上海大学

SHANGHAIUNIVERSITY

毕业论文（设计）

UNDERGRADUATE THESIS (PROJECT)

题目：基于机器学习的子宫肉瘤患者的临床病理及预后分析

学院	计算机工程与科学学院
专业	计算机科学与技术
学号	19120171
学生姓名	黄奕恺
指导教师	沈文枫
起讫日期	2023.02.21– 2023.06.03

目 录

基于机器学习的子宫肉瘤患者的临床病理及预后分析

摘 要

子宫肉瘤是源于子宫体部位的一组独立的高度恶性肿瘤，根据原发部位又分为子宫平滑肌肉瘤、子宫内膜间质肉瘤等。该肿瘤属于较为罕见的肿瘤，通常发病于45岁以上的绝经后妇女，虽然已经有一些相关文献对其进行了科普与介绍，关于其临床病理及预后分析的研究依然较少。本文基于SEER癌症数据集使用KM曲线与COX多因素比例风险回归模型对影响其患者生存率的各个因素进行了分析与总结；并根据其预后相关数据进行了多种聚类算法的使用，分析聚类后得到簇所代表的表型具有的特征，从而分析得到子宫肉瘤患者的不同表型特点。由于聚类算法属于无监督学习，难以运用在新的数据上。所以本文建立了相关预后模型，使用决策树与随机森林对SEER上的聚类得到的表型进行训练，从而得到能区分患者所对应表型的预后模型，从而更好地评估患者的生存情况。最后结合医院提供的数据对上述模型进行了验证，总结全文工作与创新点，并展望后续工作。

关键词：机器学习, 比例风险回归模型, 决策树, 随机森林, 层次聚类

Machine learning-based clinicopathological and prognostic analysis of patients with uterine sarcoma

ABSTRACT

Uterine sarcoma is an independent group of highly malignant tumors originating from the body of the uterus, which are classified into smooth muscle sarcoma and endometrial mesenchymal sarcoma according to the site of origin. It is a rare tumor that usually develops in postmenopausal women over 45 years of age. Although it has been popularized and introduced in the literature, there are still few studies on its clinicopathological and prognostic analysis. In this paper, we analyzed and summarized the factors affecting the survival rate of patients based on the SEER cancer dataset using the KM curve and COX multifactor proportional risk regression model; and used various clustering algorithms to analyze the characteristics of the phenotypes represented by clusters after clustering to analyze the different phenotypic characteristics of patients with uterine sarcoma based on their prognosis-related data. Since clustering algorithms are unsupervised learning, they are difficult to apply on new data. Therefore, in this paper, a relevant prognostic model is developed and the phenotypes obtained from clustering on SEER are trained using decision trees and random forests to obtain a prognostic model that can distinguish the phenotypes corresponding to the patients and thus better assess the survival of the patients. Finally, the above models were validated with the data provided by the hospital, summarizing the full work and innovations and looking forward to the follow-up work.

Keywords: Machine learning, proportional risk regression models, decision trees, random forests, hierarchical clustering

第1章　绪论

本章主要介绍了子宫肉瘤研究的相关研究的研究背景及意义，分析了相关课题的研究方法与现状，最后列举了本文的研究目标内容与结构。

§1.1　研究背景与意义

子宫肉瘤是一种罕见的子宫恶性肿瘤，其相关死亡率在子宫恶性肿瘤中中的比例在16%以上[1]，目前的主要治疗方式是以手术为主，手术后一般会通过化疗（少数情况使用放疗）辅助，以确保最后效果。根据SEER数据分析显示，子宫肉瘤患者的中位年龄在65岁左右，发病原因有很多，根据对文献中的数据显示，子宫肉瘤可能与某些遗传基因的突变有关。而子宫肉瘤的复发情况较为常见。子宫肉瘤的复发是指在手术和治疗完成后，术后阶段肿瘤重新生长并再次出现的一系列过程。子宫肉瘤的复发并不局限于原发病灶，也可能出现在周边组织内，包括子宫、输卵管、卵巢等等。按照文献显示，I II期患者5年生存率为59%，而III期则为22%，IV期为9%。[1]从数据中可以看出，子宫肉瘤的复发率较高，预后效果较差。在预防子宫肉瘤的复发中，一些研究表明了术后辅助治疗和放疗能够达到降低复发率的作用，并提高患者的生存率。然而作为较为少见的恶性肿瘤，目前子宫肉瘤方面尚未有比较公认的预后处理模型出现。为了处理近年各个医院积累的子宫肉瘤的随访与预后数据，利用数据对患者术后复发状况或生存率进行预测与评估，能够更好地帮助患者了解自身身体状况，也能帮助医生对于危险程度较高的患者进行重点关注，从而能够让医疗资源能够更加有效地被利用。因此，本文旨在建立一套使用患者预后数据训练而成的预后模型，并建立相应的UI界面，让医学人士能够方便地管理并使用该系统对于患者的将来的生存状况进行评估。

