在生存分析中使用基于地理位置的公共卫生特征

Navid Seidi
nseidi@mst.edu 密苏里科技大学 美国密苏里州罗拉

Ardhendu Tripathy
astripathy@mst.edu 密苏里科技大学 美国密苏里州罗拉

Sajal K. Das
sdas@mst.edu 密苏里科技大学 美国密苏里州罗拉

摘要

感兴趣事件发生前的时间通常使用生存分析方法进行建模，该方法根据输入特征估算生存评分。近年来，人们重新开始关注利用神经网络等现代工具，开发更准确的个性化医疗事件发生时间预测模型。更高质量的特征和更频繁的观察能够提升对患者的预测效果，但将患者的地理位置相关公共卫生统计数据纳入个体预测所产生的影响尚未得到研究。本文提出一种对生存分析模型的补充性改进，即将公共卫生统计数据纳入输入特征。我们表明，加入基于地理位置的公共卫生信息，能够显著提升在监测、流行病学和最终结果（SEER）数据集上评估的一致性指数；该数据集包含全国范围的癌症发病数据。这种提升在标准 Cox 比例风险模型和最先进的深度生存机器模型中均成立。我们的结果表明，基于地理位置的公共卫生特征在生存分析中具有实用价值。

CCS 概念

• **计算数学**！ **生存分析**； • **计算方法**！ **机器学习**； • **应用计算**！ **医疗保健信息系统**； • **信息系统**！ **数据挖掘**；

关键词

生存分析，机器学习，深度神经网络，智慧健康，医疗信息物理系统

1 引言

长期以来，研究某种关注事件随时间发生的概率一直是多个 fi 领域的热门课题。在健康信息学中，人们利用它来探究某种疾病完全康复（痊愈）或死亡的可能性 [21, 28, 30]。在商业研究中，人们用它来预测消费者再次光顾商店所需的时间 [22]，或游客在休闲区域停留的时间 [4]。社会学中探讨的是新婚夫妇随时间推移离婚或迎来fi第一个孩子的可能性 [12, 14, 35, 39]。犯罪学分析人员利用它来确定刚获释的囚犯再次实施 fi首次违法行为所需的时间 [6, 38]。统计学采用生存分析方法研究事件发生时间这一课题。

在医疗保健领域，研究人员希望利用过去开展研究所得的数据集，建立一个模型来预测个体距离某一事件发生的时间。这些数据集包含有关每个个体的信息，例如健康状态参数、每个个体的观察期长度，以及所需事件发生的时间。该事件可以定义为个体总体健康状况的任何变化，例如疾病诊断、疾病康复、疾病分期变化、出院或死亡。模型输入参数（特征）的选择会因所研究疾病的类型而异，涵盖从生命体征到疾病严重程度指标等内容。

本研究的主要目标是探究基于地理位置的公共卫生特征如何提升对个体患者的预测效果。正如 [20] 所述，这些特征可以从公共卫生数据库中提取。以往对医疗结果进行基于位置的分析，主要侧重于对 fi 拟合参数的空间依赖进行建模 [2, 16, 46]。本文并未显式地对空间依赖进行建模，而是将基于位置的特征纳入输入。我们假设，加入这些基于地理位置的公共卫生特征能够提升所训练模型的预测质量。我们证明，在监测、流行病学和最终结果（SEER） [18] 数据集上训练的两个不同模型均支持这一假设。我们通过以下步骤检验该假设。首先，从其中一个监测、流行病学和最终结果（SEER） [18] 数据集中提取预期生存率特征。其次，将该特征添加到 SEER 中的一个生存数据集，并使用一致性指数（C-Index） [13] 指标评估其对该数据集生存预测的影响 ff。评估过程中采用了两个重要模型：I）Cox比例风险模型（CoxPH） [7] ，这是最受欢迎且最基础的工具之一；II）深度生存机器（DSM） [30]，这是目前最新的深度神经网络模型。以下将介绍实现这一构想的步骤。

(1) 我们使用了监测、流行病学和最终结果（SEER） [18]数据集中的发病率数据集。由于该数据集唯一的地理特征是居住州，我们尝试将其与具有相似属性的预期生存和人口数据集（这两者均包含在 SEER 数据集组中）结合，以丰富该数据集。此阶段最终汇编了十个数据集，涵盖 2000年 至 2017年 期间美国十个州的癌症患者。这些州包括加利福尼亚州、康涅狄格州、佐治亚州、夏威夷州、艾奥瓦州、肯塔基州、路易斯安那州、新泽西州、新墨西哥州和犹他州。在现有的发病率 SEER 数据集中，新增了两个州级特征（预期生存率和 SEER 登记处）以及一个国家级特征（每个人的居住州）。

(2) 我们使用包含来自十个州的所有患者的数据集，对 CoxPH 方法和 DSM 模型进行了训练和测试。

(3) 一致性指数（C-Index） [13] 已选用于评估阶段。一致性指数（C-Index）是一种性能指标，用于衡量预测生存时间与观察到的生存时间之间的一致程度。

(4) 为检验关于在生存回归中加入基于地理位置的公共卫生特征是否具有显著性的假设，我们开发并实施了一种 t检验方法.

在所有州中，使用配对 t 检验发现，观察到的提升在统计学上均具有显著性。

本文组织如下：第 2 节回顾生存分析方面的先前研究，第 3 节阐述问题。第 4 节概述本研究的理论层面以及数据集的构建方式，第 5 节讨论实验阶段和验证。第 6 节介绍本项目对计算机科学和医疗保健行业的研究人员可能具有的优势，并讨论其潜在应用。第 7 节给出结论，第 8 节探讨未来研究。第 9 节说明如何访问项目 GitHub 仓库中包含的所有代码和 fi les。

2 背景与相关工作

生存分析是一种统计技术，用于分析感兴趣事件发生前的时间，例如死亡、机械部件故障或疾病复发。在生存分析中，生存函数用于估计个体在给定时间段之后仍未经历感兴趣事件的概率。Cox比例风险模型（CoxPH） [7] 是生存分析中广泛使用的一种技术，可用于估计风险率，即事件在特定时间发生的概率，并通过协变量进行建模。CoxPH模型假设风险率与基线风险函数成比例，并受一组协变量的影响，这些协变量对数线性地影响对数风险率。该模型是半参数模型，意味着它不对基线风险函数的函数形式作出假设，因此 fl具有灵活性，可适用于各种数据。近年来，深度神经网络模型在生存分析领域备受关注，因为它们能够对复杂关系进行建模，并捕捉数据中的非线性关系。

2.1 深度神经网络模型

传统生存分析模型（如 Cox比例风险模型）假设对数风险函数呈线性，因此在捕捉数据中复杂的非线性关系方面能力有限。相比之下，深度神经网络模型为建模此类复杂关系提供了强大的工具，并有望提高生存预测的准确性。近年来，研究人员开发了新的神经网络模型，并在 SEER、治疗预后、偏好、结果和风险研究（SUPPORT）以及乳腺癌分子分类国际联盟（METABRIC）等真实世界数据集上对其进行了评估，以验证这些模型在实际应用中的有效性。

在**DeepSurv** [21], 中，作者将深度学习技术与生存分析相结合，使模型能够捕捉对数风险函数中的复杂非线性关系。在他们的评估中，作者使用上述数据集和模拟数据来展示模型的性能和泛化能力。结果表明，**DeepSurv** 优于传统的Cox比例风险模型，并展现出在各个医学领域改善风险分层fi和个体化治疗建议方面的良好潜力。

DeepHit 方法 [27, 28] 提出了一种用于生存分析的新型深度学习框架，可同时对个体风险和竞争风险进行建模。该方法能够更准确地预测事件发生的时间和类型，解决了传统模型常常忽视的一个关键问题。作者使用多个数据集（包括器官共享联合网络（UNOS）数据库和 METABRIC）对**DeepHit**进行了评估，以展示其有效性和适用性。**DeepHit** 的表现优于传统生存分析方法，为深入理解复杂的事件发生时间数据和改进决策提供了有力工具。

深度生存机器（DSM） [30] 融合了深度学习与非参数方法的优点，并消除了 DeepHit 的一项限制：事件时间必须属于离散集合。DSM 能够独立学习潜在的风险模式，并在不受参数假设约束的情况下识别协变量之间的复杂关系。作者使用多个基准数据集评估了 DSM 的性能，包括 SUPPORT研究、SEER 和 METABRIC，凸显了其在各种情境下的有效性。最终，在生存分析中存在竞争风险时，DSM 的表现优于传统生存分析方法和其他深度学习模型。

