PicnicHealth

RWD24

Optimizing EHR Data Completeness: A Conceptual Framework for Bringing Real-World Data into Clinical Research through Relevant Completeness

Priyanka Ramamurthy, Ruby Maa, Colleen Goldberg

Background

Real-world data (RWD), defined as data collected during routine healthcare delivery rather than through predefined research protocols, holds transformative potential in clinical research. It enables insights derived from diverse populations in settings that are more generalizable than traditional clinical trials. However, challenges arise due to the fragmented nature of real world electronic health records (EHRs), which results in incomplete datasets.

Key Challenges with RWD Completeness

- Fragmented EHR Systems: Patients, especially those with complex, multiple comorbidities and therefore high utilization, often receive care across multiple disconnected systems, leading to gaps in data for research purposes.
- Traditional Completeness Metrics: Metrics such as "percentage of missing fields" classify fragmented datasets as inadequate, underestimating the nuanced value of RWD for clinical research.
- Increased Burden: Pursuing absolute completeness can increase the effort required for data collection without proportionate value, undermining one of RWD's key benefits—reducing research burden.

Novel Retrieval Density Framework

Retrieval Density is a novel metric designed to quantify relevant completeness rather than absolute completeness. It addresses the limitations of traditional completeness frameworks by focusing on capturing meaningful clinical encounters relevant to study objectives.

Retrieval density is defined as % of signals that lead to relevant clinical visit with the following detailed definitions:

- Signals: Signals indicate interactions with healthcare providers. They are derived from various sources like physician notes or insurance claims. See Step 1 in Methods for implementation details.
- Relevant Clinical Visits: Defined as the subset of visits identified by prioritized signals that contribute to satisfying research objectives. See Step 2 in Methods for implementation details.

Retrieval Density can help researchers balance between study efficiency and data completeness, a key component of study data quality.

Methods

This case study analyzed detailed medical records for four patients enrolled in a U.S.-based IgA nephropathy (IgAN) registry designed to study disease progression and natural history over time (Table 1).

Step 1: A clinically-trained study member conducted the following multi-step iterative process to understand the IgAN patient journey, identify signals, and retrieve records:

INITIAL CHART REVIEWS

- Disease-, treatment-, and study-specific parameters were established through manual reviews or AI-assisted chart reviews.
- These reviews served as the foundation for identifying signals relevant to care encounters.

PROSPECTIVE PREDICTION

- Synthetic signals were generated to anticipate future care encounters likely to be relevant for the research question.

PATIENT ENGAGEMENT

- Patients were directly engaged to provide provider and visit details, supplementing signal identification efforts.

RECORD RETRIEVAL

- Signals derived from above data sources (e.g., physician notes, claims data) were used to retrieve records for completed visits relevant to the study.

Step 2: Once records were retrieved, a clinically-trained central reader reviewed each record to determine its relevance to the research questions stated at the top of this section.

- **Evaluation Criteria**: The central reader assessed both the content of the record and its metadata (e.g., provider name, provider specialty).
- **Relevance Determination**: Records were classified as relevant if they contained information that contributed to understanding the progression and natural history of IgAN over time.

This systematic approach ensured comprehensive signal identification, retrieval, and classification of care encounters critical to the study's objectives, enabling the calculation of retrieval density.

Table 1: Characterization of patients reviewed

	Patient A	Patient B	Patient C	Patient D
Age at Dx	36	46	29	23
Record retrieval range	2012-2024	2014-2024	2009-2024	2015-2024
Year of Dx	2019	2014	2022	2019
Sex at birth	F	M	F	F
Location	GA	NJ	TX	TX
Total Records	147	96	90	122
Relevant	66	38	35	25
Not Relevant	81	58	55	97

