МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

Кафедра прикладної математики та моделювання складних систем

Допущено до захисту
Завідувач кафедри ПМ та МСС

_____ Коплик І.В.
(підпис)

«__» _____ 20__ р.

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня «бакалавр»

спеціальність 113 «Прикладна математика»

освітньо-професійна програма «Прикладна математика»

тема роботи **«Математичне моделювання впливу основних характеристик ракових пухлин легень на ймовірність повного одужання пацієнта»**

Виконавець

студент факультету ЕлІТ

Чухно Віра Юріївна _____

Науковий керівник

старший викладач, кандидат фіз.-мат. наук

Дворниченко Аліна Василівна _____

Суми – 2022

СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

Факультет	**електроніки та інформаційних технологій**
Кафедра	**прикладної математики та моделювання складних систем**
Рівень вищої освіти	бакалавр
Галузь знань	**11 Математика та статистика**
Спеціальність	**113 Прикладна математика**
Освітня програма	**освітньо-професійна** «Прикладна математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри ПМ та МСС
Коплик І.В.
«___»_____20___р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧА ВИЩОЇ ОСВІТИ

Чухно Віра Юріївна

1. Тема роботи Математичне моделювання впливу основних характеристик ракових пухлин легень на ймовірність повного одужання пацієнта

   Керівник роботи Дворниченко А.В., старший викладач, кандидат фіз.-мат. наук

   затверджено наказом по факультету ЕлІТ від «16»лютого 2022р. №0146-VI
2. Термін подання роботи студентом «21» червня 2022р.
3. Вихідні дані до роботи: дані пацієнтів з раком легень, отримані на базі Медичного Інституту СумДУ

4. Зміст розрахунково-пояснювальної записки (перелік питань, що їх належить розробити): Провести літературний огляд за темою: «Модель пропорційних ризиків Кокса», формування бази даних для аналізу, програмна реалізація регресії Кокса, аналіз впливу характеристик раку легень на ймовірність появи рецидиву, смерті та повного клінічного одужання.
5. Перелік графічного матеріалу: візуалізація теоретичного матеріалу; візуалізація коефіцієнтів регресії Кокса; візуалізація отриманих результатів.
6. Консультанти до проекту (роботи), із значенням розділів проекту, що стосується їх

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання «02»травня 2022р.

КАЛЕНДАРНИ ПЛАН

№ п/п	Найменування роботи, заходи	Термін виконання роботи	Примітка
1	Ознайомитися з теоретичними основами регресійного аналізу Кокса.	02.05 - 13.05	Виконано
2	Сформувати базу даних для проведення регресійного аналізу.	14.05 - 24.05	Виконано
3	Провести регресійний аналіз за методом Кокса.	25.05 - 15.06	Виконано

Здобувач вищої освіти _____ Чухно В.Ю.

Керівник роботи _____ Дворниченко А.В.

РЕФЕРАТ

Кваліфікаційна робота: 42 с., 12 рисунків, 14 формул, 18 джерел.

Мета роботи: з використанням регресії Кокса встановити вплив віку пацієнтів та основних характеристик ракових пухлин легень на ймовірність виникнення рецидиву, імовірність летального випадку та імовірність клінічного одужання.

Об'єкт дослідження: виявлення впливу факторів на появу рецидивів та смертності хворих на рак легень.

Предмет дослідження: регресійна модель Кокса.

Методи навчання: регресійна модель Кокса

Для проведення регресійного аналізу було сформовано базу даних пацієнтів з раком легень на базі Медичного Інституту СумДУ.

Ключові слова: РЕГРЕСІЙНИ АНАЛІЗ, ФАКТОРИ РИЗИКУ, АНАЛІЗ ДОЖИВАННЯ.

ЗМІСТ

ВСТУП

Рак легень, також званий карциномою легень, є типом раку, який викликає неконтрольовану швидкість росту клітин в легеневих тканинах. Він є однією з провідних причин смертності в усьому світі.

Дослідження основних характеристик ракових пухлин легень можуть допомогти розробити кращі методи лікування, підвищивши виживання та якість життя пацієнтів. Для оцінки впливу характеристик на результат, тобто появу рецидиву, смерть чи одужання пацієнта, використовують регресійні моделі. Найпопулярнішими регресійними моделями є модель пропорційних ризиків (регресія Кокса) як напівпараметричний метод, модель прискореного часу відмови як параметрична модель, оцінка Каплана-Майєра як непараметрична модель.

Щоб проаналізувати фактори ризику виживання раку легень, у цій роботі використовується регресійний метод Кокса. Цей метод відноситься до методів аналізу виживання, відомого як аналіз часу до події. Програми аналізу виживання дуже великі: наприклад, їх можна використовувати для визначення рівня виживання популяції або порівняння виживання двох або більше груп.

Регресійний аналіз Кокса є дуже популярним і широко використовуваним методом. Розроблений Девідом Коксом у 1972 році, його мета — оцінити одночасно вплив кількох факторів на виживання. Також відомий як модель пропорційних ризиків, його важливість є вирішальною і має багато застосувань у дослідженнях сучасної медицини.

Потрібно використовувати більш підходящі та вдосконалені методи, які детально можуть відобразити фактори, які можуть вплинути на виживаність пацієнтів з раком легень.

У роботі детально описано метод пропорційних ризиків Кокса, який використовувався для аналізу впливу характеристик раку легень на ризик появи рецидиву, смерті та одужання.

ПОСТАНОВКА ЗАДАЧІ

1) Ознайомитись з теоретичними основами методу аналізу визначення незалежного впливу факторів ризику на швидкість настання досліджуваної події методом регресії Кокса.
2) Сформувати базу даних для проведення регресійного аналізу на основі отриманих даних з інших джерел (Мед Інститут СумДУ)
3) Провести регресійний аналіз за методом Кокса з метою встановлення впливу віку пацієнтів та основних факторів ракових пухлин легень пацієнтів на імовірність виникнення рецидиву, імовірність летального випадку та імовірність клінічного одужання.

РОЗДІЛ 1 ОГЛЯД ЛІТЕРАТУРИ

1.1 Аналіз виживання

Аналіз виживання є важливим методом, за допомогою якого відстежується час до настання події; час називається часом виживання. Аналіз виживання використовується для оцінки цих спостережуваних часів виживання, для порівняння відмінностей між ними, для оцінки факторів ризику та для кількісної оцінки їх впливу на результуючий час виживання; у зв'язку з цим ми також можемо вирішити подальший прогноз часу виживання. Все це можна дослідити за допомогою статистичних методів, особливо регресійних моделей. Завдяки своїм особливостям і легкому тлумаченню моделі стали дуже популярними і часто використовуються в цьому плані, але, з іншого боку, необхідно усвідомлювати, що вони також мають ряд передумов, які необхідно виконати. Недотримання цього може призвести до того, що моделі та отримані з них результати будуть невірними та невірними. Таким чином, висновки, зроблені з моделі, не можна вважати остаточними, якщо ми не переконаємося, що всі припущення моделі були виконані. Для цього ми використовуємо інструменти регресійної діагностики, за допомогою яких ми можемо не тільки достовірно перевірити, чи дотрималися всі ці припущення, але й оцінити, наскільки модель підходить і чи правильно (у межах своїх можливостей) вона описує вихідні дані [1].

Оцінка часу до настання певної події має довгу історію, яка сягає 18 століття, коли швейцарський математик Даніель Бернуллі проаналізував захворюваність і смертність, щоб показати важливість та переваги вакцинації від віспи [1]. Нині аналіз часу до події або аналіз виживання став частиною багатьох галузей, приклади яких можна знайти в медицині, біології, епідеміології, економіці, демографії та технічних галузях. Основною характеристикою, яка відрізняє дані про виживання від інших типів даних, наприклад, від класичного сприйняття смертності як частки померлих пацієнтів у клінічних застосуваннях, є їх часовий компонент. Дані про виживання не тільки відображають інформацію

про кількість або частку спостережуваних подій, але й повідомляють нам, коли подія сталася. Прикладом є дві групи пацієнтів з певним захворюванням та ідентичною часткою смертей через п’ять років після встановлення діагнозу, але виживання в яких може бути абсолютно різним, причому одна група демонструє вищу смертність незабаром після початку спостереження (наприклад, початку лікування). з подальшою тенденцією до зниження. , у той час як друга група, з іншого боку, демонструє нижчу смертність на початку спостереження і збільшення з часом. Якщо просто підсумувати частку смертей в обох групах, ми будемо позбавлені інформації про те, як змінюється розвиток ризику смерті в обох групах з часом [2].

