UNIVERSITY OF KALYANI

DEPARTMENT OF STATISTICS

Performance of Distributed Cox Proportional Hazards Modelling for Summary-Level Cancer Data

SUBMITTED BY

DEBJIT SAHA

ROLL: 96/STA NO.: 230007

REGISTRATION NO.: 2320010 of 2023-2024

SUPERVISOR:

DR. SUMANTA ADHYA

ASSISTANT PROFESSOR, DEPARTMENT OF STATISTICS

WEST BENGAL STATE UNIVERSITY

YEAR: 2025

Performance of Distributed Cox Proportional Hazards Modelling for Summary-Level Cancer Data

Dissertation submitted to

Department of Statistics

University of Kalyani, West Bengal

In partial fulfilment of the requirements for the degree of

Master of Science (Two Years) in Statistics

By

Debjit Saha

Roll: 96/STA No.: 230007

Registration No.: 2320010 of 2023-2024

Under the Supervision

Of

Dr. Sumanta Adhya

Assistant Professor, Department of Statistics

West Bengal State University

Department of Statistics

University of Kalyani, Kalyani, Nadia

PIN: 741235, West Bengal

DECLARATION

I hereby declare that the project work entitled **"Performance of Distributed Cox Proportional Hazards Modelling for Summary-Level Cancer Data (2015-2024)"** submitted to Department of Statistics, University of Kalyani, Kalyani, Nadia in partial fulfilment of the requirement for the award of the degree of Master of Science (Two Years) in Statistics, is a record of original project work done by me during the period of my study under the guidance and supervision of **Dr. Sumanta Adhya, Assistant Professor, Department of Statistics, West Bengal State University**, during 2025.

Place: Kalyani
Date:

Debjit Saha
Roll: 96/STA No.: 230007
Department of Statistics
University of Kalyani

TABLE OF CONTENTS

Performance of Distributed Cox Proportional Hazards Modeling For Summary-Level Cancer Data

ABSTRACT

This study investigates the performance of distributed Cox Proportional Hazards Modelling (CPHM) using summary-level cancer data spanning 2015–2024. The dataset incorporates key risk factors such as genetic predisposition, smoking, and other relevant covariates. The work addresses a critical challenge in medical research—balancing the need for robust statistical modelling with restrictions on sharing individual-level patient data. We demonstrate that reliance solely on descriptive statistics can be misleading and insufficient for uncovering the true significance of associations in complex medical datasets. To overcome data privacy barriers, we apply a recent distributed modelling technique capable of fitting a CPHM using only aggregated summary statistics, without requiring access to primary individual records. The analysis evaluates model performance, quantifies information loss in comparison to full-data approaches, and highlights the potential of distributed survival models as a privacy-preserving yet statistically rigorous alternative for large-scale epidemiological studies.

Dataset

Patient.ID	Age	Gender	Country_Region	Year	Genetic.Risk	Air.Pollution	Alcohol.Use	Smoking.Score
PT0000000	71	Male	UK	2021	6.4	2.8	9.5	0.9
PT0000001	34	Male	China	2021	1.3	4.5	3.7	3.5
PT0000002	80	Male	Pakistan	2023	7.4	7.9	2.4	4.7
PT0000003	40	Male	UK	2015	1.7	2.9	4.8	3.5
PT0000004	43	Female	Brazil	2017	5.1	2.8	2.3	3.3
PT0000005	22	Male	Germany	2018	9.5	6.4	3.3	5.7
PT0000006	41	Male	Canada	2021	5.1	6.0	8.2	3.4
PT0000007	72	Female	Canada	2018	6.8	3.3	6.4	3.0
PT0000008	21	Male	USA	2022	4.3	3.8	8.7	5.2
PT0000009	49	Female	Canada	2016	8.1	0.8	7.8	2.7
PT0000010	21	Female	Brazil	2021	5.2	1.7	7.2	5.0
PT0000011	83	Male	Canada	2016	3.5	1.5	8.1	3.1
PT0000012	79	Female	USA	2021	8.5	9.6	3.6	6.1
PT0000013	40	Male	UK	2023	4.6	3.3	5.5	3.8
PT0000014	52	Male	Germany	2024	2.3	5.1	9.1	3.9
PT0000015	41	Male	Germany	2016	6.4	9.2	7.2	4.8
PT0000016	68	Male	UK	2023	8.4	7.4	7.8	3.7
PT0000017	78	Male	India	2023	8.3	6.8	3.5	7.9
PT0000018	61	Female	India	2023	9.4	4.6	1.9	5.9
PT0000019	19	Male	Germany	2020	6.1	7.5	6.5	3.4
PT0000020	24	Male	China	2024	9.6	6.3	0.4	3.3
PT0000021	32	Female	India	2019	9.7	1.9	4.5	4.0
PT0000022	81	Male	China	2018	6.4	9.8	2.6	4.5
PT0000023	83	Female	UK	2011	8.1	3.1	5.3	5.6
PT0000024	70	Male	Pakistan	2022	6.6	1.9	2.9	5.3
PT0000025	26	Male	USA	2022	5.2	2.5	1.8	5.4

Dataset Overview: This dataset, sourced from Kaggle, represents global cancer patient data between 2015 and 2024. It includes simulated information on age, gender, cancer type, genetic risk, environmental exposure (air pollution), and lifestyle behaviors (smoking, alcohol use, obesity). These features provide a comprehensive basis for analyzing factors influencing cancer diagnosis and survival.This data contains 33550 data points.Here we have shown some of the data points.

The total dataset: Kaggle: Global Cancer Patients (2015–2024).

Data Description

Patient Details :-

1. Patient ID

A unique identifier assigned to each patient. 50,000 unique patients are included in the dataset.

2. Age

Age of the patient at the time of diagnosis or study entry. Values range from 20 to 89 years.

3. Gender

Gender of the patient. Typically recorded as ”Male” or ”Female”.

4. Country/Region

The country or region where the patient resides. There are 10 different countries represented, including examples like the UK, China, and Brazil.

5. Year

The year in which the data was captured or last updated. Covers the period from 2015 to 2024.

Risk Factors :-

6.Genetic_Risk

`Genetic Risk` refers to a numerical score or level that estimates a patient's predisposition to developing cancer based on their inherited genetic factors.

In This Dataset:

- The **Genetic_Risk** value is a quantitative score from 0 to 10.

$$= \begin{cases} 0 & \text{if no genetic risk} \\ 10 & \text{if very high genetic risk} \end{cases}$$

- A score of 6.4 suggests the patient has a moderately high inherited susceptibility to cancer.

7. Air_Pollution

The `Air Pollution` column represents a numerical index of a patient's exposure level to air pollution.

Scale: 0 to 10

$$= \begin{cases} 0 & \text{if no or negligible exposure} \\ 10 & \text{if extremely high exposure} \end{cases}$$

- It likely reflects long-term exposure to pollutants such as $PM_{2.5}$, NO_2, ozone, etc., possibly based on the patient's region.

8. Alcohol:

`Alcohol` **c**olumn represents level or frequency of alcohol consumption.

- **Scale:** 0 to 10

$$= \begin{cases} 0 & \text{no consumption of alcohol} \\ 10 & \text{consume alcohol in extreme level} \end{cases}$$

  High values may correlate with greater cancer risk and poorer survival outcomes.

Chronic alcohol consumption is linked to multiple types of cancer, including:

- Liver ,Esophagus ,Breast ,Colorectal ,Mouth and throat

Alcohol is a GROUP 1 Carcinogen as classified by the International Agency for Research on Cancer(IARC).

9. Smoking:

`Alcohol` **c**olumn represents level or frequency of alcohol consumption.

- **Scale:** 0 to 10

$$= \begin{cases} 0 & \text{non smoker} \\ 10 & \text{heavier smoking} \end{cases}$$

  High values may correlate with greater cancer risk and poorer survival outcomes.

10. Obesity:

`Obesity Level` represents body fat level or a proxy for **BMI (Body Mass Index)**.

- Scale :- Typically between 0 and 10 (observed 0.1 to 8.7 in your data sample).
- Higher values suggest greater levels of obesity, which is a known risk factor for several cancers.

Cancer Details :-

11. Cancer_Type:

The dataset includes a categorical variable indicating the type of cancer diagnosed in each patient.It helps differentiate between various cancer subgroups for further clinical analysis and risk stratification.

This dataset contains following cancer classifications:

,Lung ,Leukemia ,Breast ,Colon ,Skin ,Cervical ,Prostate ,Liver

12. Cancer_Stage:

The dataset includes a categorical variable indicating the stages of cancer diagnosed in each patient.

This dataset contains following cancer stages:

- Stage 0: Pre-cancer or in situ (not yet invasive)
- Stage I: Small, localized tumor
- Stage II: Larger and/or spread to nearby tissue
- Stage III: Advanced local spread, possible lymph node involvement
- Stage IV: Distant metastasis (most severe)

Survival and Outcome

13. Survival_Years:

This column indicates the number of years a patient has survived after their cancer diagnosis or treatment started.

- **A higher value means longer survival.**

14. Target_Severity_Score:

This column indicates composite score that represents the estimated severity of the patient's cancer condition.It calculated from risk factors like **stage**, **lifestyle**, and **genetic predisposition**.

- **Higher scores indicate more severe disease, possibly lower expected survival.**

Survival Analysis

The **Survival Function**, denoted by $S(t)$, gives the probability that an individual or item survives beyond time t. Mathematically, it is defined as:

Survival Function:

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

where:

- T is a continuous random variable denoting the time to the event of interest (e.g., failure, death, churn).
- $F(t)$ is the cumulative distribution function (CDF): the probability that the event has occurred by time t.

This function is a **non-increasing** function of t, and it starts from 1 at $t = 0$ and decreases toward 0 as $t \to \infty$.

Interpretation:

- The larger the value of $S(t)$, the less likely the event (e.g., death, churn) has occurred by time t.
- For example, in customer retention studies, $S(t)$ denotes the probability a customer is still subscribed at time t.

Relation to the pdf:

The survival function is related to the probability density function $f(t)$ as follows:

$$S(t) = \int_t^\infty f(u)\, du$$

Differentiating both sides:

$$\frac{dS(t)}{dt} = -f(t) \quad \Rightarrow \quad f(t) = -\frac{dS(t)}{dt}$$

Here, $f(t)$ represents the **event density** at time t.

Real Life Example:

1. How long a machine will work before it breaks.
2. How long a patient will live after receiving treatment.
3. How long a customer will remain subscribed to a service.

Example Interpretation :

If we are interested in whether a person survives more than 5 years after cancer therapy, we evaluate:

$$P(T > 5) = S(5)$$

The survival function is also known as the **Survivor Function** or **Reliability Function** in engineering contexts.

Hazard Function:

The hazard function denoted by $h(t)$ is given by the formula: $h(t)$ equals the limit, as $\Delta t \to 0$, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time. This mathematical formula is difficult to explain in practical terms.

The hazard function $h(t)$ gives the **instantaneous potential per unit time** for the event to occur, **given that the individual has survived up to time** t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}$$

$P(t \leq T < t+\Delta t \mid T \geq t) = P(\text{individual fails in the interval } [t, t+\Delta t] \mid \text{survived up to time } t)$

In the hazard formula, the conditional probability gives the probability that a person's survival time T will lie in the time interval between t and $t + \Delta t$, given that the survival time is greater than or equal to t.

Because of the given sign here, the hazard function is sometimes called a **conditional failure rate**.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t \cdot P(T \geq t)}$$

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)}$$

The final equivalent expression for the hazard function is:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \cdot \frac{dS(t)}{dt} = -\frac{d}{dt} \left(\ln S(t) \right)$$

Hazard Function Alternate Names :

The hazard function is also known as:

- **Force of mortality**
- **Conditional mortality rate**
- **Age-specific failure rate**

Hazard Ratio:

A **Hazard Ratio (HR)** is a measure used in survival analysis to compare the rate (*hazard*) of an event happening in one group versus another over time.
The event is usually something like death, disease recurrence, or failure of a treatment.
It tells us how much more (or less) likely the event is to happen at any time point in one group compared to another.

