# 临床数据集数据集偏移文献综述

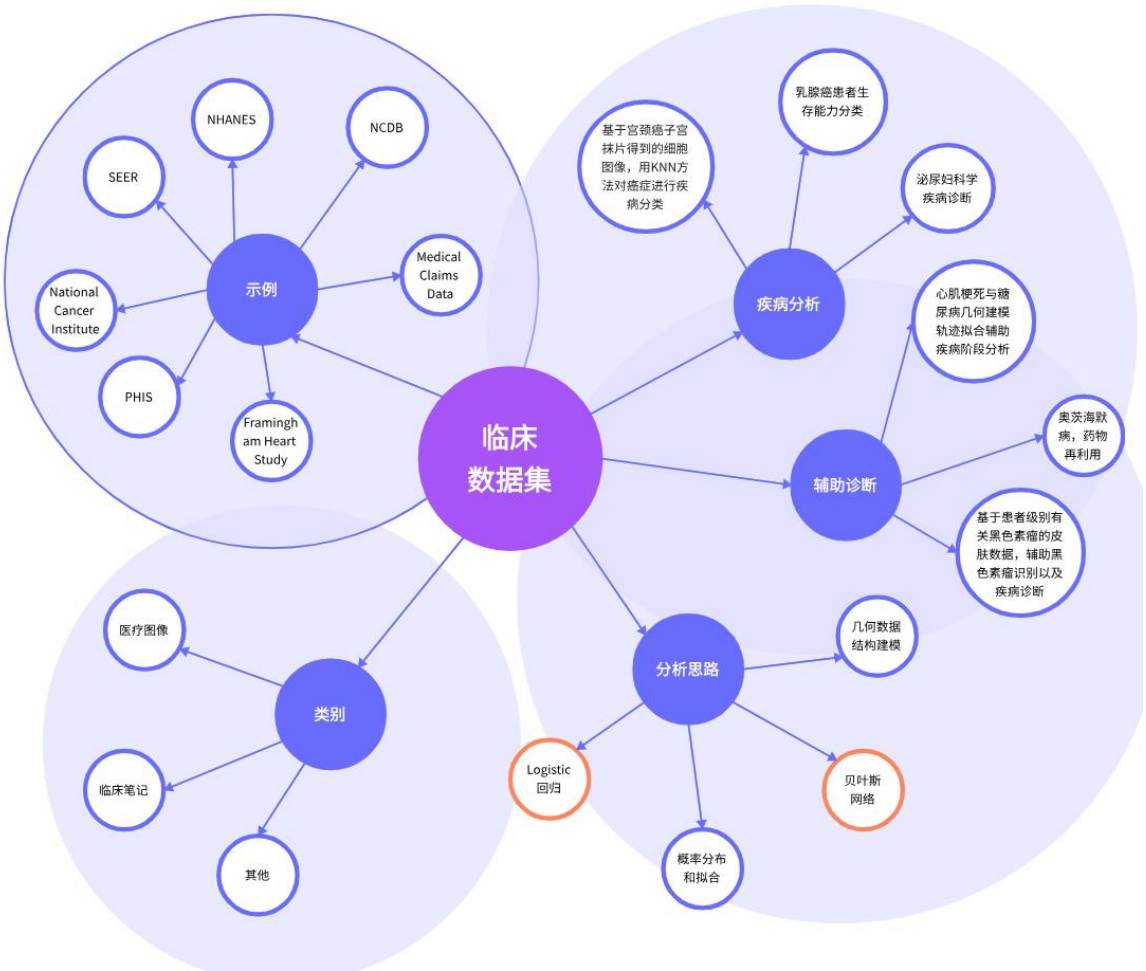
陈苇远　3210105677

【摘要】随着医疗大数据时代的来临，越来越多的高质量临床数据集由专业机构整理收集，形成了一套完整的数据库，这些数据库促进了临床资源的使用，为了人类医疗事业做出了重大的贡献。然而在使用这些医疗数据库进行建模和科研的过程中，科学家们也经常遇到使用公开医疗数据库进行训练的模型在实际应用于测试数据集时结果很差的情况，即数据集偏移的问题。临床数据集偏移是指因为技术变化、人口和环境变化或人们行为变化等原因，开发时的数据集与部署时的数据集不匹配，从而导致机器学习模型不再具有鲁棒性，严重时会影响医疗系统的安全性。通过检测校准漂移指导数据驱动的数据集更新策略、使用数据流聚类方法加权更新数据集、使用合成数据集作为替代方案等方法可以在一定程度上缓解临床数据集偏移。

【关键词】临床数据集偏移、机器学习、鲁棒性、数据驱动

## 1. 临床数据集应用现状

随着医疗大数据时代的来临，越来越多的高质量临床数据集由专业机构整理收集，形成了一套完整的数据库。这些数据集或来源于基于人群的横断面调查，如国家健康和营养检查调查（NHANES），旨在收集和反映社会上成人和儿童的普遍健康状况；或由特定疾病医疗学会牵头、基于医院登记数据，如美国国家癌症数据库（NCDB），旨在通过收集和分析治疗方案、病理、基因等临床信息，研究特定疾病的预后研究、阶段分析死亡研究等[14,28]。多样化的高质量临床数据集促进了临床资源的使用，为了人类医疗事业做出了重大的贡献。



