REGRESI *COX PROPORTIONAL HAZARD* DAN *PARTIAL LIKELIHOOD* PADA DATA AIDS

Disusun untuk memenuhi tugas Mata Kuliah: Model Survival (B)

Pengampu: Kurnia Susvitasari, S.Si., M.Sc., Ph.D.

Disusun oleh Kelompok 1:

Muhammad Fasya S.	(2206025496)
Nanda Octaviana	(2206032053)
Pandu Adjie Sukarno	(2206026826)
Siti Nur Salamah	(2206048833)
Widya Siti Ropiah	(2206048745)

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

2025

TABEL KONTRIBUSI

No.	Nama Lengkap	NPM	Kontribusi	Persentase
1.	Muhammad Fasya S	2206025496	BAB III Model *Cox*-PH, Interpretasi Model, dan Uji Asumsi Cox-PH	100%
2.	Nanda Octaviana	2206032053	Bab I dan Model *Cox*-PH	100%
3.	Pandu Adjie Sukarno	2206026826	Bab III Kontribusi *Likelihood*	100%
4.	Siti Nur Salamah	2206048833	Bab II *Pre-Processing* dan EDA	100%
5.	Widya Siti Ropiah	2206048745	Bab I Kontribusi *Likelihood* dan Kesimpulan	100%

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1. Latar Belakang

Penyakit HIV/AIDS (*Human Imunodeficiency Virus/Acquired Immunodeficiency Syndrome*) merupakan salah satu masalah kesehatan masyarakat global yang telah berlangsung selama lebih dari empat dekade. HIV menyerang sistem kekebalan tubuh, khususnya sel CD4 atau sel T-helper, yang berperan penting dalam melawan infeksi. Tanpa pengobatan, HIV dapat berkembang menjadi AIDS, suatu kondisi di mana sistem kekebalan sangat lemah dan rentan terhadap berbagai penyakit oportunistik, bahkan dapat menyebabkan kematian. Sejak awal penyebarannya, HIV/AIDS telah menimbulkan dampak besar secara medis, sosial, dan ekonomi di seluruh dunia. Oleh karena itu, pengembangan pengobatan yang efektif serta pemahaman lebih lanjut terhadap karakteristik klinis pasien menjadi sangat penting.

Kemajuan dalam terapi antiretroviral (ART) telah secara signifikan mengubah lanskap pengobatan HIV. ART terbukti memperpanjang harapan hidup pasien sekaligus meningkatkan kualitas hidup mereka secara keseluruhan. Namun demikian, efektivitas terapi ini tidak seragam pada setiap individu. Variabilitas respons pasien terhadap terapi antiretroviral dapat dipengaruhi oleh berbagai faktor klinis dan demografis, seperti usia, jumlah sel CD4 saat inisiasi terapi, gejala klinis yang menyertai, riwayat pengobatan sebelumnya, serta kondisi fungsional pasien yang dapat diukur melalui skor Karnofsky.

Untuk memahami peran dari berbagai faktor tersebut terhadap risiko kematian pasien HIV/AIDS, diperlukan pendekatan statistik yang tepat, salah satunya adalah model *Cox Proportional Hazards* (*Cox*-PH). Model ini merupakan metode regresi survival semi-parametrik yang mampu mengevaluasi pengaruh variabel-variabel prediktor terhadap waktu bertahan hidup, termasuk pada data dengan kejadian kematian yang tidak seluruhnya teramati (*censored*). Model *Cox*-PH menjadi sangat relevan dalam konteks ini karena mampu memberikan estimasi risiko kematian berdasarkan karakteristik pasien secara individual.

Data yang digunakan dalam penelitian ini bersumber dari *AIDS Clinical Trial Group Study 175*, sebuah studi klinis besar yang mengevaluasi efektivitas berbagai rejimen antiretroviral pada pasien HIV yang belum pernah menjalani terapi sebelumnya. Dataset mencakup sejumlah variabel penting seperti usia, status gejala (*symptom*), skor Karnofsky (karnof), jenis terapi (*drugs*), jumlah CD4 (z30 dan cd820), serta variabel stratifikasi lainnya.

Penelitian ini memiliki relevansi yang tinggi karena HIV/AIDS masih menjadi penyakit dengan tingkat mortalitas yang signifikan di banyak negara berkembang, termasuk Indonesia. Meskipun pengobatan semakin membaik, pemahaman tentang faktor-faktor klinis yang memengaruhi keberhasilan terapi tetap menjadi kebutuhan yang mendesak. Melalui penelitian ini, analisis survival menggunakan model Cox-PH akan diterapkan terhadap data *AIDS Clinical Trial Group Study 175* untuk mengidentifikasi variabel-variabel yang secara signifikan berpengaruh terhadap lama bertahan hidup pasien. Dengan mengetahui faktor-faktor risiko tersebut, diharapkan dapat memberikan kontribusi nyata dalam pengembangan strategi pengobatan HIV yang lebih personal, berbasis data, dan efisien.

1.2. Rumusan Masalah

1. Bagaimana pengaruh karakteristik klinis seperti usia, penggunaan obat (drugs), skor Karnofsky, jumlah sel CD4 pada hari ke-30 (z30), riwayat terapi sebelumnya (preanti), dan gejala klinis (symptom) terhadap risiko kematian pada pasien HIV/AIDS?
2. Bagaimana bentuk model *Cox*-PH terbaik?
3. Apakah terdapat pengaruh yang signifikan dari masing-masing variabel tersebut terhadap waktu bertahan hidup pasien HIV/AIDS berdasarkan model Cox Proportional Hazard?
4. Apakah asumsi proportional hazard dipenuhi dalam model ini berdasarkan uji validitas (cox.zph)?

1.3. Tujuan Penelitian

1. Menganalisis sejauh mana karakteristik klinis seperti usia, penggunaan obat, skor Karnofsky, z30, riwayat terapi, dan gejala klinis memengaruhi risiko kematian pada pasien HIV/AIDS.
2. Membuat model *Cox*-PH terbaik.
3. Menentukan variabel-variabel yang berpengaruh secara signifikan terhadap waktu bertahan hidup pasien HIV/AIDS menggunakan pendekatan model Cox Proportional Hazards.
4. Mengevaluasi pemenuhan asumsi proportional hazard dari model Cox melalui uji *cox.zph* guna memastikan validitas model dalam menggambarkan hubungan antara faktor klinis dan waktu bertahan hidup.

1.4. Manfaat Penelitian

1. Bagi Mahasiswa

   Penelitian ini memberikan manfaat edukatif bagi mahasiswa, terutama yang sedang menempuh studi di bidang statistika, kesehatan masyarakat, epidemiologi, atau bioinformatika. Dengan terlibat dalam analisis data klinis nyata seperti dataset *AIDS Clinical Trial Group Study 175*, mahasiswa dapat memahami bagaimana metode statistik seperti model Cox Proportional Hazard diterapkan dalam konteks dunia nyata untuk menilai faktor-faktor yang memengaruhi kelangsungan hidup pasien. Pengalaman ini juga mengembangkan keterampilan teknis dalam analisis survival, interpretasi hasil, serta penggunaan perangkat lunak statistik seperti R yang sangat relevan dalam penelitian dan karier profesional di bidang sains data kesehatan.
2. Bagi Kelompok Penulis

   Bagi kelompok penulis, penelitian ini menjadi wadah pengembangan kompetensi dalam melakukan analisis data kompleks berbasis medis serta menerjemahkannya ke dalam informasi yang berguna untuk pengambilan keputusan klinis. Proyek ini juga memperkuat kemampuan kerja sama tim dalam membagi peran, mengevaluasi data, serta menyusun laporan ilmiah yang sistematis. Selain itu, kegiatan ini memperkaya wawasan penulis terhadap dinamika klinis yang berkaitan dengan terapi dan kondisi pasien HIV/AIDS.
3. Bagi Pembaca

Penelitian ini memberikan pemahaman yang lebih dalam bagi pembaca umum maupun akademisi tentang bagaimana karakteristik pasien dan jenis pengobatan dapat memengaruhi prognosis dalam pengelolaan penyakit HIV/AIDS. Dengan bahasa yang mudah dimengerti, pembaca dapat memperoleh wawasan mengenai pentingnya analisis statistik dalam mengevaluasi efektivitas terapi serta faktor risiko yang berdampak terhadap kelangsungan hidup pasien. Pengetahuan ini dapat menjadi referensi berharga dalam mendukung edukasi kesehatan, penelitian lanjutan, ataupun pertimbangan dalam penyusunan kebijakan kesehatan berbasis data.

1.5. Batasan Masalah

1. Penelitian ini hanya menggunakan data dari *AIDS Clinical Trial Group Study 175,* yang merupakan uji klinis terhadap pasien HIV yang belum pernah menerima terapi antiretroviral sebelumnya. Oleh karena itu, hasil penelitian tidak dapat digeneralisasikan untuk pasien HIV yang sudah menjalani terapi sebelumnya atau memiliki komorbiditas lain.
2. Metode analisis yang digunakan adalah model Cox Proportional Hazards (Cox-PH). Asumsi proportional hazard telah diuji dan dipenuhi, namun model ini tidak mempertimbangkan kemungkinan interaksi antar variabel ataupun efek waktu (*time-dependent covariates*).
3. Penelitian ini hanya membahas hubungan antara variabel-variabel prediktor dan risiko kematian, tanpa mengeksplorasi outcome lain seperti waktu sampai infeksi oportunistik, perubahan kadar viral load, atau kualitas hidup.
4. Analisis survival dilakukan dengan asumsi bahwa data yang *censored* bersifat non-informatif, yaitu tidak ada perbedaan sistematis antara pasien yang *censored* dan yang mengalami kejadian (kematian), sebagaimana asumsi dasar dari model Cox.

BAB II
TINJAUAN PUSTAKA

2.1 Kaplan Meier

Metode Kaplan-Meier merupakan Teknik non-parametrik yang digunakan untuk memperkirakan fungsi survival, yaitu Probabilitas pasien bertahan hidup hingga waktu tertentu. Metode ini cocok digunakan pada data *censored*, yaitu data di mana kejadian (misalnya kematian) belum terjadi sampai akhir observasi. Fungsi survival Kaplan-Meier dihitung dengan rumus:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

dimana $\hat{S}(t)$ adalah peluang survival sampai waktu t, d_i adalah jumlah kejadian pada waktu t_i, dan n_i adalah jumlah individu yang masih dalam risiko pada waktu tersebut (Kleinbaum & Klein, 2010).

