Harmonizing perioperative open datasets using the OMOP-CDM via LLM and LangChain

Hyeonsik Kim[1], Hyeonhoon Lee[2,3], DongHyuk Lee[3], Hyung-Chul Lee[1,2,3]

[1] *Department of Interdisciplinary Program in Medical Informatics, Seoul National University*

[2] *Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul National University Hospital*

[3] *Department of Data Science Research, Innovative Medical Technology Research Institute, Seoul National University Hospital*

NSTRI
National Strategic Technology Research Institute

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047, RS-2024-00439677)

Trends in Medical Data Science

Open Dataset

- Publicly Available Datasets
- Open Source, Collaborations
- ICU Datasets
  - MIMIC (MIT, US)
  - K-MIMIC (SNUH, Korea)
  - eICU-CRD (MIT, US)
  - AmsterdamUMCdb (Ams Univ, Netherland)
  - HiRID (Bern Univ, Switzerland)
  - SICdb (Salzburg Univ, Austria)
- **Perioperative Datasets**
  - VitalDB, INSPIRE (SNUH, Korea)
  - MOVER (UC Irvine, US)

Federated Network

- Distributed, Restricted Data
- Moving Models and Queries
- Common Data Model: **OMOP-CDM**
- Research Networks
  - **OHDSI Network**
  - Research Border Free Zone
- Platforms
  - Mayo Clinic Platform (US)
  - FeederNet (47 hospitals, 57M pts, Korea)

H.C. Lee. (2023.10) Perioperative Data Science, Global trends and Future Directions

Perioperative Open Datasets

VitalDB

- Seoul National University Hospital, Korea, 2016-2017
- 6,388 surgeries, 6,090 patients
- High resolution waveform data, clinical information

INSPIRE

- Seoul National University Hospital, Korea, 2011-2020
- 131k surgeries, 100k patients (random 50% of all cases)
- EMR data of perioperative periods

