MSR 82

A Methodological Framework to Conduct Data Source Landscaping for Real-World Studies

Zhen Dong[1], Shiau-Han Chen[2], Anne-Frieda Taurel[1], Kong Chian Toh[1], Nora J. Kleinman[3]
[1]IQVIA Solutions Asia Pte Ltd, Singapore; [2]IQVIA Solutions Taiwan Ltd, Taiwan; [3]IQVIA Solutions Hong Kong Ltd, Hong Kong SAR China

Background

- Real-world data (RWD) are increasingly used to capture large and diverse patient populations in a cost and time effective manner to complement and support evidence generated in randomized clinical trials[1-3].
- In particular, real-world evidence can be used strategically to support products across their lifecycle, notably as part of launch excellence[4].
- The applicability of RWD to generate high-quality evidence should be carefully evaluated, especially in the context of regulatory or reimbursement purpose[5].
- Therefore, early identification of RWD sources is imperative to determine their suitability for evidence generation to support product strategy[6-7].
- We propose a methodological framework that can be adopted for data source landscaping to identify fit-for-purpose real-world secondary data sources.

Results

- A flexible five-step framework was developed to identify and select data sources, adaptable to disease areas and research questions:

1. Data source identification

Conduct a pragmatic literature search targeting publications with specific populations, diseases, and retrospective study designs

Activities

- Develop search strings based on study objectives and disease area, and apply them in search engines
- Exclude non-related topics based on abstract review; de-duplicate included publications

Case example for IED

- IED presents in many forms associated with multiple medical coding or key words, thus comprehensive search terms with multiple disease manifestations of IED (e.g., bacteremia, sepsis) were used

2. Data source extraction and categorization

Extract data sources from included publications and assess suitability for potential study objectives

Activities

- Generate a long-list of data sources by extracting information from included publications
- Categorize listed data sources into different data source types adapted for each specific study objective

Case example for IED

- Data sources were categorized based on geographic coverage (e.g., national, single site) and type (e.g., electronic medical records [EMR], surveillance) as surveillance may provide more details for infectious diseases

3. Data source screening

De-prioritize data sources for technical or operational considerations

Activities

- Generate a short-list of data sources by de-prioritization for technical considerations (e.g., small sample size, sub-population focus, inactive data sources) and operational considerations (e.g., unwilling to collaborate with industry)

Case example for IED

- Data sources focusing on uropathogenic E. coli were de-prioritized as they are a subgroup of ExPEC

4. Data source evaluation

Assess data sources based on an evaluation matrix, including key variable availability and data source accessibility

Activities

- Develop a research-specific evaluation matrix with dimensions accounting for availability of key patients' data, operational accessibility, and rank score to represent extent of variable coverage (Example: **Table 1**)
- Conduct desk research of data sources to expand understanding and assess variable coverage

Case example for IED

- Lab data was included in the evaluation matrix because it is required for a key IED diagnostic criteria

5. Data source selection

Select data sources with high scores from the evaluation matrix for deep-dive feasibility

Activities

- Develop a research specific algorithm for recommendation tier based on data source type, patient count, and variable coverage (Example: **Table 2**)
- Select data source for further investigation based on recommendation tier and, where possible, inputs from experts with experience with the data sources

Case example for IED

- Recommendation tier was upgraded if data sources could identify IED with strict definitions (i.e., E. coli present in blood and patient had at least one symptom of systemic inflammatory response syndrome criteria)

Methods

We used an inductive approach[8] to develop a methodological framework to systematically identify, characterize, and select appropriate data sources. This approach was based on detailed understanding of RWD source requirements for secondary database research and empirical cases.

Empirical Case Example

A project exploring potential data sources to conduct real-world studies for invasive Extraintestinal pathogenic E. coli (ExPEC) disease (IED) in Asia Pacific (APAC).

Table 1. Illustrative evaluation matrix of data sources

Data Source				Publication[a]			Variable Coverage[b]			Score
ID	Data source type	Geographic coverage	Open to industry studies	Disease coverage of data source	Number of patients	Study period	Patient characteristics	Treatment patterns	Lab data	
1	EMR	Multi-site	Y	CRE	109	2013-2017	N	N	N	3
2	EMR	Single-site	Y	UPEC	7,713	2012-2019	Y	N	Y	7
3	Surveillance	National	Y	E.coli	1,618	2005-Now	N	N	Y	5
4	Surveillance	Multi-site	Y	CRE	664	2015	Y	Y	Y	9
5	EMR	Regional	Not sure	CRE	504	2016-2017	N	N	Y	6

[a] Collected from publicly available information; [b] Selected variables, in total 10 categories were evaluated

Table 2. Algorithm for grouping data source into recommendation tiers

Tier	Data source type		Patient count		Score
Tier 1	National/regional	AND	>1,000	AND	Good (8-10)*
Tier 2A	National/regional	AND	<1,000	OR	Moderate (5-7)*
	Single-site EMR/others	AND	>100	AND	Good (8-10)*
Tier 2B	National/regional	AND	<100	OR	Average (3-4)
	Single-site EMR/others	AND	<100	AND	Good (8-10)
	Single-site EMR/others	AND	>100	AND	Moderate (5-7)
Tier 3	All data sources	AND	<100	AND	Low (<4)

* With potential linkage to increase variable coverage

Conclusions

- This framework covers the essential steps in a data source landscaping, which can be adapted for different disease areas and research questions.
- Following a data source landscaping, deep-dive feasibility assessments are recommended to confirm its fit for purpose for the target research question by further investigating variable quality and completeness and key operational considerations.

Acknowledgements

- The case shared in this poster was sponsored by Janssen APAC.

Abbreviation: CRE, Carbapenem-Resistant Enterobacteriaceae; UPEC, uropathogenic E. coli

Reference:
1.Hiramatsu, K., et al. Drugs – real world outcomes, 8(4), 459–480.
2. Crane, G., et al. Current medical research and opinion, 38(9), 1543–1551.
3. van Dorn A, et al. Lancet (London, England), 396(10250), 523–524.
4. IQVIA White Paper. Launch Excellence VII. The three pillars of post-Pandemic Launch Excellence. Available at: https://www.iqvia.com/library/white-papers/launch-excellence-vii.
5. National Medical Products Administration (NMPA). Center for Drug Evaluation (CDE), "Guideline on Using Real-World Data to Generate Real-World Evidence (Pilot)"
6. Davies, M., et al. International journal of environmental research and public health, 19(3), 1329.
7. Waite, S., et al. Canadian review of sociology = Revue canadienne de sociologie, 56(1), 93–117
8. Goddard, W. et al. Available at: https://books.google.com.tw/books/about/Research_Methodology.html?id=bJQJpsU2a10C&printsec=frontcover&source=kp_read_button&hl=en&redir_esc=y#v=onepage&q&f=false

Correspondence: nora.kleinman@iqvia.com

Presented at ISPOR 2023 ; Boston, MA, USA ; May 7-10 2023