F-056

一般化加法モデルを用いたシスプラチン誘発性急性腎障害の発症予測

Predicting Cisplatin-Induced Acute Kidney Injury Using Generalized Additive Models

西澤 達哉 [1] 英 彰吾 [1] 亀谷 由隆 [1]
Tatsuya Nishizawa Shogo Hanabusa Yoshitaka Kameya
髙橋 和男 [2] 坪井 直毅 [2] 水野 智博 [2]
Kazuo Takahashi Naoki Tsuboi Tomohiro Mizuno

1 はじめに

現在，がん治療のためにシスプラチンを投与する場合が多い．シスプラチンは有効な抗がん剤であるが，一方で副作用としてシスプラチン誘発性急性腎障害（Cisplatin-induced acute kidney injury，以下 Cis-AKI）のリスクがあることが知られている．予防法として積極的な水分補給が行われているが，依然として発症する確率は高い．そのため，Cis-AKI の早期発見，予測が不可欠であるといえる．

Ohkawa らは Prediction One（`https://predictionone.sony.biz/`）を用いて Cis-AKI の予測モデルを構築した [9]．更に, 英ら [4] は, 6 種類のモデルクラス（XGBoost [2], SVM, L1-SVM [3], k-NN, Simple-ECHO [4]）を用いて予測モデルを構築している．そして英らは，XGBoost を用いた予測モデルが最も高い予測精度を記録することを示した．XGBoost は勾配ブースティングの考え方を取り入れており，一般的に高い予測精度が出やすいとされている．しかし，モデル構造が複雑で予測根拠の解釈が難しいという問題点がある．特に今回のような医療分野での利用を想定する場合，医師がモデルの予測結果を信頼して良いかを判断するために，予測根拠が確認できることは重要である．先行研究では，SHAP（SHapley Additive exPlanations）[7] を使用してモデルの予測根拠の解釈を行ったが，解釈性の高いモデルを使用することでその工程が必要なくなる．

そこで本研究では，解釈性の高い一般化加法モデル（Generalized Additive Models，以下 GAM）[5] および相互作用項を含んだ一般化加法モデル（Generalized Additive Models plus Interactions，以下 GA^{2}M）[6] を用いて予測モデルの構築を行う．また，構築したモデルの解釈を行い，医学的知見との整合性を持っているかどうかを確認する．

以降，本論文では次の構成をとる．まず，2 節で GAM および GA^{2}M の導入を行う．3 節では本研究で使用するデータセットについて述べる．4 節では実験方法について述べる．実験結果は 5 節で示し，最後に 6 節で本論文のまとめを行う．

2 準備

2.1 データセットの定義

ここでは，サイズ N の学習用データセットを $D = \{(\boldsymbol{x}_i, y_i)\}_1^N$ と定義する．$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ は p 個の観測値を持つ説明変数，y_i は目的変数であるとする．

2.2 GAM

GAM は，一般化線形モデル（Generalized Linear Model，以下 GLM）の線形回帰式の部分を非線形な関数の和で置き換えたモデルを指し，以下のように定義される．

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) \tag{1}$$

ここで，$g(\cdot)$ はリンク関数と呼ばれ，モデルの推定を行えるように累積確率を変換するための関数である．この関数を用いて，モデルを回帰や分類等の問題設定に適応させる．

モデルの予測値に対する各説明変数の寄与の度合いは，非線形関数 f_j の値を調べることにより理解できるため，解釈性の高いモデルであるといえる．また，本研究では非線形関数の推定に平滑化スプライン [10] を用いる．$2r-1$ 次（r 階）の平滑化スプラインは

$$\sum_j \{y_j - f(x_j)\}^2 + \lambda \int \{f^{(r)}(x)\}^2 dx \tag{2}$$

を最小化する関数 f である．λ は平滑化パラメータと呼ばれ，データに対する当てはまりの度合いと曲線の滑らかさのバランスを調整する．λ の値を大きくすると関数は直線に近づく．本研究では，λ の選択に UBRE（Un-Biased Risk Estimator）の最小化を使用する．