2.2 空间统计模型

除了采用表达能力更强的风险函数模型外，研究人员还研究了生存预测模型之间的空间依赖。在 [45], 中，他们介绍了地理加权Cox回归模型的开发，用以解决地理分布稀疏的生存数据分析问题；由于样本量较小，无法按位置对数据进行简单分层。作者提出，根据研究对象与各位置之间的距离对其进行加权，并依据托布勒地理学第一定律 fi，通过最大化局部部分似然，为每个位置估计回归系数 fi。他们讨论了选择适当的加权函数、带宽和距离度量的重要性，并建议使用图距离，以实现稳健估计和更简便的实施。作者没有采用基于预测的带宽选择方法，而是提出一种基于似然的方法，利用经过修改的竹内信息准则（TIC） [41] 来考虑偏差—方差权衡 ff。该方法旨在更清晰地了解地理分布稀疏的生存数据中响应变量与协变量之间的关系。尽管该方法具有创新性，但由于高度依赖精确的地理坐标，考虑到人们隐私方面的限制以及《健康保险流通与责任法案》（HIPAA） [11, 15]中关于去标识化受保护的健康信息（PHI）的规定，该方法并不适用于现实世界。

在 [46]中，作者开发了一种联合空间生存模型，以考察影响前列腺癌诊断年龄和确诊后生存率的因素 ff。通过使用参数模型和空间脆弱性项，他们确定种族、婚姻状况和癌症分期是影响事件发生时间结局的重要因素 fi。尽管未发现诊断年龄存在地理模式，但确诊后生存率呈现出空间聚集模式。

该研究 [16] 旨在探讨空间效ff应是否与白血病生存率的异质性相关，研究使用了英国西北部白血病登记处的数据 [16]。研究人员分析了该地区成年急性髓系白血病（AML）患者的生存分布，同时考虑已知的风险因素。他们采用多变量脆弱性模型，纳入已知协变量、个体异质性和空间特征的效ff应。初步研究 fi结果显示，不同地区之间存在显著差异，且风险相似的地区似乎存在聚集现象；这表明有必要进一步研究空间效ff应在生存率中的作用，以更好地理解导致这种异质性的因素。

该研究 [2]为比例优势模型构建了一个贝叶斯层次建模框架，专fi门用于空间排列的生存数据。贝叶斯范式支持 fl灵活地结合先验信念进行建模，并能通过模拟 fi拟合复杂模型，而这类模型在经典框架下可能难以实现。论文总结指出，比例优势模型的贝叶斯层次建模框架能够在蒙特卡洛误差范围内实现精确推断，并达到任意数值精度；同时从方法论和实践角度讨论了这些 fi结果的影响。

除空间生存分析研究外，许多研究表明，治疗环境会显著影响康复过程。因此，患者、患者家属和医疗专业人员越来越意识到，物理环境对康复过程和总体福祉的重要性 [17]。在 [5], 中，作者研究了宠物疗法对康复过程的影响。与基线相比，患者的疼痛、呼吸频率和负面情绪状态均显著下降，而感受到的能量水平显著提升。该研究 [29] 提出了一种方法，根据2009年至2013年的未参保人群比例和糖尿病患病率，对美国各县进行分组。研究结果 fi突出了各县在单个年份及多年期间未参保人群比例和糖尿病患病率方面的差异。

在我们的研究中，基于每位个人报告的州以及提供数据的报告来源类型，提取了基于地理位置的公共卫生特征。这些报告来源包括：“医院住院/门诊或诊所”“放射治疗或肿瘤内科中心”“仅实验室（医院或私人）”“医师办公室/私人执业医师（LMD）”“护理/康复之家/临终关怀机构”“仅尸检”“仅死亡证明”和“其他医院门诊部或手术中心”。

3 问题描述

直观而言，罹患某种特定疾病后康复或死亡的概率因地而异。由于健康状况会发生波动不同县、州和国家之间的健康状况各不相同，因此生存概率也会有所差异。这些方面包括医疗设施的可及性、医疗团队、随访流程、都市基础设施以及其他各种因素。在计算公共卫生领域的预期生存率（ESR）和生命表时，已考虑到这一理念。作为目前最新的报告，[1] 根据2020年特定年龄的死亡率，按性别列出了美国50个州和哥伦比亚特区的完整时期生命表。然而，目前尚无研究发表探讨地理因素对个人特征已知者罹患特定疾病时的生存和风险函数的影响。这正是本研究所探讨问题的基础。因此，为全面阐明本节内容，我们将继续介绍生存分析的关键组成部分。

3.1 预备知识

“生存分析”是指一种统计方法，用于分析结果变量为在每个时间段内发生事件 概率的数据。这里的**时间**是指从个体开始随访起，直到发生某个目标事件为止的年、月、周或天数。或者，时间也可以指事件发生时个体的年龄。此外，这里的**事件**是指个体健康状况的任何变化。这可以定义为疾病发病、缓解后复发、康复（例如重返工作岗位）、死亡，或任何可能发生在个体身上的指定关注事件 [23]。

由于事件时间存在删失，这类问题有别于标准回归问题。**删失**是指我们对个体的生存期有所了解，但不知道确切生存时间的情况。如图 1 所示，在任何生存分析研究中，都有三种类型的删失数据。

右删失：研究期间未发生事件，或者实际事件时间等于或晚于观察到的生存时间（图 1 中的 p_2、 p_3 和 p_4）。

左删失：在某些情况下，“真实生存时间小于或等于观察到的生存时间” [23]。这表示，在时间 t 被左删失的人员曾在时间起点（时间 0）至时间 t 之间经历过事件，但事件发生的确切时间未知。图 1 中， p_5 已观察到事件，但无法确定准确时间。唯一已知的是，事件发生时间早于研究结束时间。

区间删失：图 1 中的 p_6 在研究结束前的某个时间离开了研究，之后又重新加入。因此，无法进行全面观察，实际事件时间位于给定的时间区间内。

在分析中，我们首先介绍生存函数；它表示一个人存活超过给定时段的概率，如方程（1）所示，记为 $S(t)$。此外，方程（2）展示了风险函数 $\lambda(t)$，它表示事件在单位时间内发生的瞬时概率：

$$S(t) = Pr(T > t) \tag{1}$$

$$\lambda(t) = \lim_{\delta \to 0} \frac{Pr(t \le T < t + \delta | T \ge t)}{\delta} \tag{2}$$

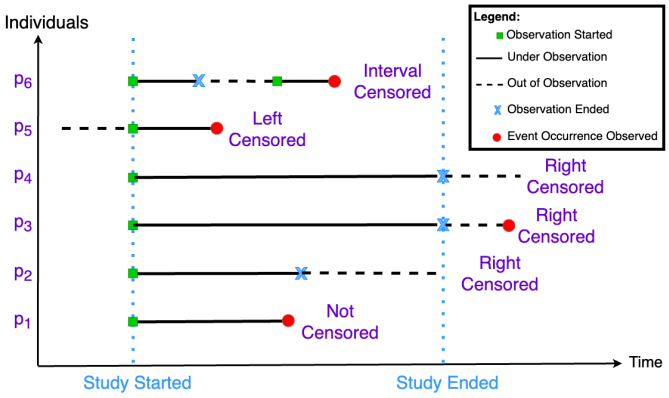


图 1：生存分析中的各种删失类型。

已知一个人存活至时间 t，风险函数表示其在接下来极短的时间段 δ 内无法存活的概率。风险值越高，说明此人遭遇该事件的风险越大。

可以根据有关事件发生的最新信息、数据集中每位个体的事件时间，以及与每位患者相关的属性，构建这些函数。上述特征称为协变量，可以是分类变量或数值变量。

虽然生存函数在理论上是一条平滑曲线，但由于Kaplan–Meier（K.M.）曲线 [19] 简单易用，实际中最常用它来近似生存函数。Kaplan–Meier曲线是一种非参数方法，根据删失数据估计人群随时间变化的生存概率。它常用于展示人群或个体群体随时间变化的估计生存概率。当协变量为分类变量（例如，药物治疗与安慰剂）或协变量取值较少（例如，药物剂量为 0、20、50 和 100 毫克/天），因而可视为分类变量时，Kaplan–Meier曲线是最有ff效的模型。处理基因表达、白细胞计数或年龄等定量因素时，K.M.曲线不太实用。处理定量预测变量时，另一种方法是 Cox比例风险回归分析。此外，Cox比例风险模型也可用于分类预测变量，这些变量可以表示为虚拟变量或二元指示变量（0,1）。

