Global Healthcare Transformation in the Era of Artificial Intelligence and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250669

Exploring the Potential of LLMs for Patient Safety Incident Reporting in Finland: Interview Insights and a Proof-of-Concept Study

Jusa ANNEVIRTA[a,b,1] and Ilkka SAARENPÄÄ[b,c]
[a] *Finnish Centre for Client and Patient Safety, Wellbeing Services County of Ostrobothnia, Vaasa, Finland*
[b] *Wellbeing Services County of Southwest Finland, Turku, Finland*
[c] *Neurocenter, Department of Neurosurgery, Turku University Hospital and University of Turku, Turku, Finland*

Abstract. This paper explores the potential of Large Language Models (LLMs) to improve patient safety incident (PSI) reporting in Finland. Through semi-structured interviews with doctors and authorities, key requirements and perspectives on AI integration were gathered. A Proof-of-Concept (PoC) study evaluated the feasibility of using a commercial LLM (GPT-4o) to generate structured PSI reports from unstructured clinical text from patient records. Interview results highlighted the need for integrated and automated reporting systems, with AI seen as a tool to reduce documenting burden and improve data analysis. The PoC demonstrated the technological capability of the LLM to generate coherent and relevant reports but also revealed challenges in completeness and distinguishing incident causality. The findings suggest promising avenues for leveraging LLMs in PSI reporting, warranting further research and development for national implementation.

Keywords. Large language models, GPT-4, patient safety, incident reporting, Finland

1. Introduction

Patient Safety Incidents (PSI) are inevitable in healthcare and addressing these incidents is critical for improving care and reducing avoidable harm. Effective PSI reporting and learning systems are essential tools for identifying risks and learning from errors [1]. Addressing errors and resulting harm from treatment accounts up to 13% of total healthcare costs with more than half of these expenses stemming from preventable incident [2]. In high-income countries, 10% of hospital patients experience adverse events, which account for up to 6% of hospital days and 2.5% of healthcare costs [1].

In Finland, PSIs in social and healthcare add up to one billion euros annually [3]. Currently, PSI-related information in Finland is fragmented across multiple voluntary systems as well as law-mandated documenting to patient records using correct codes and separate reporting to handful of authorities [4,5]. Absence of unified data and

[1] Corresponding Author: Jusa Annevirta, M.Sc. (Tech.); E-mail: jusa.annevirta@ovph.fi.

organizational boundaries lead to underutilized data, hindering efforts to improve patient safety [4]. There is a recognized need to improve the collection, processing, and utilization of PSI data [3].

Artificial Intelligence (AI), particularly Large Language Models (LLM), presents promising opportunities for improving clinical documentation and analysis [6,7]. LLMs have the potential to automate summarizing free-text into structured reports. This study aimed to explore the perspectives of Finnish healthcare professionals and authorities on the use of AI in PSI reporting process and to evaluate the feasibility of using a commercial LLM, GPT-4, for report generation.

2. Methods

Semi-structured interviews were conducted with six doctors in managerial positions from the Wellbeing Services County of Southwest Finland (Varha) and eight participants from national healthcare authorities detailed in Table 1. The interviews aimed to gather insights into the requirements for an intelligent PSI reporting system and the potential for integrating AI into the process. Thematic analysis was used to identify key themes from the transcribed and pseudonymized interview data.

Table 1. Interview participants by organization.

Group	Organization/Department	Interviewees
Doctors (Varha)	Neurosurgery	1
	Orthopedic surgery	3
	Vascular surgery	2
Authorities	STM (Ministry of Social Affairs and Health)	3
	THL (Finnish Institute for Health and Welfare)	2
	Kela (Social Insurance Institution of Finland)	2
	DigiFinland	1
Total		**14**

A Proof-of-Concept (PoC) study was conducted to evaluate the feasibility of using OpenAI's GPT-4o model [8] for generating structured PSI reports. The study focused on the feasibility of using an LLM to generate structured reports from patient records where adverse events had already been documented (indicated by ICD-10 diagnosis codes related to adverse events). The dataset consisted of pseudonymized records from 21 patients from the past five years with ICD-10 diagnosis codes related to adverse events. Issues in data access and time constraints limited the scale of this initial study. Individual patient records were combined, and the LLM was prompted with a single request to generate a structured PSI report based on the WHO's Minimal Information Model for Patient Safety Incident Reporting Systems (MIM PS) [9] (Figure 1). No iterative prompting was employed in this initial study. The generated reports were evaluated by one doctor based on criteria used for evaluating LLM summarization tasks in a medical context [6,10]. A scale from 1 (poor) to 5 (excellent) was used.

- **Completeness**: All critical and expected details are included.
- **Correctness**: Information accurately represents the facts from the source data without any errors, contradictions, or hallucinations.
- **Relevance**: Generated content is focused on clinically most important information from the source document.
- **Coherence**: Information is presented in a clear, logical, and well-organized manner, making it easy for the reader to understand.