In other words, the hazard ratio (HR) compares the hazard rates between two groups (e.g., treatment vs control in a clinical trial). It calculates *relative risk* for two different groups.

$$HR = \frac{h_1(t)}{h_0(t)}$$

(i) $HR = 1$: $\Rightarrow$ No difference between groups.

(ii) $HR > 1$: $\Rightarrow$ Higher hazard (event happens faster)
Higher risk in group 1 compared to group 0.

(iii) $HR < 1$: $\Rightarrow$ Lower hazard (event happens slower)
Lower risk in group 1 compared to group 0.

Censoring:

Most survival analysis must consider a key analytical problem called **censoring**. In essence, censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly.

In simple words, the event has occurred or not occurred up to a certain point, but we don't know the exact time it happened (or will happen).

Three reasons why censoring may occur:

(i) A person does not experience the event before the study ends.

(ii) A person is lost to follow-up during the study period.

(iii) A person withdraws from the study because of death
(if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk).

Types of Censoring:

a. Right Censoring:

- True survival time is equal to or greater than observed survival time.
- In simple words, I know the event hasn't happened yet, but I stop observing before it does.

Example: I am studying how long a light bulb lasts. One bulb is still running after 1000 hours of testing.

- I don't know when it will fail, only that it survived at least 1000 hours.
- That is right censored at 1000 hours.

b. Left Censoring:

- True survival time is less than or equal to the observed survival time.
- In simple words, you know the event has already occurred but you don't know when it actually started.

Example: You test people's blood pressure and find someone already has high blood pressure at the first test.

- You don't know when it started — just that it started before your observation.
- That's left-censored.

c. Interval Censoring:

- True survival time is within a known time interval.
- In simple words, you know the event occurred within a time interval, but not exactly when.

Example: You check a machine once every 10 days. On day 20 it's fine, and on day 30 it's broken.

- The failure happened between day 20 and day 30.
- That's interval censored.

Cox Proportional Hazard Model (CPHM):

Hazard:

In survival analysis, the hazard is the instantaneous risk of an event (like death, failure or relapse) occurring at a particular time, given that the subject has survived up to that time.

Proportional Hazard:

The Proportional Hazards assumption says that the hazard rates for different individuals are proportional over time i.e, the ratio of their risks is constant, even though their risks may change over time.

In simple words, Proportional Hazards means that the risk of something happening (like getting sick, dying or a machine breaking) is always in the same ratio between two people or things no matter how much time has passed.

Example:

Imagine two people:

- Person A has a baseline hazard.
- Person B has twice the risk of Person A.

Even as time goes on, Person B's risk remains exactly twice that of Person A – it doesn't get closer or further apart.

This is what we mean by *"proportional"* – the hazard ratios stay the same over time.

Kaplan–Meier Estimation:

Kaplan–Meier estimation is a nonparametric method of estimating the survival function. Nonparametric methods are rather simple methods which do not make any distributional assumptions in this context about the distribution of survival times observed in a study.

Let $t_1 < t_2 < \cdots < t_k$ be the observed event times and $n = n_0$ the sample size. Let d_j be the number of individuals who have an event at time t_j, m_j the number of individuals censored in the interval $[t_j, t_{j+1})$. Then the number at risk at time t_j, denoted by n_j, is given by:
where:

- n_j: $(\mathrm{m}_j + d_j)$+.....$(m_k + d_k)$
- d_j: number of dying $[t_j, t_{j+1})$ The **Kaplan–Meier** (or product-limit) estimator is a nonparametric estimator of the survival function:

$$\hat{S}(t) = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

Standard errors can be calculated using **Greenwood's formula**, which approximates the variance as:

$$\widehat{\mathrm{Var}}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j:t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

The Kaplan–Meier curves for the two groups suggest that at any time point, a smaller proportion of people with more than four lymph nodes survive beyond that point that at any time point, a smaller proportion of a people with more than four lymph nodees survive beyond that point,

Compared to those with up to four positive lymph nodes.

To estimate and plot the cumulative hazard function, the **Nelson–Aalen estimator** can be used. The Nelson–Aalen estimator is a non-parametric estimator of the cumulative hazard function,

$$H(t) = \sum_{j:t_j \le t} \frac{d_j}{n_j} = \sum_{j:t_j \le t} h_j$$

- d_j is the number of individuals who have an event at time t_j, where $j = 1, 2, \ldots, k$, and
- n_j is the number of individuals at risk just prior to t_j.

A very similar alternative is to calculate the **Kaplan–Meier estimate** of the survival function and take: $-\log \hat{S}(t)$ as an estimate of the cumulative hazard, derived by the relationship between the survival and cumulative hazard function.

Log-Rank test

The **Log-Rank Test** is a large sample **Chi-Square test(Non- paramteric)** using a test criterion a statistic that provides an overall comparison of the KM curves being compared.

Hypothesis:

- The null hypothesis is that there is no difference in survival between the groups.
  $H_0 : S_1(t) = S_2(t) \quad \forall t$
- While the alternative hypothesis is that there is a difference.
  $H_A : S_1(t) \neq S_2(t)$ for some t

Test Statistic:

The test statistic of log rank test is given by

$$\chi^2 = \frac{\left(\sum_i (O_i - E_i)\right)^2}{\sum_i \mathrm{Var}(O_i - E_i)},$$

where O_i is the observed number of events in group i,
E_i is the expected number of events under the null hypothesis,
$Var(O_i - E_i)$ is the variance.

Decision Rule:

- Compute χ^2 using the formula.
- Compare with critical value from $\chi^2_{k-1,\alpha}$.
  - If $\chi^2 > \chi^2_{k-1,\alpha}$, reject H_0.
  - Equivalently, compute the **p-value**:If $p < \alpha$ (say 0.05), reject H_0.

Assumptions for the Log Rank Test

The assumptions for the log-rank test are as follows:

- **Independence:** The survival times or event times of individuals in each group should be independent to each other.
- **Non-Informative Censoring:** Censoring should not be related to the event being studied or to the group assignment (Censored and non-censored patients do not differ in terms of their actual event times). The log-rank test assumes that the probability of censoring should be the same for all individuals within each group. In other words, censoring should not be related to the event being studied or to the group assignment.

– **Proportional Hazards:**The hazard rates (the risk of an event occurring) for the compared groups should be consistent over time. The ratio of the hazard rates should remain constant, indicating that the groups are not experiencing significantly different risks at different time points.

CPHM:

The Cox Proportional Hazards Model (CPHM) has the form:

$$h(t, X) = h_0(t, \alpha)\, e^{\beta^T X}$$

where:

- $\boldsymbol{X}$ denotes a vector of covariates.
- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the parameters of the Cox's Regression Model.
- $h_0(t, \alpha)$ is called the *Baseline Hazard Function* at time point t.

Realizations:

(i) $h_0(t, \alpha)$ is the value of $h(t, X)$ at $X = 0$ (i.e., a baseline choice of X). That's why $h_0(t, \alpha)$ is called a *Baseline Hazard Function* (BHF).

(ii) $h_0(t, \alpha)$ depends on time t, but not on covariates X.

(iii) The term $e^{\beta^T x}$ depends on covariates x, but not on time point t.

(iv) β measures the effect of covariate x on the hazard function $h(t, x)$. How? For a univariate covariate x, denoted by X, $\frac{h(t,x+1)}{h(t,x)} = e^{\beta} \Rightarrow \beta$ gauges the impact of the change in $h(t, X)$ for the unit change in the value of x.

(v) Consider two individuals with covariates x_1 and x_2. The ratio of their hazards at time t is given by,

$$\frac{h(t, x_1)}{h(t, x_2)} = \frac{h_0(t, \alpha) \cdot e^{\beta^T x_1}}{h_0(t, \alpha) \cdot e^{\beta^T x_2}} = e^{\beta^T (x_1 - x_2)}$$

which is **constant with respect to time** t.

$$\Rightarrow \quad h(t, x_1) \propto h(t, x_2) \quad \text{for all } t$$

In other words,The hazards are proportional $\forall\, t$.
Hence, the name **Proportional Hazards Model (PHM)**. *i.e* the hazard ratio between two individuals (e.g., one aged 8 years and another aged 80 years) remains the same over time.
This assumption may seem unrealistic in certain real-life situations.

(vi) In Cox's Proportional Hazard, one will be interested in parameters β, which measure the effects of covariates on survival, but usually we are not interested in α
$\Rightarrow \beta \rightarrow$ Pararmters of interests, $\alpha \rightarrow$ Nuisance Parameters (Parts of a statistical model that we don't really care about, but we have to include them to make the model work propoerly)

(vii) Hence no form is pre-specified for the basline hazard, $h_0(t, \alpha)$, the Cox's Proportional Hazard Model is thus called a **Semi − Parametric Model**.

Estimating the Covariate Parameters β: Partial Likelihood:

As here, the form of $h_0(t, \alpha)$ is not specified, so the form of the likelihood function is not known properly. $\Rightarrow$ The usual method of maximum likelihood fails.

For estimating the regression parameters β, Cox developed a non-parametric method and called it **Partial Likelihood**.

Let m be the number of individuals under study.

$$\delta_i = \begin{cases} 1 & \text{if the individual is uncensored} \\ 0 & \text{if the individual is right censored} \end{cases}$$

Define the risk set at time t_i: $R(t_i) = \{j \mid T_j \geq t_i\}$

Let $h_j(t)$ denote the hazard of the j^{th} individual at time t.

Then, given that t_i is an event time (i.e., failure or death time), the probability that individual i has that event is given by:

$$P([i] \mid t_i) = \frac{\text{Likelihood that } i \text{ has the event at } t_i}{\text{Total likelihood that anyone in } R(t_i) \text{ has the event at } t_i}$$

The probability that individual i has the event at time t_i, among all individuals at risk, is proportional to their hazard at that time:

$$\text{Likelihood that } i \text{ has the event at } t_i \propto h_i(t_i)$$

To turn this into a valid probability, we divide by the sum of all hazard rates for individuals at risk at that time:

$$P([i] \mid t_i) = \frac{h_i(t_i)}{\sum\limits_{j \in R(t_i)} h_j(t_i)} \quad \text{,where } [i] \text{ denotes individual } i.$$

As per **Cox's Proportional Hazards Model (CPHM)** assumptions,

$$h_j(t, x_j) = h_0(t, \alpha) e^{\beta^T x_j}$$

So,

$$P([i] \mid t_i) = \frac{h_0(t_i, \alpha) e^{\beta^T x_i}}{\sum\limits_{j \in R(t_i)} h_0(t_i, \alpha) e^{\beta^T x_j}} = \frac{e^{\beta^T x_i}}{\sum\limits_{j \in R(t_i)} e^{\beta^T x_j}}$$

Let the **risk score** for each individual be:

$$\phi_i = e^{\beta^T x_i}, \quad \phi_j = e^{\beta^T x_j}$$

Then the probability becomes:

$$P([i] \mid t_i) = \frac{\phi_i}{\sum\limits_{j \in R(t_i)} \phi_j}$$

$P([i] \mid t_i)$ may be called the **Risk Probability** of individual i at time point t_i.

