BAB 2
LANDASAN TEORI

2.1 Tinjauan Teori

Bab ini menjelaskan landasan teori yang menjadi dasar penelitian. Pembahasan dimulai dari diabetes melitus tipe 2 sebagai kondisi indeks, konsep komorbiditas dan multimorbiditas, serta penggunaan ICD-10 untuk merepresentasikan diagnosis klinis. Setelah itu, bab ini membahas *Association Rule Mining* (ARM), algoritma FP-Growth, analisis ketahanan hidup, indeks keparahan klinis, dan MIMIC-IV sebagai sumber data penelitian. Urutan pembahasan ini disusun agar hubungan antara teori klinis, metode komputasional, dan analisis statistik dapat dipahami secara sistematis.

2.1.1 Diabetes Melitus Tipe 2 sebagai Kondisi Indeks dalam Analisis Komorbiditas

Diabetes melitus tipe 2 (DMT2) didefinisikan sebagai kondisi hiperglikemia kronis yang diakibatkan oleh resistansi insulin pada jaringan perifer yang dikombinasikan dengan disfungsi sel beta pankreas secara progresif, yang bersama-sama menghasilkan defisiensi insulin relatif [12, 13]. Dalam konteks epidemiologi klinis, T2DM tidak muncul sebagai entitas penyakit tunggal, melainkan hampir selalu disertai oleh kondisi komorbid yang membentuk pola multimorbiditas kompleks. Salah satu studi di Republik Rakyat Cina menunjukkan bahwa 79,8% pasien T2DM memiliki setidaknya satu komorbiditas, dengan 51% mengalami dua atau lebih kondisi penyerta [14].

Beban komorbiditas pada pasien DMT2 bersifat kumulatif dan longitudinal. Studi berbasis data klaim nasional Taiwan yang mencakup lebih dari satu juta pasien DMT2 menunjukkan bahwa aDCSI (*adapted Diabetes Complications Severity Index*) dan CCI keduanya berhubungan secara signifikan dengan mortalitas semua penyebab, namun tidak ada indeks tunggal yang mampu sepenuhnya menangkap interaksi antarkomorbiditas yang berkontribusi terhadap prognosis [3]. Hal ini memperkuat argumen bahwa pendekatan berbasis pola asosiasi, yang secara eksplisit memodelkan keterkaitan antarkondisi, lebih kaya secara informasional dibandingkan pendekatan indeks tunggal.

2.1.2 Konsep Komorbiditas dan Multimorbiditas dalam Epidemiologi Klinis

Dalam literatur ilmiah, istilah *comorbidity* (komorbiditas) dan *multimorbidity* (multimorbiditas) sering kali digunakan secara bergantian (*interchangeably*) meskipun secara konseptual keduanya memiliki perbedaan yang mendasar [15]. Oleh karena itu, penelitian ini secara eksplisit mengadopsi definisi bahwa komorbiditas didefinisikan sebagai kehadiran satu atau lebih kondisi medis tambahan yang terjadi bersamaan dengan penyakit primer (*index disease*), sedangkan multimorbiditas merujuk pada koeksistensi dua atau lebih kondisi kronis tanpa penetapan penyakit primer [15, 16].

2.1.3 Sistem Klasifikasi ICD-10 dalam Analisis Komorbiditas Berbasis Data

International Classification of Diseases, Tenth Revision (ICD-10) adalah sistem klasifikasi diagnostik yang dikelola oleh WHO dan digunakan secara global untuk pengodean diagnosis klinis [17]. Dalam konteks MIMIC-IV, kode ICD-10 digunakan untuk mencatat seluruh kondisi yang relevan selama perawatan, mencakup diagnosis utama, komorbiditas penyerta, dan komplikasi yang timbul.

Untuk keperluan ARM dalam penelitian ini, kode ICD-10 dipotong hingga tiga digit pertama (*level category*), yang merepresentasikan entitas penyakit spesifik tanpa subspesifikasi klinisnya. Pendekatan ini konsisten dengan studi komorbiditas berbasis ICD-10 sebelumnya dan memungkinkan agregasi yang cukup untuk menghasilkan *support* yang memadai pada level populasi, sekaligus cukup kecil untuk mempertahankan makna klinis [4, 6, 18, 19, 20]. Kode dari blok Z (faktor yang memengaruhi status kesehatan), Y (penyebab eksternal morbiditas), dan R (gejala, tanda, dan temuan klinis abnormal) dikecualikan dari analisis karena tidak merepresentasikan kondisi patologis yang terdiagnosis secara definitif.

