RWD65

Characteristics of Real-World Databases in the United States: A Systematic Review

Nilixa Raval[1], MDS, MPH; Amit Raval[2], M.Pharm, PhD

[1] Nyra Health, Piscataway, NJ; [2] Bayer HealthCare Pharmaceuticals, Whippany, NJ

BACKGROUND

- Real-world databases (RWDs) are recognized as important sources to aid in regulatory, clinical, and reimbursement decision-making for novel medical/health technologies with early evidence of utility in rare diseases or medical device[1]
- Healthcare systems play a crucial role in shaping the RWDs in general.
  - In 1986, the Center for Medicare and Medicaid Services (CMS) introduced a bill to incentivize electronic submission of health insurance claims. It provided harmonized/ coded medical data information for reimbursement. Claims databases have been used to examine treatment outcomes in the US since early 1989[2].
  - In 2009, the US government financially incentivized the "meaningful use" of electronic health/medical records (EHR/EMRs) through the Health Information Technology and Critical Health (HITECH) Act, which led to the adoption and use of EMRs. At present, nearly all hospitals (96%) (compared to 40% in 1979) and more than three fourth (65%) of physicians utilize EMR[3]
  - The CMS implements a National Coverage Determination (NCD) policy to require the safety or effectiveness of medical technology beyond clinical trials or other national bodies to incentivize the development of surveillance systems to track disease burden, treatment, and outcomes at a population level[4]
  - The COVID-19 pandemic highlighted the utility of RWDs in informing real-time decision-making regarding the surveillance or efficacy of medications.
- A favorable health ecosystem and ever-growing demands of generating evidence on the benefits and safety of medical technology at a larger scale would further increase the adoption of RWDs.
- However, most RWDs are by-product/healthcare transaction databases or generated within a larger schema of policy implementation or fragmented healthcare insurance. RWDs hold significant potential, but their utilization requires thoughtful consideration and rigorous evaluation of data sources.

STUDY OBJECTIVES

To summarize characteristics of RWDs related to their suitability to generate RWEs within the United States (US) using a systematic literature review (SLR)

METHODS

- **Study Design:** SLR of Published RWD studies
- **Criteria for Selection of Studies:**
  - An observational study utilizing at least one real-world data source, including claims, EMRs, registry, wearables, and web application, and excluding single centers studies or no specific source data published in the past 5 years
- **Electronic Database Search:**
  - PubMed search (May 2023) using search terms related to RWDs
- **Data Screening :**
  - First-pass titles and abstracts were screened to identify studies with specific RWDs, followed by full-text screening of full-text for inclusion
- **Data Extraction & Evaluation**

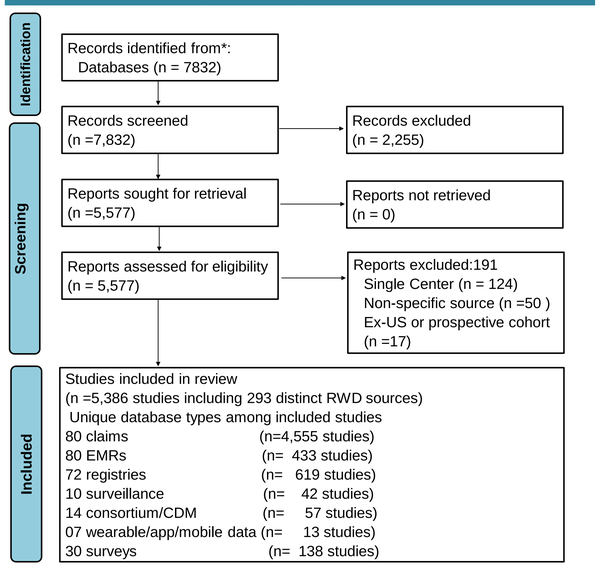
Characteristic	Extracted/Evaluated Information
RWD Types	Claims, EMRs, Registry, Consortium/Common Data Model, Wearables, App-based, Survey
Purpose	Transactional; Purpose Driven
Provenance	Payor/clearinghouses, EMR providers, Integrated Delivery Networks/hospital system, consortium, mobile app/wearables
Representativeness	Total lives covered, representativeness to the US census, social determinants health (SODH)
Coding Semantics	Single or Multiple Vocabulary i.e. ICD-9/10 only or mix of ICD-9-10 and SNOMED for medical condition
Accuracy	Ability to capture simple and complex: population, treatment, laboratory, and genomic/biomarker, imaging constructs Simple Construct: A single criterion or few codes to identify construct. For example, prostate cancer with 1 code of C61 Complex Construct: Multiple criteria, i.e., combination of conditions, laboratory, procedures, or imaging codes to identify construct: relapsing disease with failure to 2+ line of therapy
Longitudinally	Impact of long travel or rural Impact or doctor shopping
Missingness	Missing constructs/phenotypes within data or entire data
Data Access	Ease of access for research
Data Cost	Cost of Data Acquisition for research purpose
Data Lag	Data is available for a lag time of 1+ year, i.e., older data