近年来，医疗数据集的爆炸性增长导致了基于机器学习的数据驱动医学研究的激增[30]。在运用机器学习进行临床数据集分析的过程中，临床数据集被分为三类：医学图像、临床笔记和其他[1]。其中最普遍的医疗数据集是医学图像，医学图像在输入时可以被分类为矢量格式，如 2D 像素值或 3D 体素值，适用于典型的多层神经网络或卷积神经网络。深度学习方法已应用于医学图像的计算机辅助病理学检测、图像分割和疾病分类[15,16]。临床笔记，例如出院总结、测量报告和死亡证明等，则通常以文本形式呈现，在机器学习中常用自然语言处理技术和各种深度学习方法完成医疗概念提取、疾病分类等重要任务[1]。医学图像类数据集是应用最广泛的临床数据集，其次是临床笔记类临床数据集，此类数据集在某些情况下与医学图像相结合研究以提高系统性能，第三类临床数据集即其他数据集鲜少使用[1]。

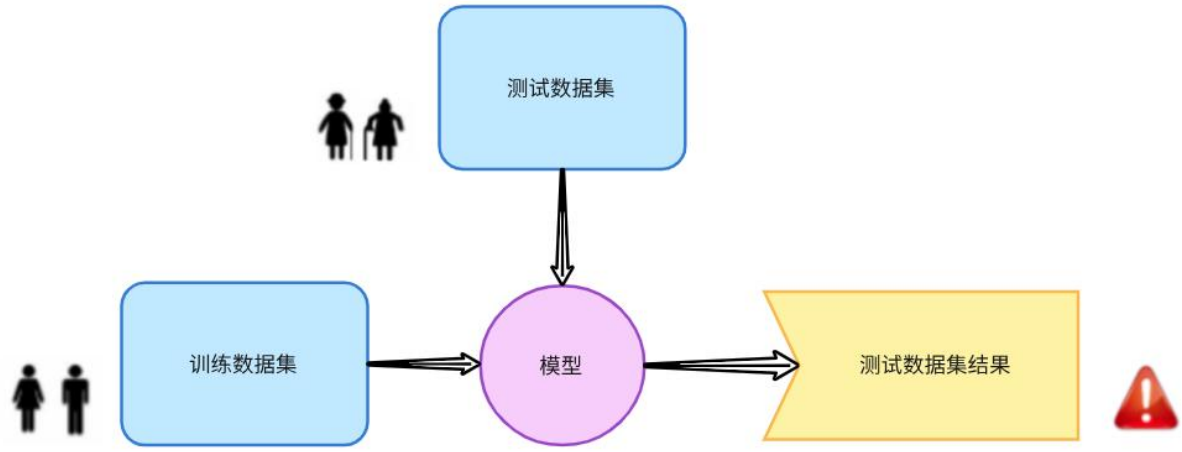
在临床数据集的应用方面，通过多样的分析方法处理医疗数据，与现实中临床诊断结合以达到更准确的诊断结果和更好的治疗效果。科学家们已经做到了通过概率分布和拟合的思想，使用贝叶斯网络对医学图像中面部、动作、形态等多模态的分析，用来比对和监测人的健康状态[2]。2014 年公开发表的研究中已经做到利用概率论与统计学中基础的 $X^2$ 检验和 t 检验方法、Logistic 回归，分析手术期死亡率和并发症辅助治疗[9,24]。2021 年科学家在研究中已经做到利用几何数据结构建模，将疾病状态典型发展路线通过数据可视化具象为弹性主图中分叉临床轨迹花束，用以辅助治疗决策[7,29]。

临床数据集分析常常应用于疾病分析和辅助诊断，一个典型的例子是基于广泛的癌症相关数据集对于癌症相关病理分析、阶段预测与辅助诊断等。2016 年 7 月，相关研究者在《印度科学技术杂志》上发表了基于宫颈抹片得到的细胞图像，分析细胞图像特征并归一化、使用 KNN 方法对癌症进行异常分类的有效方法，通过分析特定疾病患者的数据集的相关指标来预测该疾病的相关特征和阶段预测[4]。2023 年 1 月，《智能自动化与软计算》杂志上发表的研究提出基于乳腺癌患者数据库，应用基于深度学习的监督分类器，提升乳腺癌患者生存能力分类准确性的算法和工作流程[3]。