2.1.4 Association Rule Mining (ARM)

A Konsep Dasar Association Rule Mining

Association Rule Mining (ARM) adalah teknik penambangan data yang bertujuan menemukan hubungan yang menarik, berulang, dan tidak trivial antarvariabel dalam *dataset* transaksional berskala besar [21]. Dalam konteks klinis, setiap pasien diperlakukan sebagai sebuah transaksi, dan kode diagnosis ICD-10 yang dimiliki pasien tersebut diperlakukan sebagai *item* dalam transaksi.

ARM kemudian mencari pola berupa aturan implikasi berbentuk:

$$X \Rightarrow Y \tag{2.1}$$

di mana X (anteseden) dan Y (konsekuen) adalah himpunan *item* (*itemset*) yang tidak saling beririsan ($X \cap Y = \emptyset$), dan aturan tersebut menyatakan bahwa kemunculan X cenderung diikuti oleh kemunculan Y dalam transaksi yang sama [22, 23].

B Metrik Evaluasi Aturan Asosiasi

Tiga metrik utama digunakan untuk mengevaluasi kualitas aturan asosiasi [6, 22, 24, 25]:

Support mengukur seberapa sering suatu *itemset* muncul dalam *dataset*. Untuk aturan $X \Rightarrow Y$, *support* didefinisikan sebagai:

$$\text{support}(X \Rightarrow Y) = \frac{|\{t \in T : X \cup Y \subseteq t\}|}{|T|} \tag{2.2}$$

di mana T adalah himpunan seluruh transaksi (pasien) dan $|T|$ adalah jumlah total transaksi. *Support* mencerminkan prevalensi klinis dari kombinasi komorbiditas dalam populasi.

Confidence mengukur kekuatan kondisional aturan, yaitu proporsi transaksi yang mengandung X yang juga mengandung Y:

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} = P(Y \mid X) \tag{2.3}$$

Confidence tinggi mengindikasikan bahwa kemunculan anteseden X merupakan prediktor yang kuat bagi kemunculan konsekuen Y.

Lift mengukur seberapa jauh hubungan antara X dan Y menyimpang dari kemunculan acak yang independen:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{support}(Y)} = \frac{P(X \cup Y)}{P(X) \cdot P(Y)} \tag{2.4}$$

Nilai $lift > 1$ mengindikasikan bahwa X dan Y lebih sering muncul bersama

daripada yang diharapkan secara acak, yang merupakan syarat minimal bagi sebuah aturan untuk dianggap memiliki asosiasi positif yang bermakna, sedangkan nilai $lift = 1$ menunjukkan independensi, dan nilai $lift < 1$ menunjukkan asosiasi negatif [6, 22, 24, 25]. Dalam penelitian ini, ambang batas penyaringan aturan ditetapkan pada $support \geq 0{,}08$, $confidence \geq 0{,}60$, dan $lift \geq 2{,}0$.

Setelah kualitas aturan asosiasi dijelaskan melalui *support*, *confidence*, dan *lift*, tahap berikutnya adalah menentukan algoritma yang digunakan untuk menemukan *frequent itemset*. Penelitian ini menggunakan FP-Growth karena algoritma tersebut lebih efisien untuk data transaksi berukuran besar dan tidak memerlukan pembentukan kandidat secara berulang seperti Apriori. Oleh karena itu, penjelasan berikut berfokus pada prinsip kerja, tahapan, dan alasan pemilihan FP-Growth dalam konteks analisis komorbiditas berbasis ICD-10.

C Algoritma FP-Growth

Algoritma *Frequent Pattern Growth* (FP-Growth) merupakan pendekatan penambangan *frequent itemset* yang lebih efisien dibandingkan algoritma Apriori. Berbeda dengan Apriori yang secara berulang memindai seluruh basis data untuk menghasilkan kandidat *itemset*, FP-Growth menggunakan struktur data ringkas bernama FP-tree (*Frequent Pattern tree*) yang merepresentasikan seluruh transaksi dalam satu pohon terkompresi [22].

Proses FP-Growth terdiri dari dua tahap utama. Pertama, konstruksi FP-tree: basis data dipindai sebanyak dua kali. Kali pertama untuk menghitung frekuensi *item* tunggal, dan sekali lagi untuk membangun pohon dengan mengurutkan *item* berdasarkan frekuensi menurun dalam tiap transaksi. Kedua, penambangan *conditional pattern base*: untuk setiap *item* yang sering muncul, FP-Growth membangun FP-tree bersyarat yang merepresentasikan hanya pola-pola yang berakhir pada *item* tersebut. Kemudian, FP-Growth secara rekursif menelusuri pohon kondisional tersebut untuk menghasilkan seluruh *frequent itemset* [22].

