乳腺癌骨转移是一个复杂的、多步骤的生物学过程。首先，原发乳腺癌细胞增殖并侵袭周围组织，在钙黏蛋白、基质金属蛋白酶（MMPs）、整合素等因子调节下逃逸入血。癌细胞受趋化因子引导迁移至骨骼，并通过外渗离开血管到达骨组织。一旦进入骨骼，癌细胞就可以利用整合素和钙黏蛋白附着并结合到骨基质上。最终，癌细胞通过与骨微环境的相互作用得以存活、增殖和分化，从而导致骨转移。

——————

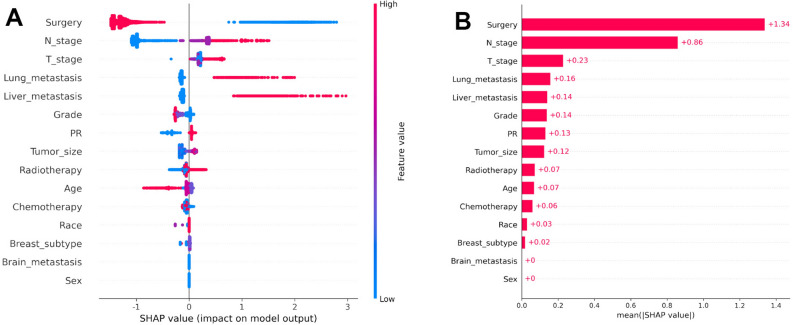
Zhong的研究表明，乳腺癌（BC）是女性最常见的恶性肿瘤，根据2022年美国癌症统计数据，BC是美国女性中最常见的新发恶性肿瘤，占新诊断肿瘤的31%，也是癌症相关死亡的第二大病因。随着乳腺癌生存率的不断提高，发生转移的患者数量也在增加，其中骨是最常见的远处转移部位，近75%的远处转移为骨转移（BCBM）。

Zhong基于SEER数据库分析了2010年至2016年间283373名BC患者，其中，骨转移发生率约为2.3%，BCBM患者主要年龄段为40-79岁，白人发病率明显高于其他种族，其中Luminal A型BC占65.7%。在BC的预后方面，虽然无转移BC患者的5年总生存率超过80%，但发生远处转移的患者5年生存率仅有25%，而BCBM患者的5年生存率仅为22.8%。另外，骨转移引起的骨折、高钙血症或脊髓压迫等骨质事件发病率约为45.1%，这些会显著地影响患者的预后。所以早期识别可能发生骨转移的高危患者并预测其生存率，从而指导后续检查和治疗对BC患者至关重要。

Zhong基于单因素和多因素分析研究了BCBM发生的危险因素，单因素分析表明，手术是最显著的保护因素，而脑转移是最显著的危险因素，其他危险因素还包括分级、T分期、N分期、肺转移和肝转移。多因素分析结果表明，BCBM的危险因素包括年龄、种族、婚姻状况、分级、手术、放疗、化疗、脑转移、肺转移、肝转移、ER和PR，其中手术是最显著的保护因素。

Zhong根据纳排指标从数据库中选取了3492例BCBM患者构建了5种机器学习模型，并采用五重交叉验证对机器学习预测模型的性能进行评价，以重复5次后得到的结果和各结果ROC曲线的平均ROC曲线作为评价指标。在其构建的模型中，XGBoost 模型在五重交叉验证中表现最好。

SHAP分析显示：手术、N分期、T分期是BCBM最显著的影响因素。手术是保护因素，而N、T分期是危险因素。



——————

Li利用MRI影像组学特征结合深度学习算法建立BC远处转移预测模型，他使用深度学习算法对DCE-MRI和1期增强MRI进行分割，纳入了96例转移组患者共100个病灶和随机抽取192例非转移组患者共197个病灶，并从病灶中提取形态学特征和灰度统计特征，通过特征降维和交叉验证建立预测模型。另外，他还比较了仅基于影像特征的预测模型与结合影像组学特征和临床特征的联合模型在准确率、敏感性、特异性和AUC方面的表现。结果表明，后者在敏感性、特异性和准确性三个方面都优于前者。

