＜特集「AIが切り拓く医療の未来」＞

データサイエンス時代の医学研究

川　上　英　良*

国立研究開発法人理化学研究所 医科学イノベーションハブ推進プログラム
健康医療データAI予測推論開発ユニット
千葉大学大学院医学研究院 人工知能（AI）医学
京都府立医科大学大学院医学研究科 統合生理学

Medical Research in the Data Science Era

Eiryo Kawakami

Medical Sciences Innovation Hub Program, RIKEN
Artificial Intelligence Medicine, Graduate School of Medicine,
Chiba University
Department of Physiology and Systems Bioscience,
Kyoto Prefectural University of Medicine Graduate School of Medical Science

抄　録

近年，計測技術の発展と人工知能技術の普及により，医学研究にデータサイエンスの導入が進んでいる．医療データ解析において，タイムスケールや階層性の違いにより，基礎生物学のモデルを直接当てはめることは困難であることが多く，特定の仮説無しに対象疾患の観察とデータ取得を行い，データに基づいて個別化モデルをつくるデータ駆動型アプローチが重要となる．データ駆動型アプローチにおいて，様々な種類の変数や変数間の複雑な依存関係を考慮しながら潜在的なパターンを抽出し，精度の高い予測を行うために機械学習が用いられる．本稿では，卵巣がんの術前診断を例として，教師あり学習と教師なし学習の適用事例を紹介する．教師あり学習によって，卵巣がんの良性・悪性，進行期，組織型といった性質が従来の統計手法に比べて高い精度で予測できることが明らかになった．また，今まで臨床上気づかれなかった術前血液検査のパターンを教師なし学習により発見できた．人間の知識発見と仮説形成をサポートする手段としての機械学習の使い方を紹介し，次世代の医療に向けた新しい医学研究の枠組みを議論したい．

キーワード：医療データ解析，データ駆動型研究，機械学習，予測医療，個別化医療．

平成31年4月29日受付　令和元年5月7日受理
*連絡先　川上英良　〒230-0045 横浜市鶴見区末広町1-7-22　国立研究開発法人理化学研究所 中央研究棟2階C213
eiryo.kawakami@riken.jp
doi:10.32206/jkpum.128.06.397

Abstract

In recent years, with the development of measurement technology and the spread of artificial intelligence technology, data science has been rapidly introduced to medical research. In medical data analysis, it is often difficult to directly apply a basic biological model due to differences in time scale and hierarchy. In such case, a data-driven approach is essential to create individualized models based on comprehensive observation of the target disease without specific hypotheses. In the data-driven approach, machine learning technique is often used to extract potential patterns of diseases considering different types of variables and dependencies among variables, and to perform accurate predictions. In this manuscript, we present supervised learning and unsupervised learning, taking preoperative diagnosis of ovarian cancer as an example. Supervised learning could predict the characteristics of benign / malignant, advanced stage, and histological types of ovarian cancer with high accuracy compared to the conventional statistical method. In addition, unsupervised learning enabled us to discover preoperative blood test patterns that were not clinically noticed until now. We introduce how to use machine learning as a means to support human knowledge discovery and hypothesis formation, and discuss new framework of medical research for the next generation of medicine.

Key Words: Medical date analysis, Date-driven research, Machine learning, Predictive medicine, Individualized medicine.

近年，次世代シーケンサーなどの計測技術の発展と人工知能技術の普及により，急速に医学研究にデータサイエンスの導入が進んでいる．従来の医学研究は，基礎生物学で仮説を積み上げて疾患のメカニズムを解明し，そのコンセプトを大規模なコホートデータで統計的に検証するという，仮説駆動型のスタイルで進展してきた．一部の疾患においては，このような基礎生物学に基づくモデルが極めて良く疾患病態を説明し，革新的な治療に繋がった例もある．しかし，基礎生物学で得られる知見は細胞やモデル生物といったコントロールされた条件のもので，現実の疾患とはタイムスケールや階層も異なるため，そのままヒトの疾患病態に当てはめるのは困難であることが多い．特に，免疫アレルギー疾患や生活習慣病といった慢性炎症性疾患は多因子疾患であり，遺伝的要因と後天的要因が複雑に絡み合って発症に至る．また，がんにおいても組織型や進行期のみならず，腫瘍組織内においてもがん細胞が多様性を持つこと（腫瘍内不均一性）が明らかになっている．このように，単一のモデルに基づく説明や治療が難しい疾患に対して，特定の仮説無しに対象疾患の観察とデータ取得を行い，データに基づいて個別化モデルをつくるアプローチを仮説駆動型研究と対比して「データ駆動型研究」と呼ぶ．