Kompleksitas waktu FP-Growth adalah $O(|T| \cdot \bar{l})$ di mana $\bar{l}$ adalah panjang rata-rata transaksi, jauh lebih efisien dibandingkan Apriori yang memiliki kompleksitas eksponensial terhadap panjang *itemset* maksimum [26]. Keunggulan ini menjadi sangat signifikan pada *dataset* klinis berskala besar seperti MIMIC-IV, di mana setiap pasien dapat memiliki hingga puluhan kode ICD-10 yang berbeda.

Secara lebih rinci, langkah kerja algoritma FP-Growth yang diimplementasikan dalam penelitian ini mengikuti struktur dua tahap yang

telah dijelaskan di atas dan disajikan dalam bentuk *pseudocode* pada Kode 2.1.

```
INPUT   : Basis data transaksi D (profil komorbiditas per
   pasien),
          ambang batas min_support
OUTPUT  : Himpunan seluruh frequent itemset F

TAHAP 1 - Konstruksi FP-tree
1.  Pindai D satu kali untuk menghitung frekuensi setiap item
    (kode ICD-10 tiga digit) secara individual
2.  Buang item dengan frekuensi < min_support x |D|
3.  Urutkan item yang tersisa berdasarkan frekuensi menurun ->
   L
4.  Inisialisasi root FP-tree sebagai node kosong
5.  FOR setiap transaksi T dalam D:
6.      Urutkan item dalam T mengikuti urutan L (frekuensi
   menurun)
7.      current_node = root
8.      FOR setiap item p dalam T (sudah terurut):
9.          IF current_node memiliki child berlabel p:
10.             Tambahkan count child tersebut sebesar 1
11.         ELSE:
12.             Buat child node baru berlabel p dengan count =
   1
13.             Tautkan node baru ke header table pada label p
14.         current_node = child berlabel p

TAHAP 2 - Penambangan Conditional Pattern Base (rekursif)
15. FUNCTION FPGrowth(Tree, alpha):
16.     IF Tree hanya memiliki satu lintasan tunggal P:
17.         FOR setiap kombinasi item beta pada lintasan P:
18.             Hasilkan itemset (beta UNION alpha),
19.                 support = nilai count minimum pada beta
20.     ELSE:
21.         FOR setiap item a_i pada header table Tree,
22.             dimulai dari item dengan frekuensi terendah
   :
23.             beta = {a_i} UNION alpha
24.             Hasilkan itemset beta dengan support = support(
   a_i)
25.             Bangun conditional pattern base dari a_i
26.                 (seluruh lintasan prefix menuju a_i pada
   Tree)
27.             Bangun conditional FP-tree Tree_beta
28.                 dari conditional pattern base tersebut
29.             IF Tree_beta bukan kosong:
30.                 CALL FPGrowth(Tree_beta, beta)
31. RETURN F = seluruh itemset yang dihasilkan

MAIN
32. CALL FPGrowth(FP-tree hasil Tahap 1, alpha = {})
```

Kode 2.1: *Pseudocode* algoritma FP-Growth

Implementasi aktual pada penelitian ini menggunakan fungsi `fpgrowth` dari pustaka `mlxtend.frequent_patterns` pada Python, yang menjalankan logika algoritmik yang setara dengan *pseudocode* di atas secara teroptimasi sebagaimana dijelaskan lebih lanjut pada Bab 3.

Meskipun FP-Growth mampu menemukan pola kemunculan bersama secara efisien, hasil yang diperoleh tetap harus ditafsirkan sesuai dengan sifat dasar ARM.

Aturan yang dihasilkan menunjukkan hubungan statistik antarkode diagnosis dalam transaksi yang sama, bukan urutan kejadian klinis atau mekanisme sebab-akibat. Oleh karena itu sebelum hasil ARM digunakan untuk membentuk kluster dan dibahas secara klinis, batas interpretasi kausal dari metode ini perlu ditegaskan terlebih dahulu.

D Keterbatasan Interpretasi Kausal pada Association Rule Mining

Meskipun notasi aturan asosiasi $X \Rightarrow Y$ (Rumus 2.1) menggunakan simbol panah yang secara visual menyerupai notasi implikasi logis atau hubungan sebab-akibat, perlu ditegaskan bahwa ARM secara fundamental merupakan teknik yang mengukur (*statistical co-occurrence*) dalam data transaksional, dan bukan merupakan metode untuk menetapkan hubungan kausal antarvariabel [21, 22].

Terdapat tiga keterbatasan mendasar yang membedakan ARM dari analisis kausal formal. Pertama, ketiga metrik evaluasi, *support*, *confidence*, dan *lift*, seluruhnya diturunkan dari frekuensi kemunculan bersama *itemset* dalam himpunan transaksi tanpa memperhitungkan urutan temporal kemunculan tiap *item*. Nilai *lift*, bahkan bersifat simetris terhadap arah aturan, yaitu $\text{lift}(X \Rightarrow Y) = \text{lift}(Y \Rightarrow X)$ sehingga notasi panah pada suatu aturan semata-mata merefleksikan arah *conditional probability* yang dihitung sebagai *confidence*, bukan arah pengaruh kausal antara X dan Y.

