Effect of (OHDSI) Vocabulary Mapping on Phenotype Cohorts

Matthew Levine, Research Associate

George Hripcsak, Professor

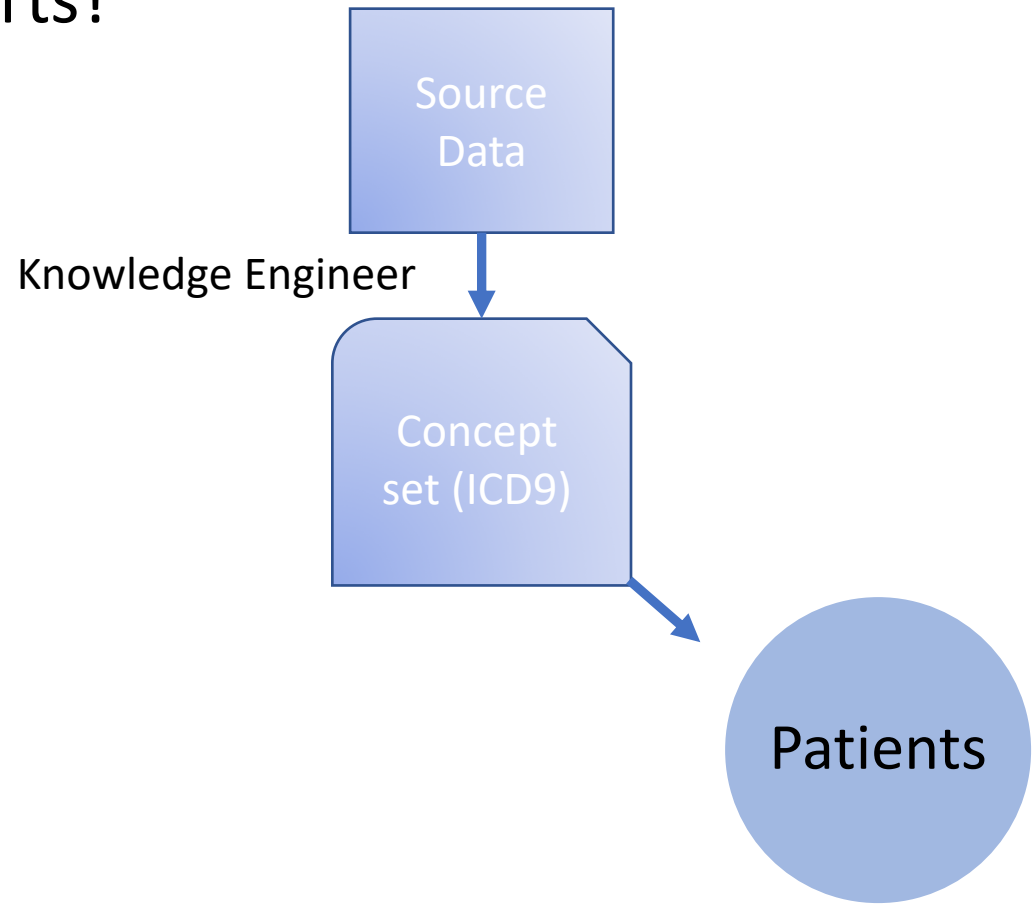
Department of Biomedical Informatics, Columbia University

Intro

- Reasons to map:
  - International (with only 5% of the world population, the US is limited in what observational research it can do)
  - join future data (ICD10-CM plus problem lists plus NLP)
  - shift away from billing codes as the primary means of specifying patients

