Knowledge-guided Generative AI For Automated

Taxonomy Learning From Drug Labels

Yilu Fang; Patrick Ryan*; Chunhua Weng*

* Equally contributed senior authors

Presenter: Yilu Fang (PhD student at Columbia University)

06/18/2024

COLUMBIA COLUMBIA UNIVERSITY IRVING MEDICAL CENTER



Issues	More Content 🔻	Submit 🗸	Purchase	Alerts	About 🔻	Journal of the Ame
Article	Contents	JO	URNAL ARTICLE			
Abstract			0	U U	0	rtificial intelligence
Introductio	on	-	r automat	eata	ixonomy learni	ng from drug labels
Methods		<mark>)</mark> Yilu	Fang, MA, Patricl	k Ryan,	PhD, Chunhua Weng, PhD	Author Notes
Results		lou	rnal of the Americ	can Mea	lical Informatics Association	n ocae105
Discussion	ı		os://doi.org/10.10			<i>ii, ocacios,</i>
Conclusior	าร	Pul	olished: 24 May 2	2024	Article history 🗸	

Scan for our paper



Drug indication

- **Definition:** the approved conditions, diseases, or states for which a drug can be safely and effectively used.
- Formal representation of drug indication knowledge:
 - treatment identification;
 - clinical knowledge management;
 - secondary use of EHR

Challenge in structuring drug indications

- **Concept mapping:** indication term -> existing ontology (e.g., SNOMED-CT)
- Potential issue:
 - mismatch granularity
 - incomplete coverage

e.g.,

ulcerative colitis \checkmark (in SNOMED-CT) mild to moderate ulcerative colitis χ (not in SNOMED-CT) moderately to severely active ulcerative colitis χ

e.g.,

metastatic castration-resistant prostate cancer ✓ metastatic castration-sensitive prostate cancer X

Challenge in structuring drug indications

- **Concept mapping:** indication term -> existing ontology (e.g., SNOMED)
- Potential issue:
 - mismatch granularity
 - incomplete coverage

e.g.,

ulcerative colitis ✓ (in SNOMED) mild to moderate ulcerative colitis X (not in SNOMED) moderately to severely active ulcerative colitis X

e.g.,

metastatic castration-resistant prostate cancer ✓ metastatic castration-sensitive prostate cancer X

Why not establish a taxonomy for drug indications directly?

Aim of this study

Integrate large language models and real-world evidence to develop an automatic workflow to

- 1. extract verbatim indication terms from free-text drug product labels
- 2. derive subsumption relations
- 3. create a drug indication taxonomy to organize drugs

Indication concepts:

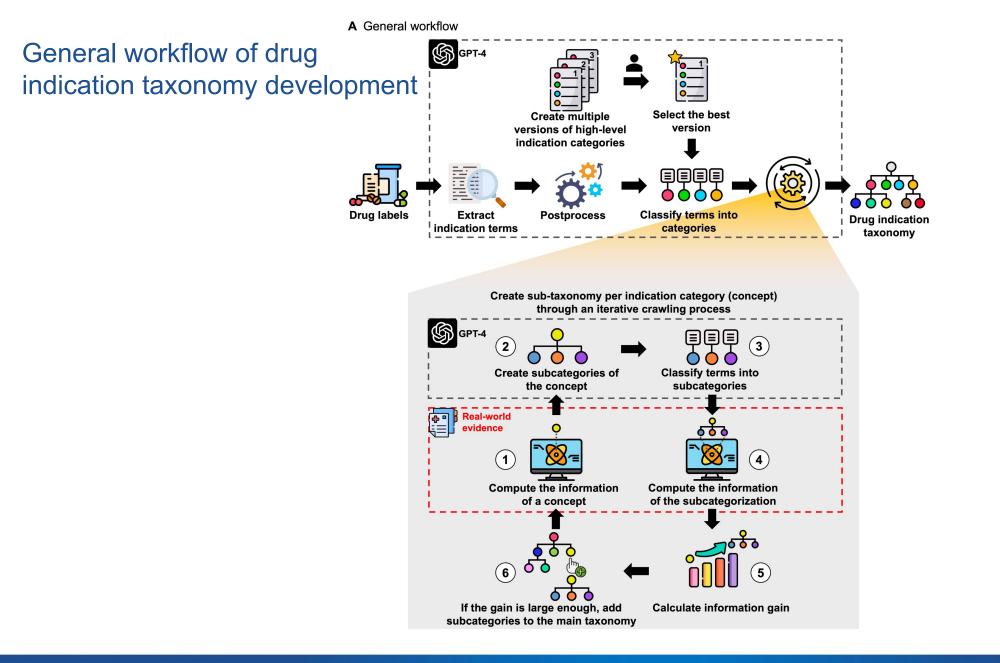
• nodes in the taxonomy

Indication terms:

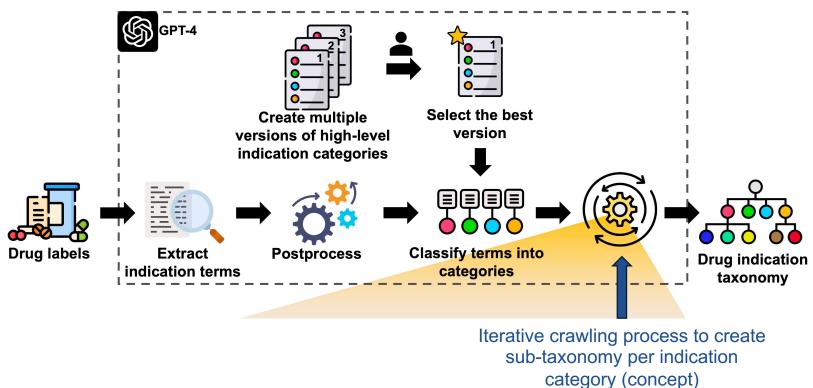
- mentions extracted from free-text drug labels
- linked to concepts representing their closest ancestors within the taxonomy

An ideal drug indication taxonomy:

- differentiate distinct indication terms
- semantically equivalent terms are linked to the same concept, no further distinction made with child nodes
- Need to check concept-to-concept and concept-to-term subsumption relations



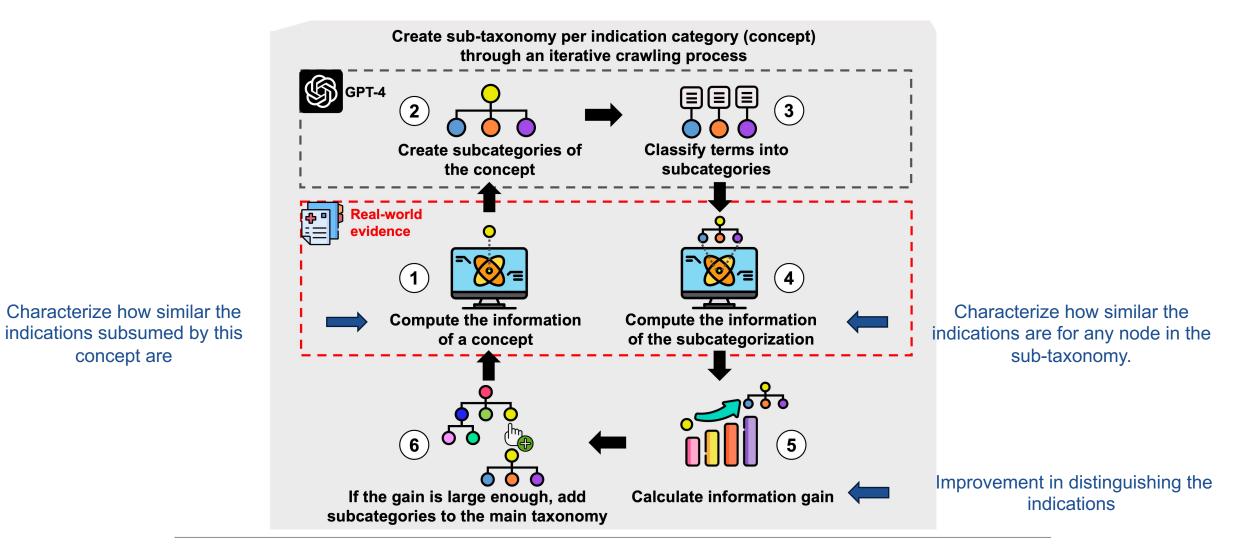
General workflow of drug indication taxonomy development



A General workflow

COLUMBIA COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

General workflow of drug indication taxonomy development



B Demonstration example

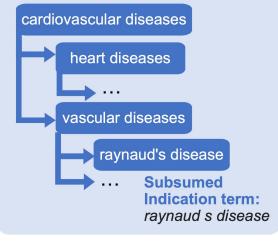


"... In peripheral vascular disease of arteriosclerosis obliterans, thromboangitis obliterans (Buerger's disease) and Raynaud's disease..."

