graticule™

Beyond Tokenization: Considerations for Linking Healthcare Data Sets for Scientific Research

Yuval Koren & Jenny Dusendang, Graticule Inc.

PHUSE EU Connect 2025
Paper ID RE03

Introduction & Background

Background

- Linking data sources can help derive novel insights

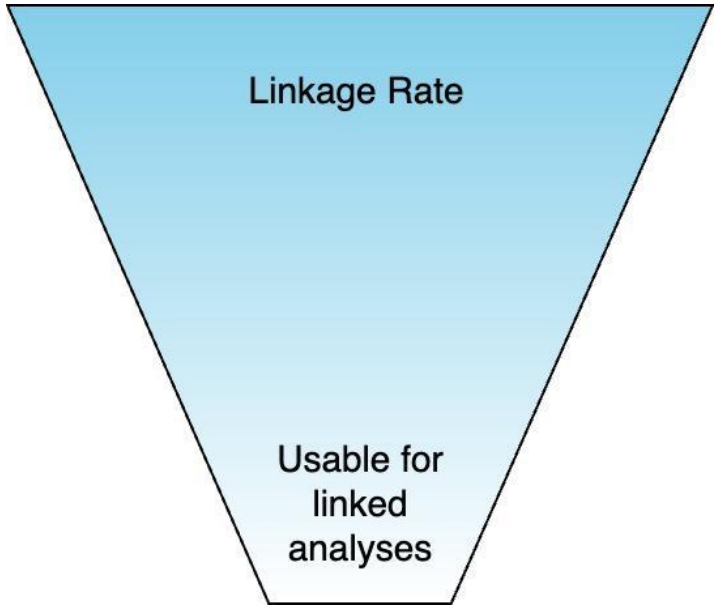
Background

- Linking data sources can help derive novel insights
- Linkage performed by Privacy-Preserving Record Linkage (PPRL)

Background

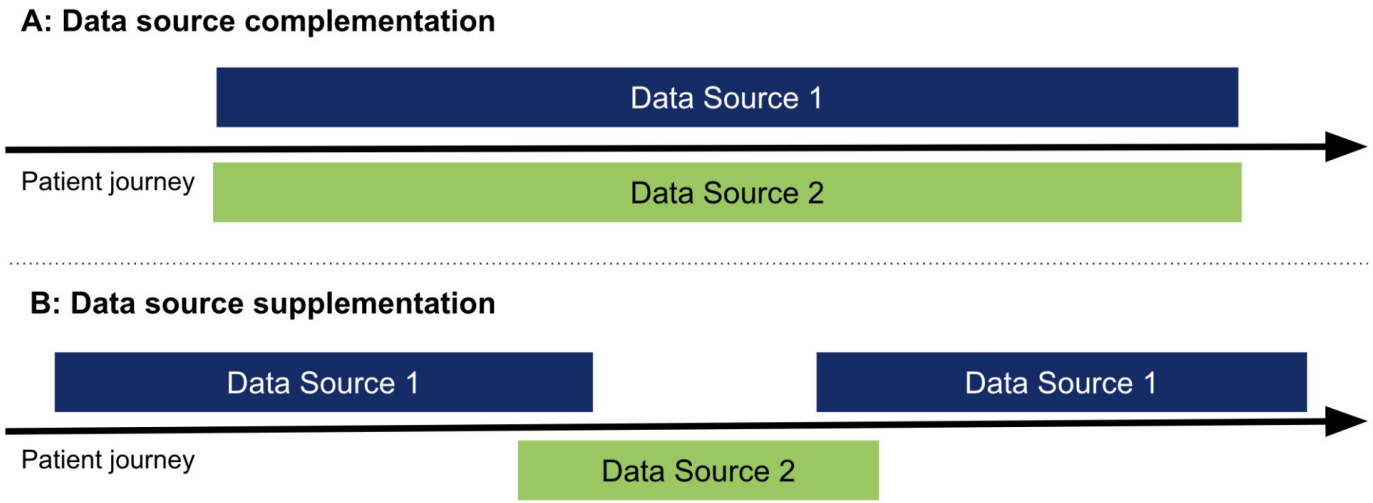
- Linking data sources can help derive novel insights
- Linkage performed by Privacy-Preserving Record Linkage (PPRL)
- Linkage Rate / Proportion: % of tokens from a primary data source that appear in a secondary data source

