Paper RE06

Scaling ECAs from ad hoc projects to the enterprise: multicenter, multisource infrastructure for external control arms

Jussi Leinonen, Bayer, Helsinki, Finland
Mayur Saxena, Droice Labs, New York, USA

INTRODUCTION

External control arms (ECAs) utilizing real world data (RWD) will become standard in clinical trials for determining drug efficacy and safety. Accordingly, processes to generate ECAs require the same scale and rigor as global multicenter trials, where evidence is generated in a harmonized, coordinated, and systematic way across geographies, sites, and data sources. However, because RWD is messy and challenging to prepare at scale, attempts to create ECAs utilize ad hoc processes that result in data processing delays, low-quality data, and ultimately lack of trust in analytical results. In this paper, we present how Bayer is leveraging Droice Hawk, an AI middleware for patient data preparation, to create enterprise infrastructure for high-quality evidence generation using ECAs, and illustrate how we have overcome the scalability and quality challenges of RWD preparation to scale up ECAs from ad hoc projects to enterprise-level machinery for clinical development.

BUILDING REAL-WORLD PATIENT PROFILES: CHALLENGES

The secondary use of RWD (e.g., EHR, claims) for clinical research requires transforming the data into accurate and fit for purpose patient profiles. Typical approaches to generating research data from RWD transform and aggregate multiple RWD sources into a common data model (CDM). However, transformation of RWD into research data typically results in data loss because information collected for billing and clinical care does not neatly align with clinical research variables such as inclusion/exclusion criteria, covariates, and outcomes. The complexity and diversity of RWD sources, such as EHRs containing tens of thousands of tables, further complicates the data transformation process, leading to human errors and omissions of important source data elements. Therefore, such data loss must be carefully considered when assessing whether RWD in a CDM is fit for purpose for a proposed study, especially when considering RWD for use in high-stakes applications such as evidence for regulatory applications. Accordingly, recent FDA guidance (Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products) cautions that the potential for missing source data elements may render RWD unfit for certain research applications when transformed into a CDM:

- "…combining many data sources, especially with the addition of data transformation into a CDM, adds a layer of complexity that should be considered… data in CDM-driven networks rarely contain all of the source information present at the individual health care sites, and the data elements chosen for a given CDM network may not be sufficient for all research purposes or questions."
- "When converting multiple data sources into a CDM, processes used for data transformation into a CDM (e.g., common terminology and structure), the comprehensiveness of the CDM (e.g., does the CDM contain the key data elements), approaches (e.g., algorithms/methods) for identification and handling of duplicate records within and across data sources, and potential impact of restricting to CDM on sample size and duration of patient follow-up or duration of drug exposure."

CURRENT INFRASTRUCTURE AND PROCESSES FOR PATIENT DATA PREPARATION

The current industry standard to prepare RWD for research is a manual process involving multidisciplinary teams of data engineers, data scientists, epidemiologists, and clinical researchers working together to find relevant patient data elements in the source data, and then map, transform, and normalize the data to build patient profiles in the target CDM schema (Figure 1). Such CDMs often serve as enterprise-level "one size fits all" RWD sources from which diverse study-specific analysis datasets are then derived. However, information loss in the transformation of RWD into CDMs can result in incomplete patient profiles that are missing critical elements required for particular studies/therapeutic areas, and can also introduce significant and unquantifiable biases in the analysis data (e.g. selection bias, misclassifications of exposures, covariates, and outcomes), rendering the CDM data unfit for the intended downstream research use cases.

