DEVELOPMENT AND CLINICAL VALIDATION OF A GENERATIVE AI ASSISTED MEDICATION-INDICATION KNOWLEDGE BASE

Florence Ma[1], Alice Beattie[1], Linxuan Zhang[1], Saniya Deshpande[1], Mattia Ficarelli[1], Edward Bonnet[1], Alistair Marsland[1], Benjamin D. Bray[1]

Corresponding author: ben.bray@lcp.uk.com

No funding was provided for this study.

Presented at the international Society for Pharmacoepidemiology (ISPE) Annual Meeting, 24-28 August 2024, Berlin, Germany

1 Health Analytics, Lane Clark & Peacock LLP. London, UK

https://www.lcp.com/en

Summary

+ Information on medication-indication relations is extensively used in biomedical research and pharmacoepidemiology. Existing knowledge bases are limited by their insufficient representation of actual clinical practice or are challenging to access and use in analytical workflows.

+ We aimed to create a medication-indication knowledge base using generative AI large language models (LLMs) and validate the accuracy of this compared to clinician knowledge.

+ 10,853 medication-indication pairs were generated. Out of the 465 pairs that were randomly sampled and had a high LLM-generated confidence score, 418 pairs (89.9%) were assessed to be clinically correct. We observed a clear relationship between the confidence score and accuracy from clinical checking. The proportion of errors detected suggested that such knowledge base should not be used in clinical practice but has potential value in biomedical research and high throughput pharmacoepidemiology research.

Background +

- Information regarding the medications and their indications is extensively utilised in biomedical research and pharmacoepidemiology. 1 However, this data is not yet readily available in a standardised format and typically requires significant manual creation and checking by clinical experts.
- Existing databases are limited by their insufficient coverage of off-label use and/or lack of machine-readable formats2-7, which are essential for enabling real-world evidence analytical workflows.
- Recent generative AI large language models (LLMs) such as GPT4 have demonstrated near-physician competency in tests of clinical knowledge. This offers a significant opportunity to develop a comprehensive, LLM-assisted medication-indication knowledge base.

Methods +

The medication-indication knowledge base was developed in four steps as outlined below (Figure 1).

Extraction of unique medication entities

Medication names were extracted from the English Prescribing Dataset, which includes information about medicines dispensed in community pharmacies in England (excluding categories such as food and nutrition products, vaccines, anaesthetics, appliances, and dressings).

Prompt engineering

Prompts were developed for use with the GPT4 LLM[8] to generate based on medications a list of indications, purpose (treatment vs prevention), and confidence scores on a scale of 0 to 1 for each of the indications identified.

Prompts were iteratively tested and optimised on a sample of medicines-indication pairs. The final prompt was uniformly applied across all medications.

Generation of medication-indication pairs using GPT-4 LLM

The resulting output went through a data cleaning pipeline to remove duplications, standardise spellings, fix spelling errors and tackle clinical synonyms.

The indications were standardised to SNOMED and ICD-10 terminologies where applicable using the MedCAT LLM, which has been trained on a large corpus of electronic health records[9].

Stratified sampling and clinical validation

Samples stratified by the model-generated confidence score

Results were clinically validated by three independaent reviewers (two physicians and a pharmacist)

