2025 年　12 月　10 日

＜研究課題＞ 交通事故に伴う脊椎骨折の見逃しを防ぐための大規模言語モデルを用いた電子カルテスクリーニングシステムの開発

代表研究者　千葉大学医学部附属病院　整形外科学　医員　北村　昂己
共同研究者　千葉大学医学部附属病院　整形外科学　助教　牧　聡
千葉大学医学部附属病院　整形外科学　特任講師　古矢　丈雄
千葉大学医学部附属病院　整形外科学　秘書　山崎　恵美

【抄録】
医師の働き方改革により救急外来で経験の浅い医師が初療を担当する機会が増え、脊椎骨折の見逃しが懸念されている。本研究は、電子カルテ自由記載を用いて脊椎骨折を自動的に判定する仕組みを構築し、感度を主要指標として性能を評価した。2013～2023 年に腰痛や背部痛を主訴として救急受診した症例のうち、MRI で圧迫骨折の有無を確認できた 334 例を対象とした。SOAP の S と O を用いてカルテテキストを整備し、複数の高性能 LLM に simple および詳細プロンプトを与えて判定させた。詳細プロンプトでは感度が各モデルで 0.814 と 0.831 に向上し、AUROC も改善した。さらに、LLM が抽出した関連因子を用いて LightGBM を構築したところ、感度 0.847～0.874、AUROC 約 0.76～0.78 の性能を示した。以上より、カルテ記載に基づく自動スクリーニングは、救急現場での見逃し防止に有用となる可能性がある。

１．研究の目的

医師の働き方改革により救急外来では時間外労働の上限規制が導入され、初期診療を経験の浅い医師が担当する場面が増えている。この状況では診療のばらつきや見逃しが生じやすく、とくに脊椎骨折の見逃しは神経障害や長期的な機能低下につながる可能性がある。本研究では、救急外来の電子カルテ自由記載を基盤とした自動スクリーニングを構築し、脊椎骨折の見逃しを最小化する体制の確立を目指した。院内ローカル環境で運用可能な大規模言語モデル(LLM)を用いて初療記載のテキストから脊椎骨折の有無を推定する方法を検証し、さらに記載内容から抽出された脊椎骨折関連因子を用いて機械学習モデルの性能を評価した。

２．研究方法と経過

2-1 対象患者とデータ作成

対象は 2013 年 4 月～2023 年 12 月に腰痛・背部痛・腰部打撲を主訴として救急外来を受診し、初期研修医が初療を行った患者 6606 例とした。このうち整形外科外来で MRI を撮影し、圧迫骨折あり 183 例、なし 151 例を解析対象とした。電子カルテ記載のうち SOAP 式記載の S と O のセクションを用い、画像所見の直接的な記述は除外した。

2-2 LLM による判定手法の比較

複数の高性能ローカル LLM を用い、①専門医としての役割をプロンプトで与え骨折の有無を判定する simple プロンプト、② ①に加えて専門知識や判断ロジックを体系的に組み込んだ内容をプロンプト内に記載した詳細プロンプトの 2 種類を比較した。各 LLM に、作成したカルテデータから脊椎骨折の有無を判断させ、その精度を算出した。

2-3 LLM 由来特徴量を用いた機械学習モデルの構築

複数のガイドラインから抽出した 13 項目の脊椎骨折関連因子の有無を LLM に判定させ、得られた特徴量を LightGBM に入力した。5 分割交差検証を実施し、AUROC、感度、特異度、Average Precision を指標として性能を評価した。

３．研究の成果

3-1 LLM 単独判断の結果

LLM による直接判定では、感度が詳細プロンプトで顕著に改善した。医学領域向けに特化して調整された LLM では simple プロンプトで 0.4 程度 から詳細プロンプトで 0.8 を超える水準へ、日本語対応の LLM では simple プロンプトで 0.5 程度 から詳細プロ

ンプトで 0.8 を超える水準まで上昇した。救急外来でのスクリーニングを想定すると、この感度の向上は実運用上の意義が大きい。一方、AUROC も詳細プロンプトで上昇し、医学領域向けに特化して調整された LLM は 0.6 程度→0.7 程度、日本語対応の LLM も 0.6 程度→0.7 程度 と全体の識別性能も改善した。特異度は感度向上に伴い低下したが、文字情報のみを用いた判定で高い感度を得られた点は評価できると考えられる。

