AI と集合知による推定精度の強化: 医師による患者の余命推定を通じた検証

Boosting estimation accuracy via AI and the wisdom of crowds: A study on physicians' patient survival estimation

本田 秀仁[1], 浜野 淳[2], 香川 璃奈[3]
Hidehito Honda, Jun Hamano, Rina Kagawa
[1]追手門学院大学, [2]筑波大学, [3]産業技術総合研究所
[1]Otemon Gakuin University, [2]University of Tsukuba, [3]National Institute of Advanced Industrial Science and Technology
hitohonda.02@gmail.com

概要
本研究では、医師による患者の余命推定課題を通じて、AI および集合知の活用によって人間の推定精度がいかに向上するかを検証した。認知実験の結果、AI の推定を参照することで医師の推定精度が向上する一方、その効果には個人差があった。さらに、計算機シミュレーションによって集合知の効果を検討した結果、AI の推定を参照した後の医師の判断を集約することで、AI 単独を超える精度が得られる可能性が示された。これらの知見は、AI と人間の協調、および集合知を活用した柔軟な推定支援の設計に示唆を与えるものである。
キーワード：医療意思決定、AI、集合知、推定

1. はじめに

近年の AI 技術の発展は、様々な分野に大きな影響を及ぼしている。中でも医療分野においては、従来から存在する多くの課題を解決する手段として、AI 技術への期待が高まっている(Topol, 2019)。特に、医師と AI の協調による医療意思決定は、その精度と効率の両面において向上の可能性を秘めており、実験的な証拠も報告されている(Reverberi et al., 2022)。

本研究では、AI の導入によって、医療文脈における医師の推定精度がどのように改善されるかを検証するとともに、行動科学の知見を活用して、さらなる精度向上のための方法論について議論を行う。本研究が注目する行動科学の知見は集合知（wisdom of crowds）である。集合知には複数の定義が存在するが、本研究では、ある数値的推定に対して、複数の人々が個別に行った推定を集約することで、得られる推定値がランダムに選ばれた個人の推定よりも正確となり、場合によっては集団内で最も優れた個人の推定をも上回る現象を指すものとする(Herzog et al., 2019; Surowiecki, 2004)。また、医療現場においても、患者の治療やケアには多様な職種が連携するチーム医療が不可欠であることが知られており、集合知の重要性は実践を通じて広く認識されているといえる。

集合知の知見を活用することで推定精度をさらに向上させることは、1)AI の適応的活用を妨げるバイアスの問題を補完する、2)医療診断における意見集約による精度向上という 2 つの知見に基づき期待されるものである。

まず 1)について、人には AI の活用を忌避する傾向があることが指摘されており(Burton et al., 2020; Dietvorst et al., 2015)、たとえ AI が優れた推定を提供したとしても、それを十分に参照しない者にとっては、AI の導入による推定精度の向上が期待しにくい。一方で、AI に過度に依存するバイアスの存在も報告されている(Bogert et al., 2021; Logg et al., 2019)。AI は常に正確であるとは限らず、人間の認知特性に基づいた単純なヒューリスティックの方が優れている場合すら存在する(Katsikopoulos et al., 2022)。これらの知見から、AI 以外の方法、特に人間の認知特性を活かす手法を併用することで、より柔軟かつ信頼性の高い推定支援が可能になると考えられる。次に 2)について、医療診断に代表される不確実性の高い判断・意思決定場面において、複数の医師による独立した判断を集約することで、診断精度が大幅に向上することが示されている(Kurvers et al., 2016; Wolf et al., 2015)。これらの研究は、医療のような複雑で判断困難な文脈においても、集合知が効果的に機能し得る可能性を示すものである。

以上を踏まえ、本研究では、AI と集合知の活用によって人間の推定精度がどの程度向上し得るのかを検証する。具体的には、医師による患者の余命推定を対象とした認知実験、および集合知に関する計算機シミュレーションを実施し、検討する。