§1.2　子宫肉瘤研究发展现状

目前，子宫肉瘤的研究尚处于开展阶段，使用的模型主要分为以下两类：图像模型、数据模型。

§1.2.1　在子宫肉瘤方面使用的图像模型

图像模型主要用于子宫肉瘤的诊断与治疗方案的提出。其中比较常用的图像源是计算机断层扫描（CT）与核磁共振成像（MRI）。图像模型主要通过人工智能技术对于图像进行分隔，提取出ROI，并用于分类模型。分类模型是用于分类和预测的模

型，目前常用于子宫肉瘤的治疗决策中。其中，Transformer与生成对抗网络（GAN）是目前比较常用的模型。Transformer可以实现多模态的医学图像分类。以往使用的深度神经网络由于是基于卷积架构形成，它在图像像素较为清晰，或内容较为复杂时，难以对于图像中结构的远端依赖性有较为明确的认知，对于复杂情况分类效果并不好。然而在使用了自注意力机制的Transformer后，它对远端结构的编码让它拥有了更强的学习表达能力，从而有了更好的分类效果。而生成对抗网络则可以生成与真实数据相似的模拟图像，通过模拟图像与真实图像的对比，生成模型与判别模型的不断迭代提升。GAN生成的图像可以为有限的图像数据添加标注后的新数据，同时其附带的判别模型也可以用于对于医生训练数据的扩充。

§1.2.2 在子宫肉瘤方面使用的数据模型

最常用的分类模型是支持向量机（SVM）和随机森林（Random Forest）。这些模型可以用于预测子宫肉瘤的侵袭能力和转移风险，从而为临床医生提供重要的治疗决策参考。

§1.3 本文的研究内容及目标

§1.3.1 研究内容

本文旨在设计并使用Vue与Flask作为前后端技术栈构建一个简易的患者病情预测平台，调用Python实现的预后模型。从而可以帮助医生预判患者病情，使得医生倾注医疗资源来为高风险患者进行进一步的病情随访，推动医学诊断的数字化，让医生能在这个更便携的平台上开展一系列工作，同时也为医疗服务的集成提供了一条较为有效的工程实践经验。

§1.3.2 研究目标

对于本文的研究内容，我制定了以下几条目标：

1） 从SEER数据库获得数据，验证各类模型的前置条件是否满足，测试各类模型方法，评估各类方法的作用、优缺点与效果。

2） 实现一个具有登录、授权功能的前后端系统，添加工单功能，同时引入数据导出与导出功能，在后端中集成Python实现的预后模型方法，并提供在线的训练与预测功能，从而能够满足医生与管理员的共同使用。

§1.4 本文组织结构

整篇论文一共分为七章。

第一张介绍了子宫肉瘤相关研究的背景与意义，阐述了当前子宫肉瘤相关研究的内容与方向，并说明了本文的研究目标与具体内容。

第二章主要介绍了本文使用的代码环境与技术，描述了相关的曲线或参数的具体含义，并阐释了为何使用这些技术与指标。

第三章是本文筛选并处理数据集的过程，描述了本文中如何从SEER癌症数据集与医院数据集中分别下载并处理数据，从而能在下面的章节中分析并使用。

第四章分析了数据集中的内容，利用数据进行了生存分析，使用KM曲线与COX比例风险模型研究各影响因素对生存的影响，并进而验证数据的有效性。

第五章首先使用无监督聚类模型对患者的表型进行分类，并分析了各表型所具有的不同特征与预后效果的不同。用分类得到的患者表型数据作为监督学习的数据源，使用决策树与随机森林模型以判断患者所处的表型，并与使用三年生存率的传统方法进行了对比，评估了两者的效果。

第六章介绍了根据模型实现的项目主要功能及实现过程中的技术细节。

第七章对全文进行了总结，归纳了本文的创新点与具体内容，并指出了本文使用的模型的局限性与改进方向。

第2章　子宫肉瘤预后模型相关技术综述

本文的子宫肉瘤预后模型使用Python搭建模型，力求模型能够具有更好的可移植性，并依据这一点构建了相关系统。系统主要使用Vue.js作为核心框架，并配合以TypeScript和Sass作为技术补充；服务端考虑到模型算法基于Python，使用轻量级并高度可定制的Flask实现，并使用Docker容器技术让后端的部署更加稳定、便携。

同时，本文亦使用了一些医学方面的常用模型与分析方法，本章主要对使用到的相关技术、生存分析方法与预后模型进行介绍。

§2.1　环境

本文使用WSL2(Windows Subsystem Linux)Arch发行版进行开发，Python版本为3.10.9，Vue.js版本为5.0.8，主要依赖为Element-plus。WSL2是基于Windows开发的Linux子系统，使用子系统可以在性能不受限的情况下使用Windows的大部分计算与存储资源，且传统的虚拟机与双启动系统的开销也不存在。相较于直接使用Windows进行开发，使用WSL进行开发能够使用apt、pacman等包管理器，且Arch还具有详细的官方WIKI与社区支持；而比较使用Linux开发，WSL的Remote连接更加稳定，能够更好地使用WIndows独占的部分软件，例如本文使用的SEERStat数据库官方软件，让开发更加便利。