データ駆動型研究では，1．バイアスのないデータ取得，2．データに基づく患者，疾患病態の分類（層別化），3．個別化予測モデルの構築，が主要なステップとなるが，疾患層別化と個別化予測モデル構築において従来の統計的手法の代わりに使われるのが機械学習である．医療データは，正規分布に従わない様々な種類の変数を含み，変数間に複雑な依存関係があるため，従来の統計的手法では，そういった複雑なパターンを抽出することが難しい．機械学習は人工知能研究の中で生まれてきた手法であり，過去のデータの潜在的なパターンをコンピューターに「学習」させ，そのパターンに基づいて新しいデータの予測を行う．機械学習モデルは，単純な統計モデルに比べて自由度が高く，様々な種類の変数を柔軟に扱うことができ，複数の変数の組み合わせに基づいた複雑なパターンを抽出することができる．本稿では，医療データに対する機械学習の適用事例を紹介し，次世代の医療に向けた新しい医学研究の枠組みを議論したい．

様々な機械学習手法

機械学習は目的によって，「教師あり学習」，

「教師なし学習」,「強化学習」に大別される．教師あり学習は複数の観測データを入力として，既知の分類や数値を予測することを目的とする．現実社会においては，メール文中に出てくる単語の頻度から迷惑メールかどうかを予測する，マーケティングデータから売上を予測するといったタスクに使われている．医療においては，画像データに基づいて皮膚がんのタイプを予測するアルゴリズム[1]や，病理画像データから肺がんのサブタイプや遺伝子変異を予測するアルゴリズム[2]が開発されている．一方，予測すべき分類や数値があらかじめ決まっていないデータの潜在的なパターンを抽出するのに使われるのが教師なし学習である．生物学研究でも頻繁に使われる主成分分析（PCA）やクラスタリングも教師なし学習に含まれるが，近年はAutoEncoderなど，ニューラルネットワークに基づいて，画像データの特徴を低次元のベクトルで表現する手法が開発されている[3]．強化学習は，ある環境内において報酬が最も多くなるような行動パターンを試行錯誤の中から学習していく手法である．囲碁や将棋といったルールが明確に決まっているが総当たりでは可能性が膨大すぎるために最適解が決まらないような問題と相性が良い．医療においては，医師の診療における意思決定プロセスを学習させるのに使われることが多い[4]．また，逆強化学習という手法によって，意思決定プロセスにおいて医師が何を重視しているのかを逆に推定するという試みも行われている[5]．

また，学習に用いられるアルゴリズムも，様々な数理科学，情報科学を背景に発展してきており，教師あり学習に用いられるアルゴリズムだけでも，カーネル法に基づくもの，決定木に基づくもの，ニューラルネットワークに基づくものなど，数多くの手法が開発・改良されている．医学研究に機械学習を導入する際は，予測したいもの，抽出したいものを明確にし，データの数や特性に合った手法を選択する必要がある．ここでは,「教師あり学習」と「教師なし学習」を取り上げ，具体的な適用事例として卵巣がんの術前診断を紹介する[6]．

教師あり機械学習による精度の高い予測

卵巣がんは女性の生殖器腫瘍の中で最も予後が悪いものの一つで，近年卵巣がんによる死亡者数は増加している．卵巣がんは，組織学的に少なくとも五つの型（高異型度漿液性がん，低異型度漿液性がん，類内膜がん，粘液性がん，明細胞腺がん）に分かれ，また世界産婦人科連合（FIGO）の進行期分類で，転移の有無などによって早期がん（ステージⅠ，Ⅱ）および進行がん（ステージⅢ，Ⅳ）に分かれる[7]．治療としては，手術による腫瘍の切除が第一選択として行われるが，化学療法への反応性も比較的良いため，手術の前後に化学療法を行うことが一般的である[8]．化学療法への反応性は，進行期や組織型によって大きく異なるのに加え，近年PARP阻害薬や抗体医薬などの有効な抗がん剤が登場してきたこともあり[9]，術前に進行期や組織型を予測し，患者ごとに適切な治療戦略を策定することが強く望まれている．

