TP ANALYSE DE SURVIE *SEMRAOUI Souhail / ZOUGUI Intissar*

Lecture et préparation des données

```
!pip install lifelines
```

```
Requirement already satisfied: lifelines in /usr/local/lib/python3.12/dist-packages (0.30.0)
Requirement already satisfied: numpy>=1.14.0 in /usr/local/lib/python3.12/dist-packages (from lifelines) (2.0.2)
Requirement already satisfied: scipy>=1.7.0 in /usr/local/lib/python3.12/dist-packages (from lifelines) (1.16.3)
Requirement already satisfied: pandas>=2.1 in /usr/local/lib/python3.12/dist-packages (from lifelines) (2.2.2)
Requirement already satisfied: matplotlib>=3.0 in /usr/local/lib/python3.12/dist-packages (from lifelines) (3.10.0)
Requirement already satisfied: autograd>=1.5 in /usr/local/lib/python3.12/dist-packages (from lifelines) (1.8.0)
Requirement already satisfied: autograd-gamma>=0.3 in /usr/local/lib/python3.12/dist-packages (from lifelines) (0.5.0)
Requirement already satisfied: formulaic>=0.2.2 in /usr/local/lib/python3.12/dist-packages (from lifelines) (1.2.1)
Requirement already satisfied: interface-meta>=1.2.0 in /usr/local/lib/python3.12/dist-packages (from formulaic>=0.2.2->life
Requirement already satisfied: narwhals>=1.17 in /usr/local/lib/python3.12/dist-packages (from formulaic>=0.2.2->lifelines)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/python3.12/dist-packages (from formulaic>=0.2.2->l
Requirement already satisfied: wrapt>=1.0 in /usr/local/lib/python3.12/dist-packages (from formulaic>=0.2.2->lifelines) (2.0
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines) (0.
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines) (11.3.
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifelines)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.12/dist-packages (from matplotlib>=3.0->lifeli
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas>=2.1->lifelines) (2025.2
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas>=2.1->lifelines) (2025
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.7->matplotlib>=3
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from lifelines import KaplanMeierFitter, NelsonAalenFitter, CoxPHFitter

# Lecture des données
df = pd.read_csv('survival_data_1000.csv')
```

```
df
```

	Age	Sex	Smoker	Comorbidities	Treatment	BMI	Physical_Activity	Time_to_Event	Event_Observed
0	73	Female	0	0	Standard	23	Moderate	63.839174	0
1	55	Male	1	1	Standard	20	Moderate	67.477398	1
2	60	Female	0	1	Standard	30	Moderate	79.181521	1
3	56	Male	0	1	Standard	28	Low	14.761940	1
4	72	Female	0	1	Standard	37	High	17.227929	1
...	...	...	...	...	...	...	...	...	...
995	60	Male	0	1	Standard	32	Low	6.494965	1
996	59	Female	0	2	Standard	25	Moderate	11.600511	0
997	54	Male	0	1	Standard	24	Low	11.820989	1
998	60	Female	1	1	Standard	18	Low	17.895198	1
999	60	Male	1	0	Standard	30	Low	0.459815	1

1000 rows × 9 columns

Next steps: Generate code with df | New interactive sheet

Transformation des variables

```
# Création de Tranche_Age
def categoriser_age(age):
    if age < 50:
        return '<50'
    elif age <= 60:
        return '50-60'
    else:
        return '>60'
```

```
df['Tranche_Age'] = df['Age'].apply(categoriser_age)

# Création de Tranche_BMI
def categoriser_bmi(bmi):
    if bmi < 18:
        return '<18'
    elif bmi <= 26:
        return '18-26'
    else:
        return '>26'

df['Tranche_BMI'] = df['BMI'].apply(categoriser_bmi)
```

```
df
```

	Age	Sex	Smoker	Comorbidities	Treatment	BMI	Physical_Activity	Time_to_Event	Event_Observed	Tranche_Age	Tranc
0	73	Female	0	0	Standard	23	Moderate	63.839174	0	>60	
1	55	Male	1	1	Standard	20	Moderate	67.477398	1	50-60	
2	60	Female	0	1	Standard	30	Moderate	79.181521	1	50-60	
3	56	Male	0	1	Standard	28	Low	14.761940	1	50-60	
4	72	Female	0	1	Standard	37	High	17.227929	1	>60	
...	...	...	...	...	...	...	...	...	...	...	...
995	60	Male	0	1	Standard	32	Low	6.494965	1	50-60	
996	59	Female	0	2	Standard	25	Moderate	11.600511	0	50-60	
997	54	Male	0	1	Standard	24	Low	11.820989	1	50-60	
998	60	Female	1	1	Standard	18	Low	17.895198	1	50-60	
999	60	Male	1	0	Standard	30	Low	0.459815	1	50-60	

1000 rows × 11 columns

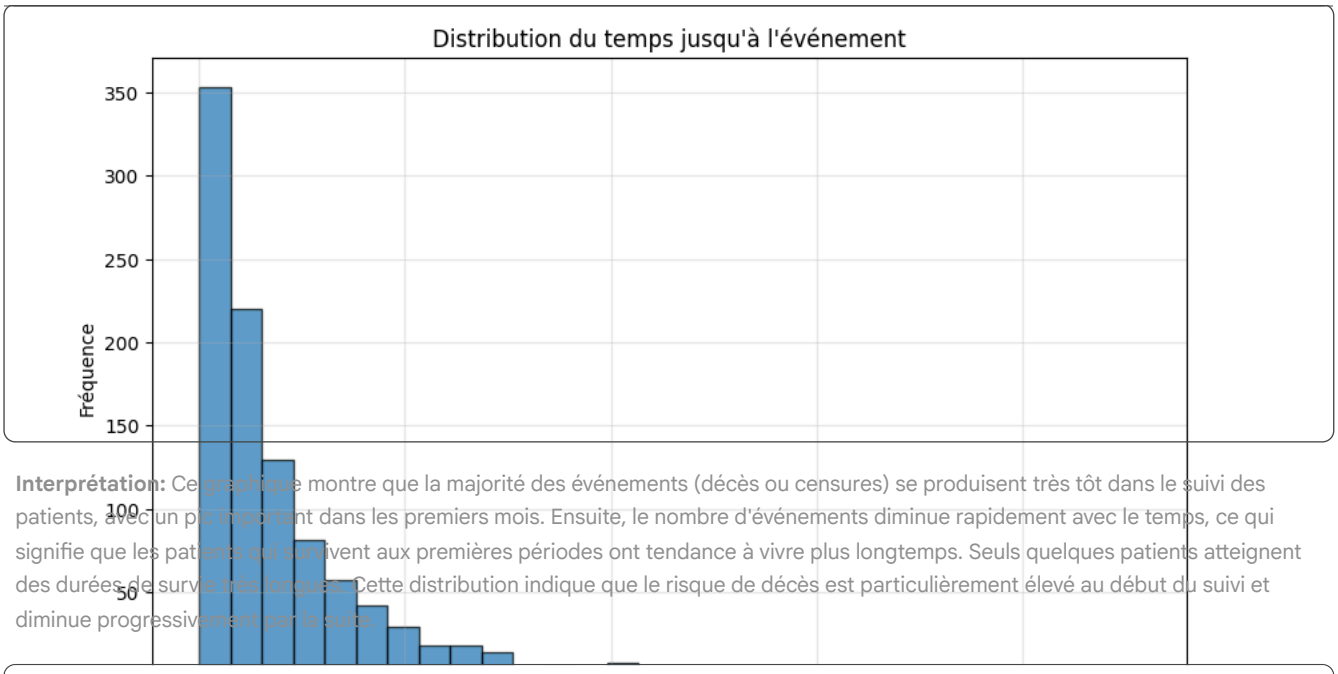
Next steps: Generate code with df | New interactive sheet

Statistiques descriptives sur la variable “time”

```
time_col = "Time_to_Event"
event_col = "Event_Observed"
desc_time = df[time_col].describe()
print(desc_time)
print("Médiane:", df[time_col].median())
print("Min:", df[time_col].min(), "Max:", df[time_col].max())
```