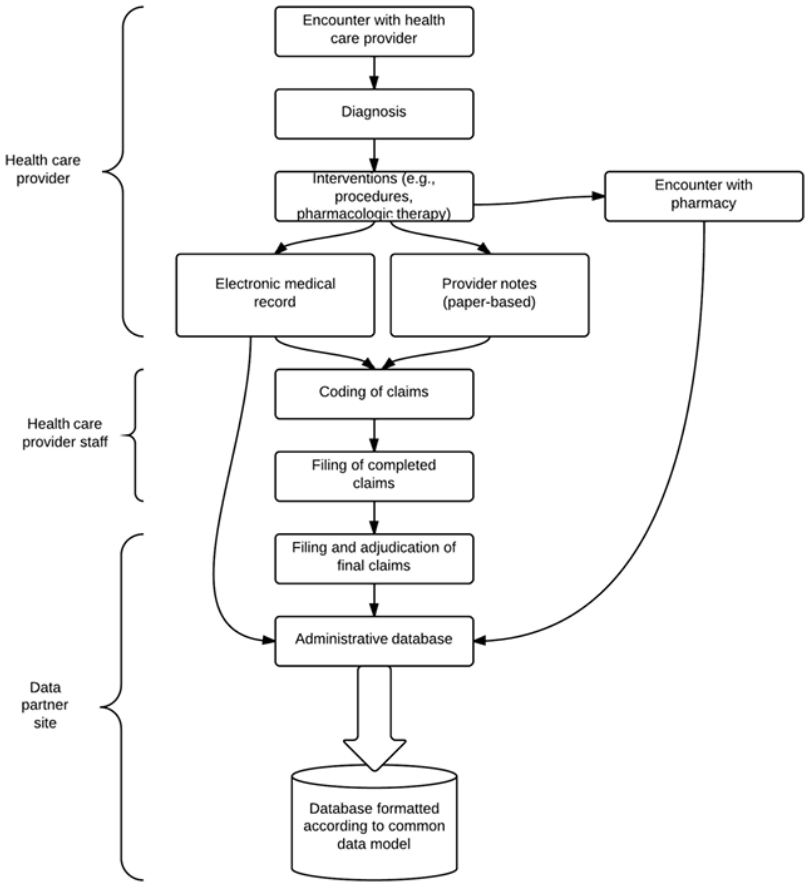
Distributed Data Network :

A distributed data network is a system where data is not stored in one central place. Instead, the data stays with the people or organizations (called *data holders*) that collected it.

Instead of moving the data around, the same piece of code is sent to each data holder. That code is run locally on their computers, and only the results are shared back – not the raw data.

Example

- Imagine 5 hospitals each have patient data.
- Instead of sending all the data to one place to study a new drug, each hospital runs the same analysis code on their own data.
- Then they send back only the results (like how many patients improved), not the patient records themselves.



Distributed Cox Proportional Hazards Regression:

This talks about a new method to analyze medical data from different hospitals or research sites without needing to share individual patient details, which helps protect privacy & reduces technical issues.

Introduction

What's the Problem?

1) We now have access to huge amounts of data from many hospitals & clinics (called records or EHRs).
2) To make stronger and more reliable results, researchers want to combine data from multiple places.
3) But sharing individual patient data is hard due to privacy, legal, and technical issues.

What's the Current Workaround?

1) Instead of sharing raw patient data, share summary-level statistics (like means, totals).
2) These summaries can still be used for analysis without risking patient privacy.
3) This idea is used in DRNs (Distributed Research Networks), where each site keeps its data but takes part in the research together.

What Kind of Data Analysis is Being Done?

1) The focus here is on **CPHM (Cox Proportional Hazards Model)** for time-to-event data (i.e., how long for a patient to be readmitted, die, etc.).
2) But unlike mean or one-shot data, CPHM often needs multiple rounds of calculations (iterations), which slows it down.

Key Terms

(a) X_i: Baseline covariate vector which may contain the exposure/treatment variable. (Information (variables) about person i – like age, gender, treatment etc.)

(b) T_i: The time we actually observe – either the event time or the time they were lost to follow-up (censored).

(c) T_i^*: Actual (true) event time.

(d) C_i: The censoring time,the point after which we stop observing the person (e.g., if they leave the study).

(e) $T_i = \min(T_i^*, C_i)$: We only know the earlier of the event or censoring.

(f) $\delta_i = \mathbb{I}[T_i^* \leq C_i] = \begin{cases} 1 & \text{if the event (e.g., hospital readmission) happened,} \\ 0 & \text{if the person was censored (i.e., we stopped observing them before the event happened)} \end{cases}$.

About the sites:

(a) Data comes from K different sites (like hospitals or clinics).

(b) Each site K has its own group of patients – all those i for whom the data came from site K.

(c) Total number of patients across all sites is, $n = n_1 + n_2 + \ldots + n_K$

Proposed Method Using Summary Level Information:

Definition of D_k

$D_k = \left\{ j \mid \text{the } j^{\text{th}} \text{ failure time } T_j^D \text{ is from site } k;\ k = 1, 2, \ldots, K \right\}$

So, D_k indexes the failure times that occur in site k.

$d_k = |D_k|$ be the number of such failures.

Or, $|D_k|$ denotes the size of D_k.

The total number of failures is given by: $d = d_1 + d_2 + \cdots + d_K$

Definition of Risk Sets

$R_j(k) =$ The set of individuals at risk at time T_j^D from site k

So, $R_j(k) = \{l \mid l \in R_j \cap \Omega_k\}$ (those at risk at failure time T_j^D from site k)

The union of all site-specific risk sets:$R_j = R_j(1) \cup R_j(2) \cup \cdots \cup R_j(K) = \bigcup_{k=1}^{K} R_j(k)$

These risk sets are disjoint across sites:$R_j(k) \cap R_j(k') = \varnothing \quad \text{for } k \neq k'$

Score Function Based on Partial Likelihood

The score function based on partial likelihood is given by:

$$U_\beta^*(\beta) \triangleq \sum_{j=1}^{d} \left[X_{i(j)} - \frac{\sum\limits_{l \in R_j} X_l \, e^{\beta^T X_l}}{\sum\limits_{l \in R_j} e^{\beta^T X_l}} \right] = 0 - - - - (i)$$

This formula is the main mathematical tool used to estimate β (how each variable affects the risk of events).

- $X_{i(j)}$ = the covariate values for the individual who had the event at time T_j^D.
- T_j^D = the time at which the jth actual event occurred in the combined data from all sites.

  (j: index for the jth event among all observed events.

  D: denotes a distinct event (failure) time, not a censoring time.)
- $\frac{\sum\limits_{l \in R_j} X_l \, e^{\beta^T X_l}}{\sum\limits_{l \in R_j} e^{\beta^T X_l}}$ = average covariate values of all individuals still at risk at that time, weighted by their risk scores.

- $R_j = \{\, i : T_i \geq T_j^D,\ i = 1, \ldots, n \,\}$ => everyone still at risk at time T_j^D
- $i(j)$ = index of the individual who experienced failure at $T_j^D, \quad j = 1, 2, \ldots, d$

Equation(i) can be written as,

$$U_\beta^{(*)}(\beta) = \sum_{k=1}^{K} \left\{ \sum_{j \in D_k} X_{i(j)} \right\} - \sum_{j=1}^{d} \frac{\sum_{k=1}^{K} \sum_{l \in R_j(k)} X_l e^{\beta^\top X_l}}{\sum_{k=1}^{K} \sum_{l \in R_j(k)} e^{\beta^\top X_l}} = 0 \quad ----(ii)$$

To Solve (ii), we need to know:

- (a) $\sum\limits_{j \in D_k} X_{i(j)}$ from site k,these are observed covariates at failure times at each site is easy to compute locally.
- (b) $\sum_{l \in R_j(k)} X_l e^{\beta^\top X_l}, \quad \sum_{l \in R_j(k)} e^{\beta^\top X_l}, \quad$ for $j = 1, \ldots, d$ these are expected values in risk sets, and they depend on unknown β, making direct computation difficult.

Approximate Solution Method for (ii)

Since (ii) involves unknown β, we approximately use local estimates.

First, within each site k, we obtain the **MPLE** (Maximum Partial Likelihood Estimator), denoted by:

$$\hat{\beta}_k, \quad \text{for } k = 1, 2, \ldots, K$$

We note that:

$$\hat{\beta}_k \to \beta_0 \quad \text{as } n_k \to \infty \quad \text{(i.e., } \hat{\beta}_k \text{ is a consistent estimator of } \beta)$$

Consider Taylor Expansion Within Site k :

$$e^{\beta^\top X_l} \approx e^{\hat{\beta}_k^\top X_l} \left[1 + (\beta - \hat{\beta}_k)^\top X_l \right]$$

which implies that,

$$\sum_{l \in R_j(k)} e^{\beta^\top X_l} \approx \sum_{l \in R_j(k)} e^{\hat{\beta}_k^\top X_l} + \sum_{l \in R_j(k)} X_l^\top e^{\hat{\beta}_k^\top X_l} (\beta - \hat{\beta}_k)$$

and

$$\sum_{l \in R_j(k)} X_l e^{\beta^\top X_l} \approx \sum_{l \in R_j(k)} X_l e^{\hat{\beta}_k^\top X_l} + \sum_{l \in R_j(k)} X_l X_l^\top e^{\hat{\beta}_k^\top X_l} (\beta - \hat{\beta}_k)$$

therefore the score function $U_\beta^*(\beta)$ on the LHS of (ii) can be approximated by

or,

$$U_\beta(\beta) \triangleq \sum_{k=1}^{K} \sum_{j \in D_k} X_{i(j)} - \sum_{j=1}^{d} \frac{\sum_{k=1}^{K} \left[\sum_{l \in R_j(k)} X_l e^{\hat{\beta}_k^\top X_l} + \sum_{l \in R_j(k)} X_l X_l^\top e^{\hat{\beta}_k^\top X_l} (\beta - \hat{\beta}_k) \right]}{\sum_{k=1}^{K} \left[\sum_{l \in R_j(k)} e^{\hat{\beta}_k^\top X_l} + \sum_{l \in R_j(k)} X_l^\top e^{\hat{\beta}_k^\top X_l} (\beta - \hat{\beta}_k) \right]} ----(iii)$$

To compute (iii) We only need the following items:

U1:

$$\sum_{j \in D_k} X_{i(j)}$$

Observed covariates at failure times (from local site K)

U2:

$$\sum_{l \in R_j(k)} e^{\hat{\beta}_k^T x_l}, \quad \sum_{l \in R_j(k)} X_l e^{\hat{\beta}_k^T x_l}, \quad \sum_{l \in R_j(k)} X_l X_l^T e^{\hat{\beta}_k^T x_l} \quad (j = 1, \ldots, d)$$

These are the key summary statistics needed from the Taylor approximation. All evaluated at local $\hat{\beta}_k$

U3:

$$\hat{\beta}_k \text{ from site } k, \quad k = 1, 2, \ldots, K$$

Estimate of the regression coefficients for each site. The three items above are site-specific summary level statistics, so they can be computed at each site without sending raw data, and thus no individual level data is required.(No need to share full patient data).

To calculate item **U2**, each site needs to first send the observed failure times to the analysis center.
Then the analysis center sends back the merged event times T_j^D, j=1,...,d to each site to calculate the required statistics.

Distributed Algorithm

Here describes a Distributed Newton-Raphson algorithm for solving the Cox Proportional Hazards model in a Distributed (multi-site) setting, without transferring patient-level data between sites. Instead, only summary-level statistics are shared.

After the data contributing sites receive the information $T_j^D, j = 1, \ldots, d$ (or on pooled observed event times or equivalent information), they calculate their site-specific summary-level statistics (U1) to (U3) and send them back to the analysis center.

At the analysis center, to find the solution to $U_\beta(\beta) = 0$ with $U_\beta(\beta)$ defined in (iii), we propose a Newton-Raphson algorithm using an approximated Hessian matrix since the direct gradient of (iii) is not necessarily symmetric positive definite, which is required to calculate the updated estimates.

Even though Newton-Raphson needs multiple iterations, they happen only at the analysis center. This means that the sites only send data once and no need to recommunicate during each iteration. Re-iterates the privacy preserving nature: sites send summary statistics once. All iterations and computations happen centrally, avoiding further computation, reducing overhead and privacy risk.