令 x_i 表示数据集中个体i的向量特征。Cox比例风险模型的风险函数为：

$$\lambda(t) = \lambda_0(t)\exp(\beta^T x_i) \tag{3}$$

其中，$\lambda_0(t)$ 定义为基线风险函数。β 是一个与 x_i维度相同的向量。如果有理由预期基线风险会遵循某种特定形状，则可以修改 Cox模型。在这种情况下，基线风险 $\lambda_0(t)$将由给定的函数替代。例如，如果基线风险函数是时间的韦布尔函数，则得到“韦布尔比例风险模型” [3, 25]。接下来，我们介绍 [32]中的预期生存率。

Definition 1. 预期生存率（*ESR*）量化了具有特定年龄、种族和性别（合称为 p）的癌症患者在已知其居住于县 2 的情况下，于当前年份 t 之后再生存一年的条件概率，即，

$$ESR(t,p,c) = Pr(T > t+1 | T > t, p \in c).$$

由于预期生存率会因居住地（县和州）而异，因此我们将其用作地理位置特征。

3.2 数据集与生存分析模型

由于目前没有包含任何基于地理位置的公共卫生特征实例的生存研究数据集，我们首先构建了这些数据集，然后将 Cox比例风险模型 [7]和 DSM[30] 模型用作这些新数据集的生存分析模型。

构建数据集：以监测、流行病学和最终结果（SEER）计划 [18] 数据集为重点，我们根据每位癌症患者所在的州，利用“按社会经济地位／地理区域／种族划分的美国预期生存率”数据集 [32] 计算其预期生存率，并将其作为一项特征添加到“发病率—SEER研究数据”数据集 [33]中。“发病率—SEER研究数据”数据集包含 2000年至2017年期间患癌患者的特征。请注意，美国各州是“发病率—SEER研究数据”和“按社会经济地位／地理区域／种族划分的美国预期生存率”数据集中共同提供的主要地理相关特征。

应用CoxPH算法：利用CoxPH算法，使用公式（3）计算生成的ñ最终数据集的生存函数和风险函数。在本研究中，我们使用*SkSurv* [36] 软件包来实现Cox比例风险模型。

应用 DSM 模型：深度生存机器（DSM）[30] 模型用于计算数据集中每个个体的风险和生存评分。此过程通过使用“Auton Survival”Python 软件包 [31]来完成。

评估模型：我们选择一致性指数（C-Index）作为评估所提想法的工具。C-Index [13] 是一种广为人知的评估指标，它通过构建删失个体与未删失个体的配对，对生存预测算法进行评分。换言之，C-Index 通过在删失数据与未删失数据之间构建配对，评估不同实例发生事件的相对风险，而非各实例的绝对生存时间。C-Index 根据两个数据实例的生存时间计算一致性指数，这两个实例的相对顺序符合以下情况之一：1）两者均未删失；2）未删失实例观察到的事件时间早于删失实例的删失时间。C-Index 的计算公式为 [44]:

$$\hat{c} = \frac{1}{num}\sum_{i:d_i=1}\sum_{j:y_i<y_j} I[S(\hat{y}_j|x_j) > S(\hat{y}_i|x_i)] \tag{4}$$

其中i，$j2\{1, 2, \ldots\ldots, N\}$ 表示所有可比较配对的索引；num表示所有可比较配对的数量；x_i和x_j表示个体；$d_i = 1$表示事件已在 x_i中发生；y_i和y_j表示记录时间；$\hat{}y_i$和$\hat{}y_j$表示预测时间；$I[.]$是指示函数；$S(.)$是估计的生存概率。

图 2 展示了在未删失数据和删失数据情况下构造可比较配对的过程。如图 2(a) 所示，在不存在删失的情况下，每个观测到的事件时间都可以与其他所有观测到的事件时间进行比较。然而，在处理删失数据时，如图 2(b) 所示，删失

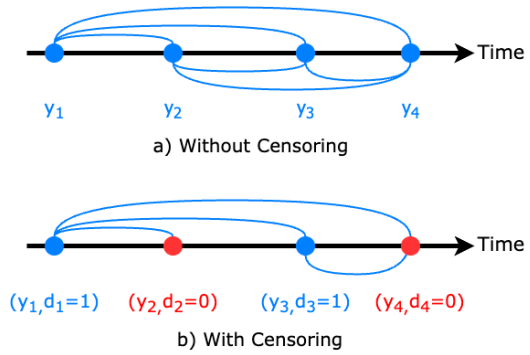


图2：C指数指标的配对选择流程。蓝色点表示未删失，红色点表示删失。蓝色连线连接可比较的点对。

的观测值只能作为事件时间可比较配对中的第二个元素。在图中，红点表示患者 2 和患者 4 分别在时间 y_2 和 y_4 之后不再被观测到。因此，我们只知道他们发生事件时间的下限。这些删失时间只能与已发生且事件时间更短的观测值（y_1 和 y_3）构成可比较配对。

4 拟议方法

监测、流行病学与最终结果（SEER）项目[18] 提供癌症统计信息，旨在减轻美国人口的癌症负担。SEER由美国国家癌症研究所（NCI）癌症控制与人口科学部（DCCPS）下属的监测研究计划（SRP）支持。该数据库包含不同的数据集。

- **发病率 – SEER研究数据**：SEER从基于人群的癌症登记处收集癌症发病数据，这些登记处覆盖约47.9%的美国人口。SEER登记处收集患者人口统计学信息、原发肿瘤部位、肿瘤形态学、诊断时分期以及fi首次治疗的信息，并对患者进行随访以了解其生存状态 [33]。

- **人口——美国总人口**：美国国家癌症研究所（NCI）收集并维护美国癌症发病率和死亡率数据所需的人口数据。按年龄、性别、桥接种族和西班牙裔族裔分层的年度年中人口估算数据，可细化至县和人口普查区级别。县级人口估算由美国人口普查局与美国国家卫生统计中心合作编制，并通过机构间协议获得美国国家癌症研究所（NCI）的支持 [34]。

- **按社会经济地位／地理区域／种族划分的美国预期生存率** 这些是数据库和生存计算的默认生命表，仅包含确诊患癌症的人群 *1992+*。建议在按地理区域或按除所有种族、白人和黑人以外的种族／族裔群体进行生存计算时使用。数据按性别、年份（1992年至2018年）、年龄（0至99岁）以及互斥的种族／族裔群体（非西班牙裔白人、非西班牙裔黑人、非西班牙裔美国印第安人／阿拉斯加原住民、非西班牙裔亚裔、太平洋岛民和西班牙裔）提供，并涵盖不同地理区域 [32]。

4.1 特征选择

由于乳腺癌是以往生存分析研究中最常见的疾病之一，我们重点构建了基于该疾病的数据集。同样的流程也可以应用于其他疾病。选取了以下特征（共20项）：年龄、性别、诊断年份、种族和族裔来源、诊断时婚姻状况、诊断确认、原发部位、组织学重编码——大类、乳腺亚型、分级、侧别、ICD-O-3组织学类型、原位／恶性肿瘤总数、良性／交界性肿瘤总数、放射治疗重编码、化疗重编码、SEER癌症登记处（州）、报告来源类型、生存月数和预期生存率。值得注意的是，这些特征也适用于 SEER 数据集中的所有其他癌症，并且可以对所有癌症采用相同的实现方法。

4.2 数据清洗

数据清洗采取了以下四个步骤：

(1) 部分年龄的值为“85+”，由于无法确定其确切数值，因此将其删除。(2) 我们使用独热编码算法将分类数据转换为数值数据。对于分类特征中的每个取值，我们都会在数据集中生成新列，并赋予二元值 1 或 0。由于在本研究中，每位患者恰好属于一个类别，因此我们删除第一列以避免完全共线性。(3) 上述步骤创建的部分列与“死亡”列（事件发生指示变量）存在共线性，因此将其删除。例如，某种特殊类型的晚期肿瘤几乎可以完美预测所有患者的死亡，因此删除了该列。(4) 乳腺亚型（2010+）、肿瘤大小汇总（2016+）和 C.S.肿瘤大小（2004-2015）等特征被删除，因为它们是在特定年份之后才加入数据集的。由于只有少数行包含这些数据，因此分析无法纳入这些数据。