ここでは，東京慈恵会医科大学産婦人科において，2010～2017年に治療された334名の悪性卵巣腫瘍患者と101名の良性卵巣腫瘍患者のデータを用い，診断時の年齢と術前にルーチンで行う血液検査データ32項目（CA125，CA19-9などの腫瘍マーカーに加え，CRP，LDH，アルブミンなど）に基づいて良性・悪性の予測を行った．教師あり学習としては，サポートベクターマシン（SVM），ランダムフォレスト（RF），ニューラルネットワーク（NN），ナイーブベイズ（NB）といった複数の手法を，従来の統計学的手法である多変量ロジスティック回帰と比較した．その結果，予測の精度の指標となるROC（Receiver Operating Characteristic）曲線のAUC（Area Under the Curve）は，従来の統計学的手法である多変量ロジスティック回帰では0.897だったのに対し，教師あり学習では軒並み0.95を越えた（Fig. 1a）．特に，ランダムフォレストを含む，多数の決定木を組み合わせて学習する方法が非常に精度良く予測できることが分かった．ランダムフォレストは，ランダムサンプリング

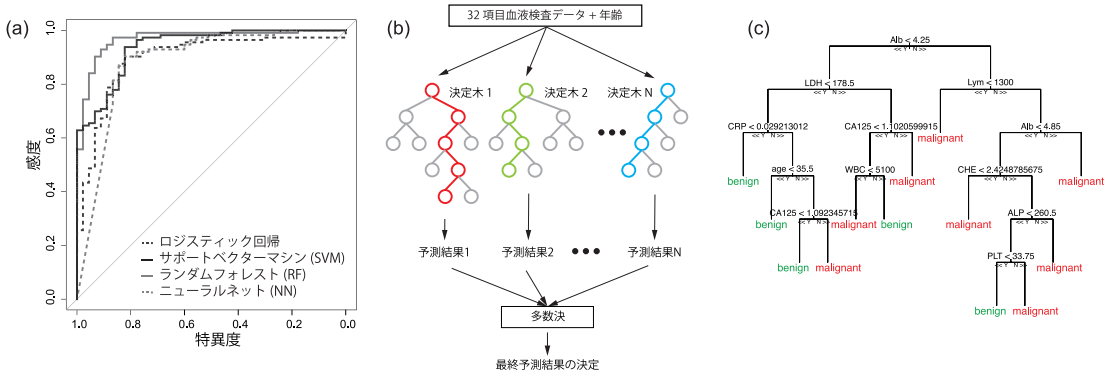

Fig. 1　教師あり学習による疾患分類予測.
(a) 卵巣腫瘍患者に対する悪性と良性の予測結果を示すROC曲線.
(b) ランダムフォレストによる教師あり分類の概略図.
(c) 卵巣がんの良性悪性鑑別のランダムフォレストで作られる決定木の例．このような決定木を数千～数万作り，最終結果を多数決により決定する.

した訓練データと説明変数を用いて，数千～数万の決定木を作り，各決定木の予測結果の多数決もしくは平均を取ることで，最終結果を決定する集団学習アルゴリズムである（Fig. 1b, c）[10]．条件分岐によってデータを分割していく決定木は，外れ値の影響を受けにくく，複数の変数の依存関係を捉えることができるという特徴を持つ．また，統計学的手法において，相関の強い説明変数を用いると，多重共線性という問題が発生して予測モデルが不安定化することが知られている．ランダムフォレストは，少数の変数をランダムに選んで使うことで，これらの問題を回避し，精度の高い予測を可能にしていると考えられる[11]．

さらに，同じ術前血液検査データに基づいて，がんの進行期（早期がんまたは進行がん）や組織型などの予測もランダムフォレストを用いて行った結果，進行期は，AUC=0.760という比較的良い精度で予測することができた（Fig. 2a）．また，既に知られている腫瘍マーカーに加えてCRPとLDHが重要であることが示され，進行期と炎症との関連が示された（Fig. 2b）．また，組織型は，高異型度漿液性がんと粘液性がんの予測精度が比較的良く（AUC=0.785，0.728）（Fig. 2c），高異型度漿液性がんはCA125とCA19-9，粘液性がんはCEAが予測のマーカーとなることが明らかになった（Fig. 2d）．明細胞腺がんと類内膜腺がんの予測精度はそれぞれAUC=0.650，0.597とほとんど他の組織型と鑑別できないことが分かった（Fig. 2c）．これらの組織型については今回の術前血液検査データに関連する指標が含まれておらず，今後術前診断のためには新たなバイオマーカーが必要になると考えられる．

