Using AI to Map HEOR Evidence Gaps: Findings from a Large-Scale LLM-Based Classification Analysis

Weicheng Ye[1]; Corrina Mau[1]; Denise Zou[1]

[1]Thermo Fisher Scientific, Waltham, MA, United States

MSR250

Background

- Rapid growth in health technologies has expanded health economics and outcomes research (HEOR) literature.
- Identifying over- or underrepresented disease areas may guide biopharma and diagnostics R&D; AI/LLMs offer scalable bibliometric analysis.

Objectives

- We developed and evaluated an artificial intelligence (AI)-assisted pipeline to:
  - Screen abstracts for intervention-focused HEOR
  - Classify studies by disease area, subtype, and intervention category
  - Generate a scalable HEOR evidence map for investment and strategic planning

Methods

Data Source

- PubMed-based HEOR dataset (2000-2025; English only).
- Constructed using predefined economic-evaluation and intervention-related search terms.

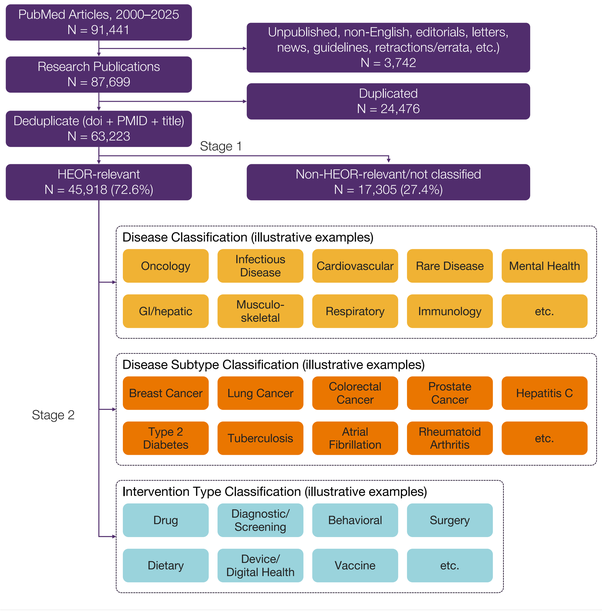
AI/Large Language Model (LLM) Pipeline

- Model: Llama 3.1 was used for scalable natural language processing (NLP) and multi-label classification of unstructured data.
- Stage 1 – HEOR relevance screening: Abstracts were screened for intervention-focused HEOR based on economic-evaluation concepts and intervention context.
- Stage 2 – Study classification: HEOR-relevant abstracts were assigned to disease areas, subtypes, and intervention categories using a predefined classification framework. Outputs included model-generated confidence scores.
- Validation: A random sample of 350 abstracts underwent blinded human review to assess classification accuracy.

Analysis

- Disease-area and intervention-type distributions were summarized.
- Publication trends were evaluated across disease areas from 2000 to 2025.
- Growth rates were estimated using log-linear regression of annual publication counts.
- Relative representation and growth patterns were used to identify over- or underrepresented disease areas.

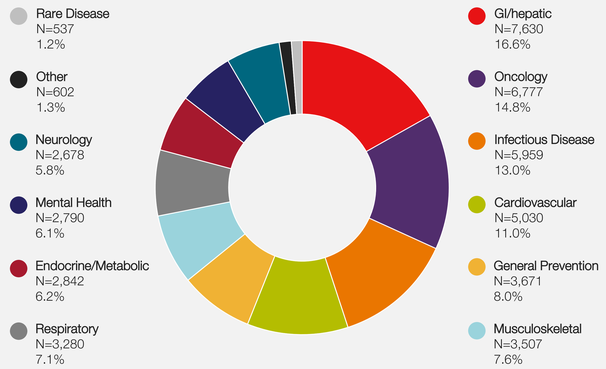
Figure 1. AI-Assisted Pipeline for HEOR Abstract Screening and Classification



Results

- Among 63,223 unique records, 45,918 (72.6%) were classified as HEOR-relevant.
- HEOR-relevant studies were classified into 42 disease areas and 4,086 subtypes, with high model confidence (median: 0.9).
- Human-review accuracy was 81% for HEOR relevance and 75% for disease-area classification.
- Oncology, infectious disease, cardiovascular, and gastrointestinal/hepatic diseases comprised the largest share of HEOR studies (**Figure 2**).
- Mental health (6.1%) and neurology (5.8%) were moderately represented; rare diseases (1.2%) and fields such as reproductive health, dermatology, urology, ophthalmology, pediatrics, hematology, and genetics each represented less than 0.2%.
- Drug interventions accounted for the largest proportion (34.7%), followed by diagnostic/screening (18.1%) and behavioral interventions (13.2%) (**Figure 3**).
- From 2000 to 2025, endocrine/metabolic disease, neurology, musculoskeletal disease, and oncology showed the highest growth rates in HEOR publication volume (**Figure 4**).

Figure 2. Distribution of HEOR Publications by Disease Area (Top 12 Categories)



Top 12 disease areas account for 98.66% of all classified studies.

