Integrating large language models and real-world evidence into an automated drug indication taxonomy development workflow

Yilu Fang[1], Chunhua Weng[1,*], Patrick Ryan[1, 2,*]
[1] Department of Biomedical Informatics, Columbia University, [2] Janssen Research and Development, *equal contribution senior authors

Background

Drug indications refer to signs, symptoms, diseases, or conditions that the medication can treat or prevent.[1] It helps healthcare professionals readily identify appropriate treatments for patients[2]. Structuring drug indications can further facilitate clinical knowledge management and support the secondary use of electronic health records (EHR) data.[3] Multiple previous efforts have been made to extract indications from drug product labels[4-7] or generate medication-indication knowledge bases[1,8]. To structure the identified indications, a common practice is to map them to existing ontologies. However, this may potentially encounter semantics issues such as mismatched granularity and incomplete coverage. An alternative approach is to establish a taxonomy for drug indications directly. However, almost none of the research has investigated the subsumption relations between indications, and no automatic process is available to create drug indication taxonomy. This brings challenges in conducting research built upon drug-indication relations. Therefore, this study aims to create an automatic process that identifies the verbatim indication terms from drug product labels, derives subsumption relations, and further creates a taxonomy to use as a basis for organizing drugs. It harnesses the power of both large language models (LLM) and real-world evidence (RWE).

Methods

The automated workflow is described in Figure 1.

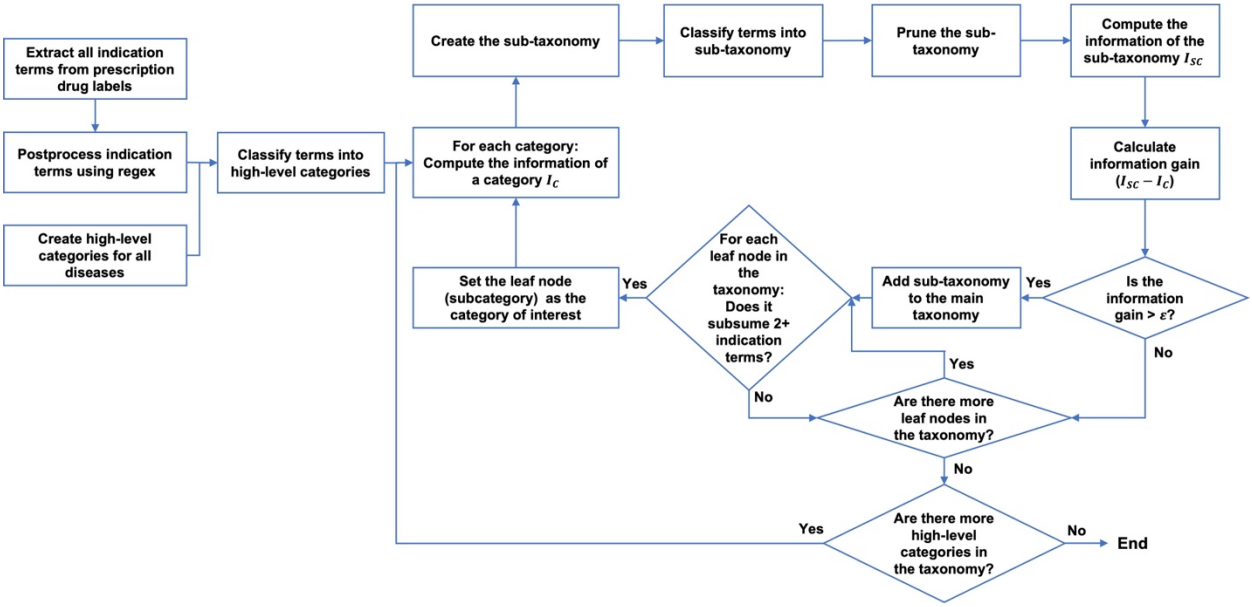

Figure 1. Automated workflow of developing the drug indication taxonomy.

Drug indication identification

DailyMed (https://dailymed.nlm.nih.gov/dailymed/) stores FDA-approved product labels, where the Indications and Usage section of drug labeling provides information on the approved indications.[2] We first retrieved all human prescription drug labels from this source and identified the active moieties of the drugs. For each active moiety, we extracted the drug indications using GPT-4. We further post-processed the indications, including lowercase conversion, punctuation removal, and suffix substitution.