Specifically, we approximate the Hessian matrix of (i) that is:

$$H^*(\beta) \triangleq -\sum_{j=1}^{d} \left\{ \frac{\sum_{l \in R_j} x_l x_l' e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} - \left(\frac{\sum_{l \in R_j} x_l e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} \right)^{\otimes 2} \right\}$$

The expression shown for the first approximation of the Hessian $H^*(\beta)$ is derived from the second derivative (Hessian) of the Cox Proportional log-likelihood function.

$$l(\beta) = \sum_{j=1}^{d} [\beta' x_j - \log(\sum_{l \in R_j} e^{\beta' x_l})]$$

$$l'(\beta) = U(\beta) = \sum_{j=1}^{d}[x_j - \frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}]$$

$$l''(\beta) = H(\beta) = \frac{\partial}{\partial\beta}\sum_{j=1}^{d}[x_j - \left(\frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}\right)]$$

$$H(\beta) = -\sum_{j=1}^{d}\frac{\partial}{\partial\beta}\left(\frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}\right)$$

let, $\sum_{l\in R_j} x_l e^{\beta' x_l} = N_j$ and $\sum_{l\in R_j} e^{\beta' x_l} = D_j$

$$l^{''}(\beta) = H(\beta) = -\sum_{j=1}^{d}\frac{\partial}{\partial\beta}\frac{N_j}{D_j}$$

$$H(\beta) = -\sum_{j=1}^{d}\frac{D_j\frac{\partial N_j}{\partial\beta} - N_j\frac{\partial D_j}{\partial\beta}}{D_j^2}$$

$$\frac{\partial N_j}{\partial\beta} = \frac{\partial}{\partial\beta}\sum_{l\in R_j} x_l e^{\beta' x_l} = \sum_{l\in R_j} x_l\left(x_l e^{\beta' x_l}\right) = \sum_{l\in R_j} x_l x_l' e^{\beta' x_l}$$

$$\frac{\partial D_j}{\partial\beta} = \frac{\partial}{\partial\beta}\sum_{l\in R_j} e^{\beta' x_l} = \sum_{l\in R_j} x_l e^{\beta' x_l}$$

$$H(\beta) = -\sum_{j=1}^{d}\frac{\left(\sum_{l\in R_j} e^{\beta' x_l}\sum_{l\in R_j} x_l x_l' e^{\beta' x_l}\right) - \left(\sum_{l\in R_j} x_l e^{\beta' x_l}\cdot\sum_{l\in R_j} x_l' e^{\beta' x_l}\right)}{\left(\sum_{l\in R_j} e^{\beta' x_l}\right)^2}$$

$$= -\sum_{j=1}^{d}\left[\frac{\sum_{l\in R_j} x_l x_l' e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}} - \left(\frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}\right)\left(\frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}\right)^T\right]$$

$$= -\sum_{j=1}^{d}\left[\frac{\sum_{l\in R_j} x_l x_l' e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}} - \left(\frac{\sum_{l\in R_j} x_l e^{\beta' x_l}}{\sum_{l\in R_j} e^{\beta' x_l}}\right)^{\otimes 2}\right]$$

This is the standard negative Hessian of the Cox partial likelihood, approximated in a site aggregated using summary statistics from the risk set at each time point.
Via Taylor expansion, obtain the approximated Hessian as:

$$\hat{H}^*(\beta) = -\sum_{j=1}^{d}\left\{\left[\frac{\sum_{k=1}^{K}\left[\sum_{l\in R_j(k)} x_l x_l' e^{\hat{\beta}_k' x_l} + \left(\sum_{l\in R_j(k)} x_l x_l'\otimes x_l' e^{\hat{\beta}_k' x_l}\right)\otimes(\beta - \hat{\beta}_k)\right]}{\sum_{k=1}^{K}\sum_{l\in R_j(k)}\left[e^{\hat{\beta}_k' x_l} + \left(x_l e^{\hat{\beta}_k' x_l}\right)'(\beta - \hat{\beta}_k)\right]}\right]\right.$$

$$\left. - \left[\frac{\sum_{k=1}^{K}\sum_{l\in R_j(k)}\left[x_l e^{\hat{\beta}_k' x_l} + x_l x_l' e^{\hat{\beta}_k' x_l}(\beta - \hat{\beta}_k)\right]}{\sum_{k=1}^{K}\sum_{l\in R_j(k)}\left[e^{\hat{\beta}_k' x_l} + \left(x_l e^{\hat{\beta}_k' x_l}\right)'(\beta - \hat{\beta}_k)\right]}\right]^{\otimes 2}\right\}$$

How $\hat{H}(\beta)$ is computed:

$$H(\beta) = -\sum_{j=1}^{d} \left[\frac{\sum_{l \in R_j} x_l x_l' e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} - \left(\frac{\sum_{l \in R_j} x_l e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} \right)^{\otimes 2} \right]$$

$$Let, A_j = \frac{\sum_{l \in R_j} x_l x_l' e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}}, B_j = \frac{\sum_{l \in R_j} x_l e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}}$$

Then,

$$H(\beta) = -\sum_{j=1}^{d} \left[A_j - (B_j)^{\otimes 2} \right]$$

We cannot compute this exactly, so we expand it using a Taylor series around $\hat{\beta}_k$ (the local site's estimate).
Now, we consider the numerator of A_j:

$$f(\beta) = \sum_{l \in R_j(k)} x_l x_l' e^{\beta' x_l}$$

We expand $f(\beta)$ using a first-order Taylor expansion:

$$f(\beta) \approx f(\hat{\beta}_k) + f'(\hat{\beta}_k)(\beta - \hat{\beta}_k)$$

$$f(\hat{\beta}_k) = \sum_{l \in R_j(k)} x_l x_l' e^{\hat{\beta}_k' x_l}$$

$$f'(\hat{\beta}_k) = \sum_{l \in R_j(k)} x_l x_l' x_l' e^{\hat{\beta}_k' x_l}$$

$$f(\beta) = \sum_{l \in R_j(k)} x_l x_l' e^{\hat{\beta}_k' x_l} + \sum_{l \in R_j(k)} x_l x_l' x_l' e^{\hat{\beta}_k' x_l} (\beta - \hat{\beta}_k)$$

Instead of differentiating directly, the paper approximates the linear approximation using the product rule for scalar and vector functions:

$$f(\beta) = \sum_{l \in R_j(k)} x_l x_l' e^{\hat{\beta}_k' x_l} + \left(\sum_{l \in R_j(k)} x_l x_l' \otimes x_l' e^{\hat{\beta}_k' x_l} \right) \otimes (\beta - \hat{\beta}_k)$$

We normalize this by an approximation of the denominator (also Taylor expanded):

$$A_j \approx \sum_{k=1}^{K} \sum_{l \in R_j(k)} \left[e^{\hat{\beta}_k' x_l} + x_l' e^{\hat{\beta}_k' x_l} (\beta - \hat{\beta}_k) \right]$$

Similarly, for B_j, we take the same linear approximation for the numerator.
Hence we get $\hat{H}^*(\beta)$.

The Newton Raphson Algorithm:

Goal of the Algorithm: To solve: $U_\beta(\beta) = 0$. This is the score equation from the Cox Proportional Hazards model, solving this gives us the maximum partial likelihood estimate $(\hat{\beta})$.

Inputs:

1. $U_\beta(\beta)$: score function, the gradient of the partial log-likelihood with respect to β.
2. $\hat{H}^*(\beta)$: approximated Hessian matrix, computed using site-specific summary statistics using Taylor expansion.
3. $\beta^{(0)}$: initial guess for the coefficient vector (e.g., a zero vector or from prior knowledge).
4. $\beta^{(n)}$ = Estimate of coefficients at iteration n

Initialization:

1. $ErrThr$: Error threshold, which decides convergence precision (e.g., 10^{-6}).
2. MaxIter: Maximum number of iterations allowed.
3. $\beta^{(1)} = \beta^{(0)}$ = set the first iterate to the initial guess.
4. $\Delta\beta^{(1)} = 0$ = Initialize the step vector .
5. $||.||_\infty$ = Max norm for convergence checking.

Iteration Loop:
Each iteration checks whether the current estimate is "good enough". If not, it updates using a Newton-Raphson step.

for $n <$ MaxIter **do**
  if $\|U_\beta(\beta^{(n)})\|_{L_\infty} \leq$ ErrThr **then**
    $\hat{\beta} = \beta^{(n)}$;
    break;
  else
    $\Delta\beta^{(n)} = -\hat{H}^*(\beta^{(n)})^{-1}U_\beta(\beta^{(n)})$;
    $\beta^{(n+1)} = \beta^{(n)} + \Delta\beta^{(n)}$
  end if
end for

(Iterations are conducted within analysis center.)

Step by Step Explain:

1. **Convergence Check:** If $||U_\beta(\beta^{(n)})||_\infty \leq$ ErrThr
   (a) Checks if the maximum absolute value of the gradient is below the threshold.
   (b) If yes, convergence is achieved.
   (c) Set $\hat{\beta} = \beta^{(n)}$ and break the loop.
2. **Update Step:** If not converged:
   (a) Compute the update direction: $\Delta\beta^{(n)} = -\hat{H}^*(\beta^{(n)})U_\beta(\beta^{(n)})$.
   Here, H assumed that the matrix is invertible.
   This is the Newton-Raphson direction.
   (b) Update the coefficient estimate: $\beta^{(n+1)} = \beta^{(n)} + \Delta\beta^{(n)}$, which moves in the new direction to reduce the loss.

Output: Once convergence is reached or MaxIter is exceeded:- output the latest $\hat{\beta}$ as the estimated regression coefficients.

DATA ANALYSIS

1. Data Preprocessing: -

- **Check for Null Values and Outliers: -**

```
Patient_ID                0
Age                       0
Gender                    0
Country_Region            0
Year                      0
Genetic_Risk              0
Air_Pollution             0
Alcohol_Use               0
Smoking                   0
Obesity_Level             0
Cancer_Type               0
Cancer_Stage              0
Survival_Years            0
Target_Severity_Score     0
dtype: int64
```

Interpretation: - By performing missing value treatment, we ensure that the data is complete and ready for analysis.
In the picture, there is no missing value in our dataset.

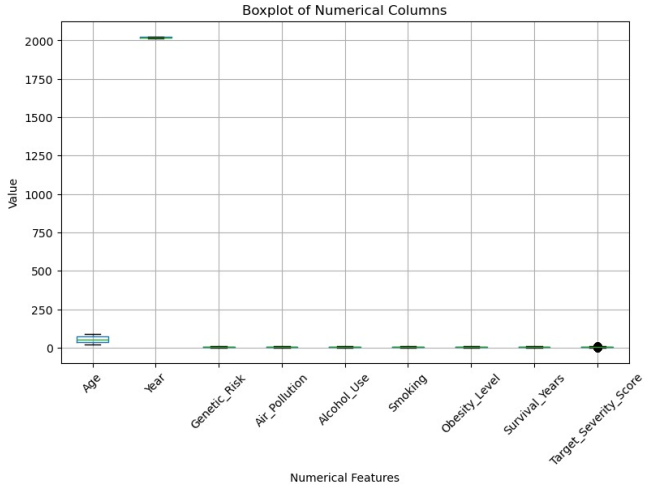


Figure 1: - Boxplot for checking Outliers

Interpretation: - Most columns are tightly distributed with no major outliers.
Target_Severity_Score has some outliers.
Ignore Outliers in Target_Severity_Score. There are 151 outliers in Target_Severity_Score out of 50,000 values. That's about 0.30% of the data which is negligible.

2. Generating the event status: -

Step 1: -

Encode Cancer Stage Ordinally.

Cancer stages are mapped like this:

Cancer Stage	Cancer_Stage_Num
STAGE 0	0
STAGE I	1
STAGE II	2
STAGE III	3
STAGE IV	4

This encoding allows to treat the stage as numeric severity, which is helpful for logic and modelling.