Kedua karena profil komorbiditas pasien dalam penelitian ini dikonstruksi dari seluruh riwayat admisi tanpa penanda urutan waktu kemunculan tiap diagnosis (*lifetime comorbidity burden*, lihat Batasan Permasalahan pada Bab 1), ARM tidak dapat membedakan apakah suatu kondisi mendahului, mengikuti ataupun muncul bersamaan dengan kondisi lainnya. Ketiga, aturan asosiasi tidak menyesuaikan terhadap variabel perancu (*confounding variables*) seperti usia, tingkat keparahan penyakit dasar, durasi penyakit atau riwayat pengobatan, yang berpotensi menjadi penyebab bersama (*common cause*) dari dua kondisi yang tampak berasosiasi kuat, tanpa salah satunya benar-benar menyebabkan yang lain.

Dengan demikian, sinyal ARM yang ditemukan bermakna secara klinis dalam penelitian ini, misalnya keterkaitan kuat antara N18 (penyakit ginjal kronis), I50 (gagal jantung), dan I13 (penyakit jantung hipertensif dengan penyakit ginjal kronis) harus diinterpretasikan sebagai bukti kemunculan bersama secara statistik yang *konsisten dengan*, bukan yang *membuktikan*, mekanisme patofisiologis yang telah didokumentasikan secara independen dalam literatur klinis. Interpretasi

kausal atas pola tersebut hanya dapat didukung secara lebih kuat melalui triangulasi dengan bukti eksperimental atau studi longitudinal yang secara eksplisit mengontrol urutan waktu dan variabel perancu, sesuatu yang berada di luar cakupan metodologis ARM itu sendiri.

2.1.5 Analisis Ketahanan Hidup (Survival Analysis)

Analisis ketahanan hidup adalah kumpulan metode statistik yang dirancang untuk memodelkan waktu hingga terjadinya suatu kejadian (*time-to-event data*) [27]. Dalam konteks penelitian ini, kejadian (*event*) yang dimaksud adalah kematian pasien, dan variabel waktu adalah durasi *follow-up* dari admisi pertama hingga kematian atau penyensoran. Fitur kunci yang membedakan analisis *survival* dari regresi standar adalah penanganan data tersensor (*censored data*), yaitu observasi di mana kejadian belum terjadi pada akhir periode pengamatan [28].

Dalam penelitian ini, analisis ketahanan hidup digunakan untuk mengevaluasi apakah kluster komorbiditas yang terbentuk dari hasil ARM memiliki perbedaan prognosis. Oleh karena itu, bagian ini menjelaskan konsep dasar fungsi *survival* dan *hazard*, estimator Kaplan-Meier, model regresi Cox, penghitungan nilai-p dan *hazard ratio*, serta pengujian asumsi *proportional hazards*. Rangkaian konsep ini menjadi dasar untuk menafsirkan hasil analisis prognosis pada Bab 4.

A Fungsi Survival dan Fungsi Hazard

Misalkan T adalah variabel acak kontinu yang merepresentasikan waktu hingga kematian, dengan fungsi distribusi kumulatif $F(t) = P(T \leq t)$. Fungsi *survival* $S(t)$ didefinisikan sebagai probabilitas seorang individu bertahan hidup melampaui waktu t:

$$S(t) = P(T > t) = 1 - F(t) \tag{2.5}$$

dengan sifat $S(0) = 1$ dan $\lim_{t\to\infty} S(t) = 0$ dan T adalah variabel acak waktu hingga kejadian dan $F(t)$ adalah fungsi distribusi kumulatif.

Fungsi *hazard* $h(t)$ mendefinisikan laju instansus terjadinya *event* pada waktu t pada kondisi individu masih bertahan hingga t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.6}$$

di mana $f(t) = -\frac{dS(t)}{dt}$ adalah fungsi densitas probabilitas waktu *event*. Hubungan antara fungsi *survival* dan *hazard* dapat dinyatakan sebagai:

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp(-H(t)) \tag{2.7}$$

di mana $H(t) = \int_0^t h(u)du$ adalah fungsi *cumulative hazard* [29].