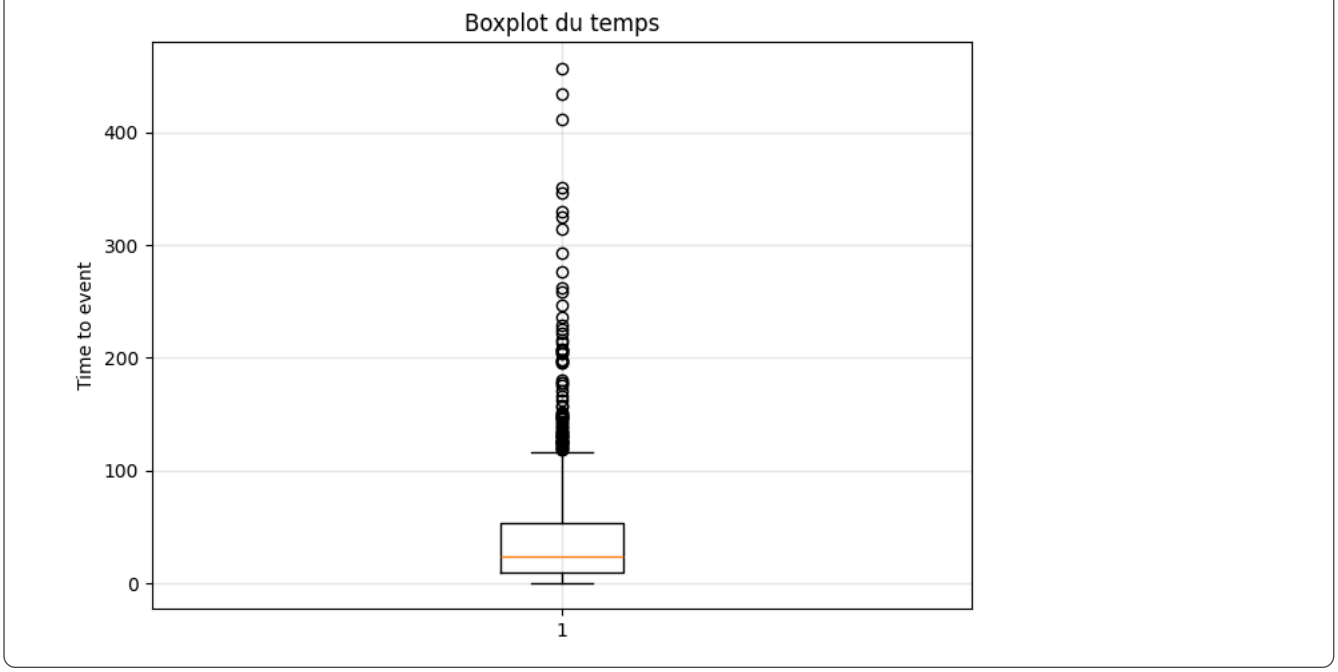
```
count    1000.000000
mean       42.420523
std        53.967320
min         0.051313
25%         9.722772
50%        23.819866
75%        52.941360
max       457.199295
Name: Time_to_Event, dtype: float64
Médiane: 23.819865500690533
Min: 0.051313492859505 Max: 457.19929457443095
```

```
plt.figure(figsize=(10, 6))
plt.hist(df['Time_to_Event'], bins=30, edgecolor='black', alpha=0.7)
plt.xlabel('Temps jusqu\'à l\'événement')
plt.ylabel('Fréquence')
plt.title('Distribution du temps jusqu\'à l\'événement')
plt.grid(True, alpha=0.3)
plt.show()
```



Interprétation: Ce graphique montre que la majorité des événements (décès ou censures) se produisent très tôt dans le suivi des patients, avec un pic important dans les premiers mois. Ensuite, le nombre d'événements diminue rapidement avec le temps, ce qui signifie que les patients qui survivent aux premières périodes ont tendance à vivre plus longtemps. Seuls quelques patients atteignent des durées de survie très longues. Cette distribution indique que le risque de décès est particulièrement élevé au début du suivi et diminue progressivement par la suite.

```
fig, ax = plt.subplots(figsize=(7, 5))

ax.boxplot(df['Time_to_Event'])
ax.set_ylabel('Time to event')
ax.set_title('Boxplot du temps')
ax.grid(True, alpha=0.3)

plt.tight_layout()
plt.show()
```



Interprétation: Ce boxplot révèle que le temps de survie médian se situe autour de 25-30 unités de temps, avec 50% des patients concentrés dans l'intervalle entre 10 et 50 unités de temps. La présence de nombreuses valeurs aberrantes au-dessus de 100 unités indique qu'une minorité de patients survit beaucoup plus longtemps que la majorité. Cette forte dispersion vers les valeurs élevées confirme l'asymétrie de la distribution observée précédemment et montre une grande hétérogénéité dans les durées de survie au sein de la population étudiée. La boîte relativement compacte en bas du graphique suggère que les temps de survie courts sont très homogènes, tandis que les survies longues sont plus variables et exceptionnelles.

4. Analyse des variables qualitatives et quantitatives

4.1 Quantitatives (Age, BMI, Comorbidities, Time_to_Event)

```
quant_cols = ["Age", "BMI", "Comorbidities", "Time_to_Event"]
print("\n------Quantitatives------")
print(df[quant_cols].describe())
```

```
------Quantitatives------
               Age          BMI  Comorbidities  Time_to_Event
count  1000.000000  1000.000000    1000.000000    1000.000000
mean     59.603000    24.534000       0.966000      42.420523
std       9.871009     5.189391       0.916343      53.967320
min      31.000000     9.000000       0.000000       0.051313
25%      53.000000    21.000000       0.000000       9.722772
50%      59.000000    25.000000       1.000000      23.819866
75%      67.000000    28.000000       1.000000      52.941360
max      87.000000    41.000000       5.000000     457.199295
```

4.2 Qualitatives (Sex, Treatment, Physical_Activity, Tranche_Age, Tranche_BMI)

```
qual_cols = ["Sex", "Treatment", "Physical_Activity", "Tranche_Age", "Tranche_BMI", "Smoker"]
for c in qual_cols:
    print("\n", c)
    print(df[c].value_counts())
    print((df[c].value_counts(normalize=True)*100).round(2))
```

```
Name: count, dtype: int64
Sex
Male      50.6
Female    49.4
Name: proportion, dtype: float64

 Treatment
Treatment
Standard        726
Experimental    274
Name: count, dtype: int64
Treatment
Standard        72.6
Experimental    27.4
Name: proportion, dtype: float64

 Physical_Activity
Physical_Activity
Moderate    490
Low         293
High        217
Name: count, dtype: int64
Physical_Activity
Moderate    49.0
Low         29.3
High        21.7
Name: proportion, dtype: float64

 Tranche_Age
Tranche_Age
>60      446
50-60    402
<50      152
Name: count, dtype: int64
Tranche_Age
>60      44.6
50-60    40.2
<50      15.2
Name: proportion, dtype: float64

 Tranche_BMI
Tranche_BMI
18-26    557
>26      350
<18       93
Name: count, dtype: int64
Tranche_BMI
18-26    55.7
>26      35.0
<18       9.3
Name: proportion, dtype: float64

 Smoker
Smoker
0    612
1    388
Name: count, dtype: int64
Smoker
```

Analyse de survise avec la méthode de Kaplan-Meier

1-Estimer la probabilité de survie et l'intervalle de confiance

```
time_col = "Time_to_Event"
event_col = "Event_Observed"

kmf = KaplanMeierFitter()
kmf.fit(
    durations=df[time_col],
    event_observed=df[event_col],
    label="Survie globale"
)

# Afficher les résultats
print("=== Résultats de l'analyse de Kaplan-Meier ===")
print(f"Temps médian de survie : {kmf.median_survival_time_:.2f}")
print(f"\nNombre total de sujets : {len(df)}")
print(f"Nombre d'événements observés : {df['Event_Observed'].sum()}")
```

```
=== Résultats de l'analyse de Kaplan-Meier ===
Temps médian de survie : 36.38

Nombre total de sujets : 1000
Nombre d'événements observés : 711
```

Interprétation: Environ 71 % des individus ont connu l'événement pendant la période de suivi, les autres étant censurés. Temps médian de survie : 36.38 → Cela signifie que 50 % des individus ont connu l'événement avant environ 36 unités de temps.