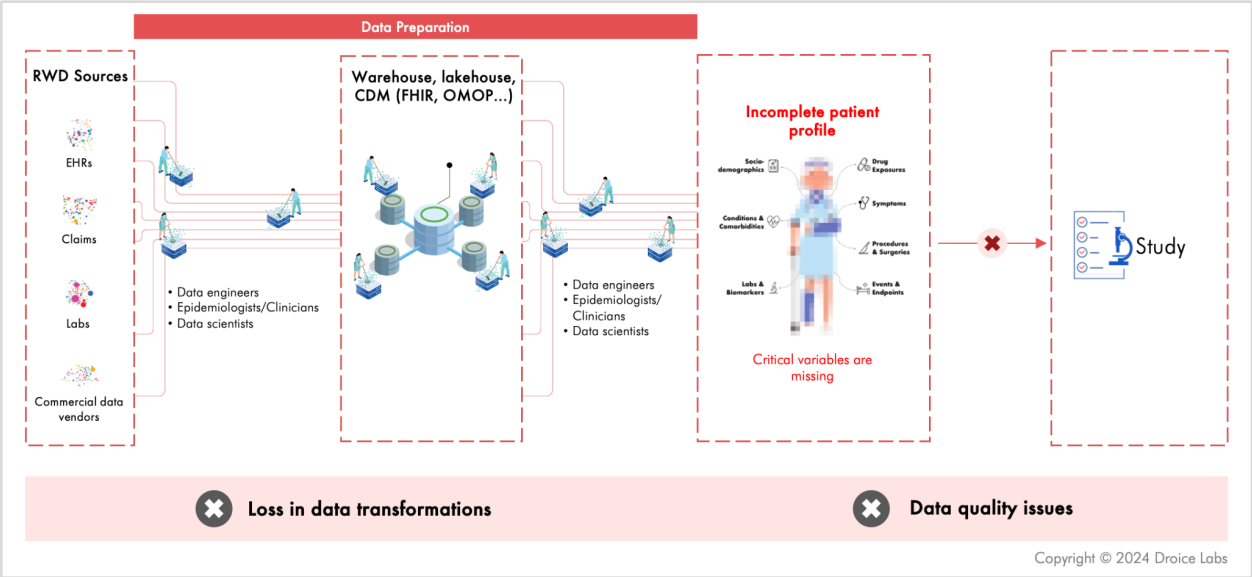

Figure 1: Traditional RWD infrastructure and manual data preparation process

SOURCES OF INFORMATION LOSS AND BIAS

Two major sources of information loss and biases impact transformation of RWD into CDMs: 1) data model limitations, where the entirety of the information in the source data cannot be stored in the analysis data model, both syntactically and semantically [1-5], and 2) data transformation loss, which occurs due to imperfect transformation algorithms [1,7,8] and/or the practical process of implementing said algorithms [2,6].

EXAMPLES:

- *Data model limitations:*
  - Syntactic normalization:
    - Inclusion/exclusion criteria (IE) and adverse events (AE) are not explicitly collected in EHR, so they must be inferred from RWD elements such as diagnosis codes, lab values, etc.
    - Important drug exposures (EC, EX) might be captured in an EHR medication list, in which important fields such as dose and frequency may not be recorded
  - Semantic normalization:
    - Mismatches between RWD terminologies (e.g., ICD-10, SNOMED) and clinical research terminologies (e.g., MedDRA, CDISC terminology) can alter the meaning/interpretation of the source data element
    - Some source RWD terminology codes may not have an appropriate corresponding code in the target research terminology, and thus some source elements are missing normalized codes in the target CDM

Impact: Decisions on how to solve data model limitation loss gaps are based on human judgment in a dataset specific context, and metadata around such decisions is often not captured, and thus such solutions are often not usable or reproducible across different data sources. Additionally, even for the same data transformation task, different individuals or teams may not come to the same conclusion about how to best address data model limitation gaps.

- *Data transformation loss:*
  - Missing data:
    - The manual process of finding relevant data elements in complex RWD sources such as EHRs, which are comprised of thousands to tens of thousands of tables and cryptic schemas, often results in missing entire tables or columns containing critical elements of the target patient profile
  - Transformation errors:
    - Incorrect source to target mappings (misclassification of the target element)

- Errors in parsing data elements from unstructured or semi-structured fields in source RWD, such as a single source field containing the a drug name, dose, route and frequency
- Imperfect algorithms (e.g., natural language processing (NLP) algorithms or computable phenotyping algorithms), which in practice cannot perfectly extract/classify patient variables
- Bugs and typos in manually generated query code

Impact:
Because the complexity of RWD source data is high, and derived study variables often involve many-to-one mappings from source data, even small error rates in transformations can lead to large numbers of mistakes and systematic errors/biases in the final analysis dataset

Overall, there is rampant information loss occurring in the transformation of RWD into CDMs and analysis datasets, typically without an understanding of what has been lost. This has major impacts:

- Lack of transparency/auditability of how output data elements were derived
- Implausible results
- Irreproducibility
- Potentially biased study findings
- Data and generated evidence are not fit for high stakes decision-making due to high degrees of and unquantified uncertainty

DROICE HAWK - RWD INFRASTRUCTURE SOLUTION FOR ECAs

Bayer has been using Droice Hawk, an AI-based middleware infrastructure for patient data preparation, to connect to multiple diverse RWD sources from the US and Europe and transform the data into analysis-ready patient RWPs for a variety of studies across several therapeutic areas (Figure 2). Hawk capabilities include:

- **Integration of complex and diverse RWD inputs.** Data infrastructure, compliance layers, connectivity pipes, and software to integrate with, ingest, and process a massively diverse multiplicity of complex RWD input sources from varied health IT infrastructures (e.g. hospitals x EHRs x vendors x site implementations x geographies x languages)
- **Handling the multiplicity of research requirements for analysis data outputs.** Capability to generate and manage an equally diverse set of output datasets (e.g. CDMs x therapeutic areas x study teams x studies x analysis datasets) to serve the full spectrum of pharma use cases
- **RWD preparation at scale.** Trained and tuned on diverse RWD of millions of patients from around the globe, Hawk losslessly and comprehensively ingests source data and transforms/harmonizes the required patient data elements into the target data model using a configurable software approach (as opposed to manual/custom query writing)
- **Auditability and measurable data quality.** SuperLineage™ - a regulatory grade, standardized lineage solution that integrates lineage across disparate data sources to enable audits and quantification of errors and biases in the data

The net result of Hawk infrastructure is seamless handling of the multiplicity encountered in the pharma RWD research environment – multiple input sources/geographies/data types, multiple study teams, and multiple studies/therapeutic areas, to scale ECA generation across the enterprise.

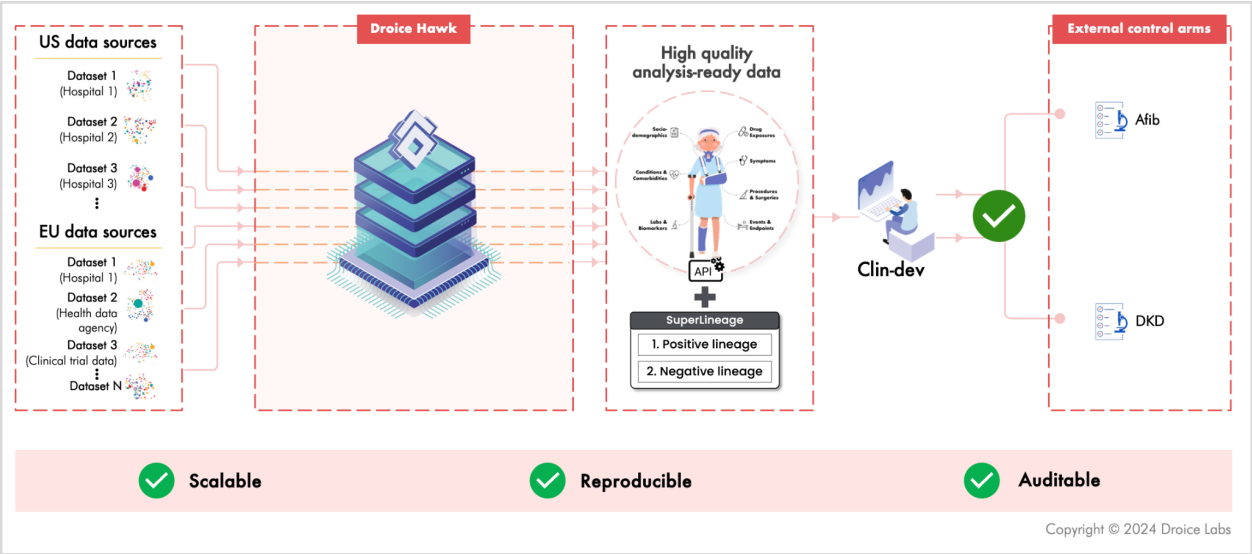

Figure 2: Droice Hawk – AI middleware RWD infrastructure for scalable preparation of high-quality analysis ready data with auditable and quantifiable data transformation performance.

CONCLUSION

RWE will be increasingly used in high stakes decision-making such as regulatory submissions, which will require scalable preparation of RWD that meets the same quality and auditability requirements as clinical trial data. The current ad hoc processes typically employed for generating analysis data from RWD will not be suitable or sufficient to handle the complexity and multiplicity involved in generating usable ECAs. Enterprise-level infrastructure for scalable, reproducible, and auditable RWD preparation with quantifiable data quality, such as enabled by Droice Hawk and SuperLineage™, will be critical for reliable RWE generation from RWD ECAs at scale.

REFERENCES

1. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products
2. Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products
3. Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey
4. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance
5. Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method
6. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform
7. Analyzing the Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms
8. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Mayur Saxena
Company: Droice, Inc.
Address: 31 West 34th Street, Suite 7118, New York City, NY 10001
Work Phone: +1 (929) 224-0049
Email: mayur@droicelabs.com
Website: https://www.droicelabs.com/