B Estimator Kaplan-Meier

Estimator Kaplan-Meier (KM) adalah metode nonparametrik untuk mengestimasi fungsi *survival* dari data yang mengandung sensor [30, 31]. Misalkan terdapat k waktu *event* yang berbeda $t_1 < t_2 < \cdots < t_k$.. Pada setiap waktu t_j, didefinisikan d_j sebagai jumlah *event* (kematian) yang terjadi dan n_j sebagai jumlah individu yang masih berisiko (*at risk*) sesaat sebelum t_j. Estimator KM didefinisikan sebagai:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.8}$$

Varian dari estimator KM dapat diperoleh menggunakan formula Greenwood:

$$\widehat{\text{Var}}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{2.9}$$

Perbandingan kurva KM antarkelompok dalam penelitian ini dilakukan menggunakan uji *log-rank*, yang menguji hipotesis nol bahwa tidak terdapat perbedaan dalam fungsi *survival* antarkelompok. Statistik uji *log-rank* diformulasikan sebagai:

$$\chi^2 = \frac{(\sum_i (O_{1i} - E_{1i}))^2}{\sum_i V_{1i}} \tag{2.10}$$

di mana O_{1i} adalah jumlah *event* yang terobservasi pada kelompok 1 di

waktu t_i, dan $E_{1i} = n_{gj} \cdot \frac{d_j}{n_j}$ adalah jumlah *event* yang diharapkan di bawah hipotesis nol. Statistik ini berdistribusi χ^2 dengan derajat bebas $G-1$ [31].

C Model Regresi Cox (Cox Proportional Hazards Model)

Model regresi Cox adalah model semiparametrik yang memungkinkan pemodelan pengaruh *covariate* terhadap *hazard* tanpa mengasumsikan bentuk parametrik tertentu untuk *baseline hazard* [30]. Model Cox diformulasikan sebagai:

$$h(t \mid \mathbf{x}) = h_0(t) \cdot \exp\left(\sum_{j=1}^{P} \beta_j x_j\right) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{x}) \tag{2.11}$$

di mana $h_0(t)$ adalah *baseline hazard function* yang tidak dispesifikasi, $x = (x_1, x_2, \ldots, x_P)^\top$ adalah vektor kovariat, dan $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^\top$ adalah vektor koefisien yang diestimasi melalui *partial likelihood* [27].

Parameter β diestimasi dengan memaksimalkan *partial likelihood* Cox:

$$\mathscr{L}(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{j \in \mathscr{R}(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_j)} \tag{2.12}$$

di mana $\delta_i \in \{0,1\}$ adalah indikator *event* (1 jika meninggal, 0 jika tersensor) dan $\mathscr{R}(t_i)$ adalah *risk set*, himpunan individu yang masih berisiko sesaat sebelum waktu t_i [27, 30].

D Uji Signifikansi dan Interval Kepercayaan Koefisien Cox

Setelah estimasi $\hat{\boldsymbol{\beta}}$ diperoleh melalui maksimisasi *partial likelihood* pada Rumus 2.12, signifikansi statistik tiap koefisien $\hat{\beta}_j$ diuji menggunakan uji Wald. Galat baku (*standard error*) $SE(\hat{\beta}_j)$ diperoleh dari akar diagonal matriks kovarians, yang dihitung sebagai invers matriks informasi Fisher (turunan kedua negatif dari *log partial likelihood* pada titik $\hat{\boldsymbol{\beta}}$). Statistik uji Wald untuk koefisien ke-j didefinisikan sebagai:

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{2.13}$$

di bawah hipotesis nol $H_0 : \beta_j = 0$ (tidak terdapat pengaruh kovariat terhadap *hazard*), statistik z_j mengikuti distribusi normal baku secara asimtotik sehingga

nilai-p dua sisi dihitung sebagai $p = 2 \cdot P(Z > |z_j|)$ dengan $Z \sim N(0,1)$. Interval kepercayaan 95% untuk $\hat{\beta}_j$ dihitung sebagai $\hat{\beta}_j \pm 1{,}96 \cdot SE(\hat{\beta}_j)$, dan interval kepercayaan 95% untuk *hazard ratio* diperoleh dengan mengeksponensialkan batas bawah dan batas atas tersebut, yaitu $\exp\left(\hat{\beta}_j \pm 1{,}96 \cdot SE(\hat{\beta}_j)\right)$ [27, 30]. Seluruh estimasi koefisien, galat baku, nilai-p, dan interval kepercayaan pada penelitian ini dihitung menggunakan pustaka `lifelines.CoxPHFitter`, yang secara default menangani observasi dengan waktu *event* yang identik (*ties*) menggunakan metode aproksimasi Efron.