2.3 GA^{2}M

先述の GAM は加法構造を維持しているため，説明変数間の相互作用を捉えることが難しい．GA^{2}M は，GAM に相互作用項を追加して説明変数間の相互作用を捉えるとともに，予測精度向上を図ったモデルである．そのため，GA^{2}M は以下のように定義される．

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j) \tag{3}$$

本研究では GA^{2}M の学習アルゴリズムの高速な実装として EBM（Explainable Boosting Machine）[8] を用いる．EBM では，勾配ブースティングとバギングを組み合わせて各説明変数に適用する関数を学習する．勾配ブースティングにおいては，説明変数の順序が問題とならないように，低い学習率かつラウンドロビン方式を採用している．ラウンドロビン方式で学習をすることにより，各説明変数に対して最適な関数を学習でき，予測値に対する寄与の度合いを確認することができる．また，EBM では FAST と呼ばれるアルゴリズムを利用し，効率良く予測への影響度が高い相互作用の検出を行う．

3 データセットの概要

予測モデルの構築には，2006 年から 2013 年にかけて藤田医科大学病院で記録された非 ICU（集中治療室に入院していない）患者データ [5] を用いる．予測モデルでは，性別（Sex），体表面積（BSA），年齢（Age），シスプラチンの 1 日最大投与量（Cisplatin dose），血清クレアチニン値（base sCr），血清アルブミン値（baseline albumine），糖尿病既往歴（DM），心

[1]名城大学大学院理工学研究科情報工学専攻
[2]藤田医科大学医学部医学科

[3]L1-SVM は L1 正則化付き損失関数で学習された線形カーネル SVM である．

[4]識別パターンを使用した規則分類器を指す．

[5]本データの使用にあたり，藤田医科大学病院倫理委員会および名城大学倫理審査委員会の承認を得ている．

表 1: 各予測モデルの予測精度.

モデルクラス	再現率	適合率	最大 F 値	AUROC	AUPRC
GAM	0.357	0.400	0.377	0.733	0.289
GA^2M	0.607	0.321	0.420	0.731	0.328
ロジスティック回帰	0.643	0.269	0.379	0.723	0.283
XGBoost	0.607	0.274	0.377	0.723	0.322

血管イベント既往歴（CVD）の 8 つの説明変数から Cis-AKI の CTCAE グレードが 1 以上になるかどうかを予測する．また，今回用いた患者データはシスプラチンの初回投与時のものである．そして，モデルの学習には 2006 年から 2012 年までに記録されたデータ（1014 件）を使用し，モデルの評価には 2013 年に記録されたデータ（226 件）を使用する．

4 実験方法

本研究では，先述のデータセットを用いて予測モデルを構築し実験を行う．モデルクラスには GAM および GA^2M を使用する．各予測モデルの構築方法については以下で述べる．また，構築した予測モデルの解釈を行い，医学的知見との整合性を持つかどうかを議論する．なお，本研究の実験は全て Ryzen Threadripper 3970X（3.7GHz）を使用して行っている．

4.1 GAM

各説明変数に適用する関数の平滑化パラメータをグリッドサーチを用いて探索し，最適な組み合わせを発見する．探索範囲は 10^{-3}，10^{-1}，10^{1}，10^{3}，10^{5} である．また，連続属性を持つ説明変数（BSA，Age，Cisplatin dose，base sCr，baseline albumine）には 3 次の平滑化スプラインを，離散属性を持つ説明変数（Sex，DM，CVD）には 1 次の平滑化スプラインをそれぞれ適用する．