Таким чином, аналіз виживання включає математико-статистичні методи оцінки часу до настання даної події, які, однак, відрізняються від стандартної статистичної оцінки, оскільки дані, що описують час до настання події, мають ряд особливостей. Іншими словами, ми вибираємо методи аналізу виживання, коли нас цікавить дослідження не лише частоти виникнення спостережуваної події, а й плину часу цього явища. Виходячи з мети обробки наших даних, методологію аналізу виживання можна розділити на три групи [3]:

1. Описові методи. Метою використання цих методів є опис часу настання події, що спостерігається, у заданій сукупності суб’єктів чи об’єктів та визначення ймовірності виживання без настання даної події в окремі моменти часу.
2. Порівняльні методи. Порівняльні методи використовуються, коли ми хочемо з’ясувати, чи відрізняється даний набір суб’єктів чи об’єктів за настанням спостережуваної події від очікуваного значення, чи відрізняються один від одного окремі групи суб’єктів.
3. Моделі виживання. Використання стохастичного моделювання в аналізі виживання допомагає нам вирішити питання про те, чи залежить тривалість виживання суб’єкта чи групи суб’єктів від однієї чи кількох спостережуваних змінних, чи ця залежність розвивається з часом. Іншими словами, мова йде про ідентифікацію змінних, які впливають

на ймовірність того, що спостережувана подія відбудеться рано чи пізно.

Ключовим елементом аналізу виживання є, звичайно, визначення події, що цікавить. Вона повинна бути чітко визначена, що також пов’язано з тим, що вона також повинна бути легко помітною або виявлена. Наша здатність якомога легше спостерігати або записувати настання спостережуваної події пов’язана з точністю результуючого часу виживання групи досліджуваних. Приклади подій, які підходять для оцінки за допомогою методів аналізу виживання, включають смерть пацієнта, прогресування раку або рецидив після попереднього безсимптомного періоду (також званого ремісією захворювання). Смерть пацієнта зазвичай відповідає вимозі легкої та правильної вимірювальності, але складніша ситуація, наприклад, із згаданим прогресуванням захворювання, де точне визначення дати прогресування може бути зовсім непростим. Неточність у визначенні настання цільової події може призвести до викривлення результатів, явища, якого ми намагаємося максимально уникати в біостатистиці, оскільки зазвичай це призводить до помилкових висновків та інтерпретацій. Не менш важливим компонентом в оцінці часу виживання є також визначення початкової точки моніторингу, з якої розраховується результуючий час виживання. Вибір конкретного моменту часу принципово впливає на інтерпретацію результатів. Для ілюстрації візьмемо інший приклад із медицини: існує велика різниця між оцінкою виживання від дати діагностики захворювання, коли пацієнт може пройти ряд методів лікування, розділених на ізольовані секції лікування, та оцінкою виживання від дати початку. конкретного препарату. У першому випадку дуже важко визначити вплив окремих препаратів або лікувальних процедур на загальний час до спостережуваної події (часто до смерті), що, таким чином, відображає якість відповідного сегмента охорони здоров’я, а не якість разова підготовка. У другому випадку, однак, ми відстежуємо час до спостережуваної події (часто до прогресування захворювання), який має чітко відповідати введенню конкретного препарату і, отже, має вказувати на його здатність лікувати захворювання [4].

Дані про виживання описують проміжок часу від моменту виникнення до кінцевої точки, що цікавить.

Характеристики даних про виживання включають:

1. відповідне визначення часу походження для кожного суб’єкта дослідження, тобто часу від початку дослідження;
2. відповідне визначення кінцевої події (невдачі), наприклад, смерть, рецидив, відновлення після операції або серцевого нападу та розвиток захворювання;
3. суб'єкти дослідження повинні бути порівнянними за моментом їх походження, тобто кожен зареєстрований індивідуум буде спостерігатися від вихідної дати (наприклад, у разі раку; дати діагнозу або дати операції) до дати смерті або припинення дослідження ;
4. дані про виживання ніколи не можуть бути негативними, оскільки це дані про час реакції [5].

Під часом ми маємо на увазі роки, місяці, тижні чи дні від початку зарахування до дослідження. Основним обмеженням даних від часу до події є можливість того, що подія не відбудеться у всіх суб’єктів протягом конкретного періоду дослідження.

Основні цілі аналізу виживання полягають у:

1. оцінка виживання та/або ризиків за наявними даними;
2. порівняння виживання та/або функцій ризику;
3. оцінка взаємозв'язку різних факторів із часом виживання [4, 5].

1.2 Методи аналізу виживання

Методами оцінки виживання – це методи для оцінки часу до події, що цікавить, наприклад, смерті, рецидиву хвороби, розвитку побічної реакції та нового захворювання. Час виживання визначається як час до цікавих подій, таких як смерть, рецидив хвороби, безробіття та завершення завдання. Основна

характеристика часу виживання полягає в тому, ми не можемо відстежити точний час виживання для тих, хто ще живий наприкінці дослідження, або для тих, кого втрачено із спостереження протягом періоду навчання.

Було розроблено багато статистичних методів для оцінки функцій виживання, порівняння кривих виживання між двома групами та моделювання даних про виживання за допомогою регресії для зв'язку з факторами ризику, такими як демографічні та клінічні предиктори.

В аналізі виживання непараметричний статистичний висновок ширше використовується для оцінки функції виживання та порівняння кривих виживання між двома або більше групами. Наприклад, як оцінка Каплана-Майєра (KM) для функції виживання, і тест логарифмічного рангу, для порівняння функцій уцілілих, отримані за допомогою непараметричного підходу. Однак, якщо припущено або заздалегідь визначено відповідний розподіл даних про виживання, параметричний підхід є більш прийнятним. Коли основний інтерес становить зв'язок часу виживання з різними факторами ризику, найпопулярнішою моделлю є регресія Кокса, заснована на напівпараметричному підході, оскільки вплив предикторів на рівень небезпеки визначається параметрично, тоді як небезпека базової лінії функція не визначена. Загалом, усі підходи до аналізу виживання повинні враховувати механізм цензури, коли робиться статистичний висновок [5, 6].

РОЗДІЛ 2 МАТЕМАТИЧНИЙ ЗМІСТ МЕТОДУ

2.1 Функція виживання та функція ризику

Дані про виживання зазвичай описуються та моделюються за допомогою двох пов’язаних функцій, а саме функцій виживання (функцій ймовірності виживання) та функцій ризику.

Час виживання представлено невід’ємною випадковою величиною T. Щоб відрізнити суб’єкти з цензурованим і фактичним часом виживання, ми записуємо дані про виживання за допомогою випадкового вектору (T,C), де C є випадковою величиною, що представляє індикатор цензури. Якщо T представляє фактичний час виживання або час до настання спостережуваної події, то випадкова величина C 0 приймає значення 1. Якщо, з іншого боку, час виживання суб'єкта T піддається цензурі, і ми не спостерігаємо спостережувану подію, то випадкова величина C набуває значення 0. Оскільки T випадкова величина, її ймовірнісна поведінка може бути описана розподілом виживання або однією з наступних функцій [6]:

1. *Функція розподілу* $F(t)$ виражає ймовірність того, що чисельна реалізація випадкової величини T не перевищує задане значення на реальній осі t, що іншими словами означає, що час виживання даного суб’єкта буде меншим або рівним значенню t. Ми можемо записати це визначення як [7]:

$$F(T) = \mathrm{P}(T < t) \tag{2.1}$$

2. *Функція щільності* вказує на ймовірність того, що спостережувана подія відбудеться з часом t, відповідно в заданий інтервал часу на реальній осі [8]. Щільність можна отримати або шляхом виведення функції розподілу, тобто як

$$f(t) = \frac{dF(t)}{dt}$$

або це може виражатися відношенням:

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t) \qquad (2.2)$$

3. *Функція виживання* $S(t)$ виражає ймовірність того, що випадкова величина T реалізується на реальній осі до заданого значення t, що означає, що час виживання суб'єкта буде більшим за вибраний час t [9]. Отже, функцію виживання можна записати так:

$$S(T) = P(T > t) = 1 - P(T \leq t) = \ 1 - F(T) \qquad (2.3)$$

Оскільки це ймовірність, функція виживання приймає значення лише від 1 до 0 (частіше виражається як 100% і 0%), коли значення 1 має функцію виживання в часі t = 0, а значення 0 - у момент виникнення останньої події, коли t теоретично може бути будь-яке число. Функція виживання завжди є функцією, що не зростає [10].