Dalam konteks HIV/AIDS, metode ini digunakan untuk membandingkan probabilitas bertahan hidup pasien berdasarkan jenis pengobatan, skor karnofsky, atau gejala klinis. Kaplan-Meier sering dikombinasikan dengan uji log-rank untuk menguji signifikansi perbedaan antar-kelompok.

Metode ini merupakan langkah awal penting sebelum analisis multivariat, seperti regresi Cox, karena memberikan gambaran visual tentang perbedaan survival antar-kelompok pasien

2.2. Uji K-Sampel

Uji K Sampel dilakukan untuk melihat apakah ada perbedaan *survival experience* dari K kelompok, yang berasal dari K populasi. Contohnya, apakah pasien kanker stadium yang berbeda (1, 2, 3, 4) memang memiliki *survival experience* yang berbeda? Dalam hal ini, karena membandingkan 4 kelompok pasien berdasarkan stadium penyakitnya, maka uji ini adalah uji 4 sampel. Berikut adalah prosedur pengujian hipotesis pada K sampel. Misalkan, K sampel masing-masing berukuran $n_1, n_2, \ldots, n_K$ diambil dari K populasi.

2.2.1 Hipotesis Uji K-Sampel

Hipotesis dari pengujian ini adalah sebagai berikut.

H_0: $H_1(t) = h_2(t) = \cdots = h_K(t)$ untuk semua $t \leq \tau$

H_1: tidak demikian

Hipotesis tersebut menyatakan bahwa fungsi hazard untuk setiap sampel pada setiap titik waktu dalam rentang pengamatan $(t \leq \tau)$ adalah sama, yang mengimplikasikan tidak ada perbedaan *survival experience* dari seluruh K populasi di mana sampel tersebut diambil. Sementara, jika ada satu atau beberapa titik waktu di mana fungsi hazard dari minimal 2 sampel adalah berbeda, maka hal tersebut sudah menyatakan dukungan terhadap hipotesis alternatif.

2.2.2 Statistik Uji

$$\chi^2_{hitung} = \big(Z_1(\tau), Z_2(\tau), \dots, Z_{K-1}(\tau)\big)\Sigma^{-1}\big(Z_1(\tau), Z_2(\tau), \dots, Z_{K-1}(\tau)\big)^t \sim \chi^2_{K-1}$$

Di mana Σ adalah matriks varian-kovarian dari $Z_1(\tau), Z_2(\tau), \dots, Z_{K-1}(\tau)$ yang berukuran $(K-1) \times (K-1)$ dengan entri-entri $\hat{\sigma}_{jj}$ dan $\hat{\sigma}_{jg}$, yaitu

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i)\left[d_{ij} - Y_{ij}\frac{d_i}{Y_i}\right], j = 1,2, \dots, K$$

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i\,, j = 1,2, \dots, K$$

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i\,; j \neq g;\ j, g = 1,2, \dots, K$$

Pada kasus ketika $K = 2$, maka statistik uji nya adalah

$$Z_{hitung} = \frac{\sum_{i=1}^{D} W(t_i)\left[d_i - Y_{i1}\frac{d_i}{Y_i}\right]}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{i1}}{Y_i}\left(1 - \frac{Y_{i1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i}} \sim N(0,1)$$

Saat n besar dibawah H_0 benar.

Apabila $W(t_i) = 1, \forall t_i$, maka uji yang dilakukan dikenal dengan nama **uji log-rank.**

2.2.3 Aturan Keputusan

Dengan taraf signifikansi α, apabila $\chi^2_{hitung} > \chi^2_{(K-1),1-\frac{\alpha}{2}}$ atau $\chi^2_{hitung} < \chi^2_{(K-1),\frac{\alpha}{2}}$, maka H_0 ditolak. Artinya paling tidak ada satu atau beberapa sampel yang memiliki perbedaan *survival experience* pada beberapa titik waktu.

Untuk $K = 2$, aturan keputusan ini ekivalen dengan jika $|Z_{hitung}| > Z_{\frac{\alpha}{2}}$ maka H_0 ditolak. Hal ini berarti kedua populasi memiliki *survival experience* yang berbeda signifikan secara statistik pada beberapa titik waktu.

2.3. Cox Proportional Hazard

Model Cox Proportional Hazard (Cox-PH) adalah metode statistik yang umum digunakan dalam analisis survival untuk menilai pengaruh beberapa faktor terhadap waktu terjadinya suatu peristiwa, seperti kematian, kegagalan mesin, atau kekambuhan penyakit. Model ini termasuk model semi-parametrik karena tidak mengharuskan bentuk tertentu untuk fungsi dasar hazard, tetapi tetap memodelkan pengaruh variabel secara parametrik. Fungsi hazard pada waktu *t* untuk individu dengan kovariat $\boldsymbol{x}$ dinyatakan sebagai:

$$h(t, \boldsymbol{x}) = h_0(t)\exp\left(\beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p\right)$$

di mana $h_0(t)$ adalah *baseline hazard* dan $\beta_1, \beta_2, \dots, \beta_p$ adalah koefisien regresi yang menggambarkan kontribusi masing – masing variabel penjelas $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p$. Asumsi utama

model ini adalah *proportional hazard*, yang berarti rasio hazard antara dua individu tetap konstan sepanjang waktu, tergantung hanya pada perbedaan nilai kovariat, bukan pada waktu itu sendiri.

Model ini dapat digunakan untuk variabel kategorik maupun numerik, serta memungkinkan adanya interaksi antar-variabel. Interpretasi koefisien β umunya dilakukan melalui *hazard ratio*, yaitu $\exp(\beta)$, yang menunjukkan seberapa besar perubahan risiko akibat perubahan satu unit pada variabel tertentu. Untuk mengecek apakah asumsi *proportional hazard* terpenuhi, dapat digunakan metode visual seperti grafik log(-log(survival)) atau uji statistik menggunakan *Schoenfeld residuals*. Estimasi parameter dalam model Cox-PH dilakukan dengan metode *partial likelihood*, yaitu:

$$L(\beta) = \prod_{j=1}^{D} \left(\frac{\exp(\beta^t \boldsymbol{x}_j)}{\sum_{l \epsilon R_j} \exp(\beta^t \boldsymbol{x}_l)} \right)$$

dengan $\boldsymbol{x}_j$ merupakan vektor variabel penjelas untuk objek yang mengalami events di waktu t_j, dan $\boldsymbol{x}_l$ merupakan vektor variabel penjelas untuk objek yang beresiko mengalami events di waktu t_j. Dengan demikian, perhitungan dengan rumus di atas hanya memperhitungkan individu yang mengalami peristiwa pada waktu tertentu, dan membandingkannya dengan seluruh individu yang masih berisiko saat itu. Jika terdapat kejadian yang terjadi pada waktu yang sama (disebut data ties), maka pendekatan seperti metode Breslow atau Efron digunakan untuk menyesuaikan perhitungan likelihood tersebut.

2.4. Kontribusi Partial Likelihood

2.4.1 Data Tanpa Ties

Misalkan terdapat D *event* dan tidak ada *ties*, maka terdapat D titik waktu terjadinya *event*, yakni $t_1 < t_2 <_{\dots} < t_D$. Maka, bentuk *likelihood* dari data ini adalah sebagai berikut.

$$L(\beta) = \prod_{j=1}^{D} \left\{ \frac{e^{\beta \boldsymbol{x}_j}}{\sum_{l \epsilon R_j} e^{\beta \boldsymbol{x}_l}} \right\}$$

Dengan $\boldsymbol{x}_j$ merupakan vektor variabel penjelas untuk objek yang mengalami *event* di waktu t_j, dan $\boldsymbol{x}_l$ merupakan vektor variabel penjelas untuk objek yang berisiko mengalami *event* di waktu t_j.

2.4.2 Data Ties

2.4.2.1 Metode Breslow

Fungsi *likelihood* dari model Cox-PH $h(t, \boldsymbol{x}) = h_0(t) e^{\beta' \boldsymbol{x}}$ di mana terjadi sebanyak D *events* pada rentang waktu pengamatan adalah sebagai berikiut.

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' \boldsymbol{Z}_i)}{\left[\sum_{l \epsilon R_i} \exp(\beta' \boldsymbol{x}_l)^{d_i} \right]}$$

Dengan $Z_i = \sum_{i \epsilon D} \boldsymbol{x}_i$, D_i adalah himpunan objek yang mengalami *event* di waktu t_i. Jika pada waktu t_i terjadi d_i *event*, kejadian-kejadian tersebut diasumsikan terjadi satu per satu. Saat salah satu objek dari objek mengalami *event*, maka $d_i - 1$ objek yang lain

diasumsikan belum mengalami *event*. Sehingga saat perhitungan fungsi hazard pada himpunan risiko, semua objek yang berada dalam himpunan risiko akan diperhitungkan.

2.4.2.2 Metode Efron

Jika pada metode *breslow* $d_i - 1$ objek yang lain diasumsikan belum mengalami *event* saat satu objek dari di objek tersebut mengalami *event,* tidak demikian pada metode *Efron*. Ada faktor koreksi yang ditambahkan. Pada objek yang pertama kali mengalami *event* (dari d_i objek di t_i), pembagi pada kontribusi *likelihood* adalah total hazard dari seluruh anggota R_i. Namun, ketika *event* kedua terjadi, maka faktor pembagi dikoreksi dengan mengurangkan hazard dari objek yang pertama mengalami *event*.

Tetapi, bagaimanakah urutan kejadian dari d_i *event* etrsebut? Tidak diketahui. Sehingga, untuk faktor koreksi cukup dikurangi dengan rata-rata terboboti dari total hazard di objek tersebut, misal $h^* = \sum_{i \in D} h(t_i)$. Jadi, saat *event* kedua terjadi, total hazard pada komponen pembagi dikurangi dengan $\frac{1}{d_i} h^*$. Saat *event* ketiga terjadi, total hazard pada komponen pembagi dikurangi dengan $\frac{2}{d_i} h^*$, dan seterusnya. Dengan demikian diperoleh hasil berikut.

$$\prod_{i=1}^{D} \frac{\exp\left(\beta' \boldsymbol{Z}_i\right)}{\prod_{j=1}^{d_i} \left[\sum_{l \in R_i} \exp\left(\left(\beta' \boldsymbol{x}_l\right)\right) - \frac{j-1}{d_i} \sum_{k \in D_i} \exp(\beta' \boldsymbol{x}_i)\right]^{d_i}}$$

2.4.2.3 Metode Diskrit

Fungsi *likelihood* dari model Cox-PH $h(t, \boldsymbol{x}) = h_0(t) e^{\beta' \boldsymbol{x}}$ di mana terjadi sebanyak D *events* pada rentang waktu pengamatan adalah sebagai berikiut.

