Natural History of Creutzfeldt-Jakob Disease: A Comparison of Manual Versus Manual plus AI-Assisted Approaches in Literature Review

Poster # MSR152

Setareh A. Williams, PhD, Richard J. Weiss, MD, Russell Vincent Becker, MA, Guy Rayford Mitchell, Jr., MPH, Charles David Williams, BS

Star Biopharma Consulting, LLC., Malvern, PA, USA

OBJECTIVES

To compare the results of a literature review on the natural history of Creutzfeldt-Jakob Disease (CJD), a rare, fatal, neurodegenerative brain disorder:

- **Conducted manually vs.**
- **Manually with AI assistance using the GPT-4 Large Language Model (LLM)**

To evaluate LLM performance, particularly in:

- **Precision**
- **Sensitivity**
- **Time required**

BACKGROUND

- Evaluation of disease burden, including natural history, is the first step towards documentation of unmet medical need and requires literature reviews
- Growing volume of published literature makes manual literature reviews **increasingly time and resource-intensive**
- **NICE** (National Institute for Health and Care Excellence) guidelines suggest that **LLMs have the potential to automate** parts of literature review including:
  - **Generation of search strategies**
  - **Study classification**
  - **Primary and full-text screening to find eligible studies**
  - **Visualization of search results**[1]
- LLM performance characteristics should be transparent, and they should be used to classify the same types of data used to train them[2]
- Validation should include comparison to manual reviewers, including specificity and precision metrics
- Current research on potential of LLM assistance for literature review includes:
  - Abstract screening and ranking for qualitative reviews
  - Informing conceptual model development for clinical outcomes assessment (COA)[3]
  - Classification of abstract types[4]
  - Article search and narrative generation for targeted literature reviews[5]

METHODS

PubMed Literature Search:

1. Initially searched with manually chosen terms: *"natural history"* or *"disease progression"* or *"worsening"* or *"clinical course"* or *"longitudinal"*

2. Added terms suggested by LLM: "*time course*" or "*disease trajectory*" or "*clinical evolution*" or "*temporal pattern*" or "*disease timeline*" or "*progressive symptoms*" or "*onset to death*" or "*early signs*" or "*late-stage features*" or "*chronology of symptoms*"

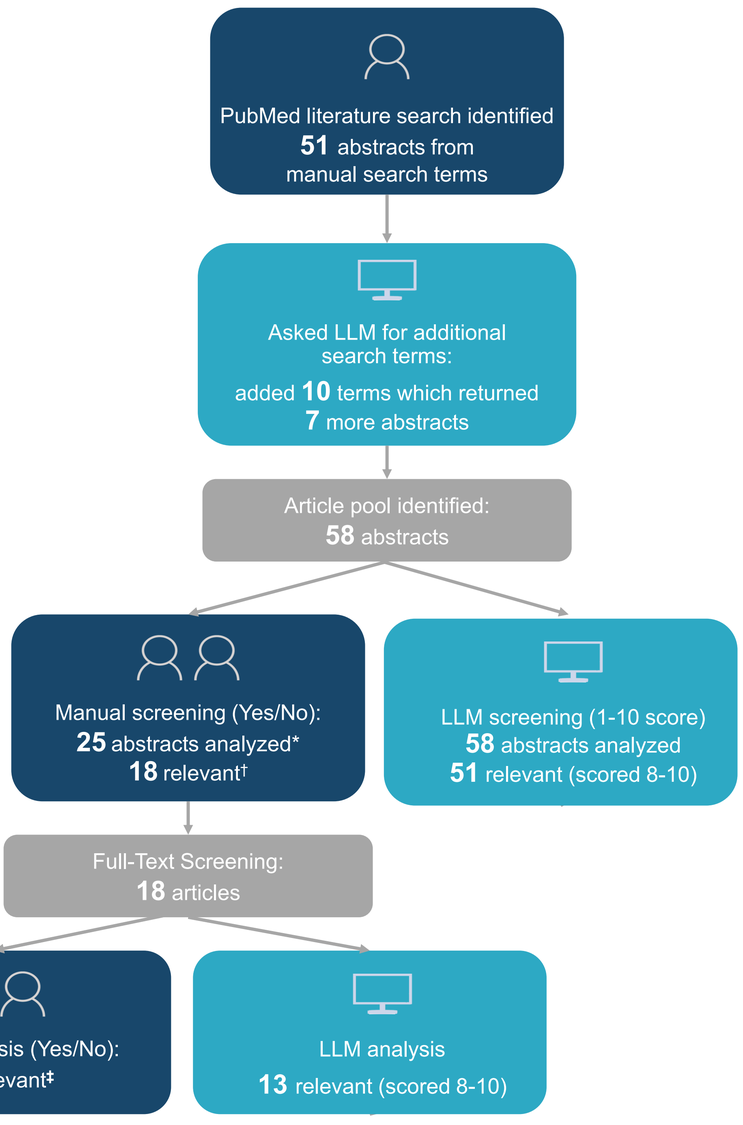
For detailed search strings, see our Supplemental Poster Information here:

Screening:

Manual: manuscripts were screened by two human reviewers on a Yes/No basis.