§2.2　技术栈

§2.2.1　Docker

Docker是一个现今广泛使用的用于开发，运输和运行容器化应用程序的开放平台。使用Docker能够让我们专注于应用程序，而不是花费大量时间调试基础架构。Docker相比较于之前的虚拟机服务，Docker不需要占用多余的磁盘IO，能够有效减少计算资源的消耗。同时，Docker自包含程序依赖，这意味着Python项目的使用中也可以像js前端框架中一样，使用类似package.json的Dockerfile记录所需要的依赖，不同点是Dockerfile不直接写明依赖，而是用requirements.txt等文件来进行存储。这些优势意味着Docker无论在开发还是后期维护中都提供着很强的便携性与健壮性。Docker的核心概念是镜像、容器与仓库。其中镜像不难理解，类似于Linux中的镜像。我们一般使用一个基础镜像，譬如Win10、Linux等，这些基础镜像中包含着能够运行容器的最低限度的底层环境，而我们则基于这个基础镜像编写相关的配置，

比如导入依赖等，然后将这些配置逐层地添加到镜像中，使用Union FS技术对其进行分层与合层记录。镜像的层化技术能让我们具体地分出环境的各层结构，并依据其共享来减少重复镜像的拉取，从而最大化资源的利用率。而容器则包含程序运行需要的一切环境，轻量化地提供可共享可复制地一致服务，容器层的一切修改都不会作用于底层环境，而容器销毁时随着其生命周期的结束所有更改也会消失，从而提供了copy on write的安全特性。仓库类似Github中的仓库，可以使用Docker命令拉取。本文使用Docker搭配代码部署平台后，程序可以在上传到Github仓库之后自动部署程序于服务器上，从而实现方便便携的开发。

§2.2.2 Vue.js

Vue.js（简称Vue）是一个用于构建用户界面的JavaScript框架。它建立于HTML、CSS、JavaScript之上，提供声明式和基于组件的编程模型。声明式表明Vue可以使用模板语法动态渲染HTML，让我们可以基于JavaScript中的各个状态动态地描述输出的HTML文本；而组件化意味着Vue的各个模块是可以高度可重用的，Vue SFC将HTML、CSS、JavaScript组合在一起并封装在一个文件中。Vue使用渐进式框架，可以根据需求使用其支持的特性，包括：去构建化的HTML增强、可以在任何页面上嵌入的组件、单页应用程序（SPA）、全栈服务器端渲染（SSR）、静态站点生成（SSG）、可以面向多种应用（包括桌面、移动、WebGL与终端）。

§2.2.2.1 Vite

Vite与Vue-Cli类似，是一个提供项目脚手架与开发服务器的构建工具。不过区别在于Vite并不是构建在Webpack上的，而是使用浏览器中的的ES模块，这让Vite项目提供了很低的延迟和很高的速度，在大型项目中，Vite拥有着远超Vue-Cli的构建与启动速度。在日常使用中，随着项目的不断增大，基于Webpack的Vue-Cli构建速度一般在20秒左右，而相同体量的Vite项目往往恒定在1秒以下，这在需要经常修改的前端项目中会提供很大的便利。这是由于Vite不绑定服务端而是使用浏览器的原生支持。

但是这样的设计也会带来一些问题，Vite的开发环境需要基于现代浏览器，也就是说至少要支持ES2015，在版本较老甚至只使用CommonJS的浏览器中兼容不全，可能会带来一定问题，不过在如今的生产环境中这个问题很少见；同时它暂时也不支持Vue2；脚手架功能中相较于Vue-Cli有一定删节；最后就是开发与构建工具不同可能会导致一些程序页面构建后与开发服务器上的内容不一致。

§2.2.2.2 Element-Plus

Element-Plus是一套基于Vue实现的常用组件库，它提供了较为丰富的PC组件。使用这些封装好的组件能够一定程度上减少开发者自己对于常用组件的再实现与重封装，从而提升开发的速度与稳定性。

Element-Plus的设计原则一共有四条：

- 一致 Consistency: 这表示不但组件的流程与逻辑与生活中使用的一致，而且所有元素和结构亦保证有一致的风格与逻辑
- 反馈 Feedback: 表示用户操作与页面状态都可以让用户感知到不同，从而让用户对网页状态有清晰的认知
- 效率 Efficiency: 说明组件的操作流程直观清晰，用户可以快速而直接地认知到各个结构的用途而不是花额外的时间回忆
- 可控 Controllability: 所有操作都交由用户自身来决策，而且用户可以对已完成或正在进行的操作撤销或终止

