Streamlining Systematic Review Feasibility Assessments With Large Language Models: A Novel AI-Driven Workflow

#144966

Tim Disher, PhD, RN | Loon | Halifax, NS, Canada

INTRODUCTION

The integration of artificial intelligence (AI) into systematic literature review workflows has predominantly focused on automated screening and data extraction. Following these stages, a detailed feasibility assessment of evidence synthesis should be performed to identify sources of heterogeneity in outcome definitions, inclusion criteria, and other trial design characteristics.

OBJECTIVES

The objectives of this study were to:

1. Describe an LLM-based approach to ITC feasibility assessments;
2. Illustrate application in a case study using on ongoing systematic review;
3. Assess the ability of the approach to simplify, improve, and accurately conduct the required tasks.

METHOD

We utilized a combination of iterative conversations and API calls with GPT-3.5 to assess outcome and inclusion/exclusion criteria feasibility.
The primary tasks included:

1. De-duplication of identical or near-identical outcome definitions;
2. Identification of key areas of clinically important variability in outcome definitions and inclusion/exclusion criteria;
3. Extraction of criteria into structured forms that capture variability in these components.

De-duplication of identical or near-identical outcome definitions

Verbatim outcome definitions can typically be similar but not identical. A key first step is the identification of duplicate components into harmonized definitions. In this pilot, this step was conducted iteratively with the main issue being that not all outcomes were assigned a unique outcome name.

a total mayo score of 2 points or lower with no individual subscore exceeding 1 point.

total mayo score of ≤2 and no individual subscore >1

mayo clinic score of 2 or lower and no subscore higher than 1

total mayo score of ≤2 and no subscore >1

→ *total mayo score of ≤2 and no subscore >1*

Identification of key areas of clinically important variability

LLM is provided with the required clinical and expert context and asked to identify aspects of variability in definitions across trials that may be important for explaining heterogeienty. Components that don't vary are de-prioritized.

≥18 years of age, received a diagnosis of ulcerative colitis at least 3 months before screening and had moderate-to-severe ulcerative colitis, defined as a total score of 6 to 12 on the mayo scale and a subscore of 2 or 3 on the endoscopic component of the mayo scale, as determined during central review of videoendoscopy. Eligible patients were required to have had an inadequate response to or unacceptable side effects from TNF antagonists, vedolizumab, or conventional (i.e., nonbiologic) therapy.

Relevant Construct

- Age
- Severity Score
- Disease Duration
- Bio-experience

Irrelevant Construct Insufficient variability

- Qualitative severity

Extraction into Structured Format

The key components from the previous step are used to create a structured json list for data extraction. These form the basis of comparison of trials in terms of variability across key components of inclusion/exclusion and outcome definitions.

Inclusion Criteria	Age Range	Disease Duration	Mayo Score Range	Prior Treatment Categories	Extent of disease	Endoscopic Confirmation
Age 16-80 years old, moderate to severely active UC (confirmed by endoscopy with >=10 cm rectal involvement and on the basis of a modified mayo score of 4-9 with a centrally read endoscopic subscore >=2 and rectal bleeding subscore >=1), a documented history of inadequate response, loss of response, or intolerance of at least one therapy approved for the treatment of UC.	16-80	NA	4-9	Biologic or Conventional Therapy failure	>=10	Yes

Conclusions

The LLM workflow substantially reduced the manual effort involved in review, data sheet development, and extraction for feasibility assessments. Although human intervention and review were necessary, agreement across tasks was generally high (~90%). Errors in outcome definitions and inclusion criteria were typically straightforward to identify and correct, resulting in significant time savings. However, evaluating exclusion criteria posed greater challenges and required increased input from clinical experts to develop an initial structured data extraction form. Subsequent extraction showed unacceptable error rates, necessitating a re-focus on a more straightforward sub-task of exclusion feasibility
Large language models can likely streamline the evidence synthesis feasibility assessment process with minimal risk, provided experts are involved at all stages. Evaluating exclusion criteria may be more complex due to greater variability in language and difficulty in understanding criteria implications.

tim@loonbio.com

loonbio.com