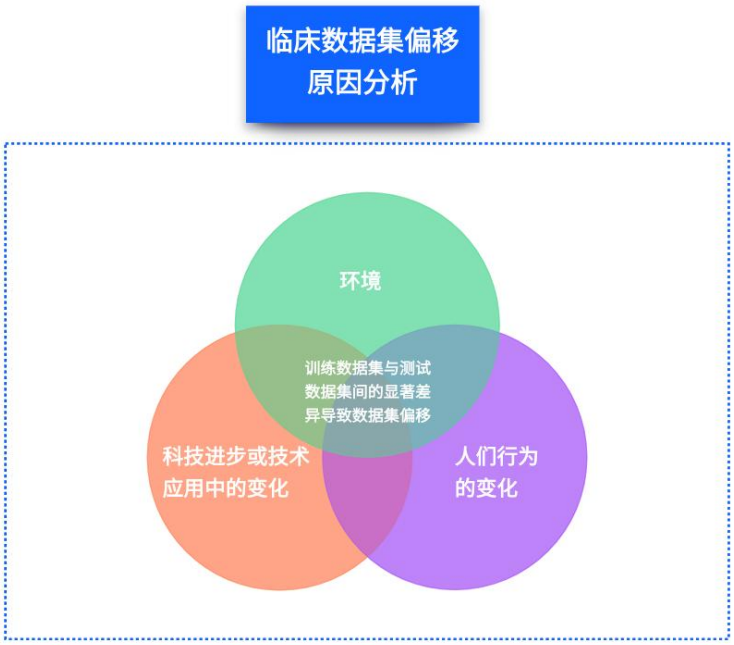
类似的利用临床数据集进行分析也被应用在一些其他的疾病中。2021 年 8 月，研究者基于 BSUG 数据集，一个泌尿妇科学方向的非强制性手术数据库，探讨了泌尿妇科学临床诊断中的数据分析应用[6]。2021 年 8 月，科学家基于患者级别的有关黑色素瘤的皮肤数据，通过深度学习方法判断黑色素瘤并且成功假阳性的概率、提高准确度[5]。2021 年 9 月，研究者通过将患者定位于特定的临床轨迹（病理场景），通过预后不确定性的定性估计来描述其延轨迹的进展程度，进而提出一种基于心肌梗死与糖尿病相关临床数据集分析大型临床数据的半监督方法[7]。2023 年 7 月，研究者介绍了基于范德堡大学医学中心（VUMC）和美国国立卫生研究院（NIH）的电子病历数据集，结合新兴的生成人工智能技术，如 ChatGPT，对治疗有限且症状严重的疾病，如奥茨海默病，进行药物识别和推荐，以达到药物再利用的结果[8]。

## 2. 临床数据集偏移及其产生原因分析

然而在使用医疗数据库进行建模和科研的过程中，由于现行的概率模型、临床 AI 机器学习系统都基于基本的概率分布和统计方法，从临床数据中学习关键模式的过程依赖于数据集的分布情况[18,41]。科学家们也经常遇到使用临床数据集进行训练的模型在实际应用于测试数据集时结果很差的情况，即数据集偏移的问题[19]。数据集偏移通常是指机器学习系统在其开发时使用的数据集与其部署时使用的数据不匹配而表现不佳的情形[18,41]。




当前的大数据临床数据集的规模通常很大，可以排除采样对于结果的扰动，但是数据生成方式的差异仍然会对结果造成影响，因此大规模的数据不代表一个临床数据集在任何实际疾病分析场景下都具有普遍性[17,41]。造成这种临床数据集偏移的原因通常有两类，一类是训练样本与测试样本在变量空间的分布差异巨大，即特征值偏移，另一类是训练数据集与测试数据集的类别分布不均衡，即类别偏移[22,23]。常见的临床数据集偏移的原因有遗传、环境和种族分布；科技进步或技术应用中的变化；人们行为的变化等[17]。

