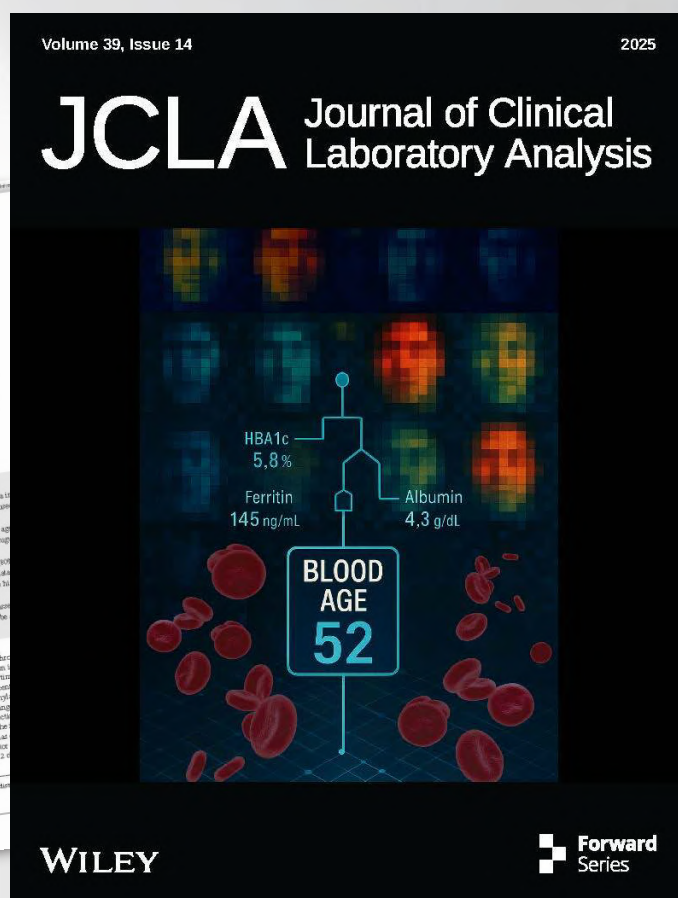


KYBメディカルサービスの 研究論文 が 国際学術誌に 掲載されました！

表紙に選出！



2025年7月

国際学術誌

JCLA Journal of Clinical Laboratory Analysis

Vol.39, Issue 14, 2025

Age Estimation From Blood Test Results Using a Random Forest Model
ランダムフォレストモデルを用いた血液検査結果からの年齢推定

著者

Satomi Kodera (小寺 理美 / 株式会社KYBメディカルサービス)*

Osamu Yokoi (横井 修 / 数理医科学研究センター)

Masaki Kaneko (金子 雅希 / 株式会社KYBメディカルサービス国士館大学体育研究所)*

Yuka Sato (佐藤 由郁 / 株式会社KYBメディカルサービス)*

Susumu Ito (伊藤 孝 / 数理医科学研究センター国士館大学ハイテクリサーチセンター)

Katsuhiko Hata (羽田 克彦 / 数理医科学研究センター国士館大学ハイテクリサーチセンター他)

*はKYBグループ所属



ランダムフォレストモデルを用いた血液検査結果からの年齢推定

Age Estimation From Blood Test Results Using a Random Forest Model

背景と目的

予防医学の観点から、本研究では血液検査で得られるスクリーニングデータを用いて年齢を推定し、加齢や健康課題におけるスクリーニングデータの役割を明らかにすることを目的とした。

あわせて、血液検査に用いられる項目のうち、年齢推定に必要な項目とその数を検証した。

対象と検査方法

2020年2月～2023年8月に全国の医療機関で健診を受けた日本人11,554名（男性3,043名、女性8,511名、0～95歳）を対象とした（表1）。

健診内容は、血液検査60項目・尿検査8項目・唾液検査2項目を含む。

（表1）

Gender	Number	Mean	SD	Q1	Actual age (y.o.)				IQR
					Median	Q3	Minimum	Maximum	
Male	3043	47.95	18.95	34.53	48.52	62.53	5.25	93.60	28.00
Female	8511	49.87	16.47	38.82	49.09	61.40	4.56	95.81	22.58
Total	11554	49.37	17.18	37.90	48.99	61.73	4.56	95.81	23.84

本研究では、複数の決定木を組み合わせて予測を行う機械学習アルゴリズムであるランダムフォレスト法を用いた。データは80% (9,243件) を学習用、20% (2,311件) を検証用に分割し、教師あり学習としてモデルを構築。解析にはPython 3.10.12を使用し、年齢を目的変数、性別および血液・尿・唾液検査の計70項目を説明変数とした。なお、計算コストの観点からハイパーパラメータは最適化せず、デフォルト設定を使用した。