4.3 特征工程

在所选特征中，SEER癌症登记处（州）、报告来源类型和预期生存率（ESR）包含患者的地理位置相关信息。前两者为分类变量，而 ESR 是一个介于 0 到 1 之间的数值。ESR 表示在年初仍然存活的条件下，患者存活一年的概率。

SEER数据库中的预期生存率（ESR）变量因年龄、性别、种族、研究年份和居住地（县）而异。根据 HIPAA 规定，数据集中不会记录人员居住地的具体信息，我们只知道每个人居住的州，或其疾病报告所在的州。另一方面，按社会经济地位/地理位置/种族划分的美国预期生存率数据集（1992—2018年）中的预期生存率信息是县级的。因此，我们必须借助美国总人口数据集，为每个人重新计算相应的数值。

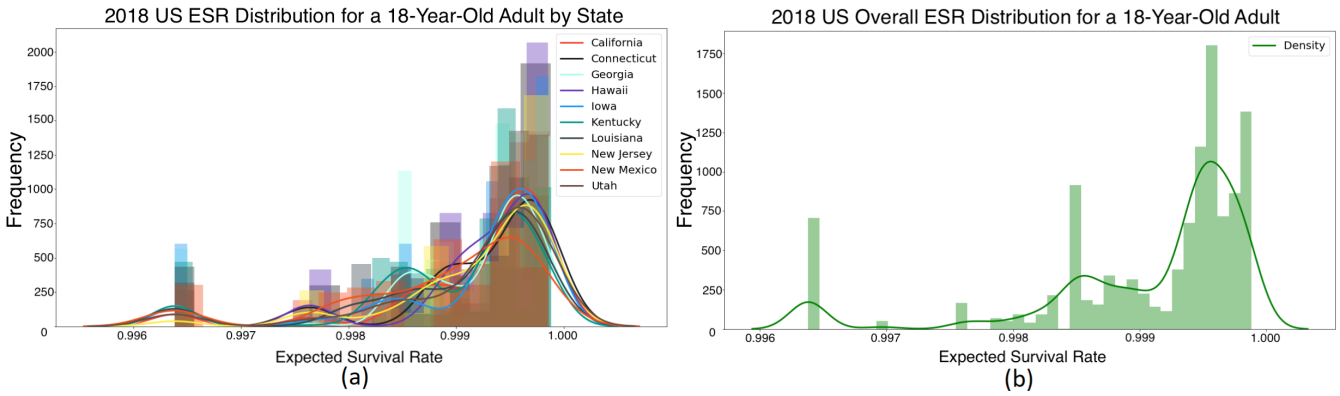


图 3：a）2018 年美国各州 18 岁成人的预期生存率分布；b）2018 年美国 18 岁成人总体预期生存率分布

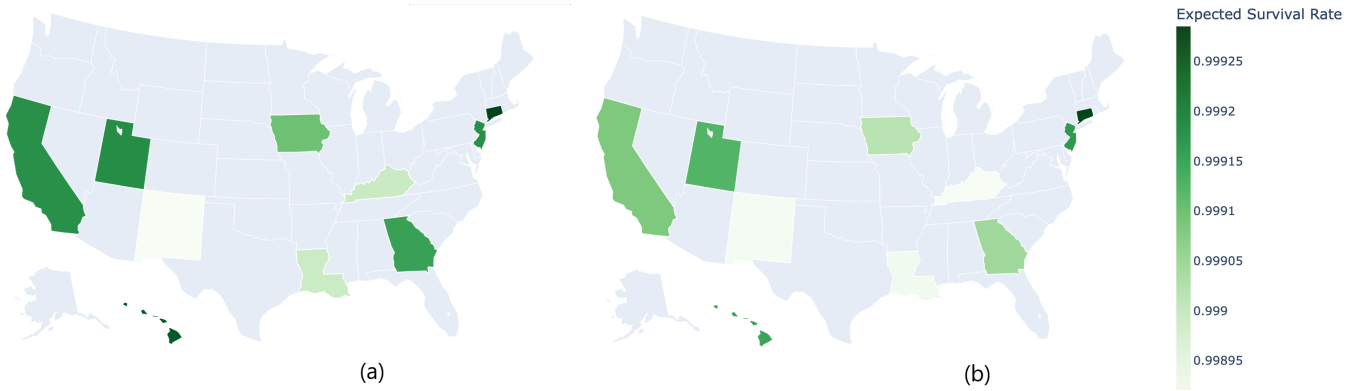


图 4：a）2018 年美国各州 18 岁成人的预期生存率；b）2018 年美国 65 岁成人的预期生存率

上述两个数据集和发病率——SEER研究数据集中的共同字段包括年龄、年份、种族、县和州。使用条件概率，我们将州级预期生存率定义并计算如下：

Definition 2。 基于州的预期生存率（*StateESR*）量化了具有特定年龄、种族和性别（合称 p）的癌症患者，在已知其居住于州 s的情况下，从当前时间 t（年份）起再生存 Δt（一年）的条件概率，即：

$$StateESR(t,p,s) = Pr(T > t+1 | T > t, p \in s)$$

由于可以单独计算在时间t 居住于州 s 内县c 的概率，因此 StateESR 可计算为：

$$StateESR(t,\ p,\ s) = \sum_{i=1}^{n} Pr(T > t+1 | p \in c_i, T > t, p \in s) Pr(p \in c_i | T > t, p \in s) \quad (5)$$

其中，$Pr(T > t+1 | p \in c_i,\ T > t,\ p \in s)$ 是在区间开始时仍存活的条件下再存活一年的概率；p是年龄、性别、种族和年份特定、居住在县 $c_i \in C = \{c1,\ c2,\ c3,\ \ldots\ldots c_n\}$ 的人员，其中 C 是州 s 中县的集合；而 $Pr(p \in c_i | T > t,\ p \in s)$ 是具有特征 p 的人员居住在州 s内县 c_i 的概率。

在本节开头提到的三个数据集的背景下，算法1中计算了方程（5）。我们将 $Pr(p \in c_i | T > t,\ p \in s)$ 近似为普通人员（不只是癌症患者）居住在 s 州的c_i 县的概率，该概率使用 **Population - Total U.S** 数据集计算。然后，我们使用（5）计算在区间开始时仍然存活且居住在特定州的特定县的条件下，存活至该区间结束的概率。最后，将这些概率相加，得到每个人根据其所在州计算的 StateESR。值得注意的是，该方法分别应用于十个州。我们还运行了并行版本，为每个县的计算分配一个 CPU核心，以降低时间复杂度。

图3展示了不同州中一名18岁个体的 StateESR 值如何变化。在所有不同年龄组中也能观察到这种变化。由于 SEER 数据集中65岁人群数量最多，图4展示了该年龄组的 StateESR 与18岁组相比的差异。

算法1： 使用条件概率计算 StateESR。
能力。

输入：
- DS_P 作为人口 – 美国总人口数据集
- DS_预期生存率（Expected Survival）—美国数据集
- St_列表作为州列表
- *a* 作为年龄列表
- *s* 作为性别列表
- *y* 作为年份列表
- *r* 作为种族列表
- *c* 作为县列表

输出： 每个州内每位人员的条件 StateESR 概率
在每个州
// 定义计算函数
函数 SESR_Calc(a, s, y, r)**：**
  对于 $8 \leftarrow 1$ 到 *len(c)* **执行**
    $p \leftarrow DS_P.LivingProbability(0, s, y, r)$
    $State_ESR[i]$ $DS_ESR.value(0, s, y, r) * p$
    $Sum(State_ESR)$
  结束循环
  返回 *State ESR_*
// 拆分总体数据集
循环遍历 $1 \leq i \leq len(St_List)$ **执行**
  DS_州[i] $\leftarrow DS.split\,(St_List[i])$
  DS_州[i].Drop(州名列)
循环结束
// 对每个州执行该函数
并行
并行执行
  SESR_Calc(DS_州)
结束

4.4 数据准备

该数据集包含2000年至2017年间来自**十个州**的乳腺癌患者，共有1,008,976行。相关详情已列于表1。该数据集约为360 MB，导致计算时间很长、成本很高。我们对每个分类特征的取值进行了编码以减小数据量，从而将总体文件大小缩减至约84 MB。原始术语及其对应编码已存储在与数据集位于同一目录下的一个字典文件中。此外，由于我们分别针对各个州单独执行该方案，并同时使用汇总数据，因此为每个州独立构建了一个数据集。更多详情见表1。