LLM: manuscripts were screened for relevancy using the following prompt: You are a biomedical research assistant. Evaluate the relevance of the following abstract to the topic: "Understanding the natural history, disease progression, and risk factors for Creutzfeldt-Jakob disease."Give a score from 0 to 10, where: 0 means completely irrelevant and 10 means highly relevant and informative on the topic

RESULTS



Reasons for exclusion:

* 33 case studies on diagnostic challenges

† 7 off topic (i.e., ALS, Alzheimer's, delirium, dementia, bovine spongiform encephalopathy, animal studies and general guidelines) (13 abstracts marked relevant + 5 more articles added to pool by second reviewer)

‡ 3 with few specific mentions of CJD

Manual vs Manual+AI Performance:

Abstract Screening	Full Text Screening
LLM Precision[a]: **0.27**	LLM Precision: **1**
LLM Recall[b] (Sensitivity): **.93**	LLM Recall (Sensitivity): **0.77**
Percentage Agreement[c]: **34.4%**	Percentage Agreement: **83%**

Most of the discrepancy between manual and manual+AI precision can be attributed to manual exclusion of case studies (33 case studies manually excluded but included by AI).

[a]Precision = (both manual and manual+AI found relevant)/(all AI found relevant)

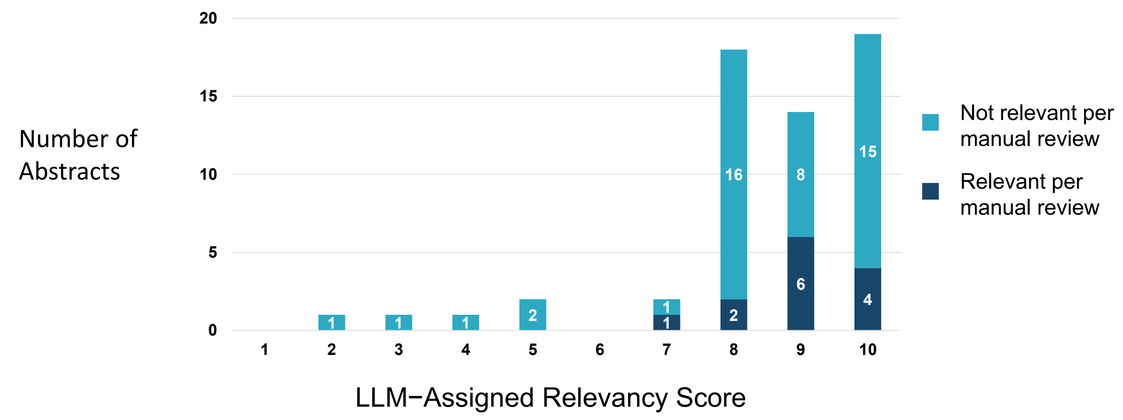
[b]Recall (Sensitivity) = (both manual and manual+AI found relevant) / (both manual and manual+AI found relevant + AI found not relevant but manual found relevant)

[c]Percentage Agreement = (both manual and manual+AI found relevant + manual+AI found not relevant but manual found relevant) /total articles X 100

Time required:

Manual review (abstract and full text):	LLM review (programming, execution):
23.5 hours	**25 minutes**

LLM Scoring is Sensitive But Not Precise



CONCLUSIONS

- **Search Strategy:** Adding 10 LLM-suggested search terms to the PubMed search resulted in one additional relevant article
- **Article Screening:** Analysis of full-text manuscripts by LLM showed higher percentage agreement than analysis of abstracts. This is mostly due to the exclusion of case reports by human reviewers.
- LLM sensitivity throughout was high (few false negatives observed) but precision was low.
- LLM offered substantial time savings
- While LLMs may have the potential to enhance literature review, it is advised to be cautious and use them to complement, not replace, human reviewers to maintain accuracy and reliability at every stage of literature review.

References: 1. National Institute for Health and Care Excellence. 2024. https://www.nice.org.uk/position-statements/use-of-ai-in-evidence-generation-nice-position-statement.
2. NICE. 2024. https://www.nice.org.uk/process/pmg20/chapter/identifying-the-evidence-literature-searching-and-evidence-submission#searches-during-guideline-recommendation-scoping-and-surveillance.
3. Burbridge C et al. *Value in Health*. 2024;27(12):S477-S477.
4. Metcalf T et al. *Value in Health*. 2024;27(12):S482-S482.
5. Baisley W et al. *Value in Health*. 2023;26(12):S402-S402.

Funding No funding was received for the development of this poster.
Acknowledgments Editorial and graphics services were provided by Catherine Hueston and funded by Star Biopharma Consulting, LLC

For more information, contact:

Setareh A. Williams setareh.williams@starbiopharmaconsulting.com

Presented at ISPOR EU | Glasgow, Scotland | 9-12 November 2025