Each patient's evaluation considered all PSI reports generated by the LLM for that individual, with the number of reports varying based on the documented incidents and the LLM's ability to summarize them. Evaluator assessed the overall output per patient, using the criteria with additional comments and general feedback.

Data processing impact assessment for the PoC was conducted and accepted at Varha regarding the use of patient data with commercial LLM and compliance the Finnish law on the secondary use of health data as well as the GDPR of the EU.

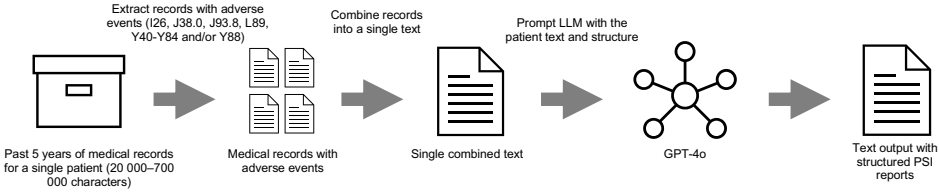

Figure 1. Process of the PSI report generation for single patient's medical records with LLM. In total, the dataset consisted of medical records for 21 patients.

3. Results

Interviews revealed key requirements for an intelligent PSI reporting process (Table 2): integrating reporting into existing systems, standardization and classification to ensure comparability, automating report generation with AI to reduce reporting burden, and providing feedback for engagement. The fragmentation of current reporting systems and procedures is a major frustration leading to redundant reporting burden. Together with lack of proper standardization and classification, the poor data quality contributes to difficulties in national use for monitoring and analysis. Interviewees acknowledged AI's potential for reporting automation, signal detection and data analysis. Doctors and authorities envisioned AI identifying patterns in large datasets, potentially revealing systemic issues and reducing manual work. Concerns about data privacy, security, and the risk of incorrect responses were mentioned, highlighting the need for human oversight.

Table 2. Results from the thematic analysis of interviews (N=14: 6 doctors, 8 authorities) with number of interviewees who mentioned each sub-theme.

Theme	Sub-themes	Interviews
Integration with existing systems	Reporting to multiple systems and authorities result in fragmentation and confusion.	3 (D3, D6, A1)
	Clinicians are overworked with reporting and have no time to report.	3 (D4, D3, A8)
Data quality, standardization and classification	Better common practices and standards for reporting needed.	3 (D6, A8, A1)
	Reporting suffers from inconsistent use, unreliability and incompleteness of ICD-10, procedure, and diagnosis codes.	3 (D3, D4, D2)
	Need for systematic classification system for incidents.	2 (D6, A1)
Artificial intelligence in PSI reporting	AI could help reduce reporting burden.	4 (D2, A3, D3, A5)
	AI has potential for improved analytics and trend identification.	5 (D2, D5, A1, A3, A5)
	AI requires ensuring privacy and security and human oversight.	8 (D2, D3, D5, D6, A1, A2, A3, A7)
Feedback to reporter	Importance of reporting not understood.	3 (D3, D6, A4)
	Lack of feedback and visible benefits on reported incidents.	3 (D3, D5, D6)

The PoC study demonstrated that GPT-4o can effectively generate structured PSI reports from free-text patient data. The LLM achieved high average scores for

correctness (4.86), relevance (4.86), and coherence (5.00), indicating its potential for generating accurate and well-structured reports without significant hallucinations (Figure 2). The completeness (4.38) of the reports scored slightly lower, suggesting some issues in capturing all expected details. This could be attributed to many factors, including limitations in the LLM's ability to extract and synthesize all relevant information from the combined patient records in a single prompt, as well as variability or incompleteness in the original clinical documentation itself. On top of the numeric metrics, a key challenge identified in the evaluation was the LLM's difficulty in consistently distinguishing between treatment-related and non-treatment-related incidents in the large amount of patient records, leading to the generation of PSI reports for events outside the study's scope.

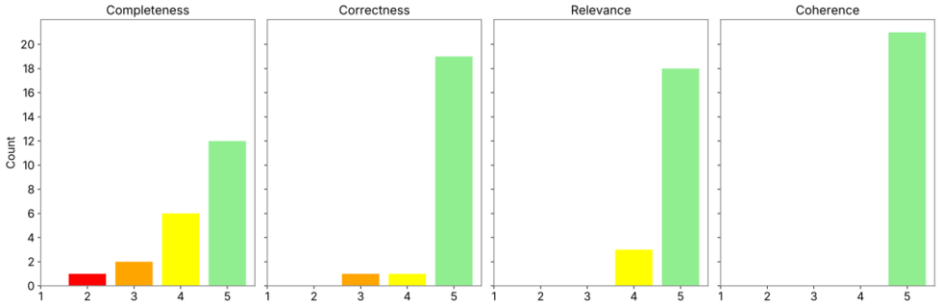

Figure 2. Score distributions for the four LLM evaluation criteria. Mean scores: completeness (4.38), correctness (4.86), relevance (4.86), and coherence (5.00).