教師なし学習で新たな疾患分類を発見する

進行期予測において，AUC=0.760とそれなりの精度は出たものの，良性・悪性の鑑別に比べて精度が良くなかった．このように，機械学習の予測精度がいまいち良くない，というときに考慮しなくてはならないのが，既知の分類が絶対的なものではないという点である．早期卵巣がんと進行卵巣がんは形態学的な分類で，卵巣がんの場合はリンパ節転移がない場合は早期がん，リンパ節転移，腹膜播種，遠隔転移を認め

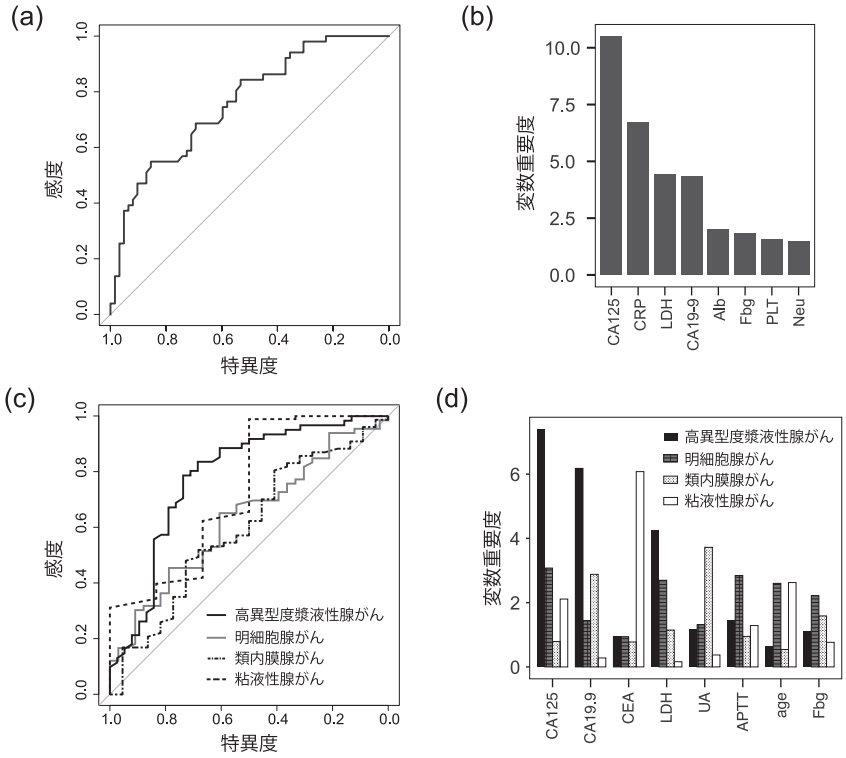

Fig. 2　教師あり学習による卵巣がん進行期および組織型予測
（a）進行期が早期がんか進行がんかの予測結果．AUC=0.760の予測精度だった．
（b）ジニ係数の減少を指標とした，進行期の予測において重要だった変数．血液検査項目のCRPとLDHが重要であることが分かった．
（c）どの組織型に属するかの予測結果を示すROC曲線．高異型度漿液性がんはAUC=0.785，粘液性がんはAUC=0.728，明細胞線がんはAUC=0.650，類内膜腺がんはAUC=0.597の予測精度だった．
（d）ジニ係数の減少を指標とした，組織型の予測において重要だった変数．高異型度漿液性がんは血液検査項目のCA125とCA19-9，粘液性がんはCEAが重要であることが分かった．