2-Le tableau des proportions de survivants

```
# Fonction de survie S(t)
survival_table = kmf.survival_function_
ci_table = kmf.confidence_interval_

result_table = survival_table.join(ci_table)
result_table.columns = ["S(t)", "IC_inf", "IC_sup"]

print(result_table.head(15))
print("\nDernières lignes:")
print(result_table.tail(10))
print(f"\nTableau complet avec {len(survival_table)} temps d'événements")

print("\nSurvie à des temps clés :")
for t in [12, 24, 36, 48, 60]:
    idx = survival_table.index.searchsorted(t)
    if idx < len(survival_table):
        survie = survival_table.iloc[idx, 0]
        print(f"  À {t} mois : {survie:.3f} ({survie*100:.1f}%)")
```

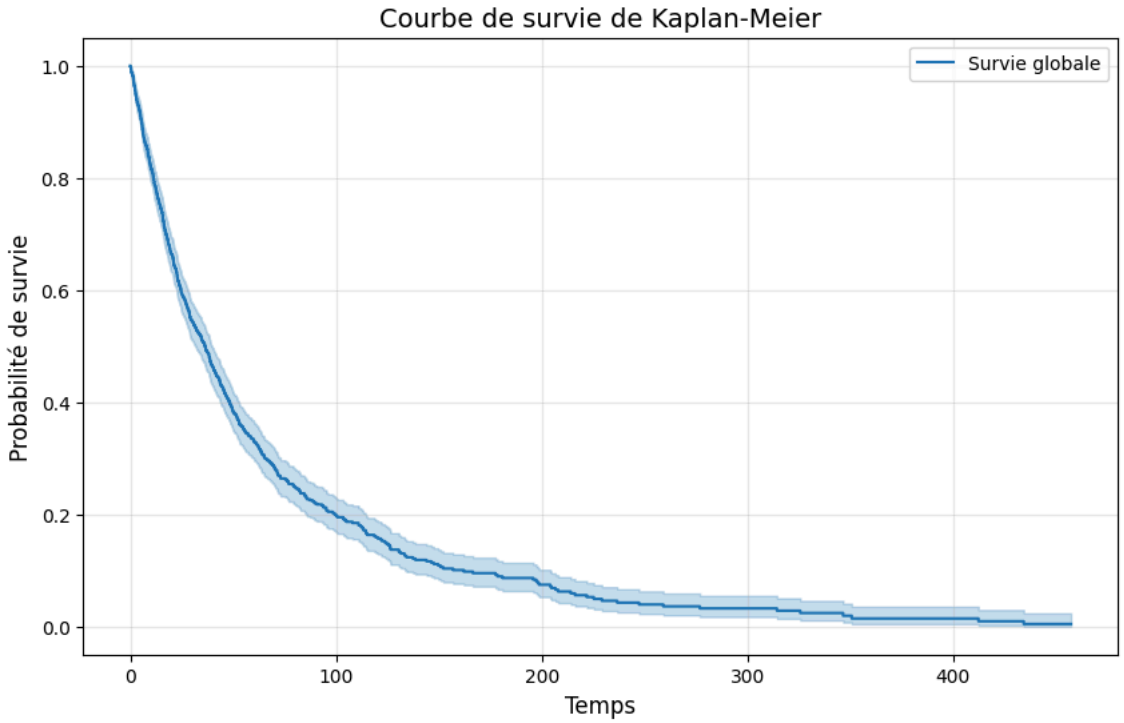
```
              S(t)    IC_inf    IC_sup
timeline
0.000000  1.000000  1.000000  1.000000
0.051313  0.999000  0.992923  0.999859
0.068754  0.998000  0.992027  0.999499
0.090925  0.998000  0.992027  0.999499
0.146223  0.998000  0.992027  0.999499
0.318730  0.996998  0.990721  0.999031
0.341391  0.996998  0.990721  0.999031
0.345882  0.995995  0.989364  0.998495
0.356617  0.995995  0.989364  0.998495
0.380436  0.995995  0.989364  0.998495
0.383972  0.995995  0.989364  0.998495
0.396064  0.994989  0.988003  0.997911
0.416682  0.993983  0.986656  0.997292
0.438470  0.992977  0.985325  0.996646
0.439082  0.991971  0.984009  0.995977

Dernières lignes:
                S(t)    IC_inf    IC_sup
timeline
276.816224  0.032759  0.017619  0.055376
293.000987  0.032759  0.017619  0.055376
314.312551  0.028664  0.014435  0.050939
325.632972  0.024569  0.011421  0.046355
329.845339  0.024569  0.011421  0.046355
346.453917  0.019655  0.007870  0.041293
350.643734  0.014742  0.004778  0.035866
412.236358  0.009828  0.002250  0.030053
434.249031  0.004914  0.000511  0.024046
457.199295  0.004914  0.000511  0.024046

Tableau complet avec 1001 temps d'événements
```

```
Survie à des temps clés :
  À 12 mois : 0.787 (78.7%)
  À 24 mois : 0.608 (60.8%)
  À 36 mois : 0.503 (50.3%)
  À 48 mois : 0.402 (40.2%)
  À 60 mois : 0.332 (33.2%)
```

Interprétation: Ce tableau présente l'estimation de la fonction de survie S(t) selon la méthode de Kaplan-Meier à différents instants du suivi, accompagnée de ses intervalles de confiance à 95%. On observe une décroissance progressive de la probabilité de survie au fil du temps, débutant à 100% au temps initial et diminuant à chaque événement observé. Les probabilités de survie à des temps clés révèlent qu'à 12 mois, 78,7% des patients sont encore en vie, puis cette proportion chute à 60,8% à 24 mois, 50,3% à 36 mois (confirmant le temps médian de survie calculé précédemment), 40,2% à 48 mois et 33,2% à 60 mois. L'élargissement progressif des intervalles de confiance au cours du temps traduit l'augmentation de l'incertitude statistique liée à la diminution du nombre de patients encore sous observation. Ces résultats démontrent une mortalité continue et importante tout au long de la période de suivi, avec une perte moyenne d'environ 20% de la population tous les deux ans.

3-La courbe de survie globale avec son intervalle de confiance

```
# Tracer la courbe
plt.figure(figsize=(10, 6))
kmf.plot_survival_function()
plt.title('Courbe de survie de Kaplan-Meier', fontsize=14)
plt.xlabel('Temps', fontsize=12)
plt.ylabel('Probabilité de survie', fontsize=12)
plt.grid(True, alpha=0.3)
plt.show()
```