Іншою важливою характеристикою даних про виживання є так звана функція ризику. Вона виражає інтенсивність настання спостережуваної події в часі за умови, що суб'єкт вижив до часу, який можна записати так [11]:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq t + \Delta t | T > t) \qquad (2.4)$$

Функція ризику як безпосередня інтенсивність спостережуваної події є дуже важливою функцією, особливо в моделюванні виживання, однак кумулятивна функція ризику (функція кумулятивної небезпеки, інтегрована небезпека) є більш практичною для практичного опису виживання в група предметів [11,12]. Ми отримуємо кумулятивний ризик, інтегруючи функцію ризику (2.4) в часі, яку ми пишемо як:

$$H(t) = \int_0^t h(x)dx \qquad 2.5)$$

Функція кумулятивного ризику (2.5) відповідає загальному ризику спостережуваної події від початку моніторингу до часу t. Оскільки це ризик, а не ймовірність, функція (2.5), на відміну від наведеної вище функції виживання (2.3), не обмежується 1 [13].

Усі визначені вище функції, що описують імовірнісну поведінку випадкової величини T, математично еквівалентні, оскільки, знаючи одну з них, можна обчислити інші [4]. Взаємні обчислювальні зв'язки можна вивести за допомогою наступних міркувань. Частиною визначення функції ризику є умовна ймовірність або ймовірність настання події, що відстежується в інтервалі $t + \Delta t$, за умови, що це не відбулося до часу t. Враховуючи, що $P(A|B)$ теорему умовної ймовірності можна виразити як $P(A \cap B)/P(B)$. Виразимо умовну ймовірність у визначенні функції ризику, використовуючи наступне співвідношення [3]:

$$P(t < T \leq t + \Delta t \,|T > t) = \frac{P(t<T\leq t+\Delta t \cap T>t)}{P(T>t)} = \frac{P(t<T\leq t+\Delta t)}{P(T>t)} \qquad (2.6)$$

Якщо подивитися на визначення функції розподілу $F(t)$ та функцію виживання $S(t)$, то співвідношення (2.6) можна записати у вигляді:

$$\frac{P(t<T\leq t+\Delta t)}{P(T>t)} = \frac{F(t+\Delta t)-F(t)}{S(t)} \qquad (2.7)$$

Підставивши рівняння (2.7) у визначення функції ризику, задане рівнянням (2.4), знаходимо, що результат відповідає визначенню густини ймовірності випадкової величини T, яку ділимо на функцію виживання. Таким чином, (2.8) результатом зв’язок між функцією ризику (2.4), функцією виживання (2.3) та щільністю ймовірності випадкової величини (2.1) [13]:

$$h(t) = \frac{f(t)}{S(t)} \tag{2.8}$$

Співвідношення (2.8) можна додатково модифікувати, застосувавши правило для виведення складеної функції та правило для виведення натурального логарифма, отримуючи таким чином таке співвідношення [9]:

$$h(t) = -\frac{d}{dt}(\ln(S(t)) \tag{2.9}$$

Якщо підставити (2.9) у визначення сукупної функції ризику, задане (2.5), то отримаємо формулу, що документує прямий зв’язок між функцією виживання та функцією кумулятивного ризику, яка має вигляд [10]:

$$H(t) = -lnS(t)$$

або можна записати у вигляді:

$$S(t) = \exp(-H(t))$$

2.2 Модель пропорційних ризиків Кокса

Щоб пояснити зв’язок між виживаністю пацієнтів і пояснювальними змінними, доцільно використовувати статистичні моделі. Найбільш використовуваними в аналізі виживання є моделі пропорційних ризиків, за допомогою яких ми можемо не тільки описати цей зв’язок, але й порівняти виживання між групами пацієнтів із різними методами лікування та передбачити виживання.

Це досить популярний кількісний метод у соціальних науках та медицині, оскільки є досить простим при використанні у статистичних пакетах, проте вкрай трудомісткий у реалізації. Регресійний аналіз допомагає визначити, вплив тієї чи іншої незалежної змінної або навіть кілька таких змінних, на іншу незалежну змінну, що цікавить нас, а так само визначити ступінь цього, можливого, впливу.

Регресія Кокса, яку називають моделлю пропорційних ризиків (Cox proportional hazards model), вивчає залежність часу дожиття (survival time) від незалежних змінних (predictor variables). Цей напівпараметричний метод передбачає прогнозування ризику настання події (hazard risk) для об'єкта, що розглядається, і оцінює вплив незалежних змінних на цей ризик. При цьому ризик настання події є залежною від часу функцією і виявляє ймовірність настання події для об'єктів, що знаходяться в групі ризику. Ніяких припущень про вид функції інтенсивності/ризику немає, у цьому полягає непараметрична частина способу. Однак усі змінні повинні лінійно впливати на логарифм функції ризику настання події, що становить параметричну компоненту методу [7].

Об'єктом дослідження може бути індивід (пацієнт), для якого прогнозується ризик наступу події. Через те, що об'єкт вже потрапив під нагляд і досліджується, він автоматично входить у групу ризику, тобто. в будь-який проміжок часу з ним може статися подія, що цікавить дослідника. Суть визначення події полягає в тому, щоб з'ясувати, чи є ризик того, що об'єкт помре, одужає або з ним станеться інша подія, що цікавить дослідника, в аналізований період часу.

Також у модель включені незалежні змінні (predictors) – характеристики об'єкта (наприклад, вік, стать пацієнта, супутні захворювання та ін.), які можуть впливати на ризик настання події [8].

Регресія Кокса має принципову особливість у контексті побудови вибірки: її обсяг і структура до настання події можуть змінюватися. Це пояснюється тим, що час не є певним, вірніше, він визначений лише у тих об'єктів, у яких настала подія. Решта учасників вибірки показник часу залишається невідомим, тому що подія може не статися. Крім того, об'єкти спостереження можуть не підлягати участі в дослідженні у зв'язку зі зміною обставин. Таким чином, з часом дані можуть містити неповну інформацію – бути цензурованими (censored). Цензурування є практично універсальною властивістю даних аналізу виживання, найбільш поширеною формою якого є правостороннє цензурування - час спостереження закінчується або об'єкт видаляється з дослідження до того, як настане подія (наприклад, пацієнти можуть бути ще живі до кінця дослідження або виходять з-під спостереження з різних причин) [9]. Лівостороннім цензурування вважається у тому випадку, якщо початок перебування у групі ризику невідомий. Зрідка цензурування може бути інтервально-цензурованим, коли дотримуються обидві перші умови разом. Крім того, в моделі виживання цензурування має бути незалежним від можливого значення ризику для об'єкта, інакше результат буде спотворений [10].

При аналізі даних виживання дві функції представляють фундаментальний інтерес – це функція виживання (survivor function) та функція ризику (hazard function) [11].

Модель Кокса визначається як модель пропорційних ризиків із використанням функції ризику. Для *i*-го пацієнта його функція ризику, включаючи вплив незалежних пояснювальних змінних, може бути виражена у вигляді

$$h(t) = h_0(t)\exp(x_i\beta_1 + x_i\beta_2 + \cdots + x_i\beta_p) = h_0(t)e^{\sum_{i=1}^{p} x_i\beta} \qquad (2.10)$$

де $x_1, \dots, x_p$ – незалежні змінні (предиктори), функція $h_0(t)$ представляє основну функцію ризику, спільну для всіх суб'єктів, $\beta_1, \dots, \beta_p$ – коефіцієнти регресії. Хоча ми розглядаємо основну функцію ризику як функцію часу, її точна форма в моделі не вказана. Причиною цього кроку, який є дуже вигідним, оскільки він звільняє нас від необхідності точно вказувати характер даних про виживання, є спеціальна процедура оцінки коефіцієнтів регресії моделі, яка не залежить від конкретної форми $h_0(t)$ [10,11].