Extract and postprocess indication terms

- peripheral vascular disease of arteriosclerosis obliterans;
- thromboangitis obliterans buerger s disease;
- raynaud s disease

Construct the taxonomy and link terms to concept nodes



COLUMBIA COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

Results: Extracted indication terms

Among 46,421 human prescription drug labels from DailyMed as of Feb 2023:

- 4,190 indication terms
- 2,909 terms after postprocessing
- 1177 RxNorm drugs that indication terms are linked to.

Results: High-level categories of the taxonomy

Index	Category	Number of indication terms	Number of drugs
0	Cardiovascular diseases	234	189
1	Respiratory diseases	209	219
2	Digestive system diseases	311	267
3	Nervous system diseases	368	303
4	Musculoskeletal diseases	135	137
5	Endocrine system diseases	269	175
6	Immune system diseases	353	296
7	Infectious diseases	521	312
8	Mental disorders	85	101
9	Neoplasms (cancer)	532	192
10	Skin diseases	265	257
11	Eye diseases	101	80
12	Ear, nose, and throat diseases	107	150
13	Genitourinary system diseases	314	260
14	Blood diseases	311	225
15	Congenital, hereditary, and neonatal diseases	267	142
16	Nutritional and metabolic diseases	206	156
17	Pregnancy complications	154	221
18	Substance-related disorders	42	36
19	Injuries, wounds, and traumas	82	101
20	Poisoning, toxicity, and environmental exposure	56	26
21	Rare diseases	880	350
22	Aging-related diseases	228	355
23	Others	250	252

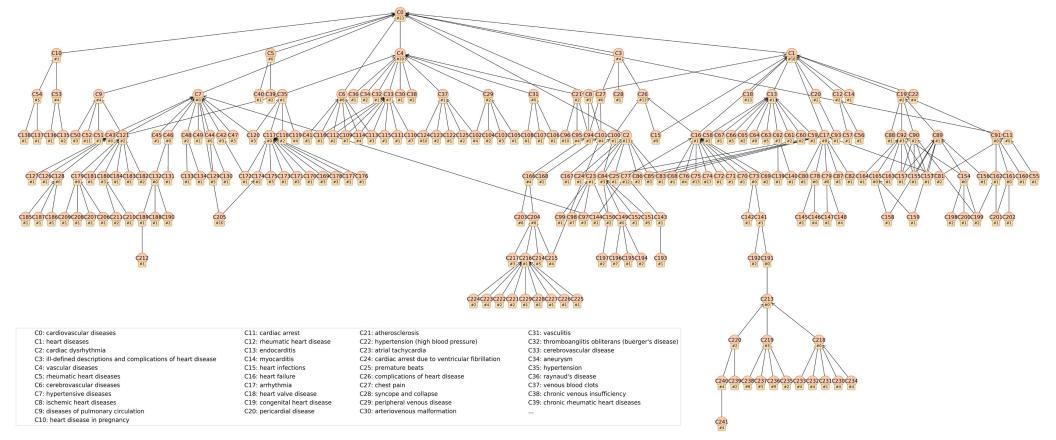
Results: Topological characteristics of the structure of the taxonomy

Index			0		5	13	
I	High-level categ		r of levels from the root	Cardiovascı diseases	-	Endocrine system diseases	Genitourinary system diseases
	Depth		e deepest node		11	14	10
the environ	har of shild	Median [Q1, Q3] Min			3 [2, 5]	3 [2, 5]	3 [2, 5]
\A/idth	ber of child a node in the				2	2	2
structure		Мах			12	14	16
		Count			242	339	516
	Indication te	rms ner	Median [Q1, Q3]		1 [1, 3]	1 [1, 2]	1 [1, 2]
	node	-	Min		0	0	0
Statistics of all unique			Мах		23	78	36
nodes	Number of nodes with 1+ RxNorm drugs				199	218	411
	Median [Q1, Q3] RxNorm drug per node Min		Median [Q1, Q3]		2 [1, 7]	2 [1, 4]	3 [1, 7]
				1	1	1	
		-	Мах		62	94	50
		Cour	nt		170	235	349
	Indication terms per		Median [Q1, Q3]		1 [1, 3]	1 [1, 2]	1 [1, 2]
	node	-	Min		1	1	1
Statistics of unique leaf			Max		17	78	33
nodes	Number of nodes with 1+ RxNorm drugs154169	325					
			Median [Q1, Q3]	2 [1, 6.75]	2 [1, 4]	2 [1, 5]
	RxNorm drug	per node	Min		1	1	1
	P		Max		62	94	50