——————

Wen-hai Zhang采用双中心回顾性研究设计，纳入239例患者，利用常规临床血液生物标志物构建预测模型。其中一个中心纳入了 176 例病例，以 8 : 2 的比例随机分为训练组（123 例）和测试组（53 例）。另一个中心提供了 63 个病例作为外部验证队列。外部验证队列来自地理位置接近但不同的医疗中心，从而验证该模型在不同机构和患者群体中的普遍性。

纳入标准：

* 明确诊断为原发性乳腺癌伴有新发骨转移
* 在治疗（放疗或化疗）或手术切除前完成临床血液生物标志物检测
* 无高血压、糖尿病或高脂血症病史
* 无肝、肾或心血管功能相关的血液指标异常史
* 无其他疾病史

排除标准：

* 治疗（手术切除或化疗）后发生远处转移
* 临床血液生物标志物数据不完整，包括肿瘤标志物（AFP、CEA、CA125、CA153、CA199）、肝肾功能检查、血脂或心血管功能检查
* 年龄<18岁
* 发生骨以外部位的转移

Zhang首先使用Spearman相关分析和lasso回归进行特征降维：

首先，应用 Z-sorce标准化对每个特征进行归一化，对数据进行预处理以拟合标准正态分布。然后，利用Spearman秩相关系数（ρ）进行统计分析衡量两个变量之间的相关性。当 ρ 接近 1 或 -1 时，表明变量之间存在很强的相关性。其选择 ρ > 0.9作为高相关性的阈值。超过阈值意味着多重共线性，需要只保留一个P最小的变量作为特征变量。然后使用lasso回归的L1正则化进行特征降维。lasso方法惩罚回归系数的绝对值，从而诱导一些系数为零，便于特征选择并生成稀疏模型。为了确定lasso的lambda值，作者对训练集采用10倍交叉验证确定lambda参数，并选择最小化均方误差的值，从而防止过拟合。

然后用LightGBM机器学习算法构建初诊骨转移的预测模型，并在训练集、内部验证集和外部验证集中进行验证。研究结果显示，LightGBM模型在训练集、内部测试集和外部验证集的AUC都表现良好，而且有较高的灵敏度、特异度和准确性。

特征重要性分析显示，CEA、肌酸激酶、白/球比值、载脂蛋白B和CA153在模型预测中起关键作用。单因素分析表明，脂蛋白a、CA153、γ-谷氨酰转移酶、α-羟丁酸脱氢酶、碱性磷酸酶和肌酸激酶与BC骨转移呈正相关，而白细胞比值和总胆固醇呈负相关。

——————

Chao Zhang进行了一项回顾性研究，纳入了6703例BC患者，根据纳排标准最终确定186例远处转移（DM）和无DM患者，并按照7:3设置训练集和验证集，通过整合这些患者的多维度数据（包括MRI和超声影像特征、临床病理特征、实验室检查指标等），构建并验证了三个Cox回归模型：远处转移预测模型、骨转移预测模型和内脏转移预测模型，并通过列线图的方式呈现。结果显示临床组学（临床元数据+影像组学）模型的预测性能显著优于仅基于临床或影像数据的模型。该模型可用于BC患者的个体化转移风险分层和指导精准的预防性治疗和随访策略。

详细的纳入和排除标准如下：

* 通过手术切除标本和穿刺活检确诊BC
* 具有诊断意义的术前MRI和US图像
* 新辅助治疗或手术切除前进行的MRI扫描和超声检查
* 至少五年的随访数据

**构建基于影像组学的模型：**

影像组学特征的确定：为每位患者共提取 2569 个影像组学特征（855 个来T2WI的特征、859 个DCE-MRI的特征和 855 个来自US的特征）。然后从所有的这些影像组学特征中确定可靠的放射组学特征。