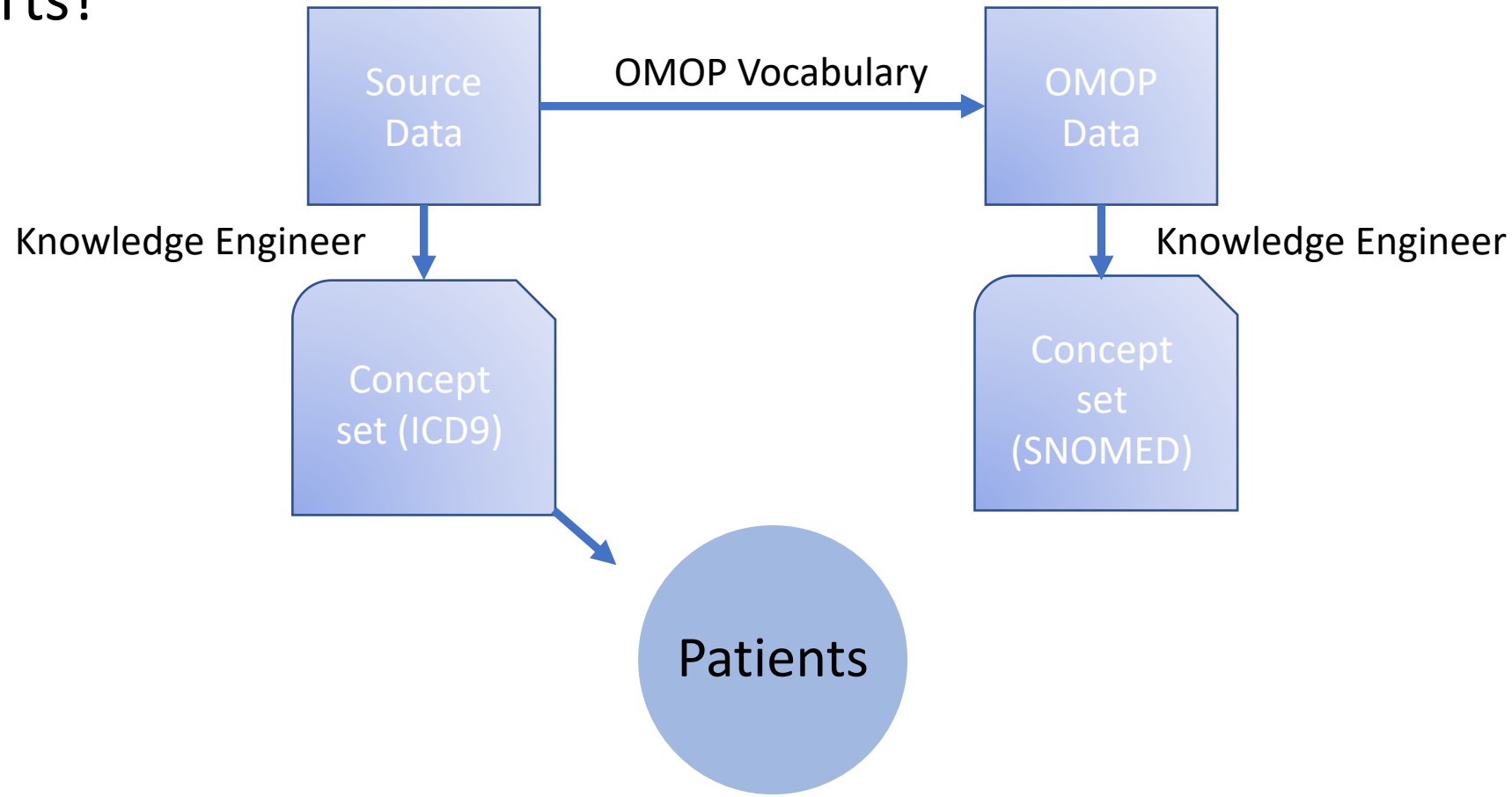
Question

To what extent does OHDSI preserve patient membership in phenotype cohorts?



