TCGA 卵巣癌コホートにおける FIGO 病期と生存の関連——単純コックス比例ハザードモデルによるハザード比推定

2122013 植村 優太
2122034 川村 悠介

2025 年 11 月 20 日

概要

本研究は、GDC 提供 TCGA Ovarian Cancer (OV) コホートの臨床サブセットを用い、FIGO 病期三群（早期～中期／ Stage IIIC ／ Stage IV）間の死亡ハザード比を単純なコックス比例ハザードモデルで推定し、その解釈可能性と限界を明らかにすることを目的とする。解析は病期欠損 4 例を除外した 577 例を対象とし、Stage IIIC を参照として二つの指示変数のみを共変量とする最小モデルを採用した。推定・基準累積ハザードの算出・比例ハザード仮定の暫定検証までを Python ライブラリ lifelines により実施し、同時発生に対しては Breslow 近似を用いた。結果、早期～中期群のハザード比は HR=0.655（95%CI 0.443 – 0.970, p=0.035）、Stage IV は HR=1.324（95%CI 1.014 – 1.729, p=0.039）であり、モデルの識別能は C-index=0.563 であった。Schoenfeld 残差に基づく検定では、早期～中期群指示変数に有意な時間依存性（p=0.003）、Stage IV 指示変数に境界的な非比例性（p=0.073）が示唆された。これらより、Stage IV では死亡ハザードの上昇が確認され、早期～中期群では相対リスクの低下が示される一方、病期のみの単純モデルは識別能に限界があり、特に早期～中期群の推定は時間変化の可能性を踏まえた慎重な解釈が必要である。本研究は、病期のみを用いた最小モデルの再現可能なハザード比推定を提示し、その解釈範囲と限界を整理するとともに、追加共変量の導入や病期順序のトレンド検定、感度解析、PH 仮定の可視化検証へとつながる拡張の出発点を示した。

1　第一章 導入

生存時間分析は、発症から死亡、装置の稼働開始から故障、契約締結から解約までといった事象の発生までの時間を対象とする統計的枠組みである。医学、公衆衛生、疫学、工学、保険数理、マーケティングなど応用分野は広く、打ち切り（censoring）や時変共変量の取り扱いを含む独自の課題に応える方法論が発展してきた。観測データは一般に、事象発生時刻あるいは観測打ち切り時刻を表す T、事象が観測されたかどうかを示す指示変数 δ、および説明変数ベクトル x から構成される。

コックス比例ハザードモデルは、ハザード関数に関する準パラメトリックな表現に基づく標準的手法であり、基準ハザード $h_0(t)$ の形状を仮定することなく、共変量の効果 β を推定できる点に特長がある（Cox, 1972）。説明変数ベクトル x に条件づけたハザード関数は、$h(t \mid x) = h_0(t)\ \exp(x^\top \beta)$ と表され、この表現は共変量がハザードに対して乗法的かつ時間に依存しない比率効果を持つことを意味する。ここで β は部分尤度により推定される。

このモデルの大きな利点は、解釈可能性と柔軟性の両立である。各成分 x_j の 1 単位増加に対応するハザード比は $\exp(\beta_j)$ であり、他の共変量を一定に保ったときの相対リスクとして直観的に理解できる。また基準ハザードは非パラメトリックに扱われるため、時間経過に伴うベースラインの変化を特定の関数形に固定しない自由度が確保される。一方で、比例ハザード（PH）仮定、すなわち任意の 2 個体間のハザード比が時刻 t によらず一定であるという前提が破れると、推定や解釈に歪みが生じうる。PH 仮定の診断には、Schoenfeld 残差の時間依存性検定や、時変係数・時変共変量の導入が用いられる（Schoenfeld, 1982）。

実データ解析では、右打ち切りの存在、共変量間の共線性、非線形効果、外れ値や影響点、欠測値の処理、高次元化に伴う過学習の抑制など、推定と推論を不安定化させる要因が複合的に現れる。コックスモデルを信頼して適用するためには、データ理解と前処理、共変量の選定、PH 仮定の検証と必要な拡張、推定結果の妥当性確認を、相互に整合的な流れで実施することが重要である。さらに、推定と検証のプロセスは再現可能性の観点から透明に記録されるべきである。

