HTA99

Matching Insights from Clinical Experts and Generative AI for JCA PICO Validation

Emma Benbow,[1] Sven Klijn,[2] Cheryl Jones,[1] Nebibe Varol,[2] Bill Malcolm,[2] Tim Reason,[1] Manoj Chevli,[3] Siguroli Teitsson[2]

[1]Estima Scientific, London, United Kingdom; [2] Bristol Myers-Squibb, Uxbridge, UK; [3] Employee of Bristol Myers-Squibb, Uxbridge, UK at the time of the study

Introduction

- The Joint Clinical Assessments (JCA) under the European Union Health Technology Assessment (EU HTA) Regulation seek to harmonize the clinical aspects of HTA for medical interventions across all European Union (EU) member states
- The JCA process uses the Patient, Intervention, Comparator, Outcome (PICO) framework and, on submission of a health technology assessment (HTA) by a health technology developer (HTD), requests that all EU member states put forward their PICO requirements for the disease of the registrational trial
- After consolidation of the PICO sets by the JCA assessors, HTDs are then informed of the final PICO sets and have up to 100 days to submit their clinical analyses (a tighter 60-day deadline applies for assessments under the accelerated procedure or for a variation to the terms of an existing marketing authorization)
- The number of PICO sets that need consideration within a JCA submission could be very large, making it potentially challenging to complete systematic literature reviews (SLRs) and analyses e.g., network meta-analyses (NMAs), within the tight timeframe
- Thus, it would be beneficial to have an automated process capable of quickly determining which populations across different PICO sets align with the pivotal trial of interest's population, and to automatically conduct SLRs for the aligned PICOs
- Automated SLRs have been described elsewhere[1]; the purpose of this work was to determine whether it is possible to automate alignment of JCA PICO populations
- Outcomes and comparators were not considered for alignment at this stage as these criteria may lead to over-exclusion of PICO sets that could be relevant

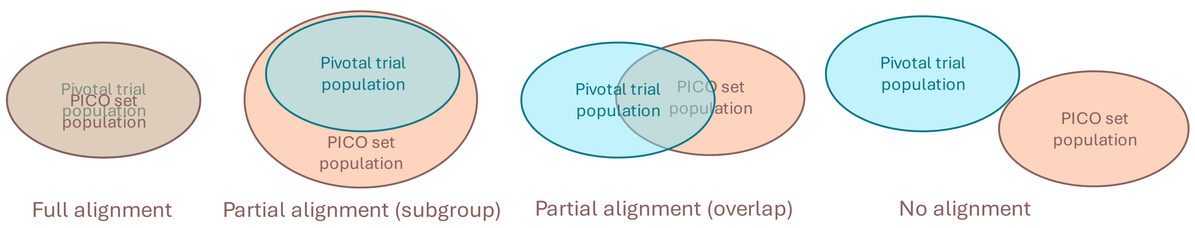
Aim

- The purpose of this study was to investigate whether large language models (LLMs) can determine the degree of alignment of JCA PICO populations, as predicted by clinical experts, with the population of a pivotal registrational trial. The degree of alignment was assessed using a case study of patients with relapsed/refractory multiple myeloma (RRMM).

Methods

- The relevant population from the trial of interest were patients with RRMM who had received at least one prior line of antimyeloma therapy, including lenalidomide.
- Based on a JCA simulation including external experts, 20 PICO sets were identified. The populations of these 20 PICO sets were assessed for their degree of alignment with the population of the pivotal trial of interest.
- The degree of alignment of each predicted PICO set population with the population from the pivotal trial of interest was defined in 4 categories (Figure 1):
  - **Full alignment:** Population characteristics of the pivotal trial of interest fully align with the population described in the PICO set
  - **Partial (subgroup):** All patients in the PICO set align with some of the patients in the pivotal trial of interest
  - **Partial (overlap):** A proportion of patients from the pivotal trial of interest align with some of the PICO set population
  - **None:** None of the pivotal trial patients align with the population of the PICO set