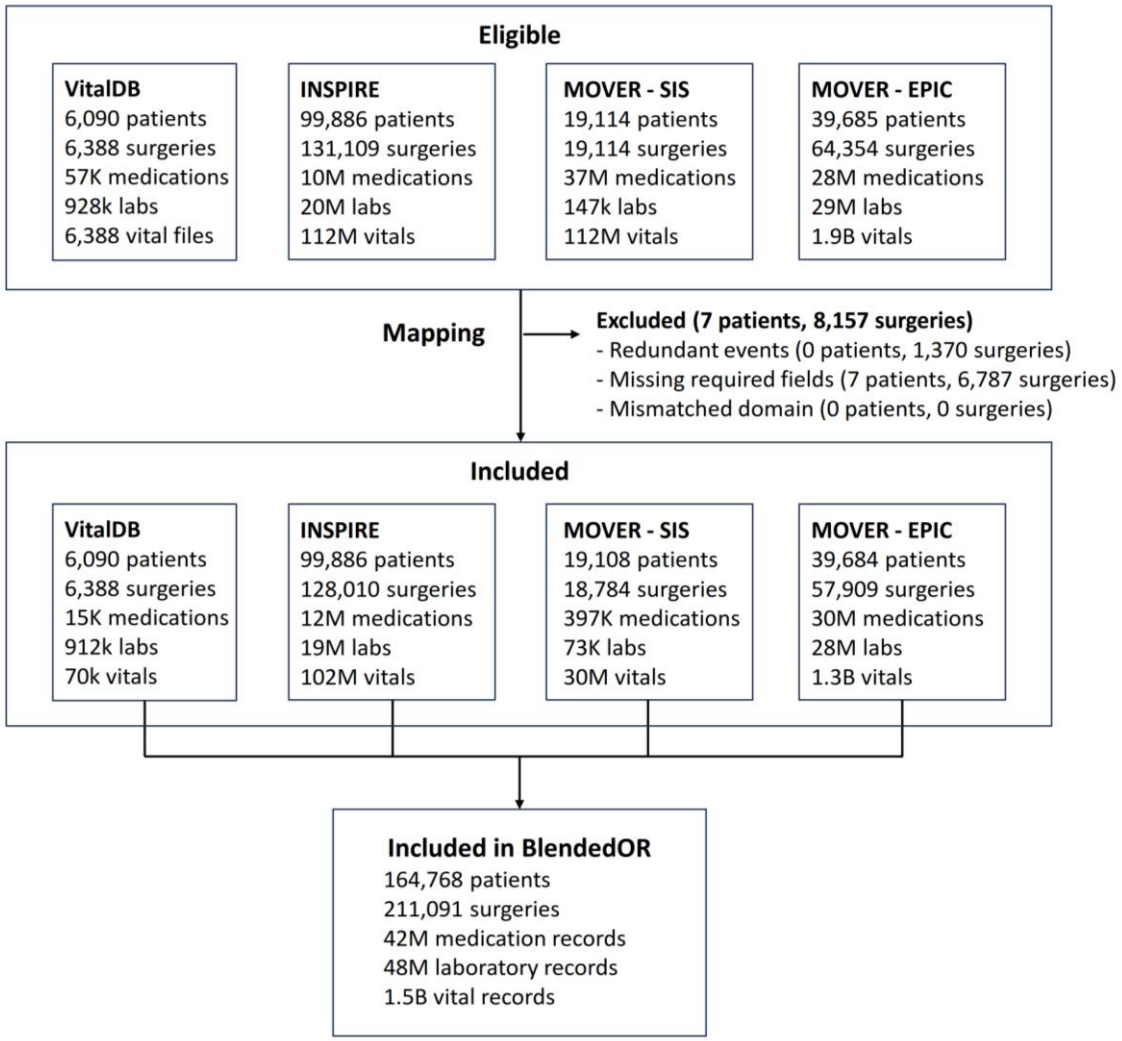
MOVER

- University of California, Irvine Medical Center, US, 2015-2022
- 83k surgeries, 59k patients
  - **MOVER-SIS** : 19k surgeries, 19k patients
  - **MOVER-EPIC** : 40k surgeries, 64k patients
- EMR data, high-fidelity waveform data

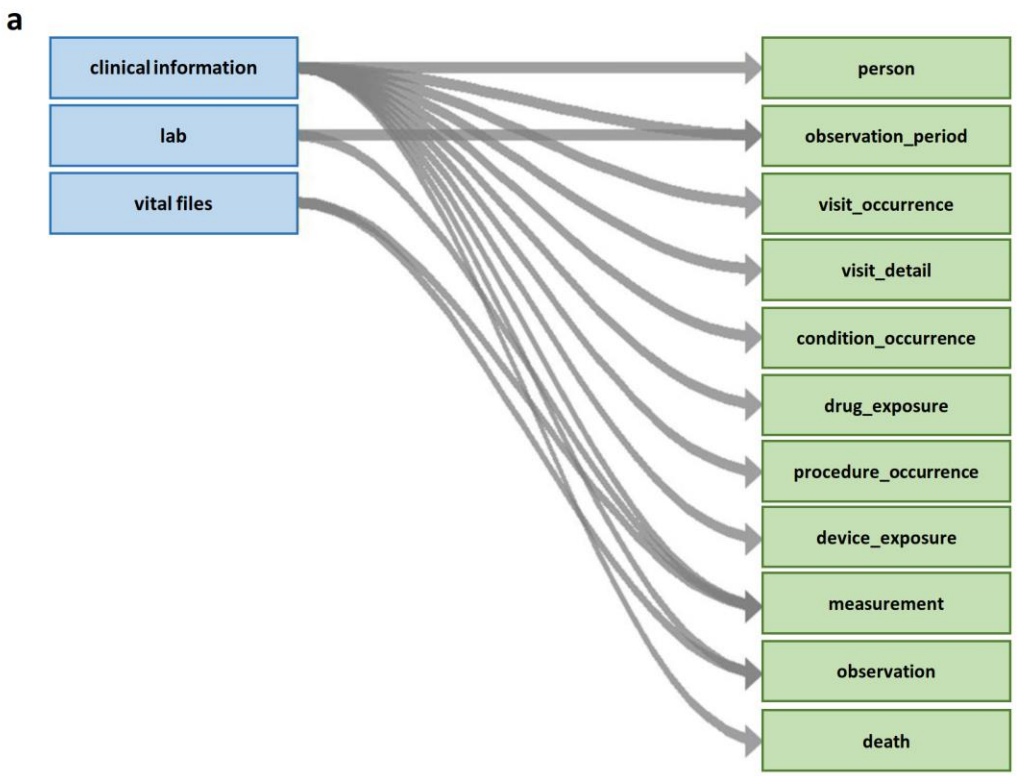
OMOP Common Data Model

INSPIRE, Informative Surgical Patient dataset for Innovative Research Environment; MOVER, Medical Informatics Operating Room Vitals and Events Repository