Вплив пояснювальних змінних на ризик спостережуваної події виражається через індивідуальні коефіцієнти регресії $\beta_k, k \in (1, \dots, p)$, які вказують на зміну ризику спостережуваної події , пов'язану зі зміною значення пояснювальної змінної *k*. Точніше, він представляє значення, на яке збільшується натуральний логарифм функції ризику, якщо значення *k*-ої мінної збільшується на одну одиницю за умови, що інші пояснювальні змінні не змінюються [12]. Позитивний знак коефіцієнта регресії означає, що ризик спостережуваної події більший у пацієнта з більшим значенням, що відповідає пояснювальній змінній. Навпаки, негативний коефіцієнт говорить нам про те, що дана пояснювальна змінна з більшим значенням має захисну дію, тобто. ризик настання спостережуваної події нижчий. Відношення β_k і функції ризику можна виразити як експоненціальне перетворення [14]:

$$exp\beta_k = \frac{h(t,x_1,x_2,\dots,x_k+1,\dots,x_p)}{h(t,x_1,x_2,\dots,x_k,\dots,x_p)} \qquad (2.11)$$

У разі двійкової змінної зі значеннями 0 і 1 формула (2.11) виражає, у скільки разів група ризику має вищий ризик виникнення, ніж референтна група (припускаючи, що обидві групи порівняні з урахуванням інших факторів)[15].

У цій моделі регресійні коефіцієнти вказують на вплив кожного предиктору на функцію ризику, і зі збільшенням значення предиктору на одиницю, якщо значення інших змінних незмінні, ризик настання події зростає в $\exp(e)$ разів [16].

Величини $\exp(\beta_i)$ називають коефіцієнтами ризику (HR). Коефіцієнт ризику вище 1 вказує на характеристику, яка позитивно пов’язна з ймовірністю події і, таким чином, негативно пов’язана з тривалістю виживання [13]. Підсумовуючи, якщо:

HR = 1: немає ефекту;

HR < 1: зменшення ризику;

HR > 1: збільшення ризику.

Таким чином, у регресії Кокса нас цікавитимуть три види показників: результат, період спостереження, предиктори. Якби нас цікавили лише результат і предиктори, ми могли б скористатися для аналізу методом логістичної регресії. Якби ми були зацікавлені в оцінці результату щодо часу, то можна було б провести аналіз методом Каплана-Мейєра або побудови таблиць дожиття. Лише за допомоги регресії Кокса ми можемо оцінити вплив безлічі предикторів на результат з урахуванням періоду спостереження [17].

Для застосування регресії Коксу необхідно дотримання цілого ряду критеріїв [18]: в основі регресії Коксу є три базові припущення щодо змінних, які будуть представлені першими у списку. Слід також враховувати умови, які є значущими для всіх методів аналізу виживання:

1. Усі предиктори незалежні. Якщо виявлено взаємний вплив незалежних змінних, то модель необхідно включити функцію взаємодії цих факторів;
2. Усі змінні лінійно впливають на логарифм функції ризику настання події;
3. Ризик настання події будь-яких двох об'єктів у будь-який інтервал часу пропорційний.
4. Момент початку та закінчення дослідження (виникнення результату або закінчення періоду спостереження) або інтервал спостереження в одиницях часу повинні бути точно визначені для кожного члена вибірки;
5. Визначення результату та момент його виникнення також мають бути чітко зафіксовані;

6. Цензуровані та нецензуровані спостереження не повинні відрізнятися за виживання один від одного;
7. Методи оцінки виживання та визначення результату однакові протягом усього дослідження;
8. Умови, що впливають на виживання, не змінюються в ході дослідження.

РОЗДІЛ 3 РЕГРЕСІЙНИЙ АНАЛІЗ

3.1 Опис даних

Проведемо регресійний аналіз за методом Кокса з метою встановлення впливу віку пацієнтів та основних факторів ракових пухлин легень пацієнтів на імовірність виникнення рецидиву, імовірність летального випадку та імовірність клінічного одужання.

Отримана база даних складається з 50 пацієнтів та 11 змінних (необхідних характеристик). Таблиця даних наведена в Додатку А. Для зручності розрахунків загальну таблицю було розділено на 3 частини, які потім були додані на певних етапах розрахунку. Опис даних подається так:

sex_patient: стать (чоловік = 1, жінка = 2)

age: вік у роках

smoking: період куріння пацієнта у роках

tumor_size: розмір пухлини

ECOG: оцінка ефективності (0 = повністю активний, здатний підтримувати всі показники до захворювання без обмежень, 1 = обмежений у фізично напруженій діяльності, але амбулаторний і здатний виконувати роботу легкого або сидячого характеру, наприклад, легку роботу по дому, роботу в офісі)

stage: стадія (1 – інвазивний рак, пухлина не проростає в сусідні органи, немає метастазів в довколишніх лімфовузлах, 2 – пухлина розміром до 5 см, з ураженням регіонарних лімфовузлів, або пухлина розміром 5-7 см, що розповсюджується на прилеглі органи й структури, 3 – пухлина будь-яких розмірів, яка проникає в сусідні органи (трахею, стравохід, серце), з метастазами в лімфовузли з обох боків шиї і грудей)

time_recurrence: час, який пройшов від дати операції до дати рецидиву в днях

status_recurrence: статус пацієнта, відносно появи рецидиву (1 = був рецидив, 2 = не було)

time_op_dead: час, який пройшов від дати операції до дати смерті в днях

dead: настання смерті (1 = помер, 2 = живий)

operation_date: дата операції

Реалізація регресії Кокса на мові Python наведена в Додатку Б.

3.2 Вплив характеристик на ризик виникнення рецидиву

Для того, щоб проаналізувати від яких характеристик залежить ризик виникнення рецидиву було обрано такі дані з головної таблиці: стать пацієнта, вік, куріння, розмір пухлини, оцінка ефективності ECOG, стадію, кількість днів від дати операції до дати рецидиву і статус пацієнта (див. рис.3.1).

	sex_patient	age	smoking	tumor_size	ECOG	stage	time_recurrence	status_recurrence
0	2	42	0	3	1	2	1378	1
1	2	57	0	5	1	3	0	2
2	1	38	23	2.5	0	1	0	2
3	1	53	20	5	1	2	119	1

Рис.3.1 – Дані для аналізу впливу характеристик на ризик виникнення рецидиву

Провівши обчислення з використанням мови Python, отримали результати, наведені на рис.3.2.

Результати регресії Кокса можна інтерпретувати так:

Стовпець, позначений «z» дорівнює відношенню кожного коефіцієнта регресії до його стандартної помилки (z=coef/se(coef)). Він показує чи β-коефіцієнт даної змінної статистично достовірно відрізняється від 0. З наведених вище результатів можна зробити висновок, що вік і куріння мають високо статистично значущі коефіцієнти.

