

臨床医学での AI と機械学習 (総説)、2023 NEJM, March 30, 2023

「僻地で世界最先端」西伊豆健育会病院早朝カンファ 2023.4 仲田和正

Artificial Intelligence and Machine Learning in Clinical Medicine, 2023(Review Article)

著者

Charlotte J.Haug, M.D., Ph.D.

University of Oslo, Norway, MSc(master of science, 科学修士), NEJM correspondents

(特派員), Adjunct Affiliate of Stanford Health Policy, Stanford University

Jeffrey M. Drazen, M.D. NEJM Editor-in-Chief

大興奮の総説が NEJM に出ました。

2022 年 11 月、OpenAI 社(米国の人工知能研究所)は Microsoft の支援を得て ChatGPT-4 (Generative Pretrained Transformer 4) を公開しました。

<https://openai.com/blog/chatgpt> (ここでダウンロードできます)

NEJM 先週号 March30, 2023 は早くも「医療関連 chatbot:チャットするロボット」に関する

3 記事を掲載、小生大興奮でした。新たな医療革命が始まったのです。

また NEJM は chatbot の大きな可能性を確信し 2024 年新たな雑誌「NEJM AI」を発行することを決定しました。2023 年夏から論文リサーチを開始し医療の AI を牽引する野心のようです。

僻地離島でも chatbot は間違いなく威力を発揮するでしょう。

今回 NEJM, March30, 2023 に掲載された chatbot 記事は次の 3 本です。本日は以下の 3 記事をまとめました。

A) 臨床医学での AI と機械学習 (総説、Review Article)

B) 医療の AI chatbot、GPT-4 の利点、限界とリスク (特別報告、Special Article)

C) 医学での AI (論説、Editorials)

NEJM, March30, 2023 の「AI と機械学習」最重要点は以下の 7 点です。

- ① 質問の答えがネット上で文書として存在し単純なロジック,計算で可能な時,回答はほぼ常に正しい。
- ② 質問(prompt)の仕方は極めて重要。ゴミを入れたらゴミしか出てこない!
- ③ GPT-4 に感情的質問や、人として接すると幻覚 (hallucination) を起こし回答を間違え。
- ④ Chatbot の回答が事実かフィクションかの区別は困難。校正はユーザーの能力を超える。
- ⑤ 相互の会話で chatbot はユーザーからも学び続け改善されていく。
- ⑥ AI 利用リサーチ:介入は定義可能,計測可能,応用可能であり必ず患者に利益があること。
- ⑦ Bing AI による NEJM 総説のまとめ

AI がどのようなものであるか、上記 B) 論文では次のような謎を GPT-4 に仕掛けています。

「父親とその息子がひどい自動車事故に会い父親は死亡した。その息子は父親が手術を受けようとしている病院へと急いだ。しかしその外科医は言った。『私には手術できない。

あなたは私の息子だ!』一体、どういうこと？」

小生、さっぱりわかりませんでした。GPT-4 の回答は次の通りです。

「その外科医はその少年の母親だ。この謎は、外科医は男性であるはずだという仮定に立っている。典型的ジェンダー (性) 観に基づくものだ」

AI は文脈から正確に判断するのです。

### 1. 質問の答えがネット上で文書として存在し単純なロジック, 計算で可能な時, 回答はほぼ常に正しい。

上記B)「医療の AI chatbot、GPT-4 の利点、限界とリスク (特別報告、Special Article)」の論文では GPT-4 に米国医師国家試験 (USMLE : U.S. Medical Licensing Examination) を解かせました。そして実にその 9 割は正解でした。

即ち「質問の答えがネット上で文書として存在し単純なロジック、数学的計算で可能な時、回答はほぼ常に正しい」のです。

### 2. 質問(prompt)の仕方は極めて重要。ゴミを入れたらゴミしか出てこない！

GPT-4 (Generative Pretrained Transformer 4) は会話形式で文脈をたどりつつ進行して便利であり自然に会話が進行します。

最初の入力を prompt (誘発、ユーザーの指示待ち記号) と言います。GPT-4 は prompt の形式、言い回し (wording) に極めて敏感であり、そのような aspect (横断的处理と実行する場所のこと) の概念を「prompt engineering」と言います。まさに芸術 (art) と科学 (science) の世界なのです。最初の prompt (質問) には、最善の結果を得るために十分な熟慮と注意が必要だと言うのです。「ゴミを入れたらゴミしか出てこない (Garbage in, garbage out)」のです。