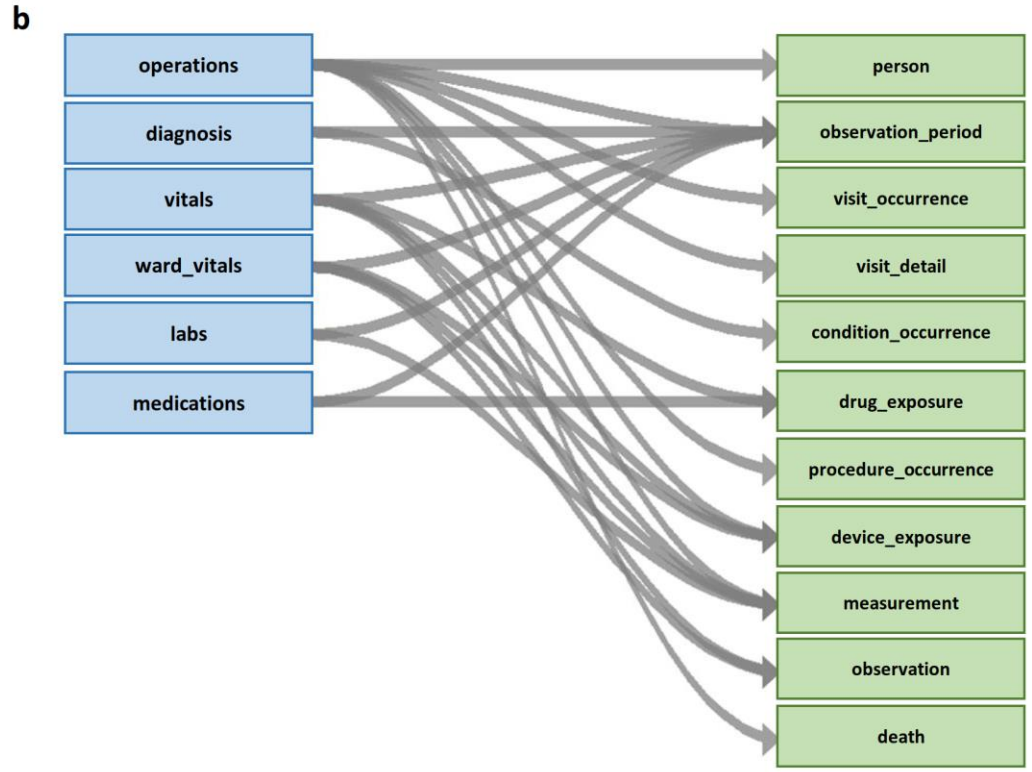
Flowchart


Eligible
VitalDB
6,090 patients
6,388 surgeries
57K medications
928k labs
6,388 vital files
INSPIRE
99,886 patients
131,109 surgeries
10M medications
20M labs
112M vitals
MOVER - SIS
19,114 patients
19,114 surgeries
37M medications
147k labs
112M vitals
MOVER - EPIC
39,685 patients
64,354 surgeries
28M medications
29M labs
1.9B vitals
Mapping
Excluded (7 patients, 8,157 surgeries)
- Redundant events (0 patients, 1,370 surgeries)
- Missing required fields (7 patients, 6,787 surgeries)
- Mismatched domain (0 patients, 0 surgeries)
Included
VitalDB
6,090 patients
6,388 surgeries
15K medications
912k labs
70k vitals
INSPIRE
99,886 patients
128,010 surgeries
12M medications
19M labs
102M vitals
MOVER - SIS
19,108 patients
18,784 surgeries
397K medications
73K labs
30M vitals
MOVER - EPIC
39,684 patients
57,909 surgeries
30M medications
28M labs
1.3B vitals
Included in BlendedOR
164,768 patients
211,091 surgeries
42M medication records
48M laboratory records
1.5B vital records

