Using Artificial Intelligence to Extract NICE Final Appraisal Documents: Evaluating the Potential Application of Large Language Models vs Human Extraction

Knott, C; Crossley, O; Stothard, C; Bodke, A; Samuels, E; Tang, M. Nexus Values, London, United Kingdom

HTA381

Introduction

- AI is a rapidly evolving field of computer science that aims to perform tasks that typically require human intelligence.[1] LLM are a type of AI algorithm that use deep learning techniques and large data sets to understand, summarise, generate, and predict new text-based content.[1]
- The emergence of LLM capable of human-level performance has already transformed various fields within healthcare[2] and presents an opportunity to revolutionise the extraction of data from large documents without the need for laborious human extraction.
- Within market access, the analysis of NICE FAD allows for identification of trends in decision-making that can inform future submissions. However, this analysis requires time-consuming human extraction.
- Therefore, this feasibility study aimed to assess the potential use of LLM in market access when extraction of data from NICE FAD is required, based on accuracy and time.

Methods

- Five NICE FADs published in 2023 were selected to represent a range of disease areas and marketing authorisation scenarios (oncology [monotherapy/combinations], add-on therapy, rare diseases/HST).[3-7]
- Six key topics were defined for extraction (Intervention, Recommendation, Clinical, Economic, Severity modifier, Differentiators). The 6 topics were broken down into subtopics to evaluate success rate.
- A script was developed to prompt the LLM (GPT-4o) to extract the pre-defined qualitative and quantitative data for each topic.
- The data extracted by the LLM was compared to double human extraction for completeness and accuracy. The time to complete each extraction was also calculated and compared between human and LLM (script development was not included within the LLM extraction time).

Results

Script Development

- The script required significant iteration via a series of trial runs before effective extraction was achieved. This was conducted on TA891.
- To achieve effective data extraction, separate scripts for the qualitative and quantitative data were required.

Quantitative Extraction Success Rate

- When compared to double human extraction, the LLM achieved a 100% success rate when extracting quantitative data across all extraction topics.

Qualitative Extraction Success Rate

- The LLM success rate for qualitative data extraction compared to double human extraction ranged from 50% to 100% across topics (Figure 1).
- An overview of the subtopics where full extraction was not completed successfully by the LLM when compared to double human extraction is presented in Tables 1-4.

Figure 1: LLM extraction success rate for qualitative data vs. double human extraction

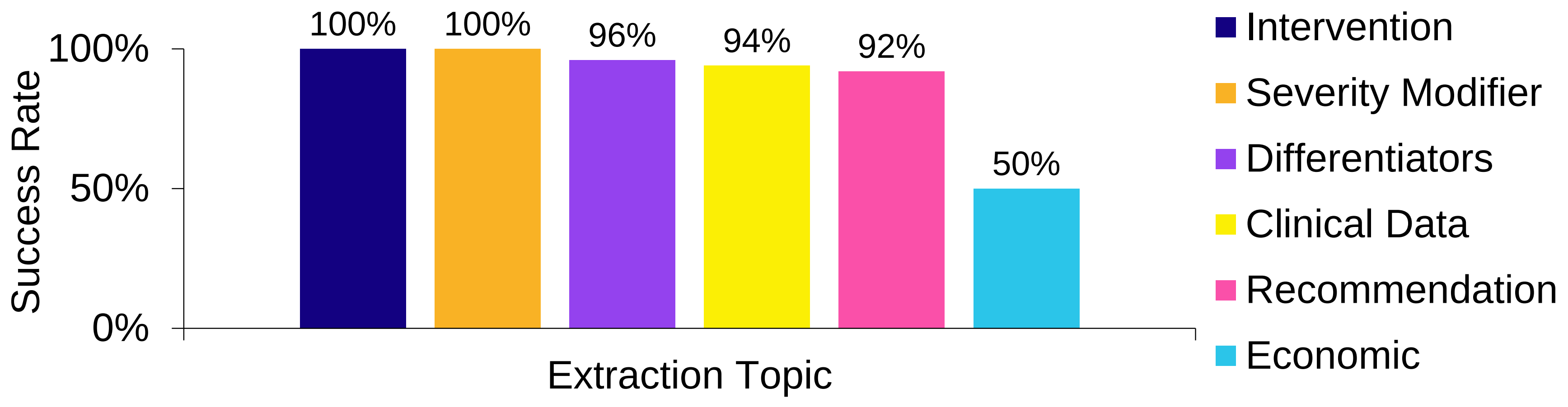

Table 1: Differentiators qualitative data extraction success rate vs. human extraction

FAD	Differentiators				
	Innovation	End of life	Equality	Patient input	Expert input
TA864	🟩	🟩	🟩	🟩	🟩
HST22	🟩	🟩	🟩	🟩	🟩
TA880	🟩	🟩	🟩	🟩	🟩
TA891	🟩	🟩	🟩	🟩	🟩
TA911	🟧	🟩	🟩	🟩	🟩