Element-Plus受到ES2018以及以上的浏览器支持，如果需要支持旧版本的浏览器，则需要用到Babel或者其他工具进行版本控制。

§2.2.3 Flask

Flask是由Python实现的轻量化Web框架，它的实现是先基于底层的HTTP与Web服务器功能的封装，再使用WSGI(Python Web Server Gateway Interface)来建立Web应用。这样的结构让Flask在传递请求前，需要先将HTTP报文转换为WSGI所需的字典、响应头部的结构体，前者包含请求的全部信息，而后者则是将要调用的函数。而请求信息则由一个显示应用对象处理，在中小型的Flask后端中，大部分的接口都可以定义在这一个对象中。使用显示对象的原因是在Python中，隐式对象只能包含一个实例，所以使用显示对象能够让应用程序集中在一个文件中。这样的设计让小型的Flask服务不会过于臃肿，而且使用者可以使用寥寥几十行代码构造服务，整体结构也相当清晰。而且相当一部分人工智能模型都在Python中有较为简单与便携的实现方式，使用Python来实现后端可以更为简单地嵌入这些模型，让前后端使用的语言数量减少，从而让系统更加健壮。

§2.3 生存分析方法

生存分析是一系列统计分析方法，用于探讨人在特定情况下的，及生存时间的分布由于时间或其他因素的变化趋势。但是生存分析并不仅仅可以用在医学领域，它还可以在商业等多种环境中使用。比如使用生存分析可以探讨会员、订阅等机制

的用户使用情况，并让厂商对于如何留住自身的客户有一定作用。其中生存时间并非单纯表示表示对象的存活时间，在医学数据中，由于数据需要得到用户的允许才能使用，往往数据的获取都是通过随访得到的。这意味着对象可能生存了更长的时间，而我们的数据只能确认对象在一段时间的存活，这种现象被称为数据的右删失。而这需要通过模型的修改或者后期调整来去除。

§2.3.1 KM曲线

Kaplan-Meier曲线是用来进行时向统计的良好方法，它可以用来评估患者群体的健康状况和治疗效果。在生存分析中，我们可以将用户群体按照特征的不同，比如年龄段、肿瘤大小等，分为多个群体，通过比较群体间生存率的区别，我们可以对特征对于患者的影响有较好的初步结论。同时，KM曲线可以用来判断数据是否能使用COX比例风险模型。

§2.3.2 Cox单/多因素分析

Cox单因素分析一般用来研究单变量中的各个值的关系，它通过类似多项式模型的方法来分析该自变量对因变量的影响，而不纳入其他变量的影响。它主要用来研究该因素对于死亡风险的影响程度。Cox模型定义了风险函数，用来表示一个实体具有相应自变量值时，因变量发生的概率。而Cox多因素得到的p值用来检验该变量对于因变量的影响是否显著，一般p值在0.05以下则表示影响显著，可以纳入。Cox多因素分析则与单因素相仿，研究多个变量对于结果变量的影响程度。

§2.4 其他预后模型

§2.4.1 决策树

决策树是一种在医学中常用的分类回归模型，它使用树形结构表示决策过程。树上的每一个节点都表示特定范围的特征值，而树的每一个叶子节点都对应一个特定的类。决策树的建立过程中，模型首先把全体数据分为大小相近的若干个子集，然后不断在这些子集中选择最佳的子集，让这些子集具有尽量多的相似特征，直到回归达到最大深度停止。决策树可以同时在分类问题与回归问题上使用。

§2.4.2 随机森林

随机森林结合了决策树与随机梯度下降算法，随机地从数据集中抽样，组合成多颗决策树。具体地来说，模型为每棵决策树划分了它使用的数据样本空间，并使用梯度下降算法让它在样本空间上训练，随后得到一系列决策树组成的森林。由于

随机森林分为多颗树训练，在样本数较大的大规模数据集中训练速度快于单棵决策树，同时也可以减小样本的方差。随机森林在NLP、医学诊断学、金融分析学中都有较为广泛的应用。

§2.4.3 聚类+分类

先使用聚类方法得到带标签的新数据，分析聚类得到的患者表型对应集群的生存表现，并根据这个无监督学习的标签分类来预测患者生存率是本文根据文献[2]使用的新方法。在这个模型中，聚类方法的引入让模型的分类更有可解释性，同时不同患者的表型数据虽然在生存率方面会有不同，但是相近的特征依然存在，从而让模型对于相近的数据会有更相似的结果，连续性更好。相比单纯使用二分类模型的概率，该模型也可以使用聚类统计得到的概率作为参考，整体上得到的分类结果更加细致。