Hazard Ratio (HR) merupakan interpretasi utama koefisien model Cox. Untuk kovariat kategorik (seperti keanggotaan kluster), HR membandingkan *hazard* suatu kelompok terhadap kelompok referensi:

$$\text{HR} = \frac{h(t \mid x_p = 1)}{h(t \mid x_p = 0)} = e^{\beta_p} \tag{2.14}$$

Nilai $HR > 1$ mengindikasikan peningkatan risiko relatif terhadap kategori referensi, sedangkan $HR < 1$ mengindikasikan efek protektif. Dalam penelitian ini, kluster Metabolik dipilih sebagai kategori referensi karena merepresentasikan pasien DMT2 dengan keterlibatan organ minimal sehingga memberikan interpretasi dasar yang paling bermakna secara klinis sebagai kluster DMT2 ”tanpa komplikasi organ mayor”.

E Uji Asumsi Proportional Hazards

Asumsi mendasar model Cox adalah bahwa rasio *hazard* antara dua kelompok bersifat konstan sepanjang waktu (*proportional hazards assumption*). Secara formal, asumsi ini menyatakan:

$$\frac{h(t \mid \mathbf{x}_1)}{h(t \mid \mathbf{x}_2)} = \exp\left(\beta^T(\mathbf{x}_1 - \mathbf{x}_2)\right) = \text{konstan terhadap } t \tag{2.15}$$

Pengujian formal dilakukan menggunakan residual Schoenfeld, yang diregresikan terhadap fungsi waktu. Jika koefisien regresi ini berbeda secara signifikan dari nol, asumsi PH dilanggar [32]. Secara matematis, residual Schoenfeld untuk kovariat j pada waktu *event* ke-i didefinisikan sebagai:

$$r_{ji} = x_{ji} - \hat{E}(x_j \mid \mathscr{R}(t_i)) \tag{2.16}$$

di mana $\hat{E}(x_j \mid \mathscr{R}(t_i))$ adalah nilai ekspektasi berbobot dari kovariat j atas seluruh subjek dalam himpunan risiko $\mathscr{R}(t_i)$ [33].

Di bawah asumsi PH, residual Schoenfeld seharusnya tidak berkorelasi dengan waktu. Korelasi yang signifikan ($p < 0{,}05$) antara residual Schoenfeld dan fungsi waktu (diuji dengan transformasi *rank* pada deret waktu) mengindikasikan pelanggaran PH [34].

2.1.6 Indeks Keparahan Klinis

Selain pola komorbiditas dan luaran mortalitas, penelitian ini juga mengevaluasi profil keparahan klinis pada setiap kluster. Evaluasi ini diperlukan agar perbedaan prognosis tidak hanya dijelaskan melalui hasil ARM dan model *survival*, tetapi juga melalui indikator klinis yang merepresentasikan beban penyakit pasien. Indeks yang digunakan dalam penelitian ini mencakup skor SOFA untuk menggambarkan disfungsi organ pada subpopulasi ICU dan CCI untuk menggambarkan beban komorbiditas kronis.

A Sequential Organ Failure Assessment (SOFA)

Skor SOFA adalah instrumen penilaian keparahan penyakit yang dirancang khusus untuk pasien ICU, menguantifikasi disfungsi pada enam sistem organ: respirasi (rasio PaO_2/FiO_2), koagulasi (jumlah trombosit), hepatik (bilirubin), kardiovaskular (kebutuhan *vasopressor*), neurologis (skala *Glasgow Coma Scale*/GCS), dan ginjal (kreatinin atau *output* urine) [35]. Skor total SOFA berkisar antara 0–24, dengan nilai yang lebih tinggi mengindikasikan keparahan disfungsi organ yang lebih besar [36]. Dalam basis data MIMIC-IV, skor SOFA tersedia sebagai tabel turunan (*derived table*) `mimiciv_3_1_derived.sofa` yang dihitung per `stay_id` ICU.

B Taksonomi Data Hilang dan Implikasinya terhadap Skor SOFA

Ketidaktersediaan data (*missing data*) dalam analisis statistik secara konvensional diklasifikasikan ke dalam tiga mekanisme menurut kerangka Rubin:

- *Missing Completely at Random* (MCAR), yaitu kondisi ketika probabilitas suatu nilai hilang sama sekali tidak berkaitan dengan nilai variabel itu sendiri maupun variabel lain yang teramati;

- *Missing at Random* (MAR), yaitu kondisi ketika probabilitas suatu nilai hilang dapat dijelaskan sepenuhnya oleh variabel lain yang teramati dalam *dataset*; dan
- *Missing Not at Random* (MNAR), yaitu kondisi ketika probabilitas suatu nilai hilang berkaitan secara langsung dengan nilai variabel tersebut seandainya dapat diamati kemungkinan besar berbeda secara sistematis dari nilai yang benar-benar teramati [29].

Klasifikasi ini penting karena menentukan apakah metode imputasi dapat diterapkan secara valid: imputasi umumnya valid di bawah asumsi MCAR atau MAR, tetapi berpotensi menghasilkan estimasi yang bias secara sistematis di bawah mekanisme MNAR, kecuali model imputasi secara eksplisit memodelkan mekanisme MNAR tersebut.

