# Personalized Disease Prediction Using a CNN-Based Similarity Learning Method

使用基于CNN的相似性学习方法进行个性化疾病预测

# 摘要

预测患者发生某些疾病的风险是医疗保健领域的重要研究课题。个性化预测建模重点针对个体患者建立特定模型，与利用全体人群训练的全球模型相比，显示出其利用异构健康数据的优势。个性化预测模型使用来自类似患者队列的信息，以捕捉特定的特征。根据个人历史记录准确识别和排列患者之间的相似性是个性化建模的关键步骤。由于缺乏适当的矢量表示，不能直接用电子健康记录（EHR）来测量患者的相似度，这些记录是不规则采样的且具有不同的患者访问长度。在本文中，我们构建了一个新的时间融合CNN框架，以同时学习患者表示并测量两两相似度。与传统的CNN相比，我们的时间融合CNN不仅可以学习本地时间关系，而且还可以学习每个时间间隔的贡献。与相似性学习过程一起，使用概率分布的输出信息来对相似的患者进行排序。利用相似性得分，我们执行个性化的疾病预测，并比较不同的向量表示和相似性学习度量的影响。

# 1.简介

准确预测疾病在公共卫生中发挥着重要作用，尤其是在早期阶段，可以让患者及时采取预防措施。随着电子健康记录（EHR）数量和可用性的不断增长，疾病进展和分析的预测建模任务越来越受到研究人员的关注。 EHR数据按患者访视时间顺序排列，每次访视表示为一组高度临床事件。由于其嘈杂，不规则和异构性质，与标准数据挖掘任务相比，采矿电子病历尤其具有挑战性。传统的疾病预测方法是单尺寸适合所有模型[1]。也就是说，利用所有可用的训练数据建立一个全球模型，然后用这个模型预测每个患者的疾病风险。应用全尺寸适合所有模型的好处是它可以捕捉整个培训人群的整体信息。但是，患者可能有不同的表型，不同的医疗条件等。使用全球模型可能会错过某些对个别患者重要的特定信息。因此，针对每位患者建立有针对性的针对患者的模型对于个性化医疗来说是紧急和重要的。

最近的研究[2-5]表明，个性化模型可以提高全局模型的预测性能。个性化预测的一般框架包含两个阶段：（1）测量患者之间的相似性;（2）使用他/她的类似队列为每个患者建立单独的模型。这个框架是由人类医生的工作过程所驱动的，即在检查或回顾了具有类似疾病或症状的被诊断患者后，医生会仔细做出决定。如果医生可以找到类似的患者，成功治愈该患者的可能性可能会提高很多。医疗数据集中已经提出了许多相似性学习方法[6-10]。然而，这些模型是为手工向量表示而开发的，如人口统计数据或平均数值，而不考虑来自不同访问的时间信息。对于纵向EHR数据，由于患者的不规则访问和不完整记录，患者访问次数差异很大。上述学习指标不能直接应用于纵向数据，因为每个患者的历史记录不会自然形成可比较的矢量。因此，衡量患者相似性的关键挑战之一是在不损失其历史信息的情况下为每位患者推导出有效的表示

**最近，深度学习方法在患者代表学习[11-18]，如自动编码器，递归神经网络（RNN）和卷积神经网络（CNN）中被广泛采用并迅速发展。 在[19]中，CNN已经显示出它在衡量患者相似性任务方面的卓越能力。 然而，传统的CNN架构的一个缺点是它不能充分利用EHR的时间和上下文信息来预测疾病。 因此，同时建模电子病历数据的时间性和内容更具挑战性。**

**针对上述挑战和问题，本文旨在解决个性化预测中的以下关键问题：如何建立一个模型，从历史记录中准确测量患者的相似度，以及如何建立一个准确的个性化预测模型学到的相似之处。为了实现这些目标，我们首先设计了一种新颖的时间融合CNN框架来解释跨越不同时间间隔的时间性。利用所提出的框架，我们可以为每个患者生成矢量表示。基于学习到的患者表示，然后引入匹配度量以获得相似性表示。考虑到实际意义，我们为框架结构增加了一个坚实的对称约束。这种相似性学习框架是端到端的，它同时学习患者表征和配对相似性。由于一对患者之间的相似概率表明两个患者发生同一疾病的风险水平，因此我们将其用作评分来排列患者之间的相似性。最后，我们使用他/她的类似队列为每位患者建立个性化模型。总之，我们的贡献如下：**

我们构建了一个框架，共同学习患者EHR表示和配对相似性，而不需要手工制作的特征聚合。 通过该框架，可以同时优化表示和相似性学习的参数，从而获得更高的准确性。

我们开发了一个时间融合CNN模型，不仅保留了相邻访问的局部时间性，还考虑了不同时间间隔的全局贡献。

我们的实验结果表明，我们的相似学习框架可以为患者的历史信息学习更好的表示向量，并提高疾病预测的准确性。 与其他常用策略相比，基于加权抽样的个性化模型提高了个性化预测的准确性。

# 4.总结

医疗保健中的个性化预测建模旨在发现个体患者的独特特征，并建立有针对性的患者特定预测。 在本文中，我们提出了一种基于时间融合CNN的框架来配对测量患者相似性，并使用三种方法来执行个性化疾病预测。 实验结果表明，我们的时间融合CNN可以更好地表示纵向EHR序列，并且我们的端到端相似性框架优于广泛使用的距离度量。 通过相似性排序，我们可以执行三种个性化预测方法，并显示加权采样可以提供稳定和高精度