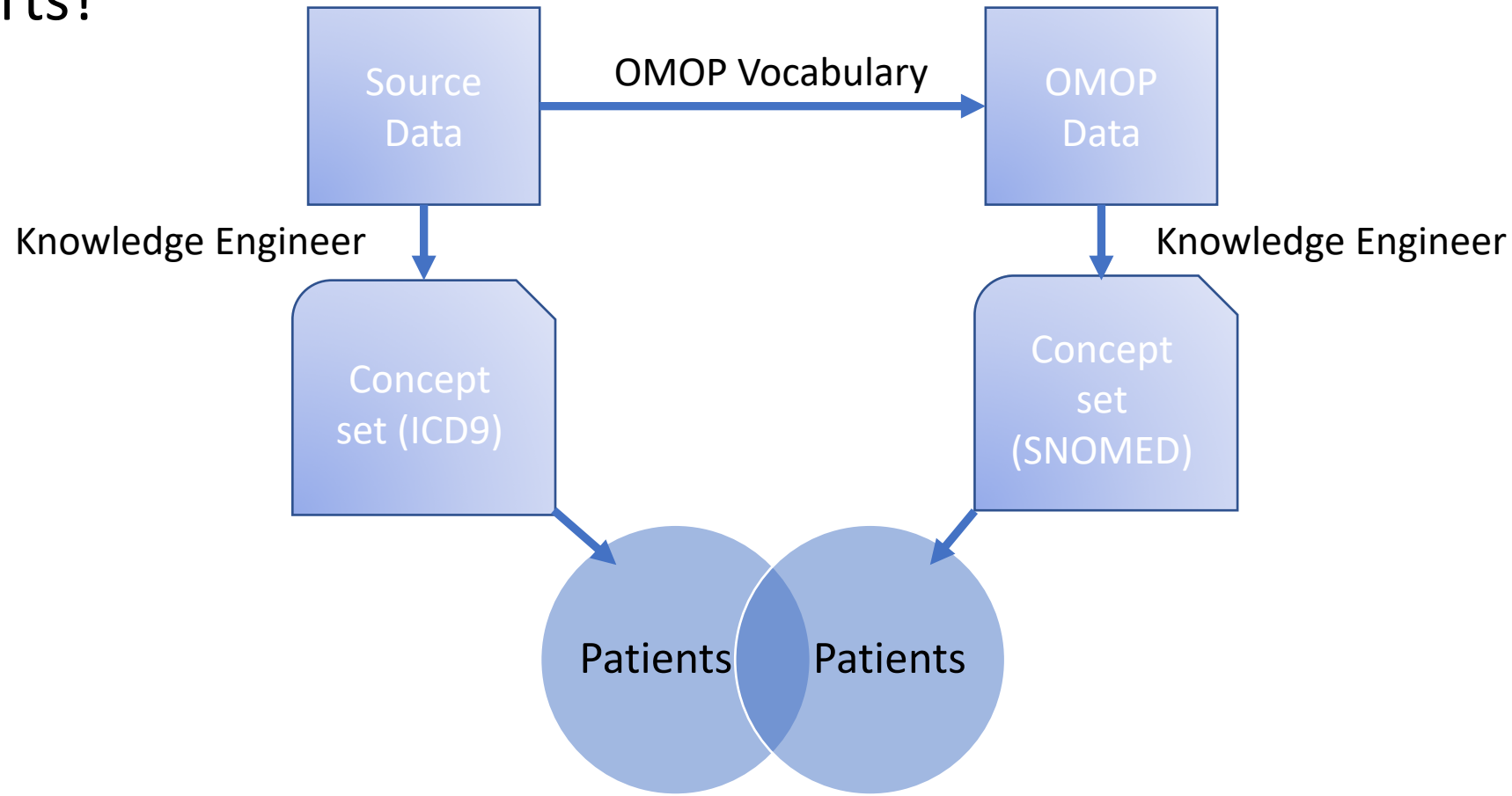
Question

To what extent does OHDSI preserve patient membership in phenotype cohorts?