2. 認知実験: 医師による患者の余命推定

2.1. 概要

本研究では、筑波大学附属病院臨床研究倫理審査委員会(承認番号：R01-108)の倫理委員会の承認を得た上

で、以下に示す課題について医師を実験参加者として、回答を求めた (n = 104)。

実験課題は、ある日の患者の症状(例：現病歴、身体所見、検査結果)を提示し、その患者の余命日数を数値として推定させる形式で構成されている。実験では、以下のような形で 2 回の推定を求めた。1 回目の推定では、医師が患者の症状に基づいて自身の判断による推定を行った。続く第 2 回目では、症状をデータとした時の AI による推定値を提示した上で、それを参照しつつ再度推定を求めた。各実験参加者は合計で 45 名の患者に対する推定が求められた。

本課題で用いたすべての患者データは和文医学論文に掲載されているデータ、すなわち実際の医療データに基づいたものである。提示された症状は匿名化した実在する患者のものであり、余命日数も実際の記録に基づくものである。また、提示された AI の予測値は、臨床現場で使用されている高精度な予測モデルである PaP、PPI、PS-PP (Baba et al., 2015; Yamada et al., 2017)によって算出されたものである。

すなわち、本実験では、実際の症状と医療現場で実際に活用されている AI の予測値を基に構成された刺激に対して、医師が推定を行うという形式で実施した。また実際の余命日数のデータを用いた上で、分析を進めた。

2.2. 分析方針

本研究では、AI の予測や集合知によって推定がどの程度改善されるかを検証することを主たる目的とする。そのため、推定の改善について、以下の指標を用いて分析を行った。まず、本研究では推定誤差は推定値と真値との差の絶対値とした。そして推定のベンチマークとして 1 回目の推定値を用いた。その推定誤差と比較対象の推定誤差を比較し、誤差の相対的な減少割合を指標とした。具体的には、この指標の値が 1 であれば誤差は等しく、1 未満であればベンチマークである 1 回目の推定値より誤差が小さく、推定精度が改善されていることを意味する。逆に、1 より大きければ誤差が増加しており、推定精度が低下していることを示す。以下において、特に断りのない限り、本指標をもって推定精度の改善度を表すものとする。

2.3. 結果

認知実験の結果については、1) AI による推定の精度、2) AI の推定値が医師の 2 回目の推定にどの程度参照されていたか、3) AI の参照度に応じて医師の 2 回目の推定精度がどの程度向上したか、3 点について報告する。

まず AI の予測の精度について、図 1 にその結果を示す。図からもわかるように、多くの患者の予測において AI の予測は医師の予測よりも有意に正確であった[$t(44) = 6.309, p < .001, d = 0.953$]。よって、AI は人間に比べて正確な予測を示していたことがわかった。

次に AI の推定値が医師の 2 回目の推定にどの程度参照されていたかについて、1・2 回目の推定値(est_1, est_2)と AI の推定値(est_{AI})の関係から、参照度を以下のように数値化した。

$$\frac{|est_1 - est_2|}{|est_1 - est_2| + |est_{AI} - est_2|}$$

この指標は、2 回目の推定値が AI の推定値からどの程度影響を受けていたかを示すものである。1 回目と 2 回目の推定値が完全に一致する場合には 0 となり、2 回目の推定値が AI の推定値と完全に一致する場合には 1 となる。また、2 回目の推定値が 1 回目の推定値と AI の推定値の中間に位置する場合には 0.5 となるように定義されている。この指標を各実験参加者(45 名)の推定データに適用し、2 回目の推定の特徴について、ブートストラップ法を用いて分析を行った。その結果を図 2(A)に示す。図中の各点は、ブートストラップ法に基づいて算出された各参加者の平均値および 95%信頼区間を表している。図からも明らかなように、AI の推定値に対する参照の度合いは、参加者ごとに大きく異なっていたことが確認された。

次に、AI の参照度に応じた医師の 2 回目の推定精度の向上について検討した。まず、各医師について AI の参照度に基づき、Self-weighted(95%信頼区間が 0.5 未満に含まれる)、Equal-weighted(95%信頼区間が 0.5 を含む)、AI-weighted(95%信頼区間が 0.5 を上回る)に分類した。そして、それぞれに含まれる医師における推定精度の変化を比較した。その結果を図 2(B)に示す。図中の各点は個々の医師のデータを示し、エラーバーを伴う点は、ブートストラップ法に基づく平均値および 95%信頼区間を表している。図から明らかなように、AI の予測をより強く参照した AI-weighted に分類された医師は推定精度の改善がより顕著であることが示された。

