响因素多从肿瘤大小、位置、浸润深度和淋巴结转移等情况入手[23]，临床实际应用中也有许多预后分期方法，如：TNM 分期、中国分期等，其中以 TNM 分期运用最为广泛，但近年来研究发现 TNM 分期系统不能对不同肿瘤特征的患者进行完全区分，使临床医师产生错误的预后评估，影响患者治疗[24]。因此运用机器学习建立胃癌预后模型提高胃癌预后预测的准确性及适用性是当前研究的重点。在机器学习算法建立预后模型中支持向量机(SVM)算法运用最广。如：CHEONG 等[25]、JIANG 等[26]、均运用了支持向量机(SVM)算法构建预后模型，有效预测胃癌患者的总体预后。当然除了应用支持向量机(SVM)算法外，也出现了其他的算法，如：LIU 等[27]基于 SEER 数据库，构建胃癌生存预测模型，为临床医生的决策和治疗方案提供了一定的理论依据；WU 等[28]基于随机森林算法，运用 DNA 甲基化数据进行胃癌预后模型建立；也有学者如 JOO M 等[29]运用深度学习的神经网络算法及 cox 模型构建生存分析模型，提高了模型的预测性能。以上基于机器学习构建胃癌预后模型的算法，都对胃癌预后预测模型的研究具有重要意义。但目前我们建立的模型适用性都较低，对于外部验证的数据集的应用依旧存在许多问题，解决起来都相对复杂繁琐，因此，未来的研究需要不断改进机器学习方法，利用先进的机器学习并结合医学领域知识，不断提高预测模型的准确性、适用性和可解释性。

5. 展望

目前机器学习在胃癌研究领域早已应用广泛，未来是一个大数据、人工智能化的时代，医学也是一门科学，机器学习将与我们的医学领域相结合创造出越来越多的技术。目前研究表明 MSI 型胃癌在免疫治疗中明显受益，如能精准筛选出 MSI 型胃癌的特征基因并建议一个可靠的预后模型那将大大提高临床医师的诊断、治疗及预后的决策能力。然而，目前机器学习对于医学研究依旧存在着许多问题，一是目前我们的原始数据均来自于临床，数据类型十分复杂并且信息量巨大，干扰因素大，大大降低了数据的纯粹性，加之临床标本量不足以及相关实验验证较少，造成机器学习数据准确性降低。二是在运用机器学习算法时，需要大量数据来进行训练和验证，对于不当使用数据集、数据缺失及低质量数据都容易导

致挖掘的信息不良以及模型的过度拟合、准确性及适用性低的问题[30]。因此，在未来的研究中，建议完善生物信息的数据收集及高质量数据的录入与管理，有效利用胃癌临床及基础的相关数据，运用先进的人工智能技术挖掘出更高水平的数据信息，建立可靠的 MSI 型胃癌特征基因筛选的标准，构建出更准确、高效的预后模型，辅助临床医师做出对病人更加有利的临床决策。

基金项目

本文系云南省教育厅科学研究基金项目“机器学习在胃癌的微卫星不稳定特征基因筛选及预后模型建立的应用”编号为：2023Y0967。

参考文献

[1] Sung, H., Ferlay, J., Siegel, R.L., *et al*. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[2] Zeng, H.M., Chen, W.Q., Zheng, R.S., *et al*. (2018) Changing Cancer Survival in China during 2003-15: A Pooled Analysis of 17 Population-Based Cancer Registries. *The Lancet Global Health*, **6**, e555-e567. https://doi.org/10.1016/S2214-109X(18)30127-X

[3] 胃癌诊治难点中国专家共识(2020 版) [J]. 中国实用外科杂志, 2020, 40(8): 869-904.

[4] 项涛, 雷慧, 谭绮琼, 等. IL-27, IL-29 及 miRNA-497 在 HER-2 阳性胃癌患者放射性粒子联合靶向治疗中的价值[J]. 重庆医学, 2017, 46(30): 4204-4206.

[5] Baretti, M. and Le, D.T. (2018) DNA Mismatch Repair in Cancer. *Pharmacology & Therapeutics*, **189**, 45-62. https://doi.org/10.1016/j.pharmthera.2018.04.004

[6] Miceli, R., An, J., Di Bartolomeo, M., *et al*. (2019) Prognostic Impact of Microsatellite Instability in Asian Gastric Cancer Patients Enrolled in the ARTIST Trial. *Oncology*, **97**, 38-43. https://doi.org/10.1159/000499628

[7] Pietrantonio, F., Miceli, R., Raimondi, A., *et al*. (2019) Individual Patient Data Meta-Analysis of the Value of Microsatellite Instability as a Biomarker in Gastric Cancer. *Journal of Clinical Oncology*, **37**, 3392-3400. https://doi.org/10.1200/JCO.19.01124

[8] NCCN (2022) NCCN Clinical Practice Guideline in Oncology, Gastric Cancer (Version 2).

[9] 严健亮, 景蓉蓉, 谢泽宇, 崔明. 机器学习在胃癌生物标志物挖掘中的应用进展[J]. 实用医学杂志, 2023, 39(6): 783-787.