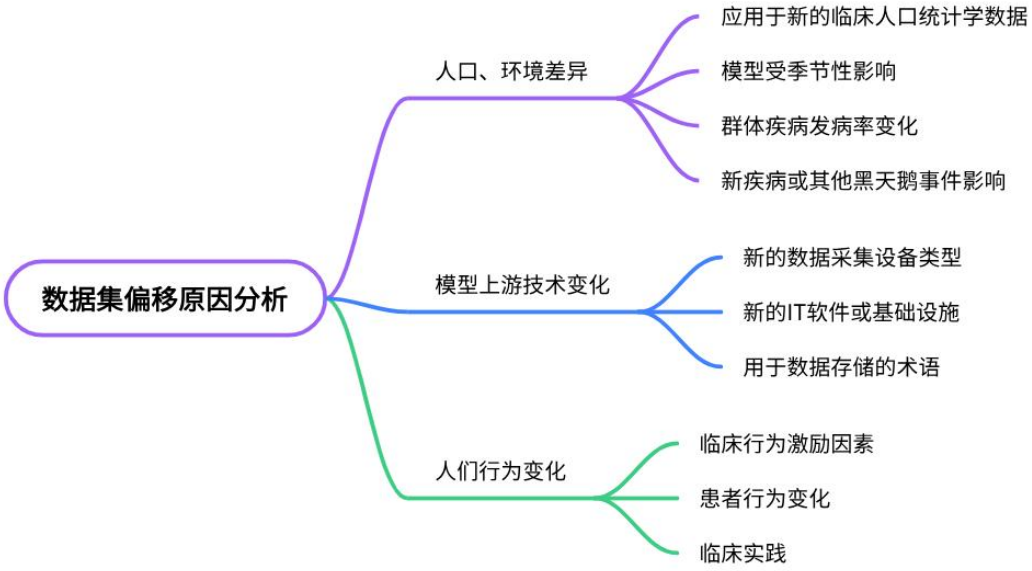



在一些情况下，多中心临床数据集具有不同的遗传、环境和种族分布，学习效果依赖于特定的人口、环境等因素，不同中心的临床数据集的学习效果会有显著差异[12,18]。一些大型临床数据集，特别是从不同临床研究中心收集的数据集，包括来自各个地理位置和中心特定特征的大量参与者[13,42]。训练数据集与测试数据集之间人口和环境的差异可能导致它们在模型中与临床结果的关联性不同，从而必然降低模型的潜在临床应用性能，例如结果预测[10,40]。一个典型的数据集偏移的例子是基于白人中产阶级样本调研产生的数据集在应用于亚洲人口、非洲人口时会出现数据集偏移导致的结果偏差问题。这可能源于心理科学一直以来的工作中的一个基本假设，即白人和中上阶层样本是中立且具有普遍性的，但实际上它们只代表了全球人口的一个狭窄部分[11]。类似地在学术或专业环境中开发的模型可能无法适用于社区使用[18]。

科技进步或技术应用中的变化，可能会导致某些依赖于现存科学技术而生成的数据集与应用了新的技

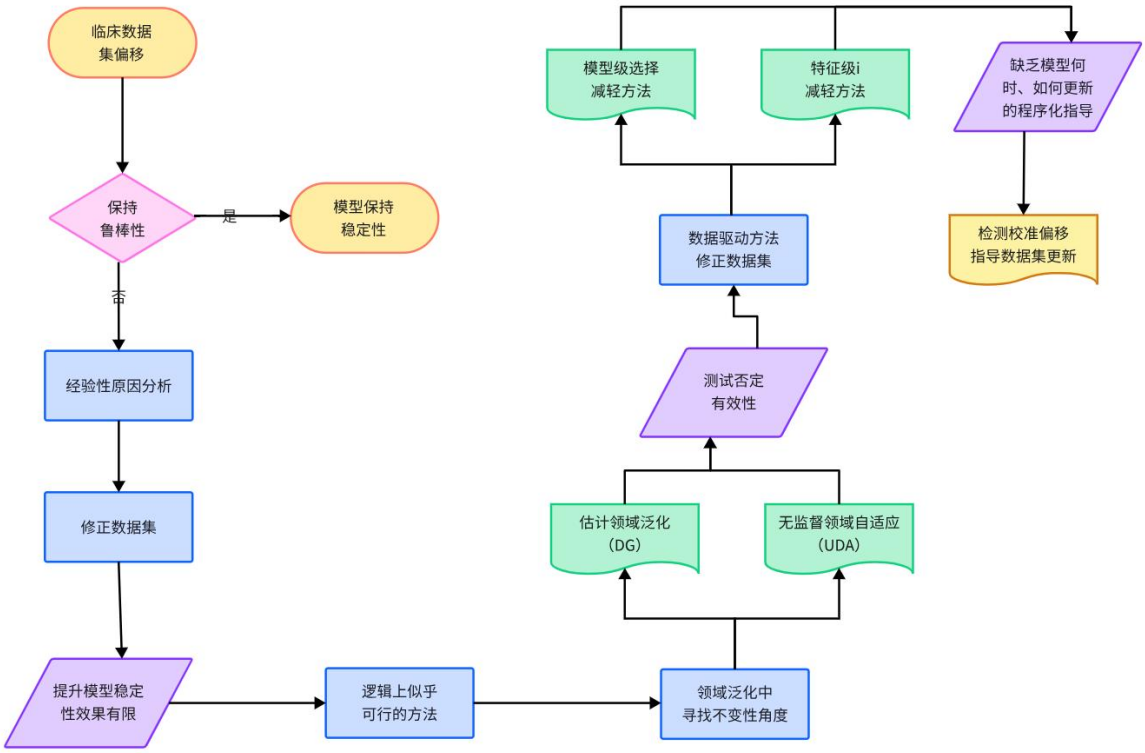
术进步而生成的数据集在某些特征上产生偏差，从而使得依赖于数据集中的相关特征的模型产生偏差甚至失效[17,18]。例如，开发用于预测髋部骨折的 CAD 模型依赖于特定的 X 光扫描仪型号和技术人员，而 X 光扫描仪的高敏感性肌钙蛋白检测方法的采用使得图像数据集产生偏移，进而改变了对可检测到的肌钙蛋白水平的临床解读；使用 ICD-9 码定义诊断的模型在采用了 ICD-10 码的医院中可能不准确，因为定义存在差异[18,40]等。

人们行为的变化也会影响数据集的生成，从而产生数据集偏移，进而影响模型的学习结果[27,39]。具体来说，临床行为激励因素、患者行为变化、临床实践都可能影响数据集的形成，从而产生数据集偏移[18,40]。临床激励因素如对败血症相较于其他死因的差异性报酬，导致败血症的诊断明显增加；患者行为变化如在高知名度名人被诊断后，患者可能会寻求具有较少或没有症状的诊断评估；临床实践如外科皮肤标记，在不同的临床环境中医嘱设置、标记时机等的差异，可能会严重影响预测模型的输出结果以及皮肤科分类器的准确性[18,39]。




