【助成 39-57】

評価表現分析で読み解く医学論文査読者の嗜好と思考：査読自動化へ向けた基盤研究

研究者　福島県立医科大学附属病院臨床研究教育推進部　副部長・特任准教授　　大前　憲史

〔研究の概要〕

論文の質を複数人の専門家で評価する査読の仕組みは、医学エビデンス創出の基盤を成す。しかし、日本では、臨床医が忙しい臨床の傍ら研究に従事することも多く、査読を学ぶ機会は劇的に不足する。また、査読そのものも、性質上閉鎖的で、実態が不透明な部分も多く、査読教育を行う上での大きな障壁となってきた。本研究は、実証的エビデンスに基づき現行の査読の実態を明らかにし、効果的な査読教育の開発に繋げることを目的とする。そのため、まず、公開査読制を採用する医学研究雑誌の査読レポートを集積し、医学研究のデザインや方法論的枠組みに着目した疫学的分析と査読特有の構造や評価表現に着目した言語学的分析から、査読者の思考や嗜好を明確化する。さらに、得られた知見をもとに、臨床家が論文を評価したり執筆したりする際に容易に参照できるような”Reviewer at Hand”といった支援システムを構築し、医学教育現場への実装を目指す。

〔研究経過および成果〕

研究の第一段階として、査読者コメントを意味的なまとまりに分類し、各パートに意味づけするアノテーションを行った。今回は、公開査読制を採用し、かつ 4 大トップ医学雑誌の１つとして知られる British Medical Journal (BMJ)に焦点を絞り、BMJ に掲載されたランダム化比較試験（RCT）論文に対する査読レポートを解析した。BMJ では一流査読者が査読するため質の高い査読レポートが入手でき、また、RCT は観察研究と比べ方法論として比較的均一で、査読レポートを構造化しやすい利点がある。BMJ で出版された RCT 論文のうち、出版時期が最近のものから適当に 10 本抽出した（うち、2 本は最初の査読結果で Reject と判断されたものを含めるようにした）。査読を受ける前の原稿とそれに対する初回の査読レポートを BMJ のウェブサイトから PDF ファイルで入手し、Python を用いて査読者のコメント部分のみ自動的に抽出できるようにした。その上で、2 本の RCT 論文に対する計 10 名の査読者コメントに対して 2 名の臨床疫学専門家が独立してアノテーションを行った。その後、互いの作業結果を照らし合わせたが想定以上に見解の相違が多く、言語学および教育学の専門家も加わって最終的な合意形成がなされた。

査読者がコメントする対象は元論文の形式に大きく依存するので、コメントの対象場所として、「論文全体」に関するものから、「タイトル」、「抄録」、「背景」、「方法」、「結果」、「考察」に加えて、「図表や補足資料」、「参考文献」や「謝辞」、「その他」に関するものに分類した。さらに、それぞれの下の階層としてより細かな分類を設定し、最も多いものでは、「方法」について「報告の形式」や「研究デザイン」、「解析手法」など、計 38 に細分類した。一方で、コメントの内容については、「挨拶」、「賛辞・評価」、「研究の強みの列挙」、「疑義・懸念」、「査読者の見解」、「統計家の見解」、「改善策の提案」、「重要情報の追記や明確化の提案・依頼」、「不適切・不要な記述の修正や削除

の提案・依頼」、「事務的報告・指示」、「その他」の計 11 の項目に集約された。次に、医学論文の執筆経験および臨床疫学の専門性を有し、上記作業には関わっていない臨床医 2 名に、査読者コメントをこれらの項目に基づき改めて分類してもらった。評価者間の一致割合は平均80％以上と比較的良好な結果が得られ、不一致部分についても評価者間の議論を通して全て解決可能であった。

次に、OpenAIのGPT 3.5-turboを用いて同様の分類作業を行った（分類理由も一緒に出力）。平均一致割合は 50％を下回ったものの、分類理由の多くは妥当と考えられ、プロンプトの工夫やスペックの向上により人の作業を補完できると考えられた。また、実際の査読作業自体での大規模言語モデル活用の可能性も示唆された。そこで、我々は査読作業における大規模言語モデルの有効性を検討した先行研究について文献レビューを実施した。2023 年 10 月に Liang らスタンフォード大学の研究チームが Nature 系雑誌 15 誌からの計 3,096 論文と International Conference on Learning Representations (ICLR)からの計 1,709 論文について GPT-4 を用いて査読を行わせ、人間の査読結果と比較した結果が報告されていた（Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783）。GPT 4 と人間の査読者による指摘の重なりは 30.9～39.2％であったが、人間の査読者間の重なり 28.6～35.3％に匹敵するものであった。さらに、研究者を対象にしたアンケート調査では、回答者の 52.4％が GPT-4 による査読が有用あるいは非常に有用と感じていた。ただし、GPT-4 のトークン数の制限などから、本研究では「意義と新規性」、「受理される可能性のある理由」、「拒否される可能性のある理由」、「改善のための提案」の4項目についてのみ評価された。