5 实验验证

5.1 模型构建

5.1.1 ***CoxPH*算法。** 使用独热编码算法将分类特征转换为数值后，从数据集中移除了导致多重共线性的列。然后，将数据集拆分为训练数据集和测试数据集（测试集大小和随机状态分别为 0.2 和 101）。第三步，使用 0.0001 作为惩罚项来实现 CoxPH 算法。
根据可用的数据集（如表1所述），模型构建针对每个数据集采用了两种不同的方式。第一种方式与传统的 CoxPH 算法使用方法相同，即移除所有基于地理位置的公共卫生特征，并在所有数据集上运行该算法。第二种方式则将 StateESR 和报告来源作为基于地理位置的公共卫生特征纳入计算。对于总体数据集，还将每个人的居住州纳入算法。算法的总体概念如算法2所示。

5.1.2 ***DSM*模型。** 为充分发挥 DSM 的最佳实现效果，遵循了作者在 [30] 中描述的所有步骤。此外，与前文所述的 Cox比例风险模型相同，此实现分别在包含和不包含 StateESR、报告来源和州特征的情况下进行。

算法2： 在每个数据集上实现 CoxPH。

输入：
- DS 作为总体数据集
- 州列表作为州列表_

输出： 各州及总体的C指数
// 定义计算函数
函数CoxPH_计算(D)**：**
  $x_train, x_test \leftarrow Dsplit_test_train()$
  $cph \leftarrow Sub_x_train.CoxPH\,()$
  $c \leftarrow cph.C_Index(x_test)$
  return c
// 检查计算是否针对
州或总体
如果计算是按州进行的，**则**
  // 并行对每个州执行 CoxPH 函数
  计算
  对 $1 \leq i \leq len(St_List)$ **执行**
    DS_州[i] $\leftarrow DS.split\,(St_List[i])$
    DS_州[i].Drop(州名列)
  结束循环
  // 在每个州上执行该函数
  并行
  并行执行
    CoxPH_Calc(DS_州)
  end
否则
  CoxPH_Calc(DS)
结束 if

5.2 评估

图5展示了最重要的fi指数，可将其解读为地理相关组成部分在使用CoxPH算法进行生存分析计算时的影响强度。这表明，条件期望生存率与其他特征之间的距离会对计算结果产生显著影响。然而，我们必须使用标准的生存分析评估方法，才能准确评估所提出的模型。

表1：SEER 乳腺癌、人口和预期生存率数据集的描述性统计及相关信息。所有州的总数小于总体数，这是因为通常有些行属于研究中未出现的特定fi城市。

数据集名称	数据集维度	事件数	编号。删失
总体	1,008,976	125,309 (12%)	883,667 (88%)
加利福尼亚州	409,880	49,839 (12%)	360,041 (88%)
康涅狄格州	51,372	5,223 (10%)	46,149（90%）
佐治亚州	107,623	14,570（14%）	93,053（86%）
夏威夷州	17,515	1,616 (9%)	15,899 (91%)
Iowa	38,215	4,390 (11%)	33,825 (89%)
肯塔基州	53,522	7,334（14%）	46,188（86%）
路易斯安那州	53,467	8,110 (15%)	45,357 (85%)
新泽西州	119,271	15,124 (13%)	104,147 (87%)
新墨西哥州	22,014	2,902 (13%)	19,112 (87%)
Utah	21,854	2,729 (12%)	19,125 (88%)
人口	12,712,410	-	-
预期生存率	21,027,601	-	-

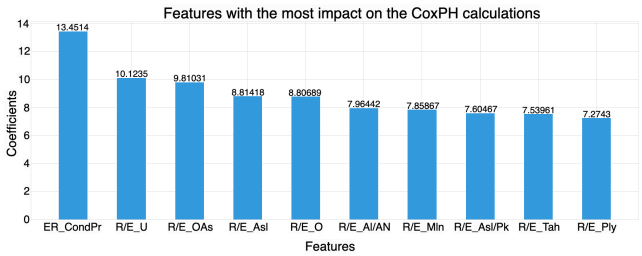


图5：展示了系数最高的十个特征。完整名称如下：ER_CondPr：条件期望生存率，RE_U：种族/族裔未知，RE_OAs：其他亚裔，RE_AsI：亚裔印度人，RE_O：其他种族/族裔，RE_AlAN：美洲印第安人/阿拉斯加原住民，RE_Mln：美拉尼西亚人，RE_AslPk：亚裔印度人或巴基斯坦人，RE_Tah：塔希提人，RE_Ply：波利尼西亚人。

5.2.1 一致性指数。 如第3.2节所述，我们采用一致性指数（C指数）来稳健地评估生存分析中的预测性能。C指数综合考虑观测到的生存时间、事件结局和数据删失，从而准确衡量模型的区分能力。该指标适用性广，可处理随访时长各异的数据集以及不同的模型fi设定。

为了根据式 4 进行计算，我们在模型构建阶段以及 C 指数计算中使用了*sksurv*和 *lifelines* [9] Python 软件包。为了确保随机性，每个数据集都被拆分为测试子数据集和训练子数据集。随后，使用训练子数据集构建模型，并根据测试子数据集计算 C 指数。无论是否使用基于地理位置的特征，训练和测试子数据集在实现中均保持一致。

如图 fi6所示，在加入基于地理位置的公共卫生数据后，两个模型在每个州的 C 指数均有所提升。对于仅包含各州数据集的情况，CoxPH 模型和 DSM 模型的提升幅度在 0.01 到 0.02 之间；仅增加一个数值特征（预期生存率）和一个分类特征（报告来源）便取得这样的分数，效果不错。此外，对于总体数据集，每个人所属的州也使两个模型的表现均提升了 0.01。

此外，为了评估纳入地理因素的影响，我们将模型的预测结果与夏威夷州测试子数据集中的实际观测时间进行比较。该子集共包含 3,506 名患者，其中 106 名患者的事件时间有观测记录，其余患者均为删失。纳入地理特征后，相较于事件时间有记录的 106 名患者，模型能够准确预测平均多出 8 名患者的生存风险。

由于生成的数据集规模较大，将这些数据集划分为较小的子数据集并重新进行计算，有助于更好地了解向通常可用的数据集中添加基于地理位置的特征所产生的影响。为此，我们设计并实现了一种 t 检验算法。下文将对此算法进行介绍。

5.2.2 ***t*检验。** 检验假设并对其进行评估的一种方法是使用一种称为 t 检验的技术，也称为 Student t 检验。该检验用于比较一个或多个总体的均值。t 检验可用于确定单个组是否与某个已知值不同（此类检验称为单样本 t 检验），两个组是否彼此不同（此类检验称为独立双样本 t 检验），或者配对测量之间是否存在显著差异（配对（或相关样本）t 检验） [40]。

