Guiding Record Linkage Method (RL) Selection in HEOR: A Targeted Review and Structured Decision Framework

Bruce Morrison, MSEd[1]
[1]BeOne, Ltd, San Carlos

CONCLUSIONS

This framework reflects the reviewed literature, but subject-matter expertise should take precedence, particularly where linked files may contain duplicate entity records.
Further guidance for applying RL methods in HEOR and RWD contexts is needed

BACKGROUND

- Real-world oncology data are inherently fragmented across administrative, registry, and clinical sources, each capturing an incomplete picture of patient disease history, treatment, and outcomes[1,4,8]
- Linking registry data with electronic health records can substantially improve completeness across disease, treatment, and health-status variables underrepresented in any single source[1]
- Linkage can shape cohort definition, outcome capture, covariates, and effect estimation; therefore, false and missed links pose direct threats to validity of Health Economics and Outcomes (HEOR) and real-world-evidence (RWE) analyses[3,8]

OBJECTIVES

- **The objective of this study was to develop a tool that can help users assess record linkage (RL) candidacy by respective use-case, with a focus on HEOR and RWE**

METHODS

- A structured targeted review was conducted to identify sources relevant to linkage decisions in HEOR-oriented real-world data settings
- Sources (n=20) were selected to cover foundational linkage methods, preprocessing and data quality, linkage error and bias, privacy and regulatory constraints, and applied healthcare and oncology linkage use-cases
- Search concepts combined record-linkage terms with healthcare, oncology, bias, machine-learning, Bayesian linkage, privacy preservation, and regulatory themes, and were supplemented by citation chaining from methods papers and official guidance documents
- The resulting evidence base was synthesized into a decision framework

RESULTS

- The earlier branches of the decision tree (**Figure 1**) evaluate linkage permissibility, availability of usable identifiers or privacy-preserving tokens, likely analytic consequences of linkage error, and field fitness following post-preprocessing[3,8,12,15-19]
  - Preprocessing (encompassing identifier standardization, missing value handling, blocking design, and dataset alignment) must be completed before method selection can proceed[2,6,9,12,14]
- For approx. 1:1 linkage tasks, the lower branches guide method selection based on three sequential considerations: whether privacy constraints permit direct identifier sharing, whether labeled training data are available and supervised modeling overhead is acceptable, and whether the dataset structure supports a canonical Fellegi-Sunter workflow or warrants a fully Bayesian RL[5,7,9-11,13,18-20]
- While the decision framework presented generally captures the literature reviewed, subject-expertise should always take precedence over the decision tree framework, especially when ≥1 files considered for linkage can be expected to contain ≥1 of the same entity[9,11,13,19]

Method Classes Captured Within This Review:

1 Privacy-Preserving Record Linkage (PPRL)

BEST SUITED FOR

- ***Settings where data governance restricts identifier sharing; a material tradeoff among privacy, linkage quality, and scalability typically applies***

- Performs without direct exchange of raw identifiers, usually achieved by encoding identifier fields and using a privacy-preserving protocol[18,19]

2 Supervised Machine Learning Linkage

BEST SUITED FOR

- ***Settings where labeled data are available and improved predictive performance justifies additional tuning and implementation overhead***[5,9,19]

- Uses labeled or adjudicated pairs to train classification rules for potential record pairs. Best when suitable training data exist, and performance gains outweigh tuning overhead[5,9]

3 Bayesian Probabilistic Linkage

BEST SUITED FOR

- ***Settings requiring 1:1 constraints should apply, shared fields are sparse or noisy, or Fellegi-Sunter weights would be unreliable***[10,11,13]

- Models the latent 1:1 structure directly, representing linkage uncertainty through posterior inference. Incorporates constraints during estimation