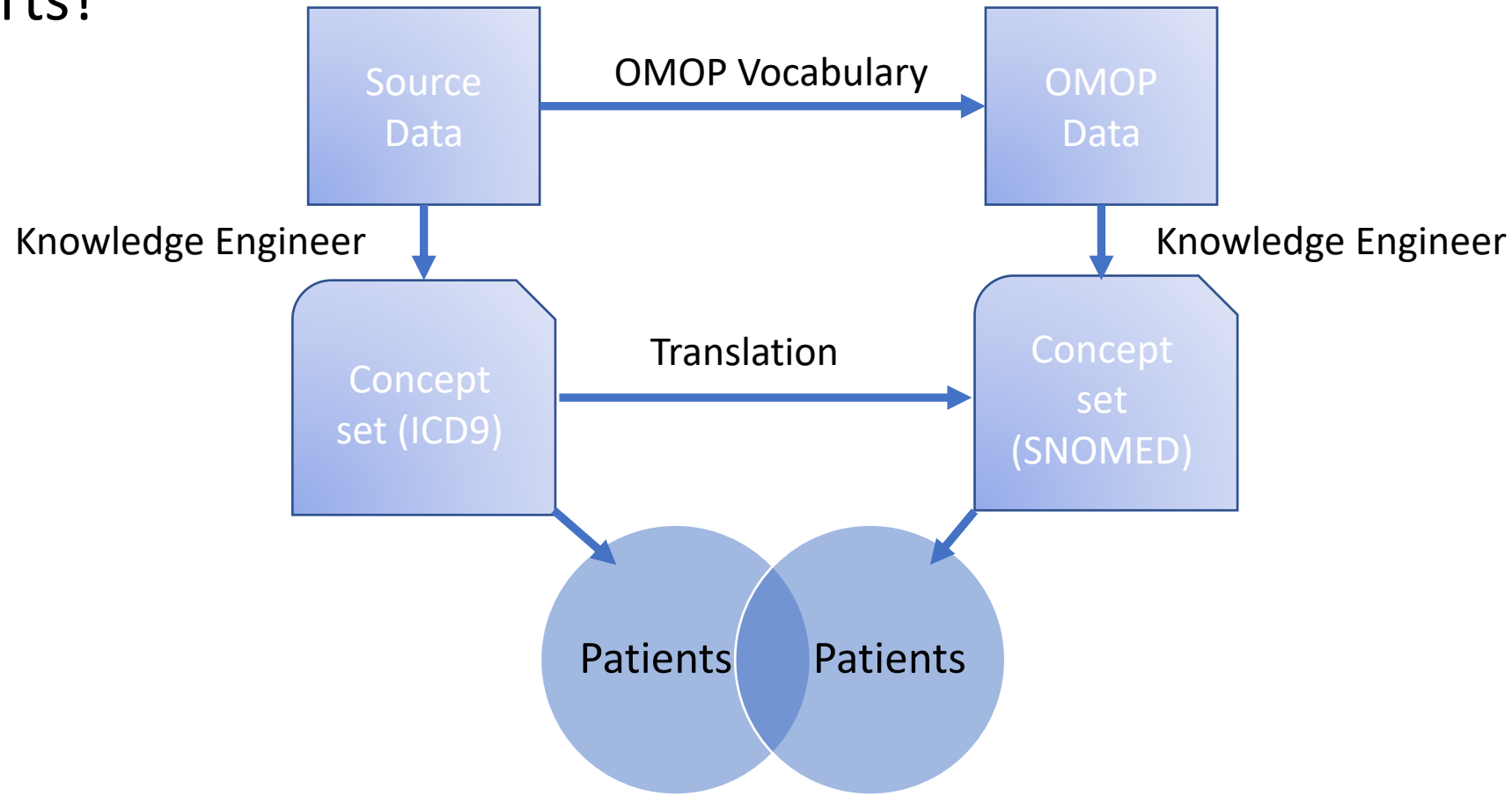
Question

To what extent does OHDSI preserve patient membership in phenotype cohorts?