本研究の目的は、GDC 提供 TCGA Ovarian Cancer (OV) コホート（Genomic Data Commons Portal, 2025）の臨床サブセットを用い、FIGO 病期三群（早期～中期 / Stage IIIC / Stage IV）間の死亡ハザード比を単純なコックス比例ハザードモデルで推定し、その解釈と限界を整理することである。対象データは 581 例（FIGO 病期欠損 4 例）から成り、生存日数（days_to_death）と死亡事象指標（vital_status）を主要変数とする。欠損病期は除外後 577 例を解析対象とし、Stage IIIC を参照カテゴリとして二つの指示変数のみを共変量とする最小構造のモデルを適用する。本稿ではハザード比推定値とその信頼区間の臨床的整合性を中心に記述する。

本論文の貢献は、進行期偏在を示す TCGA OV コホートに対して病期のみを用いた最小モデルのハザード比推定を再現可能な形で提示したこと、単純モデルの識別能（C-index）の限界を明示し病期以外の要因を導入しない場合の解釈可能性とその範囲を整理したこと、さらに今後の拡張（追加共変量、トレンド検定、感度解析、PH 仮定の検証）への接続点を明確化したことである。

さらに暫定的な比例ハザード仮定の検定（Schoenfeld 残差）を実施し、早期～中期群指示変数で有意な時間依存性（p=0.003）、Stage IV 指示変数では境界的（p=0.073）な結果を得た。これにより早期～中期群のハザード比は観測期間を通じて一定とは限らない可能性が示唆され、単純モデルによる解釈は該当係数について時間平均的な指標として慎重に位置づける。正式な可視化や時変効果導入は本稿では扱わず、今後の拡張課題とする。

本論文の構成は次のとおりである。第 2 章では、モデル、記号、仮定、および推定法（部分尤度）を導入し、最小モデル採用方針と診断項目の位置付けを述べる。第 3 章では、実データ解析として GDC 提供 TCGA OV コホート（FIGO 病期・生存時間）を対象に、前処理、モデル当てはめ、暫定的診断、結果の解釈を示す。最後に、まとめとして主要な知見と限界、今後の課題を述べる。

2　第二章 方法

解析対象は n 個体の観測からなり、i 番目の個体の事象時刻を T_i、事象が観測されたかどうかの指示を δ_i、p 次元の共変量ベクトルを x_i と表す。ハザード関数は

$$h(t \mid x_i) = h_0(t) \exp(x_i^\top \beta)$$

と定め、基準ハザード $h_0(t)$ は未知の非負関数として特定化しない。対応する生存関数は

$$S(t \mid x_i) = \exp\{-H_0(t) \exp(x_i^\top \beta)\}$$

であり、ここで

$$H_0(t) = \int_0^t h_0(u)\, du$$

と表される。

係数 β は部分尤度の最大化により推定する。観測された K 個の事象時刻を昇順に $t_{(1)} < \cdots < t_{(K)}$ とし、各時刻におけるリスク集合を $R_k = \{i : T_i \geq t_{(k)}\}$ と定義する。単一事象が逐次発生する場合、部分尤度は

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(x_{(k)}^\top \beta)}{\sum_{j \in R_k} \exp(x_j^\top \beta)}$$

と書け、その対数部分尤度 $\ell(\beta)$ を最大化して推定量 $\hat{\beta}$ を得る。数値的にはニュートン–ラフソン反復を用い、スコア

$$U(\beta) = \partial \ell / \partial \beta$$

と観測情報

$$I(\beta) = -\partial^2 \ell / \partial \beta \partial \beta^\top$$

を計算して更新を行う。推定量の分散共分散行列は $I(\hat{\beta})^{-1}$ で近似され、各係数についてスコア検定と部分尤度の比較（尤度比相当）、および情報行列に基づく正規近似による信頼区間構成を用いる（Cox, 1972）。

同時発生（ties）が存在する場合、部分尤度には Breslow 法を用いる。$\hat{\beta}$ のもとでの基準累積ハザードは Breslow 推定量

$$\hat{H}_0(t) = \sum_{k : t_{(k)} \leq t} \frac{d_k}{\sum_{j \in R_k} \exp(x_j^\top \hat{\beta})}$$