これらの結果を踏まえ、我々は、査読において最も重要だが評価が難しい部分である「研究意義の評価」に特化した研究者支援ツールを、作業効率が改善されマルチモーダル対応となった ChatGPT 4o を用いて作成する試みを行った。これまでの調査結果に基づき、①潜在性、②興味深さ、③独創性、④新奇性、⑤効果の大きさ、⑥倫理性、⑦切実性という 7 つの観点で研究の意義を評価し、それぞれの項目における評価とその理由、さらにそれらを踏まえ、論文全体としての良い点と課題についても示されるようにし、査読経験の浅い研究者の利用を想定して簡潔で見やすい出力形式となるよう工夫した。チャット機能を通じ、それぞれの項目について追加で深く掘り下げることも可能である。対象論文をウェブ上で upload し、評価を指示するプロンプトを入れるだけで短時間に結果が返却されるので、操作感について利用者の評価は非常に高く、出力形式や内容の再現性が高いとは言えないという課題は残るものの、内容的の質についても利用者のニーズをある程度満たすものであるとの評価も得た。一方で、重要な課題として、現状、先行研究を独自で調べて現行のエビデンスギャップを踏まえた解釈までは行えず、あくまで論文での記述内容の範囲内での出力に留まる。研究意義を高いレベルで評価するには、既存のエビデンスが不十分であることを把握した上で、その文脈の中で対象論文がどう位置づけられるかというところまで評価できる必要性があり、目下、開発が進む、エージェントフレームワークなどの仕組みが必要となるため、今後は、GPT だけでなく、様々な大規模言語モデルの新たな機能を活用しながら、より高いレベルで出力できるよう

なモデルの開発に向け、引き続き研究を進めていく。

〔成果物〕

GPTs『論文査読用:PIONEER チェック』

https://chatgpt.com/g/g-67c538f9f3d48191adea27379570bcbf-lun-wen-cha-du-yong-pioneertietuku

〔発表論文〕

1. 大前憲史. Cutting Edge 論文解説:Longitudinal Associations between Concurrent Changes in Phenotypic Frailty and Lower Urinary Tract Symptoms among Older Men. 排尿障害プラクティス(メディカルレビュー社) 第31巻第2号. 2023年12月.
2. 大前憲史. 効果的な論文査読を行うために知っておきたい重要な視点. 老年看護学(日本老年看護学会) 第27巻第2号. 2023年1月.

〔発表学会・講演・セミナー〕

1. 大前憲史. DAG(Directed Acyclic Graph)とDAGitty を用いた作図. エコチル調査福島ユニット学術ワーキンググループ勉強会(第 53 回) 2025年2月. オンライン.
2. 大前憲史. まるわかり!医学論文査読のお作法その③ 要旨をまとめ効果的に概要を把握する. Dr's Prime Academia. 2024年11月. オンライン.
3. Ferreira M, Furukawa TA, Tajika A, Omae K. Using the smallest worthwhile difference to interpret clinical evidence. Global Evidence Summit 2024. 2024年9月. Prague.
4. 大前憲史. ミニ講義. 第 11 回臨床研究てらこ屋 in 福島. 2024年9月. オンライン.
5. 大前憲史. 研究デザインの視点から考える医学論文査読の勘所. 第 70 回日本呼吸器学会中国・四国地方会/第62回日本肺癌学会中国・四国支部学術集会 2024年7月. 米子.
6. 大前憲史. まるわかり!医学論文査読のお作法その① 査読ってなに? Dr's Prime Academia. 2024年2月. オンライン.
7. 大前憲史. データベース研究初級者が学ぶ Clinical Question から Research Question へ. 第1回疫学初級ウェビナー. 2024年2月. オンライン.
8. 大前憲史. リサーチクエスチョンにおいて何が最も重要なのか?: Feasible, Interesting, Novel, Ethical, Relevant. 日本臨床疫学会第 6 回年次学術大会. 2023年11月. 東京.
9. 大前憲史. ミニ講義. 第 11 回臨床研究てらこ屋 in 福島. 2023年9月. オンライン.
10. 大前憲史. 量的研究論文、私はこう書く. PCR Connect 2022. 2022年12月. オンライン.
11. 大前憲史. 研究論文のどのパートが最も重要なのか? Introduction, Methods, Results, Discussion. 日本臨床疫学会第5回年次学術大会. 2022年11月. 東京.
12. 大前憲史. 効果的な論文査読を行うために知っておきたい重要な視点. 日本老年看護学会第27回学術集会. 2022年6月. オンライン.