Figure 3. Distribution of HEOR Publications by Intervention Type

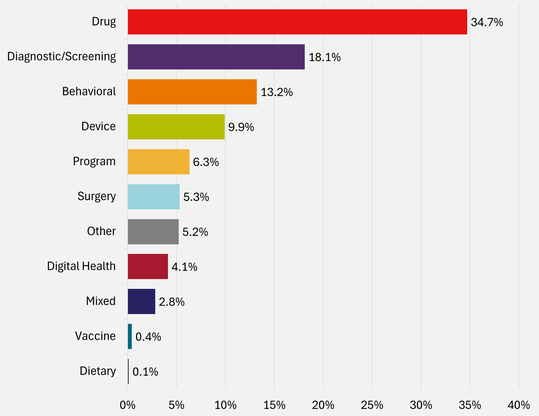
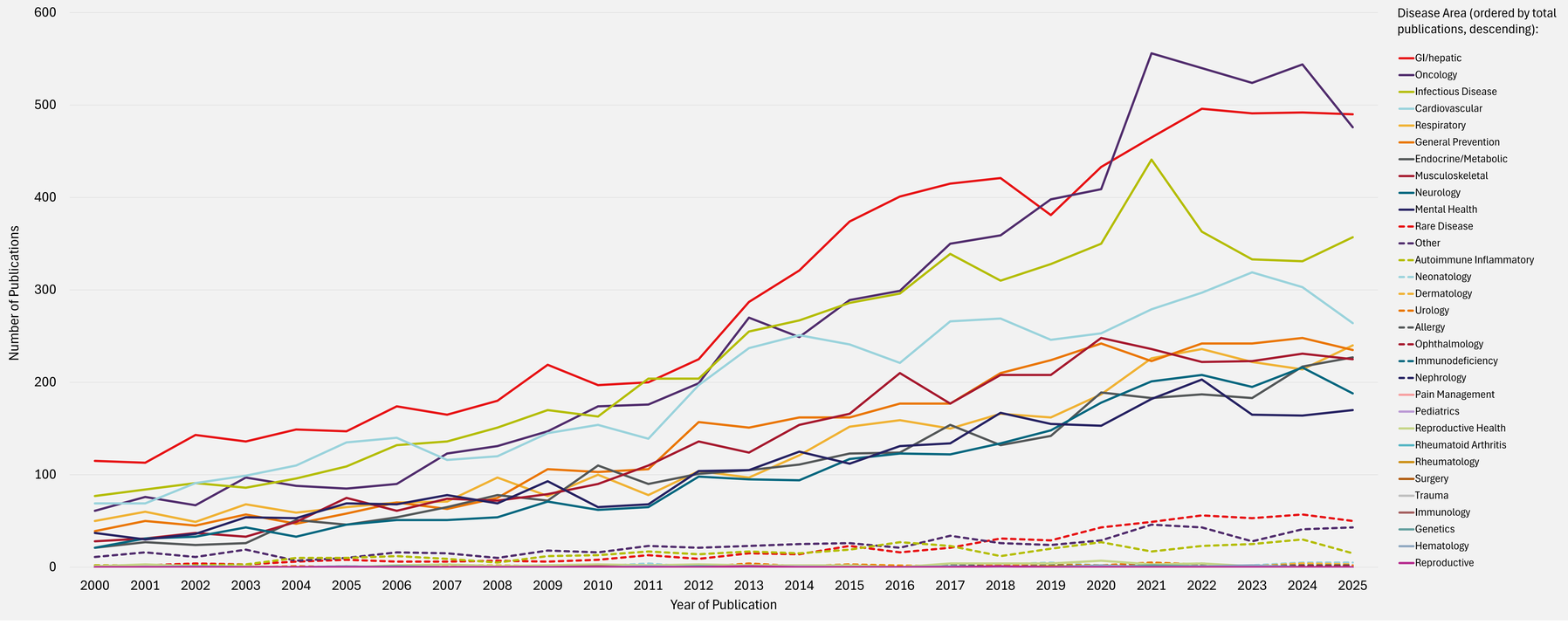


Figure 4. Trends in HEOR Publications by Disease Area, 2000 to 2025



Limitations

- Model performance was dependent on prompt design, and results may vary with different prompting strategies or task formulations.
- A general-purpose LLM was used and not specifically trained or fine-tuned for HEOR applications.
- The lack of a well-defined gold standard for disease-area and intervention classification limits the ability to fully benchmark model performance.
- Classification accuracy was moderate, indicating potential misclassification at both the screening and disease-area levels.

Conclusions

- This AI-assisted pipeline demonstrated the feasibility and value of large-scale HEOR screening and classification, highlighting disease areas with different representation.
- Although classification accuracy was moderate, results support opportunities for improved prompting, domain-specific LLMs, and human-guided workflows.
- Human-guided review remains important for interpreting disease-area and intervention classifications.

Disclosures

WY, CM, and DZ are employees of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific. Funding provided by Thermo Fisher Scientific.

Acknowledgments

Editorial and graphic design support were provided by Karissa Calara and Shani Berger of Thermo Fisher Scientific.