1. 筛选结局相关特征：采用Wilson检验鉴定与生物标志物高度相关（P<0.05）的特征
2. 避免多重共线性：采用Pearson相关系矩阵评价特征间的相关性，相关系数大于0.8的意味着多重共线性，为冗余特征，只保留一个P值最低的特征。
3. 采用lasso回归选择特征的最佳预后组合：通过计算径向评分，计算每个患者所选特征的线性组合，并按其各自的系数加权得到RadSorce（基于影响组学的危险值）。最后进行十次交叉验证以评估模型的真实诊断潜力。

最终选定了8个影像学特征。并利用这些特征建立了组合影像组学的Cox模型。通过对比组合影像组学模型、T2WI影像组学模型、US影像组学模型和DCE-MRI影像组学模型在测试集的AUC的得出结论：组合影像组学模型最佳。

**构建基于临床特征的模型：**

采用单因素分析评价训练集中的临床病理因素。将单因素分析P<0.05的变量纳入Cox回归模型，从而建立临床病理列线图。

筛选出来的远处转移相关的临床特征：淋巴结转移、CA153升高、CEA升高、CA125升高。

作者通过这些临床病理学特征构建出了远处转移的临床病理学Cox模型，并且进一步与组合影像组学模型结合构建出了远处转移的临床组学Cox模型。

另外，作者分层分析了骨转移相关因素：生育史、淋巴结转移、CA153升高、CEA升高、CA125升高。还有内脏转移相关因素：淋巴结转移、ER阳性、PR阳性、CA153升高、CEA升高、内分泌治疗。并且结合影像组学模型分别构建了骨转移模型和内脏转移模型。

作者通过对比临床病理学模型、组合影像组学模型和临床组学模型术后1，2，3年的AUC得出结论：临床组学模型最佳。

文末，作者将远处转移预测模型、骨转移预测模型和内脏转移Cox预测模型通过列线图的方式呈现，并给出了训练集和测试集的校准曲线和ROC。

——————

Gao基于SEER数据库开发了多种机器学习模型来预测BC患者发生BCBM的风险，并通过SHAP框架识别了关键预测特征（包括肿瘤大小、分级、T/N分期、ER/PR/HER2及脑肝肺转移），其中MLPNN模型表现最佳，RF模型在排除骨转移方面最佳。

Gao采用Boruta算法筛选BCBM的重要预测因素，结果表明，性别、肿瘤大小、多灶肿瘤、组织学分级、T分期、N分期、ER、PR、HER2以及脑转移、肝转移和肺转移是BCBM的显著预测特征。其中肺转移是最显著的特征。

——————  
Fabio指出，乳腺癌骨转移的危险因素主要包括肿瘤本身的病理学特征和患者的基本情况。在病理学方面，肿瘤大小超过5厘米、较高的肿瘤分级、小叶癌亚型以及较多的阳性淋巴结数量都会增加骨转移风险，特别是当有4个或以上腋窝淋巴结转移时，2年骨转移累积发生率可达14.9%，10年时更高达40.8%。在临床方面，年龄超过60岁和血红蛋白水平异常也是重要的危险因素。另外，癌症症状持续时间较长的患者骨转移风险也更高。  
骨转移是实体肿瘤最常见的转移部位，其中乳腺癌占所有骨转移病例的36%。根据SEER数据库的回顾性分析，不同乳腺癌分子亚型在骨转移发生率和转移模式上有明显差异。管腔型乳腺癌有最高的骨转移倾向，其中ER+/HER2-的骨转移发生率为58.52%，ER+/HER2+为47.28%。相比之下，ER-/HER2+更倾向于肝转移，发生率达31.72%，而三阴性乳腺癌则以肺转移为主，发生率为32.09%，这两种亚型的骨转移率分别为34.49%和36.39%。另外，Molnar等人的研究发现luminal A型（ER+/HER2-）乳腺癌不仅骨转移发生率高，而且在59%的病例中表现为孤立性骨转移，这类患者发生骨转移时往往不伴有其他器官的转移。

——————

相关文献：

[1] Zhong, X., Lin, Y., Zhang, W., & Bi, Q. (2023). Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning. *Scientific reports*, *13*(1), 18301. https://doi.org/10.1038/s41598-023-45438-zIF: 3.9 Q1