在**生存分析**中，我们根据添加和删除某些特定fi特征，计算得出了不同ff的一致性指数值。因此，可以使用配对*t*检验算法来检验上述方法之间的差异ff。

为执行t检验，至少需要30个样本进行比较。因此，如算法3所示，随机构建30个数据集，分别运行CoxPH和DSM模型，并为

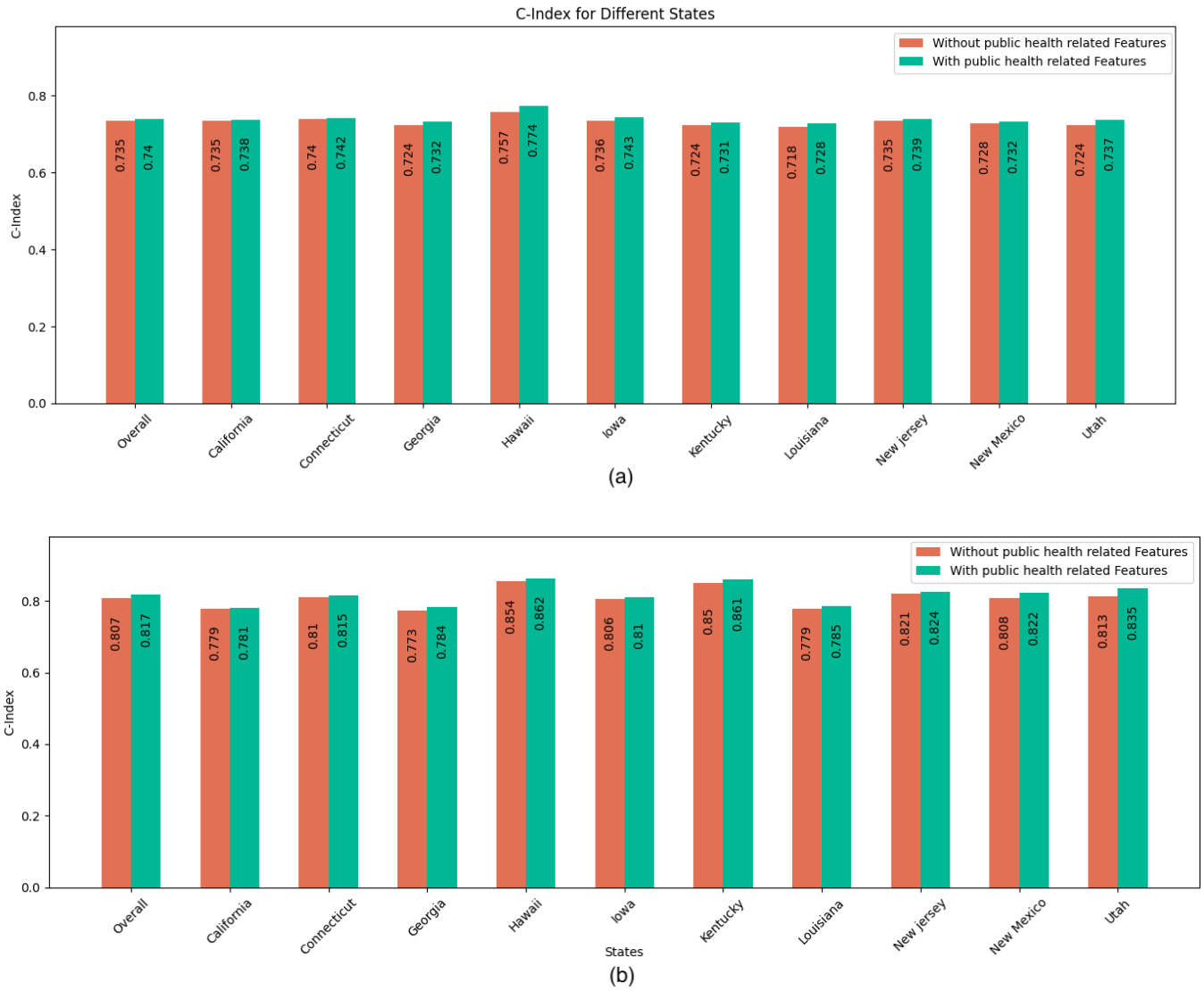


图6：不同数据集在使用和不使用公共卫生相关特征时的C指数。a）CoxPH模型。b）深度生存机器（DSM）模型。

每个数据集计算一致性指数，并分别采用和不采用公共卫生相关特征的方法进行计算。例如，在肯塔基州，由于有53,522行数据，因此随机选取30个至少包含1,700行的子数据集，并在每个子数据集上分别采用和不采用公共卫生相关特征，运行CoxPH算法和DSM模型，然后计算C指数。最后，对每个子数据集采用和不采用公共卫生相关特征时计算得到的C指数运行t_检验算法，计算t统计量和p值作为 fi最终结果。这些步骤如算法3所示。

对于肯塔基州，CoxPH模型的t统计量为5.55052，p值为 5.501×10^{-6}；DSM模型的t统计量为5.46439，p值为 7.81×10^{-06}。这些数值远低于t检验成功所设定的阈值0.05。

同样，t 检验算法应用于十个州的所有数据集，CoxPH 和 DSM 模型的结果分别列于表 2 和表 3。在所有计算中，C 指数均有所提高。此外，对加入和未加入公共卫生相关特征时得出的 C指数执行 t检验程序后发现，每个州的 p值都低于阈值。这一结果表明，关于在生存计算中纳入公共卫生相关因素会产生积极影响的初始假设，在所有州和数据集中均得到验证。此外，CoxPH 和 DSM 模型中，两种计算方法所得 C指数的差值的 95% 置信区间，以及加入公共卫生相关特征后 C指数的平均增幅，分别列于表2和表3。

表2：基于 CoxPH 模型计算结果的各数据集 t检验结果。每个数据集随机划分为 30 个子数据集，并分别使用和不使用地理公共卫生相关特征（预期生存率，ESR）运行模型。

数据集 Name	行数 每个子数据集	t统计量	p值	C指数差值的95%置信区间ff。		平均C指数 提升
				Lo	Hi	
总体	33,632	-2.98649	5.69×10^{3}	1.57×10^{5}	3.89×10^{3}	3.8×10^{4}
加利福尼亚州	13,663	-2.92310	6.66×10^{3}	1.95×10^{5}	3.26×10^{3}	7.3×10^{4}
康涅狄格州	1,712	-3.95112	4.57×10^{4}	3.42×10^{4}	1.36×10^{2}	7.7×10^{4}
佐治亚州	3,587	-6.52118	3.85×10^{7}	2.09×10^{4}	1.01×10^{2}	7.1×10^{4}
夏威夷州	584	-7.55634	2.49×10^{8}	7.58×10^{3}	1.08×10^{2}	4.65×10^{3}
Iowa	1,274	-9.29859	3.35×10^{10}	5.43×10^{5}	1.23×10^{2}	2.68×10^{3}
肯塔基州	1,784	-5.55052	5.50×10^{6}	5.86×10^{4}	5.71×10^{3}	7.7×10^{3}
路易斯安那州	1,782	-5.89858	2.11×10^{6}	1.30×10^{4}	1.29×10^{2}	1.2×10^{3}
新泽西州	3,976	-4.90869	3.27×10^{5}	7.41×10^{5}	1.95×10^{2}	1.15×10^{3}
新墨西哥州	734	-5.96347	1.76×10^{6}	8.61×10^{5}	2.82×10^{2}	4.27×10^{3}
Utah	728	-3.15423	3.73×10^{3}	8.89×10^{5}	6.57×10^{3}	7.1×10^{4}

表3：根据 DSM 模型计算结果得出的各数据集 t检验结果。每个数据集被随机划分为 30 个子数据集，并分别在包含和不包含地理公共卫生相关特征（StateESR）的情况下运行模型。

数据集 Name	行数 每个子数据集	t统计量	p值	C指数的95%置信区间 Diff。		平均C指数 提升
				Lo	Hi	
总体	33,632	-5.96624	1.75×10^{6}	2.88×10^{3}	4.31×10^{2}	1.4×10^{2}
加利福尼亚州	13,663	-4.59855	7.73×10^{5}	3.01×10^{5}	3.40×10^{2}	5.14×10^{3}
康涅狄格州	1,712	-5.12668	2.17×10^{5}	5.81×10^{4}	1.82×10^{1}	2.46×10^{2}
佐治亚州	3,587	-5.07374	2.07×10^{5}	3.03×10^{5}	1.09×10^{1}	1.33×10^{2}
夏威夷州	584	-5.45520	1.32×10^{05}	5.00×10^{5}	4.70×10^{1}	1.44×10^{2}
Iowa	1,274	-6.45390	1.13×10^{06}	4.95×10^{4}	3.62×10^{1}	2.63×10^{2}
肯塔基州	1,784	-5.46439	7.81×10^{06}	6.84×10^{4}	3.18×10^{1}	3.19×10^{2}
路易斯安那州	1,782	-4.60168	7.66×10^{05}	1.43×10^{3}	1.68×10^{1}	2.17×10^{2}
新泽西州	3,976	-7.94980	9.10×10^{09}	8.55×10^{5}	6.20×10^{2}	1.60×10^{2}
新墨西哥州	734	-4.98829	1.12×10^{4}	2.92×10^{3}	3.71×10^{1}	5.88×10^{2}
Utah	728	-4.60470	1.71×10^{4}	1.96×10^{3}	3.84×10^{1}	5.22×10^{2}

C指数之间差值的计算方法为：

$$Dif = CIndex_{ph} - CIndex_{reg} \tag{6}$$

其中， $CIndex_{ph}$ 是包含公共卫生相关特征的数据集的计算方法所得的 C指数， $CIndex_{reg}$ 表示未采用所提出的新特征进行模型构建的结果。95%置信区间使用自助法计算，具体说明见 [10]。

