大家好！我叫陈苇远，我的研究课题是临床数据集偏移。

我将从临床数据集应用现状、临床数据集偏移及其原因分析、临床数据集偏移的不良影响和临床数据集偏移的缓解四个角度介绍展开我的课题研究。

首先是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，或由特定疾病医疗学会牵头、基于医院登记数据，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。

这样的数据集有公开大型肿瘤数据库SEER、美国国家癌症数据库NCDB、美国家庭营养和健康数据库NHANES、病原与宿主相互作用的基因数据库PHI等，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。

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。其中最普遍的医疗数据集是医学图像。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。临床笔记，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，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。

通过多样的分析方法处理临床数据集，与现实中临床诊断结合可以达到更准确的诊断结果和更好的治疗效果。目前已经实现的临床数据集方法有通过传统概率和统计思想中的概率分布和拟合、贝叶斯网络、逻辑斯蒂回归等方法，近年来也出现了许多新的分析思路和角度，如2021年科学家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。

临床数据集分析常常应用于疾病分析和辅助诊断，具体来说，基于临床数据集的分析可以帮助的促进疾病相关病理分析、阶段预测与辅助诊断等。目前，在乳腺癌、宫颈癌、泌尿妇科学、心肌梗死与糖尿病、黑色素瘤等疾病都已经实现了基于临床数据集分析的疾病分析和辅助诊断。

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。引起临床数据集偏移的原因有技术变化、人口和环境变化或人们行为变化等。

首先，我们来看看数据集偏移的定义：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，因而导致基于训练数据集得到的模型应用于测试数据集时表现不佳的现象。

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。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。

接下来介绍的第一类临床数据集偏移是人口、环境造成的数据集偏移。

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。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。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。尤其是因为目前大部分公开数据集的调查对象主要是发达国家的白人中产阶级，基于这些数据集训练的模型在应用于发展中国家、其他人种的测试数据集时容易导致数据集偏移。

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。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，而相关新技术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，因为定义存在差异等。

人们行为的变化，如临床行为激励因素、患者行为变化、临床实践，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。例如，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。

由以上论述，临床数据集偏移是普遍存在的，数据集偏移可能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。

第三部分主要读PPT

那么通过什么样的方法才能够缓解临床数据集偏移呢？我接下来主要介绍三类方法：

具体来说，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。

第四部分主要读PPT