Subsumption relation derivation

We generated various interpretations of the high-level categories within a disease taxonomy by GPT-4 and the best one was selected through expert evaluation. We classified indication terms into categories by GPT-4. Then for each category, we derive subcategories and the subsumption relations between them to distinguish the related indication terms recursively. We define the information of a category as the similarity among M subsumed indications, characterized by representations of related drugs based on real-world evidence[9]. After calculating the information of the category, we created the subcategories for it and further classified indication terms using GPT-4. This hierarchical structure can be viewed as a sub-taxonomy where the root node is the category of interest. We then computed the information of the subcategories I_{SC}, defined as the average information of nodes in the sub-taxonomy. We calculated the information gain from the inclusion of those subcategories. If the gain was larger than the threshold ϵ, we added the subcategories into our overall hierarchical structure. Further, for each leaf node in this sub-taxonomy, if it subsumes two or more indications, we set it to be the category of interest and iteratively searched for its subcategories.

Results

The current workflow is under development, and the results are considered preliminary.

2,560 distinct active moieties were identified from 46,421 human prescription drug labels from DailyMed. The median number of drug labels per active moiety is 4. We extracted 4,190 indication terms. After post-processing, we had 2,909 indications left. The median number of indication terms per active moiety is 2. We linked drug labels (structured product labels (SPL)) to RxNorm, where 2,219 out of 2,909 indication terms had corresponding RxNorm drugs (1,177 distinct drugs in total). For the remaining 690 indication terms lacking a matched RxNorm drug, we set the similarity they involved to 0.85.

24 high-level categories of the indication taxonomy and their primary axes were identified, and the related summary statistics are presented in Table 1.

We selected the category 'genitourinary system diseases' as an example for demonstration. 314 indication terms fall under this category, and 261 RxNorm drugs are linked to at least one of these indications. The hierarchical structure of this category contains 684 nodes (subcategories), with a depth of 10 levels. The second level of this sub-taxonomy consists of 11 nodes. There are 482 leaf nodes in total. The median number of indications subsumed by a leaf node is 1. An illustrative portion of this sub-taxonomy is displayed in Figure 2.

Table 1. 24 high-level categories of the indication taxonomy and related summary statistics.

Index	Category	Number of indications	Number of drugs
0	cardiovascular diseases	234	190
1	respiratory diseases	209	220
2	digestive system diseases	311	268
3	nervous system diseases	368	304
4	musculoskeletal diseases	135	138
5	endocrine system diseases	269	176
6	immune system diseases	353	297
7	infectious diseases	521	313
8	mental disorders	85	102
9	neoplasms (cancer)	532	193
10	skin diseases	265	258
11	eye diseases	101	81
12	ear, nose, and throat diseases	107	151
13	genitourinary system diseases	314	261
14	blood diseases	311	226
15	congenital, hereditary, and neonatal diseases	267	143
16	nutritional and metabolic diseases	206	157
17	pregnancy complications	154	222
18	substance-related disorders	42	37
19	injuries, wounds, and traumas	82	102
20	poisoning, toxicity, and environmental exposure	56	27
21	rare diseases	880	351
22	aging-related diseases	228	356
23	others	250	253

*Statistics on drugs are based on the available 1,177 distinct RxNorm terms that are linked to the indications in the high-level category of interest.