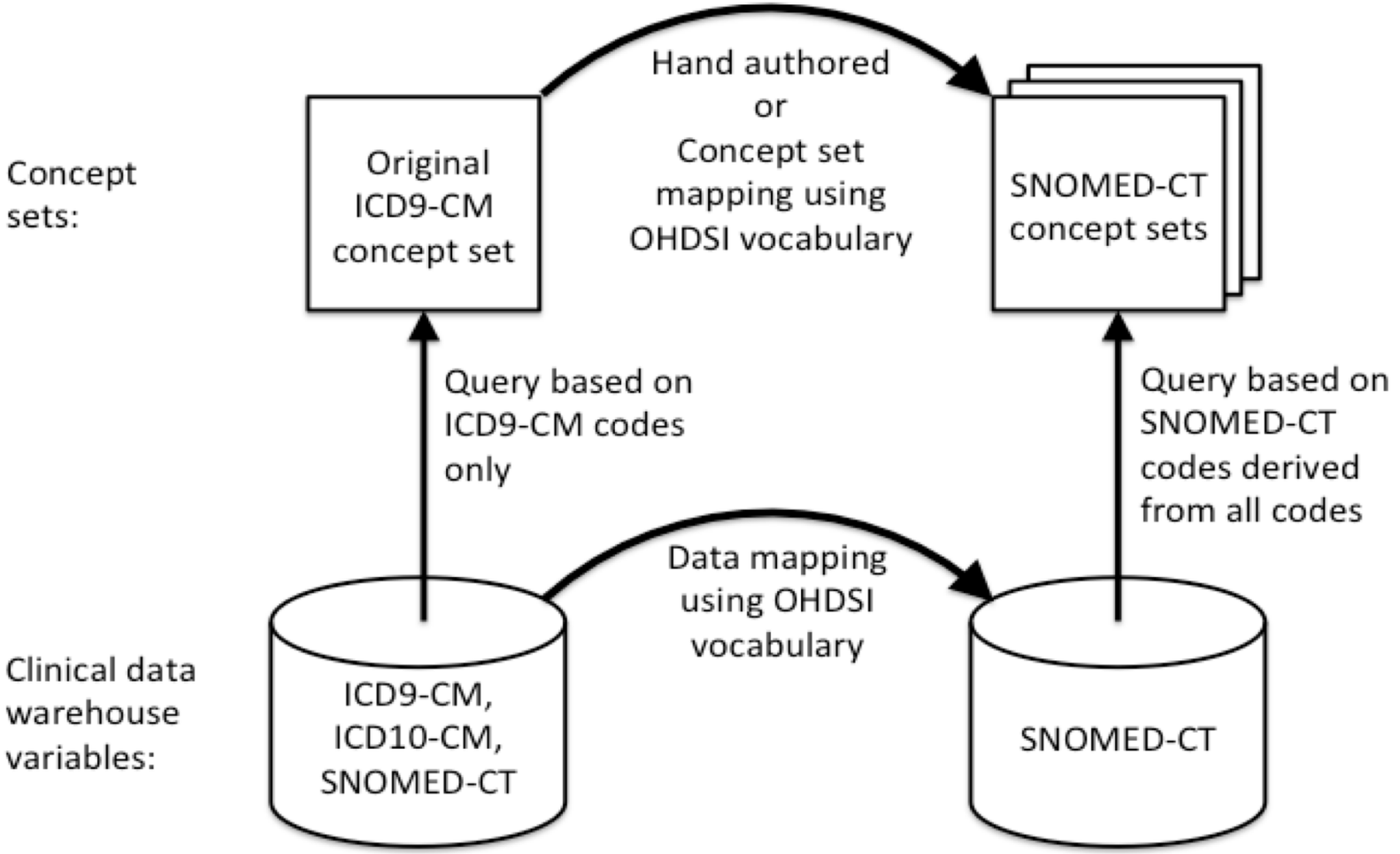


Question

To what extent does OHDSI preserve patient membership in phenotype cohorts?



Experimental Design: Compare patient cohorts defined by a) original concept sets on unmapped data, b) mapped concept sets on mapped data



Original ICD9-CM concept set: no mappings

Algorithm (from eMERGE)	Original ICD9-CM concept set‡
Heart failure (HF) [1]	428.*
Heart failure as exclusion diagnosis (HF2) [2]	428.*
Type-1 diabetes mellitus (T1DM) [3]	250.x1, 250.x3
Type-2 diabetes mellitus (T2DM) [4]	250.x0, 250.x2
Appendicitis (Appy) [6]	540.*
Attention deficit hyperactivity disorder (ADHD) [5]	314, 314.0, 314.01, 314.1, 314.2, 314.8, 314.9
Cataract (Catar) [7]	366.10, 366.12, 366.13, 366.14, 366.15, 366.16, 366.17, 366.18, 366.19, 366.21, 366.30, 366.41, 366.45, 366.8, 366.9
Crohn's disease (Crohn) [8]	555, 555.0, 555.1, 555.2, 555.9
Rheumatoid arthritis (RA) [9]	714, 714.0, 714.1, 714.2 (M05*, M06*)

What: Original ICD9-CM concept set generated by the phenotype author.

How: Run against patients' original ICD9-CM terms.

Why: Show what would have happened before either data or concept sets were mapped.