第3章　子宫肉瘤预后模型的数据提取

§3.1　数据源介绍

数据提取过程如图3.1所示：

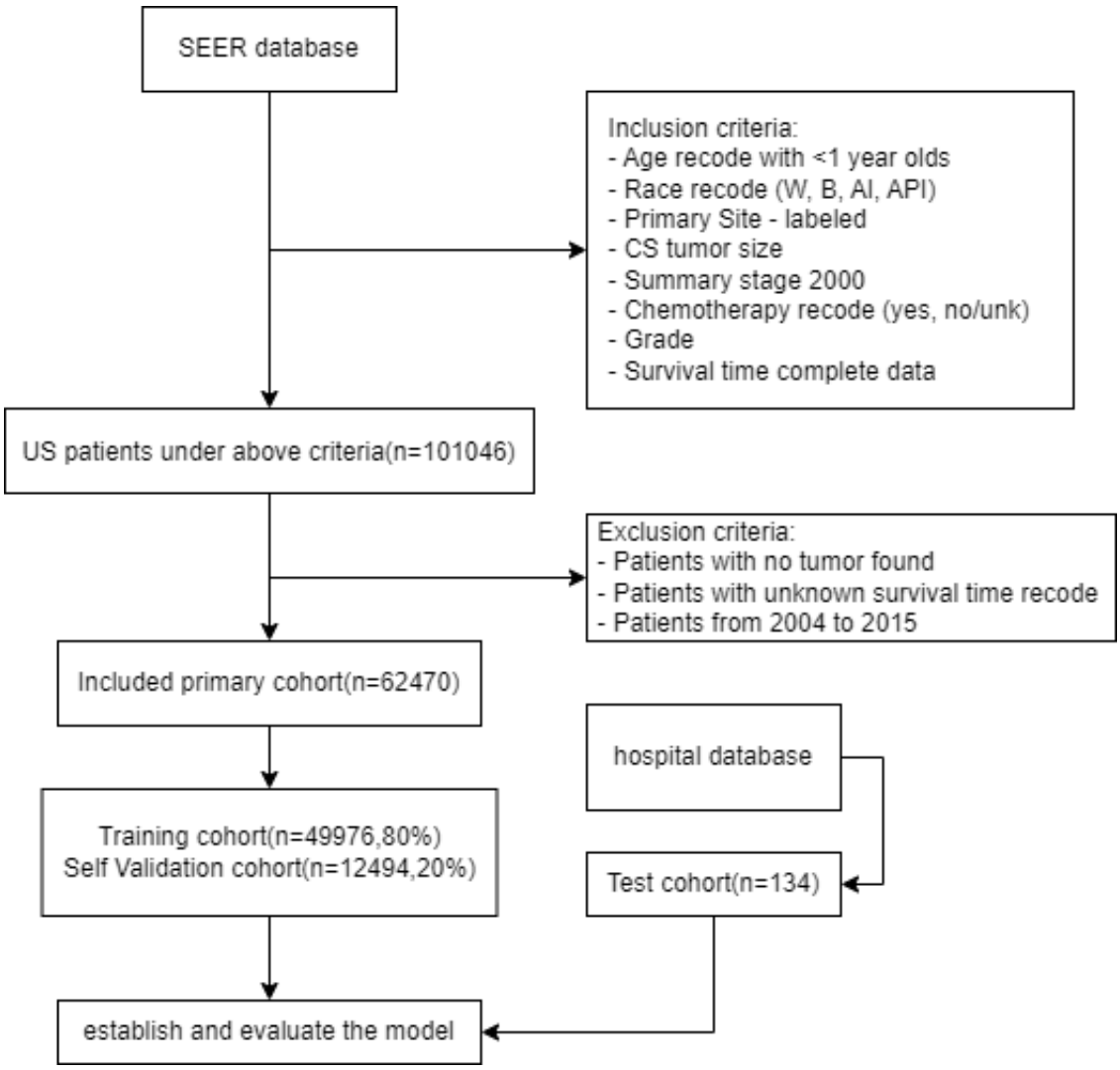


图 3.1　数据提取流程图

由图可见，在SEER数据库中，首先提取从SEER数据库中提取数据，各字段信息将在后面的章节阐释如何筛选得到，最终得到数据101046条。然后通过排除标准，将SEERStat软件中中不能进行排除的部分不需要的数据进行脚本审查，最终分离得到需要的62470条训练用数据。将这些数据进行裁切，最后得到的是49976条训练数据与其附带的自验证数据12494条。

另一方面，由于SEER数据库中的数据存在一定局限性，其只包含被国立医学研究院认可并被SEER接收的数据。这意味着SEER的数据可能由于各地医疗情况与政策差异，不包含某些地区与特定人群的数据，如果文章使用SEER数据库进行验证，无疑会生成部分不够准确的研究内容，甚至影响将来的研究方向。因此，我从医院数据库中进行采集、梳理和标记，最后筛选得到134条验证数据，以供之后验证使用。这些数据可以验证或补充 SEER 数据，以确保研究结果的可靠性和准确性。而且从医院得到的数据也能帮助将来的研究者更好地了解特定疾病的治疗方案与效果，这些信息也会对将来的医学研究与实践具有重要的意义。