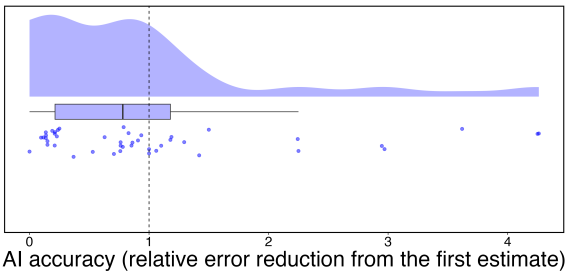

図 1. AI の推定精度。

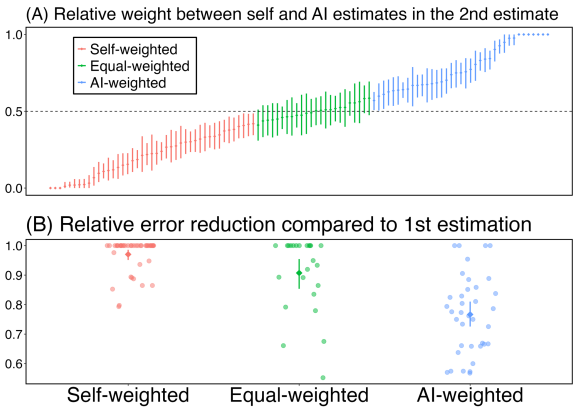

図 2. (A) 各実験参加者の 2 回目の推定における AI 推定の参照度の分布. (B) AI 推定の参照度に応じた 2 回目の推定の精度の分布。

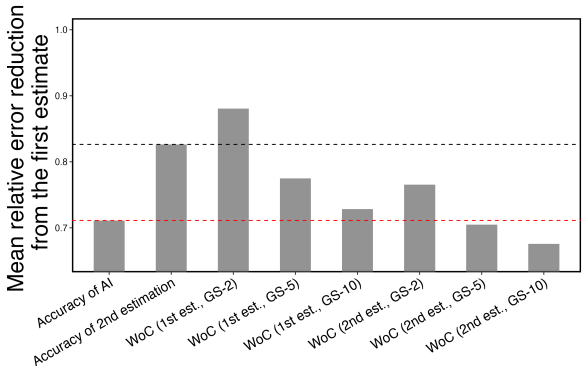

図 3. 集合知の分析. 図中の GS はグループサイズを意味する.

3. 計算機シミュレーション: 集合知が推定精度に与える影響の分析

続いて、集合知によって推定精度がどの程度改善されるかについて分析を行った。ここでは、ある患者に対する推定精度を以下の手続きに基づいて評価した。

まず、集団のグループサイズとして 2 人、5 人、10 人の 3 条件を設定し、それぞれについて医師をランダムに抽出した。抽出された医師の推定値の平均を、そのグループにおける推定値とし、推定誤差を算出した。この手続きを 1 万回繰り返し、その平均値を各グループサイズにおける推定誤差と見なした。

さらに、推定の改善度を評価するために、医師の 1 回目の推定誤差(各医師の 1 回目の推定誤差の平均値)をベンチマークとし、集団の推定値がこのベンチマークと比べてどの程度誤差を減少させるかを検討した。なお、集団の推定値としては、医師の 1 回目の推定値を用いて集約した場合と、2 回目の推定値を用いて集約した場合の双方について比較検討を行った。

結果を図 3 に示す。集合知による推定精度の改善を評価するにあたり、比較対象として AI による推定精度（赤の点線）および医師による 2 回目の推定精度（黒の点線）もあわせて図中に示している。前節で示したように、AI の予測は高い精度を有しており、それを参照することによって医師の 2 回目の推定は改善されるものの、依然として AI の予測精度には及ばない。一方で、推定値を集約することにより、推定精度は大きく改善されることが明らかとなった。たとえば、各医師の 1 回目の推定値であっても、10 名による集約を行うことで、推定精度は AI の予測レベルに近づくことが確認された。さらに、2 回目の推定値を用いて集約を行う場合には、推定精度は一層向上し、5 名あるいは 10 名の推定値を集約することで、AI による予測精度を上回る結果が得られた。