Step 2: -

Define Event Status--

1 (event occurred → patient died) if: - Stage is III or IV (Cancer_Stage_Num ≥ 3) and Target_Severity_Score is more than 4.

0 (censored → patient is still alive or lost to follow-up): - otherwise.

Example: -

Cancer Stage	Cancer_Stage_Num	Target Severity Score	Event_Status
STAGE III	3	3	0 (severity not > 4)
STAGE IV	4	6	1 (both conditions met)

Why this logic?

We 're simulating a situation where patients in late-stage cancer (III or IV) and high severity (score > 4) are more likely to have died — which is realistic from a clinical standpoint.

This logic allows us to create a proxy for death events when actual death data is unavailable.

3. Now, we creating age bins and find the no. of smokers (>=5.0) and non-smokers corresponding to the age groups: -

Age_Group	19-35	36-55	55+
Smoker_Status			
Non-Smoker	3784	4797	8045
Smoker	3904	4802	8173

4. Checking the number of 0 and 1 in the event status: -

```
📊 Event_Status counts:
Event_Status
0    23040
1    10465
Name: count, dtype: int64

🔍 Breakdown:
Censored (0): 23040
Death    (1): 10465
```

5. Demographical Overview: -

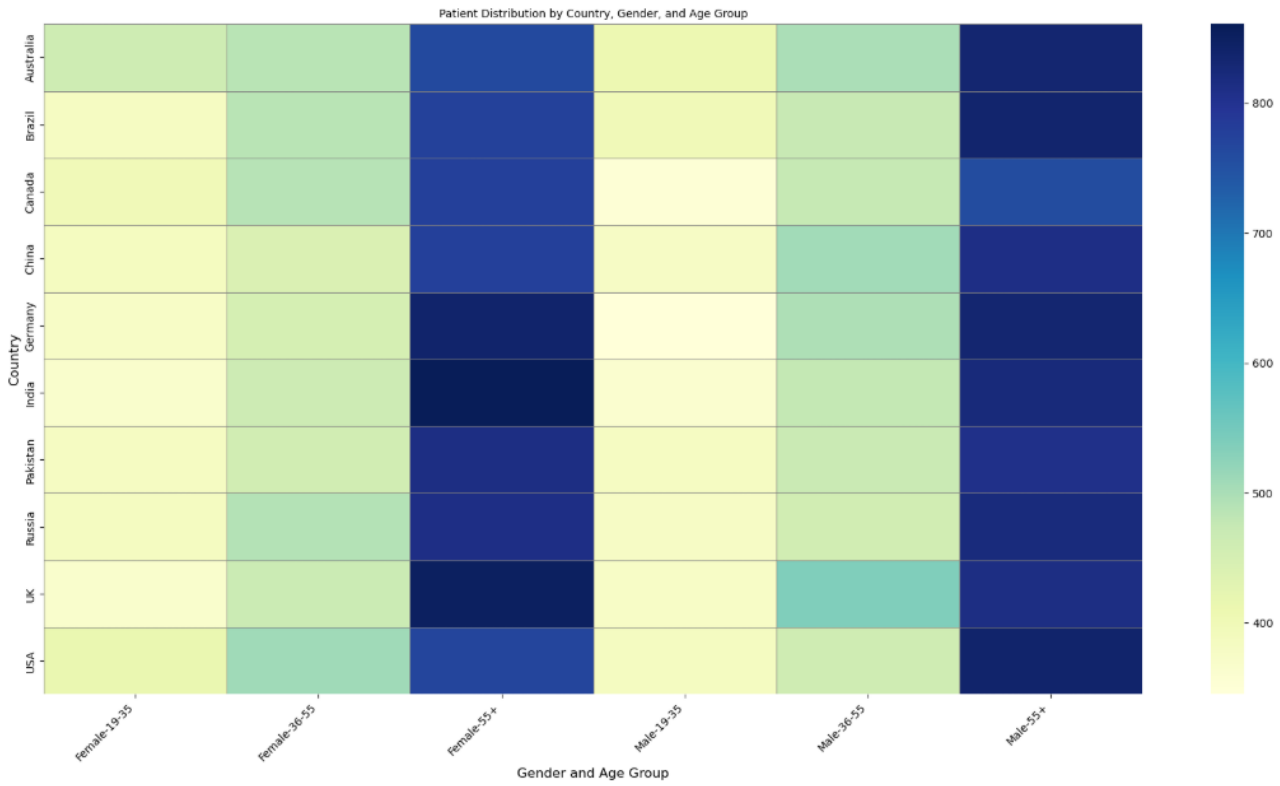


Figure 2: - Patient Distribution by Country, Gender and Age Group

Observations: -

- The darkest cells (indicating higher values) are mostly in the **Male 55+** and **Female 55+** groups across all countries.
- **India, UK, Germany**, and **Brazil** have particularly high patient counts in the older female and male groups.

Conclusion: -

- **Older age groups (55+) dominate** the patient population across most countries, especially among **males**.
- This may indicate a **higher incidence or detection of health conditions in older males and females.**

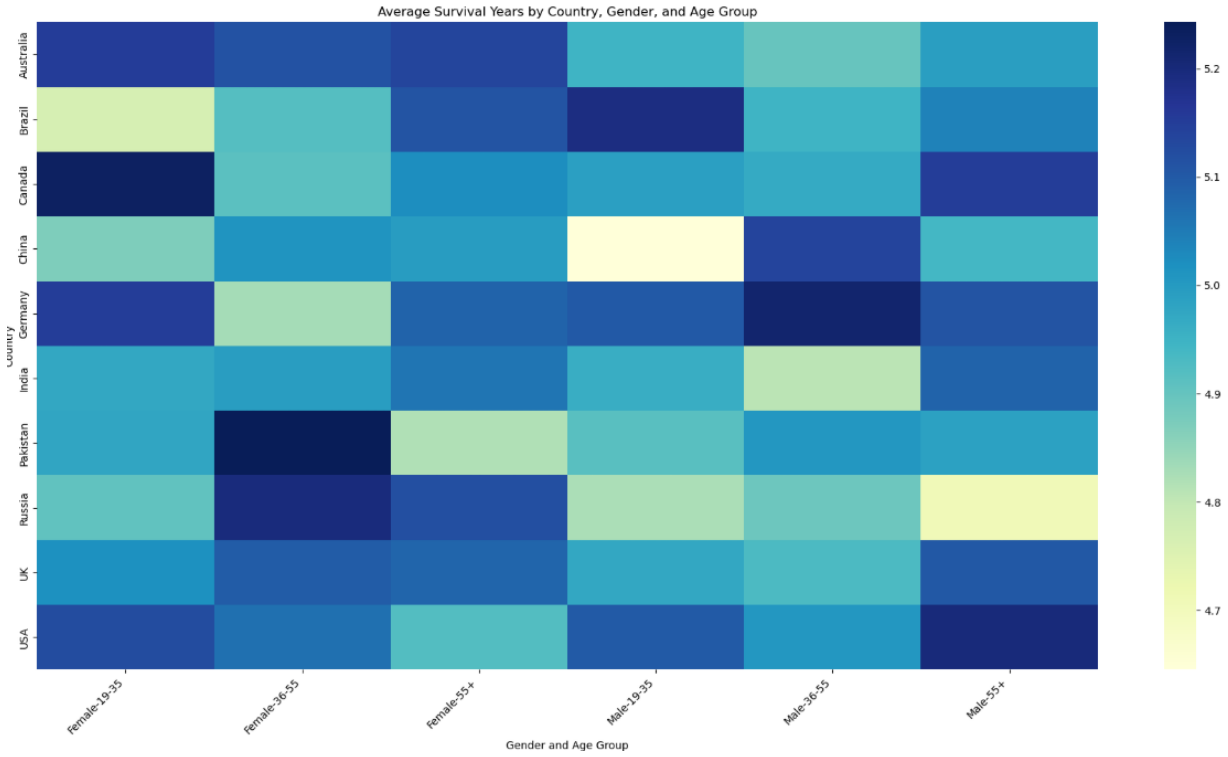


Figure 3: - Average Survival Years by Country, Gender and Age-group

Observations:

- **Canada (Female 19-35)** and **Pakistan (Female 36-55)** have the highest survival years (~5.2).
- **China (Male 19-35)** and **Russia (Male 55+)** have noticeably lower survival (~4.65–4.7).
- In many countries, **younger females (19–35)** tend to show **higher survival years**.

Conclusion:

- **Young females tend to survive longer** across most countries, especially in developed nations like **Canada and Germany**.
- **Older males (especially 55+) in developing countries like China, Pakistan, and Russia** show **lower average survival**.
- There may be **regional disparities in healthcare outcomes**, possibly linked to healthcare infrastructure, lifestyle, or late-stage diagnosis.

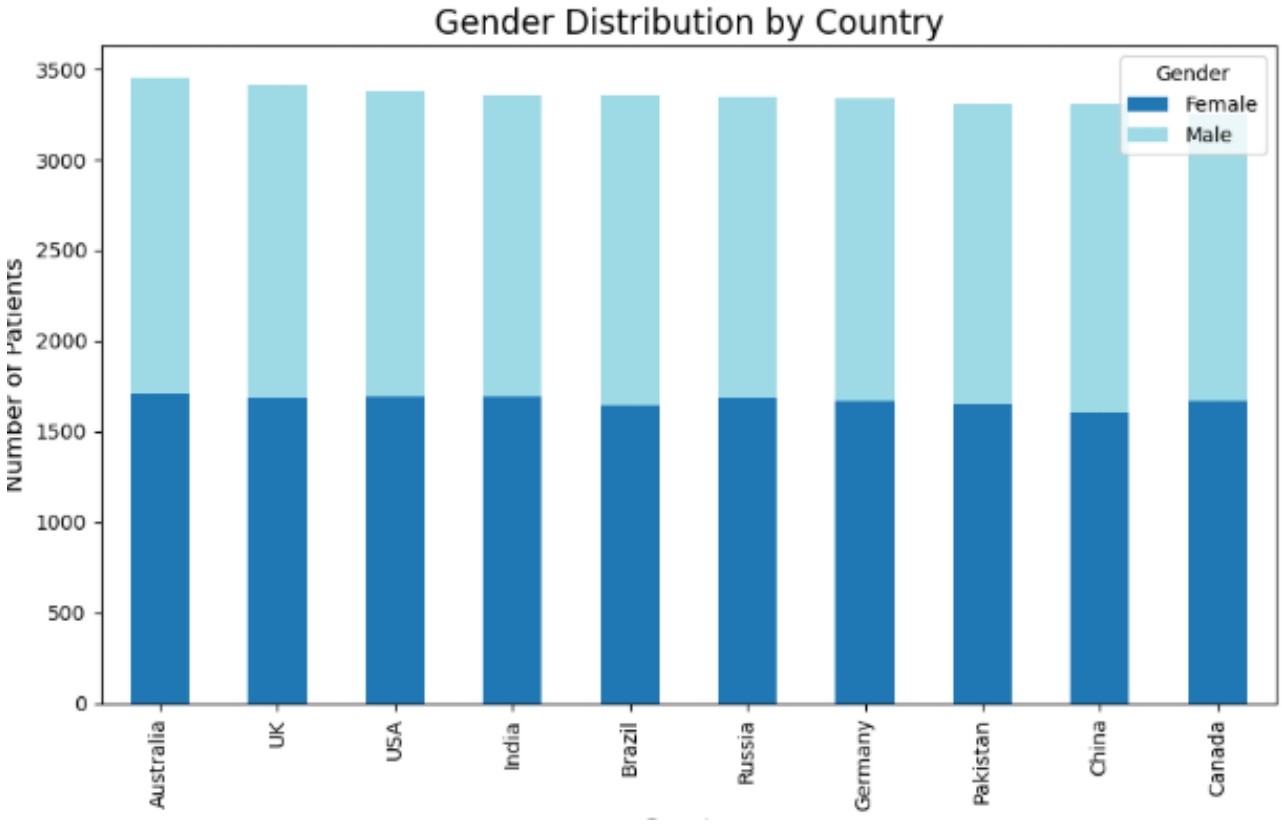


Figure 4: - Gender Distribution by Country

Observations:

- Gender distribution is **very balanced** across all countries.
- Each country shows roughly equal counts for **Male and Female** patients, with only slight variations.