**3. 临床数据集偏移的不良影响**

用机器学习模型进行数据分析对疾病状态进行预测和诊断的核心是泛化，理想的情况是在模型应用于数据集中不存在的新情况时，模型的预测能够准确无误，特别是模型对数据生成方式的扰动（即数据偏移）具有稳定性[17,27]。然而，虽然大部分时候，基于机器学习模型的泛化是具有鲁棒性的，有时候由于数据集偏移，模型对预先指定类型的转移不再具有鲁棒性，将其应用于新的数据集时，性能下降从而影响到模型的安全性[25,26]。




临床实践中，时间数据集漂移的结果是机器学习模型用于决策支持的一道障碍[31,42]，而现有的对于提升模型鲁棒性和稳定性的方法有限，当前的方法主要是经验性的。如 2018 年研究者在一个数据集上训练和验证了一个肺炎诊断模型，并比较了将其应用于来自新医疗中心的数据时的性能[17,20]；2009 年研究者在 2001 年从一个医院收集的数据上训练了一个死亡预测模型，并在随后的几年中收集的数据上评估了其性能[17,21]。模型低稳定性的原因通常在于模型方法很难捕捉到从不同临床环境中收集的数据集中特定中心特征与结果之间的变化关系，因此评估模型稳定性和鲁棒性的方法面临的困难主要在于确定性能下降的确切原因，即确定是否对特定的转移具有鲁棒性[10,27]。

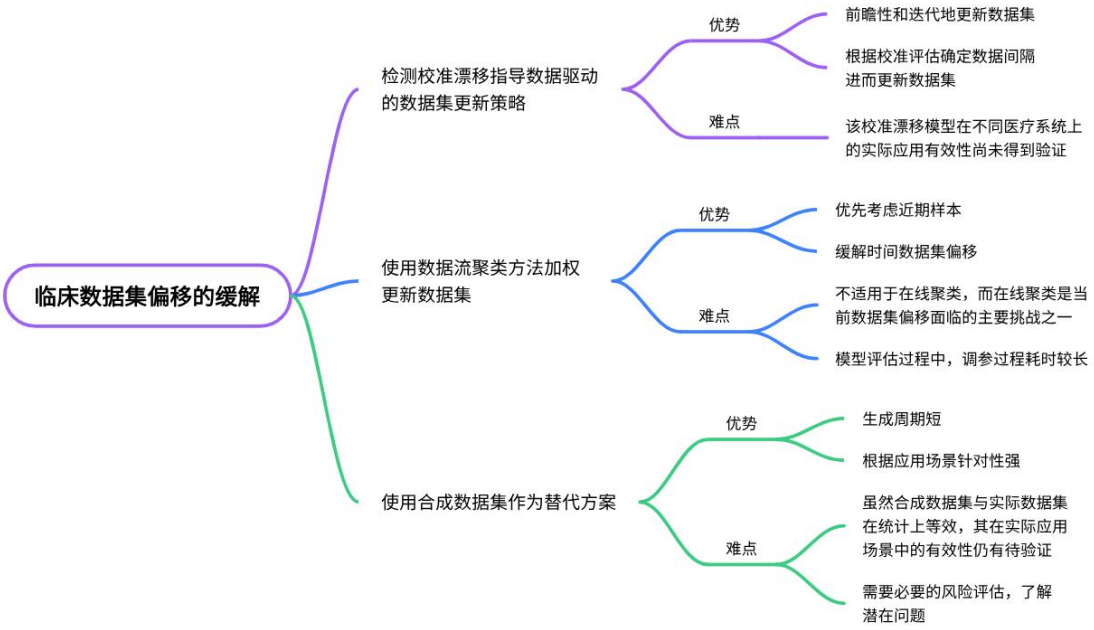
一些逻辑上似乎行之有效的方法在缓解时间数据集偏移的问题上也不具有常识上应有的效果[31,37]。2022 年，科学家从领域泛化中寻找不变属性的角度，提出了通过估计领域泛化（DG）和无监督领域自适应（UDA）中的不变性属性来学习稳健模型的算法，并在针对重症监护病房和急诊科的数据库 MIMIC-IV 中 53150 名患者的数据上进行了充分的测试[38]，用基准（经验风险最小化 ERM）、DG、UDA 三种算法进行试验，最终否定了 DG、UDA 算法主动缓解数据集偏移的有效性[37]。

目前不同的研究者对于通过评估模型稳定性和鲁棒性而减轻数据集偏移，从不同的角度提出了不同的解决方法[32,33]，然而这些方法大多是经验和数据驱动的，缺乏程序化确定模型何时以及如何进行更新的方法[39,42]。2021 年研究者从数据集的角度提出了模型级、特征级的数据集偏移识别和减轻方法，模型级选择减轻方法分为概率校准、模型重拟合、模型更新等，主要涉及使用统计检验来选择一组策略中的最佳减轻策略；而特征级减轻方法在模型拟合之前处理特征，并分为数据驱动、专业知识驱动两种驱动方式[31,34]。这些不同的数据集偏移识别和减轻方法都在一定范围内能够成功保持校准，但在保持判别方面并不一致成功[35,36]。选择最佳的缓解策略取决于数据集转换的类型和严重程度，然而还没有一种标准方法将特定环境中的数据集转换类型映射到特定的缓解策略，即无法有效的确定模型何时以及如何进行数据驱动的更新策略以及重新拟合[39,42]。因此仅仅通过数据驱动角度分析数据集偏移的识别和减轻方法仍然有局限[31,32]。