ETL : Field Mapping

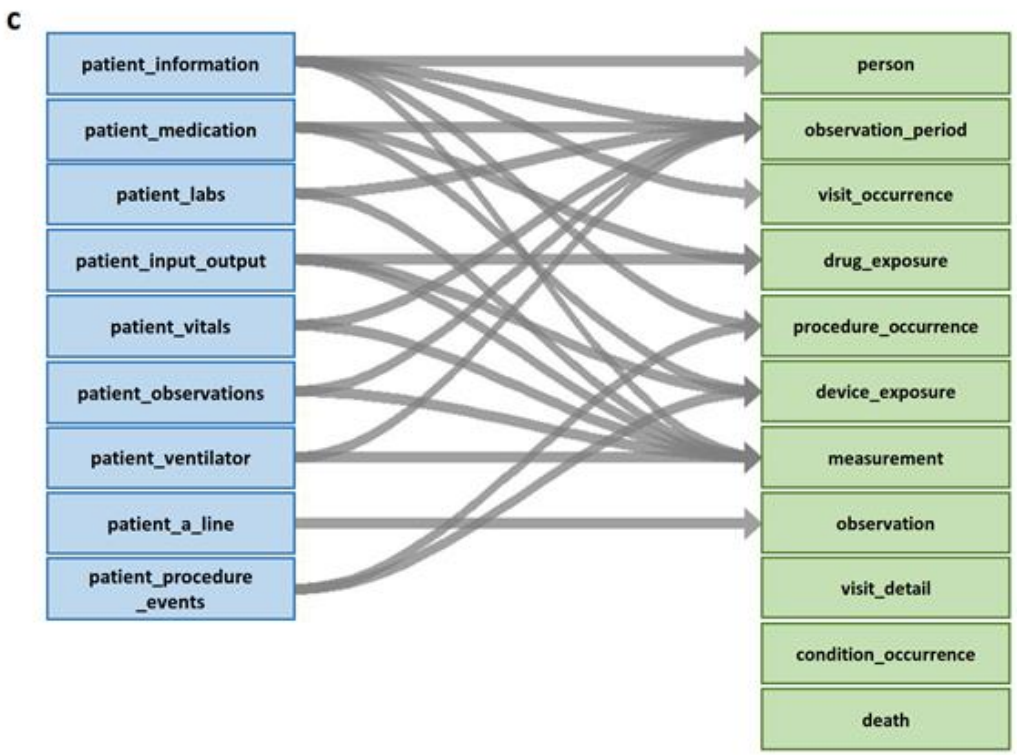


VitalDB

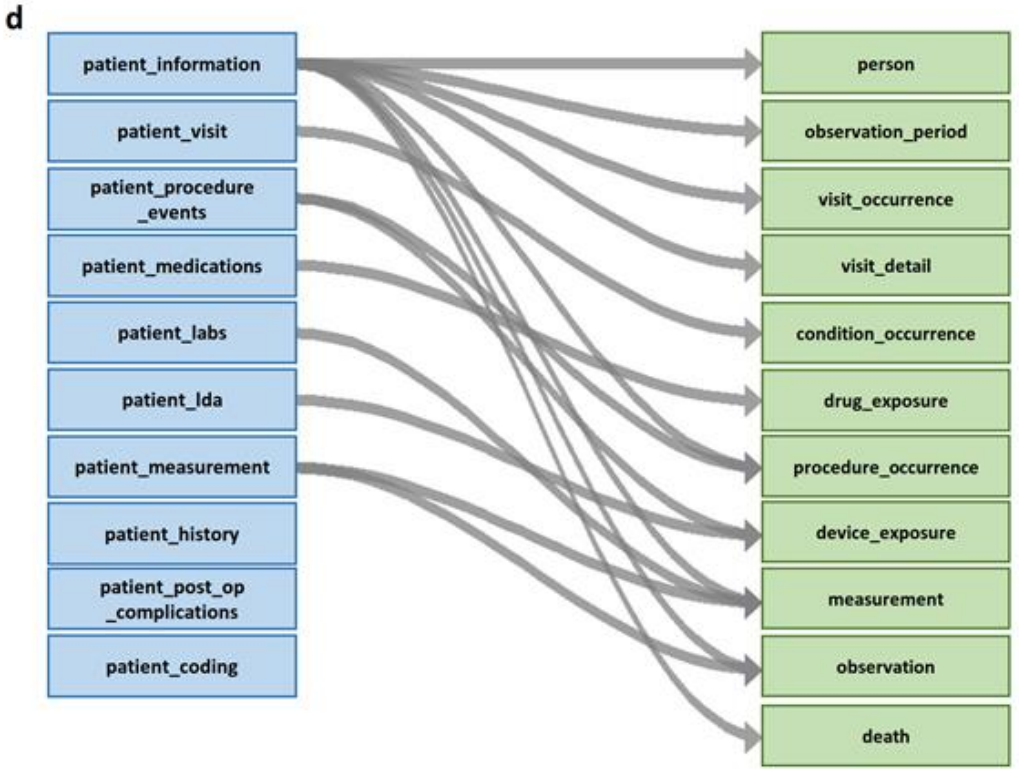


INSPIRE

ETL : Field Mapping



MOVER-SIS



MOVER-EPIC

ETL : Concept Mapping

Concept Domain	VitalDB	INSPIRE	MOVER-SIS	MOVER-EPIC	Standard Vocabulary
Procedure	Free text	ICD-10-PCS	Free text	Free text	ICD-10-PCS
Number of texts	359	0	16,281	1,768	
Condition	Free text *	ICD-10-CM	NA	ICD-9-CM, free text	SNOMED (+ICD-10-CM)
Number of texts	842	0	0	10,073	
Drug	Free text *	Free text, ATC *	Free text *	CPT, free text *	RxNorm, RxNorm Extension
Number of texts	11	59	211	4344	
Route	NA	Free text	NA	Free text	SNOMED
Units	NA	NA	Free text	Free text	UCUM
Device	Free text	Free text *	Free text *	Free text *	SNOMED
Number of texts		11	14	97	
Measurement, Observations	Free text *	Free text *	Free text *	LOINC, free text	LOINC, SNOMED
Number of texts	242	126	56	121	
Units	Free text	Free text	NA	Free text	UCUM

Results

Table	VitalDB	INSPIRE	MOVER-SIS	MOVER-EPIC	Total
Person	6,090	99,886	19,108	39,684	164,768
Observation Period	6,090	99,886	19,108	39,682	164,766
Visit Occurrence	6,090	126,673	19,108	52,385	204,256
Visit Detail	4,990	14,188	NA	18,435	37,613
Condition Occurrence	6,134	2,209,021	NA	100,069	2,315,224
Drug Exposure	14,549	12,040,789	396,595	29,568,720	42,020,653
Procedure Occurrence	6,388	128,010	48,990	117,926	301,314
Device Exposure	481	376,978	11,045	329,575	718,079
Measurement	67,245,344	122,487,072	30,059,576	1,235,132,946	1,429,824,103
Observation	3,367,303	777,850	2,947	104,433,032	108,581,132
Deaths	48	933	NA	420	1,401