### 3. GPT-4 に感情的質問や、人として接すると幻覚 (hallucination) を起こし回答を間違ふ。

ユーザーが懸念や怒りの感情を示すと GPT-4 はそれを推察して会話をそれに合わせようとし極めて興味深い反応をします。GPT-4 が答えられぬような質問をしたり GPT-4 が人間と仮定したりして質問すると、混乱し間違った回答 (false response) をし、これは「hallucination (幻覚)」と言われます。従って感情 (懸念、怒りなど) を示さぬように質問する必要があります。

この回答エラーは医療では極めて危険です。間違いは些細であっても GPT-4 は断定的に回答するため人間はこれを「真実、veracity」と受け止めかねないと言うのです。従って回答 (output) のチェックと事実確認は重要です。回答が本当かわからぬ時「これって本当？」と質問しなおすと、GPT-4 は自分の答えからエラーを発見できることがあります。

上記 B)「医療の AI chatbot、GPT-4 の利点、限界とリスク (特別報告、Special Article)」の論文では GPT-4 を使用してカルテ書き (medical note-taking task)、米国国家試験 (USMLE : U.S. Medical Licensing Examination)、そして curbside consult (マクドナルドのドライブスルー

みたいな気軽な医療相談)を試みました。これらを GPT-4 公開前の試用版と、公開後の 2023 年 3 月の 2 時点で回答(B 論文の Appendix に記載されています)を確認比較したところ最新版では hallucination (幻覚)は生じませんでした。

上記 B)で最後に「AI システムは bias(偏見、先入観)を持ち、医療的判断に悪影響するかもしれない。同意しますか?」と尋ねたところ GPT-4 は感心にも次のように回答しています。

「そう、GPT-4 のような AI システムはバイアスを持ち得ます。これはバイアスのあるトレーニングデータ、アルゴリズム(計算)などに依ります。ですから医療での AI システム使用には偏見の確認と軽減のため、注意深い評価が必要です。」

まとめますと GPT-4 に感情的質問や、人として接すると幻覚(hallucination)を起こし回答を間違えます。

#### 4. Chatbot の回答が事実かフィクションかの区別は困難。校正はユーザーの能力を超える。

問題なのは回答が適切な事実(appropriate fact)に基づくかどうか判らないことです。

医師は chatbot の回答を校正(proofread)する責任(onus)があります。

しかし難しいのはその校正がユーザーの能力を超えている点です。

Chatbot の医療相談が「curbside consult」つまりマクドナルドのドライブスルーのサービスのよう  
に気軽に使用された時、それが真実(veracity)であるかどうか判らないのです。

特に chatbot に診断や治療を尋ねた時、臨床経験のないユーザーにとって事実とフィクションを区別するのは困難です。

例えば治療方法が RCT(randomized control trial)から引用されているならエビデンスレベル A で問題  
ありませんが専門家の意見(Expert opinion)のみとなると大問題です。エビデンス的には専門家の意見  
なんて最低ランクの E です。引用論文が何かは是非確認したいところです。

まとめますと Chatbot の回答が事実かフィクションかの区別は困難です。校正はユーザーの能力を超えます。

#### 5. 相互の会話で chatbot はユーザーからも学び続け改善されていく。

それでも chatbots は医療で今後重要なツールとなると思われます。

上記 B)の論文では GPT-4 を使用して医師と患者との会話からカルテ書き(medical note taking task)を  
させ、実際これは SOAP システムで見事に完成させています。

Chatbots は私たちの業務を改善するでしょうが適切に使われないと損害を被るでしょう。

重要なのは Chatbots はユーザーからも学び続ける点です。

したがって改善にはツールとユーザー間で今後適応期間が必要です。

AI と機械学習は今後、医師がより人に関係する時間を増加させられると思われます。

まとめますと相互の会話で chatbot はユーザーからも学び続け改善されていきます。

6. AI 利用リサーチ: 介入は定義可能, 計測可能, 応用可能であり必ず患者に利益があること。

臨床トライアルで AI と機械学習を利用すれば、より簡便に、またスピードアップできるポテンシャルがあります。より効果的に参加者をリクルート、マッチングしてデータの包括的分析ができると思われまます。薬剤の使用については医師は十分な安全性を要求します。しかし AI と機械学習の場合、介入と記述が正確であるのかははっきりしません。