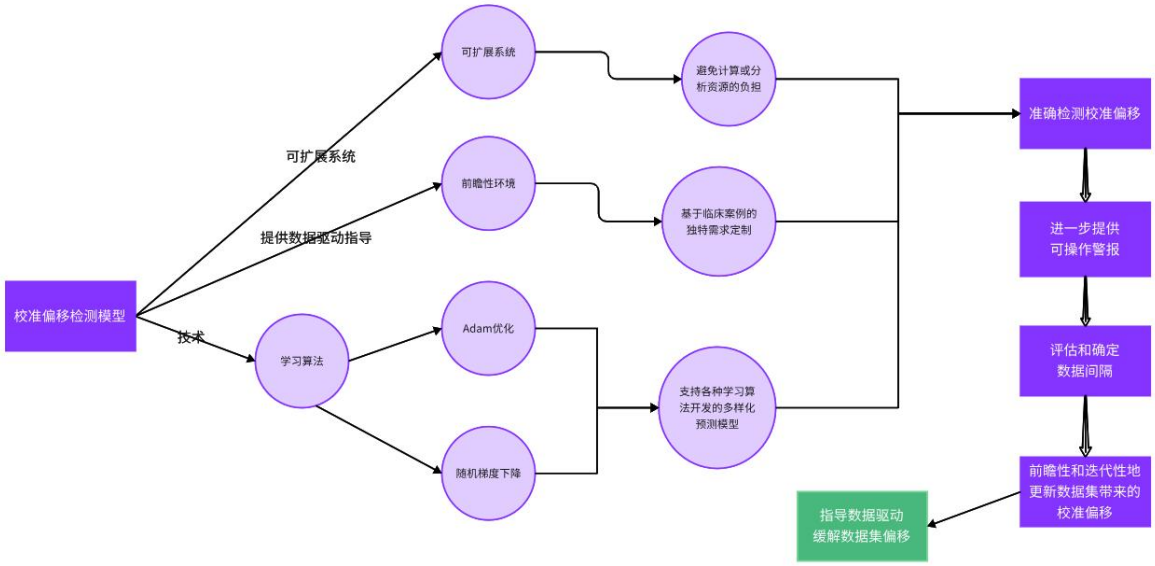
## 4. 临床数据集偏移的缓解

针对数据集偏移的不良影响，数据驱动的更新策略可以通过根据观察到的性能漂移的时间、程度和形式来定制更新，解决定期重塑的限制[45]。这一类策略需要确定模型应该如何以及何时进行更新的方法[43]。

目前研究者已经提出可以解决前一个问题的测试程序和模型，并提供数据驱动的指导，以在重塑和重新校准之间进行选择[43,44,45]。应用这样的测试建议的定期更新，比起预定义的重塑方法，可以随着时间的推移提高校准精度[46]。




2020 年研究者设计了一个可扩展的系统，用于前瞻性的生产环境中，通过提供数据驱动的指导，监控一组风险预测模型的更新时机和用于模型更新的数据[42,45]。模型实现基于 Adam 优化、随机梯度下降等技术的动态校准曲线方法，支持使用各种学习算法开发的多样化预测模型，避免计算或分析资源的负担并且基于临床使用案例的独特需求定制[43,46]。在模拟和案例研究中，能够准确地检测到校准漂移，并进一步提供可操作的警报，包括可能适用于模型更新的最近数据窗口的信息[42,44]。该模型能够基本达到随着时间的推移，前瞻性和迭代地更新数据积累的模型校准评估；评估连续校准评估中的显著漂移和确定应该用来更新现有模型的数据间隔[42,45]。然而该校准漂移模型在不同医疗系统上的实际应用有效性尚未得到验证，因此通过检测校准漂移指导数据驱动缓解临床数据集偏移的实际可行性仍然有待验证[42]。




另外一种解决数据集偏移的角度是将数据集偏移视为静态聚类任务中的一个问题，将数据集理解为一个长时间段内收集的数据，使用数据流聚类的相关方法缓解数据集偏移[52,53]。2022 年，科学家提出了一种

包含时间相关加权过程的 k-means 变体，即时间加权核 k-means，并通过引入有序加权平均（IOWA）运算符具体实现[50,51]。加权过程充当逐渐遗忘的机制，在聚类算法中优先考虑最近的样本而不是过时的样本，定期更新聚类算法获得的数据结构，以缓解时间数据集偏移[19,50,51]。计算实验表明，加权核 k-means 在处理数据集时优先考虑最近的样本，使用时间加权核 k-means 减轻数据集偏移在不断变化的环境中具有潜力[50,51]。然而应用这种方法解决临床数据集偏移仍然存在难点和局限，首先，它不适用于在线聚类，而在线聚类是当前面临数据集漂移的主要聚类挑战之一；其次，在模型评估过程中，需要调整大量参数，可能面临耗时较长的问题[51]。