Table 2: Relevant visits type

Visit Type	Relevant (n=164) n (%)	Not Relevant (n=291) n (%)
Adolescent Medicine	0 (0.00%)	1 (0.34%)
Allergist	0 (0.00%)	2 (0.69%)
Cardiology	2 (1.22%)	4 (1.37%)
Dentistry	0 (0.00%)	7 (2.41%)
Dermatology	0 (0.00%)	1 (0.34%)
Emergency Room (ER) and Inpatient	0 (0.00%)	5 (1.72%)
Emergency Room (ER)	8 (4.88%)	13 (4.47%)
Endocrinology	0 (0.00%)	1 (0.34%)
Family Medicine	2 (1.22%)	45 (15.46%)
Gastroenterology	0 (0.00%)	4 (1.37%)
Hand Surgery	0 (0.00%)	3 (1.03%)
Inpatient Visit	3 (1.83%)	4 (1.37%)
Internal Medicine	6 (3.66%)	2 (0.69%)
Laboratory	61 (37.20%)	1 (0.34%)
Labs	1 (0.61%)	0 (0.00%)
Nephrology	54 (32.93%)	5 (1.72%)
Neurology	0 (0.00%)	28 (9.62%)
Neurosurgery	0 (0.00%)	4 (1.37%)
Nutrition	0 (0.00%)	3 (1.03%)
OB/GYN	1 (0.61%)	36 (12.37%)
Ophthalmology	0 (0.00%)	2 (0.69%)
Orthopedics	0 (0.00%)	14 (4.81%)
Multispecialty Outpatient	17 (10.37%)	72 (24.74%)
Pediatric Surgery	0 (0.00%)	1 (0.34%)
Pediatrics	0 (0.00%)	9 (3.09%)
Pharmacology	0 (0.00%)	5 (1.72%)
Physical Therapy	0 (0.00%)	12 (4.12%)
Pulmonology	0 (0.00%)	3 (1.03%)
Radiology	1 (0.61%)	0 (0.00%)
Rheumatology	4 (2.44%)	1 (0.34%)
Urgent/Acute Care	2 (1.22%)	3 (1.03%)
Urology	2 (1.22%)	0 (0.00%)

Table 3: Summary of relevance

	Relevant (n)	% of Relevant	% of Total
Nephrology	54	32.9%	11.9%
Labs	61	37.2%	13.4%
Emergency Room	8	4.9%	1.8%
Multispecialty Outpatient	17	10.4%	3.7%
	140	**85.4%**	**25.3%**

Results

Across these four patients

- 64% (291 out of 455) of records had no impact on the study outcomes ("Not Relevant Records"), representing more than half of the records retrieved (Table 2)
- The visit types that had the highest percentage of relevant records were: nephrology, laboratory, ER, and multispecialty outpatient.
- Nephrology visits represented 32.9% of relevant records and 11.9% of all records. Laboratory records represented 37.2% of relevant records and 13.4% of all records. ER visits represented 4.9% of relevant records and 1.8% of all records. Multispecialty outpatient represented 10.4% of relevant records and 3.7% of total records. (Table 3)
- **25% of records available convey the majority of information relevant to the research question.** (Table 3)

Having determined where the highest concentration of valuable data exists, we can improve prospective record retrieval and data capture efficiency. Under this model, the missingness values that truly matter are nephrology and lab encounters that are known about but not incorporated into the dataset.

Conclusion

Achieving complete EHR data is not a zero-sum game. By prioritizing relevance, this framework enables researchers

1. **Focused Record Retrieval** Targeting key records based on disease-specific and study objective drivers for improved record retrieval and abstraction efficiency
2. **Improved Missing Data Differentiation** Enhancing analytical accuracy by distinguishing between types of completeness for better understanding of RWD value

This approach also acknowledges the complexity of modern healthcare delivery by integrating diverse data sources—from patient engagement to administrative signals—to create a nuanced patient care journey map. It opens pathways for future research to refine data retrieval methods and further address the challenges of incomplete RWD data.

Disclosures

Authors are employees of PicnicHealth