る場合は進行がんと分類される．リンパ節転移は小さいものだと見つけにくいこともあるため，早期がんと進行がんというラベリングが不正確である可能性もある．また，術前血液検査という切り口で見たときに早期卵巣がんと進行卵巣がんでパターンが近い症例がある可能性も考えられる．そこで，患者間で術前血液検査のパターンの類似性を評価するために，教師なし機械学習を行った．術前血液検査には，CA125やCA19-9といった腫瘍マーカーやCRPのように，一部の患者が非常に高い値を示す項目が含まれている．このようなデータに対して，PCAなどのユークリッド距離に基づく手法を用いると，一部の患者が外れ値のように離れて分布してしまうという問題が生じる（Fig. 3a）．ここでは，変数の分布に影響を受けず，様々な項目の依存関係を考慮しつつパターンの類似性を評価する手法として，教師なしランダムフォレストという手法を用いる[12]．教師なしランダムフォレストでは，元データを項目ごとにシャッフルする

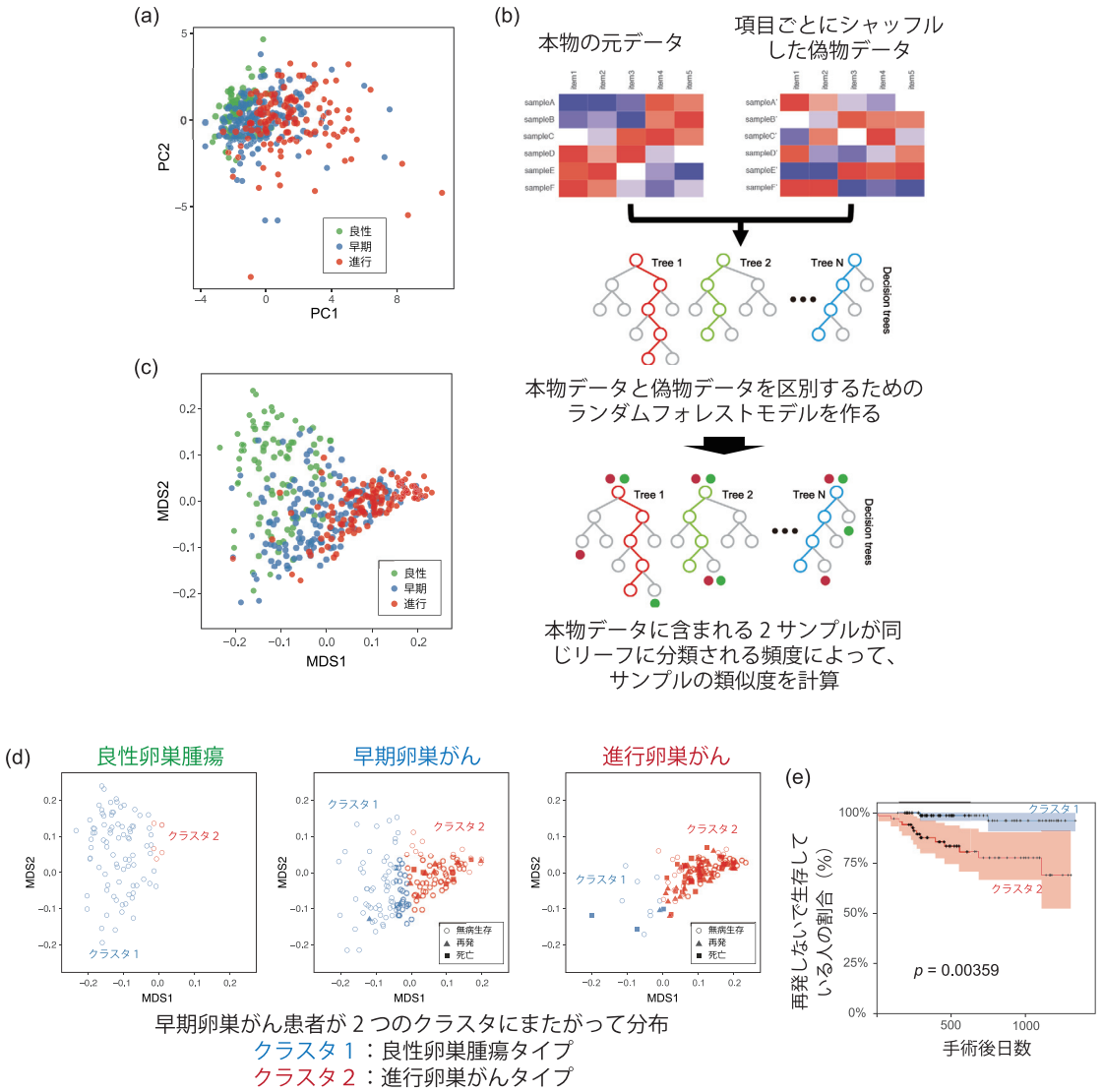

Fig. 3　教師なし学習による新たな疾患分類の発見

（a）主成分分析（PCA）による診断時年齢と術前血液検査データの二次元プロット．一部の進行がん症例が，腫瘍マーカーなどの高値に影響されて外れ値のような分布をとっている．