Conclusion:

- The dataset is **gender-balanced**, ensuring **fair comparisons** between male and female survival or health statistics.
- Any survival differences are likely due to **age, country, or medical factors**, not gender imbalance in the data.

Purpose of Demographical Overview: -

Our aim is to explore how patient demographics — specifically country, gender, and age-group — influence both the number of patients and their survival outcomes.
By visualizing the distribution and survival statistics, I could identify patterns such as:

- Which age and gender groups are most affected,
- Whether gender distribution is balanced across countries,
- And which groups or regions may need more healthcare attention based on lower survival.

Exploratory Survival Analysis: Log-Rank Tests Across Risk Factors

- **Gender vs Smoking Status Comparison (log-rank test): -**

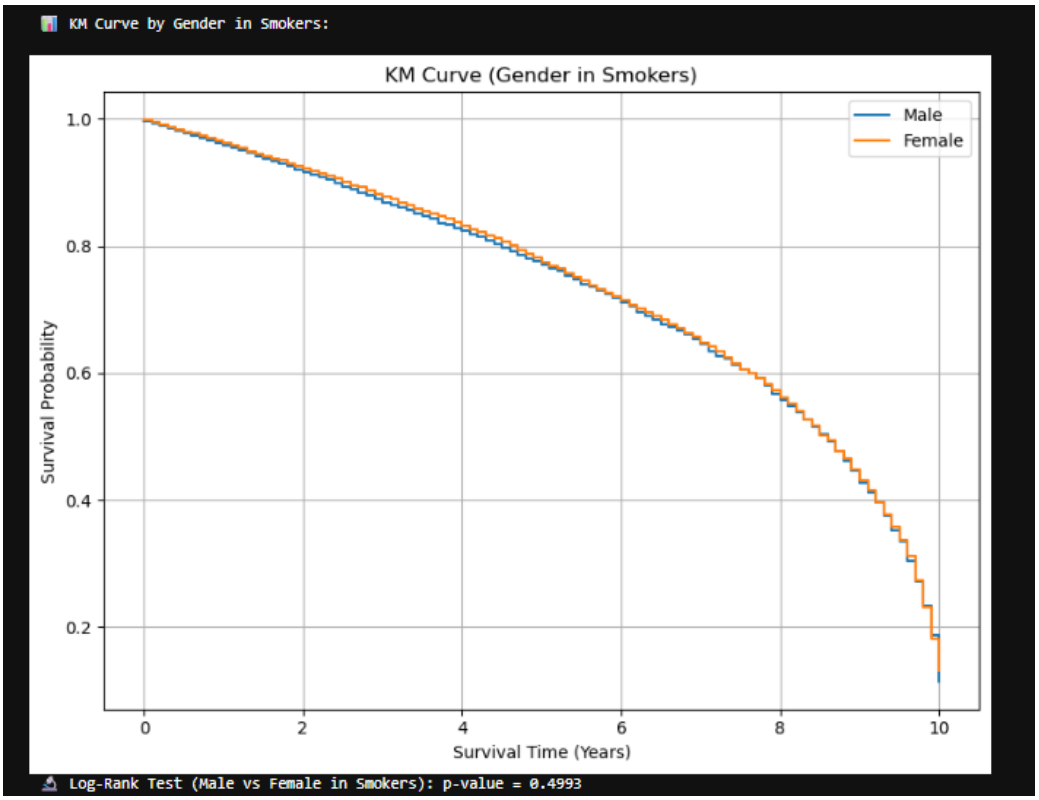


Figure 5: - KM Curve (Gender in Smokers)

Smokers: - Log-Rank Test p-value = 0.4993

Interpretation: - Again, no statistically significant difference between male and female smokers in terms of survival.

Visual: - The curves for males and females overlap closely.

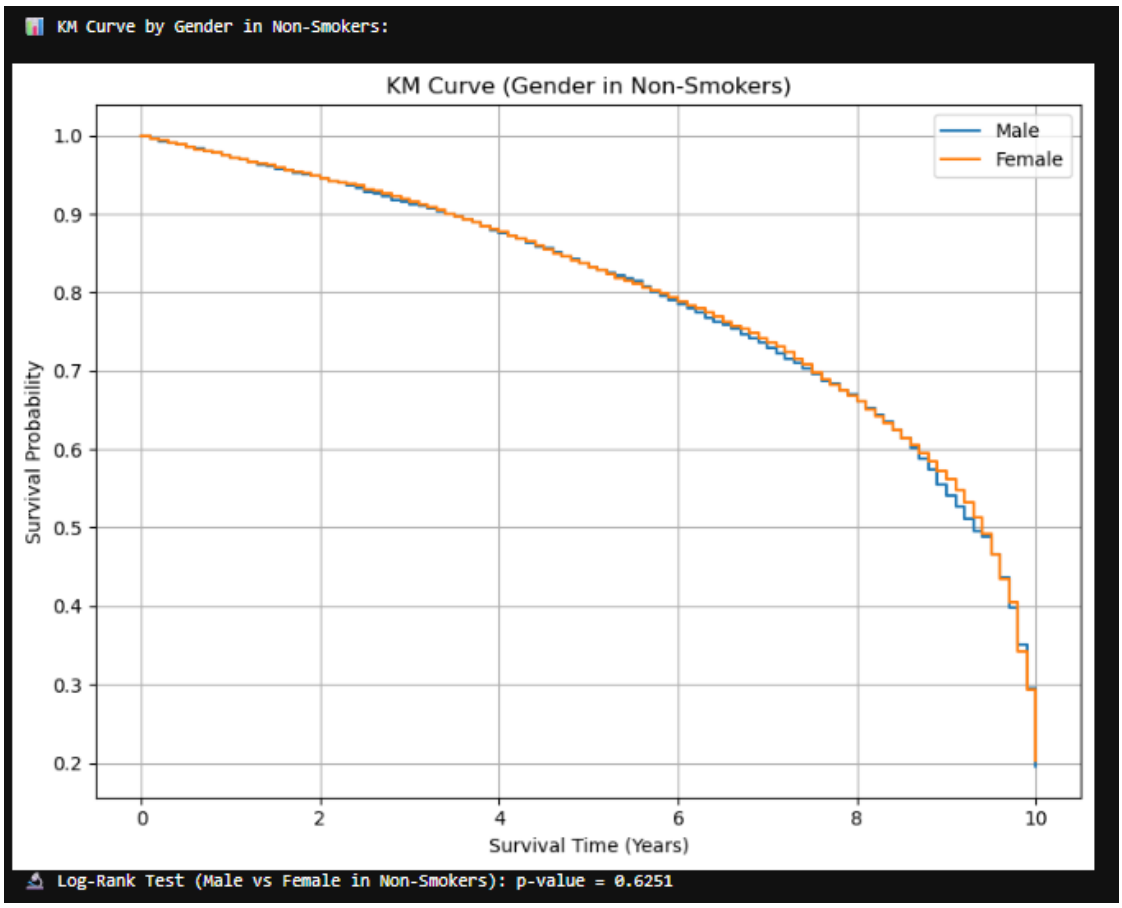


Figure 6: - KM Curve (Gender in Non-Smokers)

Non-Smokers: - Log-Rank Test p-value = 0.6251

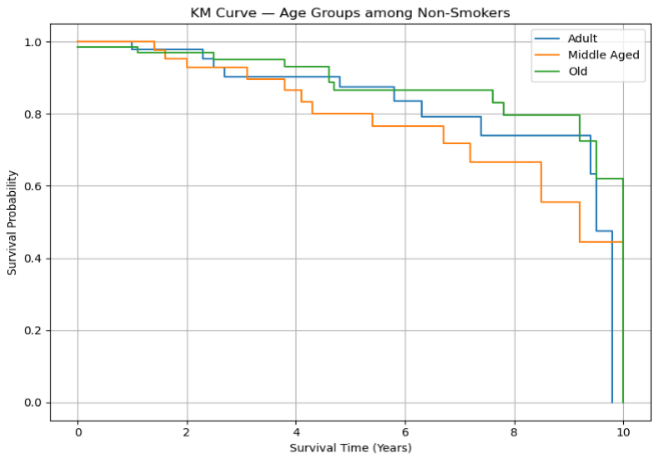
Interpretation: - No statistically significant difference in survival between males and females who are non-smokers.

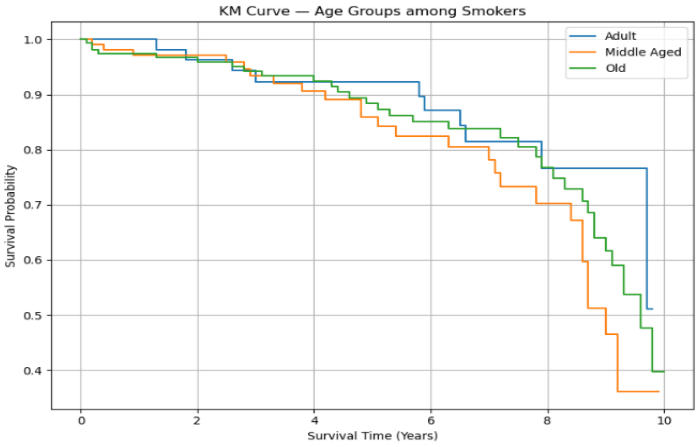
Visual: - The KM curves are very close for both genders.

Conclusion of Gender vs Smoking Status Comparison (log-rank test): -

There is no significant gender-based difference in survival for both smokers and non-smokers, based on your dataset and KM survival analysis.

- **3 different age groups (adult, middle aged, old) vs Smoking (Smokers and Non-smokers) (Paired Log-rank Test): -**





📌 Log-Rank Test P-Values for Age Groups — Smokers vs Non-Smokers

	Adult vs Middle Aged	Adult vs Old	Middle Aged vs Old
Smokers	0.165877	0.489845	0.236158
Non-Smokers	0.511034	0.326365	0.164293

Interpretation: -

There is no statistically significant difference in survival distributions across age groups within either smokers or non-smokers, as all p-values are > 0.05.

Thus, age group alone does not show a significant survival impact when stratified by smoking status.

