REAL- WORLD DATABASES IN CHINA: REGIONAL LANDSCAPE, INTEGRATION CHALLENGES, AND METHODOLOGICAL PRACTICE

Adele Li, MBA , David Wang, MBA, Yixuan Zhou, MSc.

Background

China's real-world data (RWD) ecosystem is highly regionalized and heterogeneous, spanning city- and region-level electronic health records (EHRs), insurance or claims-like databases, and disease or quality registries. These sources differ substantially in geographic representativeness, coding standards, longitudinal continuity, refresh cycles, and clinical depth. Understanding their respective characteristics and integration challenges is essential for generating fit-for-purpose real-world evidence (RWE) in China.

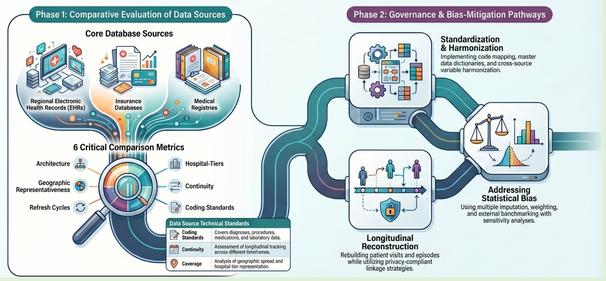
Objective

To characterize the regional landscape of major real-world databases in China, identify key integration and interoperability challenges, and summarize practical methodological approaches for improving the validity, credibility, and decision relevance of RWE studies using these fragmented data sources.

Methods

We reviewed representative regional EHRs, insurance databases, and registries in China, comparing their data architecture, geographic representativeness, refresh cycles, hospital-tier coverage, longitudinal continuity, and coding standards across diagnoses, procedures, medications, and laboratory data. Drawing on multi-project practical experience, we synthesized common governance and bias-mitigation pathways, including code mapping and master data dictionaries, visit and episode reconstruction, variable harmonization across sources, handling missingness and unequal follow-up through multiple imputation and weighting, privacy-compliant linkage strategies where permissible, and external benchmarking with prespecified sensitivity analyses.

Navigating China's Health Data: A Framework for Data Governance and Bias Mitigation



The supporting example of the Shanghai Insurance Big Data Platform illustrates the scale and potential of a city-level integrated source. This database covers the medical information of 25 million Shanghai residents and approximately 50 million patients over the past five years, capturing EHR data from around 3,000 medical institutions, including 80 Tier 3 hospitals and 200 Tier 2 hospitals.

Results

Regional EHRs were found to capture local clinical practice with relatively timely updates, typically on a bi-weekly or monthly basis depending on provider systems. Insurance databases offer systematic views of medication use and healthcare costs, while disease and quality registries provide more focused clinical granularity and quality indicators but may be selective and center-biased. Together, these data sources are complementary, but none alone provides complete national representativeness.

MAJOR REAL-WORLD DATABASES IN CHINA USED IN THIS STUDY | 研究使用的中国主要真实世界数据库

Tianjin rEHR ENLIGHTN · North China	Chongqing rEHR Southwest China	Xiamen rEHR Southeast China	Jinan rEHR North China	CRDS EMR National · 18 provinces
• **~16M** population • **82 hospitals** (43 grade 3, 39 grade 2) • Since **2015** · Every 3–6 months • Inpatient records + public health system Covers one city; tier 2/3 only	• **~39M** population • **323 hospitals** incl. 38 grade 1 • Since **2018** · Monthly updated • Inpatient + community physical exam records Covers one city; partial tier 1	• **~5M** population • **59 hospitals** (15 grade 3, 41 grade 1) • Since **2015** · Bi-weekly updated • Outpatient + inpatient + physical exam Drug amount info requires further application	• **~9M** population • **~80 hospitals** • Since **2016** (EMR until 2020) · Monthly • Inpatient medical records Only 4 years data coverage (2016–2020)	• **~7M** patients • **19 hospitals** · 100% grade 3 • **2016–2023** · Every 2 years • National multi-centre EMR database Sparse hospital distribution; limited data continuity

North China · Southwest China · Southeast China · North China · National (18 provinces)

rEHR = regional electronic health record · EMR = electronic medical record · RWD = real-world data

Base: combined population coverage >70M · Time coverage: 2015–2023 · Hospital range: grade 1–3 across 4 geographic regions

数智研 is one of China's most comprehensive real-world data platforms, covering 13 provinces and cities, ~320 million patient records, nearly 100,000 hospitals, and 10+ years of longitudinal follow-up. It integrates dual data streams — Health Commission clinical records and NHSA claims data — supporting both commercial analytics (market sizing, treatment patterns) and clinical/HEOR research (outcomes modeling, epidemiology) across all four macro-regions of China.