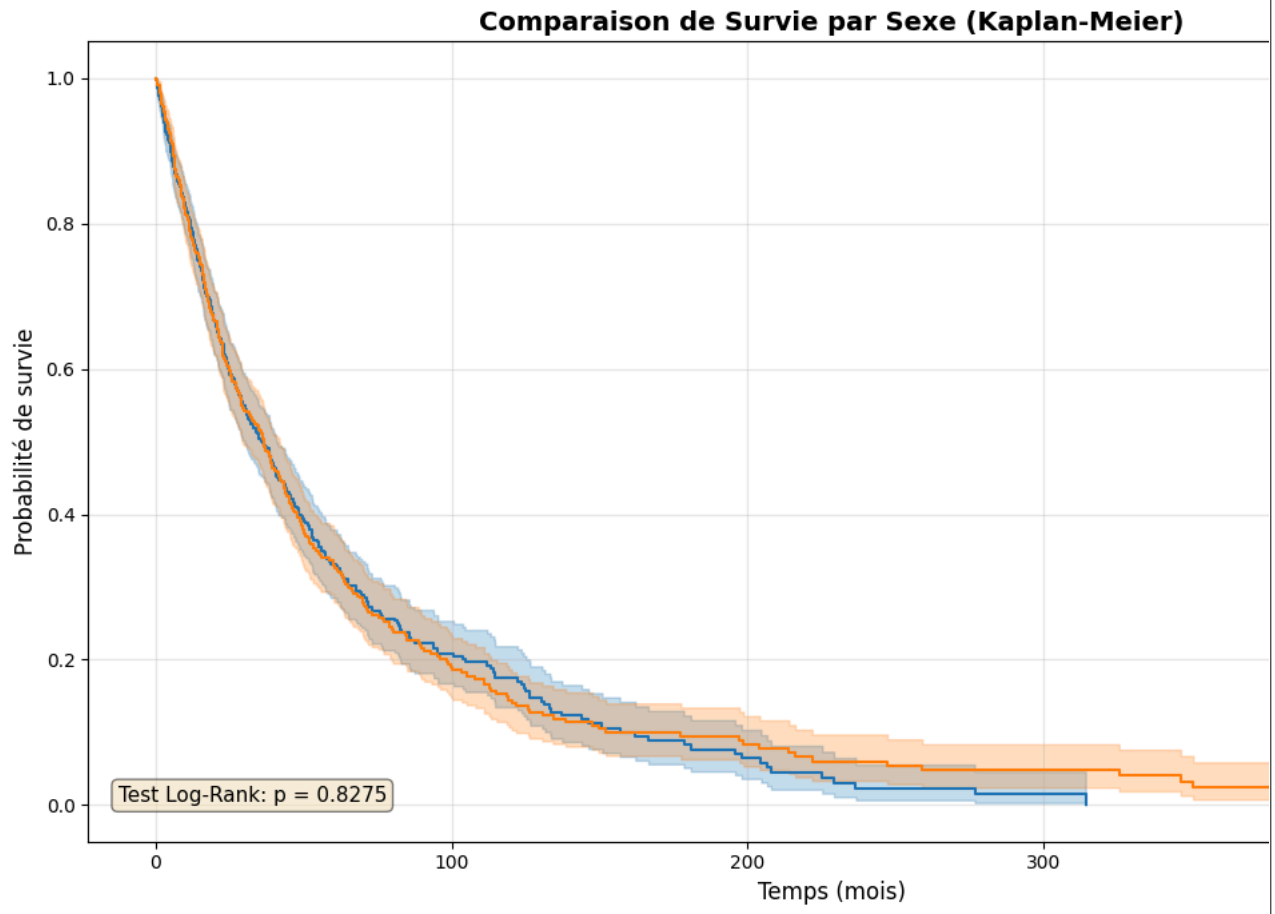
Interprétation : La courbe de Kaplan-Meier montre l'évolution de la probabilité de survie au fil du temps pour l'ensemble de la population étudiée. On observe une décroissance rapide et marquée au début, avec une chute brutale de la probabilité de survie de 100% à environ 60% dans les 50 premières unités de temps, confirmant que la période initiale est la plus critique pour les patients. Cette diminution se poursuit de manière progressive mais ralentie entre 50 et 200 unités de temps, où la probabilité de survie passe d'environ 60% à 10%. Au-delà de 200 unités, la courbe s'aplatit presque complètement, indiquant que les quelques patients qui ont survécu jusqu'à ce stade ont une probabilité de survie qui se stabilise proche de zéro. La bande d'intervalle de confiance (zone bleue claire autour de la courbe) s'élargit au fil du temps, reflétant une incertitude croissante des estimations due à la diminution du nombre de patients encore sous observation. Cette courbe illustre clairement un pronostic défavorable avec un risque de décès particulièrement élevé dans les phases précoces de la maladie.

4-Comparaison de la survie en focntion du sexe

```
from lifelines.statistics import logrank_test

# Séparer les données par sexe
df_male = df[df['Sex'] == 'Male']
```

```
df_female = df[df['Sex'] == 'Female']

print(f"\nNombre d'hommes : {len(df_male)}")
print(f"Nombre de femmes : {len(df_female)}")

# Ajuster les modèles KM pour chaque sexe
kmf_male = KaplanMeierFitter()
kmf_female = KaplanMeierFitter()

kmf_male.fit(durations=df_male['Time_to_Event'],
             event_observed=df_male['Event_Observed'],
             label='Hommes')

kmf_female.fit(durations=df_female['Time_to_Event'],
               event_observed=df_female['Event_Observed'],
               label='Femmes')

print(f"\nTemps de survie médian :")
print(f"  Hommes : {kmf_male.median_survival_time_:.2f} mois")
print(f"  Femmes : {kmf_female.median_survival_time_:.2f} mois")

# TEST LOG-RANK
print("\n--- TEST LOG-RANK ---")
results = logrank_test(durations_A=df_male['Time_to_Event'],
                       durations_B=df_female['Time_to_Event'],
                       event_observed_A=df_male['Event_Observed'],
                       event_observed_B=df_female['Event_Observed'])

print(f"Statistique de test : {results.test_statistic:.4f}")
print(f"p-value : {results.p_value:.4f}")

print("\n--- CONCLUSION ---")
if results.p_value < 0.05:
    print("Différence SIGNIFICATIVE entre les sexes (p < 0.05)")
    print("Le sexe a un effet significatif sur la survie")
else:
    print("Différence NON significative entre les sexes (p >= 0.05)")
    print("Le sexe n'a PAS d'effet significatif sur la survie")
    print("Les courbes de survie sont similaires pour les deux sexes")

# Tracer les courbes
plt.figure(figsize=(12, 7))
kmf_male.plot_survival_function(ci_show=True)
kmf_female.plot_survival_function(ci_show=True)
plt.title('Comparaison de Survie par Sexe (Kaplan-Meier)', fontsize=14, fontweight='bold')
plt.xlabel('Temps (mois)', fontsize=12)
plt.ylabel('Probabilité de survie', fontsize=12)
plt.text(0.02, 0.05, f'Test Log-Rank: p = {results.p_value:.4f}',
         transform=plt.gca().transAxes, fontsize=11,
         bbox=dict(boxstyle='round', facecolor='wheat', alpha=0.5))
plt.grid(True, alpha=0.3)
plt.legend()
plt.tight_layout()
plt.show()
```

```
Nombre d'hommes : 506
Nombre de femmes : 494

Temps de survie médian :
  Hommes : 36.24 mois
  Femmes : 36.58 mois

--- TEST LOG-RANK ---
Statistique de test : 0.0475
p-value : 0.8275

--- CONCLUSION ---
Différence NON significative entre les sexes (p >= 0.05)
Le sexe n'a PAS d'effet significatif sur la survie
Les courbes de survie sont similaires pour les deux sexes
```



