Transforming a Large-Scale Prostate Cancer Outcomes Dataset to the OMOP Common Data Model—Experiences with a Regional Health Record Database in China

Qiliang Cai[1], Yang Xie[2], Jing Li[2], Xiaoyu Lin[2], Chih-Chi Yang[2], Zheng Yin[2], Yi-Chun Yeh[2], Shiyun Huang[3], Zhe Xu[3], Yuanjie Niu[4]

[1]Department of Urology, Second Hospital of Tianjin Medical University, Tianjin, China.
[2]Real World Solutions, IQVIA.
[3]Medical Affairs, Pharmaceuticals, Bayer Healthcare Company Ltd, Beijing, China.
[4]Department of Urology, Tianjin Medical University General Hospital, Tianjin, China.

41st Annual Meeting
Walter E. Washington Convention Center
August 22-26, 2025
ispe International Society for Pharmacoepidemiology
Washington, DC, USA

Background

- Regional electronic health record (rEHR) databases in China provide comprehensive, longitudinal inpatient and outpatient data across multiple sites in a region, supporting high-quality real-world evidence (RWE) research.
- However, achieving semantic interoperability remains a challenge due to varied vocabularies used across sites.
- To address this, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) may be a potential solution.
- Taking the Tianjin rEHR database* as an example, this study aims to assess and prepare the Tianjin rEHR for OMOP conversion to support global prostate cancer (PC) research.

*Note: Tianjin City is a municipality in northern China. Tianjin rEHR database is one of the most established rEHR systems in China, integrates data from 82 sites and covers ~16 million residents.

Objective

- To evaluate the Tianjin rEHR database for its structure, features, and suitability in supporting RWE studies focused on PC.
- To standardize medical terminology—including diagnoses, drugs, procedures, and lab tests—using global vocabularies aligned with the OMOP-CDM.

Methods

Data source

- As of 2024, the Tianjin rEHR database includes over 110 billion records from public medical institutions, with 29 billion records processed through data governance, covering over 16 million residents since 1995.
- It integrates data from 43 tertiary hospitals and 39 secondary hospitals, along with public health records—such as immunizations and maternal care—covering ~12 million people since 2000.
- For this study, data from 2017 to 2021 were extracted and transformed into the OMOP-CDM schema, with research-grade data quality available since 2015.

Sample selection

- Included patients had at least one diagnosis of PC recorded during outpatient visits or hospitalizations in the Tianjin rEHR database.
- PC diagnoses were identified using ICD-10 code C61 or diagnosis names containing “前列腺” along with terms like “癌”, “原位癌”, “CA”, “恶性肿瘤”, “PCA”, “CRPC”, or “HSPC”.
- Records were excluded if diagnosis names included terms such as “OPCA”, “PCAD”, “前列腺增生”, or “肾癌” to ensure accurate case identification.

OMOP-CDM conversion

- The OMOP-CDM conversion process includes data profiling, Extract, Transform, Load (ETL) mapping, vocabulary standardization, and quality assurance. (**Figure 1**)

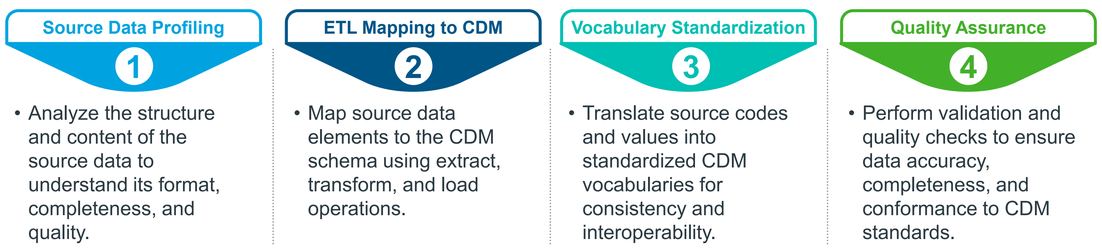

Figure 1. OMOP-CDM Conversion Workflow

- The Tianjin rEHR database contains all clinical information required for the OMOP-CDM structure. The mappings between tables in the Tianjin rEHR database and those in the OMOP-CDM are summarized in **Figure 2**.