Author's INTENT source concept set: no mappings

- We extended original ICD9-CM concepts to include similar ICD10 and SNOMED CT codes
- This allows us to acquire a cohort, using unmapped data/queries, that would reflect the author's intent under a broader availability of terminologies
- Also corrected obvious errors in original concept set

Knowledge engineered concept set (map data only)

- What: By-hand SNOMED CT concept sets
- How: Run against OHDSI-mapped data in the form of SNOMED CT terms
- 2 intentions of concept set mapping:
  1. SNOMED ***"mimic"***
     - Designed to mimic the original ICD9-CM concept set as much as possible, ignoring data from other vocabularies
  2. SNOMED ***"optimize"***:
     - Designed to carry out phenotype author's intent to ICD9-CM, ICD10-CM, and SNOMED-CT

Knowledge engineered concept set: SNOMED mimic

- Mimic original ICD9-CM concept set as much as possible
- Create a SNOMED concept set expected to ONLY find people who had the original ICD9 codes
- E.g. Does not try to find patients who might have a related ICD10-CM code

Knowledge engineered concept set: SNOMED optimize

- Interpret and extend author's original intent (e.g. Find people with appendicitis)
- Find patients with relevant ICD9-CM, ICD10-CM, and SNOMED-CT codes (e.g. that would reflect author's intent, as demonstrated by their ICD9 concept set)

Automatically generated concept sets (map data AND concept sets)

- What: generated automatically from the original ICD9-CM set using OHDSI vocabulary mappings
- How: Run against OHDSI-mapped data in the form of SNOMED CT terms.
- 4 granularities for concept set mapping:
  1. SNOMED "no descendants"
     - OHDSI mappings from ICD9-CM to SNOMED-CT, ***without*** SNOMED hierarchy.
  2. SNOMED "all descendants"
     - includes descendants of mapped terms
  3. SNOMED "descendants x child"
     - includes descendants of mapped terms **only if none of the term's CHILDREN are also in the concept set.** (limited descendants)
  4. SNOMED "descendants x descendants"
     - includes descendants of mapped terms **only if none of the term's DESCENDANTS are also in the concept set.** (more limited descendants)

Results – code level

- Some knowledge engineered mappings just worked
- Multiple source codes (ICD) to one standard code (SNOMED CT)
- One source code (ICD) to multiple standard codes (SNOMED CT)
- Missing OMOP codes
- Information gain

Results – code level

- Some knowledge engineered mappings just worked
  - Acute appendicitis – 1 code and descendants
  - Crohn’s disease – 2 codes and descendants
  - Heart failure as an exclusion diagnosis – 3 codes and descendants
  - Heart failure as an inclusion diagnosis – 29 code and some descendants

Results – code level

- Multiple source codes (ICD) to one standard code (SNOMED CT)
  - Biggest challenge
  - Causes ambiguity so that either need to gain or lose patients
  - Attention deficit hyperactivity disorder
    - Includes ICD9-CM 314.0 “Attention deficit disorder of childhood” but excludes its child, 314.00 “Attention deficit disorder without mention of hyperactivity”
    - Both of these terms map to SNOMED CT 192127007 “Child attention deficit disorder”
  - Type-2 diabetes mellitus
  - Type-1 diabetes mellitus
  - Rheumatoid arthritis
  - Cataract

Results – code level

- One source code (ICD) to multiple standard codes (SNOMED CT)
  - Mostly happens with compound ICD9-CM terms
  - Can solve with conjunction in SNOMED
  - Type-1 diabetes mellitus
    - More of an oversight; did not need conjunction
  - Type-2 diabetes mellitus
    - More of an oversight; did not need conjunction

Results – code level

- Missing OMOP codes
  - Generally new codes that have not been added to OHDSI yet
  - Type-1 diabetes mellitus
    - Not used for patient data yet so no consequence
  - Type-2 diabetes mellitus
    - Not used for patient data yet so no consequence

Results – code level

- Information gain
  - Superior hierarchy of SNOMED CT versus strict hierarchy of ICD9-CM can improve the concept set
  - E.g., find codes in different areas of the ICD9-CM hierarchy
  - Heart failure as an exclusion diagnosis
    - Added terms

Results – patient cohort level

- Mapped Cohorts vs Original ICD9 cohort
- Mapped Cohorts vs Author’s Intent
- Automated batch analysis
  - 122 eMERGE concept sets
  - original ICD9 cohort vs AUTO mappings