4.2 GA^2M

Optuna [1] を用いてハイパーパラメータの探索を行う．Optuna はハイパーパラメータの最適化を自動で行うフレームワークで，目的関数を最大化または最小化するようなハイパーパラメータの探索を指定回数または指定時間分行う．本研究では，目的関数に最大 F 値[6]を設定し，実行時間（シングルスレッド）を 1 日で固定した．また，独立試行を 10 回行い，その中で最も目的関数の値が高くなったハイパーパラメータの組み合わせをモデルの構築に使用した．なお，目的関数の算出には層化 10 分割交差検証を使用している[7]．

5 実験結果

5.1 予測精度

本研究では，モデルの予測精度の評価に再現率，適合率，最大 F 値，AUROC（ROC 曲線下の面積），AUPRC（適合率-再現率曲線下の面積）の 5 種類の指標を用いる．各予測モデルの予測精度を表 1 に示す．なお，比較のためにロジスティック回帰と XGBoost の結果[8]も記載している．結果より，GAM を用いた場合の方が適合率は高く，Cis-AKI の誤検出が少なくなっていることが分かる．一方で，GA^2M を用いた場合では再現率が高くなり，Cis-AKI の見逃しが少なくなっていることが分かる．また，先行研究で高い予測精度を記録した XGBoost と比較して，ロジスティック回帰を含むいずれのモデルも同等かそれ以上の最大 F 値を記録することが確認できた．このことから，今回対象としたデータセットに対しては加法構造を持つモデルの方が相性が良い可能性があるといえる．そして，GAM の最大 F 値がロジスティック回帰と比較して劣っている要因として，今回ロジスティック回帰では L1 正則化を使用しており，予測値に対して貢献度の低い説明変数の係数が 0 になっているため，これらの説明変数への過学習が起きなかったからだと考えられる．

表 2: 各関数に設定された平滑化パラメータの値.

説明変数名	平滑化パラメータの値
Sex	10^5
BSA	10^5
Age	10^5
Cisplatin dose	10^5
base sCr	10^1
baseline albumine	10^5
DM	10^3
CVD	10^1

なお，医療分野では AUROC を予測精度の評価に用いることがある．Fischer ら [3] は AUROC について，0.9 以上が「精度が高い」，0.7–0.9 が「精度が中程度」，0.5–0.7 が「精度が低い」と述べている．実験結果では GAM，GA^2M 共に AUROC が 0.7–0.9 の範囲であり，精度が中程度であったといえる．また，AUPRC においては GA^2M と XGBoost の予測性能が全般的に高いことがうかがえる．

5.2 各予測モデルの解釈

GAM の各説明変数に適用した関数をプロットした結果を図 1，設定した平滑化パラメータの値を表 2 に示す．また，GA^2M の各説明変数及び相互作用項の寄与度を可視化した結果を図 2 に，各説明変数に適用した関数および相互作用項をプロットした結果を図 3 に示す．各説明変数に適用した関数は，縦軸の値が正である場合，Cis-AKI を発症するという予測に寄与していることを表している．

図 1 と図 3 の結果を見ると，シスプラチン投与量は GAM では 130mg/day を超えたあたりから，GA^2M では 100mg/day を超えたところで Cis-AKI の発症予測に寄与していることが分かる．臨床的には，シスプラチン投与量は 100mg/day が多量投与であると感じる閾値とされている [4]．また，血清アルブミン値は GAM と GA^2M の両方で 3.5mg/dl を下回ったあたりから Cis-AKI の発症予測に寄与していることが確認できる．臨床的には，3.5mg/dl を下回ると Cis-AKI を発症する可能性が高いとされている．以上から，これらの特徴量については GAM と GA^2M の両方の場合で概ね医学的な知見との整合性を持っているといえる．一方で，血清クレアチニン値については GAM と GA^2M で振る舞いがやや異なる点が見られた．この原因については医療的な解釈を含め，今後検討する必要がある．

加えて，GAM では今回非線形な関数として平滑化スプラインを使用したが，表 2 から血清クレアチニン値と心血管イベント既往歴を除いて平滑化パラメータの値が大きく，データを直線で表現していることが分かる．ロジスティック回帰でも GAM 以上の予測精度を記録していることから，今回は GAM では平滑化スプラインのような非線形関数を使用する必要はなかった可能性がある．