13.5. : kidney diseases
['acute glomerulonephritis', 'acute nephrosis', 'acute pyelonephritis', 'advanced renal cell carcinoma rcc', 'anemia associated with chronic kidney disease ckd', …, + 52 terms]
13.5.1. : chronic kidney disease
['chronic kidney disease associated with type 2 diabetes', 'chronic kidney disease at risk of progression', 'chronic kidney disease stage 5', 'chronic kidney disease stages 3 and 4', …, +17 terms]
13.5.1.1. : lupus nephritis
['severe active lupus nephritis']
13.5.1.2. : nephrotic syndrome
['diabetic nephropathy with albuminuria greater than 300 mg day', 'nephrotic syndrome without uremia of the idiopathic type or that due to lupus erythematosus']
13.5.1.2.1. : membranous nephropathy
['nephrotic syndrome without uremia of the idiopathic type or that due to lupus erythematosus']
13.5.1.2.2. : secondary nephrotic syndrome
['diabetic nephropathy with albuminuria greater than 300 mg day']
13.5.1.2.3. : idiopathic nephrotic syndrome
['nephrotic syndrome without uremia of the idiopathic type or that due to lupus erythematosus']
13.5.1.3. : uremic syndrome
['chronic kidney disease stage 5', 'end stage kidney disease eskd', 'renal failure', 'states of diminished renal function']
13.5.1.4. : stage 3 ckd (moderate decrease in gfr)
['stage 3 chronic kidney disease', 'states of diminished renal function']
13.5.1.4.1. : stage 3a ckd (gfr 45-59 ml/min/1.73 m2)
['stage 3 chronic kidney disease']
13.5.1.5. : renal cysts
['polycystic kidney disease']
13.5.1.6. : stage 4 ckd (severe decrease in gfr)
['chronic kidney disease at risk of progression', 'renal failure', 'stage 4 chronic kidney disease', 'states of diminished renal function']
13.5.1.7. : stage 5 ckd (kidney failure)
['chronic kidney disease stage 5', 'end stage kidney disease eskd', 'renal failure']
13.5.1.7.1. : end-stage renal disease (esrd)
['chronic kidney disease stage 5', 'end stage kidney disease eskd', 'renal failure']
13.5.1.7.2. : uremia
['chronic kidney disease stage 5']
13.5.1.7.3. : complications due to kidney transplant
['renal failure']
13.5.1.8. : diabetic nephropathy
['chronic kidney disease associated with type 2 diabetes', 'diabetic nephropathy with albuminuria greater than 300 mg day']
13.5.1.9. : glomerulonephritis
['severe active lupus nephritis']
13.5.1.10. : polycystic kidney disease
['polycystic kidney disease']
13.5.2. : acute kidney injury
['hepatorenal syndrome with rapid reduction in kidney function']

Figure 2. Illustrative portion of the sub-taxonomy for the high-level category: 13. genitourinary system diseases. Indication terms that fall under the same category are enclosed in brackets.

Conclusion

Large language models (LLM) can be used for various taxonomy development activities, including medical term identification from free text, axes identification for medical concept subcategorization, sub-

taxonomy construction, and subsumption relation determination. However, LLM does not fully support an end-to-end process such as directly mapping terms to codes in the existing vocabularies or directly generating a complete taxonomy for the medical concept of interest. In summary, we proposed an automatic process integrating both LLM and RWE to generate an effective taxonomy, optimized to distinguish between drug indications and further organize the drugs. It enables large-scale phenotyping based on the similarity of patients who had drugs for diseases in the same placement within the taxonomy.

References

1. Li Y, Xiao C. Developing a data-driven medication indication knowledge base using a large scale medical claims database. *AMIA Summits on Translational Science Proceedings*. 2019;2019:741.
2. Indications and Usage Section of Labeling for Human Prescription Drug and Biological Products — Content and Format Guidance for Industry. https://www.fda.gov/files/drugs/published/Indications-and-Usage-Section-of-Labeling-for-Human-Prescription-Drug-and-Biological-Products-%E2%80%94-Content-and-Format-Guidance-for-Industry.pdf
3. Bhatt A, Roberts R, Chen X, et al. Dice: a drug indication classification and encyclopedia for ai-based indication extraction. *Frontiers in Artificial Intelligence*. 2021;4:711467.
4. Ursu O, Holmes J, Knockel J, et al. DrugCentral: online drug compendium. *Nucleic acids research*. 2016:gkw993.
5. Khare R, Li J, Lu Z. LabeledIn: cataloging labeled indications for human drugs. *Journal of biomedical informatics*. 2014;52:448-456.
6. Khare R, Wei C-H, Lu Z. Automatic extraction of drug indications from FDA drug labels. American Medical Informatics Association; 2014:787.
7. Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. *Journal of the American Medical Informatics Association*. 2013;20(3):482-488.
8. Salmasian H, Tran TH, Chase HS, Friedman C. Medication-indication knowledge bases: a systematic review and critical appraisal. *Journal of the American Medical Informatics Association*. 2015;22(6):1261-1270.
9. Bohn J, Gilbert JP, Knoll C, Kern DM, Ryan PB. Large-scale empirical identification of candidate comparators for pharmacoepidemiological studies. *medRxiv*. 2023:2023.02. 14.23285755.