- **Statistical Analysis**
  - Narrative synthesis to describe the characteristics pertaining to quality of RWDs

RESULTS

- Of 7,831 retrieved citations, 5836 studies with 293 distinct sources met the study inclusion criteria (**Figure 1**); claims and EMRs were most frequently utilized RWDs, followed by registries, surveys, consortium, surveillance, or app/wearables device data (**Table 1**)

Figure 1. PRISMA 2020 flow diagram for new systematic reviews,



Notes: Numbers in the abstract and posters varied due to post-hoc changes in exclusion criteria

Table 2. Characteristics of Real-World Databases

Characteristics	Claims		EMR		Registries		Surveys/Surveillance
Origin	Single Payor	MP/APCD	Structured Only	NLP/H Curated	Static/Single Point	Dynamic	Static
Purpose	Transactional	Transactional	Transactional	Purpose driven	Purpose driven	Purpose driven	Purpose driven
# of Unique Sources	34	18	34; 7+ Oncology	Varies; 3 Genomics	<10; 11 Oncology	10+	44
Representation	15-60%	State-by-State	Varies	Varies	Varies	Limited	National*
Annual Population Size	15+ million	30+ million	5+ million	<1+ million	1000-1+ million	<1 million	<100,000+
Ability to Accurately Capture Treatment and Outcomes (1-least 5-most)							
Simple Treatment	5	5	5	5	3-5	5*	1*
Complex Treatment (Adherence/Persistence)	5	5	3	4	5	3-5*	1*
Complex Clinical	2-3	2-3	2-3	4-5	4-5	4-5	1*
Simple Lab-based	0-2*	0-2*	4-5	4-5	1-3	4-5*	Varies*
Simple Imaging-based	0	0	1-3	3-4	3-4*	0-3*	0
Simple HCRU & Cost	5	5	4	4	1	3-5	Varies*
Mortality	3-5	3-5	2-4	2-4	2-5	3-5	Varies*
Ability to Capture Events Longitudinally							
Impact of Travel, Rural, or doc shopping	No	No	Yes	Yes	Yes	Yes	Yes
Missingness at Data Level or Within the Data							
Imaging, Lab, or Genomics at Data Level	Yes	Yes	Limited	Limited	Limited	Varies	Yes
Imaging, Lab or Genomics within Data	N/A	N/A	Yes	Yes	Yes	Varies	Yes
Logistics (1-least resource intensive; 5 most resource intensive)							
Ease of Access	2	4-5	4-5	4-5	4-5	4-5	1
Cost of Access	4-5	4-5	1-3	4-5	1-3	1-3	1
Issue of Data Lag (delayed by 1+ year)	No	No	No	Yes	Yes	Yes	Yes