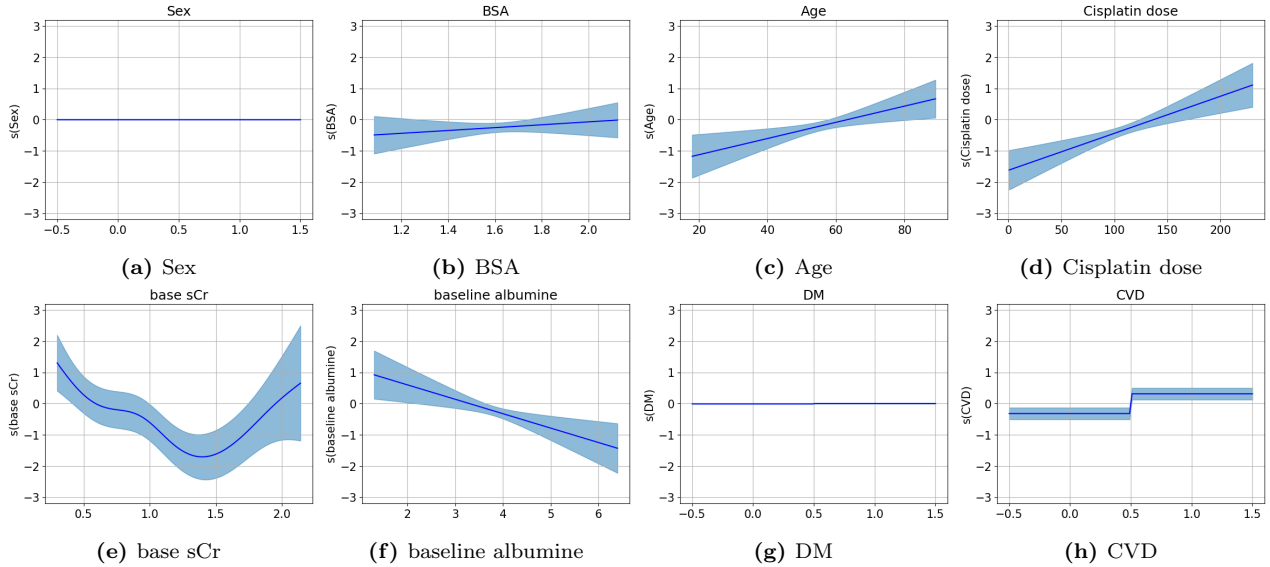
さらに，GA^2M の相互作用項については，図 2 より一定の寄与が確認できる．具体的には，

- シスプラチン投与量と血清アルブミン値：栄養状態が良好であり，全身状態が良好であることが示唆されるため，シスプラチン投与量が増えてリスクが上がる（図 3 (j)）

[6]モデルの決定閾値を変化させた時にとり得る F 値の最大値.

[7]GAM は `https://github.com/dswah/pyGAM`，GA^2M は `https://github.com/interpretml/interpret`，Optuna は `https://github.com/optuna/optuna` で提供されている実装を用いた.

[8]ロジスティック回帰は新たに Optuna で正則化手法と正則化の強さ，最適化手法のチューニングを行った結果で，XGBoost は英ら [4] の実験での結果である.