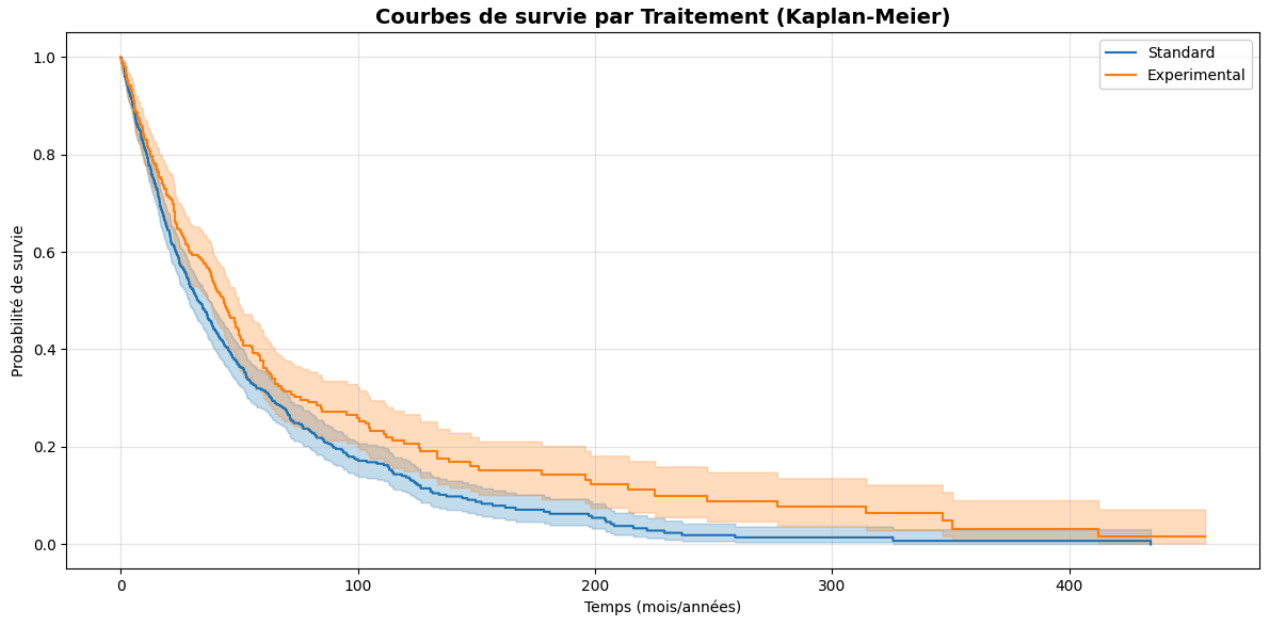
Interprétation: Cette analyse compare la survie entre les hommes et les femmes à l'aide de courbes de Kaplan-Meier distinctes pour chaque sexe. L'échantillon comprend 506 hommes et 494 femmes, avec des temps de survie médians très proches : 36,24 mois pour les hommes et 36,58 mois pour les femmes. Le test de Log-Rank a été réalisé pour évaluer si cette différence est statistiquement significative, donnant une statistique de test de 0,0475 et une p-value de 0,8275. Puisque la p-value est largement supérieure au seuil de significativité de 0,05, on conclut qu'il n'existe pas de différence significative de survie entre les deux sexes. Les courbes de survie présentées confirment visuellement cette conclusion, montrant une superposition quasi-parfaite des trajectoires de survie pour les hommes et les femmes tout au long de la période d'observation. Cela suggère que le sexe n'est pas un facteur déterminant de la survie dans cette population étudiée, et que d'autres variables explicatives devront être explorées pour identifier les facteurs de risque réellement influents.

Comparaison par traitement

```
print("\nComparaison de survie par TRAITEMENT:")
print("-" * 80)

fig, ax = plt.subplots(figsize=(12, 6))

for treatment in df['Treatment'].unique():
    mask = df['Treatment'] == treatment
    kmf_treat = KaplanMeierFitter()
    kmf_treat.fit(durations=df[mask]['Time_to_Event'],
                  event_observed=df[mask]['Event_Observed'],
                  label=treatment)
    kmf_treat.plot_survival_function(ax=ax, ci_show=True)
```

```
plt.title('Courbes de survie par Traitement (Kaplan-Meier)', fontsize=14, fontweight='bold')
plt.xlabel('Temps (mois/années)')
plt.ylabel('Probabilité de survie')
plt.legend()
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```

```
Comparaison de survie par TRAITEMENT:
--------------------------------------------------------------------------------
```



Interprétation: Cette analyse compare la survie des patients selon le type de traitement reçu : traitement standard (en bleu) versus traitement expérimental (en orange). Les courbes de Kaplan-Meier révèlent une légère différence entre les deux groupes, avec le traitement expérimental semblant offrir un avantage de survie, particulièrement visible dans la phase intermédiaire du suivi (entre 100 et 300 mois). On observe que la courbe orange (traitement expérimental) se maintient systématiquement au-dessus de la courbe bleue (traitement standard), suggérant des probabilités de survie légèrement supérieures à différents moments du suivi. Cependant, les deux courbes restent relativement proches et leurs intervalles de confiance se chevauchent largement, ce qui indique que cette différence pourrait ne pas être statistiquement significative. Un test de Log-Rank serait nécessaire pour confirmer si l'écart observé entre les deux traitements est réellement significatif ou s'il relève simplement de la variabilité d'échantillonnage. Cette analyse préliminaire suggère néanmoins que le traitement expérimental pourrait avoir un effet bénéfique modéré sur la survie des patients.

Estimation de la fonction de risque(Hazard rate)

a-Estimation de la fonction de risque cumulé - Nelson-Aalen

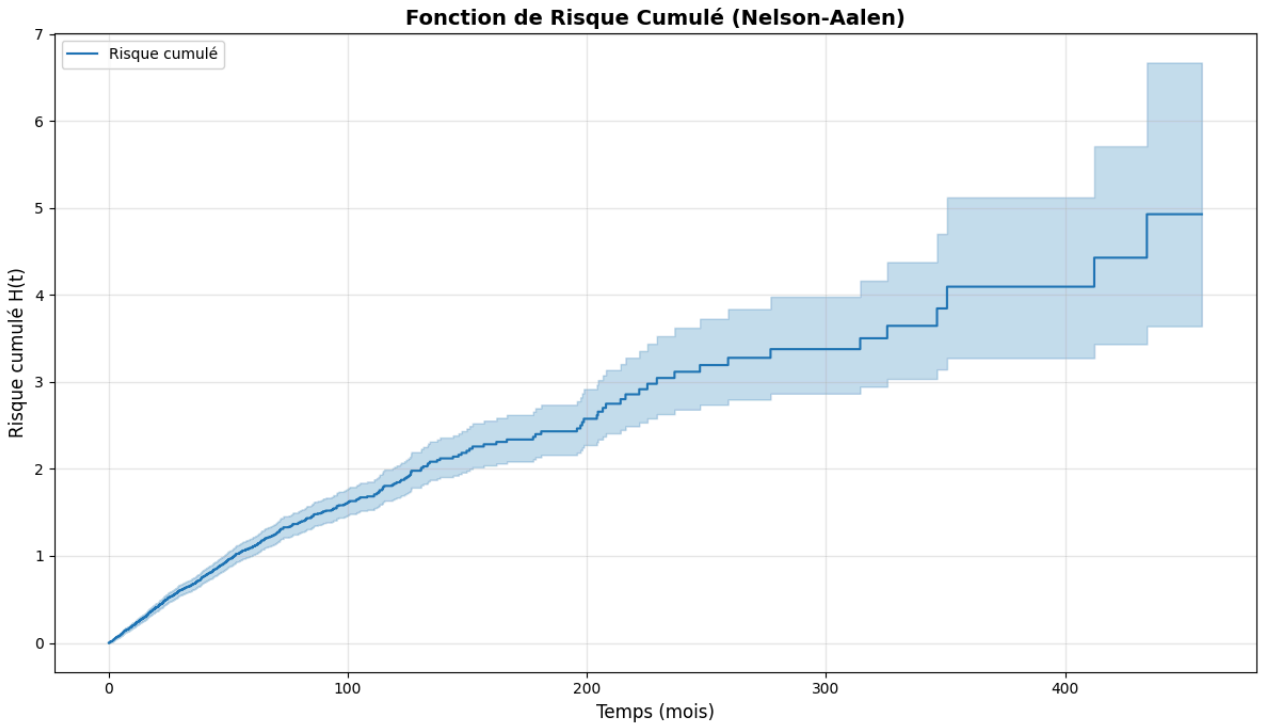
```
# Créer l'objet Nelson-Aalen
naf = NelsonAalenFitter()

# Ajuster le modèle
naf.fit(durations=df['Time_to_Event'],
        event_observed=df['Event_Observed'],
        label='Risque cumulé')

print("\nModèle Nelson-Aalen ajusté")
print("\n10 premières valeurs du risque cumulé :")
print(naf.cumulative_hazard_.head(10))
```

```
Modèle Nelson-Aalen ajusté

10 premières valeurs du risque cumulé :
          Risque cumulé
timeline
0.000000       0.000000
```