Results: Topological characteristics of the structure of taxonomy

	Index		0	5	13	
I	High-level category name		Cardiovascular diseases	Endocrine system diseases	Genitourinary system diseases	
	Depth			14	14 10	
	Median [C	(1, Q3]	3 [2, 5]	3 [2. 5]	3 [2. 5]	
Width	Min		2	the most com	olex hierarchical structur	
	Max		12	14	16	
	Cour	nt	242	339	516	
	Indication terms per node	Median [Q1, Q3]	1 [1, 3]	1 [1, 2]	1 [1, 2]	
		Min	0	0	0	
Statistics of all unique		Мах	23	78	36	
nodes	Number of nodes with 1+ RxNorm drugs		199	218	411	
		Median [Q1, Q3]	2 [1, 7]	2 [1, 4]	3 [1, 7]	
	RxNorm drug per node	Min	1	1	1	
		Max	62	94	50	
	Cour	nt	170	235	349	
	Indication terms per	Median [Q1, Q3]	1 [1, 3]	1 [1, 2]	1 [1, 2]	
	node	Min	1	1	1	
Statistics of unique leaf		Max	17	78	33	
nodes	Number of nodes with 1+ RxNorm drugs		154	169	325	
		Median [Q1, Q3]	2 [1, 6.75]	2 [1, 4]	2 [1, 5]	
	RxNorm drug per node	Min	1	1	1	
		Max	62	94	50	

Results: Sub-taxonomy for high-level category: cardiovascular diseases



Meet the desiderata for controlled medical taxonomy

- Various levels of code granularity; Depth comparable to the counterparts in SNOMED-CT;
- Polyhierarchy
- Avoid the use of 'not elsewhere classified'
- Maintain consistency with no acyclicity

Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*. 1998;37(04/05):394-403.

Results: Accuracy of concept-to-concept and concept-to-term subsumption relations from 3 evaluators and their inter-rater reliability

Relationship type	High-level category	Number Accuracy				Gwet's AC1 coefficient (95%	
type	category	cases	Rater 1	Rater 2	Rater 3	CI)	
	Cardiovascul ar diseases	49	0.857	1	0.898	0.84 (0.733 0.947)]
Concept-to- concept	Endocrine system diseases	73	0.863	0.986	0.712	0.72 (0.596, 0.844)	Good to very good
	Genitourinary system diseases	75	0.84	0.907	0.787	0.857 (0.784, 0.93)	reliability
	Cardiovascul ar diseases	64	0.562	High accu	nacy 0.672	0.462 (0.285, 0.64)	
Concept-to- term	Endocrine system diseases	85	0.6	0.8	0.329	0.234 (0.086, 0.381)	
	Genitourinary system diseases	105	0.81	0.819	0.467	0.43 (0.293, 0.566)	

LLMs are good at constructing their own hierarchy for drug indications

COLUMBIA COLUMBIA UNIVERSITY IRVING MEDICAL CENTE

Results: Accuracy on concept-to-concept and concept-to-term subsumption relations from 3 evaluators and their inter-rater reliability

Relationship	High-level	Number of test		Accuracy	Gwet's AC1 coefficient (95%	
type	category	cases	Rater 1	Rater 2	Rater 3	CI)
	Cardiovascul ar diseases	49	0.857	1	0.898	0.84 (0.733 0.947)
Concept-to- concept	Endocrine system diseases	73	0.863	0.986	0.712	0.72 (0.596, 0.844)
	Genitourinary system diseases	75	0.84	0.907	0.787	0.857 (0.784, 0.93)
	Cardiovascul ar diseases	64	0.562	0.703	0.672	0.462 (0.285, 0.64)
Concept-to- term	Endocrine system diseases	85	0.6	0.8	0.329	0.234 (0.086, 0.381)
	Genitourinary system diseases	105	0.81	0.819	0.467	0.43 (0.293, 0.566)

fair to moderate reliability

Accuracies vary

Concept-to-term subsumption relation determination is complex.

Conclusion

- We proposed an automated pipeline to generate an effective taxonomy, optimized to distinguish between drug indications and further organize the drugs. (LLM+RWE)
- LLMs are **good** at deriving their **own concept hierarchies**
- LLMs fall short in determining the subsumption relations between concepts and terms in unregulated language from free-text labels (also a hard task for humans).
- The framework for taxonomy development can be applied beyond the context of drug indications.

Thank you!

(yf2505@cumc.columbia.edu)



Scan for our paper

Fang, Y., Ryan, P., & Weng, C. (2024). Knowledge-guided generative artificial intelligence for automated taxonomy learning from drug labels. *Journal of the American Medical Informatics Association*, ocae105