Друга особливість, на яку слід звернути увагу в результатах моделі Кокса, — це знак коефіцієнтів регресії (coef). Позитивний знак означає, що ризик смерті вищий, а отже, і прогноз гірший для суб'єктів з вищими значеннями цієї змінної. У цих даних β-коефіцієнт для статі дорівнює -0,07 вказує на те, що жінки мають

нижчий ризик появи рецидиву, ніж чоловіки. Аналогічно, оскільки β-коефіцієнт для оцінки ефективності ECOG дорівнює -0,34, то це означає що пацієнти, які обмежені у фізично напруженій діяльності мають більший ризик виникнення рецидиву порівняно з безсимптомними пацієнтами.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
sex_patient	-0.07	0.94	0.50	-1.04	0.91	0.35	2.48	0.00	-0.13	0.89	0.16
age	0.03	1.03	0.02	-0.00	0.06	1.00	1.06	0.00	1.74	0.08	3.60
smoking	-0.01	0.99	0.01	-0.04	0.01	0.96	1.01	0.00	-1.36	0.17	2.52
tumor_size	0.00	1.00	0.06	-0.12	0.12	0.89	1.13	0.00	0.03	0.98	0.03
ECOG	-0.34	0.71	0.60	-1.52	0.83	0.22	2.30	0.00	-0.57	0.57	0.81
stage	-0.18	0.84	0.25	-0.67	0.31	0.51	1.36	0.00	-0.72	0.47	1.09

Concordance	0.73
Partial AIC	303.04
log-likelihood ratio test	5.92 on 6 df
-log2(p) of ll-ratio test	1.21

Рис. 3.2 – Результат регресії Кокса для оцінки впливу характеристик на ризик появи рецидиву

Використаємо нашу підібрану модель, щоб побачити, як змінюється виникнення рецидиву, коли ми змінюємо значення коваріати. Тут я використала метод ***plot_partial_effects_on_outcome()***, щоб побачити, яке виникнення рецидиву для вікової групи пацієнтів 20, 30, 40, 50 і 60 років порівняно з їх базовою функцією. Отримали такі результати:

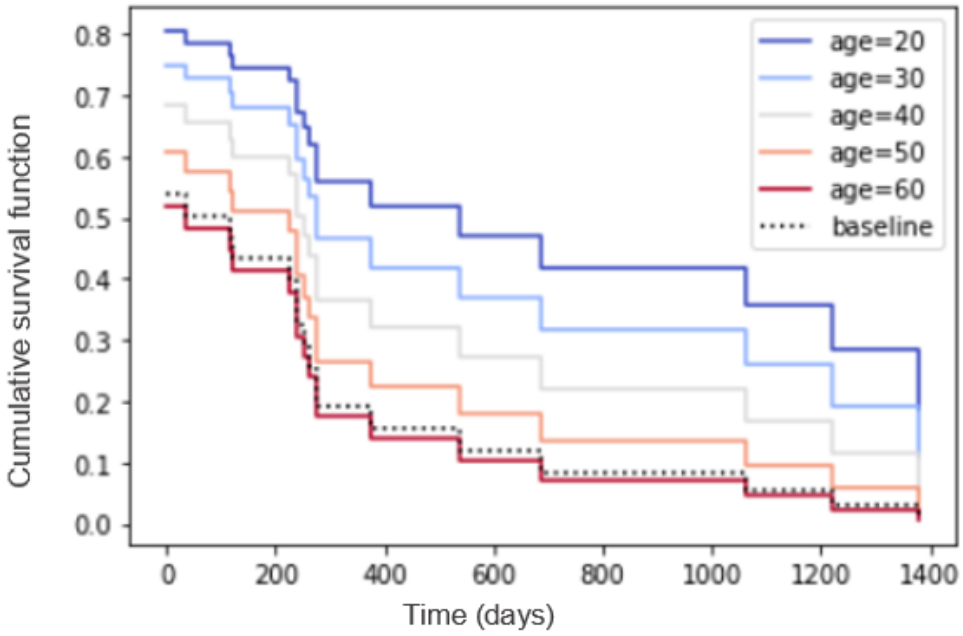

Рис.3.3 – Графік впливу віку пацієнта на результат

Очевидно, що пацієнти віком 50-60 років мають більшу ймовірність появи рецидиву.

Згідно з результатами рис.3.3 можемо затвердити, що на ризик появи рецидиву розмір пухлини взагалі не впливає, тому використаємо регресію Кокса для наших даних, але не враховуючи значення стовбця tumor_size. Отримавши дані, які показано на рис.3.4, можна зробити висновок, що коваріанта tumor_size є незначущою.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
sex_patient	-0.07	0.94	0.50	-1.04	0.90	0.35	2.47	0.00	-0.14	0.89	0.16
age	0.03	1.03	0.02	-0.00	0.06	1.00	1.06	0.00	1.74	0.08	3.60
smoking	-0.01	0.99	0.01	-0.04	0.01	0.96	1.01	0.00	-1.36	0.17	2.52
ECOG	-0.34	0.71	0.59	-1.50	0.83	0.22	2.28	0.00	-0.57	0.57	0.82
stage	-0.18	0.84	0.25	-0.66	0.31	0.52	1.36	0.00	-0.72	0.47	1.09

Concordance	0.72
Partial AIC	301.04
log-likelihood ratio test	5.92 on 5 df
-log2(p) of ll-ratio test	1.67

Рис.3.4 – Результат регресії Кокса для оцінки впливу характеристик на ризик появи рецидиву, не враховуючи розмір пухлини

3.3 Вплив характеристик на ризик смерті

Для того, щоб проаналізувати від яких характеристик залежить ризик смерті до таблиці, яка використовувалася: стать пацієнта, вік, куріння, розмір пухлини, оцінка ефективності ECOG стадію, кількість днів, які минули від дати операції до дати смерті і статус пацієнта, живий чи мертвий (див. рис.3.5).

	sex_patient	age	smoking	tumor_size	ECOG	stage	time_op_dead	dead	status_recurrence
0	2	42	0	3.0	1	2	0	2	1
1	2	57	0	5.0	1	3	0	2	2
2	1	38	23	2.5	0	1	0	2	2
3	1	53	20	5.0	1	2	0	2	1
4	1	65	50	1.0	0	1	1493	1	2

Рис.3.5 – Дані для аналізу впливу характеристик на ризик смерті

Після обчислення були отримані такі результати:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
sex_patient	0.03	1.03	0.47	-0.89	0.95	0.41	2.58	0.00	0.06	0.95	0.07
age	-0.01	0.99	0.02	-0.04	0.03	0.96	1.03	0.00	-0.36	0.72	0.48
smoking	-0.00	1.00	0.01	-0.03	0.02	0.97	1.02	0.00	-0.39	0.70	0.51
tumor_size	-0.06	0.94	0.06	-0.19	0.06	0.83	1.06	0.00	-1.03	0.30	1.72
ECOG	1.46	4.30	0.77	-0.04	2.96	0.96	19.28	0.00	1.91	0.06	4.14
stage	0.16	1.18	0.21	-0.25	0.58	0.78	1.79	0.00	0.77	0.44	1.19
status_recurrence	1.10	3.02	0.40	0.32	1.89	1.37	6.63	0.00	2.75	0.01	7.38

Concordance	0.91
Partial AIC	295.24
log-likelihood ratio test	15.71 on 7 df
-log2(p) of ll-ratio test	5.16

Рис. 3.6 – Результат регресії Кокса для оцінки впливу характеристик на ризик виникнення смерті

Згідно з результатами, які зображені на рис.3.6, значення коефіцієнта, пов’язаного з оцінкою ефективності ECOG, exp(1,46) = 4,30. Оскільки, значення коефіцієнта більше 1, то це вказує на те, що ризик смерті становить 4,30 рази для пацієнтів, які обмежені у фізично напруженій діяльності, порівняно з безсимптомними пацієнтами. Аналогічно, значення коефіцієнта, пов’язаного із наявністю рецидиву, exp(1,10) = 3,02. Це означає, що у пацієнтів, які мали рецидив ризик смерті значно більший ніж для тих, у кого його не було (див. рис.3.8).

На рис.3.7 зображена гістограма, яка показує кількість смертей пацієнтів в залежності від значення оцінки ефективності ECOG.