(d_k は $t_{(k)}$ における事象数）により与えられる（Breslow, 1974）。個体 i の生存確率推定は

$$\hat{S}(t \mid x_i) = \exp\{-\hat{H}_0(t)\exp(x_i^\top \hat{\beta})\}$$

と推定される。

比例ハザード仮定の診断には、Schoenfeld 残差の時間依存性検定を用いる。事象時刻 $t_{(k)}$ におけるリスク集合を R_k とし、重み $w_j(\beta) = \exp(x_j^\top \beta)$ を用いて

$$S^{(m)}(t_{(k)};\beta) = \sum_{j \in R_k} x_j^{\otimes m}\, w_j(\beta) \quad (m = 0, 1, 2),$$
$$\bar{x}(t_{(k)};\beta) = \frac{S^{(1)}(t_{(k)};\beta)}{S^{(0)}(t_{(k)};\beta)}$$

を定義する（$x^{\otimes 0} = 1$, $x^{\otimes 1} = x$, $x^{\otimes 2} = xx^\top$）。同時発生がない場合、k 番目の事象個体 $i(k)$ の Schoenfeld 残差は

$$r_k = x_{i(k)} - \bar{x}(t_{(k)};\hat{\beta})$$

で与えられる。Breslow 近似の下で D_k を $t_{(k)}$ の事象集合、$d_k = |D_k|$ とすると、時点 k の（集約）残差は

$$r_k = \sum_{i \in D_k} x_i - d_k\, \bar{x}(t_{(k)};\hat{\beta}).$$

これらはスコアの時点別寄与に一致し、推定量 $\hat{\beta}$ では $\sum_k r_k = 0$ を満たす。真の β_0 と PH 仮定の下で条件付き期待は $\mathbb{E}[r_k \mid R_k] = 0$ となる。成分 j について $\{r_{kj}\}$ を時間の関数 $g(t_{(k)})$ に回帰したときの傾きの検定は、$\beta_j(t) = \beta_j + \gamma_j g(t)$ を導入したモデルにおける $\gamma_j = 0$ のスコア検定に等価であり、比例ハザードの時定数性を評価する根拠となる（Schoenfeld, 1982; Cox, 1972）。本稿では当該検定に基づく確認に留め、関数形の調整や罰則法、欠測値多重代入などの高度な手法は扱わない。

観測デザインは右打ち切りのみを前提とし、病期（三群：早期～中期［Stage IA～IIIB/IIC/IIIA/IIIB を一括］・Stage IIIC・Stage IV）から成る二つの指示変数だけを共変量とする最小コックスモデルを当てはめた。欠損病期 4 例を除外後に解析し、推定・ハザード比算出・基準累積ハザード推定・比例ハザード仮定の暫定検証までを Python ライブラリ lifelines（CoxPHFitter および proportional_hazard_test）の機能を用いて実行した。係数推定は部分尤度最大化、係数の 95% 信頼区間は観測情報行列に基づく正規近似で構成し、モデルの識別能として C-index を報告した。正式な残差プロット等の可視化や追加共変量による調整、トレンド検定・感度解析は範囲外とし、手順はデータ前処理とともに再現可能に記録した。

本章の流れに基づき、次章では前処理、モデル当てはめ、比例性の検証、結果の要約と解釈を順に実施する。

3　第三章 実データ解析

本章では GDC 提供 TCGA Ovarian Cancer (OV) コホート（Genomic Data Commons Portal, 2025）の臨床サブセット（FIGO 病期および全生存時間）を用いた単純なコックス比例ハザードモデルによる病期群のハザード比推定結果を述べる。原データは 581 例、うち FIGO 病期欠損が 4 例のため、病期欠損 4 例を除外した 577 例を解析対象とした。利用変数は case_id（識別用）、figo_stage（病期）、days_to_death（観測開始から事象または右打ち切りまでの日数）、vital_status（1=死亡事象、0=右打ち切り）。

病期分布は進行期偏在が顕著で、Stage IIIC が 410 例、Stage IV が 89 例、早期～中期（Stage IA, IB, IC, IIA, IIB, IIC, IIIA, IIIB をまとめた群）が 78 例であった。イベント（死亡）数は Stage IIIC が 249 例、Stage IV が 69 例、早期～中期群が 28 例。全体イベントは 346 例、右打ち切りは 231 例。全体生存日数中央値は 1004 日（最小 8 日、最大 5481 日）で、進行した病期ほど中央値が短い傾向を示した。