4. Discussion and Conclusions

The results highlight a strong desire among healthcare professionals and authorities in Finland for more integrated and automated PSI reporting systems. AI and LLMs are seen as potential tools to alleviate the burden of reporting and enable better analysis. The use of LLM could facilitate more integrated and automated reporting as well as improve the data quality through better standardization and classification. The PoC study provides empirical evidence of the technological feasibility of using a commercial LLM (GPT-4o) to generate structured PSI reports from unstructured clinical text. The LLM's performance in coherence, correctness, and relevance suggests its ability to transform unstructured text into standardized formats. Challenges identified in the PoC regarding completeness and the accurate identification of treatment-related incidents highlight areas for further refinement. These issues may be partly attributed to the PoC's design of processing multiple patient records and PSIs simultaneously, and the inherent variability in the original clinical texts. While this PoC focused on generating reports from already documented adverse events, the potential of LLMs for the initial identification of adverse events from clinical data presents a promising avenue for future research.

This study provides insights into the potential of LLMs for enhancing PSI reporting. The interview results show a need and positive attitude towards utilizing AI to improve reporting processes. The findings support the need for a single-entry reporting process integrated within existing systems, where LLMs can assist in automatically structuring data during the documentation of adverse events. The PoC study demonstrates the promising technological capability of a commercial LLM to generate structured PSI reports from clinical text. However, the identified challenges underscore the importance

of further research focusing on refining prompting strategies, handling complex clinical scenarios, and evaluating the real-world integration of LLMs within existing healthcare systems. While the technological feasibility is promising, the broader implementation of LLMs in PSI reporting requires careful consideration of ethical, legal, and organizational aspects, including data privacy and the necessity of human oversight. This study contributes to the lack of existing research on using LLM for structured PSI report generation. While the PoC shows promising results on using LLM for report generation, its small scale limits the conclusions to preliminary findings requiring more future research in larger-scale study settings.

Acknowledgements

The work was funded by the Finnish Ministry of Social Affairs and Health (STM) as a part of an initiative led by the Finnish Centre for Client and Patient Safety on improving national patient safety reporting in Finland. This paper is derived from my master's thesis study at Aalto University [11], conducted as a part of the initiative in collaboration with Varha. Portions of the text resemble or may replicate the original text from my unpublished thesis.

AI (ChatGPT) was used for writing assistance on this paper for the following tasks: formatting, spelling and grammar, rephrasing and summarizing text. While AI provided support in these areas, all critical thinking, analysis, and content of this paper are original work of the authors.

References

[1] World Health Organization. Global Patient Safety Action Plan 2021-2030: Towards Eliminating Avoidable Harm in Health Care. 1st ed. Geneva: World Health Organization; 2021.

[2] Slawomirski L, Klazinga N. The economics of patient safety: From analysis to action [Internet]. OECD Publishing; 2022. Report No.: 145. Available from: https://dx.doi.org/10.1787/761f2da8-en.

[3] Valtiontalouden tarkastusvirasto. Tarkastuskertomus 7/2021 Potilas- ja asiakasturvallisuuden ohjaus ja seuranta [Internet]. Helsinki: Valtiontalouden tarkastusvirasto; 2021. Available from: http://urn.fi/urn:isbn:978-952-499-507-8.

[4] Virkki M, Leskelä R-L, Ikonen T, et al. Potilas- ja asiakasturvallisuuden tilannekuva ja seurantamenettelyt : Ehdotus seurannan mittaristoksi [Internet]. Helsinki: Valtioneuvoston kanslia; 2021. Available from: https://julkaisut.valtioneuvosto.fi/handle/10024/163632.

[5] Ikonen T, Halinen M, Laukkavirta M. Näin toimin, kun epäilen potilasturvallisuuden vaarantuneen. Duodecim. 2023;139(19):1554–1562.

[6] Van Veen D, Van Uden C, Blankemeier L, et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Res Sq. 2023;rs.3.rs-3483777.

[7] Ferrara M, Bertozzi G, Di Fazio N, et al. Risk Management and Patient Safety in the Artificial Intelligence Era: A Systematic Review. Healthcare (Basel). 2024;12(5):549.

[8] OpenAI. Hello GPT-4o [Internet]. 2024 [cited 2025 Mar 6]. Available from: https://openai.com/index/hello-gpt-4o/.

[9] World Health Organization. Minimal information model for patient safety incident reporting and learning systems: User guide [Internet]. Geneva: World Health Organization; 2016. p. 20. Report No.: WHO/HIS/SDS/2016.22. Available from: https://iris.who.int/handle/10665/255642.

[10] López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, et al. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study. International Journal of Medical Informatics. 2024;187:105443.

[11] Annevirta J. Intelligent Patient Safety Incident Reporting – Process Design and Feasibility of Utilizing LLM for Report Generation [Internet] [master's thesis]. [Helsinki]: Aalto University; 2025. Available from: https://urn.fi/URN:NBN:fi:aalto-202503182886.