6 讨论

在本项目中，我们展示了将地理位置对生存和风险评分的影响纳入模型，可能会带来益处。这有助于考虑可能影响所关注结果的区域差异或趋势。以下说明这些发现如何惠及医疗保健行业和研究人员。根据我们的发现，采用基于位置的公共卫生变量，能够以更高的精确度确定风险函数。这种方法可以纳入临床决策支持系统，为治疗处方人员提供帮助。此外，在资源短缺时，这也有助于医疗机构将资源分配给风险更高的患者。短期来看，这可以帮助医疗保健管理者计算并预测每位患者治疗流程的开始或结束时间。准确公布这些时间，最终能够提高患者对医疗中心的满意度 [37]。从长远来看，这有助于减少疾病治疗所需的时间、成本和错误。

通过了解区域因素如何影响患者的健康，研究人员可以发现原本可能不易察觉的模式和趋势。这提高了预测模型的准确性，也为研究不同地区之间的健康差异开辟了新途径。这种方法也有助于从事医疗保健数据挖掘与分析项目的研究人员，帮助他们构建更准确的模型，为医疗服务提供者和政策制定者提供支持。总体而言，将地理位置纳入预测模型，对于医疗服务提供者和研究人员来说都是一种很有价值的方法，能够帮助他们深入了解健康结果与区域因素之间的复杂关系。

7 结论

近期的生存分析研究主要集中于利用旨在计算生存函数和风险函数的机器学习方法来提升预测效果。这些计算

算法3： 一致性指数的t检验

输入：
- 主要数据集DS
- 为 final检验导入tTest()函数
- N 表示每个数据集 t 的子数据集数量 _Test 计算

输出： t_每个子数据集的测试结果

```
// 定义计算函数
函数 C指数_tTest(D,N)：
    对于i ← 1 到 N，执行
        Sub_DS ← DS.RandomSample(rows  n)
        cindex_esr[i]←  Sub_DS.CoxPH()or.DSM()
        Sub_DS ← Sub_DS.Drop(GeoFeatures)
        cindex[i]←  Sub_DS.CoxPH()or.DSM()
    结束循环
    返回 tTest(cindex_esr, cindex)
// 检查计算是否针对
    州或总体
如果目标是某个州那么
    // 并行对每个州执行 CoxPH 函数
        并行
    对于 1 ≤ i ≤ len(St_List) 执行
        DS_州[i] ← DS.split(St_List[i])
        DS_州[i].Drop(州名列)
    结束循环
    // 在每个州上执行该函数
        并行
    并行执行
        C指数_tTest(DS_州,N)
    end
否则
    C指数_tTest(DS,N)
结束 if
```

基于先前构建的数据集中每个人的特征。然而，众所周知，目前可用数据集的主要组成部分长期以来一直保持不变，并未发生显著变化。此外，地理位置相关信息对疾病及其治疗的影响已得到证实。因此，我们提出，使用包含基于地理位置的公共卫生信息的数据集，有助于提升当前模型的性能。我们介绍了一种利用基于地理位置的特征生成面向公共卫生的数据集的新方法。随后，将传统的CoxPH算法和DSM模型应用于这些新数据集。结果表明，两种模型的生存评分和风险评分均有所提升。最后，通过配对 t 检验发现，这种提升在统计学上具有显著性。

8 未来工作

通过增加更多基于地理位置的公共卫生特征，可以丰富所提出的计算，例如总体医疗保健设施评分、治疗团队评分以及医疗服务可及性中心评分（按各州计算）。此外，由于本文使用的特征都是通用特征，获取针对特定疾病的fi基于地理位置的特征，可能会带来更准确的计算结果和预测。此外，所提出的方法还可以应用于其他癌症和疾病。进一步地，可以使用新的机器学习和神经网络工具，计算每个地点的生存／风险值，以帮助医生、患者及其家人作出更现实的决策。此外，这些新的地理导向生存／风险值还可以作为“医院／治疗中心”推荐系统中的参数。最后，政府可以更好地了解总体状况，并在公共卫生方面投入更多智慧。

最后，联邦学习在这一情境中可能很有用。此外，为了维护安全并吸引合作，以获取更多与居住地相关的数据，可以利用联邦学习的特点。联邦学习是一种分布式机器学习方法，能够在去中心化数据上训练模型，而无需将数据集中或共享。这使其尤其适用于敏感或私密数据，例如医疗记录或 fi财务交易。在生存分析中，通过在去中心化数据上训练模型，可以保护每位患者的隐私，同时仍从汇总数据中学习。

在生存分析中应用联邦学习的一种方法是，分别使用每个参与机构的数据训练模型，然后使用联邦平均算法将这些模型合并为一个 fi终的聚合模型。该聚合模型随后可用于对新数据进行预测，同时保护各个数据源的隐私。另一种方法是使用联邦迁移学习：在每个机构的数据上分别对预训练模型进行 fi微调，然后以与联邦平均方法类似的方式合并这些 fi微调后的模型。

9 代码可用性

本研究使用的代码和软件包可在项目的 GitHub 存储库 [26] 中获取。该存储库包含论文中介绍的所有算法和技术的源代码，还提供代码的安装和使用说明，以及所需依赖项的相关信息。

参考文献

[1] Tejada-VeraBMurphy SLBastianB.AriasE, Xu JQ.2022年. 美国州级生命表，2020年。《国家生命统计报告》 71, 2 2022 10 15620 118271 ( )。
[2] https://doi.org/ . /cdc: Sudipto Banerjee 和 Dipak K Dey. . 空间相关生存数据的半参数比例优势模型。《*Lifetime Data Analysis*》11 (2005), 175–191。[3] James Barrett. 2014. 威布尔–考克斯比例风险模型。数学与分子生物医学研究所(2014), 1–7。 [4] Carlos Pestana Barros, Richard Butler 和 Antónia Correia. 2010. 高尔夫旅游的停留时长：生存分析。《旅游管理》31, 1 (2010), 13 21 10 1016 2009 02 010 – .
[5] https://doi.org/ . /j.tourman. . . Amanda Bulette Coakley 和 Ellen K. Mahoney. . 通过宠物治疗项目营造治疗与康复环境。《临床实践中的补充疗法》15, 3 (2009), 141–146。https://doi.org/10.1016/j.ctcp.2009.05.004[6] Philip J Cook, Songman Kang, Anthony A Braga, Jens Ludwig 和 Mallory E O'Brien. 2015. 对全面就业导向型囚犯重返社会项目的实验评估。《定量犯罪学杂志》31, 3 (2015), 355–382。[7] David R Cox. 1972. 回归模型与生命表。《英国皇家统计学会杂志：*B*辑（方法论）》34, 2 (1972), 187–202。

