厚生労働科学研究費補助金（政策科学総合研究事業）
分担研究報告書

実臨床下における人工知能の性能を評価するための研究に関する研究

研究分担者　井上 永介
聖マリアンナ医科大学　医学部医学教育文化部門（医学情報学）　教授

研究要旨

人工知能の開発を行った臨床現場とその実臨床応用を行う現場は乖離することが多く、報告されている診断・予測性能が必ずしも適切なものとは言えない。実臨床下で人工知能の性能を評価した臨床研究を調査し、効率的に実証実験を行う方法を検討する。

Nagendran ら（2020）が行ったシステマティックレビューよると、人工知能を利用したランダム化比較試験は 10 件、日常診療下で評価されたものは 7%であった。このことから、実臨床下で行う人工知能の有用性評価は不十分であり、人工知能開発者と臨床研究の専門家が協業できる場が必要であろうと考えられた。

A．研究目的

既に医療現場で得られているデータを用いてトレーニングされた人工知能（Artificial Intelligence, AI）を臨床現場で利用しようとするとき、実際に有用かどうかを確認するための研究（実証実験など）は大切である。様々な理由のうち、中でも影響が大きいものとして、予測しようとしているラベルのサンプリングに恣意性が入っていることと、臨床現場の時代変化を考慮していないことがある。一つ目は、非常に稀な臨床的イベントを予測しようとするとき、イベントあり群となし群のサンプリングを独立に行っている研究が相当数ある。これにより、イベントが稀という情報が無視されてしまい、AI の学習にうまく取り込めていない可能性がある。結果として、報告された AI の予測性能は実際よりも高く見積もられている可能性がある。

次に、時代変化については、主として AI が対象とする予測問題は日々洗練されていることが原因で生じる。その時点での最良な治療が行われる日常診療は変化し続けており、数年前とは異なった観点で医療が行われることもある。つまり、AI 構築に用いたデータを得た環境と実際に利用される環境が乖離しており、AI として満足できる予測性能が得られていないかもしれない。これらの問題が生じていないかを確認するために、AI を構築したデータとは別の独立した研究を行うことが大切であるが、実際にどの程度実施されているか、どの程度重要視されているかは不明である。

本研究では、日常診療下で行われた臨床研究がどの程度あるか既報をもとに確認する。また、日常診療において AI の有用性評価を行うための効率的な研究デザインを模索することを副次的な目的とする。

B．研究方法

AI の予測性能を実臨床下で評価した論文を Pubmed で検索する。AI 研究に関するシステマティックレビューがあればそれを中心に実態把握を行う。

続いて、検索された文献をもとに、研究現場でどのような問題が生じているかを推測し、前向き、後ろ向き、介入の有無など、適切な研究デザインを検討する。

最後に、既存データを利用して構築された AI の予測性能と、それを実臨床下で運用した場合の予測性能を評価した論文を検索し、効率的に AI を開発するための方策を検討する。

（倫理面への配慮）
本研究には、倫理的考慮を必要とする内容は含まれていない。

C．研究結果

Pubmed 文献検索により、令和元年度末の時点で公表されたシステマティックレビューが存在した（Nagendran et al., BMJ 2020）。本論文の主旨は医療画像に基づく診断能力を、医師と AI で比較した文献のシステマティックレビューであり、実際の臨床現場での比較を行っているかも確認されていた。よって、本論文を中心に検討を進める。

本論文の結果概要を記す。期間は 2010 年から 2019 年までで、データベースとして Embase、Cochrane、WHO trial registry を利用していた。代表的な AI アルゴリズムである深層学習を利用した画像診断の臨床試験が検索されていた。検索の結果、ランダム化比較試験は 10 件、うち 8 件は現在進行中であった。非ランダム化比較試験は 81 件、うち 6 件（7%）のみ実際の臨床現場での評価であった。なお、61 件（75%）が医師と AI の診断性能は同等であると結論していた。加えて、58 件（72%）はバイアスが大きいと判断され、解釈に注意が必要との結果であった。報告の標準形式が利用されていない実態が明白であった。

D．考察

システマティックレビューにより、医療 AI の臨床試験の最新動向を報告した論文（Nagendran et al., BMJ 2020）によると、深層学習分野のランダム化比較試験はわずか 10 件であった。これだけ深層学習が注目されている中で、エビデンスレベルが高いランダム化比較試験の実施数がわずか 10 件という状況は、何か難しい状況があることを伺わせる。実臨床現場での評価が行われているものが 10%に満たないことからも同じことが言える。加えて、報告された論文のバイアスが大きく、解釈に注意が必要なことも問題である。つまり、AI を実臨床下で評価しようとした研究は少なく、かつ質も低いと言える。

これだけ臨床試験が少ない原因の一つとして、AI 構築を進める研究者と臨床試験を進める研究者の背景の違いがあるだろう。AI の専門家は臨床試験に明るいわけではなく、単にバリデーションの一環として現場評価を効率よく行うことを考えている可能性がある。一方、臨床試験の専門家は様々な法律・規制等制約のもと研究を進めなければならないため、効率重視では研究が失敗に終わることを経験してきている。このような背景の相違から、実臨床下での AI 評価があまり進まないのだろうと考えられる。

このような状況下で AI の医学研究を進めるためには、両者が共に研究できる環境を整備し、かつ臨床試験専門家から見て効率のよい研究を推進するべきである。法律などの規制は守るべき要件の最低ラインであり、ここをおろそかにして発展は望めない。また、必ずランダム化比較試験が必要な分野はそう多くないことから、まずは効率を考えて観察研究を中心に進めるのが良いだろう。

E．結論

臨床現場に則した AI 評価を推進するため、AI の研究者と臨床試験の専門家が協業できる環境を整備することが必要である。

F．健康危険情報

該当なし

G．研究発表

1. 論文発表

特になし

2. 学会発表

（発表誌名巻号・頁・発行年等も記入）

特になし。

H．知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得　なし
2. 実用新案登録　なし
3. その他　なし