4 Probabilistic Fellegi-Sunter Linkage

BEST SUITED FOR

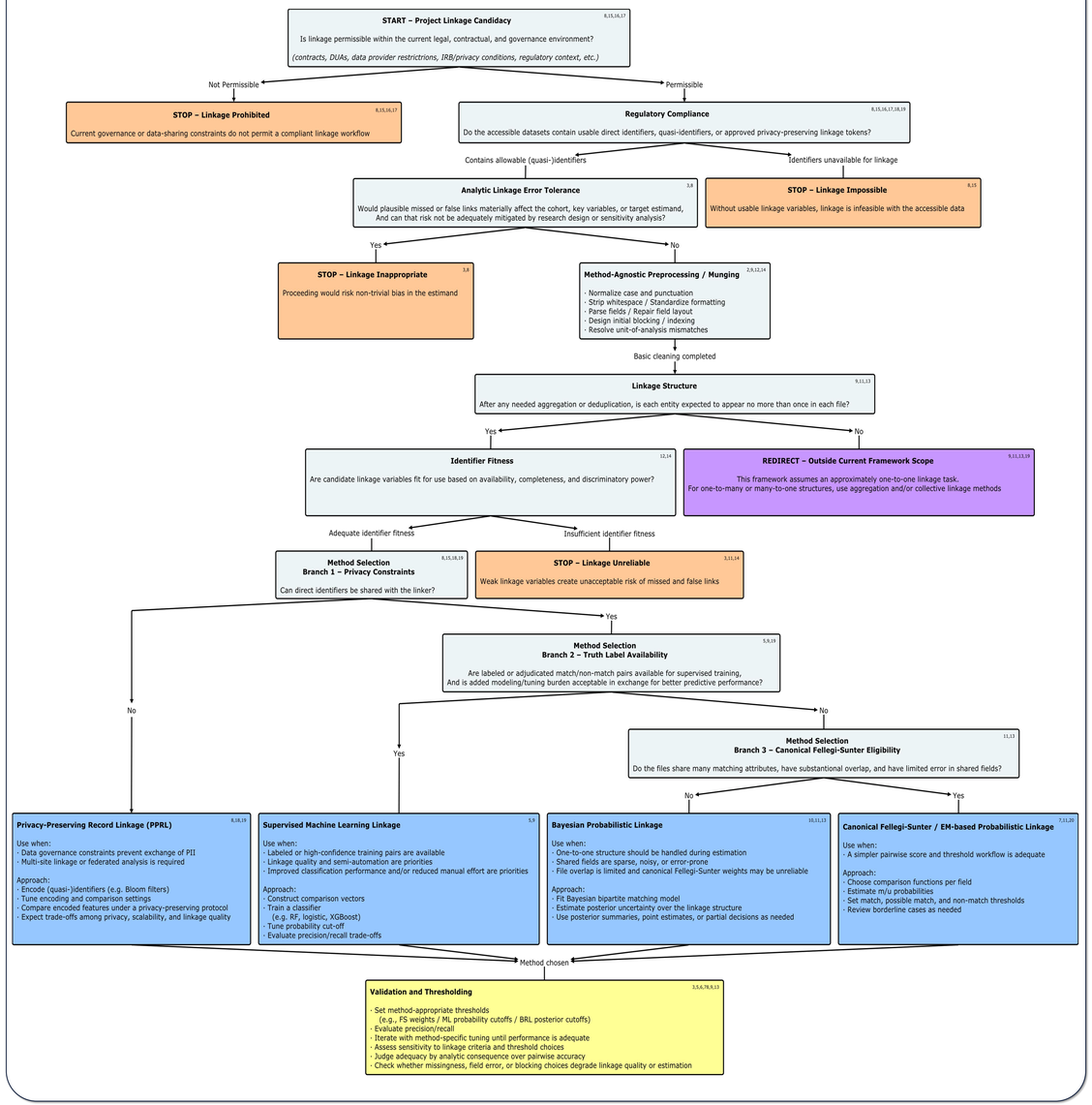
- ***Pairwise-score-and-threshold workflow is adequate, files share many matching attributes, and shared fields carry limited error***[6,7,11,20]

- Estimates match likelihoods from identifier comparison patterns, assigns match, possible-match, and non-match decisions under threshold rules

Validation and Interpretation:

- Validation is a necessary component of any record linkage workflow regardless of method chosen[3,8]. Typical steps include threshold setting, manual review of ambiguous pairs, and data quality checks calibrated to the specific use case[7,8,9,20]
- Missing data and measurement error can substantially affect both method performance and estimation accuracy, making use-case-specific validation essential[6,12,14]
  - While precision and recall are important benchmarks, the primary determinant of acceptable linkage performance is whether design decisions and their limitations could materially affect the estimand or downstream analysis results[3,6,8]

Figure 1: Structured Decision Framework



REFERENCES

1. Charlton, et al., *JCO Clinical Cancer Informatics*. 2022. [2]Christen, P., *Sixth IEEE International Conference on Data Mining Workshops*. 2006. [3]Doidge, J.C., Harron, K.L., *International Journal of Epidemiology*. 2019. [4]Eisinger-Mathason, T. S., et al., *Communications Medicine*. 2025. [5]Ellum, R., et al., *International Journal of Population Data Science*. 2023. [6]Enamorado, T., et al., *American Political Science Review*. 2019. [7]Fellegi, I. P., Sunter, A. B., *Journal of the American Statistical Association*, 1969. [8]Harron, K., et al., *Big Data & Society*. 2017. [9]Mason, L., *Bureau of Labor Statistics*. 2018. [10]McVeigh, B. S., Murray, J. S., *Practical Bayesian Inference for Record Linkage*. 2017 [11]McVeigh, B. S., et al., *Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking*. 2020. [12]Ong, T.C., et al., *Journal of the American Medical Informatics Association*. 2024. [13]Sadinle, M., *Bayesian Estimation of Bipartite Matchings for Record Linkage*, 2016. [14]Sariyar, M., et al., *Journal of the American Medical Informatics Association*. 2012. [15]U.S. Department of Health and Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule*. 2012. [16]U.S. Food and Drug Administration. *Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products: Guidance document*. 2024. [17]U.S. Food and Drug Administration, *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices*. 2025. [18]Vatsalan, D., Christen, P., *Multi-Party Privacy-Preserving Record Linkage using Bloom Filters*. 2016. [19]Vatsalan, D., et al., *Information Systems*. 2013. [20]Winkler, W., *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods, American Statistical Association. 1990.

Presented at 2026 ISPOR | Philadelphia, PA | May 17-20