【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

（中間報告）

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、査読に特化した膨大な言語表現データベース（コーパス）を構築する。さらに、コーパスを利用して、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質と量の両面から明確化する。

〔研究経過および成果〕

本研究の計画当初の最も大きな目的は、医学論文査読に特化したコーパスの構築であった。しかし、2022 年 11 月に米国の Open AI 社が ChatGPT をリリースしたことを皮切りに、膨大なテキストデータを用いて訓練された先進的な自然言語処理モデルが次々と登場し、比較的安価で簡単に利用できるようになった。訓練に用いられたテキストデータには医学研究や論文に関するものも多く含まれており、医学分野での応用事例もいくつか報告され、さらに自身の使用感覚も踏まえ、その性能の高さからコーパス構築の必要性は大きくないと考えるに至った。一方で、このような大規模言語モデルを本研究の中でテキストデータの言語解析に活用したり、さらには、実際の査読作業の一部のタスクを担ったりすることもできるのではないかという仮説に発展した。

我々はまず、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 大トップジャーナルの1つとして知られる The 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 では一流の査読者が査読を行うため質の高い査読レポートの入手が期待できた。また、観察研究と比較して、RCT では研究の方法論がある程度決められており、査読レポートも構造化しやすいことが想定される。さらに、RCT は医学以外の領域でも広く用いられる研究デザインであり、得られた知見の汎用性も期待できる。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に合意形成がなされた。

そもそも査読者がコメントする対象は元論文の IMRAD 形式に依存するので、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。さらに、同様の分類作業を、今度は Open AI の API を利用して ChatGPT（モデル：GPT-3.5-turbo）を用い、行った。プロンプト上で査読者としての役割を明示し、分類とともにその理由まで記載するよう指示した。平均一致割合では 50%を下回ったが分類理由の多くは妥当と考えられるものが多く、プロンプトの更なる工夫や利用の仕方により、人の作業を補完できる可能性が高いことがわかった。

この結果からは、実際の査読作業自体においても最新の大規模言語モデル活用の可能性が示唆された。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。その結果、医学研究領域では、2023 年 9 月に、少なくとも比較的簡潔な症例報告の査読では ChatGPT の利用が有効である可能性が Biswas らにより報告されていた（ChatGPT and the Future of Journal Reviews: A Feasibility Study. Yale J Biol Med. 2023;96(3):415-420）。さらに研究領域を広げ、医学以外も含めると、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 2023 年 10 月に報告されている（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ていた。使用したプロンプトも公開されており、我々は医学論文においても同様の作業が行えることを実際に確認した。

これまでの研究で得られた成果は、下記の論文や学会、講演などの中で報告することができた。今後は、引き続き GPT-4 をはじめとする大規模言語モデルを活用しながら、さらに査読者コメントの言語解析を進め、特に統計査読者や患者査読者など、査読者の立場の違いがコメントの対象や内容にどのような影響を与えるかに着目していく。さらに、特に日本の臨床家が論文執筆の際、簡単ですぐに参照できるような Reviewer at Hand といった仕組みを実装できるよう引き続き研究を進めていく。

〔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. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.