Understanding the Data Quality Issues in Real-World Data Through Real-World Examples

Introduction

In the evolving world of health research, real-world data (RWD) has emerged as a transformative force, offering insights beyond the confines of controlled clinical trials. It captures the intricacies of patient experiences, treatment outcomes and healthcare patterns in a dynamic and evolving environment.

This dynamic treasure trove, encompassing electronic health records (EHRs), health insurance claims (HICs), patient surveys, wearables and patient registries, holds the promise of reshaping our understanding of healthcare outcomes. Together, these elements contribute to the dimensions of the RWD landscape and help us understand treatment effectiveness in real life and the socioeconomic and environmental factors affecting health outcomes.

However, the journey to harness the full potential of RWD has its challenges. This blog post takes a closer look at the experiences shared within the dedicated Working Group in exploring the types of data quality hurdles encountered when using RWD.

Challenges Faced in the RWD Landscape

Using RWD comes with its own set of challenges. Issues such as data fragmentation, scarcity, errors, temporal gaps, dropouts and governance emerge as obstacles that demand careful navigation.

The Working Group discussed two RWD use cases and focused on the data quality issues encountered. These issues were classified into standardised data quality definitions. The objective was to help readers identify types of data quality issues, balance expectations, and pose relevant questions to data vendors during data feasibility discussions.

In Use Case-1, the project focused on using EHR from approximately 70 million patients, derived from linking ambulatory services across large practices/physician networks in the USA. The data was organised in an OMOP common data model and included patient demographics, diagnostic procedures, labs and prescribed medications. Challenges include the lack of socioeconomic data, inconsistencies in lab data compared to clinical trial standards, missing claims data on prescribed medication, an unknown degree of data completeness, and difficulties in auditing source data.

In Use Case-2, the project aimed to use data to identify patients with a disease of interest who lacked disease-specific ICD10 codes. This project used claims data from approximately 120 million patients and lab data from approximately 75 million patients within the USA's healthcare system. The data, stored in a proprietary common data model, encompasses categories linked by patient ID, including family history, vitals and clinical observations, as well as medical and pharmacy claims data. Challenges include interpreting physician notes, addressing temporal gaps, managing provider dropouts and reconciling inaccuracies in reimbursement-related claim codes.

Standardised Data Quality Definitions

- **Data Fragmentation**
  This refers to a quality issue where data is scattered across multiple sources or systems, making it challenging to integrate and analyse effectively. Common data fragmentation issues include anonymisation, variations in data formats, naming conventions, and coding systems. For instance, in Use Case-1, inconsistencies in lab data compared to clinical trial data standards highlight standards misalignment, which impacts on data reliability. Similarly, the absence of claims data on prescribed medication points to differences in coding systems across sources, which hinders seamless data integration. The lack of socioeconomic data illustrates the challenge of linking anonymised healthcare data to disparate sources containing socioeconomic data.

- **Data Sparsity**
  This occurs when certain data points (categories) are underrepresented or missing in a dataset, limiting the ability to draw meaningful conclusions or conduct reliable analyses. The unknown degree of data completeness in Use Case-1 highlights challenges in adequately representing certain medical conditions, treatments, or patient demographics, thus contributing to data sparsity.

- **Data Error**
  This refers to inaccuracies or inconsistencies within the dataset which can arise from coding errors, transcription mistakes or incomplete records. For instance, reimbursement-related claim codes may introduce inaccuracies, which then impacts on data reliability. In Use Case-2, lack of clarity in physician notes and missing unstructured text highlight challenges in attaining a comprehensive patient health journey view.

- **Data Temporal Gaps**
  This occurs when there are interruptions or missing data points in the temporal sequence of events, which then hinders the ability to track changes or trends over time accurately. Irregular data collection practices include sporadic patient visits, missed appointments and/or delays in the recording process. Moreover, seeking specialised care from professionals not under contract with the data vendor introduces additional temporal gaps. Identifying and bridging these gaps should be expected, and steps taken to mitigate their impact. In Use Case-2, temporal gaps within EMR data arose from missing information during specialty visits, which highlights the challenges associated with accurately capturing the temporal aspects of healthcare outcomes.

- **Data Dropout**
  This refers to instances where patients or healthcare providers opt out of data collection or sharing, leading to incomplete or biased datasets. For example, provider dropouts pose obstacles in constructing analytical datasets that reflect the entire patient journey, as seen in Use Case-2.

- **Data Governance and Documentation**
  This encompasses the policies, standards and processes governing data management, including data collection, storage and sharing. Lack of clear data governance policies, metadata and documentation can introduce challenges related to the completeness of the data. Without proper oversight, tracking data quality issues in healthcare data becomes challenging, hindering the assurance that the data remains complete and accurate over time. Use Case-1 illustrates an example where the complexity of auditing source data aligns with the absence of clear data governance policies, metadata and documentation, thereby presenting challenges to achieving data completeness.

Looking Ahead

In conclusion, unlocking the full potential of RWD requires a comprehensive understanding of expected data quality issues. In this blog post, the PHUSE RWE Working Group have endeavoured to provide valuable insights and a starting point for identifying data quality issues. It's essential to recognise that data quality in RWD is crucial from both a sponsor and a data consumer perspective, for ensuring data quality directly impacts on decision-making processes, regulatory approvals, healthcare interventions, policy-making and patient care.

Join us in this transformative journey, where the group aims to develop comprehensive guidelines for sponsors and data consumers, as a framework to effectively navigate the complexities of RWD.

Disclaimer – This blog post provides a general summary of experiences that users have encountered in using real-world data for their specific projects. This blog post does not intend to provide review on specific vendors and tools.