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta' \boldsymbol{Z}_i\right)}{\left[\sum_{q \in Q_i} \exp\left(\beta' \boldsymbol{Z}_q^*\right) \right]^{d_i}}$$

Dengan $Z_i = \sum_{i \in D} \boldsymbol{x}_i$, D_i adalah himpunan objek yang mengalami *event* di waktu t_i dan $\boldsymbol{Z}^* = \sum_{j \in D_i} X_{qj}$.

BAB III
HASIL DAN PEMBAHASAN

3.1. Sumber Data

Dataset yang digunakan adalah *AIDS Clinical Trials Group Study 175* yakni data uji klinis yang dipublikasikan pada tahun 1996 dan berisi informasi medis serta data kategorikal dari 2139 pasien AIDS. Tujuan utama pengumpulan data ini adalah untuk mengevaluasi efektivitas dan keamanan berbagai jenis pengobatan AIDS, yaitu zidovudine (ZDV), didanosine (ddI), kombinasi ZDV+ddI, dan ZDV+zalcitabine (ddC), pada pasien dengan jumlah CD4 antara 200–500 sel/mm³. Berikut detail sumber data yang digunakan:

- Sumber data: AIDS Clinical Trials Group Study 175 Dataset
- Jumlah data: 2139 observasi dan 24 variabel

Tabel 1. *Informasi Dataset*

No	Nama Variabel	Tipe	Keterangan
1.	time	Numerik	Waktu hingga kegagalan atau penyensoran (Hari).
2.	trt	Kategorik	Indikator *treatment*: 0=ZDV saja 1=ZDV+ddI 2=ZDV+Zal 3=ddI saja
3.	age	Numerik	Usia saat *baseline* (Tahun)
4.	wtkg	Numerik	Berat badan saat *baseline* (Kg)
5.	hemo	Kategorik	Status hemofilia: 0=Tidak 1=Ya
6.	homo	Kategorik	Orientasi seksual,aktivitas homoseksual: 0=Tidak 1=Ya
7.	drugs	Kategorik	Riwayat penggunaan narkoba tipe IV: 0=Tidak 1=Ya
8.	karnof	Numerik	Skor Karnofsky (status fungsional) (skala 0-100)
9.	oprior	Kategorik	Terapi antiretroviral non-ZDV sebelum studi: 0=Tidak 1=Ya
10.	z30	Kategorik	Penggunaan ZDV 30 hari sebelum studi: 0=Tidak 1=Ya
11.	zprior	Kategorik	Penggunaan ZDV sebelum studi: 0=Tidak 1=Ya

12.	preanti	Numerik	Jumlah hari terapi antiretroviral sebelum studi (Hari)
13.	race	Kategorik	Ras: 0=Kulit putih 1=Non-kulit putih
14.	gender	Kategorik	Jenis kelamin: 0=Perempuan 1=Laki-laki
15.	str2	Kategorik	Riwayat antiretroviral: 0=Naif 1=Berpengalaman
16.	strat	Kategorik	Stratifikasi riwayat antiretroviral: 1= Naif antiretroviral 2= >1 tapi ≤ 52 minggu terapi sebelumnya 3= >52 minggu
17.	symptom	Kategorik	Indikator simptomatik: 0=Asimptomatik 1=Simptomatik
18.	treat	Kategorik	Indikator perlakuan/terapi: 0=ZDV saja 1=Lainnya
19.	offtrt	Kategorik	Indikator berhenti terapi sebelum 96±5 minggu: 0=Tidak 1=Ya
20.	cd40	Numerik	Jumlah CD4 saat baseline (Sel/mm³)
21.	cd420	Numerik	Jumlah CD4 pada 20±5 minggu (Sel/mm³)
22.	cd80	Numerik	Jumlah CD8 saat *baseline* (Sel/mm³)
23.	cd820	Numerik	Jumlah CD8 pada 20±5 minggu (Sel/mm³)
24.	Label	Kategorik	Indikator penyensoran: 0=Censored 1=Event (Death)

Dalam penelitian ini, seluruh variabel yang tersedia akan disertakan untuk dianalisis lebih lanjut pada tahap-tahap berikutnya, seperti preprocessing, *Exploratory Data Analysis* (EDA), dan tahapan analisis lanjutan lainnya.

3.2. Preprocessing

Preprocessing merupakan tahap awal yang sangat penting dalam analisis data, di mana data mentah dipersiapkan agar siap untuk dianalisis lebih lanjut. Proses ini mencakup berbagai langkah seperti pembersihan data, transformasi, dan penganganan nilai-nilai yang tidak sesuai. Penjelasan lebih detail mengenai tahapan-tahapan tersebut akan dibahas pada subbab selanjutnya.

3.2.1 Tipe Data

Berikut ini merupakan hasil tipe data yang dibaca dari setiap variabel:

```
Data columns (total 24 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   time     2139 non-null   int64
 1   trt      2139 non-null   int64
 2   age      2139 non-null   int64
 3   wtkg     2139 non-null   float64
 4   hemo     2139 non-null   int64
 5   homo     2139 non-null   int64
 6   drugs    2139 non-null   int64
 7   karnof   2139 non-null   int64
 8   oprior   2139 non-null   int64
 9   z30      2139 non-null   int64
 10  zprior   2139 non-null   int64
 11  preanti  2139 non-null   int64
 12  race     2139 non-null   int64
 13  gender   2139 non-null   int64
 14  str2     2139 non-null   int64
 15  strat    2139 non-null   int64
 16  symptom  2139 non-null   int64
 17  treat    2139 non-null   int64
 18  offtrt   2139 non-null   int64
 19  cd40     2139 non-null   int64
 20  cd420    2139 non-null   int64
 21  cd80     2139 non-null   int64
 22  cd820    2139 non-null   int64
 23  label    2139 non-null   int64
dtypes: float64(1), int64(23)
memory usage: 401.2 KB
None
```

Gambar 1. *Tipe Data*

Berdasarkan hasil tipe data, semuanya sudah sesuai sehingga tidak perlu adanya perubahan yang dilakukan pada tipe data setiap variabel.

3.2.2 Missing Value

Missing Value adalah nilai kosong yang ada pada data. Dalam langkah awal *preprocessing* data, kami melakukan pengecekkan apakah ada data nilai kosong atau tidak. Setelah dilakukan pengecekan, didapat bahwa tidak ada *missing value* di semua kolom data kami. Berikut tampilan hasil pengecekan *missing value*:

```
time  trt  age  wtkg  hemo  homo  drugs  karnof  oprior  z30  zprior  \
   0    0    0     0     0     0      0       0       0    0       0

preanti  race  gender  str2  strat  symptom  treat  offtrt  cd40  cd420  \
      0     0       0     0      0        0      0       0     0      0

cd80  cd820  label
   0      0      0
```

Gambar 2. *Hasil Pengecekan Missing Value*

3.2.3 Cek Baris Duplikat

Saat melakukan pengecekan baris duplikat pada data, kami mendapatkan bahwa tidak terdapat data duplikat pada data. Berikut hasilnya:

```
Jumlah baris duplikat: 0
```

Gambar 3. *Hasil Baris Duplikat*

3.2.4 Kategorisasi Variabel

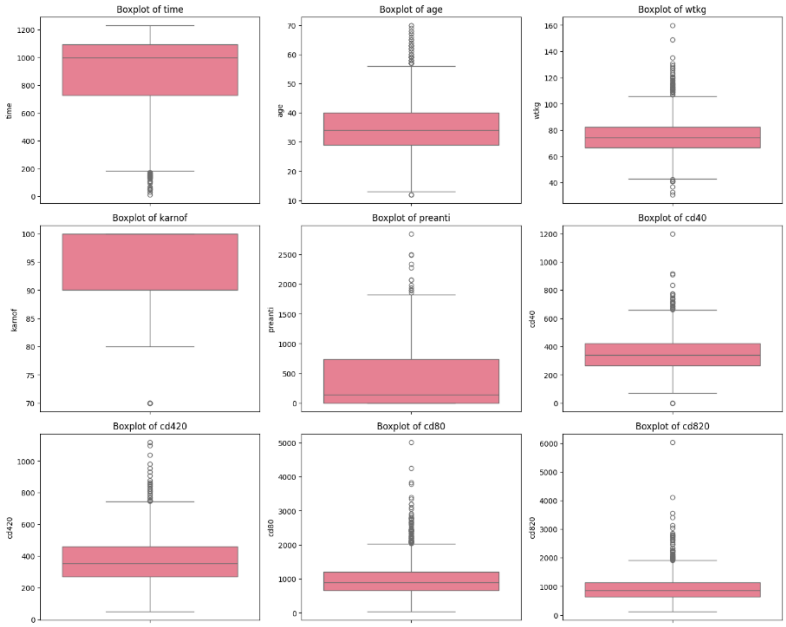
Pada tahap ini, kami akan mengelompokkan variabel ke masing-masing kategori yaitu variabel kategorik dan kontinu untuk analisis lanjutan. Berikut hasilnya:

Tabel 2. *Kategorisasi Variabel*

Variabel Kategorik	['trt', 'hemo', 'homo', 'drugs', 'oprior', 'z30', 'zprior', 'race', 'gender', 'str2', 'strat', 'symptom', 'treat', 'offtrt', 'label']. **Total: 15**
Variabel Kontinu	['time', 'age', 'wtkg', 'karnof', 'preanti', 'cd40', 'cd420', 'cd80', 'cd820']. **Total: 9**

3.2.5 Deteksi *Outlier* Variabel Kontinu

Pada tahap ini akan dilihat keberadaan *outlier* untuk data numerik menggunakan boxplot. Berikut hasilnya:



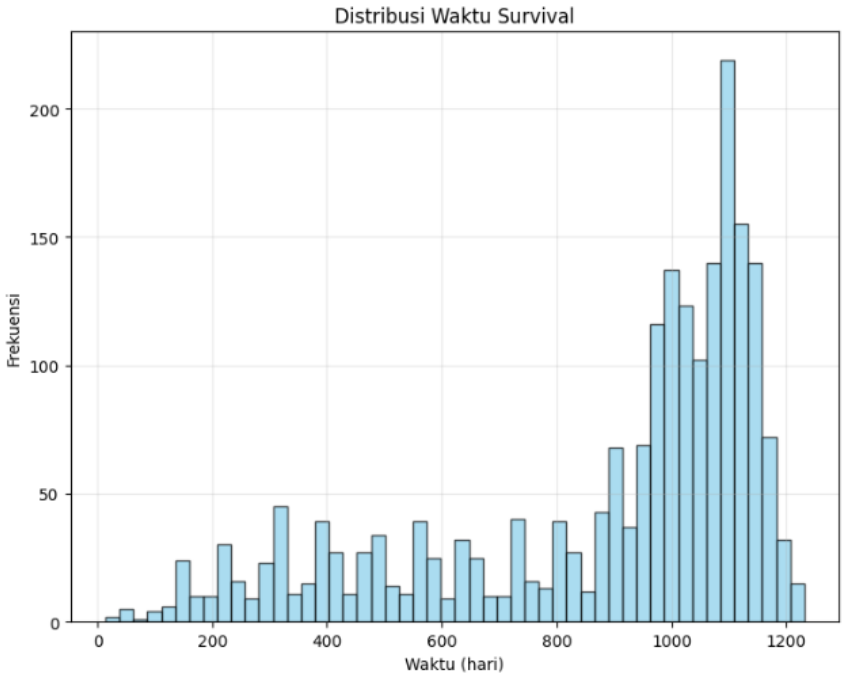
Gambar 4. *Outlier Variabel Kontinu Menggunakan Visualisasi Boxplot*

Dari hasil visualisasi boxplot, terlihat bahwa hampir seluruh variabel kontinu seperti time, age, wtkg, preanti, cd40, cd420, cd80, dan cd820 menunjukkan keberadaan *outlier* yang ditandai dengan titik-titik di luar batas bawah dan atas boxplot. Misalnya, variabel preanti dan cd820 memiliki *outlier* yang cukup ekstrem. Keberadaan *outlier* ini mencerminkan variasi alami dari kondisi pasien dan dapat mengandung informasi penting yang berkaitan dengan kondisi klinis serta waktu survival mereka. Oleh karena itu, dalam penelitian ini, *outlier* tidak akan dihapus ataupun diubah karena mengeliminasi data ekstrem tersebut dapat menyebabkan hilangnya informasi penting. Selain itu, mempertahankan *outlier* juga membantu menjaga representasi yang lebih realistis terhadap distribusi data,

khususnya dalam konteks data klinis yang memang memiliki keragaman tinggi antarindividu.

3.3. Exploratory Data Analysis (EDA)

3.3.1 Distribusi Waktu Survival



Gambar 5. *Visualisasi Distribusi Waktu Survival*

Gambar di atas menunjukkan distribusi waktu survival (dalam satuan hari) dari seluruh partisipan dalam studi klinis. Secara umum, distribusi waktu survival tampak positif *skewed* (condong ke kanan) yang berarti sebagian besar peserta memiliki waktu survival yang relatif lama.

Distribusi menunjukkan bahwa frekuensi waktu survival meningkat secara bertahap, dengan puncaknya berada sekitar 1000–1150 hari. Hal ini mengindikasikan bahwa banyak peserta bertahan hidup hingga lebih dari 2,5 tahun setelah awal studi. Pola distribusi ini mencerminkan bahwa sebagian besar peserta dalam studi memiliki durasi hidup yang cukup panjang, yang bisa menjadi indikasi efektivitas pengobatan atau karakteristik populasi yang cukup stabil secara klinis.

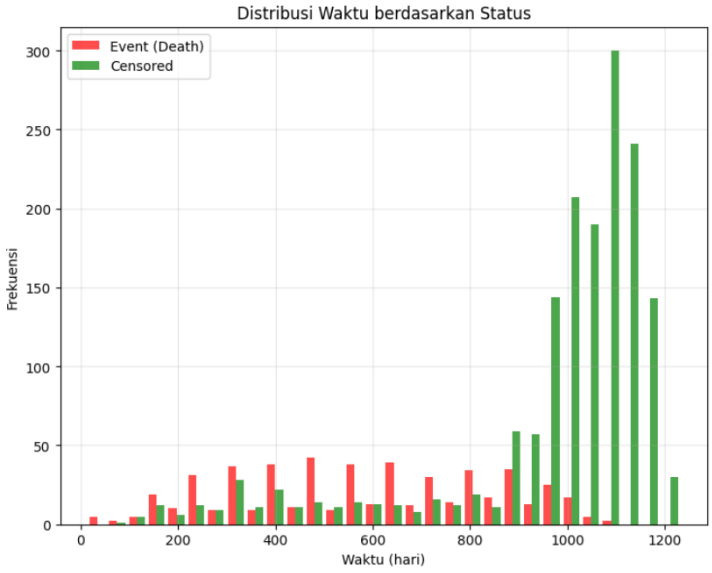
3.3.2 Statistik Deskriptif Waktu Survival

Tabel 3. *Statistik Deskriptif Waktu Survival*

Mean	879.10 hari
Median	997.00 hari
Min	14 hari
Max	1231 hari
Standard Deviation	292.27 hari

Statistik deskriptif waktu survival menunjukkan bahwa rata-rata waktu survival peserta adalah sekitar 879 hari dengan nilai median sebesar 997 hari yang berarti separuh partisipan bertahan hidup setidaknya selama hampir 2,5 tahun. Nilai minimum sebesar 14 hari dan maksimum 1231 hari mengindikasikan adanya rentang waktu survival yang cukup luas, dari kurang dari satu bulan hingga lebih dari tiga tahun. Standar deviasi sebesar 292 hari mencerminkan adanya variasi yang cukup tinggi dalam waktu bertahan antar partisipan.

3.3.3 Distribusi Waktu Survival berdasarkan Status

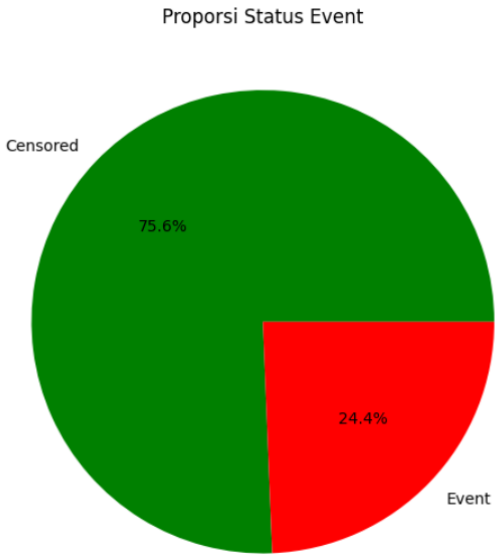


Gambar 6. *Visualisasi Distribusi Waktu berdasarkan Status*

Visualisasi di atas menunjukkan distribusi waktu survival berdasarkan status akhir partisipan, yaitu apakah individu mengalami kejadian (*event/death*) atau tersensor (*censored*). Warna merah mewakili partisipan yang mengalami *event* (kematian), sedangkan warna hijau mewakili partisipan yang tersensor (misalnya karena keluar studi, hilang *follow-up*, atau studi berakhir sebelum terjadi *event*).

Terlihat bahwa sebagian besar kejadian (*event*) terjadi pada awal hingga pertengahan waktu observasi, khususnya pada rentang 200 hingga 900 hari, dengan intensitas relatif merata. Sebaliknya, frekuensi data yang tersensor meningkat tajam setelah hari ke-900 dan mencapai puncaknya di sekitar 1000 hingga 1150 hari. Ini mengindikasikan bahwa banyak partisipan masih hidup atau belum mengalami *event* saat studi berakhir sehingga data mereka dikategorikan sebagai tersensor.

3.3.4 Proporsi Data Tersensor dan Tidak Tersensor

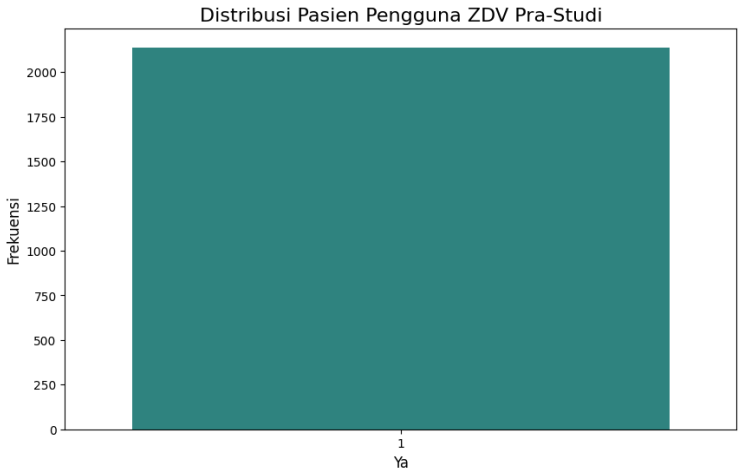


Gambar 7. *Proporsi Data Tersensor dan Tidak Tersensor*

Berdasarkan diagram lingkaran di atas, dapat diketahui bahwa dari seluruh partisipan dalam studi, sebanyak 75,6% (1618 observasi) data merupakan data tersensor, sedangkan hanya 24,4% (521 observasi) data yang mengalami kejadian (*event*). Ini menunjukkan bahwa mayoritas partisipan tidak mengalami kematian atau kejadian selama periode observasi berlangsung. Dalam konteks analisis survival, hal ini merupakan kondisi yang umum, terutama pada studi klinis jangka panjang, di mana tidak semua peserta mengalami *outcome* yang diamati. Tingginya proporsi data tersensor memberikan informasi bahwa sebagian besar estimasi fungsi survival akan sangat dipengaruhi oleh data yang belum mengalami *event*.