（b）教師なしランダムフォレストによる疾患パターン抽出の概略図．

（c）教師なしランダムフォレストで計算した症例の類似度を多次元尺度法（MDS）によって二次元に描画したプロット．

（d）良性卵巣腫瘍，早期卵巣がん，進行卵巣がん別のプロット．早期卵巣がん患者は，クラスタ 1（良性腫瘍と良く似た術前血液検査パターンを示す）とクラスタ 2（進行がんに良く似た術前血液検査パターンを示す）に分かれた．

（e）早期卵巣がん患者における，クラスタ 1 とクラスタ 2 の無再発生存曲線．クラスタ 1 ではほとんど再発がなかったのに対し，クラスタ 2 では再発・死亡率が高かった．

ことで，因子のパターンをなくした偽物データを作り，本物データと偽物データを区別するためのランダムフォレストモデルを作る．このランダムフォレストモデル中で，本物データに含まれる2サンプルが同じリーフ（樹形図の末端）に分類される頻度によって，サンプルの類似度を計算する（Fig. 3b）．そして，サンプルの類似度に基づいて似ているサンプルが近くに来るように，MDS（多次元尺度法）やtSNEといった手法によって二次元平面に分布描画する．この手法を診断時の年齢と術前にルーチンで行う血液検査データ32項目に適用したところ，PCAで見られたような外れ値症例がなくなった（Fig. 3c）．このプロットの中で，良性腫瘍は左側の領域に，進行がんは右側の領域に分布しており，良性腫瘍と進行がんは明らかに異なった術前血液検査データのパターンを示すことが分かった（Fig. 3d）．一方，早期がんは広範な分布を示し，「良性腫瘍によく似たパターンを示す集団（クラスタ1）」と「進行がんによく似たパターンを示す集団（クラスタ2）」に分かれた．そして，早期がんの中でクラスタ1に属する症例では再発がほとんどなかったのに対して，クラスタ2では再発率と死亡率が高いという，予後との強い関連を示した（Fig. 3e）．この早期卵巣がんのクラスタは，既に知られている進行期分類（ステージⅠ，Ⅱ）やがんの異型度分類とは異なるもので，術前血液検査データという患者の全身状態を見ることで見つかった，全く新しい分類であった．

ま　と　め

このように，教師あり学習を用いて術前の血液検査データから高い精度で卵巣腫瘍の良性・悪性，進行期，組織型を予測できるようになったことで，手術前に治療方針を決めるのに役立つ情報が得られる．また，予測の際に重要な変数の組み合わせをみることで，卵巣腫瘍の進行期や組織型を特徴づける要素が明らかになった．今後，この研究成果に基づいて，卵巣腫瘍の進行期や組織型ごとの性質を調べる基礎研究や創薬研究が進むことも期待できる．

もう一つ重要なのは，教師あり学習で精度の高い予測が難しかった進行期分類に関連して，今まで臨床上気づかれなかった術前血液検査のパターンを教師なし学習により発見できたという点である．機械学習は，「今までの人間の知識体系を学習する」という使い方をされることが多いが，「今まで臨床医も気づかなかった複雑なパターンを発見する」こともできるのである．今後，機械学習は単なる予測ツールではなく，人間の知識発見と仮説形成をサポートするツールとしてますます活躍していくと考えられる．