当训练数据集和测试数据集之间存在的特征差异无法通过上述方法缓解，导致测试数据集对于模型不可用的情况，可以考虑使用合成数据集作为替代方案[48,50]。2022 年，研究者已经成功用机器学习开发一种生成与真实临床数据集在统计上相等的逼真合成数据集，用于验证医疗保健应用程序[47,48,49]。这些使用机器学习中的生成对抗网络、Tensorflow 框架生成的数据集和真实数据集相比，有生成周期短的优点，不需要经历冗长的准备和批准过程，因此可以根据现实应用场景针对性生成，不会出现实际数据集的时间数据集偏移问题，因此合成数据集是为了解决数据集偏移问题的一个潜在替代方案[47,50]。尽管合成数据集在统计上与真实数据集等效，但还需要更多的工作来验证它在缓解临床数据集偏移应用上的作用，并且需要必要的风险评估，以了解潜在问题和如何解决这些问题[49,50]。

**参考文献：**


1. Ying Yu, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang, Clinical Big Data and Deep Learning: Applications, Challenges and Future Outlooks, Big Data Mining and Analytics, December 2019

2. Cristina Palmero, Maria Inés Torres, Anna Esposito, Sergio Escalera, Guest Editorial: Special issue on computer vision and machine learning for healthcare applications, Pattern Analysis and Applications, 2022

3. E. Jenifer Sweetlin and S. Saudia, A New Hybrid Feature Selection Sequence for Predicting Breast Cancer Survivability Using Clinical Datasets, Intelligent Automation & Soft Computing, 2023

4. Meenakshi Sharma*, Sanjay Kumar Singh, Prateek Agrawal and Vishu Madaan, Classification of Clinical Dataset of Cervical Cancer using KNN, Indian Journal of Science and Technology, July 2016

5. Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, A patient-centric dataset of images and metadata for identifying melanomas using clinical context, Scientific Data, 2021

6. Fiona Bach MBBS MRCOG, Philip Toozs-Hobson MD FRCOG,*Jeremy Purnell BSc (Hons) MSc， Making the best use of clinical datasets: with examples from urogynaecology, The Obstetrician & Gynaecologist, August 2022

7. Sergey E. Golovenkin, Jonathan Bac, Alexander Chervov, Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data, GigaScience, 2020

8. Wei-Qi Wei, Chao Yan, Monika Grabowska, Leveraging Generative AI to Prioritize Drug Repurposing Candidates: Validating Identified Candidates for Alzheimer's Disease in Real-World Clinical Datasets, Research Square, July 2023

9. Mahmoud Malas, MD, MHS; Isibor Arhuidese, MBBS, MPH; Umair Qazi, MD, MPH; James Black, MD; Bruce Perler, MD, MBA; Julie A. Freischlag, MD, Perioperative Mortality Following Repair of Abdominal Aortic Aneurysms Application of a Randomized Clinical Trial to Real-World Practice Using a Validated Nationwide Data Set, Original Investigation Research

10. Jiebin Chu a , Jinbiao Chen a , Xiaofang Chen a , Wei Dong b , Jinlong Shi c , Zhengxing Huang, Knowledge-aware multi-center clinical dataset adaptation: Problem, method, and application, Journal of Biomedical Informatics, September 2020

11. Luz Maria Alliende*, Teresa Vargas, and Vijay Anand Mittal, Representation Challenges in Large Clinical Datasets, Schizophrenia Bulletin, 2023
12. Jonathan Waringa, Charlotta Lindvallc, Renato Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artificial Intelligence in Medicine, February 2020
13. David Riañoa, Mor Pelegb, Annette ten Teije, Ten years of knowledge representation for health care (2009–2018): Topics, trends, and challenges, Artificial Intelligence in Medicine, September 2019
14. Farah Shamout , Tingting Zhu , and David A. Clifton, Machine Learning for Clinical Outcome Prediction, IEEE Reviews in Biomedical Engineering, vol 14, 2021
15. Livia Faes*, Siegfried K Wagner*, Dun Jack Fu, Xiaoxuan Liu, Edward Korot, Joseph R Ledsam, Trevor Back, Reena Chopra, Nikolas Pontikos, Christoph Kern, Gabriella Moraes, Martin K Schmid, Dawn Sim, Konstantinos Balaskas, Lucas M Bachmann, Alastair K Denniston, Pearse A Keane, Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study, Lancet Digital Health, September 2019
16. Garima Suman a, Anurima Patra a, Panagiotis Korfiatis a, Shounak Majumder b, Suresh T. Chari c, Mark J. Truty d, Joel G. Fletcher a, Ajit H. Goenka, Quality gaps in public pancreas imaging datasets: Implications & challenges for AI applications,vol 21, August 2021
17. Adarsh Subbaswamy, Suchi Saria, From development to deployment: dataset shift, causality, and shift-stable models in health AI, Biostatistics, vol 21, 2020
18. Samuel G. Finlayson, Ph.D., Adarsh Subbaswamy, B.S., Karandeep Singh, M.D., M.M.Sc., John Bowers, B.A., The Clinician and Dataset Shift in Artificial Intelligence, The new England Journal of Medicine, July 2021
19. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D. (editors), Dataset Shift in Machine Learning. Cambridge, 2009, pp. 131–160.
20. Zech, J. R. Badgeley, M. COSTA, A. B. Titano, J. J. and Oermann, E. K., Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Medicine, 2018 vol 15
21. Zhang, K., Scholkopf, Muandet, K. and Wang Z., Domain adaptation under target and conditional shift., Proceedings of the 30th International Conference on Machine Learning,2013
22. Jose García Moreno-Torres, José A. Sáez, and Francisco Herrera, Study on the Impact of Partition-Induced Dataset Shift on k-fold Cross-Validation, IEEE Transactions on Neural Networks and Learning Systems, 2012 vol23
23. José A. Sáez and José L. Romero-Béjar, Impact of Regressand Stratification in Dataset Shift Caused by Cross-Validation, Mathematics, July 2022
24. Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In Advances in Neural Information Processing Systems,2015 pages 1576–1584.
25. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems detected by uncertain probabilities. Management Science, 2013 59(2):341–357.
26. Bertsimas, D., Gupta, V., and Kallus, N. Datadriven robust optimization. Mathematical Program, 2018 167(2):235–292.
27. Adarsh Subbuswamy, Suchi Saria, Evaluating Model Robustness and Stability to Dataset Shift, Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021
28. T. Willem, S. Krammer, L.E. French, D. Hartmann, T. Lasser, A. Buyx, Risks and benefits of dermatological machine learning health care applications—an overview and ethical analysis, JEADV, 2022 vol 35
29. Emad A Mohammed, Behrouz H Far and Christopher Naugler, Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends, BioData Mining, 2014 vol 7
30. Shounak Pal, Baidyanath Biswas, Rohit Gupta, Ajay Kumar, Shivam Gupta, Exploring the factors that affect user