Рис.3.7 – Кількість смертей пацієнта в залежності від значення оцінки ефективності ECOG

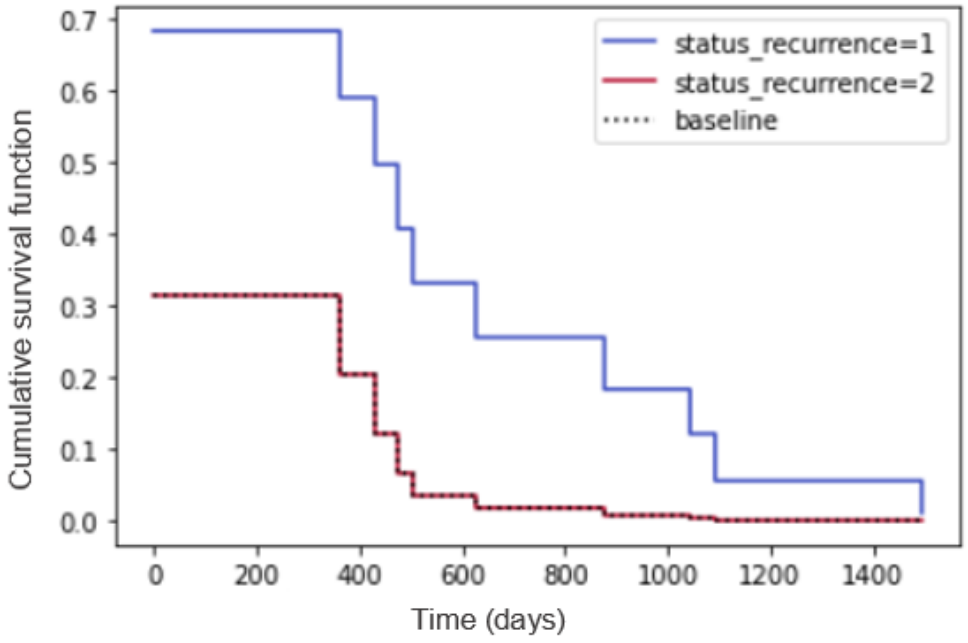

Рис.3.8 – Графік впливу наявності/відсутності рецидиву у пацієнта на результат

При аналізі характеристик раку легень на ризик виникнення рецидиву стадія раку була незначною характеристикою, проте, на ризик смерті вона впливає. Оскільки, exp(0,16) = 1,18 > 1, то пацієнти із вищою стадією мать більший ризик смерті.

3.4 Вплив характеристик на клінічне одужання пацієнта

Для того, щоб проаналізувати від яких характеристик залежить ймовірність одужання пацієнта були обрані такі дані: стать пацієнта, вік, куріння, розмір пухлини, оцінка ефективності ECOG, наявність чи відсутність рецидиву, дата операції, а також статус пацієнта (живий чи помер). В якості кінцевої дати дослідження було обрано 3.05.2022 рік.

Клінічним одужанням пацієнта вважається одужання, коли протягом 5 після операції у пацієнта не було рецидиву. У нашому випадку протягом 5 років з дати смерті та кінцевої дати дослідження.

На першому етапі було розраховано кількість днів між датою операції і кінцевою датою. Отримані дані було занесено до таблиці під стовбцем із назвою *day_after_operation*. Наступним кроком було видалення із таблиці даних про людей у яких відбувся рецидив і настала смерть. Для того, щоб дізнатися про статус пацієнта, одужав чи ні, я порівняла дані з стовбця *day_after_operation:* якщо кількість днів, які пройшли з моменту операції до кінцевої дати була більша ніж 1826 (ніж 5 років), то в створений стовбець під назвою *status* було занесено 1 - одужав, а якщо менша, то 2 – не одужав.

Для аналізу були використані дані з рис.3.9.

	sex_patient	age	smoking	tumor_size	ECOG	stage	day_after_operation	status
1	2	57	0	5.0	1	3	1489	1
2	1	38	23	2.5	0	1	1806	1
6	1	51	35	4.0	1	1	2001	2
7	1	62	25	5.5	1	2	2114	2

Рис.3.9 – Дані для аналізу впливу характеристик на повне клінічне одужання

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
sex_patient	1.73	5.62	0.64	0.47	2.98	1.61	19.66	0.00	2.70	0.01	7.17
age	0.04	1.04	0.03	-0.02	0.10	0.98	1.10	0.00	1.22	0.22	2.18
smoking	0.05	1.05	0.02	0.00	0.09	1.00	1.10	0.00	1.97	0.05	4.37
tumor_size	-0.18	0.83	0.10	-0.38	0.01	0.68	1.01	0.00	-1.86	0.06	3.99
ECOG	1.12	3.07	0.87	-0.59	2.84	0.55	17.04	0.00	1.28	0.20	2.32
stage	-0.39	0.68	0.28	-0.94	0.17	0.39	1.19	0.00	-1.36	0.17	2.52

Concordance	0.73
Partial AIC	166.42
log-likelihood ratio test	15.69 on 6 df
-log2(p) of ll-ratio test	6.01

Рис 3.10 – Результат регресії Кокса для оцінки впливу характеристик на повне клінічне одужання

Згідно з результатами, які зображені на рис.3.10, значення коефіцієнта, пов’язаного з оцінкою ефективності ECOG, exp(1,12) = 3,07. Оскільки, значення коефіцієнта більше 1, то це вказує на те, що ймовірність повного одужання безсимптомного пацієнта вища ніж пацієнта, який обмежений у фізично напруженій діяльності.

Оскільки β-коефіцієнт для стадії дорівнює -0,39, то це вказує на те, що пацієнти з 1 і 2 стадією мають більшу ймовірність на повне одужання, ніж пацієнти з 3 (див. рис.3.11). Аналогічно і з розміром пухлини, чим менший розмір пухлини, тим більша ймовірність одужання(див. рис.3.12).

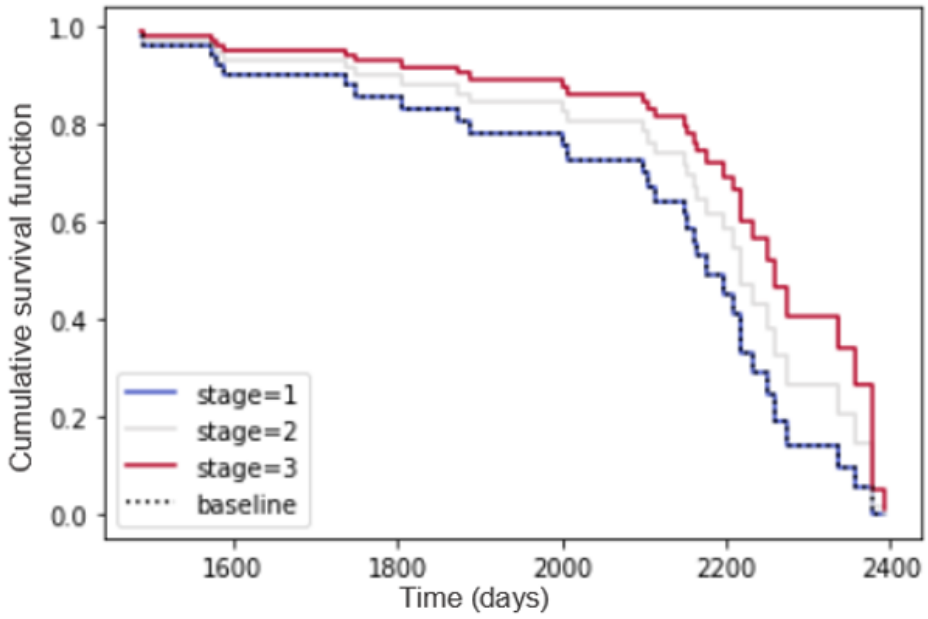

Рис.3.11 – Графік впливу стадії раку легень пацієнта на одужання пацієнта

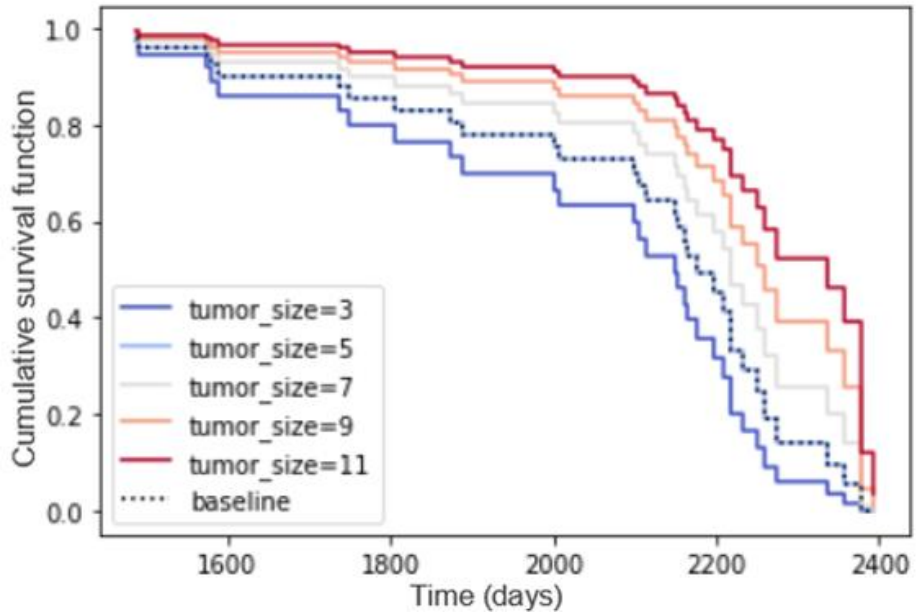

Рис.3.12 – Графік впливу розміру пухлини пацієнта на одужання пацієнта

ВИСНОВКИ

У роботі наведена основна інформація про аналіз виживання, а також про існуючі методи оцінки виживання.