```
0.051313       0.001000
0.068754       0.002001
0.090925       0.002001
0.146223       0.002001
0.318730       0.003005
0.341391       0.003005
0.345882       0.004011
0.356617       0.004011
0.380436       0.004011
```

b-La courbe du risque cumulé

```
plt.figure(figsize=(12, 7))
naf.plot_cumulative_hazard(ci_show=True)
plt.title('Fonction de Risque Cumulé (Nelson-Aalen)', fontsize=14, fontweight='bold')
plt.xlabel('Temps (mois)', fontsize=12)
plt.ylabel('Risque cumulé H(t)', fontsize=12)
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```



c-Explication de l'évolution du risque: Dans notre étude, la courbe de risque cumulé présente une forme concave avec une pente initialement forte qui s'atténue progressivement. Le risque augmente rapidement durant les 100 premiers mois (pente forte), indiquant un taux de décès élevé dans cette période critique. Au-delà de 100 mois, la pente devient plus faible et quasi-linéaire, suggérant que le risque instantané diminue pour les patients qui ont survécu à la phase initiale. Cela traduit un pattern typique où les patients les plus fragiles décèdent précocement, tandis que ceux qui survivent aux premières phases présentent une meilleure résistance à long terme.

Le risque cumulé H(t) est une fonction CROISSANTE :

- Il augmente continuellement avec le temps
- Il ne peut jamais diminuer

La PENTE de la courbe indique l'intensité du risque :

- Pente FORTE ⟶ risque élevé de décès à cette période
- Pente FAIBLE ⟶ risque faible de décès à cette période

Forme de la courbe :

- CONCAVE (ralentit) ⟶ le risque diminue avec le temps ⟶ Les survivants à long terme sont plus résistants
- LINÉAIRE ⟶ risque constant dans le temps ⟶ Pas de période plus critique qu'une autre

- CONVEXE (accélère) → le risque augmente avec le temps → La maladie s'aggrave progressivement)

```
# Analyser la forme
h_values = naf.cumulative_hazard_.values.flatten()
t_values = naf.cumulative_hazard_.index.values

if len(h_values) > 10:
    n = len(h_values)
    pente_debut = (h_values[n//10] - h_values[0]) / (t_values[n//10] - t_values[0]) if t_values[n//10] != t_values[0] else
    pente_fin = (h_values[-1] - h_values[-n//10]) / (t_values[-1] - t_values[-n//10]) if t_values[-1] != t_values[-n//10] e

    print(f"\nANALYSE DE NOS DONNÉES :")
    print(f"Pente au début : {pente_debut:.6f}")
    print(f"Pente à la fin : {pente_fin:.6f}")

    if pente_debut > pente_fin * 1.2:
        print("\n→ Courbe CONCAVE : le risque DIMINUE avec le temps")
        print("→ Période critique initiale, puis amélioration pour les survivants")
    elif pente_fin > pente_debut * 1.2:
        print("\n→ Courbe CONVEXE : le risque AUGMENTE avec le temps")
        print("→ La maladie s'aggrave progressivement")
    else:
        print("\n→ Courbe LINÉAIRE : risque relativement CONSTANT")
        print("→ Pas de période particulièrement critique")
```

```
ANALYSE DE NOS DONNÉES :
Pente au début : 0.019126
Pente à la fin : 0.009251

→ Courbe CONCAVE : le risque DIMINUE avec le temps
→ Période critique initiale, puis amélioration pour les survivants
```

d-Estimation de la survie pour un temps saisi en mois

```
def estimer_survie(temps_mois):
    """
    Estime la probabilité de survie à un temps donné (en mois)
    """
    try:
        # Trouver l'index le plus proche
        idx = kmf.survival_function_.index.searchsorted(temps_mois)
        if idx >= len(kmf.survival_function_):
            idx = len(kmf.survival_function_) - 1

        survie = kmf.survival_function_.iloc[idx, 0]
        ci_lower = kmf.confidence_interval_survival_function_.iloc[idx, 0]
        ci_upper = kmf.confidence_interval_survival_function_.iloc[idx, 1]

        return survie, ci_lower, ci_upper
    except:
        return None, None, None

# Demander le temps à l'utilisateur
temps = float(input("\nSaisissez le temps (en mois) pour estimer la survie : "))

# Calculer la survie
survie, ci_lower, ci_upper = estimer_survie(temps)

# Afficher les résultats
if survie is not None:
    print(f"\n--- RÉSULTATS POUR T = {temps} MOIS ---")
    print(f"Probabilité de survie : {survie:.4f} ({survie*100:.2f}%)")
    print(f"Intervalle de confiance à 95% : [{ci_lower:.4f}, {ci_upper:.4f}]")
    print(f"Entre {ci_lower*100:.2f}% et {ci_upper*100:.2f}%")
else:
    print("Erreur lors du calcul de la survie")
```

```
Saisissez le temps (en mois) pour estimer la survie : 12

--- RÉSULTATS POUR T = 12.0 MOIS ---
Probabilité de survie : 0.7872 (78.72%)
Intervalle de confiance à 95% : [0.7596, 0.8120]
Entre 75.96% et 81.20%
```

A-Que représente la fonction de risque cumulée?

La fonction de risque cumulée H(t) estimée par Nelson-Aalen représente :

1. Définition: H(t) = somme des taux de risque instantanés de 0 à t C'est le RISQUE TOTAL ACCUMULÉ de décès depuis le début jusqu'au temps t

2. Interprétation:

- Mesure l'intensité CUMULÉE du risque de décès au fil du temps
- Plus H(t) est élevé, plus le risque accumulé est important

3. Relation avec la survies: S(t) = exp(-H(t))

Donc :

- Si H(t) = 0 ⟶ S(t) = 1 (survie parfaite)
- Si H(t) ⟶ ∞ ⟶ S(t) ⟶ 0 (décès certain)

4. Propriétés:

- H(0) = 0 (pas de risque au temps zéro)
- H(t) est CROISSANTE (ne diminue jamais)
- H(t) ⟶ ∞ quand t ⟶ ∞ (si tous décèdent)

5. Avantage: Nelson-Aalen est plus précis que Kaplan-Meier pour estimer S(t) quand il y a beaucoup de données censurées)

B-Comment évolue le risque cumulatif avec le temps? Que peut-on en déduire sur la maladie étudiée?

1. Evolution générale:

Le risque cumulé H(t) est STRICTEMENT CROISSANT

- Il augmente continuellement
- Il ne peut jamais diminuer

2. Analyse de la pente:

La pente de H(t) = taux de risque instantané h(t)

- Pente FORTE ⟶ risque ÉLEVÉ de décès (période critique)
- Pente FAIBLE ⟶ risque FAIBLE de décès (période stable)

**3. Forme de la courbe et implications: **

a) Courbe CONCAVE (ralentit avec le temps) : ⟶ Le risque DIMINUE au fil du temps ⟶ Les patients qui survivent longtemps ont un risque décroissant ⟶ Implications : * Surveillance intensive au Début * Espoir croissant pour les survivants à long terme * Le traitement initial est crucial

b) Courbe LINÉAIRE (constante) : ⟶ Risque CONSTANT dans le temps ⟶ Pas de période plus critique qu'une autre ⟶ Implications : * Surveillance constante nécessaire * Risque uniforme, indépendant de la durée de survie

c) Courbe CONVEXE (accélère avec le temps) : ⟶ Le risque AUGMENTE au fil du temps ⟶ La maladie s'aggrave progressivement ⟶ Implications : * Surveillance accrue au fil du temps * Importance du suivi à long terme * Risque croissant avec la progression