Key strengths include geographic representativeness, long-term patient-level tracking, and continuous data refresh. For pharma researchers, 数智研 delivers the scale, source diversity, and longitudinal depth needed for robust, commercially actionable RWD insights in China.

Legend: Market Insights + Medical Research | Market Insights | ○ Not Yet Covered | Core Data Node

#	Province / City	Tier-3 Hospitals	Data Source	Supported Research
1	Shanghai	~80	Health Commission	Market Insights · Medical Research
2	Fujian Province	~90	NHSA / Health Commission	Market Insights · Medical Research
3	Chongqing	~10	Health Commission	Market Insights · Medical Research
4	Jiangsu Province	~240	Health Commission	Market Insights · Medical Research
5	Nanjing City	~40	Health Commission	Market Insights · Medical Research
6	Hainan Province	~40	Health Commission	Market Insights · Medical Research
7	Beijing	~30	NHSA	Market Insights
8	Tianjin	~40	Health Commission	Market Insights · Medical Research
9	Shandong Province	~200	NHSA / Health Commission	Market Insights · Medical Research
10	Sichuan Province	~110	Health Commission	Market Insights
11	Zunyi City, Guizhou	~30	NHSA	Market Insights
12	Baotou City, Inner Mongolia	~10	Health Commission	Market Insights
13	Nanning City, Guangxi	~10	Health Commission	Market Insights

As of August 2025, continuously updating

Data source abbreviations: NHSA = National Healthcare Security Administration (医保); Health Commission = National/Provincial Health Commission (卫健)

Several recurring limitations were identified across database types: limited national representativeness, uneven hospital coverage, heterogeneous data definitions and units, discontinuous follow-up, data sparsity, and restricted cross-institution linkage. These issues create risks of selection bias, information bias, and confounding. The supporting deck emphasizes that regional databases are usually representative only of their specific geography, and that additional regional databases are needed to provide a more comprehensive national picture. It also highlights practical mitigation strategies such as validated algorithms to reduce misclassification, DAG-informed confounder selection, and adjusted versus unadjusted Cox models to assess robustness.

From a methodological perspective, effective RWE generation in China requires fit-for-purpose study design, prespecified statistical analysis plans, transparent data provenance, and layered quality control reporting. Standardized coding systems such as ICD, ATC, and local terminologies, together with robust cohort construction, episode linking, interrupted-care handling, and sensitivity analyses, are critical to mitigate bias and enhance study credibility.

Potential biases and controls — regional RWD database

Three bias categories and corresponding mitigation strategies for China regional real-world data studies

	RISK	CONTROL STRATEGY
Selection bias — Representativeness	Database reflects only the specific region's patient population — findings may not generalize to national level.	Complement regional data with additional provincial databases to build a more nationally representative composite picture. Multi-regional data integration · Geographic expansion
Information bias — Misclassification	RWD is collected for administrative, not research purposes — leading to potential misclassification of diagnoses, exposures, or outcomes.	Apply previously published and validated diagnostic/phenotyping algorithms to minimize misclassification and resulting bias. Validated algorithms · Phenotyping protocols · Sensitivity analyses
Confounding — Covariate control	Unmeasured or uncontrolled variables may distort observed associations between exposure and outcome.	Fit Cox models with and without confounder adjustment. Use directed acyclic graphs (DAGs) to guide systematic confounder selection. Cox regression · DAG-guided selection · Adjusted + unadjusted models

Conclusions

Regional EHRs, insurance databases, and registries collectively provide a strong foundation for impactful RWE generation in China, but fragmentation and heterogeneity require rigorous integration and explicit bias-mitigation strategies. Deep knowledge of source-specific strengths, limitations, and complementarities, combined with localized analytical expertise and strong data governance, is essential to improve scientific validity and decision relevance as China ' s RWD infrastructure and interoperability continue to evolve.

Reference

1. Real-world databases in China: regional landscape, integration challenges, and methodological practice. ISPOR Poster RWE_04 supporting deck, including Shanghai city-level database example and bias-control framework.

Acknowledgement

The authors thank collaborators and database partners involved in the development, governance, and application of regional real-world databases in China, as well as project teams contributing methodological experience in database integration, quality control, and RWE study execution. Support from regional platform partners, including the Shanghai city-level database example, is gratefully acknowledged