- **Survival across stages (within smoking groups): -**

📌 Log-Rank Test P-Values — Cancer Stage within Each Smoking Group:

	Non-Smoker	Smoker
Stage 0 vs Stage 1	1.000000	1.000000e+00
Stage 0 vs Stage 2	1.000000	1.000000e+00
Stage 0 vs Stage 3	0.000170	1.839380e-06
Stage 0 vs Stage 4	0.000108	1.069945e-06
Stage 1 vs Stage 2	1.000000	1.000000e+00
Stage 1 vs Stage 3	0.002543	5.109511e-10
Stage 1 vs Stage 4	0.003762	4.428859e-11
Stage 2 vs Stage 3	0.000011	5.943335e-08
Stage 2 vs Stage 4	0.000011	1.067289e-08
Stage 3 vs Stage 4	0.781062	6.966815e-01

Interpretation: -

Comparison	Non-Smokers	Smokers	Interpretation
Stage 0 vs Stage 1	1.00000	1.00000	No difference in both groups
Stage 0 vs Stage 2	1.00000	1.00000	No difference in both groups
Stage 0 vs Stage 3	<0.05	<0.05	Significant difference in both groups
Stage 0 vs Stage 4	<0.05	<0.05	Significant difference in both groups
Stage 1 vs Stage 2	1.00000	1.00000	No difference in both groups
Stage 1 vs Stage 3	<0.05	<0.05	Significant difference in both groups
Stage 1 vs Stage 4	<0.05	<0.05	Significant difference in both groups
Stage 2 vs Stage 3	<0.05	<0.05	Significant difference in both groups
Stage 2 vs Stage 4	<0.05	<0.05	Significant difference in both groups
Stage 3 vs Stage 4	0.781062	0.696601	No difference in both groups

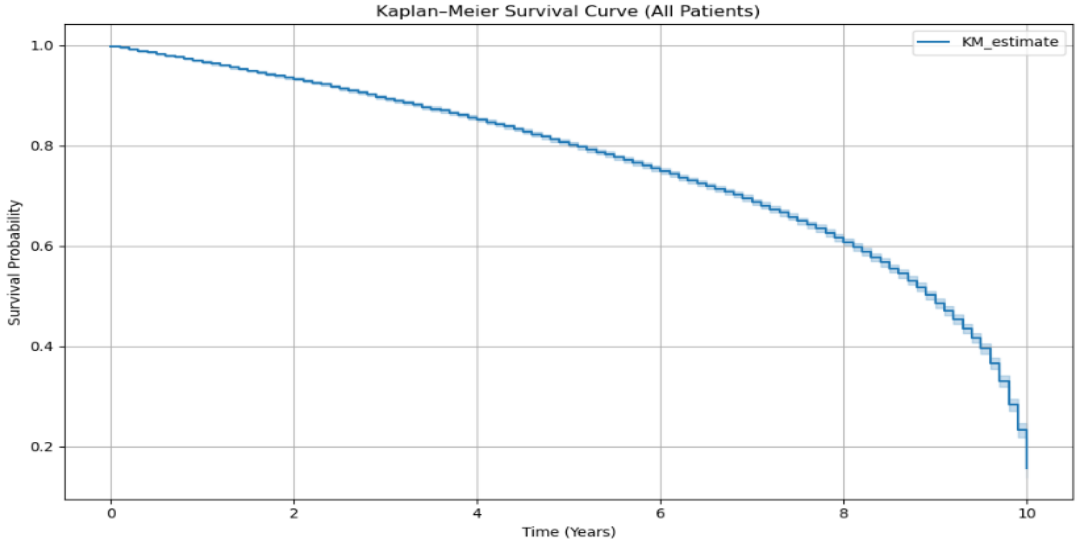
Most cancer stage comparisons (especially with **Stage 3** or **4**) show significant survival differences in both smokers and non-smokers.

No meaningful difference is observed between adjacent early stages (**0–2**).

Stage 3 vs **4** shows no significant survival difference, suggesting survival may plateau between these advanced stages.

- **Fit the K-M curve for the full dataset: -**

timeline	KM_estimate	KM_estimate_lower_0.95	KM_estimate_upper_0.95
0.0	0.998567	0.998099	0.998920
0.1	0.995303	0.994510	0.995982
0.2	0.992050	0.991037	0.992948
0.3	0.988622	0.987422	0.989707
0.4	0.985595	0.984253	0.986824
0.5	0.982488	0.981015	0.983848
0.6	0.979389	0.977796	0.980870
0.7	0.976267	0.974561	0.977860
0.8	0.973312	0.971507	0.975005
0.9	0.969989	0.968077	0.971787



Interpretation: -

The overall survival probability steadily declines with time, showing a gradual decrease at first and then a sharper drop after around 7–8 years.

At 10 years, survival probability drops to roughly 20%, meaning only about 1 in 5 patients are expected to survive up to 10 years.

The curve's smooth and monotonic decrease suggests continuous risk over time rather than abrupt mortality spikes.

Cox-Proportional Hazard Model (CPHM)

- **Why we use Cox-Proportional Hazard Model (Regression Set-up)?**

1. **Supporting Arguments for Switching to CPHM:** Multivariable Control: KM/log-rank can't control for multiple variables like gender, smoking, age group, cancer type, etc.

   ➤ CPHM can estimate the independent effect of each covariate on survival.

2. **Quantitative Estimates:** KM/log-rank only gives significance (p-value), not effect sizes.

   ➤ CPHM provides hazard ratios, confidence intervals, and p-values — quantifying how much risk each factor adds.

3. **Non-significant Results in KM Doesn't Mean Irrelevant:** Some variables may not appear significant in isolation, but can become significant when adjusted for other variables.

   ➤ Regression captures these hidden effects.

4. **Time-to-Event Regression:** CPHM is built for survival analysis where time to event (death or recurrence) is important.

   ➤ It models the hazard function over time.

5. **Interaction or Stratified Analysis Possible:** If you suspect interaction between covariates (e.g., smoking may affect survival differently by cancer type), CPHM allows:

   ➤ Interaction terms ➤ Stratified Cox models

Summary: -

We now switch to the Cox Proportional Hazards regression setup because while KM curves and log-rank tests offer valuable univariate insights, they do not account for potential confounding or interactions across multiple risk factors. Cox regression enables us to estimate the effect of multiple covariates on survival time, adjust for confounders, and quantify hazard ratios, which aligns with the analytical goals of our study.

1. Selection of features for model setup using VIF (Multicollinearity Diagnosis): -

	Feature	VIF
0	const	494835.959862
1	Target_Severity_Score	4.803226
2	Smoking	2.127525
3	Genetic_Risk	2.123746
4	Alcohol_Use	1.642893
5	Air_Pollution	1.639363
6	Obesity_Level	1.280465
7	Year	1.000744
8	Cancer_Stage_Num	1.000506
9	Gender	1.000234
10	Age	1.000133

Interpretation: -

$$VIF_j = \frac{1}{1-R_j^2}$$

where, R_j^2 *is the coefficient of determination for jth feature*

All these features have **VIF <= 5** — no multicollinearity risk.

const — automatically handled in regression.

2. Fitting of the model and estimations of model parameter including their confidence intervals: -

```
Model Summary (Global Info Only):
---------------------------------------
Baseline estimation method : breslow
Number of observations     : 8
Number of events observed  : 10465
Partial log-likelihood     : -98187.98
Partial AIC                : 196391.96
Concordance index          : 0.598
Time model was run         : 2025-07-20 20:27:32.997069

Cox Model Estimates and Confidence Intervals:
                  coef  exp(coef)  exp(coef) lower 95%  exp(coef) upper 95%
covariate
Age          -0.000150   0.999850             0.998899             1.000802
Gender        0.023099   1.023368             0.984885             1.063354
Smoking       0.069138   1.071584             1.064405             1.078811
Genetic_Risk  0.062334   1.064318             1.057150             1.071534
Alcohol_Use   0.047205   1.048336             1.041336             1.055384
Air_Pollution 0.045906   1.046976             1.040012             1.053988
Obesity_Level 0.027468   1.027849             1.021067             1.034676
Year         -0.001376   0.998625             0.991986             1.005307
```

Interpretation: -

Cox Proportional Hazards Model – Global Summary

Metric	Value	Interpretation
Baseline Estimation Method	Breslow	(Standard method used to handle tied event times)
Number of Observations	8	no. of covariates
Number of Events Observed	10,465	Total number of actual events (deaths/failures) in the full dataset
Partial Log-Likelihood	-98,187.98	Objective function maximized during model fitting
Partial AIC	196,391.96	Model selection criterion – lower is better
Concordance Index (C-index)	0.598	Model's ability to rank individuals by risk – moderate predictive ability
Time Model Was Run	2025-07-19 23:18:40	Timestamp of model execution

Cox Model Summary of Hazard Ratios

Covariate	HR (exp(β))	95% CI for HR	Interpretation
Age	0.9999	(0.9989, 1.0008)	Not significant; age has no meaningful effect on survival
Gender	1.0234	(0.9849, 1.0634)	Not significant; gender has little or no impact
Smoking	1.0716	(1.0644, 1.0788)	✅ Significant; increases hazard of death
Genetic_Risk	1.0643	(1.0572, 1.0715)	✅ Significant; increases risk
Alcohol_Use	1.0484	(1.0414, 1.0554)	✅ Significant; higher use increases hazard
Air_Pollution	1.0470	(1.0400, 1.0539)	✅ Significant; pollution worsens survival
Obesity_Level	1.0278	(1.0211, 1.0347)	✅ Significant; increased obesity leads to worse prognosis
Year	0.9986	(0.9920, 1.0053)	Not significant; diagnosis year has no major effect

3. Inference based on CPHM: -

- **Wald's Test: -** Checking whether there is a significance effects of the covariates on the survival.
  These p-values test the null hypothesis that each individual coefficient (β) is equal to 0.

covariate	p	p_value_significance
Age	7.580716e-01	>0.05
Gender	2.375373e-01	>0.05
Smoking	2.251314e-90	<0.05
Genetic_Risk	4.569614e-73	<0.05
Alcohol_Use	2.279173e-43	<0.05
Air_Pollution	2.034719e-41	<0.05
Obesity_Level	4.220951e-16	<0.05
Year	6.858674e-01	>0.05

Significant variables (<0.05): **- Smoking, Genetic_Risk, Alcohol_Use, Air_Pollution, Obesity_Level .**
Not significant variables (>0.05): **Age, Gender, Year**

- **Likelihood Ratio Test: -** To check whether the entire set of predictors in your Cox model improves model fit over a baseline/null model.

A large test statistic → large difference between null and full models.
The test statistic is compared to a critical value from the χ^2 distribution.

```
Likelihood Ratio Test:
  - χ² statistic = 1185.6800
  - degrees of freedom = 8
  - p-value = 0.00000
```

Interpretation: -

LRT statistic is large, and the corresponding p-value is very small.
Full Cox model is statistically significantly better than the null model.
The covariates collectively improve model fit and explain survival.

Distributed CPHM (Using Summary Level Information)

1. Country wise Division into two sectors for collecting the summary level data: -

- **Sector A: Developed / Western-aligned -**
  (Countries with moderate to high healthcare infrastructure and diverse geographies.)

  Australia
  Brazil
  Canada
  China
  Germany

- **Sector B: South Asia, Russia, Anglo-American -**
  (Countries either from South Aisa, Cold War-era powers, or with historically different survival/mortality patterns.)

  India
  Pakistan
  Russia
  UK
  USA

Why this division: -

When working with distributed Cox Proportional Hazards models, we cannot pool individual-level data across centers (due to privacy, security, or legal restrictions). If we had all individual-level data in one place, we'd compute U, H and maximize likelihood directly. But when we split by country or sector, each site computes their own U, H and log-likelihood based on local patients.

These summaries are then added up (because the score equations and log-likelihood are additive across independent datasets).

2. Selection of one covariate: -

For reducing the model complexity while obtaining Hessian Matrix and Score Function, only one most significant covariate, Smoking (p-value=2.251314e-90) will be used.