3.3.5 Penghapusan Variabel

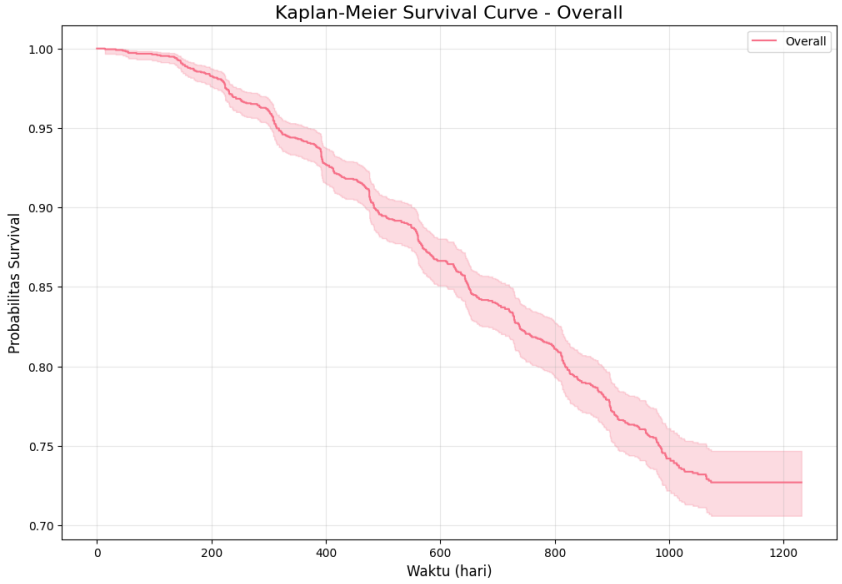
Pada data, kolom ‘zprior’ menunjukkan kondisi penggunaan ZDV pasien sebelum studi dilakukan (1 jika ya, 0 jika tidak).



Gambar 8. *Kelompok Pasien yang menggunakan ZDV pra-studi.*

Berdasarkan plot tersebut, seluruh pasien dalam studi menggunakan ZDV. Maka dari itu, variabel ini tidak memberikan kontribusi berarti dalam model nanti. Karenanya variabel ini tidak digunakan dalam pembuatan model nanti.

3.3.6 Kurva Survival Kaplan-Meier Survival – *Overall*

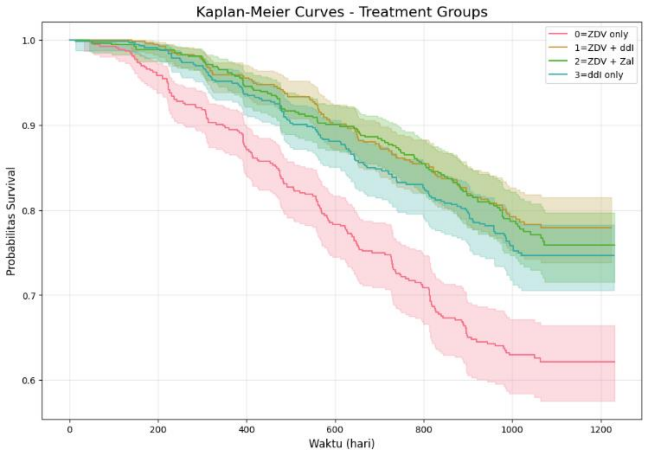


Gambar 9. *Kurva Survival Kaplan-Meier - Overall*

Kurva Kaplan-Meier di atas menyajikan estimasi probabilitas survival secara keseluruhan sepanjang periode observasi. Pada awal waktu (hari ke-0), probabilitas survival dimulai dari 1 (atau 100%), yang secara bertahap menurun seiring bertambahnya waktu. Penurunan kurva menunjukkan terjadinya kejadian (kematian) pada peserta studi. Dari grafik terlihat bahwa penurunan kurva terjadi secara konsisten, dengan probabilitas survival akhir berada di kisaran 0.75 setelah lebih dari 1200 hari, yang berarti sekitar 75% peserta diperkirakan masih bertahan hidup pada akhir masa observasi.

Area berwarna merah muda di sekitar kurva menunjukkan *confidence interval* 95% yang menggambarkan tingkat ketidakpastian dari estimasi probabilitas survival. Semakin lebar area ini, semakin besar ketidakpastian terhadap estimasi survival di waktu tersebut. Pada grafik ini, interval cenderung melebar pada waktu-waktu akhir yang merupakan indikasi semakin sedikitnya partisipan yang tersisa (*at risk*) sehingga estimasi survival menjadi kurang stabil.

3.3.7 Kurva Survival Kaplan-Meier – Treatment



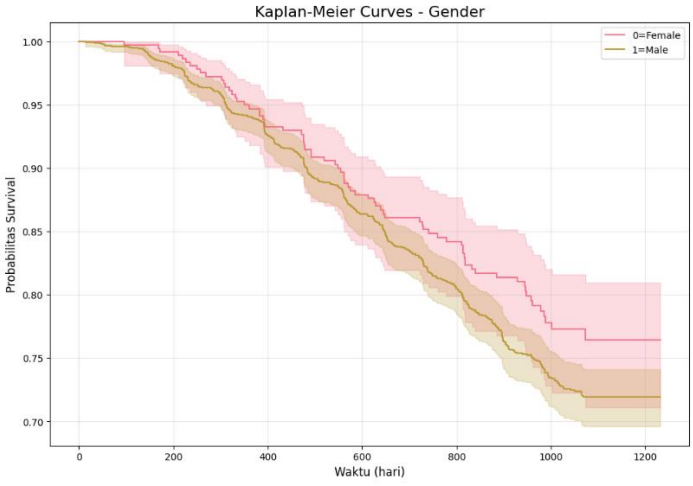
Gambar 10. *Kurva Survival Kaplan-Meier – Treatment*

```
Log-rank k sampel p-value: 0.000000
✓ Terdapat perbedaan survival yang signifikan antar kelompok (p < 0.05)
```

Gambar 11*. Uji K-Sampel Pada Kelompok Treatment*

Berdasarkan kurva Kaplan-Meier untuk empat kelompok perlakuan, terlihat bahwa probabilitas survival berbeda antar kelompok. Kelompok yang hanya menerima ZDV (kode 0) menunjukkan penurunan probabilitas survival yang paling tajam dibandingkan kelompok lainnya, menandakan efektivitas pengobatan yang lebih rendah. Sebaliknya, kelompok dengan kombinasi ZDV+ddl, ZDV+zal, dan ddl saja menunjukkan kurva yang lebih landai, mencerminkan tingkat survival yang lebih tinggi dalam jangka waktu yang sama. Hasil uji log-rank k sampel menghasilkan p-value sebesar 0.000000, yang jauh lebih kecil dari 0.05, sehingga dapat disimpulkan bahwa terdapat perbedaan survival yang signifikan secara statistik antar keempat kelompok perlakuan.

3.3.8 Kurva Survival Kaplan-Meier – Gender



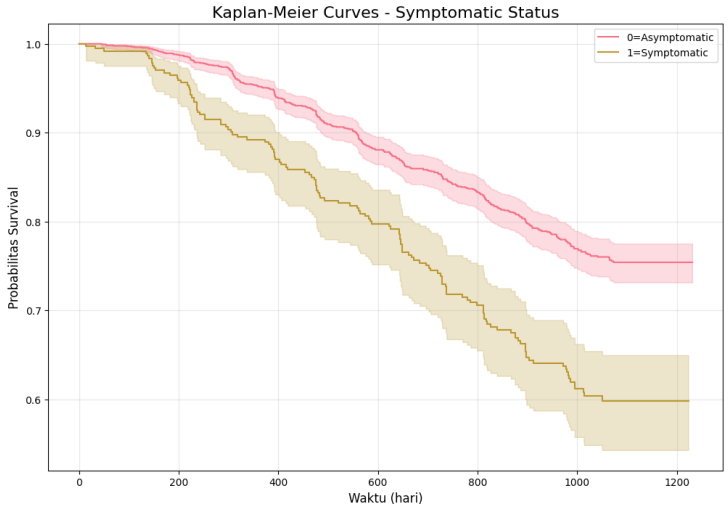
Gambar 12. *Kurva Survival Kaplan-Meier – Gender*

```
Log-rank k sampel p-value: 0.085650
✗ Tidak terdapat perbedaan survival yang signifikan antar kelompok (p ≥ 0.05)
```

Gambar 13. *Uji K-Sampel Pada Kelompok Gender*

Berdasarkan kurva Kaplan-Meier yang membandingkan survival antara jenis kelamin laki-laki dan perempuan, terlihat bahwa perempuan cenderung memiliki probabilitas survival yang sedikit lebih tinggi dibandingkan laki-laki sepanjang periode pengamatan. Namun, perbedaan ini tidak terlalu mencolok dan cenderung paralel. Hasil uji log-rank menunjukkan p-value sebesar 0.085650 yang lebih besar dari ambang signifikansi 0.05. Oleh karena itu, secara statistik tidak terdapat perbedaan survival yang signifikan antara laki-laki dan perempuan dalam data ini.

3.3.9 Kurva Survival Kaplan-Meier – Symptomatic Status



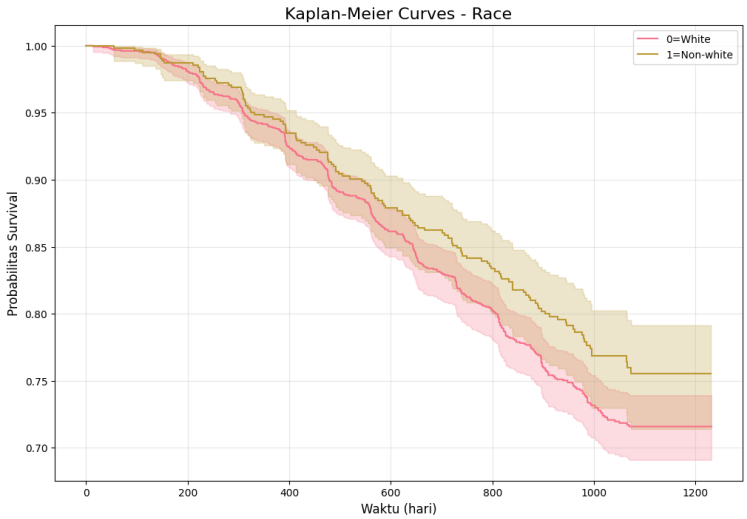
Gambar 14. *Kurva Survival Kaplan-Meier– Symptomatic Status*

```
Log-rank k sampel p-value: 0.000000
✓ Terdapat perbedaan survival yang signifikan antar kelompok (p < 0.05)
```

Gambar 15. *Uji K-Sampel Pada Kelompok Symptomatic Status*

Berdasarkan kurva Kaplan-Meier yang membandingkan status simptomatik, terlihat bahwa individu asimptomatik memiliki probabilitas survival yang secara konsisten lebih tinggi dibandingkan individu simptomatik selama periode pengamatan. Selisih antara kedua kurva cukup jelas, menunjukkan perbedaan yang mencolok dalam risiko kejadian. Hasil uji log-rank menghasilkan p-value sebesar 0.000000 yang jauh lebih kecil dari 0.05, sehingga dapat disimpulkan bahwa terdapat perbedaan survival yang signifikan secara statistik antara kelompok simptomatik dan asimptomatik.

