LLMs can support automated mapping of ICD to SNOMED concepts

Title: Large Language Model–Based Classification of ICD-10-CM to SNOMED Mappings for Improved Semantic Fidelity in OHDSI

Background: Concept mapping in OHDSI can lose/corrupt clinical meaning when source and target terminologies don't align perfectly. LLMs are proposed as a solution to define type of semantic equivalence of mappings.

Methods

1. Data retrieval. Candidate SNOMED CT targets and synonyms retrieved for each ICD-10-CM source concept with 1:1 mapping (tested on a subset yet).

2. ICD10CM descriptions normalization. Non-specific ICD-10-CM modifiers ("unspecified", "NOS", "NEC", etc.) removed from source names using GPT-4o, preserving clinically meaningful attributes.
"AND" in selected ICD10CM concepts is replaced with "OR"

3. Mapping classification steps
3.1. text based match: Pairs with identical bags of words (after lowercasing, hyphen normalization, and stop-word removal) labelled "equal" ;
3.2. remaining pairs classified by GPT-o3 reasoning model into five labels**: equal, uphill, downhill, up&downhill, or incorrect**, with confidence levels.
3.3. Low/moderate-confidence labels validated through adversarial re-check* and majority-vote tiebreaker; all-differ cases flagged for manual review.

4. Output preparation. Final mappings split into a consistent set (targets in agreement or ≥1 "equal") and an inconsistent set (conflicting labels)

5. Quality check:
5.1. rerunning the model several times on a same subset and expecting the same results
5.2. review by a medical professional

*built-in validation first call prompt part:
"A previous classification assigned the label '<label>' with <confidence> confidence. Carefully reconsider. Apply all rules strictly and return your final label and confidence."
2nd validation call is on contrary independent from the previous calls.

Results: see output examples below

concept code	orig_source_name	source_norm_final	target_concept_name	outcome step1	confidence step1	outcome step2	confidence step2	outcome step3	confidence step3	Final outcome	Final confidence	decision method
G02	Meningitis in other infectious and parasitic diseases classified elsewhere	Meningitis in infectious or parasitic diseases	Infective meningitis	equal	moderate	equal	high	NA	NA	**equal**	high	LLM (validated)
G04.00	Acute disseminated encephalitis and encephalomyelitis, unspecified	Acute disseminated encephalitis or encephalomyelitis	Acute disseminated encephalomyelitis	NA	NA	NA	NA	NA	NA	**equal**	high	LLM
G04.01	Postinfectious acute disseminated encephalitis and encephalomyelitis (postinfectious ADEM)	Postinfectious acute disseminated encephalitis or encephalomyelitis (postinfectious ADEM)	Acute disseminated encephalomyelitis following infectious disease	downhill	moderate	downhill	moderate	NA	NA	**downhill**	moderate	LLM (validated)
G05.3	Encephalitis and encephalomyelitis in diseases classified elsewhere	Encephalitis or encephalomyelitis	Encephalitis, myelitis and encephalomyelitis	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G12.1	Other inherited spinal muscular atrophy	Inherited spinal muscular atrophy	Spinal muscular atrophy	uphill	moderate	equal	moderate	uphill	high	**equal**	high	(tiebreak)
G31.09	Other frontotemporal neurocognitive disorder	Frontotemporal neurocognitive disorder	Frontotemporal dementia	NA	NA	NA	NA	NA	NA	**equal**	high	LLM
G31.83	Neurocognitive disorder with Lewy bodies	Neurocognitive disorder with Lewy bodies	Senile dementia of the Lewy body type	downhill	moderate	equal	moderate	equal	high	**equal**	high	(tiebreak)
G31.84	Mild cognitive impairment of uncertain or unknown etiology	Mild cognitive impairment of uncertain or unknown etiology	Minimal cognitive impairment	equal	moderate	equal	moderate	NA	NA	**equal**	moderate	LLM (validated)
G40.219	Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with complex partial seizures, intractable, without status epilepticus	Localization-related epilepsy or epileptic syndromes with complex partial seizures, intractable	Focal epilepsy	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G40.5	Epileptic seizures related to external causes	Epileptic seizures related to external causes	Epileptic seizure	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G40.8	Other epilepsy and recurrent seizures	Epilepsy or recurrent seizures	Epilepsy	downhill	moderate	equal	moderate	equal	moderate	**equal**	moderate	(tiebreak)
G40.824	Epileptic spasms, intractable, without status epilepticus	Epileptic spasms, intractable	Refractory infantile spasms	NA	NA	NA	NA	NA	NA	**downhill**	high	LLM
G04.3	Acute necrotizing hemorrhagic encephalopathy	Acute necrotizing hemorrhagic encephalopathy	Acute hemorrhagic leukoencephalitis	NA	NA	NA	NA	NA	NA	**incorrect**	high	LLM
G40.11	Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with simple partial seizures, intractable	Localization-related (focal) (partial) symptomatic epilepsy or epileptic syndromes with simple partial seizures, intractable	Focal onset aware epileptic seizure	NA	NA	NA	NA	NA	NA	**incorrect**	high	LLM
G40.A0	Absence epileptic syndrome, not intractable	Absence epileptic syndrome, not intractable	Absence seizure	uphill	moderate	incorrect	high	incorrect	moderate	**incorrect**	high	LLM (tiebreak)
G43.B0	Ophthalmoplegic migraine, not intractable	Ophthalmoplegic migraine, not intractable	Recurrent painful ophthalmoplegic neuropathy	uphill	moderate	uphill	high	NA	NA	**uphill**	high	(validated)

Out of 715 1:1 mappings of entire ICD10CM F chapter, 6 were labeled for manual review, 7 as incorrect, 45 as downhill, 4 as up&downhill, 154 as uphill, 510 as equal. After review by medical professional 17 labels were considered wrong - model was mostly confused by *harmful use* vs *abuse* vs *disorder related to use of psychoactive substance*

Conclusion: The automated mapping classification results are promising and mostly consistent. Non-equal mappings detected will be replaced with better target concept equivalents (decision to be made: what to do if no equivalent target exists)

OHDSI

Dmytro Dymshyts, Anna Ostropolets, Martijn Schuemie

Johnson & Johnson