Figure 1. Degree of population alignment definitions



Abbreviations: *PICO, Patient, Intervention, Comparator, Outcome*

- Four LLMs (Claude 3 Opus [29/02/2024], Claude 3.5 Sonnet [20/06/2024], GPT-4 [03/14/2024], GPT-4o [06/08/2024]), accessing their application programming interfaces (APIs) through Python, were provided with prompts and contextual information (Table 1) to assess the degree of alignment of the populations within each PICO set and the characteristics of the population of the pivotal trial of interest.
- Prompts were developed by non-clinical experts using an iterative process. The key to the successful assessment of the degree of alignment between the two populations was to provide the LLMs with sufficient context, see Table 1.
- A self-consistency approach[2] was used, whereby the LLM is instructed to repeat the same task multiple times, removing the non-systematic errors, and select the most frequently occurring (modal) answer. This ensured the highest degree of accuracy was achieved. Accuracy of alignment categorisation for the populations was determined by comparing the LLM outputs to alignment categorisation by clinical experts.

Table 1. Context provided to the LLMs

Criteria	Context Provided
Definitions	The definition of RRMM is relapsed or refractory MM, i.e., not all patients with RRMM have relapsed MM and not all are refractory.
	A patient is triple class refractory if they are refractory to a proteasome inhibitor, an immunomodulatory agent, and a monoclonal antibody treatment.
	If a patient is sensitive to a treatment, this means that they are not refractory to it.
Treatment exposure and class	Patients can be exposed to any treatment, or treatment class, unless specifically excluded in the population description.
Number of relapses	A patient cannot have more relapses than they have had lines of treatment; for example, a patient who has only had one prior line of treatment may be non-relapsed or may be at their first relapse.
	Similarly, a patient who has had 2 lines may be non-relapsed, at their first relapse, or at their second relapse.
Number of lines or classes of treatment	The number of lines of treatment that a patient has received is not indicative of the number of classes of treatment that they have received.

Abbreviations: RRMM: relapsed refractory multiple myeloma

Results

- All 20 populations were classed as having "some alignment" by either the humans or the LLMs i.e., all populations needing to be taken further within the JCA process were classed as such by the LLMs
- Human classification of the alignment of the 20 populations was partial (subgroup) ("PS") for three populations and partial (overlap) ("PO") for seventeen
- Claude 3 Opus and Claude 3.5 Sonnet were correct for 18/20 (same two populations were wrongly classified by the two Anthropic LLMs).
- GPT-4 and GPT-4o were also correct for 18/20 (same two populations were wrongly classified)
- One population was misclassified by the Anthropic LLMs (Claude), one by the OpenAI LLMs and one by all four LLMs. The three populations misclassified by the LLMs are highlighted in Table 2.

Table 2. Populations assessed by the LLMs

Populations having partial overlap (PO) with SUCCESSOR-1 population

- RRMM at first relapse, after treatment with bortezomib, lenalidomide and dexamethasone; lenalidomide refractory
- RRMM at first relapse, after treatment with daratumumab, lenalidomide and dexamethasone; lenalidomide sensitive and refractory
- RRMM at first relapse, after treatment with daratumumab, lenalidomide and dexamethasone; with t(11;14) chromosomal abnormalities
- RRMM at first relapse, After treatment with bortezomib, lenalidomide and dexamethasone; bortezomib sensitive, with t(11;14) chromosomal abnormalities
- RRMM at first relapse, after treatment with lenalidomide-based regimen; lenalidomide and bortezomib sensitive
- RRMM at second or subsequent relapse, lenalidomide refractory; proteasome inhibitor sensitive
- RRMM at second or subsequent relapse, lenalidomide refractory. proteasome inhibitor sensitive, with t(11;14) chromosomal abnormalities
- RRMM; lenalidomide refractory; anti-CD38 mAb naive
- RRMM after 1 prior line, extramedullary disease
- RRMM after 1 prior line with lenalidomide: early vs. late relapses
- RRMM after at least 1 prior line, renal impairment (if lenalidomide exposed & CCL)
- RRMM after 1 prior line, High risk cytogenetic abnormalities if lenalidomide exposed
- RRMM, not refractory to lenalidomide or anti-CD38 mAb
- RRMM, not refractory to lenalidomide but anti-CD38 mAb-refractory
- RRMM, lenalidomide refractory but not refractory to anti-CD38 mAb
- RRMM , lenalidomide and anti-CD38 mAb-refractory
- RRMM, after 1 prior line of therapy, including lenalidomide