AI と機械学習によるリサーチの条件には 3 つあります。

- i) リサーチの目的が医療者の行動変容を起こし患者にとって利益、改善があること。
- ii) 介入は定義可能 (definable)、計測可能 (scalable)、現在の問題に応用可能 (applicable) であること。
- iii) アウトカムが該当患者すべてに利益があること。 例えば希少な資源を使用するような公衆衛生的問題をも含むべきかの疑問もある。

まとめますと AI を利用したリサーチに当っては介入は定義可能, 計測可能, 応用可能であり必ず患者に利益がなければなりません。

7. Bing AI による NEJM 総説のまとめ

小生、GPT-4 に今回の NEJM の「臨床医学での AI と機械学習 (総説)、2023」の要約をするよう prompt を書いたところ「GPT-4 は 2021 年 9 月までのデータで回答するため未来のことは判りません」と断られました。

そこで Microsoft の「Bing AI (Bing の検索技術と GPT-4 を組み合わせたチャット AI サービス)」で同じことを要求したところつぎのような回答を得ました。

下記のように引用論文を示してくれるのはとてもありがたいと思いました。

ただ面食らったのは小生のまとめとずいぶん違うなあとと言う点です。

小生は論文を読むとき「自分 (医師) にとって新たな知識、役立つことは何か？」を常に考えてまとめています。Bing AI のまとめだと総論すぎて役に立たないよなあとと思いました。

#### 【Bing AI による NEJM 総説の要約】

「2023 年 3 月 30 日の New England Journal of Medicine に掲載された Artificial Intelligence and Machine Learning in Clinical Medicine, 2023 についてですね。

この論文では、医療現場における人工知能 (AI) と機械学習 (ML) の利用について、その現状と今後の展望が述べられています<sup>4</sup>。論文の要約は以下の通りです。

- AI と ML は、医療現場での診断や治療において、有望な技術である。
- AI と ML は、医療現場での診断や治療において、従来の方法より正確性が高く効率的。
- AI と ML は、医療現場で医療従事者をサポートすることができる。
- AI と ML は、患者のプライバシーやセキュリティを保護することが必要である。

このように、AIとMLは医療現場において非常に有望な技術であり、今後ますます重要性が高まっていくことが予想されます。」

ソース: Bing との会話 2023/4/1

- (1) Artificial Intelligence and Machine Learning in Clinical .... <https://www.nejm.org/doi/full/10.1056/NEJMra2302038>
- (2) Machine Learning in Medicine | NEJM – New England Journal .... <https://www.nejm.org/doi/full/10.1056/NEJMra1814259>
- (3) The New England Journal of Medicine: Archive 2023. <https://www.nejm.org/medical-archives/2023>
- (4) The New England Journal of Medicine (日本国内版). <https://nejm.jp/>
- (5) Artificial Intelligence and Machine Learning in Clinical ....  
[https://www.researchgate.net/publication/369647120\\_Artificial\\_Intelligence\\_and\\_Machine\\_Learning\\_in\\_Clinical\\_Medicine\\_2023](https://www.researchgate.net/publication/369647120_Artificial_Intelligence_and_Machine_Learning_in_Clinical_Medicine_2023)

それでは NEJM, March30, 2023 の「AIと機械学習」最重要点7の怒涛の反復です。

- ① 質問の答えがネット上で文書として存在し単純なロジック,計算で可能な時, 回答はほぼ常に正しい。
- ② 質問(prompt)の仕方は極めて重要。ゴミを入れたらゴミしか出てこない！
- ③ GPT-4 に感情的質問や、人として接すると幻覚(hallucination)を起こし回答を間違えう。
- ④ Chatbot の回答が事実かフィクションかの区別は困難。校正はユーザーの能力を超える。
- ⑤ 相互の会話で chatbot はユーザーからも学び続け改善されていく。
- ⑥ AI 利用リサーチ: 介入は定義可能, 計測可能, 応用可能であり必ず患者に利益があること。
- ⑦ Bing AI による NEJM 総説のまとめ