4. 総合討論

本研究では、AI および集合知が人間の推定精度をいかに高めるかについて、医師による患者の余命推定を題材とした認知実験および計算機シミュレーションを通じて検証を行った。実験の結果、AI の予測精度は医師単独による推定よりも高く、また医師が AI の推定を参照することにより、その精度が向上することが示された。加えて、AI を積極的に参照する傾向のある医師ほど、その推定精度の改善が顕著であることが明らかとなった。これにより、AI は単なる補助的情報源ではなく、医療判断における意思決定を実質的に改善し得る存在であることが確認された。一方で、AI の活用効果には個人差が大きかった。

さらに、計算機シミュレーションによって、集合知の効果を定量的に検証した結果、医師の推定を集約することで、個人の判断よりも高い精度が得られることが

示された。特に、AI の推定を参照した後の 2 回目の推定を集約した場合には、AI 単独の精度をも上回る推定が得られることが明らかとなり、医療判断場面のような非常に不確実性の高い課題においても、集合知が有効に機能する可能性を示している。この結果は高性能な AI が医療現場に今後ますます普及したとしても複数の医師・医療従事者による意見の集約が重要であることを示唆している。

本研究の知見は、AI と人間の協働による判断支援の在り方に対して、以下の 2 点の含意を持つと考えられる。第 1 に、AI による支援は、個人の特性や判断傾向に応じて柔軟に運用されるべきであり、単なる介入ではなく、人間の意思決定プロセスに適合したデザインが求められる点である。第 2 に、集合知の枠組みを活用することで、AI に対するバイアスの影響を相殺しつつ、個人では達成困難な推定精度を実現する道が開かれるという点である。

今後の課題としては、推定の集約方法や AI 情報の提示方法が人間のメタ認知など、推定に与える要因について、さらなる検討が必要である。また、医療現場における実装可能性や意思決定プロセスへの統合の方法についても、実証的・制度的検討が求められる。

謝辞

本研究の一部は JST さきがけ (JPMJPR23I3) および科研費(23K16249)の支援を受けた。

文献

Baba, M., Maeda, I., Morita, T., Inoue, S., Ikenaga, M., Matsumoto, Y., Sekine, R., Yamaguchi, T., Hirohashi, T., Tajima, T., Tatara, R., Watanabe, H., Otani, H., Takigawa, C., Matsuda, Y., Nagaoka, H., Mori, M., Tei, Y., Hiramoto, S., … Kinoshita, H. (2015). Survival prediction for advanced cancer patients in the real world: A comparison of the Palliative Prognostic Score, Delirium-Palliative Prognostic Score, Palliative Prognostic Index and modified Prognosis in Palliative Care Study predictor model. *European Journal of Cancer*, *51*(12), 1618–1629.

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*, *11*(1), 8028.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126.

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. H. (2019). The ecological rationality of the wisdom of crowds. In R. Hertwig, T. J. Pleskac, T. Pachur, & The Center for Adaptive Rationality (Eds.), *Taming uncertainty* (pp. 245–262). MIT Press.

Katsikopoulos, K. V., Şimşek, Ö., Buckmann, M., & Gigerenzer, G. (2022). Transparent modeling of influenza incidence: Big data or a single data point from psychological theory? *International Journal of Forecasting*, *38*(2), 613–619.

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., GI Genius CADx Study Group, & Cherubini, A. (2022). Experimental evidence of effective human-AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Doubleday.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS One*, *10*(8), e0134269.

Yamada, T., Morita, T., Maeda, I., Inoue, S., Ikenaga, M., Matsumoto, Y., Baba, M., Sekine, R., Yamaguchi, T., Hirohashi, T., Tajima, T., Tatara, R., Watanabe, H., Otani, H., Takigawa, C., Matsuda, Y., Ono, S., Ozawa, T., Yamamoto, R., … Yamamoto, N. (2017). A prospective, multicenter cohort study to validate a simple performance status-based survival prediction system for oncologists: PS-Based Survival Prediction System. *Cancer*, *123*(8), 1442–1452.