* Varies within the data types due to original purpose of database

Table 1. Commonly Utilized Real-World Data Sources by Data Types

#	Claims	N	EMR	N	Registries	N	Surveillance	N	Consortium/Common Data Model	N
1	Medicare	2058	VHA	134	SEER	211	FDA Adverse Event Reporting System (FAERS)	25	Sentinel/Mini-Sentinel Distributed Database (MSDD)	18
2	Marketscan	880	Optum	29	National Trauma Data bank (NTDB)	174	Manufacturer and User Facility Device Experience	5	Observational Health Data Sciences and Informatics (OHDSI)/CDM	12
3	Medicaid	553	Flatiron	28	USRDS/UNOS	43	National Electronic Injury Surveillance System (NEISS)	2	PCORI/PCORNet	9
4	Optum	379	Cerner	22	NCDR-PINNACLE (Practice Innovation and Clinical Excellence)	19	National Respiratory and Enteric Virus Surveillance System	1	Accelerating Data Value Across a National Community Health Center Network	4
5	IQVIA (Pharmetricx Plus)	104	EPIC	19	Cancer Registry/National Program of Cancer	19	Peace Corps Epidemiologic Surveillance System	1	National COVID-19 Clinical Care Consensus (N3C) Database	3
6	AllPayor Claims Database (APCD)	104	Explorys	15	Vascular Quality Initiative Registry 14 vascular procedures	18	Pediatric Health Information System	1	Healthjump	2
7	National Inpatient Sample (NIS)	56	TrinetX	11	Department of Defense Trauma Registry	14	Post licensure Rapid Immunization Safety Monitoring	1	Mental Health Research Network (MHRN)	2
8	Humana	49	Mid-Atlantic	10	American College of Cardiology Transcatheter Valve Therapy	12	Spinal Cord Injury Model Systems Database (SCIMS)	1	Burkitt lymphoma (BL) CDM	1
9	Veteran Administration (VHA)	39	OCHIN EMR	8	Get With The Guidelines-Heart Failure	12	Vaccine Adverse Event Reporting System (VAERS)	2	Academic Thoracic Oncology Medical Investigator's Consortium (ATOMIC)	1
10	Symphony Health	37	Kaiser	8	Get with the Guideline- COVID-19 Cardiovascular Disease	10	Vascular Implant Surveillance and Interventional Outcomes	3	Primary Care Practice Based Research Network (PBRN)	1

- Claims databases had large population size with the ability to accurately capture healthcare resource use and cost, immune to the impact of travel or change in providers, easy to access, and recent data with less data latency; however, with limited ability to accurately capture complex clinical constructs or lack of lab, imaging or genomic data and with varying level of access and cost issues for private and public insurance databases (Table 2)
- EMRs represented majorly structured data with a fragmented population with lab data/SODH enrichment, impacted by travel, residence in rural areas, and provider changes to accurately capture long-term outcomes, with limited ease of access and potentially high costs
- Major registry data were derived for surveillance; therefore, they were only available at incident conditions (static), limiting their use in evaluating long-term outcomes and often having a long lag period. The data access cost was nominal for research purposes; however, the process of data acquisition may be more resource-intensive. Clinical and operational constructs were only accurate for pre-defined registry constructs and often required linkage with other RWDs.
- Claims-linked registries were the most frequently used data resource, with Medicare limited to older publicly insured populations with no information on young individuals or private insurance providers.
- Claims-linked surveys were the second most linked data source with enrichment on social determinants of health and patient-reported outcomes. Still, they represented only a fraction of individuals within the claims database.
- A common database or consortium would overcome the limitation of representativeness but also be prone to the limitations of source data for the longitudinal identification of complex constructs.

CONCLUSION

- To our knowledge, this is the most up-to-date comprehensive review evaluating characteristics of US RWDs related to fit-for-RWE generation in general. This review highlights that the fragmented healthcare system translated into fragmented RWDs with limited ability to capture the continuum of care accurately
- With availabilities of several similar types of data sources and possibly overlapping individuals, there is a need to overcome limitations with respect to representation, ease and cost for access, or care fragmentation:
  - Benchmarking studies to establish the representativeness of the databases against the US census or within similar data sources;
  - Transparent framework of access and costs for research purposes many rich clinical data through registries
  - Linked data not only to overcome limitations of one type of RWDs but also to establish benchmarks and accuracy of individual data to identify clinical constructs in a longitudinal manner
- With precision medicine and targeted therapy approaches, it would add more challenges to have more robust RWDs potentially registries with accurate genomic, biomarkers, or imaging data in the future

Reference: 1. Ramsey et al. Journal of Clinical Oncology. 2024;42: 977-980. 2. Wennberg et al. JAMA. 1987;257(7):933-936. 3. Alexandre et al. PLoS One. 2024 25;19(1):e0295435. 4. Chambers et al. Int J Technol Assess Health Care. 2015 Jan;31(5):347-54.

Poster Presentation at ISPOR 2024, May 5- 8, 2024, Atlanta, GA