[2] Li, L., Tian, H., Zhang, B., Wang, W., & Li, B. (2022). Prediction for Distant Metastasis of Breast Cancer Using Dynamic Contrast-Enhanced Magnetic Resonance Imaging Images under Deep Learning. *Computational intelligence and neuroscience*, *2022*, 6126061. <https://doi.org/10.1155/2022/6126061>

[3] Zhang, W. H., Tan, Y., Huang, Z., Tan, Q. X., Zhang, Y. M., & Wei, C. Y. (2024). Development and validation of an artificial intelligence model for predicting de novo distant bone metastasis in breast cancer: a dual-center study. *BMC women's health*, *24*(1), 442. <https://doi.org/10.1186/s12905-024-03264-z>

[4] Zhang, C., Qi, L., Cai, J. *et al.* Clinicomics-guided distant metastasis prediction in breast cancer via artificial intelligence. *BMC Cancer* **23**, 239 (2023). <https://doi.org/10.1186/s12885-023-10704-w>

[5] Gao, Y., Liu, L., Wang, S., Tao, W., Wang, J., Duan, R., Xie, H., Takahashi, H., Hao, J., & Gao, M. (2025). SEER-based machine learning prediction of bone metastasis in breast cancer: model development and validation. *Gland surgery*, *14*(7), 1366–1378. <https://doi.org/10.21037/gs-2025-168>

[6]Marazzi, F., Orlandi, A., Manfrida, S., Masiello, V., Di Leone, A., Massaccesi, M., Moschella, F., Franceschini, G., Bria, E., Gambacorta, M. A., Masetti, R., Tortora, G., & Valentini, V. (2020). Diagnosis and Treatment of Bone Metastases in Breast Cancer: Radiotherapy, Local Approach and Systemic Therapy in a Guide for Clinicians.Cancers, 12(9), 2390. https://doi.org/10.3390/cancers12092390

——————

**乳腺癌骨转移预测模型的回顾性研究方案**

**一、研究背景与目的**

**1.1 研究背景**

乳腺癌（Breast Cancer，BC）是女性最常见的恶性肿瘤之一。BC的发病率占全身各种恶性肿瘤的7-10%，同时也是癌症相关死亡的第二大病因。尽管近年来BC的诊疗水平不断提升，患者总体生存率显著提升，但与此同时，远处转移的患者也持续增加，这已成为影响BC患者长期预后的关键因素。

在BC的各类远处转移中，骨转移（Bone Metastasis，BCBM）最为常见，约占所有远处转移的75%。BCBM对患者预后的影响极为显著。相比于无转移BC患者超过80%的5年总生存率，发生BCBM患者的5年生存率仅有22.8%。此外，骨转移常伴发病理性骨折、高钙血症、脊髓压迫等骨相关事件，这些并发症不仅会严重降低患者生活质量，还会显著影响患者预后。

因此，早期识别骨转移高危人群、准确预测患者生存风险具有重要的临床价值。通过建立精准的预测模型，可以帮助临床医生在疾病早期筛选出高危患者，及时开展针对性的影像学监测和预防性治疗干预，从而改善BC患者的整体预后。对于优化BC患者的个体化诊疗策略、提高临床决策效率具有重要的现实意义。