Table 2: Clinical qualitative data extraction success rate vs. human extraction

FAD	Clinical						
	Trial name	Design	Primary outcome	Secondary outcome	Population description	Additional evidence	Trial insight
TA864	🟩	🟩	🟩	🟩	🟩	🟩	🟩
HST22	🟩	🟩	🟩	🟩	🟩	🟩	🟩
TA880	🟩	🟩	🟩	🟩	🟩	🟧	🟧
TA891	🟩	🟩	🟩	🟩	🟩	🟩	🟩
TA911	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Key: 🟩 Full extraction 🟧 Data not captured

Table 3: Recommendation qualitative data extraction success rate vs. human extraction

FAD	Recommendation				
	Indication	Therapy line	Dependent on	Previous review	Reasons for recommendation
TA864	🟩	🟩	🟩	🟩	🟩
HST22	🟩	🟩	🟩	🟧	🟩
TA880	🟩	🟩	🟩	🟩	🟩
TA891	🟩	🟩	🟩	🟩	🟩
TA911	🟩	🟩	🟩	🟧	🟩

Table 4: Economic qualitative data extraction success rate vs. human extraction

FAD	Economic			
	ITC	ITC limitations	Model structure	Model critiques
TA864	🟩	🟧	🟩	🟧
HST22	🟩	🟧	🟧	🟧
TA880	🟩	🟧	🟩	🟩
TA891	🟧	🟧	🟩	🟩
TA911	🟧	🟩	🟩	🟧

Qualitative Text Not Captured by the LLM vs. Human Extraction

- Qualitative text missed by the LLM vs. double human extraction is presented in Table 5.

Table 5: Summary of the key qualitative text missed by the LLM vs. human extraction

Extraction topic	Summary of the qualitative text missed by the LLM
Differentiators	• The technology was considered “innovative” by the Committee
Clinical	• Company updated their clinical outcomes and presented additional clinical trial and post-hoc subgroup analysis data
Recommendation	• FAD reviewed existing trial data from a previous FAD plus additional evidence submitted in the updated report
	• Technology is recommended for another disease
Economic	• Model used by the company was specific to the target disease population
	• Committee reported uncertainty around the survival data
	• Company had to update their model to reflect the progression of the disease
	• Committee would have preferred to use a range of plausible curves provided by the Committee’s clinical experts

LLM vs. Human Extraction Time

- Per extraction, the average time for human extraction was 40 minutes vs. 13 minutes for LLM.

Conclusions

- This feasibility study suggests that LLM, like GPT-4o, could be used to accurately and efficiently extract all quantitative data and most qualitative data from NICE FAD. However, human intervention is still currently required when extracting some qualitative data, especially economic qualitative data.
- The feasibility of using current generation LLM to extract FAD data using a written script could result in time savings when analysing market access trends in NICE decision-making when extraction of large numbers of FAD is required.
- However, any efficiency gains are currently offset by the time required to develop and test the script, meaning human extraction is still most efficient where smaller numbers of FAD are being analysed.
- For future NICE FAD extraction, the same script could be used as developed in this study and cross-checked with human intervention, which would reduce the time taken to conduct this research.

Abbreviations: AI: artificial intelligence; FAD: final appraisal documents; GPT-4o: Generative Pre-trained Transformer 4; HST: highly specialised technologies; ITC: indirect treatment comparison; LLM: large language models; NICE: National Institute for Health and Care Excellence; TA: technology appraisal

References: 1. Alowais, S.A., et al., (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education, 23(1), p.689.
2. Bekbolatova, M., et al., (2024), Transformative potential of AI in Healthcare: definitions, applications, and navigating the ethical Landscape and Public perspectives. In Healthcare (Vol. 12, No. 2, p. 125).
3. NICE (2023).TA864. https://www.nice.org.uk/guidance/ta864/resources/nintedanib-for-treating-idiopathic-pulmonary-fibrosis-when-forced-vital-capacity-is-above-80-predicted-pdf-82613612686021;
4. NICE (2023). HST22. https://www.nice.org.uk/guidance/hst22/resources/ataluren-for-treating-duchenne-muscular-dystrophy-with-a-nonsense-mutation-in-the-dystrophin-gene-pdf-50216315955397;
5. NICE (2023). TA880. https://www.nice.org.uk/guidance/ta880/resources/tezepelumab-for-treating-severe-asthma-pdf-82613726899909
6. NICE (2023). TA891. https://www.nice.org.uk/guidance/ta891/resources/ibrutinib-with-venetoclax-for-untreated-chronic-lymphocytic-leukaemia-pdf-82613789045701
7. NICE (2023). TA911. https://www.nice.org.uk/guidance/ta911/resources/selpercatinib-for-untreated-ret-fusionpositive-advanced-nonsmallcell-lung-cancer-pdf-82615482098629

Nexus Values

For further details contact: charles.knott@nexusvalues.com

See our other research at ISPOR EU 2024: HTA53, HTA130, HTA214, HTA302