因此，使用 SEER 数据和从医院得到的数据都是重要的，以确保医学研究的结果更准确、全面和可靠。两者结合使用可以提高研究结果的可靠性和准确性，并为医学研究和实践提供更准确和有用的信息。

§3.1.1 SEER癌症数据库

SEER（Surveillance, Epidemiology, and End Results）计划提供癌症统计信息，目的是对癌症数据提供完整的监控数据，减少医疗系统中的癌症负担。SEER由NCI癌症控制和人口科学部（DCCPS）的监测研究计划（SRP）支持。SEER数据被认为是美国癌症研究的重要资源之一，因为它包括了来自不同地区的多种癌症类型的发病率和死亡率，以及提供了大量患者的年龄、性别、地理位置和治疗信息等详细信息，同时它也是国际公认的最大最成体系的癌症数据库之一。它起源于1971年美国国家癌症法（NCA）对于建立一个体系化数据库来收集、储存、分析和分发癌症相关数据，以用于支持、预防、诊断和治疗癌症的研究。SEER的病例收集从1973年1月1日开始，在美国的几个地理区间上进行诊断和提取。50年来SEER数据收集范围越来越大，同时内容也不断完整化、规范化，如今收集到的总数据已占美国人口的一半。大量来自不同种族和年龄段的癌症数据为研究者详细分析癌症提供了很大的便利。该数据集的优点是包含了多个癌症类型，且数据集中的样本数量较大，这使得研究人员可以更好地研究不同癌症类型的特点和趋势。

近年来，越来越多的研究人员开始使用该数据集来研究癌症分类和预测，每年大量有关癌症分类和预测的文章在各类期刊上发表，医学人士往往使用列线图等来提供可以在临床上进行定量使用的R语言模型，该数据集也被逐渐用于研究不同癌症类型的生物学特征和发病机制。

§3.1.2 医院验证数据库

医院数据由我在医生老师们的指导下得到，如下图所示：

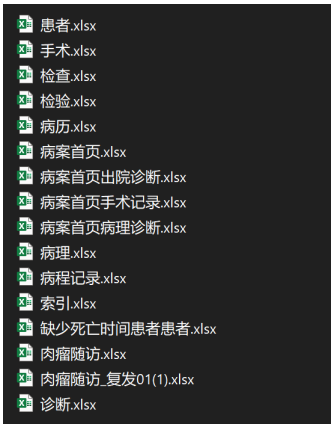


图 3.2　医院原数据图

该数据库提供了患者的病例、检查等数据，在对其中字段内容进行遍历后，我比较了了训练数据中与验证数据中共同存在的一部分内容，并按照第4章分析结果确定最终纳入的变量。

§3.2　数据字段对应编码与含义

从SEER数据集中，根据分析因素对生存率的相关性选择了下列字段：

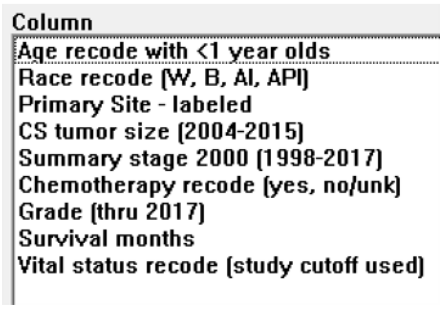


图 3.3　SEERStat中选择的变量数据图

其中方括号内的时间段是由于编码方式改变产生，在不同的时间段有不同的编码方式，所以文章最终取交集2004-2015年范围间。关于字段的解释如下：

- 年龄（例：60-64 years）
- 种族（白人、黑人、美国印第安人/阿拉斯加原住民、亚洲人或太平洋岛民）
- 原发部位（子宫内膜、子宫肌层等）
- 肿瘤大小（最大直径）
- 肿瘤分期（Localized、Regional、Distant）
- 是否化疗（是、否/未知）
- 肿瘤分级（I、II、III、IV）
- 生存时间相关信息

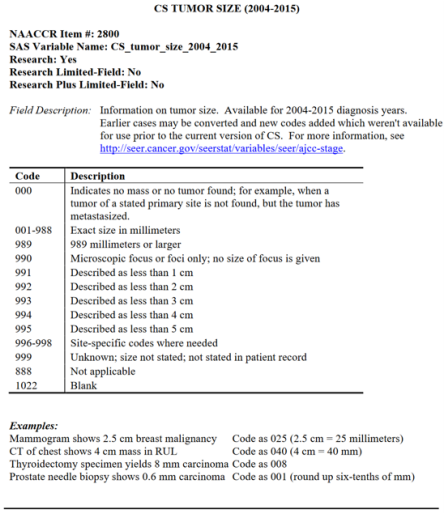
文章中使用到的字段主要如上述列表所示。

首先，年龄使是以5年为一个单位所表示的，年龄重码变量是基于诊断时的年龄（单年年龄），使用的分组是由年龄决定的，年龄重新编码变量中使用的分组是由患者数据中的年龄分组决定的。这个重新编码在年龄上有19个年龄组重码变量中有19个年龄组（¡1岁，1-4岁，5-9岁，...，85岁以上）。