Детально описано метод пропорційних ризиків Кокса, який використовувався для подальшого аналізу впливу характеристик раку легень на ризик появи рецидиву, смерті та одужання.

Провівши аналіз впливу коваріант на ризик появи рецидиву, смерті та одужання, можна зробити такі висновки: на появу рецидиву пацієнта найбільше впливає його вік, пацієнти віком 50-60 років мають найбільшу ймовірність рецидиву; пацієнти, які мають 1-2 стадію, а також пацієнти жіночої статі мають менший ризик появи рецидиву; смертність від раку найбільше залежить від оцінки ефективності ECOG, від наявності рецидиву, а також, пацієнти, у яких 3 стадія раку частіше помирають; молоді пацієнти мають менший ризик смерті, ніж пацієнти похилого віку. Ймовірність одужання повністю активного пацієнта більша ніж пацієнта, який обмежений у фізично напруженій діяльності. Також, важливу роль в одужанні відіграють стадія та розмір пухлини: пацієнти з 1 і 2 стадією частіше одужують, ніж пацієнти з 3 стадією; пацієнти з розміром пухлини менше 5 см мають більшу ймовірність одужання.

Враховуючи те, що кількість чолові в нашому досліді більша за кількість жінок в 4 рази (40 чоловіків, 10 жінок), можна сказати, що рак легень більше поширений у пацієнтів чоловічої статі. Тому, β-коефіцієнт змінної *sex_patient* протягом всього дослідження дуже високий, але не значущий.

Таким чином, модель пропорційних ризиків Кокса є зручним інструментом для медичних досліджень, тому що її використання дає можливість зробити відносно точний аналіз виживання в порівнянні з іншими методами у зв'язку з включенням набору незалежних змінних, що впливають на ризик настання події.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Blower S, Bernoulli D. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. 1766. Rev Med Virol. 2004 Sep-Oct; 14(5): 275-88.
2. Parmer M, Machin D (1995) Survival analysis. UK: John Wiley and Sons Ltd [Google Scholar]
3. Collet D. Modelling Survival Data in Medical Research. 2003, Chapman & Hall/CRC, London.
4. David G. Kleinbaum, Mitchel Klein. Survival Analysis Second Edition M.: Springer 2005. 590 с.
5. Klein JP, Moeschberge ML, Gail M, Samet JM, Tsiats A. Statistics for Biology and Health. New York: Springer; 2003.
6. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. Br J Cancer. 2003; 89:431–6.
7. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. 2nd ed. New York: Springer; 2010. [Google Scholar]
8. George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol. 2014; 21:686–94.
9. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. 2nd ed. New York: John Wiley and Sons; 2011. [Google Scholar]
10. Cox D.R. Regression models and life tables (with discussion) // J. R. Statist. Soc., Series B. 1972. N 2. P. 187- 220.
11. NCSS statistical software. Cox regression: [site]. URL: http://ncss.wpengine.netdna-cdn.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Cox_ Regression.pdf
12. Abraira V, Muriel A, Emparanza JI, Pijoan JI, Royuela A, Plana MN, et al. Reporting quality of survival analyses in medical journals still needs

improvement. A minimal requirements proposal. J Clin Epidemiol. 2013; 66:1340–6.

13. Peat J., Barton B. Medical statistics: a guide to data analysis and critical appraisal (1st ed.) // NY: Blackwell Publishing, 2005. 324 p.
14. Cox D.R. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological). 1958; 20 (2): 215–242.
15. Stel V.S., Dekker F.W., Tripepi G., Zoccali C., Jager K.J. Survival analysis II: Cox regression. Nephron CliPract. 2011; 119 (3): c255–260. DOI: 10.1159/000328916.
16. Prentice R.L., Kalbfleisch J.D., Peterson A.V.Jr., Flournoy N., Farewell V.T., Breslow N.E. The analysis of failure times in the presence of competing risks. Biometrics. 1978 Dec.; 34 (4): 541–554.
17. Cox Proportional-Hazards Model - Easy Guides - Wiki - STHDA. STHDA - Accueil. URL: http://www.sthda.com/english/wiki/cox-proportional-hazards-mode l (date of access: 7.05.2022).
18. Norusis M.J. SPSS 15.0 advanced statistical procedures companion. New Jersey, 2007. 418 p.

Додаток А

Таблиця А.1

Дані для аналізу

sex_patient	*age*	*smoking*	*tumor_size*	*ECOG*	*stage*	*time_recurrence*	*status_recurrence*	*time_op_dead*	*dead*	*operation_date*
2	42	0	3	1	2	1378	1	0	2	19.07.2017
2	57	0	5	1	3	0	2	0	2	05.04.2018
1	38	23	2.5	0	1	0	2	0	2	23.05.2017
1	53	20	5	1	2	119	1	0	2	12.09.2017
1	65	50	1	0	1	0	2	1493	1	04.03.2016
1	57	42	3.5	1	2	1219	1	0	2	01.11.2016
1	51	35	4	1	1	0	2	0	2	09.11.2016
1	62	25	5.5	1	2	0	2	0	2	19.07.2016
1	29	20	3.5	1	3	118	1	0	2	13.09.2017
1	58	40	3.5	1	2	252	1	505	1	15.05.2018
2	61	0	4.5	1	1	0	2	0	2	27.12.2017
2	65	0	5	1	1	0	2	0	2	04.01.2018
1	67	42	7	1	2	33	1	0	2	25.11.2016
1	65	50	3.5	1	3	0	2	0	2	01.08.2017
1	51	38	11	1	2	686	1	878	1	29.08.2017
1	42	22	2	1	1	539	1	1042	1	20.10.2016
2	71	0	5	1	2	0	2	0	2	04.04.2018
2	51	0	5.5	1	1	0	2	0	2	04.11.2016
1	58	25	5.9	1	1	0	2	0	2	24.02.2016
1	62	38	2	1	1	0	2	0	2	18.07.2017
1	56	30	2	0	1	0	2	0	2	09.12.2015
1	54	38	4.5	0	1	274	1	476	1	13.06.2017
1	77	20	4.2	1	1	223	1	430	1	30.12.2015
1	64	0	4	1	2	276	1	363	1	17.05.2016
1	60	42	9	1	3	238	1	627	1	16.05.2017
1	57	39	3.4	1	2	0	2	0	2	10.01.2018
1	58	40	7	1	1	0	2	0	2	17.03.2017
1	58	45	2.5	1	3	239	1	0	2	22.08.2017
1	54	36	10	1	2	1064	1	1093	1	28.11.2017
1	66	51	5	1	2	260	1	0	2	27.12.2017
1	67	50	1.1	1	2	373	1	0	2	10.10.2017
1	65	25	4	1	3	0	2	0	2	05.04.2016
1	53	20	2	1	1	0	2	0	2	14.06.2016
1	57	40	7	1	1	0	2	0	2	28.02.2017
1	75	18	3.2	1	3	0	2	0	2	31.05.2016
1	79	20	8.5	1	2	0	2	0	2	15.04.2016
2	44	15	6	1	2	0	2	0	2	03.03.2016
2	61	12	1.5	1	2	0	2	0	2	29.05.2016