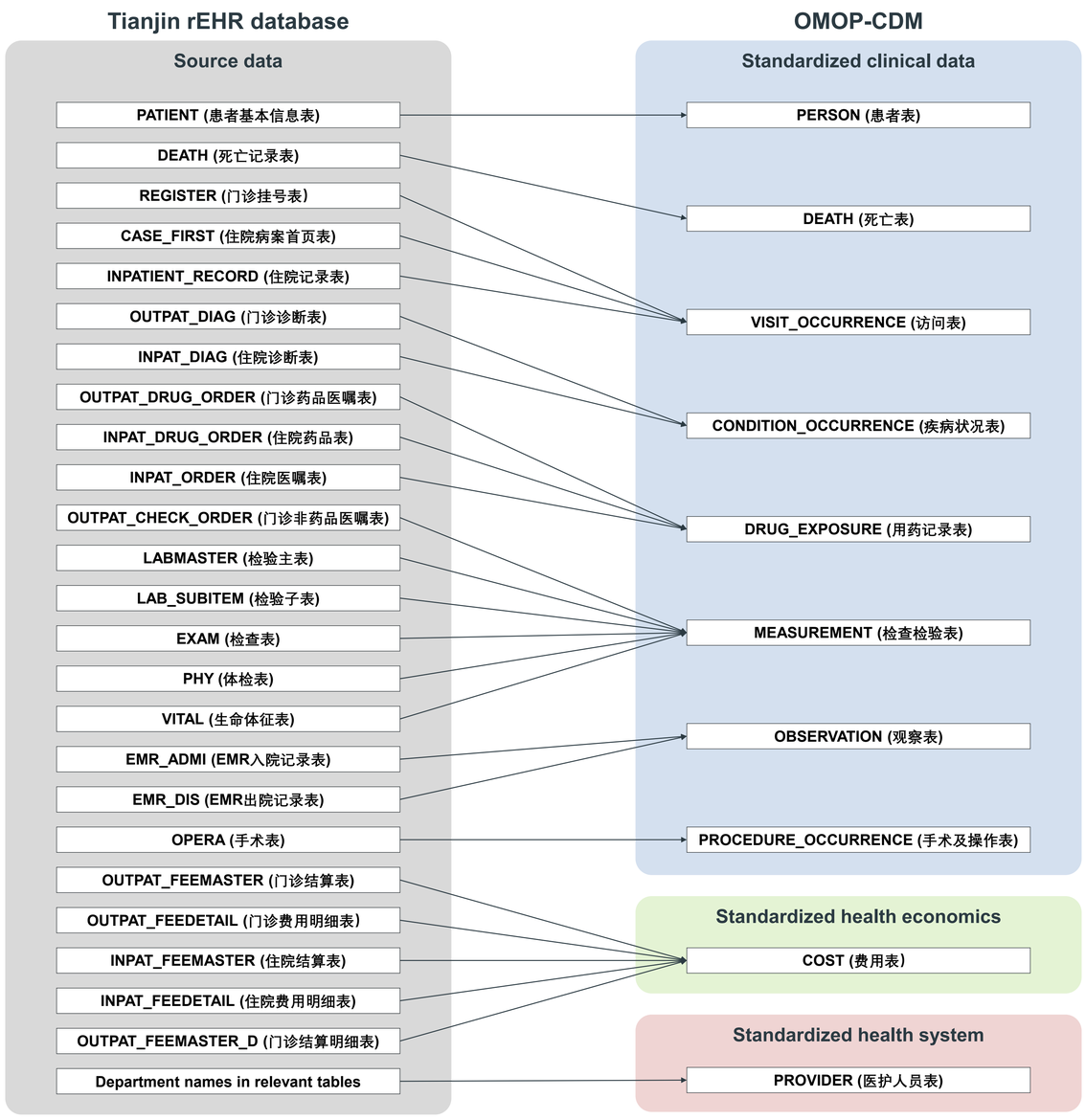

Figure 2. Transformation of the Tianjin rEHR Database into OMOP-CDM Tables

Results

- Data profiling of the Tianjin rEHR database revealed that the dataset includes approximately 310,000 clinical visits, encompassing both outpatient and inpatient records. Around 70% of these visits occurred between 2019 and 2021.
- A total of 18,745 male PC patients met the selection criteria, with 87.57% aged 65 years or older at the time of diagnosis. Among them, over 50% had comorbid urinary system diseases, followed by orthopedic system diseases (44.13%) and hypertension (39.75%). The most frequently used concomitant medication was pantoprazole (11.03%), followed by hydrocortisone (10.16%). (**Table 1**)