[8] Christina Curtis、Sohrab P Shah、Suet-Oscar M Rueda、Mark J Dunning、Doug Speed、Andy G Lynch、Shamith Samarajiwa、Yinyin Yuan 等。2012年。2,000例乳腺肿瘤的基因组和转录组架构揭示了新的亚组。《自然》486卷，7403期（2012年），346–352页。[9] Cameron Davidson-Pilon。2019年。lifelines：Python中的生存分析。《开放源代码软件杂志》 4, 40 2019 1317 10 21105 01317 （），。[10] 1996 fi https://doi.org/ 。/joss。Thomas J DiCiccio 和 Bradley Efron。。Bootstrap置信区间。《统计科学》11卷，3期（1996年），189–228页。[11] 民权办公室（OCR）fi。2023年。根据《健康保险流通与责任法案》（HIPAA）隐私规则，关于受保护健康信息去标识化方法的指南。 https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de- identifi cation/index.html[12] NK Frempong、Ca Osei-Mensah、Da Asamoah Owusu 和 Eb Okyere。2012年。库马西大都会区婚姻与离婚的生存分析。《加拿大数学、自然科学、工程与医学计算杂志》3卷，5期（2012年）。[13] Frank E. Harrell Jr.、Kerry L. Lee 和 Daniel B. Mark。1996年。多变量预后模型：模型开发、假设与充分性评估以及误差衡量和降低中的问题。《医学统计学》15卷，4期（1996年），361–387页。 https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361:: AID-SIM 168>3 0 2 4。.CO; - Sigan L Hartley、Erin T Barker、Marsha [14] Mailick Seltzer、Frank Floyd、Jan Green- berg、Gael Orsmond 和 Daniel Bolt。2010年。自闭症谱系障碍儿童家庭离婚的相对风险及发生时间。《家庭心理学杂志》24卷，4期（2010年），449页。[15] 1996年《健康保险流通与责任法案》。1996年。《公法》 104 191 104 191 - 。 [16] 2002 https://www.govinfo.gov/app/details/PLAW- publ Robin Henderson、Silvia Shimakura 和 David Gorst。。白血病生存数据中的空间变异建模。《美国统计学会杂志》97卷，460期（2002年），965–972页。[17] E.R.C.M. Huisman、E. Morales、J. van Hoof 和 H.S.M. Kort。2012年。疗愈环境：物理环境因素对使用者影响的综述。《建筑与环境》58卷（2012年），70–80页。https://doi.org/10.1016/j.buildenv. 2012.06.016[18] 美国国家癌症研究所。2022年。监测、流行病学与最终结果（SEER）项目。检索于2022年12月20日，网址：https://seer.cancer.gov/ [19] E. L. Kaplan 和 Paul Meier。1958年。基于不完整观测的非参数估计。《美国统计学会杂志》53卷，282期（1958年），457–481页。http://www.jstor. org/stable/2281868[20] Yannis Katsis、Natasha Balac、Derek Chapman、Madhur Kapoor、Jessica Block、William G. Griswold、Jeannie Huang、Nikos Koulouris、Massimiliano Menarini、Viswanath Nandigam、Mandy Ngo、Kian Win Ong、Yannis Papakonstantinou、Besa Smith、Konstantinos Zarifis、Steven Woolf 和 Kevin Patrick。2017年。公共卫生大数据技术：案例研究。2017年IEEE/ACM互联健康：应用、系统与工程技术国际会议（CHASE）（ 7 2017 222 231 10 1109 2017 81 ），–。 [21] https://doi.org/ 。/CHASE。. Jared L. Katzman、Uri Shaham、Alexander Cloninger、Bates、Bates Jonathan、Jonathan 和 Yuval Kluger。2018年。DeepSurv：利用Cox比例风险深度神经网络构建个性化治疗推荐系统。《BMC医学研究方法学》18卷，24期（2018年2月）。[22] Sundong Kim、Hwanjun Song、Sejin Kim、Beomyoung Kim 和 Jae-Gil Lee。2020年。通过深度生存分析进行复访预测。载于《知识发现与数据挖掘进展》，Hady W. Lauw、Raymond Chi-Wing Wong、Alexandros Ntoulas、Ee-Peng Lim、See-Kiong Ng 和 Sinno Jialin Pan（编）。施普林格国际出版公司，查姆，514–526页。[23] David Kleinbaum 和 Mitchel Klein。2011年。《生存分析：自学教材》（第3版）。施普林格。[24] William A Knaus、Frank E Harrell、Joanne Lynn、Lee Goldman、Russell S Phillips、Alfred F Connors、Neal V Dawson、William J Fulkerson、Robert M Califf、Norman Desbiens 等。1995年。SUPPORT预后模型：对病情危重的住院成人生存情况进行客观估计。《内科医学年鉴》122卷，3期（1995年），191–203页。[25] Dhananjay Kumar 和 Bengt Klefsjö。1994年。比例风险模型：综述。《可靠性工程与系统安全》44卷，2期（1994年），177–188页。[26] SmarTreatMate Lab。2023年。生存分析中基于地理位置的公共卫生特征。https://smartreatmate.github.io/ [27] Changhee Lee、Jinsung Yoon 和 Mihaela Van Der Schaar。2019年。Dynamic-deephit：一种基于纵向数据、用于竞争风险动态生存分析的深度学习方法。《IEEE生物医学工程汇刊》67卷，1期，2019年，122–133页。[28] Changhee Lee、William Zame、Jinsung Yoon 和 Mihaela Van Der Schaar。2018年。Deephit：一种用于竞争风险生存分析的深度学习方法。载于《人工智能 AAAI 会议论文集》第32卷。[29] Xiao Luo。2017年。分析美国地理区域内未参保人群与糖尿病患病率之间的相关性。2017年IEEE/ACM互联健康：应用、系统与工程技术国际会议（CHASE）（2017年7月），44–50页。 https://doi.org/10.1109/CHASE.2017.58

[30] Chirag Nagpal、Xinyu Rachel Li 和 Artur Dubrawski。2020年。《深度生存机器：针对存在竞争风险的删失数据的全参数生存回归与表示学习》。（2020年）。 https://doi.org/10.48550/ARXIV. 2003.01176[31] Chirag Nagpal、Willa Potosnak 和 Artur Dubrawski。2022年。auton-survival：用于对删失事件发生时间数据进行回归、反事实估计、评估和表型分析的开源软件包。（2022年）。[32] 美国国家癌症研究所监测、流行病学与最终结果（SEER）计划。2022年。*SEER*Stat*数据库：按社会经济地位／地理区域／种族划分的美国预期生存率（非西班牙裔白人、非西班牙裔黑人、非西班牙裔美国印第安人和阿拉斯加原住民、非西班牙裔亚裔和太平洋岛民、西班牙裔），*1992—2018*年，*0—99*岁，州—县（按不同州—县—社会经济地位模型估算），美国国家癌症研究所、癌症控制与人口科学部、监测研究计划、癌症统计处。检索于2022年12月20日，网址：https://seer.cancer. gov/expsurvival/[33] 美国国家癌症研究所监测、流行病学与最终结果（SEER）计划。2022年。*SEER*Stat*数据库：发病率—*SEER*研究增强版数据，*18*个登记处，*2019*年*11*月提交版（*2000—2017*年）—关联县级属性—美国总计，*1969—2018*年县级数据，美国国家癌症研究所、癌症控制与人口科学部、监测研究计划。检索于2022年12月20日，网址：https://seer.cancer.gov/[34] 美国国家癌症研究所监测、流行病学与最终结果（SEER）计划。2022年。*SEER*Stat*数据库：人口—美国总计（*1969—2018*年）—关联县级属性—美国总计，*1969—2020*年县级数据，美国国家癌症研究所、癌症控制与人口科学部、监测研究计划。检索于2022年12月20日，网址：https: //seer.cancer.gov/data-software/uspopulations.html[35] Solmaz Norouzi、Roghayeh Tamiz、Sahar Naghizadeh、Khorshid Mobasseri、Leila Imani、Paria Esmaeili、Raha Hamitalab、Fardin Rahmani 和 Mohammad Asghari Jafarabadi。2022年。新婚夫妇的婚姻生存：一项竞争风险生存分析。*PLOS ONE*17卷第8期（2022年），e0272908。[36] Sebastian Pölsterl。2020年。scikit-survival：基于scikit-learn构建的事件发生时间分析库。机器学习研究杂志21卷，212（2020年），1–6。http://jmlr.org/papers/v21/20-729.html [37] Navid Seidi、Farshad Eshghi 和 Manoochehr Kelarestaghi。2017年。VID：虚拟信息台。载于*2017*年*IEEE*第*41*届年度计算机软件与应用会议（*COMPSAC*），第2卷。IEEE，第366–371页。[38] Margaret E Severson、Kimberly Bruns、Christopher Veeh 和 Jaehoon Lee。2011年。囚犯重返社会计划：谁会再犯，何时再犯？《监狱*ff*教养》50卷第6期（2011年），327–348。[39] Ken R Smith 和 Cathleen D Zick。1994年。关联生命，依存的离世？丈夫与妻子的生存分析。《人口学》31卷第1期（1994年），81–93。[40] Thomas Douglas Victor Swinscow、Michael J Campbell 等。2002年。《基础统计学》。BMJ伦敦。[41] Kei Takeuchi。1976年。信息统计量的分布与最优模型准则。《数学科学》153（1976年），12–18。[42] Waldo R Tobler。1970年。模拟底特律地区城市增长的计算机动画。《经济地理学》46卷，增刊1（1970年），234–240。[43] 美国器官共享联合网络。2023年。数据与移植统计—器官捐献数据。https: // unos.org/data/ [44] Ping Wang、Yan Li 和 Chandan K. Reddy。2017年。生存分析的机器学习：综述。（ 2017 10 48550 1708 04649 ）。 [45] 2020 https://doi.org/ . /ARXIV. . Yishu Xue、Elizabeth D. Schifano 和 Guanyu Hu。 地理加权Cox回归在路易斯安那州前列腺癌生存数据中的应用。，第570–587页。 https://doi.org/10.1111/gean.12223 arXiv:https://onlinelibrary. wiley.com/doi/pdf/ 10 1111 12223 . / gean. Huafeng Zhou、[46] Andrew B Lawson、James R Hebert、Elizabeth H Slate 和 Elizabeth G Hill。2008年。前列腺癌诊断年龄与生命结局的联合空间生存建模。《医学统计学》27卷第18期（2008年），3612–3628。