Results: Appendicitis

-0 patient loss with any query

Pheno	#Cases	ICD9 set‡		SNOMED mimic		SNOMED optimize		SNOMED no desc		SNOMED all desc	
		Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss
VS original	9,887	0	0	0	0	0	0	0	0	0	0

‡ This column is used as the gold standard and therefore must have perfect performance

Results: Appendicitis

-0 patient loss with any query (vs original)
-same queries are extendable to capture author's intent

Pheno	#Cases	ICD9 set‡		SNOMED mimic		SNOMED optimize		SNOMED no desc		SNOMED all desc	
		Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss
VS original	9,887	0	0	0	0	0	0	0	0	0	0
VS intent	9,920	0	33	0	0	0	0	0	8	0	0

‡ This column is used as the gold standard and therefore must have perfect performance

Results: ADHD

-no perfect mapped query
-tradeoff between gain and loss
-automated don't perform well
-0.1% loss from knowledge engineering is probably better than 9.4% gain with auto-mapping
-KE queries are extendable to capture author's intent

Pheno	#Cases	ICD9 set‡		SNOMED mimic		SNOMED optimize		SNOMED no desc		SNOMED all desc	
		Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss
VS original	14,399	0	0	0	19	0	19	1362	0	1362	0
VS intent	14,547	0	148	0	39	0	19	1359	19	1359	0

Results: Mapped Cohorts vs Original ICD9 cohort

-0 patient loss from automatic mappings
-MIMIC does a good job a mimicing the patient cohort

Pheno	#Cases	ICD9 set‡		SNOMED mimic		SNOMED optimize		SNOMED no desc		SNOMED all desc	
		Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss
HF	75,312	0	0	0	0	0	0	0	0	1262	0
HF2	75,312	0	0	0	0	0	0	0	0	1262	0
T1DM	27,861	0	0	0	23	0	23	108	0	943	0
T2DM	125,342	0	0	3	30	3	30	34	0	1318	0
Appy	9,887	0	0	0	0	0	0	0	0	0	0
ADHD	14,399	0	0	0	19	0	19	1362	0	1362	0
Catar	50,879	0	0	50	0	74	0	50	0	2491	0
Crohn	4,679	0	0	0	0	0	0	0	0	0	0
RA	9,655	0	0	0	16	0	16	0	0	25103	0

‡ This column is used as the gold standard and therefore must have perfect performance

Results: Mapped Cohorts vs Author's Intent

- Author's Intent brings in new patients, compared to only original ICD9 codes.

Pheno	#Cases	ICD9 set	
		Gain	Loss
HF	75,626	0	314
HF2	76,958	0	1646
T1DM	27,935	0	74
T2DM	126,828	0	1486
Appy	9,920	0	33
ADHD	14,547	0	148
Catar	50,953	0	194
Crohn	4,679	0	0
RA	9,655	0	0