3.3.10 Kurva Survival Kaplan-Meier – Race



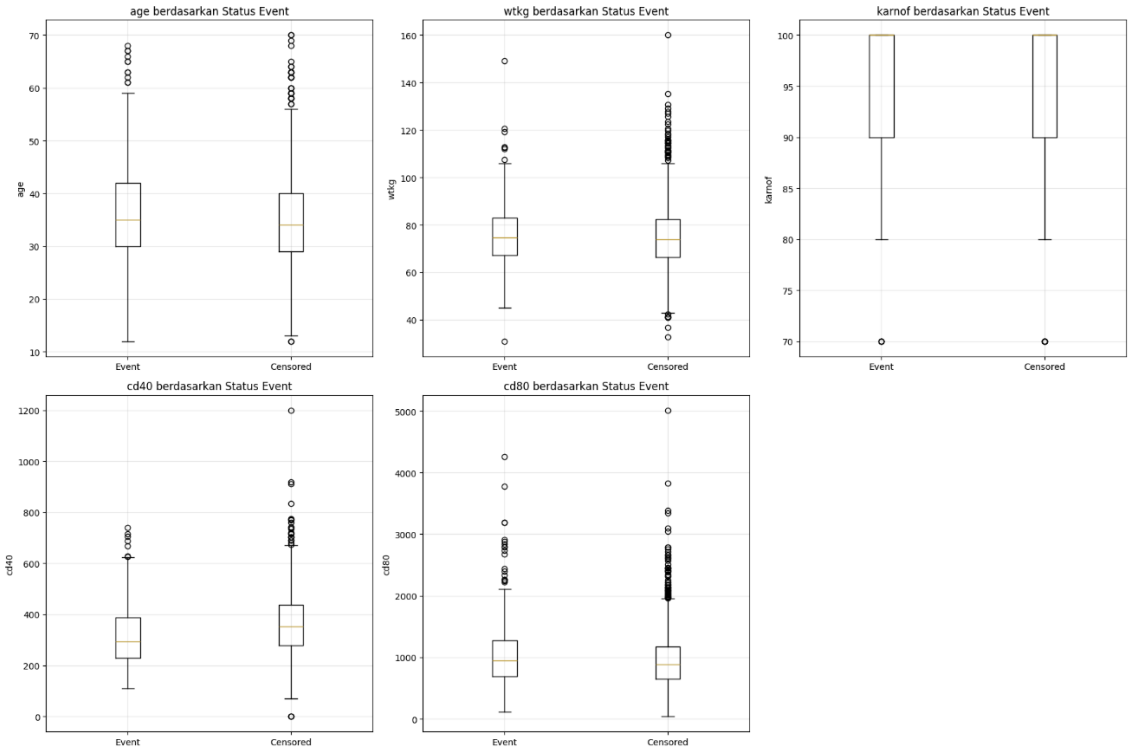
Gambar 16. *Kurva Survival Kaplan-Meier – Race*

```
Log-rank k sampel p-value: 0.061879
✗ Tidak terdapat perbedaan survival yang signifikan antar kelompok (p ≥ 0.05)
```

Gambar 17. *Uji K-Sampel Pada Kelompok Race*

Berdasarkan kurva Kaplan-Meier pada Gambar 3.12 yang membandingkan kelompok ras, terlihat bahwa individu dari kelompok non-white memiliki probabilitas survival yang sedikit lebih tinggi dibandingkan individu white selama periode pengamatan. Meskipun perbedaan antara kedua kurva mulai tampak setelah sekitar 400 hari, hasil uji log-rank menunjukkan p-value sebesar 0.061879 yang lebih besar dari 0.05. Hal ini menunjukkan bahwa tidak terdapat perbedaan survival yang signifikan secara statistik antara kelompok ras white dan non-white, sehingga perbedaan visual pada kurva kemungkinan disebabkan oleh variasi acak atau faktor lain di luar variabel ras.

3.3.11 Distribusi Variabel Kontinu berdasarkan Status



Gambar 18. *Distribusi Variabel Kontinu Berdasarkan Status*

Secara umum, distribusi nilai untuk variabel seperti age, wtkg, karno, cd40, dan cd80 tampak relatif serupa antara kedua kelompok, meskipun terdapat beberapa perbedaan kecil dalam median dan sebaran data. Tidak tampak perbedaan mencolok yang konsisten antara kelompok event dan censored, serta terdapat sejumlah outlier pada hampir semua variabel, terutama pada cd80.

3.3.12 Korelasi antar Variabel Kontinu

Gambar 19. *Korelasi Antar Variabel Kontinu*

Hasilnya memperlihatkan bahwa tidak ada pasangan variabel yang memiliki korelasi kuat (semuanya < 0.3), dengan korelasi tertinggi antara cd40 dan cd80 (0.21). Hal ini mengindikasikan bahwa multikolinearitas bukan masalah signifikan pada data ini.

3.4. Pemodelan Regresi Cox-PH

3.4.1 Model Lengkap

Pemodelan Cox-PH akan menggunakan semua variabel yang ada (kecuali variabel 'zprior') sebagai dasar untuk pemilihan variabel terbaik. Hal ini karena semua variabel tersebut memiliki kontribusi tertentu dalam menentukan kondisi terjangkitnya seseorang terhadap virus AIDS. Pada *R* digunakan fungsi coxph sebagai berikut.

```
> summary(full_model)
Call:
coxph(formula = Surv(time, label) ~ trt + age + wtkg + hemo +
    homo + drugs + karnof + oprior + z30 + preanti + race + gende
r +
    str2 + strat + symptom + treat + offtrt, data = data)

  n= 2139, number of events= 521

              coef  exp(coef)   se(coef)      z Pr(>|z|)
trt       0.0872243  1.0911414  0.0667688  1.306  0.19143
age       0.0118969  1.0119680  0.0052530  2.265  0.02353 *
wtkg      0.0022716  1.0022742  0.0035237  0.645  0.51914
hemo      0.0014010  1.0014020  0.2166172  0.006  0.99484
homo     -0.0037932  0.9962140  0.1617658 -0.023  0.98129
drugs    -0.4301607  0.6504046  0.1532318 -2.807  0.00500 **
karnof   -0.0217574  0.9784776  0.0071518 -3.042  0.00235 **
oprior    0.0871444  1.0910542  0.2742058  0.318  0.75063
z30       0.3196946  1.3767073  0.2225922  1.436  0.15094
preanti   0.0003120  1.0003120  0.0001551  2.011  0.04434 *
race      0.0033036  1.0033090  0.1105711  0.030  0.97616
gender    0.1144706  1.1212796  0.1862744  0.615  0.53887
str2      0.0320782  1.0325982  0.3022177  0.106  0.91547
strat    -0.0529606  0.9484173  0.1596111 -0.332  0.74003
symptom   0.5153539  1.6742309  0.1034310  4.983 6.27e-07 ***
treat    -0.7693403  0.4633186  0.1665648 -4.619 3.86e-06 ***
offtrt    0.7391152  2.0940819  0.0914157  8.085 6.21e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Concordance= 0.681  (se = 0.012 )
Likelihood ratio test= 202.4  on 17 df,   p=<2e-16
Wald test            = 217.9  on 17 df,   p=<2e-16
Score (logrank) test = 225.7  on 17 df,   p=<2e-16
```

Gambar 20. *Hasil Output Model Lengkap*

Berdasarkan output tersebut dapat dilihat bahwa tidak semua variabel data signifikan secara statistik untuk model walaupun model secara statistik signfikan ditandai dengan nilai *Likelihood ratio test, Wald test,* dan *logrank test*. Namun, model lengkap tersebut adalah berdasarkan asumsi bahwa semua variabel pada data mempengaruhi kondisi terjangkitnya penyakit AIDS pada seorang individu. Menilai *Goodness of Fit* dari model, dilihat bahwa nilai C-index (*concordance*) pada model yaitu 0.681 di mana secara umum model tersebut dapat dikatakan "biasa saja" yaitu lebih baik dari tebakan acak, namun masih tidak cukup baik dan diharapkan masih ada ruang untuk perbaikan.

3.4.2 Model Stepwise

Pemilihan variabel sedemikian sehingga variabel pada model regresi signifikan secara statistik digunakan metode *stepwise* berdasarkan AIC dari model lengkap sebelumnya. Metode ini dapat digunakan pada *R* melalui fungsi stepAIC.

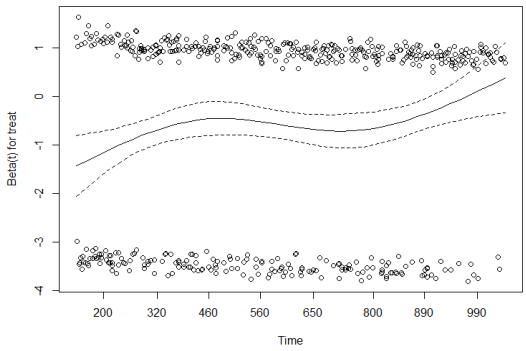
```
Call:
coxph(formula = Surv(time, label) ~ age + drugs + karnof + z30 +
    preanti + symptom + treat + offtrt, data = data)

  n= 2139, number of events= 521

              coef  exp(coef)   se(coef)      z Pr(>|z|)
age       0.0125393  1.0126183  0.0050764  2.470  0.01351 *
drugs    -0.4538370  0.6351862  0.1468711 -3.090  0.00200 **
karnof   -0.0222325  0.9780128  0.0070703 -3.144  0.00166 **
z30       0.2759252  1.3177493  0.1179666  2.339  0.01933 *
preanti   0.0002827  1.0002828  0.0001116  2.533  0.01131 *
symptom   0.5138016  1.6716340  0.1016368  5.055 4.30e-07 ***
treat    -0.5938851  0.5521778  0.0922803 -6.436 1.23e-10 ***
offtrt    0.7416409  2.0993776  0.0910029  8.150 3.65e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Concordance= 0.678  (se = 0.012 )
Likelihood ratio test= 199  on 8 df,   p=<2e-16
Wald test            = 214.5  on 8 df,   p=<2e-16
Score (logrank) test = 221.8  on 8 df,   p=<2e-16
```