Ketidaktersediaan skor SOFA dalam penelitian ini bersifat struktural, bukan MCAR maupun MAR konvensional: skor tersebut secara definitif tidak ada (*undefined*), bukan sekadar tidak terukur, pada pasien yang tidak pernah menjalani perawatan ICU. Hal ini terjadi karena komponen penyusun SOFA seperti kebutuhan *vasopressor*, rasio PaO_2/FiO_2 dari analisis gas darah arteri, dan skor GCS berkala merupakan parameter yang secara rutin hanya dipantau pada tingkat perawatan intensif dan tidak dikumpulkan pada perawatan rawat inap nonICU. Dengan demikian, absennya nilai SOFA bukan disebabkan oleh kegagalan pengukuran pada pasien yang seharusnya memiliki nilai tersebut, melainkan karena nilai tersebut secara konseptual tidak terdefinisi pada konteks perawatan pasien yang bersangkutan. Kondisi inilah yang mendasari keputusan metodologis untuk tidak melakukan imputasi, termasuk *multiple imputation* terhadap skor SOFA yang hilang: melakukan imputasi pada pasien nonICU akan berarti membangkitkan nilai bagi suatu kuantitas yang secara fisiologis tidak pernah diukur maupun didefinisikan bagi pasien tersebut sehingga nilai imputasi yang dihasilkan tidak akan merepresentasikan realitas klinis pasien, melainkan artefak dari model imputasi itu sendiri. Implikasi metodologis dari keputusan ini terhadap interpretasi profil keparahan antarkluster dibahas lebih lanjut pada Bab 3 dan Bab 4.

C Charlson Comorbidity Index

Charlson Comorbidity Index (CCI) adalah indeks komorbiditas berbobot yang dikembangkan untuk memprediksi mortalitas satu tahun berdasarkan ada tidaknya 17 kondisi kronis, masing-masing diberi bobot mulai dari angka 1

berdasarkan kekuatan asosiasinya terhadap mortalitas [37, 38]. Total skor CCI dihitung sebagai:

$$\text{CCI} = \sum_{k=1}^{17} w_k \cdot I_k \tag{2.17}$$

di mana w_k adalah bobot kondisi ke-k dan $I_k \in \{0,1\}$. Sebagai contoh, infark miokard dan gagal jantung kongestif masing-masing diberi bobot 1, sedangkan penyakit hati sedang/berat diberi bobot 3, dan metastasis tumor atau AIDS diberi bobot 6 [3]. Dalam MIMIC-IV, CCI tersedia melalui tabel `mimiciv_3_1_derived.charlson` yang dihitung per hadm_id.

2.1.7 MIMIC-IV sebagai Sumber Data Penelitian

MIMIC-IV (*Medical Information Mart for Intensive Care*, versi 4) adalah basis data perawatan kritis *open-access* yang dikembangkan oleh MIT Laboratory for Computational Physiology bekerja sama dengan *Beth Israel Deaconess Medical Center* (BIDMC), Boston, Massachusetts [9]. Basis data ini tersedia melalui platform PhysioNet setelah mendapatkan kredensial akses melalui platform PhysioNet dan dapat diakses menggunakan Google BigQuery. MIMIC-IV v3.1 mencakup 364.627 pasien, 546.028 admisi rawat inap, dan 94.458 admisi ICU, dengan data yang mencakup rentang tahun 2008–2022. Modul utama yang relevan untuk penelitian ini adalah:

- `mimiciv_3_1_hosp`: Berisi data administratif dan klinis rawat inap termasuk tabel `diagnoses_icd` (diagnosis per admisi), `admissions` (informasi admisi), `patients` (demografi termasuk tanggal kematian `dod`), `prescriptions` (medikasi), dan `labevents` (hasil laboratorium).
- `mimiciv_3_1_icu`: Berisi data spesifik ICU termasuk tabel `icustays` (informasi masa rawat inap ICU per `stay_id`).
- `mimiciv_3_1_derived`: Berisi tabel turunan yang telah dikomputasi, termasuk `sofa` (skor SOFA per `stay_id`) dan `charlson` (CCI per `hadm_id`).

Waktu *survival* dalam penelitian ini dihitung menggunakan konvensi MIMIC sebagai selisih antara tanggal kematian (`dod`) atau tanggal sensor (`COALESCE(dod, '2200-01-01')`) dengan tanggal admisi pertama (`admittime`) menggunakan fungsi `DATE_DIFF`, dalam satuan hari.