Validation : Data Quality Dashboard

a

VITALDB

DataQualityDashboard Version: 2.6.1
Results generated at 2024-10-10 14:16:02 in 4 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	400	5	405	99%	290	1	291	100%	690	6	696	99%
Conformance	509	1	510	100%	64	0	64	100%	573	1	574	100%
Completeness	269	4	273	99%	10	0	10	100%	279	4	283	99%
Total	1178	10	1188	99%	364	1	365	100%	1542	11	1553	**99%**

577 out of 1542 passed checks are Not Applicable, due to empty tables or fields.
4 out of 11 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 99% (965/972).

b

INSPIRE

DataQualityDashboard Version: 2.6.1
Results generated at 2024-10-10 13:35:54 in 7 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	403	2	405	100%	291	0	291	100%	694	2	696	100%
Conformance	508	2	510	100%	64	0	64	100%	572	2	574	100%
Completeness	268	5	273	98%	10	0	10	100%	278	5	283	98%
Total	1179	9	1188	99%	365	0	365	100%	1544	9	1553	**99%**

540 out of 1544 passed checks are Not Applicable, due to empty tables or fields.
4 out of 9 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 100% (1004/1009).

c

MOVER_SIS

DataQualityDashboard Version: 2.6.1
Results generated at 2024-10-10 12:57:24 in 3 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	399	6	405	99%	289	2	291	99%	688	8	696	99%
Conformance	508	2	510	100%	64	0	64	100%	572	2	574	100%
Completeness	269	4	273	99%	10	0	10	100%	279	4	283	99%
Total	1176	12	1188	99%	363	2	365	99%	1539	14	1553	**99%**

784 out of 1539 passed checks are Not Applicable, due to empty tables or fields.
4 out of 14 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 99% (755/765).

d

MOVER_EPIC

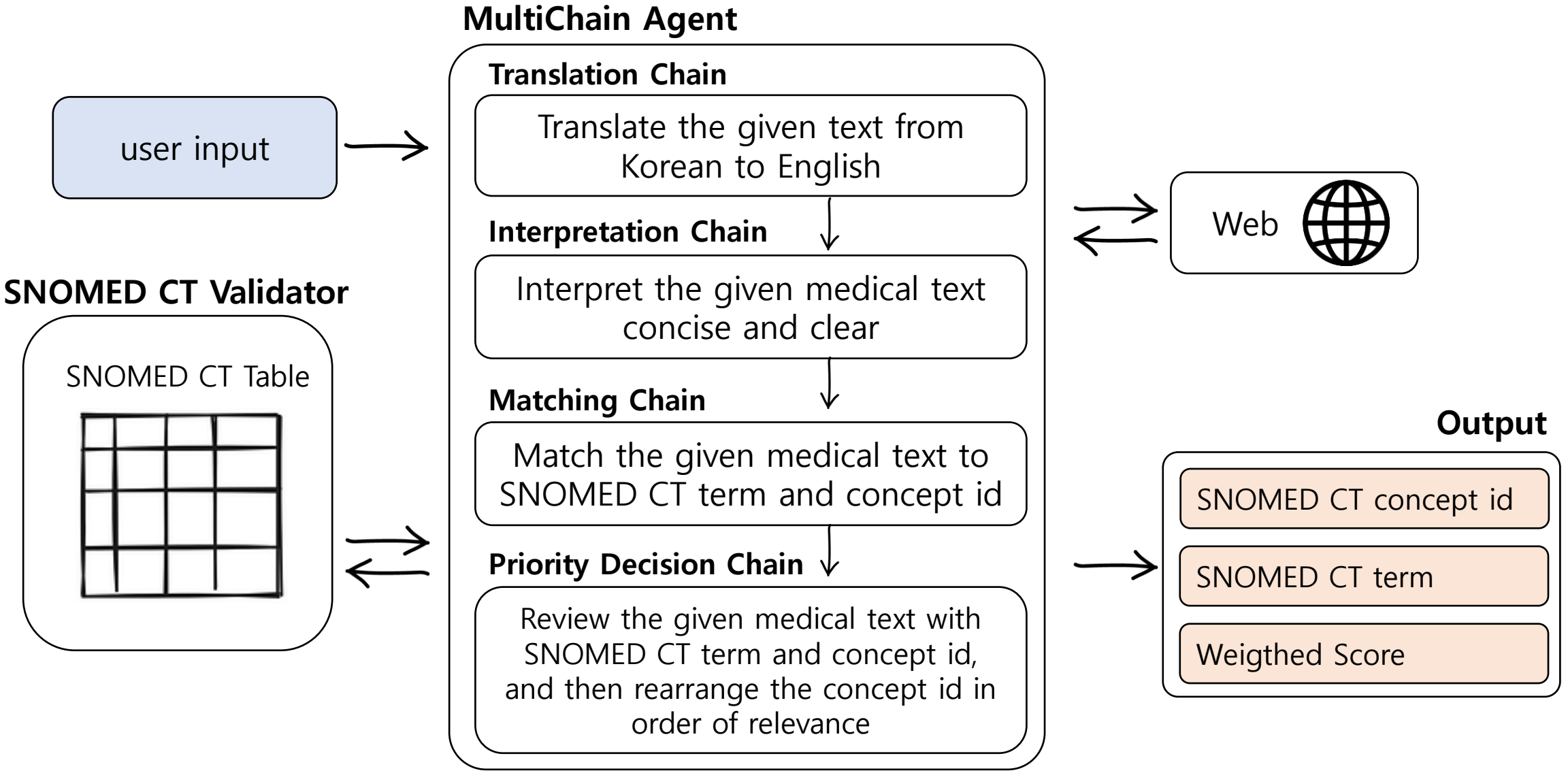
DataQualityDashboard Version: 2.6.1
Results generated at 2024-10-10 09:54:22 in 1 hours