***Gambar 21**. Hasil Output Model Stepwise*

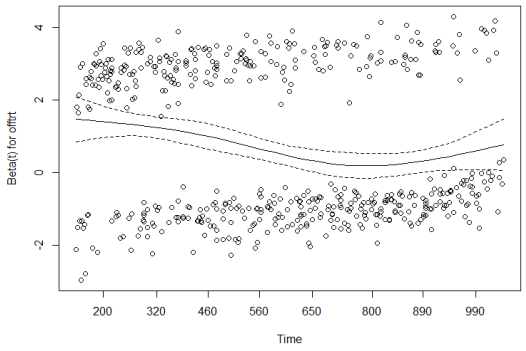
Berdasarkan output model, nilai C-index sedikit turun dari sebelumnya, yaitu pada 0.678. Lalu, secara statistik model signifikan dengan seluruh variabel pada model masih signifikan juga. Lalu, **langkah berikutnya adalah menguji apakah model tersebut memenuhi asumsi *proportional hazard*.** Untuk itu digunakan fungsi cox.zph.

```
            chisq df       p
age       1.0483  1 0.30590
drugs     0.6218  1 0.43037
karnof    0.7321  1 0.39219
z30       0.0363  1 0.84885
preanti   0.6202  1 0.43097
symptom   1.4351  1 0.23093
treat     8.7189  1 0.00315
offtrt   18.9019  1 1.4e-05
GLOBAL   29.5236  8 0.00026
```

***Gambar 22.** Hasil Uji Asumsi Cox-PH Pada Model Stepwise*



Gambar 23. *Plot Residual Schoenfeld variabel 'treat' (kiri)*



Gambar 24. *Plot Residual Schoenfeld variabel 'offtrt' (kanan)*

Pada pengujian asumsi *proportional hazard*, secara global uji statistik mengatakan bahwa adanya asumsi yang dilanggar karena *p-value* global kurang dari 0.05, melihat pada uji asumsi masing-masing variabel dapat diihat bahwa **variabel yang tidak memenuhi asumsi adalah variabel 'treat' dan 'offtrt'.** Oleh karena itu model kembali di-*fit* tanpa menggunakan kedua variabel tersebut.

```
Call:
coxph(formula = Surv(time, label) ~ age + drugs + karnof + z30 +
    preanti + symptom, data = data)

  n= 2139, number of events= 521

              coef  exp(coef)   se(coef)      z Pr(>|z|)
age       0.0095685  1.0096145  0.0050557  1.893   0.0584 .
drugs    -0.3533822  0.7023087  0.1461687 -2.418   0.0156 *
karnof   -0.0304554  0.9700037  0.0070747 -4.305 1.67e-05 ***
z30       0.2765219  1.3185358  0.1183317  2.337   0.0194 *
preanti   0.0002655  1.0002655  0.0001131  2.348   0.0189 *
symptom   0.5728085  1.7732402  0.1011071  5.665 1.47e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Concordance= 0.614  (se = 0.013 )
Likelihood ratio test= 93.72  on 6 df,   p=<2e-16
Wald test            = 98.42  on 6 df,   p=<2e-16
Score (logrank) test = 100.6  on 6 df,   p=<2e-16
```

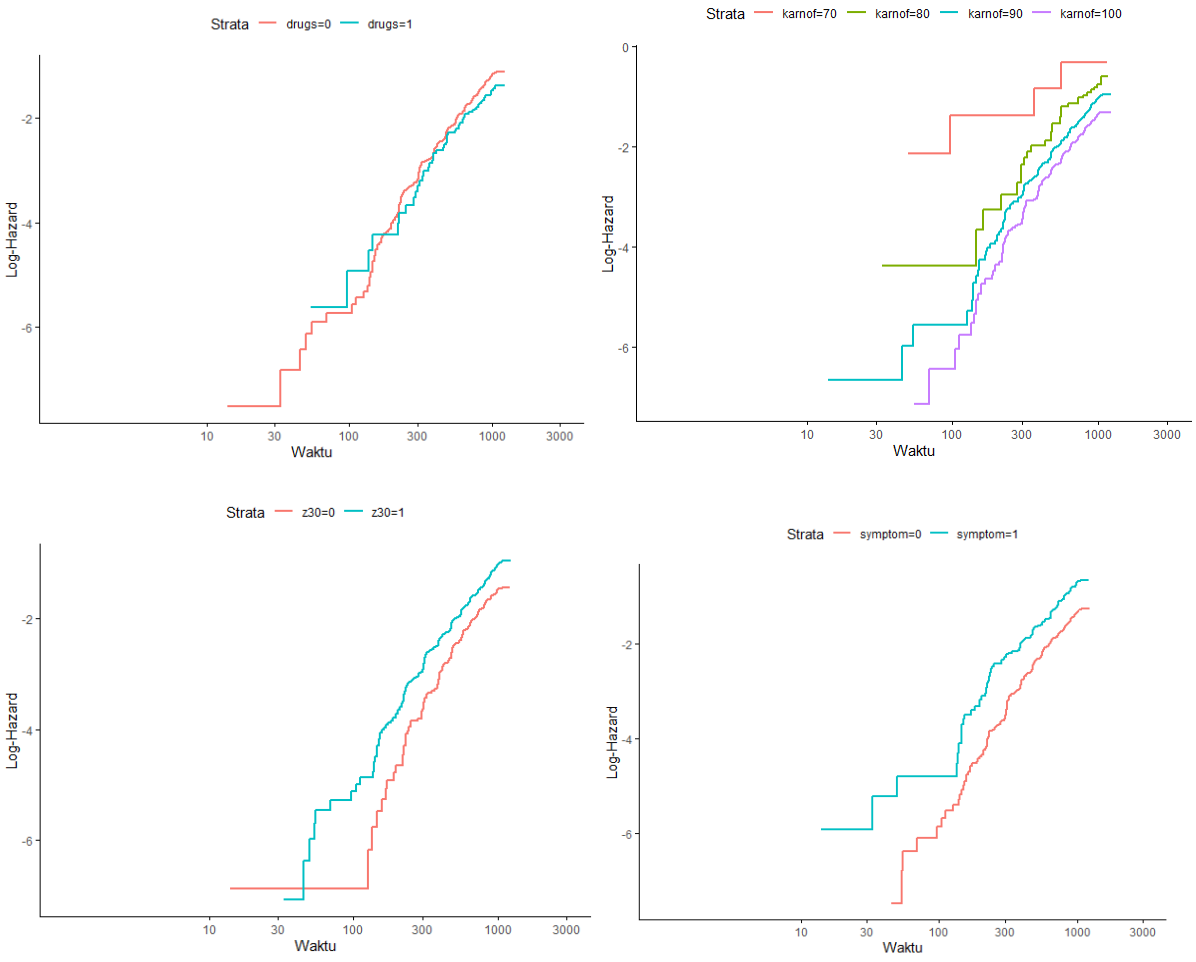
Gambar 25. *Hasil Output Model Stepwise Tanpa Variabel 'treat' dan 'offtrt'*

Model baru ini memiliki penurunan nilai C-index dan pada taraf signifikansi 0.05 memiliki variabel tidak siginifikan 'age'.

	chisq	df	p
age	0.93243	1	0.33
drugs	0.55356	1	0.46
karnof	0.74401	1	0.39
z30	0.00472	1	0.95
preanti	0.32246	1	0.57
symptom	1.29430	1	0.26
GLOBAL	3.54098	6	0.74

Gambar 26. *Hasil Output Model Stepwise Tanpa Variabel 'treat' dan 'offtrt'*

Namun, output pada pengujian asumsi *proportional hazard* menyimpulkan bahwa tidak ada dari variabel-variabel pada model baru yang melanggar asumsi. Maka model ini adalah **model terbaik** jika dipandang secara asumsi *proportional* hazard.



Gambar 27. *Plot waktu vs log(H(t)) untuk variabel 'drugs', 'karnof', 'z30', dan 'symptom'*

Setelah menguji asumsi *proporsional hazard* dilakukan uji multikolinearitas pada variabel model dengan menggunakan fungsi vif.

age	drugs	karnof	z30	preanti	symptom
1.023123	1.017114	1.038558	1.634816	1.648942	1.021848

Gambar 28. *Uji Multikolinearitas Variabel Model Stepwise dengan VIF*

Nilai VIF < 5 menandakan bahwa setiap variabel tidak memiliki multikolinearitas dengan variabel lainnya.

3.5. Interpretasi Model Cox-PH

Berdasarkan output model regresi Cox-PH hasil *stepwise* dan pemilahan variabel, dapat dilakukan analisis deskriptif dari koefisien rasio hazard model. Rasio hazard mengukur perubahan risiko kejadian realtif terhadap perubahan unit pada masing-masing variabel. Berikut adalah interpretasinya.

1) Variabel 'age'
   Koefisien bernilai 0.0096, HR 1.0096, p-value 0.0584. Artinya, setiap peningkatan satu tahun usia meningkatkan risiko kejadian sebesar 1.01%. p-value lebih besar dari 0.05, yang menunjukkan bahwa variabel ini tidak signifikan pada tingkat 0.05, tetapi mendekati batas signifikansi.
2) Variabel 'drugs'
   Koefisien bernilai -0.3534, HR 0.7023, p-value 0.0156. Artinya, penggunaan obat mengurangi risiko kejadian sebesar 29.77%. P-value < 0.05, artinya pengaruhnya signifikan.
3) Variabel 'karnof'
   Koefisien bernilai -0.0305, HR 0.9700, p-value 1.6e-05. Artinya, setiap kenaikan satu poin pada skor Karnofski mengurangi risiko kejadian sebesar 3.00%. P-value sangat kecil (< 0.001), yang menunjukkan bahwa pengaruhnya sangat signifikan.
4) Variabel 'z30'
   Koefisien bernilai -0.0305, HR 0.9700, p-value 1.6e-05. Artinya, setiap peningkatan satu unit pada z30 meningkatkan risiko kejadian sebesar 31.85%. P-value < 0.05, artinya pengaruhnya signifikan.
5) Variabel 'preanti'
   Koefisien bernilai 0.0003, HR 1.0003, p-value 0.0189. Artinya, peningkatan sedikit pada preanti mengurangi risiko kejadian, meskipun pengaruhnya sangat kecil (HR hampir 1). P-value < 0.05, artinya pengaruhnya signifikan, meskipun kecil.
6) Variabel 'symptom'
   Koefisien bernilai 0.5728, HR 1.7732, p-value 1.47e-08. Artinya, keberadaan gejala meningkatkan risiko kejadian sebesar 77.32%. P-value sangat kecil (< 0.001), menunjukkan bahwa pengaruh gejala sangat signifikan terhadap kejadian.