開示すべき潜在的利益相反状態はない．

文　　献

1）Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542: 115, 2017.
2）Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature medicine, 24: 1559, 2018.
3）Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. Science, 313: 504-507, 2006.
4）Gottesman, O. et al. Guidelines for reinforcement learning in healthcare. Nature medicine, 25: 16-18, 2019.
5）Asoh, H. et al. in ECMLPKDD2013 workshop on reinforcement learning with generalized feedback.
6）Kawakami, E. et al. Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers. Clinical Cancer Research, 2019.
7）Kurman, R. J., Carcangiu, M. L., Herrington, S. & Young, R. H. WHO classification of tumours of female reproductive organs.（IARC, 2014）.
8）Vergote, I. et al. Neoadjuvant chemotherapy or pri-

mary surgery in stage IIIC or IV ovarian cancer. New England Journal of Medicine, 363: 943-953, 2010.
9) Grunewald, T. & Ledermann, J. A. Targeted therapies for ovarian cancer. Best Practice & Research Clinical Obstetrics & Gynaecology, 41: 139-152, 2017.
10) Breiman, L. Random forests. Machine learning, 45: 5-32, 2001.
11) Kleinberg, E. An overtraining-resistant stochastic modeling method for pattern recognition. The annals of statistics, 24: 2319-2349, 1996.
12) Shi, T. & Horvath, S. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15: 118-138, 2006.

著者プロフィール

川上　英良　Eiryo Kawakami

所属・職：理化学研究所 医科学イノベーションハブ推進プログラム
健康医療データAI予測推論開発ユニット・ユニットリーダー，
千葉大学大学院医学研究院 人工知能（AI）医学・教授，
京都府立医科大学大学院医学研究科統合生理学・客員教授

略　　歴：2007年 3 月　東京大学医学部医学科卒業，医師免許取得
2011年 3 月　東京大学大学院医学系研究科病因病理学専攻博士課程修了，博士（医学）
2011年 4 月　ERATO河岡感染宿主応答ネットワークプロジェクト研究員
2013年10月　理化学研究所　統合生命医科学研究センター疾患システムモデリング研究グループ　特別研究員
2016年 7 月　理化学研究所　医科学イノベーションハブ推進プログラム　疾患機序研究グループ 上級研究員
2017年10月　現職
2019年 1 月　千葉大学大学院医学研究院・人工知能（AI）医学 教授（クロスアポイント）

専門分野：システム医学，予測・個別化医療，機械学習
最近興味のあること：リザーバコンピューティングやバイオコンピューティングなど，新しい機械学習・数理手法
主な業績：

1. Kawakami E, Tabata J, Yanaihara N*, Ishikawa T, Koseki K, Iida Y, Saito M, Komazaki H, Shapiro JS, Goto C, Akiyama Y, Saito R, Saito M, Takano H, Yamada K, Okamoto A., Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers., ***Clin Cancer Res***., DOI: 10.1158/1078-0432.CCR-18-3378, 2019.
2. Kawakami E, Adachi N, Senda T, Horikoshi M., Leading role of TBP in the Establishment of Complexity in Eukaryotic Transcription Initiation Systems., ***Cell Reports***, **21**: 3941-3956, 2017.
3. Kawakami E*, Nakaoka S, Ohta T, Kitano H., Weighted enrichment method for prediction of transcription regulators from transcriptome and global chromatin immunoprecipitation data., ***Nucleic Acids Res***., **44**: 5010-5021, 2016.
4. Kawakami E, Singh VK, Matsubara K, Ishii T, Matsuoka Y, Hase T, Kulkarni P, Siddiqui K, Kodilkar J, Danve N, Subramanian I, Katoh M, Shimizu-Yoshida Y, Ghosh S, Jere A, Kitano H*., Network analyses based on comprehensive molecular interaction maps reveal robust control structures in yeast stress response pathways., ***npj Syst. Biol. App***., **2**: 15018, 2016.
5. Watanabe T, Kawakami E, Shoemaker JE, Lopes TJ, Matsuoka Y, Tomita Y, Kozuka-Hata H, Gorai T, Kuwahara T, Takeda E, Nagata A, Takano R, Kiso M, Yamashita M, Sakai-Tagawa Y, Katsura H, Nonaka N, Fujii H, Fujii K, Sugita Y, Noda T, Goto H, Fukuyama S, Watanabe S, Neumann G, Oyama M, Kitano H, Kawaoka Y., Influenza virus-host interactome screen as a platform for antiviral drug development., ***Cell Host Microbe***, **16**: 795-805, 2014.
6. Kawakami E, Watanabe T, Fujii K, Goto H, Watanabe S, Noda T, Kawaoka Y*., Strand-specific real-time RT-PCR for distinguishing influenza vRNA, cRNA, and mRNA., ***J. Virol. Methods***, **173**: 1-6, 2011.