4. Ce qu'on peut déduire sur la maladie:

-Type de pathologie (aiguë vs chronique) : La forme concave de notre courbe de risque cumulé suggère une pathologie avec une phase aiguë critique en début de suivi. Le risque élevé initial suivi d'une stabilisation indique que la maladie étudiée présente des complications précoces sévères plutôt qu'une évolution chronique progressive. Les patients qui franchissent le cap des 100 premiers mois ont un pronostic relativement stable. **-Efficacité du traitement dans le temps :** La diminution progressive du taux de risque instantané (pente décroissante) suggère que les traitements administrés sont efficaces pour stabiliser les patients qui survivent à la phase initiale. Cela peut également indiquer que les interventions thérapeutiques précoces sont cruciales pour améliorer la survie à long terme. Les patients répondeurs au traitement montrent une meilleure résistance au fil du temps. **-Identification des périodes critiques :** La période critique se situe clairement dans les 100 premiers mois de suivi, où la pente de la courbe de risque cumulé est la plus forte. Cette phase nécessite une surveillance médicale intensive et des interventions thérapeutiques agressives. Au-delà de cette période, le risque se stabilise, permettant un suivi moins intensif. **-Pronostic pour les survivants à long terme :** Les patients qui survivent au-delà de 100 mois ont un pronostic favorable avec un risque de décès nettement réduit. Cette observation est encourageante car elle indique qu'une fois la phase critique dépassée, les patients peuvent espérer une survie prolongée avec un risque relativement faible. Cela justifie des efforts thérapeutiques intensifs en début de traitement pour maximiser les chances de survie à long terme.

Dans notre cas, la courbe de risque cumulé présente une forme concave, avec une pente forte dans les 100 premiers mois qui s'atténue progressivement par la suite. Cela indique que la maladie étudiée présente une phase critique en début de suivi, avec un risque de décès particulièrement élevé durant cette période. Les patients qui survivent au-delà de cette phase initiale voient leur risque instantané diminuer, suggérant qu'ils sont plus résistants ou que le traitement devient plus efficace à long terme. Cette observation souligne l'importance cruciale d'une surveillance et d'une intervention thérapeutique intensive durant les premiers mois de suivi.

Régression de COX

A-Ajuster un modèle de COX

```
from lifelines import CoxPHFitter


#Sélection des variables demandées
cox_vars = [
    "Sex",
    "Age",
    "Smoker",
    "Treatment",
    "Physical_Activity",
    "Time_to_Event",
    "Event_Observed"
]

df_cox = df[cox_vars].dropna().copy()

#Encodage des variables catégorielles
df_cox = pd.get_dummies(
    df_cox,
    columns=["Sex", "Treatment", "Physical_Activity"],
    drop_first=True
)

#Ajustement du modèle de Cox
cph = CoxPHFitter()
cph.fit(
    df_cox,
    duration_col="Time_to_Event",
    event_col="Event_Observed"
)

#Résumé du modèle
cph.print_summary()
```

model	lifelines.CoxPHFitter
duration col	'Time_to_Event'
event col	'Event_Observed'
baseline estimation	breslow
number of observations	1000
number of events observed	711
partial log-likelihood	-4126.67
time fit was run	2026-02-03 16:43:10 UTC

	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)
Age	0.03	1.03	0.00	0.03	0.04	1.03	1.04	0.00	8.60	<0.005	56.83
Smoker	0.39	1.48	0.08	0.24	0.55	1.27	1.73	0.00	4.96	<0.005	20.44
Sex_Male	-0.03	0.97	0.08	-0.18	0.12	0.83	1.12	0.00	-0.44	0.66	0.60
Treatment_Standard	0.34	1.41	0.09	0.17	0.51	1.19	1.67	0.00	3.99	<0.005	13.90
Physical_Activity_Low	0.73	2.08	0.11	0.52	0.94	1.69	2.57	0.00	6.83	<0.005	36.77
Physical_Activity_Moderate	0.38	1.46	0.10	0.19	0.57	1.21	1.77	0.00	3.88	<0.005	13.23

Concordance	0.64
Partial AIC	8265.33
log-likelihood ratio test	148.19 on 6 df

Interprétation:

-Age

- HR = 1.03
- p-value < 0.005

Interprétation : Chaque augmentation d'une unité d'âge est associée à une augmentation du risque d'environ 3 %, toutes choses égales par ailleurs. Facteur de risque significatif, effet progressif mais cumulatif.

-Smoker

- HR = 1.48
- IC 95 % ≈ [1.27 ; 1.73]
- p-value < 0.005

Interprétation : Les individus fumeurs présentent un risque de survenue de l'événement environ 48 % plus élevé que les non-fumeurs, à caractéristiques comparables. Facteur de risque majeur et significatif.

-Sex_Male

- HR = 0.97
- p-value = 0.66

Interprétation : Le sexe masculin n'est pas significativement associé au risque de survenue de l'événement après ajustement sur les autres variables du modèle. Aucun effet significatif du sexe, ce qui est cohérent avec le chevauchement observé dans les courbes de Kaplan–Meier.

-Treatment_Standard

- HR = 1.41
- IC 95 % ≈ [1.19 ; 1.67]
- p-value < 0.005

Interprétation : Les individus fumeurs présentent un risque de Les individus recevant le traitement standard présentent un risque de survenue de l'événement environ 41 % plus élevé que ceux recevant le traitement expérimental (modalité de référence).Le traitement expérimental apparaît comme protecteur, ce qui confirme les résultats observés avec Kaplan–Meier.

-Physical_Activity_Low

- HR = 2.08
- IC 95 % ≈ [1.69 ; 2.57]
- p-value < 0.005

Interprétation : Une faible activité physique est associée à un risque de survenue de l'événement plus que doublé par rapport à une activité physique élevée (référence).Facteur de risque très important, avec l'effet le plus fort du modèle.

-Physical_Activity_Moderate

- HR = 1.46
- IC 95 % ≈ [1.21 ; 1.77]
- p-value < 0.005

Interprétation : Une activité physique modérée est également associée à une augmentation significative du risque (≈ 46 %) par rapport à une activité élevée. Gradient clair d'effet selon le niveau d'activité physique.

Conclusion Le modèle de Cox multivarié met en évidence plusieurs facteurs significativement associés au risque de survenue de l'événement. L'âge, le tabagisme, le traitement standard et un faible niveau d'activité physique augmentent significativement le risque, tandis que le sexe n'apparaît pas comme un facteur explicatif indépendant après ajustement. L'activité physique se distingue comme le déterminant le plus fortement associé au risque, avec un effet graduel selon le niveau d'intensité.

B-Vérifier les hypothèses du modèle de Cox

Hypothèse de proportionnalité des risques

```
cph.check_assumptions(
    df_cox,
    p_value_threshold=0.05,
    show_plots=True
)
```

```
   Bootstrapping lowess lines. May take a moment...

   Bootstrapping lowess lines. May take a moment...

   Bootstrapping lowess lines. May take a moment...

   Bootstrapping lowess lines. May take a moment...

   Bootstrapping lowess lines. May take a moment...

   Bootstrapping lowess lines. May take a moment...
```

The ``p_value_threshold`` is set at 0.05. Even under the null hypothesis of no violations, some covariates will be below the threshold by chance. This is compounded when there are many covariates. Similarly, when there are lots of observations, even minor deviances from the proportional hazard assumption will be flagged.

