

Using a Large Language Model to Detect Missed Documentation of Appendicular Musculoskeletal Fractures in the Emergency Department: A Pilot Study

Dr Kasim Alimohamed, Dr Sabeeh Syed, Dr Aimen Hassan, Dr Hassan Ahmad, Dr Saranya Baleswaran, Mr Mohammed Jafer Ali,
Dr Ravindran Karthigan

Guy's and St Thomas' NHS Foundation Trust

Introduction

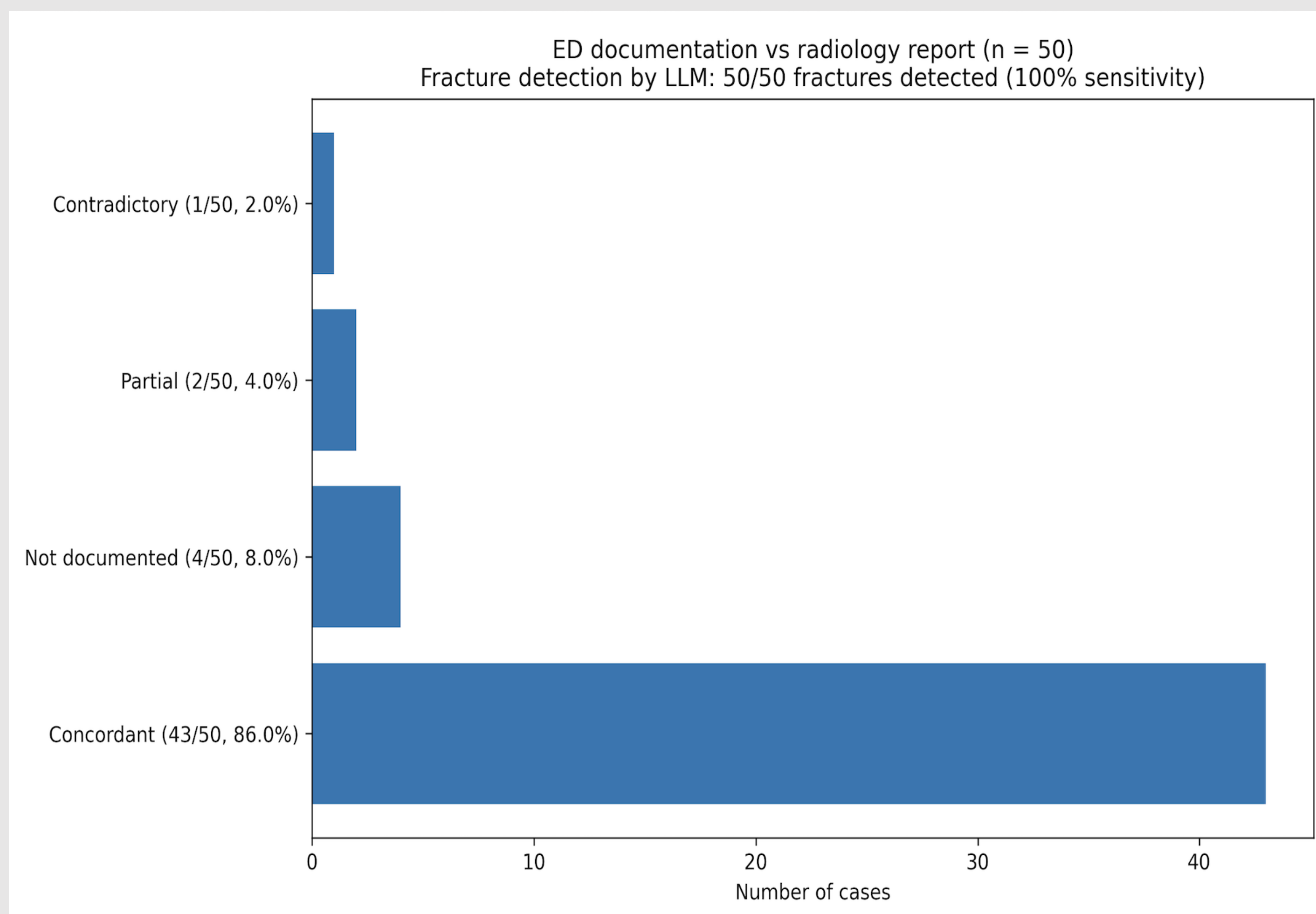
Missed or undocumented appendicular musculoskeletal (MSK) fractures in the emergency department (ED) can delay treatment and compromise patient safety. Fractures are among the most commonly missed injuries on plain radiographs and are a recognised source of diagnostic error in the ED.¹ This study evaluated whether a large language model (LLM) could (1) detect acute appendicular fractures from radiology reports and (2) identify missed, incomplete, or contradictory documentation of these fractures in ED clinical notes.

Methods

We retrospectively identified 50 consecutive appendicular X-ray reports over a 4-day period in which an acute fracture was present. For each case, ChatGPT (GPT-5.1 Thinking, OpenAI) was provided with: (1) the radiology report to determine whether an acute fracture was present, and (2) the corresponding ED notes to assess whether the fracture was documented and whether appropriate management had been initiated. The model was instructed to recommend alerting the ED team if it believed a fracture had been missed or incompletely documented. All radiology reports, ED documentation, and model outputs were independently reviewed by a radiologist as the reference standard.

Results

ChatGPT correctly identified the presence of an acute fracture in all 50/50 cases (100% sensitivity; approximate 95% CI 92.9–100%). ED documentation was fully concordant with the radiology report in 43/50 cases (86%). In the remaining 7/50 cases (14%), potential documentation issues were identified: 4/50 (8%) fractures were not mentioned in the ED notes at all, 2/50 (4%) had only partial transcription of the radiology findings, and 1/50 (2%) explicitly documented “no fracture” despite a positive X-ray report. ChatGPT achieved 100% agreement with the radiologist reference standard in determining whether the fracture was documented in the ED notes and whether appropriate management had been initiated. It flagged all 4/4 (100%) completely undocumented fractures and all 3/3 (100%) cases with incomplete or contradictory documentation as requiring attention.



Conclusion

This pilot study suggests that an LLM (ChatGPT, GPT-5.1 Thinking) can accurately detect appendicular MSK fractures from radiology reports and identify missing, incomplete, or inconsistent ED documentation, including overlooked and contradictory entries. Such tools may provide an automated safety net for missed fractures and reduce the need for manual review of ED notes by radiologists. Larger, prospective studies are warranted to validate generalisability and explore integration into clinical workflows.

References

1. Pinto A, Berritto D, Russo A, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. *Acta Biomed*. 2018;89(1-S):111–123.

Contact

Dr Kasim Alimohamed
Guy's and St Thomas NHS Foundation Trust
Email: Kasim.Alimohamed@nhs.net
Website: <http://www.guysandstthomas.nhs.uk>