* 1. **研究目的**

构建基于临床元数据和影像学数据的乳腺癌骨转移预测模型，识别乳腺癌骨转移的独立危险因素，为临床决策提供依据。

**二、研究设计**

**2.1 研究类型**

回顾性队列研究

**2.2 研究对象**

**纳入标准:**

* 通过手术切除标本或穿刺活检诊断为BC患者
* 临床资料完整
* 影像学资料完整，包括钼靶X线、MRI和超声检查
* 无肝、肾或心血管功能相关的血液指标异常

**排除标准:**

* 诊断时已经存在 BCBM
* 临床资料或影像学资料不完整
* 合并其他恶性肿瘤
* 随访资料缺失

**2.3 样本量估算**

目标样本量需要150-225例患者（阳性模型约占20-40%）。

**三、数据收集**

**3.1 临床元数据**

* + 1. **人口学特征**

年龄、性别、身高、体重、BMI、文化水平、吸烟史、饮酒史、种族、婚姻史、生育史、绝经状态、家族史

**3.1.2 肿瘤特征**

* 病理类型
* 组织学分级
* 临床分期（TNM分期）
* 分子分型：ER、PR、HER-2、Ki-67

**3.1.3 实验室指标**

* 白细胞、血红蛋白、血小板、血清总蛋白、血清白蛋白、丙氨酸氨基转移酶、天冬氨酸氨基转移酶、血清肌酐、血尿素氮、血清钠、血清氯、血清钾、血糖、纤维蛋白原、D-二聚体和血氧饱和度、动脉血气pH值
* 酸性磷酸酶、碱性磷酸酶、乳酸脱氢酶、血钙、球和白蛋白、肌酸激酶、载脂蛋白
* CEA、CA153、CA125

**3.1.4 治疗信息**

* 手术方式
* 化疗方案
* 放疗方案

**3.2 影像学数据**

* 乳腺钼靶X线
* 乳腺超声
* 乳腺MRI

**3.3 结局指标**

主要结局：骨转移的发生

次要结局：无骨转移生存时间、总生存时间、其他部位转移情况

**四、统计分析方法**

**4.1 数据预处理**

影像学数据：对所有图像进行统一的预处理，包括脱敏、重采样、灰度归一化、降噪和去除伪影，以确保数据一致性。利用3D Slicer提取影像组学特征。

临床元数据：

首先进行数据导入和质量检查，删除重复记录，评估各变量的缺失情况。对于缺失率超过20%的变量予以删除；缺失率在5%-20%之间的变量采用多重插补法处理，生成插补数据集；缺失率低于5%的变量，连续变量用中位数填补，分类变量用众数填补。处理完缺失值后，对连续变量进行异常值检测，结合临床实际判断是否为录入错误并进行修正或删除。  
分类变量采用独热编码，二元变量采用0/1编码。对所有连续变量进行Z-score标准化处理，消除量纲影响，提高模型收敛速度和预测性能。

对变量进行多重共线性检验，处理存在严重的多重共线性变量。计算连续变量之间的Pearson相关系数矩阵，对于包含分类变量的多因素模型，计算方差膨胀因子(VIF)，一般认为VIF>10表示存在严重多重共线性，只保留P值最小的变量。

根据单因素和多因素分析确定对BCBM发生具有临床意义的变量，根据Boruta算法确定对BCBM影响较大的变量。两者取交集确定最终的特征变量。

**4.2 机器学习模型**

逻辑回归、RF、Xgboost、Svm、LightGBM

**4.3 模型验证**

数据集划分：训练集(70%)和验证集(30%)

交叉验证：10折交叉验证

**4.4 模型性能评估**

* 区分度：以敏感度为纵坐标，1-特异度为横坐标绘制ROC曲线，曲线下面积（AUC）用以评估模型的区分度，对比训练集和验证集的AUC-ROC
* 校准度：分别绘制训练集和验证集的校准曲线，验证模型的稳健性和可靠性。使用Hosmer-Lemeshow拟合优度检验确定模型的拟合度
* 有效性：分别绘制训练集和验证集的决策曲线，并与全预测和无预测进行对比，确定模型在大部分阈值概率范围内的临床应用价值

对比模型的AUC、灵敏度、特异度、准确率、 精确率、F1-score

**4.5 模型解释**

应用SHAP框架，计算每个特征对每个样本预测结果的贡献值。

通过SHAP摘要图，展示特征的全局重要性排序。

**五、研究时间安排**

第1-4个月：方案设计、伦理申请、数据收集与整理

第5-8个月：数据清理、统计分析、模型构建与验证

第9-12个月：论文撰写与投稿

**六、预期结果与意义**

**6.1 预期结果**

构建一个准确、稳定的BCBM预测模型

整理研究方法、模型构建过程与验证结果并发表论文

**6.2 研究意义**

为临床医生早期识别骨转移高危患者提供工具

指导个体化随访

为骨转移预防性干预提供决策依据

为后续前瞻性研究奠定基础