Populations for which SUCCESSOR-1 population is a subgroup

- RRMM; lenalidomide exposed or refractory
- RRMM after at least 1 prior line of therapy: fit vs. unfit
- RRMM after at least 1 prior line of therapy: proteasome inhibitor exposed vs. naïve

RRMM Population Description	Human Assessment	Claude 3 Opus	Claude 3.5 Sonnet	GPT-4	GPT-4o
After 1 prior line of therapy, including lenalidomide	PO	Fully aligned	Fully aligned	PO	PO
After at least 1 prior line of therapy: fit vs. unfit	PS	PO	PO	PO	PO
After at least 1 prior line of therapy: proteasome inhibitor exposed vs. naïve	PS	PS	PS	PO	PO

Abbreviations: *PO, partial (overlap); PS, partial (subgroup); RRMM, relapsed/refractory multiple myeloma*

- Potential ambiguity in the population definition of the three populations was likely to have caused the misclassification of the degree of alignment. Indeed, it was difficult for a human to correctly classify the populations defined as "RRMM after at least 1 prior line of therapy: X vs. Y"
- Changing *"RRMM after 1 prior line of therapy, including lenalidomide"* to read *"RRMM after 1 prior line of therapy containing lenalidomide"* allowed Claude 3 Opus to correctly assess the degree of alignment as "partial alignment (overlap)".
- Similarly, for example, changing *"After at least 1 prior line of therapy: fit vs. unfit"* to either 2 separate populations: *"After at least 1 prior line of therapy, fit patients"* and *"After at least 1 prior line of therapy, unfit patients"*; or just removing the subgroups from the description *"After at least 1 prior line of therapy"*, allowed all four LLMs tested to assess the degree of alignment as either "partial alignment (overlap)" for the two separate populations, or "partial alignment (subgroup)" for the combined population.

Conclusion

- If appropriate context is provided, LLMs are capable of understanding complex epidemiological concepts and categorizing the alignment of two populations. Thus, LLMs could be used to automate the classification of PICOs within the JCA process.
- Given the volume of potential analyses required to complete a JCA submission and the tight timelines, the automation of specific tasks could offer substantial benefits in terms of time and costs saved.
- One potential challenge of the process is understanding the degree of context the LLM requires in order to make appropriate judgements on the degree of alignment between the PICO population and the trial population. It may be recommended to seek clinical inputs to aid with the development of accurate contextual information to pass to the LLM. With that said, the patient population tested for this study was particularly complex, therefore, for a simpler patient population, the amount of context required may be of less importance, however this would require further testing.
- The results of this study suggest a key step (assessing the degree of alignment of predicted PICOs with the pivotal trial population) within the JCA process could be automated using LLMs. Automation for other tasks including the mass extraction of PICOs from clinical abstracts[1], screening and assessing bias for systematic literature reviews[3], data extraction and[4,5] analysis[5] have also been demonstrated. Further research is currently ongoing to investigate whether these elements can be tied together in a fully automatic toolchain.

References

1. Reason, T., Langham, J. & Gimblett, A. Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study. Pharm Med(2024). https://doi.org/10.1007/s40290-024-00539-6
2. MSR18 Improving the Performance of Generative AI to Achieve 100% Accuracy in Data Extraction. Klijn, S et al. Value in Health, Volume 27, Issue 6, S262 - S263
3. MSR80 AI-Enabled Risk of Bias Assessment of RCTs in Systematic Reviews: A Case Study. Langham, J. et al. Value in Health, Volume 26, Issue 12, S408
4. Wu, et al., Generative AI: A Novel Approach to Data Extraction for NMAs in EU JCA. Value in Health: ISPOR EU 2024
5. Reason, T., Benbow, E., Langham, J. et al. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. PharmacoEconomics Open 8, 205–220 (2024). https://doi.org/10.1007/s41669-024-00476-9

Acknowledgments

- This study was supported by Bristol Myers Squibb

Presented at ISPOR Europe 2024; November 17-20, 2024; Barcelona, Spain

Email: <emma.benbow@estima-sci.com>