第二个字段是种族，在SEER数据库中，主要有六个选项，他们分别是白种人、黑种人、美国印第安人/阿拉斯加原住民、亚洲人或太平洋岛民、其他未说明的（1991年以上）。考虑到在测试数据集中，我们使用的是中国人作为主要的测试患者，一开始我只使用了亚洲人或太平洋岛民的数据，但是在之后的模型建立中，我发现如果只使用亚洲人和太平洋岛民作为训练数据的话模型的效果并不好。可能的原因是亚洲人和太平洋岛民中存在着显出的差异，另一种可能的解释是数据量的不足。考虑到该字段频数的差异我最后使用了白种人黑种人和亚洲人或太平洋岛民作为训练集中开始算的参数选项。

文章使用的第三个字段是原发部位（Primary Site），这个字段主要表示病发的部位。这个字段提供了ICD-O-3规范的主要部位代码和一个描述性的主要部位标签该标签是首选的ICD-O-3加粗的名称，其他部位或子部位包含在代码中，但没有反映在代码中，本文使用的子宫肉瘤数据都包含ICD-O-3标签。编码中可能还包括其他部位或子部位，但没有反映在首选标签中。诊断年份在1992年之前的病例从早期版本转换为了ICD-O-3。这里亦只选取了四个选项，见图3.5中。

第四个字段是肿瘤大小，肿瘤大小指的是肉眼所见肿瘤的最大直径。肿瘤大小适用于2004-2015年的诊断年份。早期的病例部分会被转换，并增加新的编码以加以匹配，这些编码在目前的CS版本之前是不可用的，而且在当前版本的CS之前无法使用的新编码。关于肿瘤大小的详细说明见图3.4。



CS TUMOR SIZE (2004-2015)

NAACCR Item #: 2800
SAS Variable Name: CS_tumor_size_2004_2015
Research: Yes
Research Limited-Field: No
Research Plus Limited-Field: No

Field Description: Information on tumor size. Available for 2004-2015 diagnosis years. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS. For more information, see http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage.

Code	Description
000	Indicates no mass or no tumor found; for example, when a tumor of a stated primary site is not found, but the tumor has metastasized.
001-988	Exact size in millimeters
989	989 millimeters or larger
990	Microscopic focus or foci only; no size of focus is given
991	Described as less than 1 cm
992	Described as less than 2 cm
993	Described as less than 3 cm
994	Described as less than 4 cm
995	Described as less than 5 cm
996-998	Site-specific codes where needed
999	Unknown; size not stated; not stated in patient record
888	Not applicable
1022	Blank

Examples:

Mammogram shows 2.5 cm breast malignancy	Code as 025 (2.5 cm = 25 millimeters)
CT of chest shows 4 cm mass in RUL	Code as 040 (4 cm = 40 mm)
Thyroidectomy specimen yields 8 mm carcinoma	Code as 008
Prostate needle biopsy shows 0.6 mm carcinoma	Code as 001 (round up six-tenths of mm)

图 3.4　SEER数据库中CS Tumor Size字段的详细说明图

肿瘤分期是指对恶性肿瘤进行分期分类的过程，它用于描述肿瘤的扩散程度和位置等特征，以便更好地评估肿瘤的严重程度和治疗前景。肿瘤分期通常由国际癌症联合会制定，其分期系统根据不同的肿瘤类型和特定癌症的组织学和生物学特征而有所不同。SEER主要使用的是TNM分期系统。TNM分期系统是最常用的肿瘤分期系统之一，它根据肿瘤的大小（T）、淋巴结的受累情况（N）以及转移情况（M）来分期。这里主要分为三种情况，分别表示局部区域与远端三种类型。其中局部表示肿瘤仅限于原始的发生位置，没有扩散到周围组织或器官。区域表示肿瘤扩散到周围组织或病变器官，但没有扩散到远处的器官或淋巴结。远端则表示肿瘤已经扩散到其他远隔的器官或淋巴结。肿瘤分期表示了肿瘤扩散的进展过程，是对患者生存率有较大影响的一个影响因子。

化疗与否标识了患者是否使用了化疗作为辅助治疗手段。这里要主要的一点是由于患者隐私和机构编码原因，使用的两个编码分别是“是”和“否/未知”，这样的编码会对准确率产生一定的影响。