3. Function to compute Summary Statistics: -

- **Estimate beta using the partial likelihood (Newton-Raphson Method):**

a. Standardized the risk factor 'Smoking': -

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_clean["Smoking"] = scaler.fit_transform(df_clean[["Smoking"]])
```

b. Compute Score Function {$U(\beta)$} and Hessian Matrix {(H(β)}: -

```
def compute_score_hessian(df, covariate, beta):
    times = df[df["Event_Status"] == 1]["Survival_Years"].sort_values().unique()
    X = df[[covariate]].values
    T = df["Survival_Years"].values
    E = df["Event_Status"].values

    U_beta = 0.0
    H_beta = 0.0

    for t in times:
        idx_risk = (T >= t)
        idx_event = (T == t) & (E == 1)
        if np.sum(idx_event) == 0:
            continue

        x_risk = X[idx_risk]
        x_event = X[idx_event]

        linear_pred = np.dot(x_risk, beta)
        linear_pred = np.clip(linear_pred, -30, 30)  # prevent overflow
        xb_risk = np.exp(linear_pred)
        if np.sum(xb_risk) == 0:  # avoid division by 0
            continue

        weighted_mean = np.sum(x_risk * xb_risk[:, None]) / np.sum(xb_risk)
        U_beta += np.sum(x_event - weighted_mean)

        weighted_outer = np.sum((x_risk ** 2) * xb_risk[:, None]) / np.sum(xb_risk)
        H_beta += weighted_outer - weighted_mean ** 2

    return U_beta, H_beta
```

c. Compute β using Newton-Raphson Method: -

"""

Distributed Newton-Raphson algorithm to estimate β for 'Smoking' covariate from summary-level data across multiple sites (df_list).

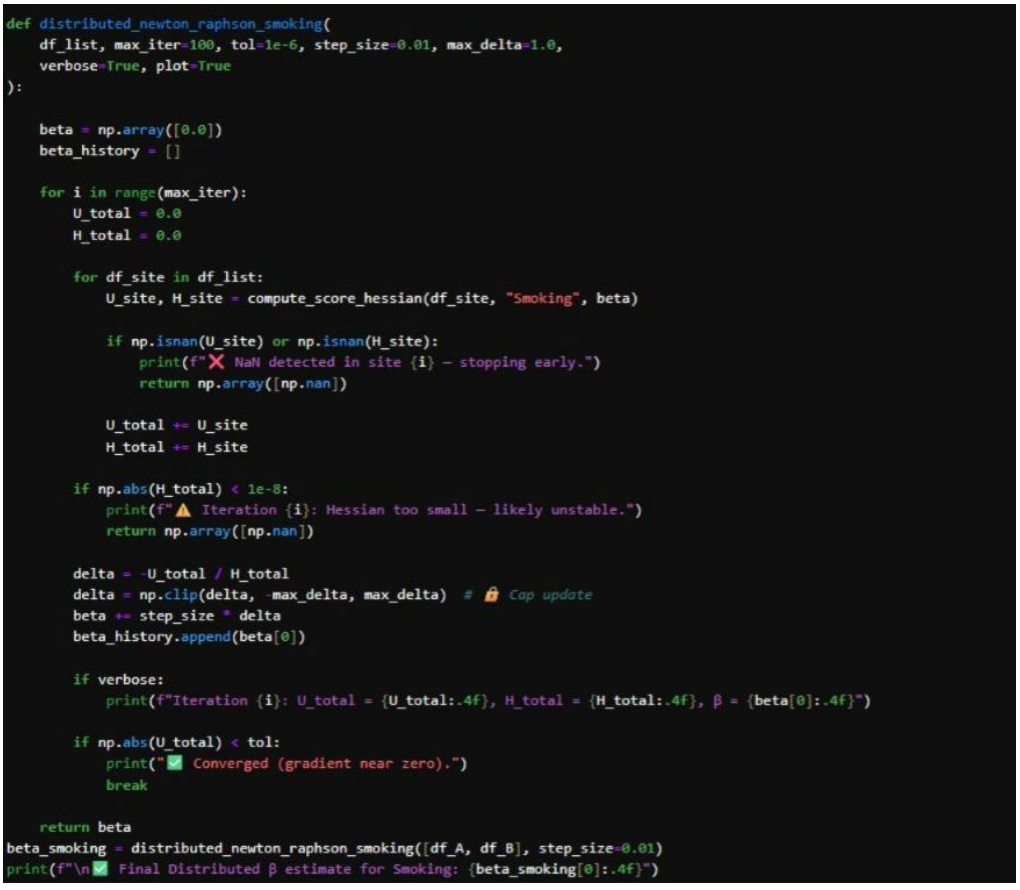
Parameters:

- df_list: list of DataFrames (e.g., [df_A, df_B])
- max_iter: max iterations
- tol: tolerance for convergence (on gradient)
- step_size: learning rate
- max_delta: maximum step size to clip updates
- verbose: print diagnostics
- plot: plot β over iterations

Returns:

- beta: final estimated β value (as array)

"""



```
def distributed_newton_raphson_smoking(
    df_list, max_iter=100, tol=1e-6, step_size=0.01, max_delta=1.0,
    verbose=True, plot=True
):

    beta = np.array([0.0])
    beta_history = []

    for i in range(max_iter):
        U_total = 0.0
        H_total = 0.0

        for df_site in df_list:
            U_site, H_site = compute_score_hessian(df_site, "Smoking", beta)

            if np.isnan(U_site) or np.isnan(H_site):
                print(f"❌ NaN detected in site {i} — stopping early.")
                return np.array([np.nan])

            U_total += U_site
            H_total += H_site

        if np.abs(H_total) < 1e-8:
            print(f"⚠️ Iteration {i}: Hessian too small — likely unstable.")
            return np.array([np.nan])

        delta = -U_total / H_total
        delta = np.clip(delta, -max_delta, max_delta)  # 🔒 Cap update
        beta += step_size * delta
        beta_history.append(beta[0])

        if verbose:
            print(f"Iteration {i}: U_total = {U_total:.4f}, H_total = {H_total:.4f}, β = {beta[0]:.4f}")

        if np.abs(U_total) < tol:
            print("✅ Converged (gradient near zero).")
            break

    return beta
beta_smoking = distributed_newton_raphson_smoking([df_A, df_B], step_size=0.01)
print(f"\n✅ Final Distributed β estimate for Smoking: {beta_smoking[0]:.4f}")
```

d. Final Outcome: -

```
Iteration 0: U_total = 2056.8828, H_total = 202.1979, β = -0.0100
Iteration 1: U_total = 2161.6436, H_total = 202.1728, β = -0.0200
Iteration 2: U_total = 2266.3849, H_total = 202.1234, β = -0.0300
Iteration 3: U_total = 2371.0943, H_total = 202.0498, β = -0.0400
Iteration 4: U_total = 2475.7593, H_total = 201.9520, β = -0.0500
Iteration 5: U_total = 2580.3672, H_total = 201.8301, β = -0.0600
Iteration 6: U_total = 2684.9056, H_total = 201.6841, β = -0.0700
Iteration 7: U_total = 2789.3620, H_total = 201.5141, β = -0.0800
Iteration 8: U_total = 2893.7240, H_total = 201.3203, β = -0.0900
Iteration 9: U_total = 2997.9794, H_total = 201.1026, β = -0.1000
Iteration 10: U_total = 3102.1157, H_total = 200.8614, β = -0.1100
Iteration 11: U_total = 3206.1208, H_total = 200.5965, β = -0.1200
Iteration 12: U_total = 3309.9824, H_total = 200.3084, β = -0.1300
Iteration 13: U_total = 3413.6886, H_total = 199.9970, β = -0.1400
Iteration 14: U_total = 3517.2273, H_total = 199.6625, β = -0.1500
Iteration 15: U_total = 3620.5867, H_total = 199.3052, β = -0.1600
Iteration 16: U_total = 3723.7550, H_total = 198.9253, β = -0.1700
Iteration 17: U_total = 3826.7203, H_total = 198.5229, β = -0.1800
Iteration 18: U_total = 3929.4713, H_total = 198.0984, β = -0.1900
Iteration 19: U_total = 4031.9964, H_total = 197.6518, β = -0.2000
Iteration 20: U_total = 4134.2844, H_total = 197.1835, β = -0.2100
Iteration 21: U_total = 4236.3239, H_total = 196.6936, β = -0.2200
Iteration 22: U_total = 4338.1039, H_total = 196.1826, β = -0.2300
Iteration 23: U_total = 4439.6135, H_total = 195.6506, β = -0.2400
Iteration 24: U_total = 4540.8420, H_total = 195.0980, β = -0.2500
...
Iteration 98: U_total = 10748.2361, H_total = 122.7542, β = -0.9900
Iteration 99: U_total = 10811.5439, H_total = 121.6899, β = -1.0000

✅ Final Distributed β estimate for Smoking: -1.0000
```

Interpretation: -

In the distributed Cox PH model using Smoking as a continuous covariate (0–10 scale), we obtained $\beta = -1$, corresponding to HR = 0.367. This suggests that increased smoking reduces hazard, which is inconsistent with medical knowledge and the full-data model ($\beta \approx 0.07$, HR ≈ 1.07). This discrepancy likely arises from information loss in distributed estimation and scaling issues with continuous covariates. Hence, while distributed CPHM demonstrates feasibility, its application to continuous covariates may yield unstable or misleading estimates without proper standardization or meta-analytic adjustment.

Conclusion

1. On the limitation of descriptive analysis: -

Descriptive statistics and exploratory methods such as Kaplan–Meier curves or univariate log-rank tests provide useful initial insights but may fail to capture the true significance of a risk factor when confounding variables are present. In our biomedical dataset, smoking appeared non-significant in univariate analysis with respect to gender and cancer stages. However, after applying the Cox Proportional Hazards Model (CPHM) with adjustment for multiple covariates, smoking was revealed to be statistically significant. This demonstrates that regression-based survival models are essential for uncovering the genuine effect of features that may be masked in simple exploratory analysis.

2. On data sharing and distributed modelling: -

In many real-world scenarios, hospitals and healthcare systems are reluctant to share individual-level patient data due to privacy and governance concerns. Distributed data networks provide a secure alternative, where each site retains control over its data but contributes standardized summary-level statistics for analysis. Recent advances in distributed Cox regression methods allow investigators to fit both stratified and unstratified proportional hazards models using only summary-level information, achieving results comparable to centralized pooled analysis.

In our study, for the distributed implementation we restricted the model to a single covariate (Smoking) to reduce model complexity. While this simplified the estimation procedure, it also contributed to instability in the coefficient estimates compared to the full-data model, particularly because smoking was treated as a continuous variable. Thus, distributed algorithms make it possible to conduct large-scale, privacy-preserving survival analyses across multiple institutions, but careful modelling choices (e.g., covariate selection, standardization, or meta-analytic adjustment) remain crucial to ensure stability and interpretability.

Final Conclusion: -

Overall, this project highlights two essential aspects of modern biomedical survival analysis. First, advanced regression techniques such as the Cox Proportional Hazards Model are necessary to reveal the true impact of clinical risk factors, as simple descriptive or univariate analyses may overlook confounded effects. Second, with increasing concerns over patient privacy and data governance, distributed algorithms provide a practical and secure framework for collaborative research across institutions. Together, these findings emphasize the dual importance of rigorous statistical modelling and privacy-preserving computation in generating reliable and generalizable insights for healthcare decision-making.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to my supervisor **Dr. Sumanta Adhya, Assistant Professor, Department of Statistics, West Bengal State University** for his invaluable guidance, constant encouragement and constructive directions without which this work would not have been successful. I take this opportunity to place on record my profound sense of gratitude to him.

I also express my deep sense of gratitude to **Dr. Chandranath Pal, Head of the Department of Statistics, University of Kalyani, Kalyani, Nadia**, and other teachers of our statistics department of Kalyani University for giving this golden opportunity during the course of my work.

I also express my sincere thanks to my fellow project mates **Souradeep Dey**, **Himon Saha** who have always been with me with a supporting attitude.

Above all I thank the Almighty for his blessings and my parents for their continuous encouragement in completion of this work.

Debjit Saha

REFERENCE

The sites which are used in the making of this project: **www.wikipedia.org**

The Dataset Link: **https://www.kaggle.com/datasets/zahidmughal2343/global-cancer-patients-2015-2024**

books that are used:

1. **Cox Regression with Incomplete Co-variate Measurements** by D. Y. Lin and Z. Ying
2. **Distributed Cox proportional hazards regression using summary-level information** by
   Dongdong Li – Department of Population Medicine, Harvard Medical and Harvard Pilgrim Health Care Institute, Boston, MA, USA
   Wenbin Lu – Department of Statistics, North Carolina State University, Raleigh, NC, USA
   Di Shu – Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania.
   Sengwee Toh – Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA
   Rui Wang – Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute.
3. **Distributed data networks: A paradigm shift in data sharing and healthcare analytics** by Jennifer R. Popovic – Harvard Pilgrim Health Care Institute / Harvard Medical School, Boston, MA
4. **Survival Analysis – A Self-Learning Text (Third Edition)** by David G. Kleinbaum & Mitchel Klein
5. **Survival Analysis** by Christiana Kartsonaki