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	391	14	405	97%	289	2	291	99%	680	16	696	98%
Conformance	505	5	510	99%	64	0	64	100%	569	5	574	99%
Completeness	269	4	273	99%	10	0	10	100%	279	4	283	99%
Total	1165	23	1188	98%	363	2	365	99%	1528	25	1553	**98%**

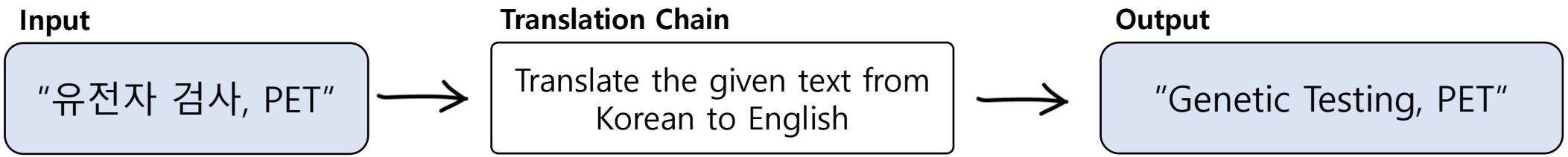
483 out of 1528 passed checks are Not Applicable, due to empty tables or fields.
4 out of 25 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 98% (1045/1066).