肿瘤分级是指对恶性肿瘤进行分期的的一种方法，它可以根据肿瘤的大小、形状、密度和边缘等特征来评估肿瘤的恶性程度。肿瘤分级通常由医生进行视觉评估，也可以通过计算机辅助断层扫描（CT）、磁共振成像（MRI）和其他影像学技术来进行辅助评估。这里一个分为四个等级：

- Ⅰ级（G1）：肿瘤细胞和组织看起来最像健康的细胞和组织，称为分化良好的肿瘤。肿瘤被认为是低级别的，恶性程度低。
- Ⅱ级（G2）：肿瘤细胞和组织有些异常，看起来不像正常的细胞和组织，并且比正常的细胞生长更快，称为中度分化的肿瘤。肿瘤被认为是中等级别的，恶性程度相对较高。
- Ⅲ级（G3）：肿瘤细胞和组织看起来非常异常，称为低分化的肿瘤。肿瘤被认为是高等级的，恶性程度更高。
- Ⅳ级（G4）：肿瘤细胞和组织看起来最异常，称为未分化的肿瘤。这类肿瘤被认为是最高等级的，恶性程度最高，生长和扩散更快。GX，表示医生无法评估等级，也称为未定等级。

以上数据中，肿瘤大小、肿瘤分期、肿瘤分级与生存相关内容外的数据都可以直接使用SQL语句或Python脚本简单提取，这里我使用了ipynb文件作为提取方案，以适应研究过程中的大量更改。生存相关数据在数据库中并没有记录，所以我使用脚本筛选了所有提供联系方式的患者的信息，在医院方面进行随访后对得到的数据进行提取。

而肿瘤大小、肿瘤分期与分级数据没有直接字段提供，只能从病理分析中的诊断文本中提取，我首先筛选了其他字段记录完整的患者，并在这些数据的诊断信息

中提取。肿瘤大小这里选取的是左附件区和右附件区中的最大肿瘤的最大直径，所以我编写了相应的Python 简单NLP算法，文本的左附件右附件的划分，并在这两段中分别找到用以描述肿瘤的大小。由于还有关于回声区和输卵管长度等的干扰，必须严格找到肿瘤对应的大小。最后在得到的所有肿瘤大小中找到最大值并返回。无匹配内容的描述只能得到空值作为回应，所以最后我对所有数据进行了人工校验。而肿瘤分期与分级同理，先在文本中检索是否有直接指明的描述，如果没有则按照上文中的定义来确定相应的内容。

通过上述的处理，初步筛选得到的300条数据在去除上述内容的缺失后一共是134条测试数据。

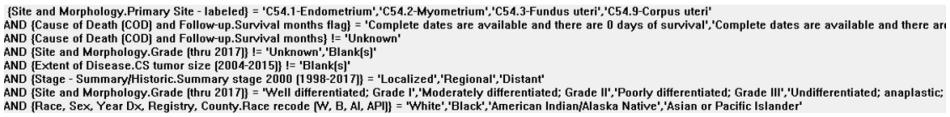


图 3.5　SEERStat中选择的变量排除条件图

第4章　参考文献

114

致　谢

衷心感谢导师沈文枫教授对本人的精心指导。感谢上海大学开源社区提供的Latex模板。

参考文献

[1] 中国医师协会微无创医学专业委员会妇科肿瘤专业委员会(学组)中国优生科学协会生殖道疾病诊治分会, 中国优生科学协会肿瘤生殖学分会. 子宫癌肉瘤诊治中国专家共识(2020年版) [J]. 中国癌症防治杂志, 2020, 12 (6): 599–605.

[2] Zhou X, Nakamura K, Sahara N, et al. Exploring and Identifying Prognostic Phenotypes of Patients with Heart Failure Guided by Explainable Machine Learning [J/OL]. Life, 2022, 12 (6). https://www.mdpi.com/2075-1729/12/6/776.

附录 A　经典不等式

论文中用到的经典不等式.

(Hölder Inequality) 设 $a_i \geq 0, b_i \geq 0, i = 1, 2, \cdots, n$, 且 $p > 1, q > 1$ 满足 $1/p+1/q = 1$. 则有

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q \right)^{\frac{1}{q}},$$

等号成立当且仅当存在一个常数 c 满足 $a_i^p = c b_i^q$.

(PM Inequality) 设 $x_1, x_2, \ldots, x_n$ 是 n 个非负实数. 如果 $0 < p < q$, 那么

$$\left(\frac{x_1^p + x_2^p + \cdots + x_n^p}{n} \right)^{\frac{1}{p}} \leq \left(\frac{x_1^q + x_2^q + \cdots + x_n^q}{n} \right)^{\frac{1}{q}},$$

等号成立当且仅当 $x_1 = x_2 = \cdots = x_n$.

(AM-GM Inequality) 设 $x_1, x_2, \ldots, x_n$ 是 n 个非负实数. 则有

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n},$$

等号成立当且仅当 $x_1 = x_2 = \cdots = x_n$.