(a) Sex **(b)** BSA **(c)** Age **(d)** Cisplatin dose

(e) base sCr **(f)** baseline albumine **(g)** DM **(h)** CVD

図 1: GAM で各説明変数に適用された関数をプロットした図．青色で塗りつぶされた領域は 95%信頼区間を表す．

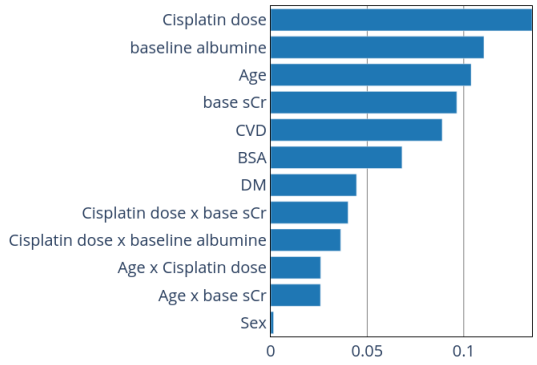


図 2: 各説明変数及び相互作用項の寄与度を可視化した図．

- 年齢とシスプラチン投与量：高齢者にシスプラチン投与量を増やし過ぎるとリスクが上がる（図 3 (k)）
- 年齢と血清クレアチニン値：若年者にも関わらず腎機能が低い患者はリスクが上がる（図 3 (l)）

の間で一部相互作用が確認できるため，これらを考慮できていることが予測精度の向上につながっていると考えられる．医療的な観点からの解釈は引き続き進めていきたい．

6 おわりに

本研究では，GAM および GA^2M を用いて Cis-AKI の発症を予測するモデルを構築し，評価を行った．その結果，両モデルが XGBoost と同程度あるいはそれ以上の予測精度を得て，予測根拠が Cis-AKI に関する医学的知見と概ね一致することを確認した．今後は，モデルの学習に使用するデータを増やすことで更なる予測精度の向上を目指したい．

参考文献

[1] T. Akiba et al.: Optuna: A Next-Generation Hyperparameter Optimization Framework, Proc. of KDD-19, 2019.

[2] T. Chen and C. Guestrin: XGBoost: A Scalable Tree Boosting System, Proc. of KDD-16, 2016.

[3] J. E. Fischer, L. M. Bachmann, and R. Jaeschke: A Readers' Guide to the Interpretation of Diagnostic Test Properties: Clinical Example of Sepsis, Intensive Care Medicine, Vol. 29, No. 7, pp. 1043–1051, 2003.

[4] 英彰悟, 亀谷由隆, 水野智博: シスプラチン誘発性急性腎障害の発症を予測する機械学習モデルの構築と予測根拠の分析, 第 12 回日本医療情報学会「医用人工知能研究会」・人工知能学会「医用人工知能研究会」合同研究会予稿集, SIG-AIMED-012-04, 2022.

[5] T. Hastie and R. Tibshirani: Generalized Additive Models: Some Applications, J. of the American Statistical Association, Vol. 82, No. 3, 1987.