With that in mind, it's best to use a combination of statistical tests and visual tests to determine the most serious violations. Produce visual plots using ``check_assumptions(..., show_plots=True)`` and looking for non-constant lines. See link [A] below for a full example.

null_distribution	chi squared
degrees_of_freedom	1
model	<lifelines.CoxPHFitter: fitted with 1000 total...
test_name	proportional_hazard_test

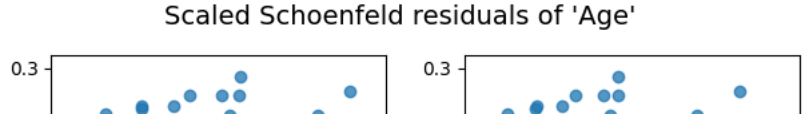
		test_statistic	p	-log2(p)
Age	**km**	0.18	0.67	0.57
	rank	0.09	0.76	0.40
Physical_Activity_Low	**km**	2.54	0.11	3.17
	rank	2.35	0.13	2.99
Physical_Activity_Moderate	**km**	4.64	0.03	5.00
	rank	4.20	0.04	4.63
Sex_Male	**km**	0.01	0.92	0.12
	rank	0.08	0.78	0.36
Smoker	**km**	0.05	0.82	0.28
	rank	0.03	0.86	0.21
Treatment_Standard	**km**	1.43	0.23	2.11
	rank	1.15	0.28	1.82

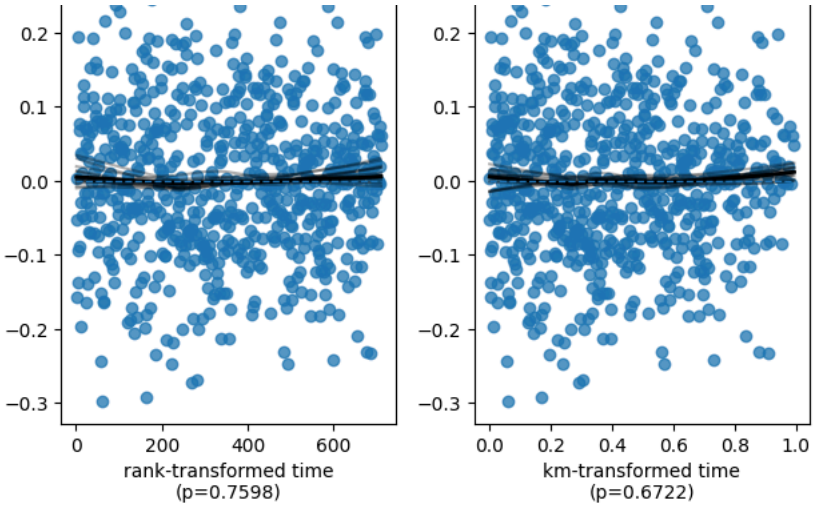
1. Variable 'Physical_Activity_Moderate' failed the non-proportional test: p-value is 0.0313.

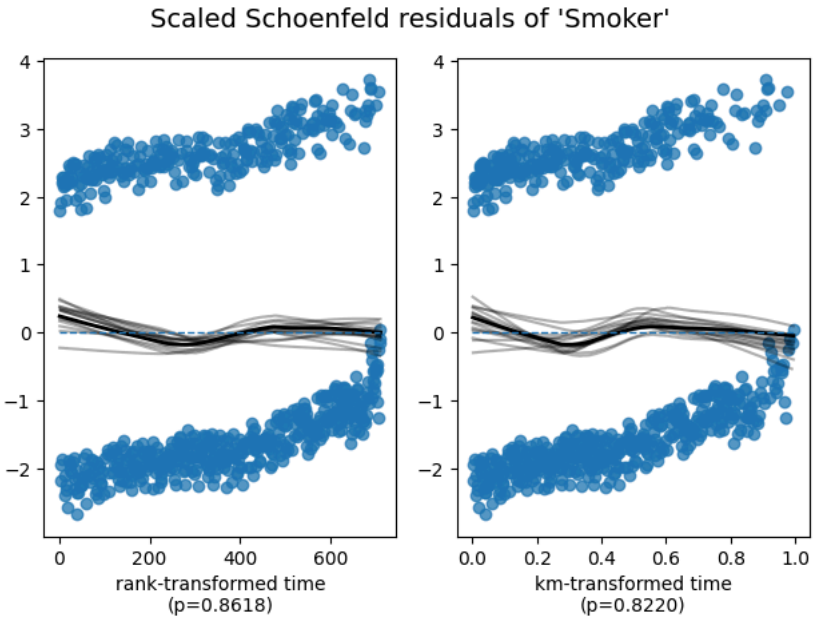
   Advice: with so few unique values (only 2), you can include `strata=['Physical_Activity_Moderate', ...]` in the call in `.fit`. See documentation in link [E] below.

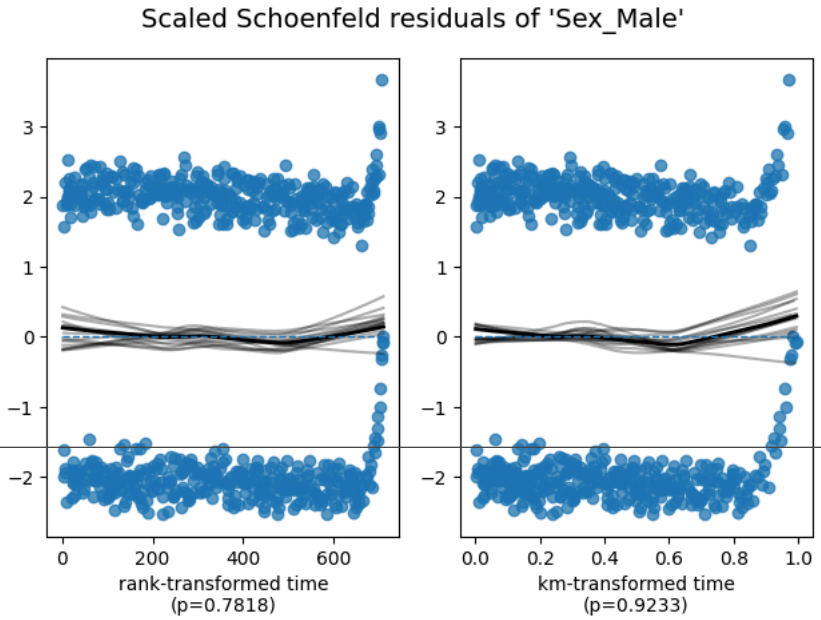
[A] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html
[B] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Bin-variable-and-s
[C] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Introduce-time-var
[D] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Modify-the-functio
[E] https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Stratification

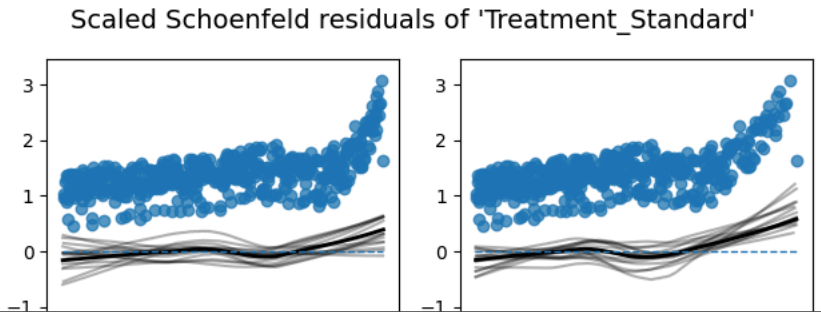
```
[[<Axes: xlabel='rank-transformed time\n(p=0.7598)'>,
  <Axes: xlabel='km-transformed time\n(p=0.6722)'>],
 [<Axes: xlabel='rank-transformed time\n(p=0.8618)'>,
  <Axes: xlabel='km-transformed time\n(p=0.8220)'>],
 [<Axes: xlabel='rank-transformed time\n(p=0.7818)'>,
  <Axes: xlabel='km-transformed time\n(p=0.9233)'>],
 [<Axes: xlabel='rank-transformed time\n(p=0.2825)'>,
  <Axes: xlabel='km-transformed time\n(p=0.2313)'>],
 [<Axes: xlabel='rank-transformed time\n(p=0.1256)'>,
  <Axes: xlabel='km-transformed time\n(p=0.1113)'>],
 [<Axes: xlabel='rank-transformed time\n(p=0.0405)'>,
  <Axes: xlabel='km-transformed time\n(p=0.0313)'>]]
```




rank-transformed time
(p=0.7598)
km-transformed time
(p=0.6722)


Scaled Schoenfeld residuals of 'Smoker'
rank-transformed time
(p=0.8618)
km-transformed time
(p=0.8220)


Scaled Schoenfeld residuals of 'Sex_Male'
rank-transformed time
(p=0.7818)
km-transformed time
(p=0.9233)


Scaled Schoenfeld residuals of 'Treatment_Standard'