病期を三群（早期～中期群、Stage IIIC、Stage IV）に再分類し、Stage IIIC を参照カテゴリとする二つの指示変数（早期～中期 vs IIIC、IV vs IIIC）を共変量とする基本コックスモデルを適用。追加共変量は導入せず、β の推定とハザード比の解釈に焦点を絞った。推定は部分尤度最大化、lifelines を使用。得られたハザード比は、早期～中期群が HR=0.655（95%CI 0.443 – 0.970, p=0.035）、Stage IV が HR=1.324（95%CI 1.014 – 1.729, p=0.039）であった。

係数符号は臨床的方向性（進行度上昇に伴うハザード上昇、Stage IV の追加的増加）と整合。

モデルの予測性能指標として C-index=0.563。病期のみの単純モデルでは識別能が限定的であり、症例間ばらつき（残存腫瘍量、治療反応、遺伝子変異など）を十分説明していない可能性を示す。比例ハザード仮定は Schoenfeld 残差に基づく簡易検定で、早期～中期群指示変数に有意な時間依存性（p=0.003）、Stage IV 指示変数に境界的非比例性（p=0.073）。従って早期～中期群の HR=0.655 は期間平均の相対リスクとして解釈を限定し、時間的変化の可能性を付記する。時変係数導入や残差プロット可視化は範囲外とし、今後の拡張で補う。

解釈として、Stage IIIC から Stage IV へ進行した症例では死亡ハザードが約 1.3 倍に増加し、Stage IV 病期の予後悪化が確認された。一方、早期～中期群は進行例（IIIC）に比べハザードが約 0.66 倍に低下。ただし早期群の症例数は少なく信頼区間が広いため、過度の細分化や追加共変量調整を伴う精緻な結論は避ける。限界として、病期以外の臨床因子（年齢、残存腫瘍量、治療レジメ）およびゲノム情報を考慮していないことによる交絡の可能性、C-index の低さに起因する個別予測精度の制約、早期群サンプル不足による推定不確実性が挙げられる。今後、必要最小限の臨床共変量を追加した調整モデルと、病期順序の連続的傾向を扱うトレンド検定を補助的解析として検討する。

4　まとめ

本稿では、TCGA OV コホートを用いた病期三群による最小コックスモデルを構築し、死亡ハザード比を推定した。Stage IIIC を参照とした結果、Stage IV は HR=1.324（95%CI 1.014 – 1.729）と増加し、早期～中期群は HR=0.655（95%CI 0.443 – 0.970）と低下し、臨床的方向性と整合した。Schoenfeld 残差に基づく暫定的検定では、早期～中期群で時間依存性が示され、当該係数の解釈は観測期間にわたる平均的指標として位置づけることが適切である。モデル全体の識別能は C-index=0.563 にとどまり、病期のみでは個別予後の判別が限定的であることが確認された。

手法面では、右打ち切りのみを前提に、部分尤度最大化と Breslow 近似を用いて推定を行い、推定・基準累積ハザード算出・比例性の暫定検証までを Python ライブラリ lifelines により実装した。データ前処理と推定手順は再現可能性を重視して記録した。

限界として、病期以外の臨床因子や分子情報を導入していないことによる交絡の可能性、早期～中期群の症例数不足に伴う推定不確実性、低い C-index による個別予測精度の制約が挙げられる。今後は、年齢や残存腫瘍量など必要最小限の臨床共変量を追加した調整モデルの検討、病期順序に対する傾向の検定、Schoenfeld 残差の可視化や時変係数の導入による比例性の詳細検証、簡易な感度解析などを段階的に行うことで、解釈の堅牢性と予測性能の双方を高めることが望まれる。

参考文献

[1] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187 – 220.

[2] Breslow, N. E. (1974). Covariance analysis of censored survival data. Biometrics, 30(1), 89 – 99.

[3] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239 – 241.

[4] Genomic Data Commons Portal (2025). GDC Data Portal (accessed 2025-11-14). https://portal.gdc.cancer.gov/projects/TCGA-OV

表 1　実データ解析主要集計表

項目	早期〜中期群	Stage IIIC (参照)	Stage IV
症例数 (n)	78	410	89
事象数 (死亡)	28	249	69
打ち切り数	50	161	20
事象割合 (%)	35.9	60.7	77.5
ハザード比 HR	0.655	1.000	1.324
95% CI	0.443 – 0.970	—	1.014 – 1.729
p 値	0.035	—	0.039
Schoenfeld 残差 p 値 (PH 検定)	0.003	—	0.073