Dengan karakteristik tersebut, MIMIC-IV v3.1 tidak hanya berfungsi sebagai sumber data klinis, tetapi juga sebagai basis yang memungkinkan integrasi antara informasi diagnosis, derajat keparahan pasien, terapi, biomarker laboratorium, dan luaran mortalitas. Kode diagnosis ICD-10 dapat digunakan untuk membentuk representasi komorbiditas pasien, sedangkan data admisi, tanggal kematian, skor SOFA, CCI, hasil laboratorium, dan medikasi dapat digunakan untuk mengevaluasi relevansi klinis serta prognostik dari pola komorbiditas yang ditemukan. Oleh karena itu, MIMIC-IV v3.1 menyediakan kerangka data yang sesuai untuk menghubungkan analisis pola asosiasi komorbiditas dengan analisis ketahanan hidup dan interpretasi klinis pasien DMT2.

Namun, ketersediaan data yang komprehensif belum secara otomatis menjawab bagaimana pola komorbiditas pada pasien DMT2 terbentuk, bagaimana pola tersebut dapat dikelompokkan menjadi profil klinis tertentu, dan sejauh mana pola tersebut berkaitan dengan prognosis pasien. Untuk memperjelas posisi penelitian ini, bagian berikutnya membahas penelitian-penelitian terdahulu yang relevan serta kesenjangan penelitian yang masih belum terjawab.

2.1.8 Penelitian Terdahulu dan Kesenjangan Penelitian

Berdasarkan landasan teori dan karakteristik data yang telah dijelaskan sebelumnya, penelitian mengenai komorbiditas pasien DMT2 dapat dikembangkan melalui beberapa pendekatan, yaitu identifikasi pola diagnosis berbasis ICD-10, pembentukan kelompok pasien berdasarkan pola komorbiditas, serta evaluasi prognosis menggunakan analisis ketahanan hidup dan indikator keparahan klinis. Sejumlah penelitian terdahulu telah menggunakan pendekatan tersebut secara terpisah, misalnya melalui *association rule mining*, analisis jaringan penyakit, indeks komorbiditas, biomarker klinis, maupun regresi Cox. Namun, belum semua penelitian mengintegrasikan pendekatan-pendekatan tersebut dalam satu kerangka analisis pada pasien DMT2 berbasis MIMIC-IV v3.1. Ringkasan penelitian terdahulu yang relevan dengan penelitian ini disajikan pada Tabel 2.1.

Tabel 2.1. Kesenjangan penelitian

Penelitian	Dataset	Metode Penelitian	Kesenjangan Penelitian
S. M. Bramesh et al. [6]	MIMIC-III dan MIMIC-IV	ARM dengan Apriori, FP-Growth, dan FP-Max	Mengidentifikasi pola penyakit, tetapi belum mengevaluasi hubungan pola komorbiditas dengan *survival*, mortalitas, dan profil keparahan klinis.
M. Wang et al. [7]	MIMIC-IV, eICU, COHD	Apriori, *disease network construction*, dan analisis *organ crosstalk*	Menunjukkan koeksistensi diagnosis, tetapi tidak menguji implikasi prognostik melalui Kaplan-Meier atau regresi Cox.
X. Zheng et al. [8]	MIMIC-IV	Kaplan-Meier, *restricted cubic splines*, dan regresi Cox	Berfokus pada satu biomarker dan tidak menggunakan ARM untuk menemukan pola komorbiditas berbasis ICD-10.
S. Cha & S. S. Kim [10]	KNHDS	ARM dengan Apriori	Menggunakan ARM untuk pola komorbiditas, tetapi tidak mengembangkan kluster klinis dan analisis *survival*.
R. Jiao et al. [11]	HIS dari tiga pusat medis PLA General Hospital	Apriori, korelasi Spearman, dan regresi kuantil terstratifikasi usia	Memberikan bukti asosiasi *cross-sectional*, tetapi belum memakai data MIMIC-IV v3.1 dan belum mengaitkan pola komorbiditas dengan mortalitas.

Berdasarkan Tabel 2.1, dapat disimpulkan bahwa penelitian sebelumnya belum secara terpadu menggabungkan analisis pola asosiasi komorbiditas berbasis ICD-10, pembentukan kluster klinis, evaluasi prognosis menggunakan analisis ketahanan hidup, serta interpretasi keparahan klinis menggunakan CCI, skor

SOFA, biomarker laboratorium, dan pola medikasi intensif pada pasien DMT2 dalam MIMIC-IV v3.1. Oleh karena itu, penelitian ini diarahkan untuk mengisi kesenjangan tersebut dengan menempatkan ARM sebagai metode eksploratif untuk menemukan pola asosiasi, kemudian mengevaluasi relevansi klinis dan prognostiknya melalui analisis *survival*.