3-2 LLM 抽出由来の特徴量を用いた LightGBM

LLM に関連因子の抽出を行わせ、LightGBM による機械学習モデルを構築した結果、複数の設定で高感度を維持できた。
医学領域向けに特化して調整された LLM 由来の特徴量では、SelectFromModel 7 変数モデルの感度が 0.8 程度 と最も高く、AUROC も 0.7 程度 を示した。日本語対応の LLM では、13 変数モデルで感度 0.8 を超える水準と最も高く、AUROC は 0.7 程度 であった。特徴量の削減により特異度が上昇する設定もあり、感度を優先しつつ、用途に応じたバランス調整が可能であることが示された。

3-3 総合評価

LLM 単独判定では詳細プロンプトが最も高い感度を示し、スクリーニングとしての適性が確認できた。一方で、詳細プロンプトでは LLM が独自に判断過程を形成するため、判定の根拠が明確に把握できない点が課題として残る。これに対し、LLM に関連因子を抽出させて機械学習モデルに入力する方法では、感度 0.8 を超える水準の高い性能が得られたうえ、どの因子が判定に寄与したかを確認できるため、判断基準の可視化という利点がある。AUROC は 0.7 程度 を維持し、識別性能も十分であった。以上より、文章記載に基づく自動スクリーニングは感度を重視する救急現場に適しており、特徴量抽出を組み合わせた手法は、性能と説明可能性の両面を満たす点で実装上有用と考えられる。

４．今後の課題

カルテ記載の質や内容にばらつきがあり、関連因子の抽出精度に影響する可能性がある。画像所見を含めない運用は救急現場での簡便性に寄与する一方で、脊椎骨折の一部は文章のみでは判断が困難である。また、ローカル LLM の推論速度や運用負荷の評価も必要であり、実臨床ワークフローに組み込む際の検証が求められる。

５．研究成果の公表方法

本研究成果は、脊椎脊髄病学会での発表を検討しており、さらに Journal of Orthopaedic Science への論文投稿も視野に入れている。

以上

Development of an Electronic Medical Record Screening System Using Large Language Models to Prevent Missed Vertebral Fractures Associated with Traffic Accidents

Primary Researcher:
Takaki Kitamura, M.D.　Orthopaedic Surgeon,Department of Orthopaedic Surgery, Chiba University Hospital
Co-researchers:
Satoshi Maki, M.D., Ph.D. Assistant Professor, Department of Orthopaedic Surgery, Chiba University Hospital
Takeo Furuya, M.D., Ph.D. Specially Appointed Lecturer, Department of Orthopaedic Surgery, Chiba University Hospital
Megumi Yamazaki, Administrative Staff, Department of Orthopaedic Surgery, Chiba University Hospital

Abstract
Recent changes in working regulations for physicians have increased the likelihood that less-experienced doctors provide initial care in emergency departments, raising concerns about variability in clinical evaluation and missed diagnoses. Missed vertebral fractures can lead to neurological deficits and long-term functional impairment. This study aimed to develop an automated screening approach based on free-text entries in electronic medical records to minimize missed vertebral fractures. We analyzed 334 patients who were presented with low back pain, back pain, or lumbar contusion between 2013 and 2023 and subsequently underwent MRI to confirm the presence or absence of compression fractures. Text from the subjective and objective sections of the records was processed and evaluated using two locally implemented large language models (LLMs), Japanese-capable large language models and medically specialized large language models, with both simple and detailed prompts. The detailed prompts resulted in higher sensitivity (Both are in the 0.8 range) and improved AUROC values compared with simple prompts. In addition, vertebral fracture–related factors extracted by the LLMs were used to train Light GBM models, which achieved sensitivities of approximately 0.8 and AUROC values of approximately 0.7. These findings suggest that an automated screening system based on routine clinical documentation may help reduce missed vertebral fractures in emergency care. The combination of LLM-based factor extraction and machine learning offers both high sensitivity and improved transparency in the decision-making process.