experience in mobile-health applications: A text-mining and machine-learning approach, Journal of Business Research
31. Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Jose Posada, Scott Lanyon Fleming, Catherine Aftandilian, Nigam Shah, Lillian Sung, Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine, Applied Clinical Information, 2021 vol.12
32. Challener DW, Prokop LJ, Abu-Saleh O. The proliferation of reports on clinical scoring systems: issues about uptake and clinical utility. JAMA 2019;321(24):2405–2406
33. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18
34. Sendak MP, Balu S, Schulman KA. Barriers to Achieving Economies of Scale in Analysis of EHR Data. A Cautionary Tale. Appl Clin Inform 2017;8(03):826–831
35. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012;45(01):521–530
36. Challen R, Denny J, Pitt M, Gompels L, Edwards T, TsanevaAtanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28(03):231–237
37. Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, CatherineAftandilian, Nigam Shah & Lillian Sung, Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine, Scientific Reports, 2022
38. Wilson, G. & Cook, D. J. A survey of unsupervised deep domain adaptation. ACM Trans. Intell. Syst. Technol. 11, 1–46 (2020)
39. Suchi Saria, Adarsh Subbaswamy, Tutorial: safe and reliable machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, January 29–31, 2019. New York: Association for Computing Machinery, 2019.
40. H Echo Wang, Matthew Landers, Roy Adams, Adarsh Subbaswamy, Hadi Kharrazi, Darrell J Gaskin, Suchi Saria, A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models, Journal of the American Medical Informatics Association, Volume 29, Issue 8, August 2022
41. Subbaswamy, Adarsh, and Suchi Saria. "Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms." UAI. 2018.
42. Sharon E. Davis, Robert A. Greevy Jr., Thomas A. Lasko, Colin G. Walsh, Michael E. Matheny, Detection of calibration drift in clinical prediction models to inform model updating, Journal of Biomedical Informatics, October 2020
43. Ruben Amarasingham, Rachel E. Patzer, Marco Huesch, Nam Q. Nguyen, and Bin Xie Implementing electronic health care predictive analytics: considerations and challenges, Health Aff. 33 (7) (2014) 1148–1154.
44. Matheny, Michael, et al. "Artificial intelligence in health care: The hope, the hype, the promise, the peril." Washington, DC: National Academy of Medicine 10 (2019).
45. M.E. Matheny, D. Whicher, Israni S. Thadaney, Artificial Intelligence in health care: a report from the national academy of medicine, JAMA (2019).
46. K.G. Moons, A.P. Kengne, M. Woodward, et al., Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker, Heart 98 (9) (2012) 683–690.
47. Theodoros N Arvanitis, Sean White, A method for machine learning generation of realistic synthetic datasets for validating healthcare applications, Health Informatics Journal, 2022
48. Buczak AL, Babin S and Moniz L. Data-driven approach for creating synthetic electronic medical records, BMC Med Inform Decis Mak, 2010
49. Walonoski J, Kramer M, Nichols J, et al. An approach, method, and software mechanism for generating

synthetic patients and the synthetic electronic health care record, J Am Med Inform Assoc, 2018
50. Buczak AL, Babin S and Moniz L. Data-driven approach for creating synthetic electronic medical records, BMC Med Inform Decis Making 2010; 10: 59
51. Sebastián Maldonadoa, Ramiro Saltos, Carla Vairetti, José Delpiano, Mitigating the effect of dataset shift in clustering, Pattern Recognition, 2023
52. K. Song, X. Yao, F. Nie, X. Li, M. Xu, Weighted bilateral k-means algorithm for fast co-clustering and fast spectral clustering, Pattern Recognition, 2021
53. Y. Zheng, R. Hu, S. fu Fung, C. Yu, G. Long, T. Guo, S. Pan, Clustering social audiences in business information networks, Pattern Recognition, 2020