[10] Cancer Genome Atlas Research Network (2014) Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature*, **513**, 202-209. https://doi.org/10.1038/nature13480

[11] 郑瑞, 聂明明. 微卫星不稳定性在胃癌治疗作用中的研究进展[J]. 中国临床医学, 2022, 29(5): 864-869.

[12] 袁玥, 沈存芳. 微卫星不稳定型胃癌的研究进展[J]. 世界最新医学信息文摘, 2018, 18(28): 84-85+93.

[13] 施维, 薛均, 潘璀然, 任元凯, 倪正杰, 张远鹏, 王理, 吴辉群, 蒋葵, 董建成. 机器学习在肿瘤早期诊断与预后预测中的应用[J]. 医学信息学杂志, 2016, 37(11): 10-14+22.

[14] 刘欢, 辛彦. 微卫星不稳定性(MSI)与胃癌关系的研究进展[J]. 现代肿瘤医学, 2018, 26(1): 124-127.

[15] 李立立, 王艳军, 安有志. 微卫星不稳定型胃癌的研究进展[J]. 癌症进展, 2022, 20(15): 1519-1524.

[16] 王雅, 吕佳乐, 梁路等. MSI 检测在胃癌治疗中的研究进展[J]. 胃肠病学和肝病学杂志, 2022, 31(6): 691-695.

[17] 陈凯, 朱钰. 机器学习及其相关算法综述[J]. 统计与信息论坛, 2007(5): 105-112.

[18] Golub, T.R., Slonim, D.K., Tamayo, P., *et al*. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531-537. https://doi.org/10.1126/science.286.5439.531

[19] Wang, Y., Makedon, F.S., Ford, J.C., *et al*. (2005) HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification Using Microarray Gene Expression Data. *Bioinformation*, **21**, 1530-1537. https://doi.org/10.1093/bioinformatics/bti192

[20] Lin, T.C., Liu, R.S., Chao, Y.T., *et al*. (2013) Classifying Subtypes of Acute Lymphoblastic Leukemia Using Silhouette Statistics and Genetic Algorithms. *Gene*, **518**, 159-163. https://doi.org/10.1016/j.gene.2012.11.046

[21] Pavithra, D. and Lakshmanan, B. (2017) Feature Selection and Classification in Gene Expression Cancer Data. 2017 *IEEE International Conference on Computational Intelligence in Data Science* (*ICCIDS*), Chennai, 2-3 June 2017, 1-6.

https://doi.org/10.1109/ICCIDS.2017.8272668

[22] Radovic, M., Ghalwash, M., Filipovic, N., *et al*. (2017) Minimum Redundancy Maximum Relevance Feature Selection Approach for Temporal Gene Expression Data. *BMC Bioinformatics*, **18**, Article No. 9. https://doi.org/10.1186/s12859-016-1423-9

[23] 魏晟宏, 陈路川, 叶再生, 林振孟, 王益, 严明芳. 胃癌肿瘤大小的临床病理特征及预后分析(附 753 例报告) [J]. 福建医药杂志, 2017, 39(6): 88-91.

[24] 陕飞, 李子禹, 张连海, 李双喜, 贾永宁, 苗儒林, 薛侃, 李浙民, 高翔宇, 王胤奎, 闫超, 李沈, 季加孚. 国际抗癌联盟及美国肿瘤联合会胃癌 TNM 分期系统(第 8 版)简介及解读[J]. 中国实用外科杂志, 2017, 37(1): 15-17.

[25] Cheong, J.H., Wang, S.C., Park, S., *et al*. (2022) Development and Validation of a Prognostic and Predictive 32-Gene Signature for Gastric Cancer. *Nature Communications*, **13**, Article No. 774.

[26] Jiang, Y., Xie, J., Han, Z., *et al*. (2018) Immunomarker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit. *Clinical Cancer Research*, **24**, 5574-5584. https://doi.org/10.1158/1078-0432.CCR-18-0848

[27] Liu, D., Wang, X., Li, L., *et al*. (2022) Machine Learning-Based Model for the Prognosis of Postoperative Gastric Cancer. *Cancer Management and Research*, **14**, 135-155. https://doi.org/10.2147/CMAR.S342352

[28] Wu, J., Xiao, Y., Xia, C., *et al*. (2017) Identification of Biomarkers for Predicting Lymph Node Metastasis of Stomach Cancer Using Clinical DNA Methylation Data. *Disease Markers*, **2017**, Article ID: 5745724. https://doi.org/10.1155/2017/5745724

[29] Joo, M., Park, A., Kim, K., *et al*. (2019) A Deep Learning Model for Cell Growth Inhibition IC50 Prediction and Its Application for Gastric Cancer Patients. *International Journal of Molecular Sciences*, **20**, Article No. 6276. https://doi.org/10.3390/ijms20246276

[30] 徐嘉昕, 钱凯, 蒋立虹. 机器学习算法在肺癌临床诊断及生存预后分析中的应用[J]. 中国胸心血管外科临床杂志, 2022, 29(6): 777-781.