3.6. Kontribusi Likelihood

3.6.1 Kontribusi Partial Likelihood pada Data Tanpa Ties

Akan dilakukan konstruksi partial likelihood untuk model Cox-PH pada subset data menggunakan variabel prediktor numerik age (x_1) dan variabel prediktor kategorik gender (x_2) dengan total 30 observasi. Berikut adalah subset data yang digunakan.

```
> print(subdf)
   time age gender label
1    14  30      1     1
2    33  36      1     1
3    45  38      1     1
4    50  43      1     1
5    54  46      1     1
6    55  38      1     1
7    62  41      1     0
8    69  33      1     1
9    96  40      0     1
10  105  33      1     1
11  109  33      1     0
12  111  39      1     1
13  125  31      1     0
14  126  43      1     1
15  133  30      1     0
16  133  42      1     0
17  134  51      0     0
18  135  44      1     1
19  137  37      1     1
20  138  35      1     0
21  139  25      1     0
22  140  39      1     1
23  140  34      1     0
24  140  39      1     1
25  142  45      1     1
26  143  27      0     0
27  145  27      1     0
28  146  30      1     1
29  146  57      1     1
30  147  38      1     0
```

Gambar 29. *Subset Data*

Misalkan, kita ingin mengetahui kontribusi *partial* likelihood pada $t_1 = 14$. Pada waktu tersebut hanya terjadi satu event dengan total risiko dari seluruh pasien yang ada pada setiap waktu t_j sebanyak 30. Perhatikan bahwa risiko pasien mengalami event pada waktu t_1 adalah

$$h\big(14\big|x_1 = (30,1)\big) = h_0(14)e^{30\beta_1+\beta_2}$$

dan risiko relatif dari seluruh objek yang ada pada setiap waktu t_1 adalah

$$h(14|R_1) = h_0(14)e^{30\beta_1+\beta_2} + h_0(14)e^{36\beta_1+\beta_2} + \cdots + h_0(14)e^{38\beta_1+\beta_2}$$

Sehingga partial likelihood pada $t_1 = 14$ adalah

$$L(14) = \frac{h_0(14)e^{30\beta_1+\beta_2}}{h_0(14)e^{30\beta_1+\beta_2} + h_0(14)e^{36\beta_1+\beta_2} + \cdots + h_0(14)e^{38\beta_1+\beta_2}}$$

$$= \frac{e^{30\beta_1+\beta_2}}{e^{30\beta_1+\beta_2} + e^{36\beta_1+\beta_2} + \cdots + e^{38\beta_1+\beta_2}}$$

Untuk $t_2 = 33$ dapat dilakukan konstruksi partial likelihood yang serupa. Pada waktu ini terjadi satu event dan total risiko dari seluruh pasien $|R_2| = 29$. Sehingga risiko pasien mengalami event pada waktu t_2 adalah

$$h\big(33\big|x_2 = (36,1)\big) = h_0(33)e^{36\beta_1+\beta_2}$$

dan risiko relatif dari seluruh objek yang ada pada setiap waktu t_2 adalah

$$h(33|R_2) = h_0(33)e^{36\beta_1+\beta_2} + \cdots + h_0(33)e^{38\beta_1+\beta_2}$$

Sehingga partial likelihood pada $t_2 = 33$ adalah

$$L(33) = \frac{h_0(33)e^{36\beta_1+\beta_2}}{h_0(33)e^{36\beta_1+\beta_2} + \cdots + h_0(33)e^{38\beta_1+\beta_2}}$$

Untuk semua titik waktu yang tanpa ties, bisa dilakukan konstruksi yang serupa. Namun perhatikan bahwa pada subset data yang diambil terdapat data dengan ties, sehingga fungsi partial likelihood keseluruhan tidak dapat dibentuk.

3.6.2 Kontribusi Likelihood pada Data Ties

Pada subset data yang digunakan, terjadi dua event pada $t = 146$. Dengan menggunakan metode Breslow, asumsikan kedua event terjadi satu per satu. Ketika satu objek mengalami event, maka objek lainnya diasumsikan belum mengalami event sehingga masih menjadi bagian dari himpunan objek berisiko.

Pada subset data di titik waktu $t = 146$, terdapat 3 anggota himpunan risiko. Ketiganya memiliki gender laki-laki. Maka partial likelihood pada waktu tersebut adalah

$$L(146) = \frac{\exp(30\beta_1 + \beta_2) \cdot \exp\,(57\beta_1 + \beta_2)}{[\exp(30\beta_1 + \beta_2) + \exp(57\beta_1 + \beta_2) + \exp(38\beta_1 + \beta_2)]^2}$$

Untuk data dengan ties yang lain dapat dikonstruksi dengan cara serupa. Sehingga fungsi partial likelihood keseluruhan dapat dikonstruksi dengan mengalikan kontribusi likelihood pada semua waktu. Misalkan R_i^+ himpunan objek berisiko dengan gender laki-laki dan R_i^- himpunan objek berisiko dengan gender perempuan pada waktu t_i, maka

$$L(\beta) = L(14) \cdot L(33) \cdot \ldots \cdot L(146)$$

$$= \frac{\exp(30\beta_1 + \beta_2)}{\sum_{x_1 \in R_1^+} \exp(\beta_1 x_1 + \beta_2) + \sum_{x_1 \in R_1^-} \exp(\beta_1 x_1)} \cdot \frac{\exp(36\beta_1 + \beta_2)}{\sum_{x_1 \in R_2^+} \exp(\beta_1 x_1 + \beta_2) + \sum_{x_1 \in R_2^-} \exp(\beta_1 x_1)} \cdot \ldots \cdot \frac{\exp(30\beta_1 + \beta_2) \cdot \exp\,(57\beta_1 + \beta_2)}{[\exp(30\beta_1 + \beta_2) + \exp(57\beta_1 + \beta_2) + \exp(38\beta_1 + \beta_2)]^2}$$

Selanjutnya estimasi parameter β_1 dan β_2 dapat dicari dengan memaksimumkan fungsi likelihood tersebut. Dengan menggunakan R, didapatkan $\widehat{\beta_1}$=0.02084 dan $\widehat{\beta_2} = 0.60612$

```
> breslow <-coxph(Surv(time, label)~age+gender, data=subdf, ties="breslow")
> summary(breslow)
Call:
coxph(formula = Surv(time, label) ~ age + gender, data = subdf,
    ties = "breslow")

  n= 30, number of events= 18

          coef exp(coef) se(coef)     z Pr(>|z|)
age    0.02084   1.02106  0.02669 0.781    0.435
gender 0.60612   1.83330  1.03476 0.586    0.558

       exp(coef) exp(-coef) lower .95 upper .95
age        1.021     0.9794    0.9690     1.076
gender     1.833     0.5455    0.2412    13.932

Concordance= 0.558  (se = 0.08 )
Likelihood ratio test= 1.05  on 2 df,   p=0.6
Wald test            = 0.98  on 2 df,   p=0.6
Score (logrank) test = 0.99  on 2 df,   p=0.6
```

Gambar 30. *Hasil Estimasi Parameter Menggunakan Metode Breslow*

BAB IV
KESIMPULAN

Dari hasil dan pembahasan di bab III, dapat disimpulkan bahwa:

1. Terdapat perbedaan *survival* experience antara kelompok ‘treatment’ dan ‘symptom’.
2. Model *stepwise* adalah model terbaik berdasarkan uji asumsi *Cox*-PH karena semua variabel yang berkontribusi pada model ini memenuhi asumsi kecuali variabel ‘age’, walaupun nilai C-*index* cenderung cukup kecil 0.614.
3. Variabel-variabel yang signifikan dalam mempengaruhi waktu terjadinya *event* pasien HIV/AIDS adalah ‘age’, ‘drugs’, ‘karnof’, ‘z30’, ‘preanti’, dan ‘symptom’.
4. Setiap peningkatan satu tahun usia meningkatkan risiko kejadian sebesar 1.01%.
5. Penggunaan obat mengurangi risiko kejadian sebesar 29.77%.
6. Setiap kenaikan satu poin pada skor Karnofski mengurangi risiko kejadian sebesar 3.00%.
7. Setiap peningkatan satu unit pada z30 meningkatkan risiko kejadian sebesar 31.85%.
8. Peningkatan sedikit pada preanti mengurangi risiko kejadian, meskipun pengaruhnya sangat kecil.
9. Keberadaan gejala meningkatkan risiko kejadian sebesar 77.32%.

Berikut adalah model *Cox*-PH stepwise.

$$h(t,x) = h_0(t)[\beta_1(age) + \beta_2(drugs) + \beta_3(karnof) + \beta_4(z30) + \beta_5(preanti) + \beta_6(symptom)$$

‘drugs’ = 0 jika tidak menggunakan narkoba tipe IV dan 1 jika menggunakan narkoba tipe IV. ‘z30’ = 0 jika tidak menggunakan ZDV 30 hari sebelum studi dan 1 jika menggunakan ZDV 30 hari sebelum studi. ‘symptom’ = 0 jika asimptomatik dan 1 jika simptomatik.

DAFTAR PUSTAKA

Abdullah, S., 2022. *Analisis Survival: Konsep dan Aplikasi dengan R.* Bumi Aksara.

Alimonti, J. B., Ball, T. B., & Fowke, K. R. (2003). Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. Journal of general Virology, 84(7), 1649-1661. https://doi.org/10.1099/vir.0.19110-0

Barry, M., Howe, J. L., Ormesher, S., Back, D. J., Breckenridge, A. M., Bergin, C., ... & Nye, F. (1994). Pharmacokinetics of zidovudine and dideoxyinosine alone and in combination in patients with the acquired immunodeficiency syndrome. British journal of clinical pharmacology, 37(5), 421-426. https://doi.org/10.1111/j.1365-2125.1994.tb05708.x

Barry, M., Mulcahy, F., & Back, D. J. (1998). Antiretroviral therapy for patients with HIV disease. British journal of clinical pharmacology, 45(3), 221. https://doi.org/10.1046/j.1365-2125.1998.00673.x

Collett, D. (2015). Modelling Survival Data in Medical Research. Chapman and Hall/CRC. https://doi.org/10.1201/b18041

Kleinbaum, D.G. and Klein, M., 2010. *Survival analysis: A-self learning text.* (Vol. 3). New York: Springer.

LAMPIRAN

Link Code:

ristek.link/LampiranCodeKelompok7