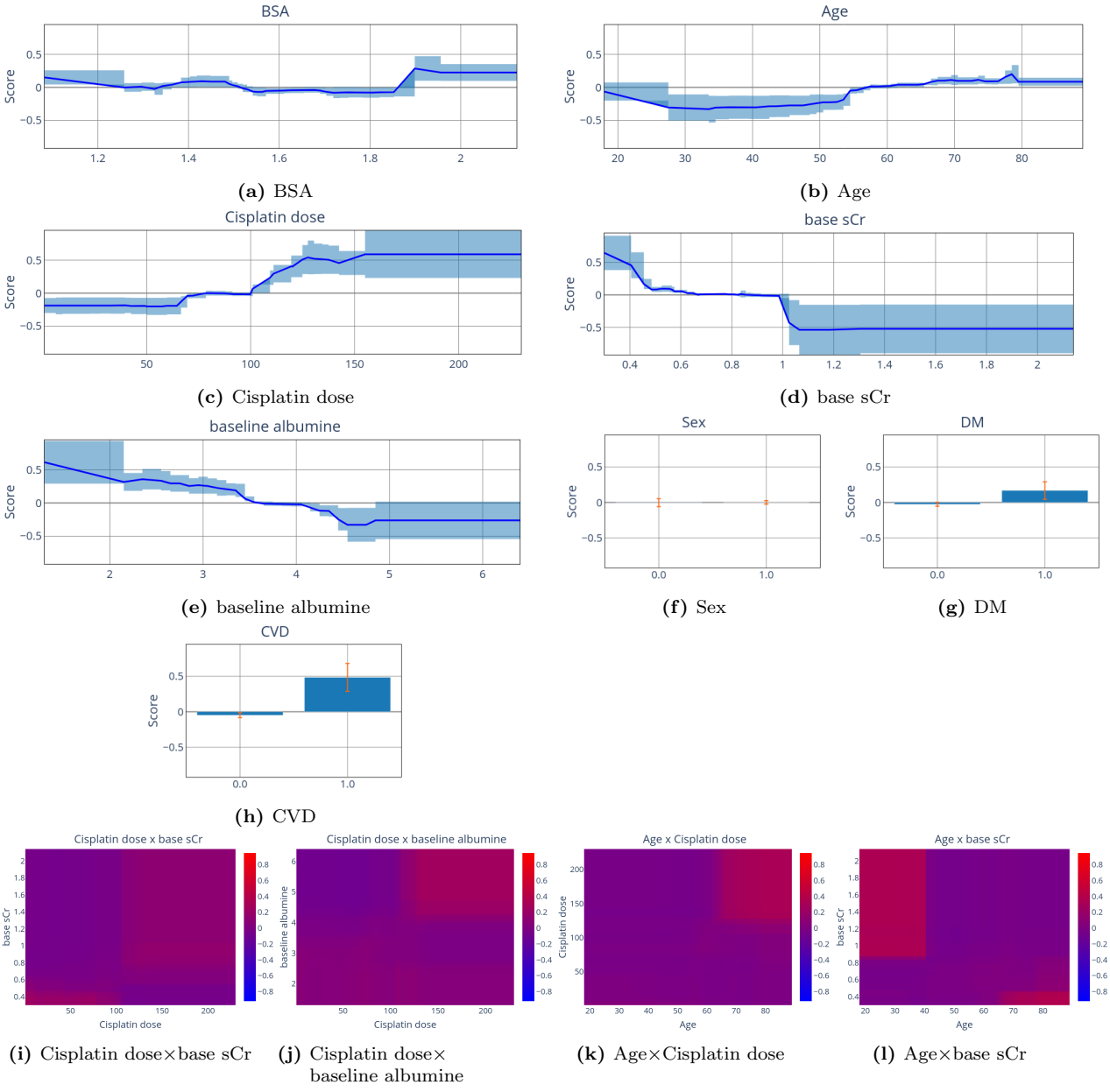
[6] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker: Accurate Intelligible Models with Pairwise Interactions, Proc. of KDD-13, 2013.

[7] S. M. Lundberg and S.-I. Lee: A Unified Approach to Interpreting Model Predictions, Proc. of NIPS-17, 2017.

[8] H. Nori, S. Jenkins, P. Koch, and R. Caruana: InterpretML: A Unified Framework for Machine Learning Interpretability, arXiv:1909.09223, 2019.

[9] T. Okawa et al.: Prediction Model of Acute Kidney Injury Induced by Cisplatin in Older Adults Using a Machine Learning Algorithm, PLOS ONE, Vol.17, No.1, 2022.

[10] C. H. Reinsch: Smoothing by Spline Functions, Numerische Mathematik, Vol. 10, No. 3, pp. 177–183, 1967.



(a) BSA **(b)** Age **(c)** Cisplatin dose **(d)** base sCr **(e)** baseline albumine **(f)** Sex **(g)** DM **(h)** CVD **(i)** Cisplatin dose×base sCr **(j)** Cisplatin dose× baseline albumine **(k)** Age×Cisplatin dose **(l)** Age×base sCr

図 3: GA^{2}M で各説明変数に適用された関数および相互作用項をプロットした図．相互作用項はヒートマップを用いてプロットしている．連続属性を持つ説明変数での青色で塗りつぶされた領域と，離散属性を持つ説明変数でのオレンジ色のバーは，それぞれエラーバーを表している．エラーバーの長さは，GA^{2}M のバギングで得られる複数の予測値から計算される標準偏差である．