SNOMED-CT Mapping using LLM

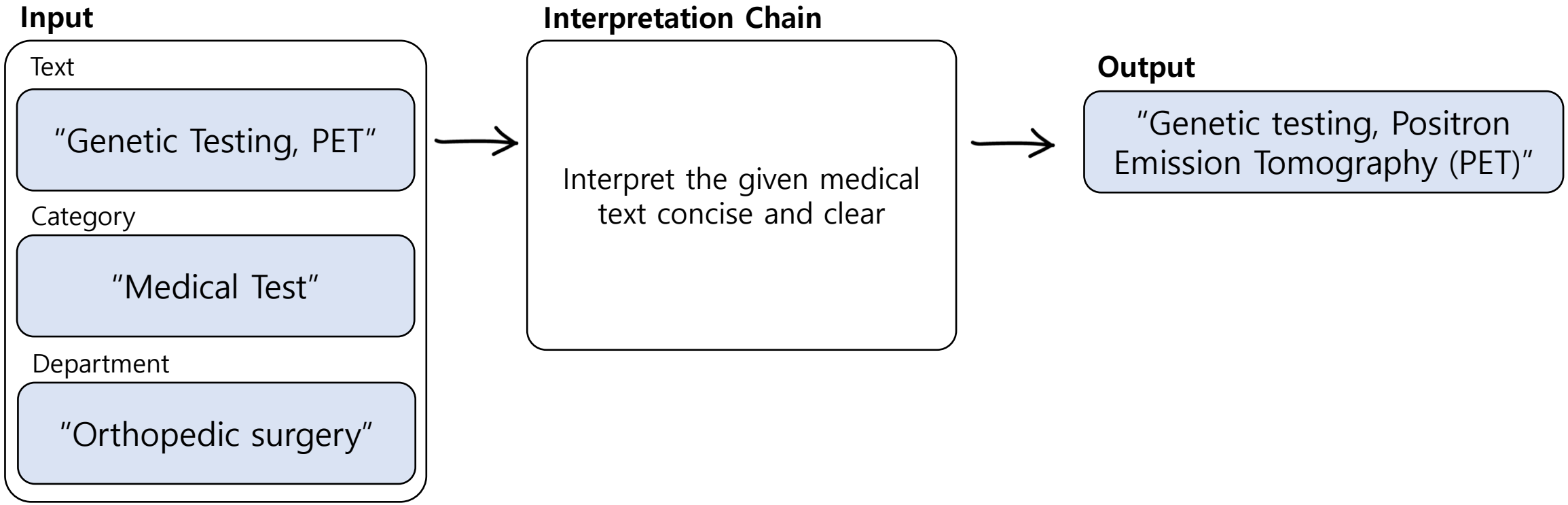
Architecture



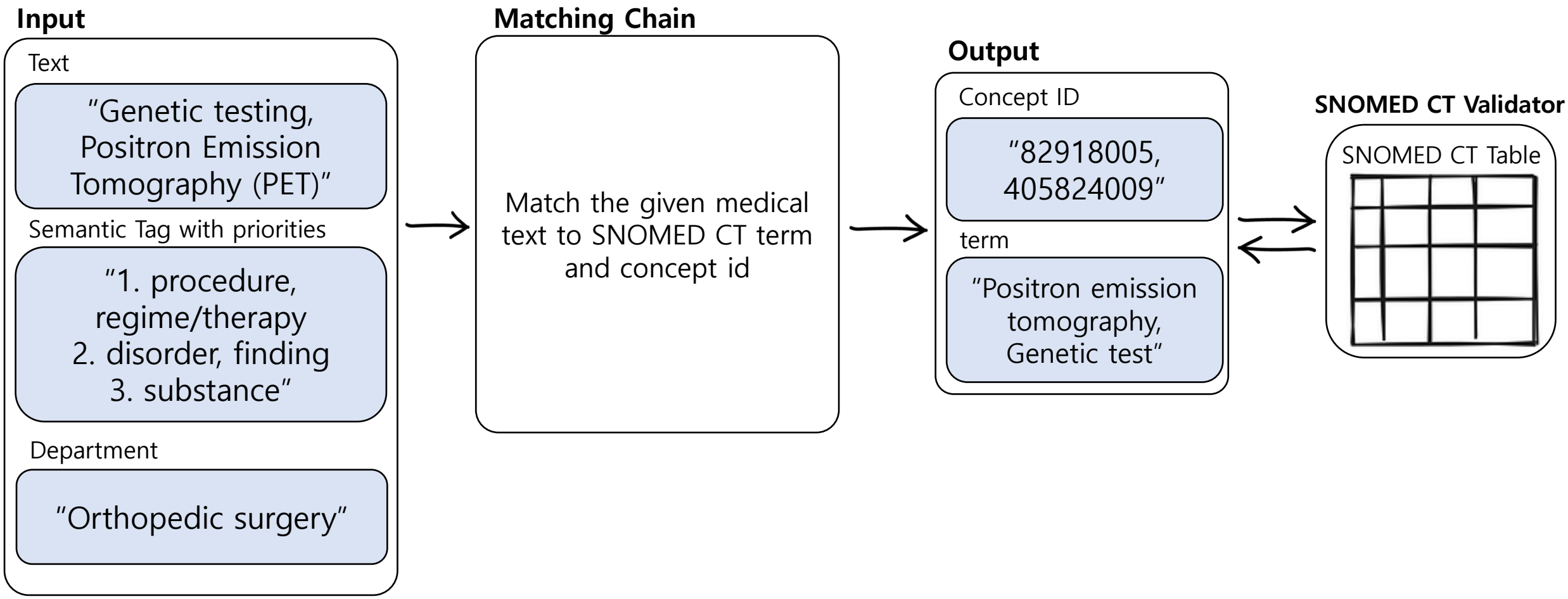
Translation Chain



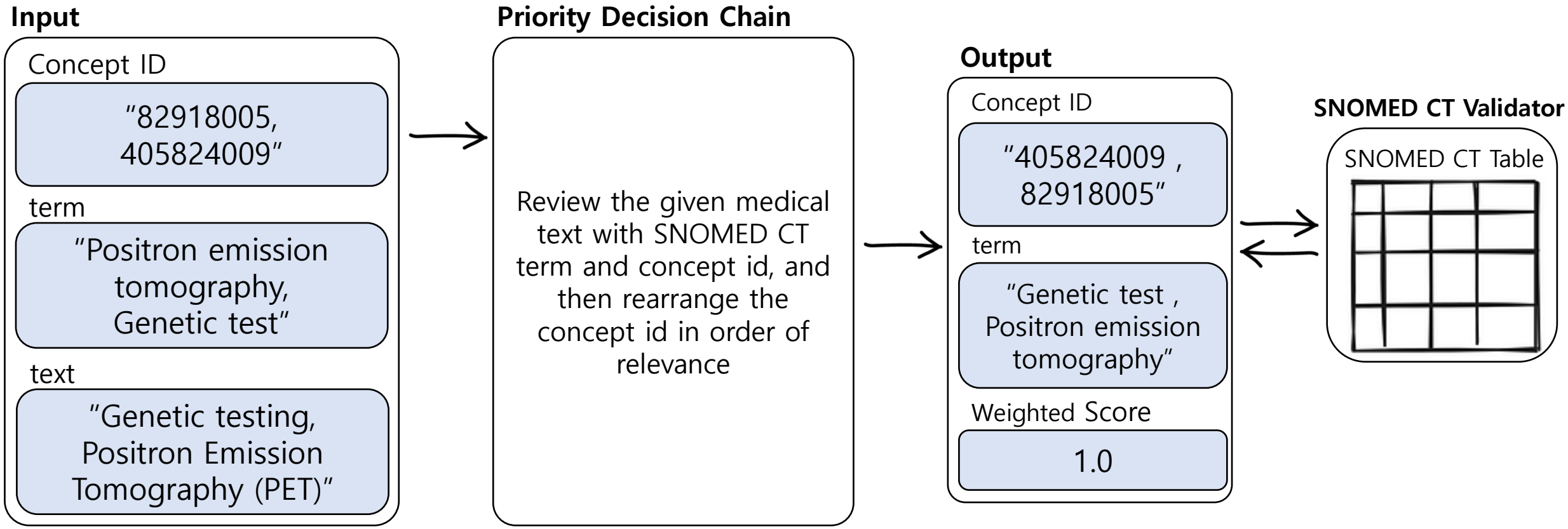
Interpretation Chain



Matching Chain



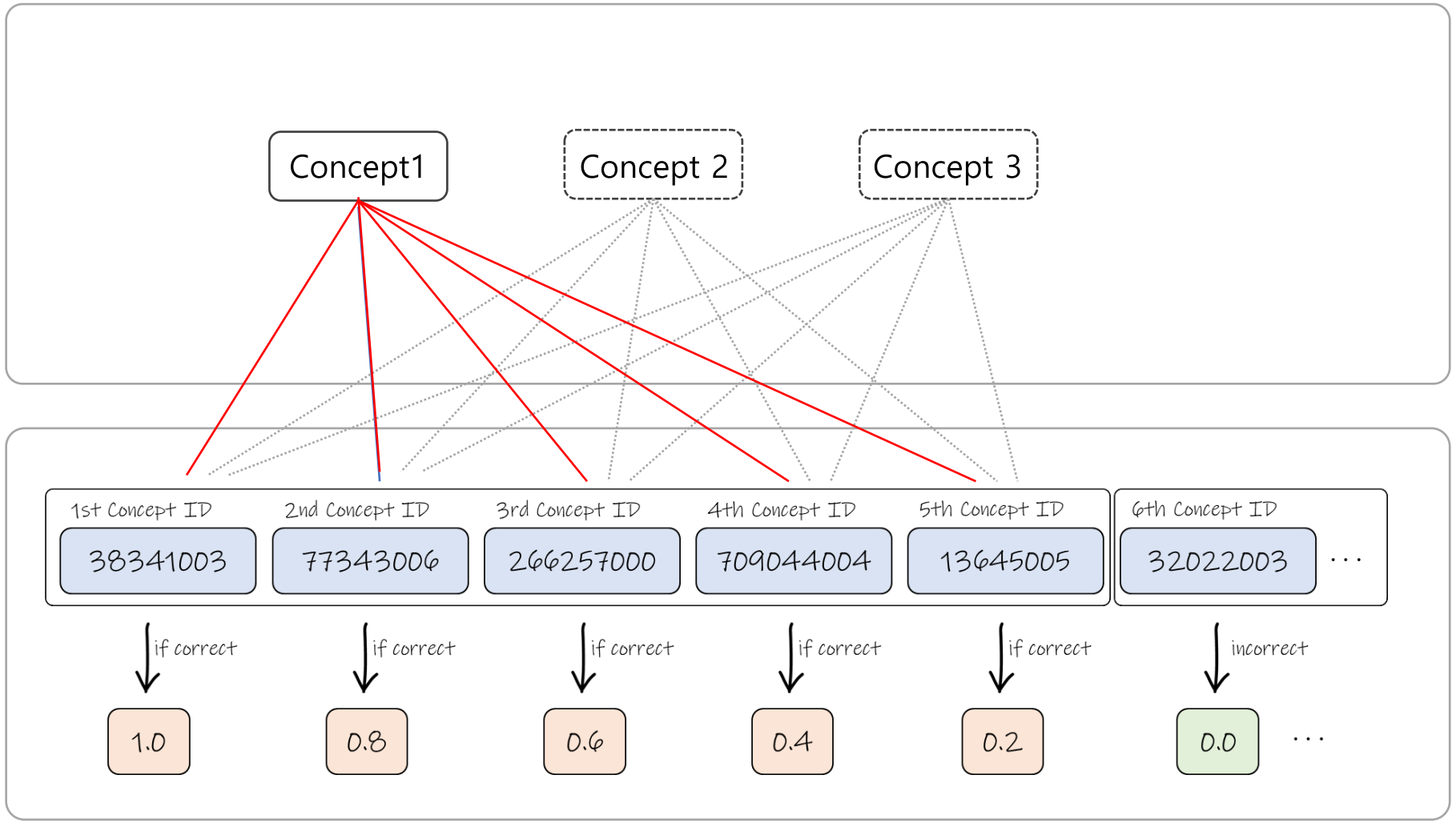
Priority Decision Chain



Evaluation

Manual mapping

LLM Automated mapping



Conclusion

- Anticipate seamless integration of BlendedOR with various OHDSI tools

- Harnessing LLMs for concept mapping in medical text is both promising and essential

Thank you for your attention!