結果

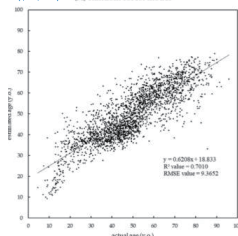
ランダムフォレストモデルと線形回帰モデルの予測精度を比較した結果

ランダムフォレスト： $R^2 = 0.7010$ 、 $RMSE = 9.37$

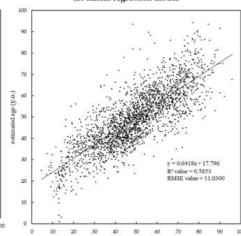
線形回帰： $R^2 = 0.5853$ 、 $RMSE = 11.03$

ランダムフォレストの方が高い精度で年齢推定が可能であった（図1）。これらの結果に基づき、データ数・特徴量の種類・性別差の影響について、ランダムフォレストモデルを用いて詳細に検討した。

（図1） (A) Random forest model



(B) Linear regression model



（データ数の影響）

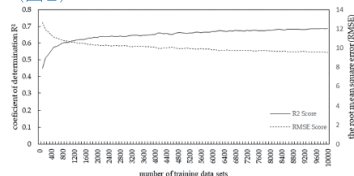
学習データ数を100から10,000まで段階的に増加させて精度を検証した。

100件： $R^2 < 0.45$ 、 $RMSE = 12.65$

800件以上： $R^2 > 0.60$ 、 $RMSE = 10.77$

8,000～10,000件： $R^2 = 0.68 \sim 0.69$ で安定

（図2）



精度の高い推定には8,000件以上のデータが十分であることが示された（図2）。

（特徴量の重要性）

尿検査・唾液検査を除外し、血液検査のみでモデルを構築したところ、 $R^2 = 0.6963$ 、 $RMSE = 9.44$ となり、全項目使用時（ $R^2 = 0.7010$ 、 $RMSE = 9.37$ ）とほとんど差はなかった。さらに、特徴量の重要度を算出し、重要度の低い項目を順次除外した。

15項目まで減少： $R^2 = 0.6937$ （性能ほぼ維持）

10項目まで減少： $R^2 = 0.6570$ （やや低下）

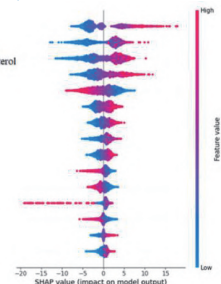
5項目まで減少： $R^2 = 0.5673$ （大きく低下）

年齢推定には、15項目程度で十分な精度が確保可能であることが示された。上位15項目は以下に示す（表2、図3）。

（表2）

Feature	Importance
HbA1c	0.163994
Ferritin	0.134489
Free cholesterol	0.115059
Urea nitrogen (BUN)	0.095047
Albumin (ALB)	0.075884
Alkaline phosphatase (ALP)	0.071181
Mean corpuscular volume (MCV)	0.046920
Red blood cell count (RBC)	0.045560
Copper (Cu)	0.043293
Blood sugar (BS)	0.041645
Free fatty acids (FFA)	0.036925
Alpha 2 globulin (A2G)	0.035531
Platelet count (PLT)	0.035518
Triglycerides (TG)	0.029847
Lactate dehydrogenase (LDH)	0.029100

（図3）



（性別・更年期の影響）

データを男女別に解析したところ、精度は同程度であった（男性： $R^2 = 0.711$ 、女性： $R^2 = 0.695$ ）。一方、女性において50歳前後の散布図に微妙な差異が見られたため、閉経群と月経あり群で比較を行った。T検定の結果、閉経後女性は推定年齢が有意に高く算出される傾向が認められた（ $p = 4.97 \times 10^{-50}$ ）。特に45～55歳の中央値比較では、閉経後女性は8.26歳高く推定された。

考察、まとめ

本研究では、血液検査データを用いて「血液年齢(blood age)」を推定し、既存研究と比較検討した。先行研究において重要とされたアルブミン、グルコース、ALP、尿素、赤血球などは本研究でも上位に入り、さらにフェリチンや遊離コレステロールなど鉄代謝・脂質代謝関連の指標も高い重要度を示した。これらは代謝症候群や慢性炎症、認知症など加齢関連疾患との関連が指摘されている。今後は、生活習慣や遺伝的要因など外的要素も組み込むことで、さらに精度の高い生物学的年齢推定が可能になると考えられる。推定年齢と実年齢の乖離を詳細に解析することで、健康状態や疾患リスク評価に応用できる可能性が期待される。

参考文献

- Hayflick L, Moorhead PS. The serial cultivation of human diploid cell strains. Exp Cell Res. 1961;25:585–621.
 - Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
 - López-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. Cell. 2023;186:243–278.
- ※その他の参考文献については論文本文をご参照ください。