Results: Mapped Cohorts vs Author's Intent

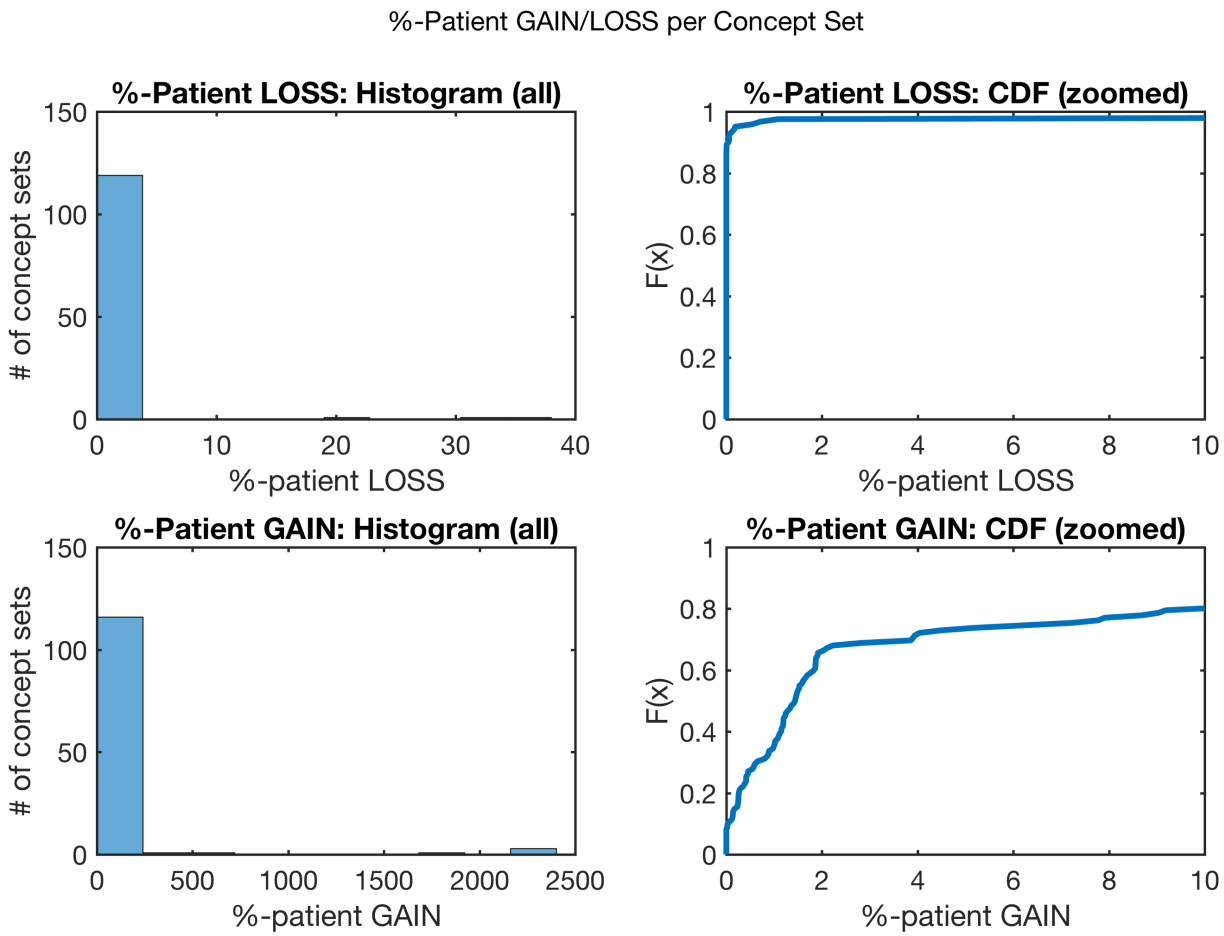
- Author's Intent brings in new patients, compared to only original ICD9 codes.
- OPTIMIZE does a good job of finding patients that reflect the author's intent
  - Generally <1:1000 (worst is RA at 0.013)

Pheno	Cases#	ICD9 set		SNOMED mimic		SNOMED optimize		SNOMED no desc		SNOMED all desc	
		Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss	Gain	Loss
HF	75,626	0	314	0	0	0	0	0	0	1332	0
HF2	76,958	0	1646	0	1332	0	0	0	1332	0	0
T1DM	27,935	0	74	0	23	0	23	108	67	943	0
T2DM	126,828	0	1486	3	1412	3	30	34	1486	1317	0
Appy	9,920	0	33	0	0	0	0	0	8	0	0
ADHD	14,547	0	148	0	39	0	19	1359	19	1359	0
Catar	50,953	0	194	39	26	39	2	39	26	2451	0
Crohn	4,679	0	0	0	0	0	0	0	0	0	0
RA	9,655	0	0	113	0	113	16	113	0	25289	0

Results: Patient Gain/Loss using only automatically mapped queries

Results: Patient Gain/Loss using only automatically mapped queries

- Patient loss was almost always exactly 0.
  - Exceptions occurred when invalid ICD9 codes were specified
- Patient gain was usually <4% (70% of concept sets).
- Patient gain was sometimes very large (e.g. > 100%)
- Concpet sets with largest %-gain had obvious flaws:
  - Missing icd9 codes that probably should be included
  - Typos



Discussion 1

- 5/9 concept sets had code mapping issues
- But effect on patient cohort is minimal
  - 8/9 concept sets produced error < 1/700, 9th was 1.3%
  - Small compared to coding error 2% to 50%
- Mapping process can improve queries
  - 2/9 mapping revealed codes that were clearly intended but missed from the original list

Discussion 2

- Automated mappings did not perform well consistently (versus
- OHDSI retains the source data so can always go back to th needed