Продовження таблиці А.1

П1	66	0	6	1	3	0	2	0	2	29.10.2015
1	57	0	6	1	2	0	2	0	2	30.10.2015
1	63	22	5.5	1	1	0	2	0	2	19.11.2015
2	54	0	5	1	1	0	2	0	2	26.04.2016
1	55	20	9.5	1	2	0	2	0	2	15.10.2015
1	40	12	3.5	1	1	0	2	0	2	10.06.2016
1	65	20	5	1	1	0	2	0	2	27.07.2016
1	63	20	4	1	1	0	2	0	2	06.04.2016
1	51	25	3.5	1	2	0	2	0	2	19.05.2016
2	56	0	6	1	1	0	2	0	2	04.08.2016
1	56	30	2	1	2	0	2	0	2	22.03.2016
1	64	25	7.5	1	3	0	2	0	2	09.02.2016

Додаток Б

Програмна реалізація регресії Кокса для встановлення впливу характеристик на ризик появи рецидиву

```
#Встановлення бібліотеки аналізу виживання
!pip install lifelines

#Імпортуємо бібліотеки
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from lifelines import CoxPHFitter

#Імпортуємо і переглядаємо датасет
data_rec = pd.read_csv("../input/lang-cancer-dyplom/recurrence.csv",
encoding="Windows-1251",  sep=';', dtype='unicode')
data_rec.head()

#Перевіряємо тип даних в датасеті
data_rec.dtypes

#Переводимо дані таблиці в цілий (int64) та дробовий (float64) типи
data_rec["sex_patient"] = data_rec["sex_patient"].astype("int64")
data_rec["age"] = data_rec["age"].astype("int64")
data_rec["smoking"] = data_rec["smoking"].astype("int64")
data_rec["tumor_size"] = data_rec["tumor_size"].astype("float64")
data_rec["ECOG"] = data_rec["ECOG"].astype("int64")
data_rec["stage"] = data_rec["stage"].astype("int64")
data_rec["time_recurrence"] = data_rec["time_recurrence"].astype("int64")
data_rec["status_recurrence"] = data_rec["status_recurrence"].astype("int64")
```

```
#Створюємо об’єкт класу CoxPHFitter. Викликаємо метод .fit() і надаємо дані, стовпець тривалості та стовпець події.
#Показуємо отриману зведену таблицю оцінки моделі
coxreg_rec = CoxPHFitter()
coxreg_rec.fit(data_rec, duration_col = 'time_recurrence', event_col = 'status_recurrence')
coxreg_rec.print_summary()

# Побудуємо рафік впливу віку пацієнта на результат
coxreg_rec.plot_partial_effects_on_outcome(covariates = 'age',values = [20, 30, 40, 50, 60], cmap = 'coolwarm')

#Приберемо з таблиці data_rec дані про розмір пухлини
data_rec_without = data_rec
data_rec_without = data_rec_without.drop(['tumor_size'], axis = 1)
data_rec_without.head()

#Перевіримо, як зміняться значення коефіцієнтів в регресії Кокса із зміною таблиці data_rec
coxreg_rec_without = CoxPHFitter()
coxreg_rec_without.fit(data_rec_without, duration_col = 'time_recurrence', event_col = 'status_recurrence')
coxreg_rec_without.print_summary()
```

Додаток В

Програмна реалізація регресії Кокса для встановлення впливу характеристик на ризик смерті

```
#Імпортуємо і переглядаємо датасет
data_dead = pd.read_csv("../input/dyplom-dead/dead.csv", encoding="Windows-1251", sep=';', dtype='unicode')
data_dead.head()

#Переводимо дані таблиці в цілий (int64) та дробовий (float64)
data_dead["sex_patient"] = data_dead["sex_patient"].astype("int64")
data_dead["age"] = data_dead["age"].astype("int64")
data_dead["smoking"] = data_dead["smoking"].astype("int64")
data_dead["tumor_size"] = data_dead["tumor_size"].astype("float64")
data_dead["ECOG"] = data_dead["ECOG"].astype("int64")
data_dead["stage"] = data_dead["stage"].astype("int64")
data_dead["time_op_dead"] = data_dead["time_op_dead"].astype("int64")
data_dead["dead"] = data_dead["dead"].astype("int64")

#Додаємо до нашого датасету колонку даних status_recurrence з датасету data_rec, який використовувся для оцінки ризику рецидиву
recurrence = data_rec['status_recurrence']
data_dead_with_status_rec = data_dead
data_dead_with_status_rec['status_recurrence'] = recurrence
data_dead_with_status_rec.head()

# Створюємо об’єкт класу CoxPHFitter. Викликаємо метод .fit() і надаємо дані, стовпець тривалості та стовпець події.
coxreg_dead = CoxPHFitter()
coxreg_dead.fit(data_dead, duration_col = 'time_op_dead', event_col = 'dead')
```

```
coxreg_dead.print_summary()

#Побудуємо гістограму, щоб показати кількість смертей пацієнтів в залежності від значення ECOG
ecog1 = data_dead.loc[data_dead['ECOG'] == 1]
ecog_1 = ecog1.loc[ecog1['dead'] == 1]
ecog0 = data_dead.loc[data_dead['ECOG'] == 0]
ecog_0 = ecog0.loc[ecog0['dead'] == 1]
ecog1_dead = len(ecog_1.index)
ecog0_dead = len(ecog_0.index)
ECOG_dead = {'ECOG = 1': ecog1_dead,'ECOG =0': ecog0_dead }

plt.bar(ECOG_dead.keys(), ECOG_dead.values(), width = 0.1, color='g')
plt.title('Кількість смертей пацієнтів в залежності від значення ECOG')
plt.ylabel('Кільість пацієнтів')
plt.show()

#Розглянемо як на смертність пацієнтів впливає наявність чи відсутність рецидиву, яке в нашому випадку становить 1 і 2
coxreg_dead.plot_partial_effects_on_outcome(covariates = 'status_recurrence',values = [1,2], cmap = 'coolwarm')
```

Додаток Г

Програмна реалізація регресії Кокса для встановлення впливу характеристик на ймовірність повного одужання пацієнта

```
#Встановимо бібліотеку для читання файлу Excel
!pip install openpyxl

#Імпортуємо і переглядаємо дані про дату операції
data_op = pd.read_excel('../input/operation-date/operation_date.xlsx')
data_op.head()

#Кінцеву дату записуємо типом  datetime64
new_date = pd.to_datetime({'year':[2022], 'month':[5], 'day':[3]})
new_date

#Знаходимо різницю між кінцевою датою та датою операції
n = pd.to_datetime(data_op['operation_date'])
nm = [new_date - i for i in n]
m = [str(i) for i in nm ]
m = [i[4:8] for i in m ]

m = [int(i) for i in m ]
data_op['day_after_operation'] = m #Додаємо отримані дані в таблицю data_op
data_op.head()

#Додаємо до нашого датасету колонку даних date_after_operation з датасету
data_op
operation = data_op['day_after_operation']
data_operation = data_dead
data_operation['day_after_operation'] = operation
```

```
data_operation.head()

#Для того щоб оцінити як коваріанти впливають на одужання пацієнта, необхідно видалити з таблииці пацієнтів, які вже померли nf vfkb htwblbd
data_operation_wihout_dead = data_operation.loc[data_operation['dead'] == 2]
dt = data_operation_wihout_dead.drop(['dead','time_op_dead'], axis = 1)
dt1 = dt.loc[dt['status_recurrence'] == 2]
data_living = dt1.drop(['status_recurrence'], axis = 1)
data_living.head()

#Визначаємо статус пацієнта (2 – одужав, 1 – не одужав)
data_living['status'] = [2 if i >= 1826 else 1 for i in data_living['day_after_operation']]
data_living.head()

# Створюємо об’єкт класу CoxPHFitter. Викликаємо метод .fit() і надаємо дані, стовпець тривалості та стовпець події
coxreg_living = CoxPHFitter()
coxreg_living.fit(data_living, duration_col = 'day_after_operation', event_col = 'status')
coxreg_living.print_summary()

#Побудуємо рафік впливу стадії пацієнта на результат
coxreg_living.plot_partial_effects_on_outcome(covariates = 'stage',values = [1, 2, 3], cmap = 'coolwarm')

#Побудуємо рафік впливу розміру пухлини пацієнта на результат
coxreg_living.plot_partial_effects_on_outcome(covariates = 'tumor_size',values = [3, 5, 7, 9, 11], cmap = 'coolwarm')
```