Figure 1: Process of developing a medication-indication knowledge base

Results +

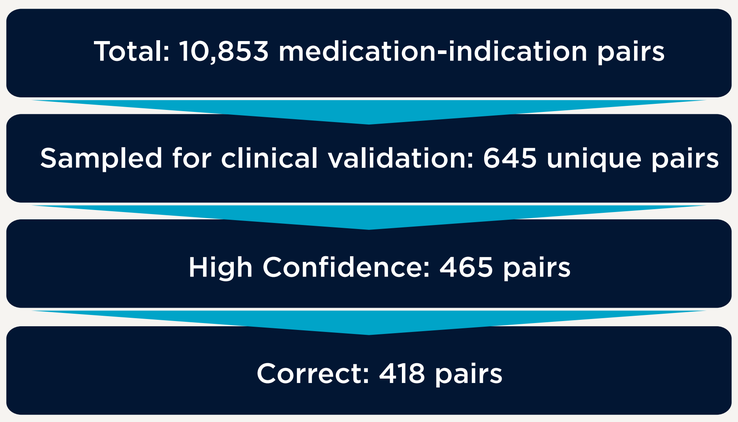

Figure 2: Attrition diagram of clinical validation

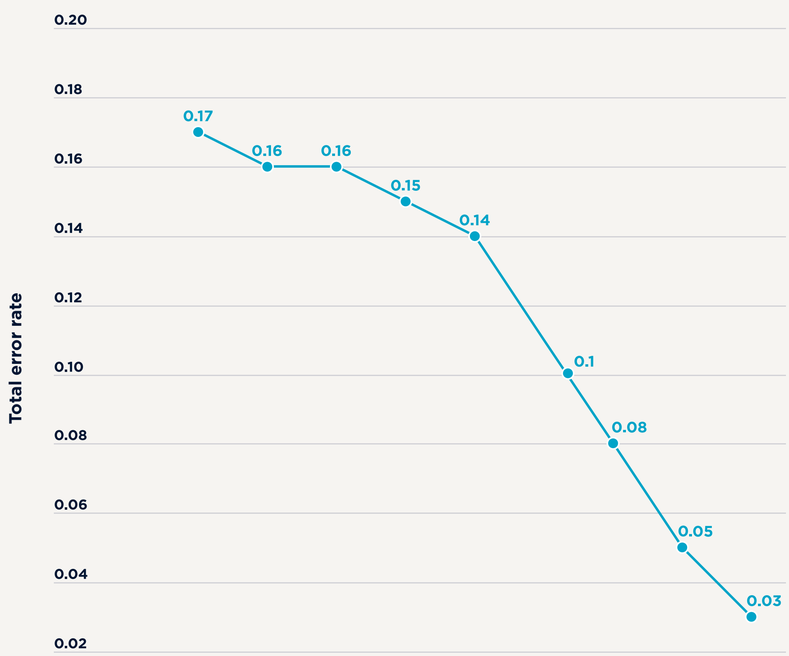

Figure 3: Total error rates of the knowledge base with each confidence score threshhold

- 10,853 medication-indication pairs with associated confidence scores were generated based on 1,540 unique medicine entities.
- Clinical validation was conducted on 645 unique pairs (5.94% of 10,853 pairs) by stratified random sampling. 465 pairs had a confidence score of at least 0.75 (where 1.0 = maximum confidence). 418 out of the 465 pairs were assessed to be clinically correct, demonstrating a precision rate of 89.9% (Figure 2).
- Both licensed and off-label indications were included in the output. We observed various types of hallucinations with erroneous indications in the output, especially at low confidence scores. There was a clear relationship between the GPT4-assessed confidence scores and the accuracy from clinical checking (Figure 3).
- For outputs that were clinically correct, additional data cleaning and standardisation was needed for the majority of the medication-indication pairs (Table 1).

Medication	Example Medication	Indication(s)
Clinically correct, but inconsistent terminology	Alendronic acid/ colecalciferol	“Paget’s disease”, “Paget’s disease of the bone”
	Benozoyl peroxide	“Acne”, “Acne vulgaris”
	Exemestame	“Advanced breast cancer”, “Advanced-stage breast cancer”
Clinically incorrect indication	Co-tenidone	“Cardiac arrythmias”
	Alprazolam	“Depression”

Table 1: Examples of medication-indication pairs in the knowledge base that were either clinically correct but lacked standardisation, or where clinically incorrect

Conclusions +

Using a combination of a general-purpose foundation LLM (GPT-4) and a health data specific model optimised for natural language processing of electronic health records (MedCAT), we have developed a medication-indication knowledge base suitable for use in health data analytics applications and high throughput pharmacoepidemiological research.

The large number of medication-indication pairs generated, and the high precision rate observed during clinical validation highlighted LLM’s potential value in biomedical research. The proportion of errors detected suggested that such knowledge base should not be used in clinical practice.

Future work will focus on expanding the dataset and improving model accuracy. Furthermore, our approach can be adapted to an extended range of data sources to maintain the relevance and accuracy of the knowledge base.

Further reading:

1. Zhu Y, Elemento O, Pathak J, Wang F. Drug knowledge bases and their applications in biomedical informatics research. Briefings in Bioinformatics. 2019 Jul;20(4):1308–21.
2. Khare R, Li J, Lu Z. LabeledIn: Cataloging Labeled Indications for Human Drugs. Journal of Biomedical Informatics. 2015 Dec;52:448–56.
3. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research. 2015 Oct;44(D1):D1075–9.
4. Bhatt A, Roberts R, Chen X, Li T, Connor S, Hatim Q, et al. DICE: A Drug Indication Classification and Encyclopedia for AI-Based Indication Extraction. Frontiers in Artificial Intelligence [Internet]. 2021;4. Available from: https://www.frontiersin.org/articles/10.3389/frai.2021.711467
5. Moodley K, Rieswijk L, Oprea T, Dumontier M. InContext: curation of medical context for drug indications. Journal of Biomedical Semantics [Internet]. 2021;12. Available from: https://doi.org/10.1186/s13326-021-00234-4
6. Jung K, LePendu P, Chen WS, Iyer SV, Readhead B, Dudley JT, et al. Automated Detection of Off-Label Drug Use. Sarkar IN, editor. PLoS ONE. 2014 Feb 19;9(2):e89324.
7. Joint Formulary Committee. British National Formulary (BNF) [Internet]. NICE; 2024. Available from: https://bnf.nice.org.uk/
8. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 Technical Report [Internet]. arXiv; Available from: https://doi.org/10.48550/arXiv.2303.08774
9. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artificial Intelligence in Medicine. 2021 Jul;117:102083.