Table 1. Characteristics of Prostate Cancer Patients

Characteristics	%	Characteristics	%
Sex		**Concomitant medication (Cont.)**	
Male	100.00%	Clonazepam	0.88%
Age group at diagnosis		Alprazolam	0.88%
< 65 years old	12.43%	Morphine	0.70%
≥ 65 years old	87.57%	Vitamins	0.70%
Medical insurance type		N-acetylcysteine	0.70%
UEBMI	10.51%	Zopiclone	0.70%
URBMI	2.63%	Naproxen sodium	0.70%
NRCMS	0.00%	Statins (Calcium tablets)	0.70%
Other BMI	24.17%	Imatinib	0.53%
Off-site BMI	0.00%	Recombinant human granulocyte colony-stimulating factor injection	0.53%
Self-paid	4.20%		
Unspecified insurance type	58.49%	Tramadol hydrochloride	0.53%
Year of diagnosis		Atorvastatin calcium tablets	0.35%
2019	76.18%	Borneol	0.35%
2020	23.82%	Iodixanol	0.35%
Comorbidity		Ibuprofen	0.35%
Urinary system disease	50.09%	Lactulose oral solution	0.35%
Orthopedic system diseases	44.13%	Methylprednisolone acetate	0.35%
Hypertension	39.75%	Clopidogrel bisulfate tablets	0.35%
Cerebral disease	33.80%	Procaine hydrochloride	0.35%
Respiratory system disease	25.74%	Diphenhydramine	0.18%
Digestive system disease	25.57%	Finasteride	0.18%
Kidney disease	23.99%	Fluvastatin sodium	0.18%
Diabetes	18.04%	Lornoxicam sodium	0.18%
Hyperlipemia	0.70%	Glucose sodium chloride injection	0.18%
Heart disease	0.00%	Cimetidine	0.18%
Concomitant medication		Lactate dehydrogenase	0.18%
Pantoprazole	11.03%	Sildenafil	0.18%
Hydrocortisone	10.16%	Diclofenac sodium hyaluronate	0.18%
Corticosteroids	8.58%	Mushroom polysaccharide	0.18%
Levofloxacin	7.53%	Levetiracetam	0.18%
Diphosphate	6.48%	Sodium phospholine iodide	0.18%
Methylprednisolone tablets	3.50%	Budesonide suspension	0.18%
Furosemide	3.33%	Dextromethorphan	0.18%
Tamsulosin	2.98%	Fexofenadine	0.18%
Cephalosporin	2.80%	Compound Aztine enteric-coated tablets	0.18%
Mosapride citrate	2.45%	Methyldopate hydrochloride	0.18%
Piperacillin tazobactam	1.93%	Lidocaine	0.18%
Insulin	1.93%	Cyclandelate	0.18%
Nifedipine	1.75%	Silybin	0.18%
Isosorbide mononitrate	1.40%	Sucrose iron	0.18%
Calcium hydroxybenzene sulfonate	1.05%		

Abbreviations: UEBMI: Urban Employee Basic Medical Insurance; URBMI: Urban Residents Basic Medical Insurance; NRCMS: New Rural Cooperative Medical Care System; Other BMI: Other Basic Medical Insurance; Off-site BMI: Off-site Basic Medical Insurance.

- All clinical information required for the OMOP-CDM structure was contained in the Tianjin rEHR database, with over 85% of records successfully mapped using the specified logic. (**Table 2**)
- The transformed database passed all 126 quality checks, including validations for patient information, diagnoses, medications, procedures, and laboratory measurements.

Table 2. Domain Distribution and Record-Level Mapping Rate

Domain	Mapping logic	Mapping rate (records level)
Diagnosis (inpatient + outpatient)	• Mapping was based on raw diagnosis names, as diagnosis codes were considered unreliable. • Raw diagnosis names were first mapped to ICD10CN codes by matching Chinese names. Remaining unmatched names were mapped to OMOP concepts where feasible. All mapped names will ultimately be standardized to OMOP condition concepts. • For raw diagnosis entries containing multiple diagnoses, mapping to multiple concepts was performed where applicable.	89%
Drug (inpatient + outpatient)	• Raw drug names were mapped to ATC codes, and any remaining names were mapped to OMOP concepts where feasible. All mapped names will ultimately be standardized to OMOP drug ingredient concepts.	85%
Procedure	• Raw operation names were mapped to ICD9ProcCN codes by matching Chinese names. Remaining names were mapped to OMOP concepts where feasible. All mapped names will ultimately be standardized to OMOP procedure concepts.	89%
Lab tests	• Raw lab names were mapped to LOINC CN codes by matching Chinese names. Remaining names were mapped to OMOP concepts where feasible. All mapped names will ultimately be standardized to OMOP measurement concepts.	87%
Lab test units	• Raw lab test units were mapped directly to OMOP unit concepts.	98%

Abbreviations: ICD10CN: International Classification of Diseases, 10th Revision, China Version; OMOP: Observational Medical Outcomes Partnership; ATC: Anatomical Therapeutic Chemical Classification System; ICD9ProcCN: International Classification of Diseases, 9th Revision, Procedure Codes, China Version; LOINC CN: Logical Observation Identifiers Names and Codes, China Version.

Conclusion

- The Tianjin rEHR database was successfully converted to the CDM, demonstrating that **standardizing regional EHR data in China is feasible with minimal information loss**.
- This success highlights the **potential to scale the OMOP-CDM approach** to other regional or standalone EMR systems across China.
- This standardization effort supports **enhanced data interoperability**, aligns with national health data integration goals, and **reinforces the value of applying this framework in the current study**.

Disclosure

This research was financially supported by Bayer Healthcare Company Ltd., Beijing, China.