Top of Attrition Funnel: Linkage Rate

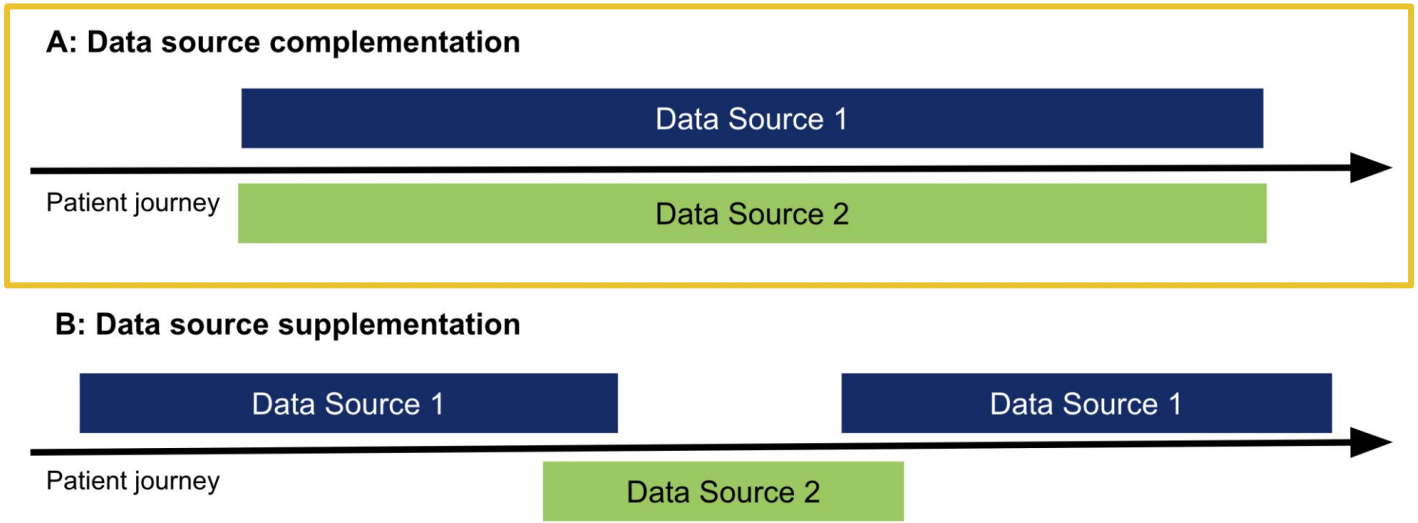


Types of Linkage Studies

Types of Linkage Studies

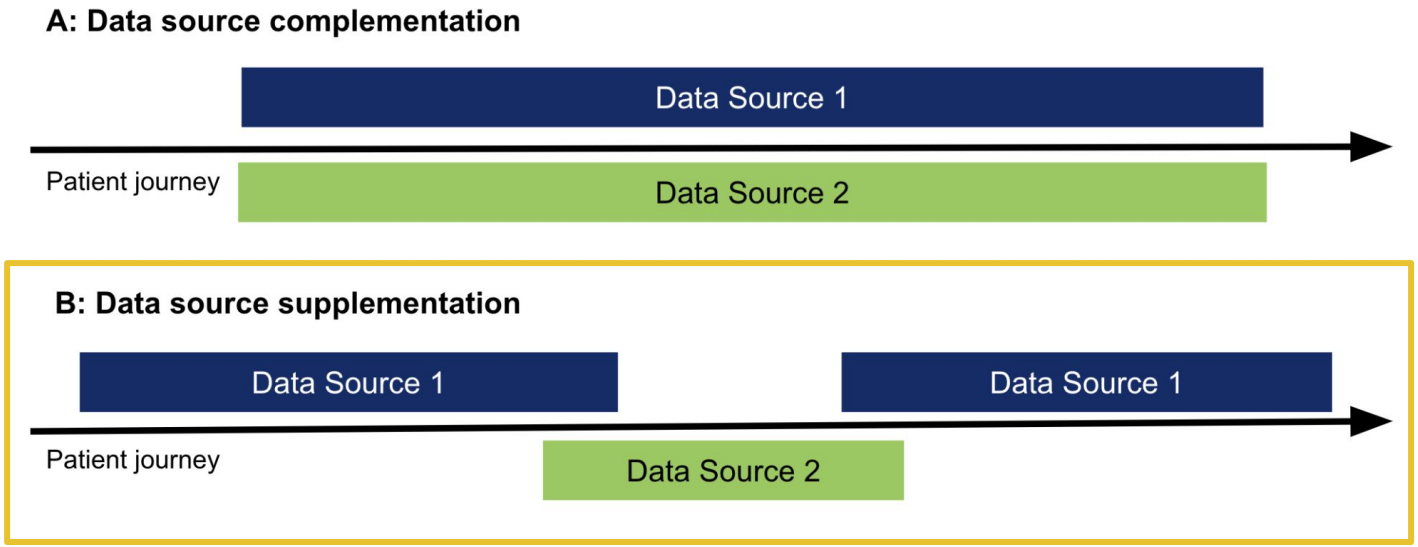


Types of Linkage Studies



Different data elements not available in original data

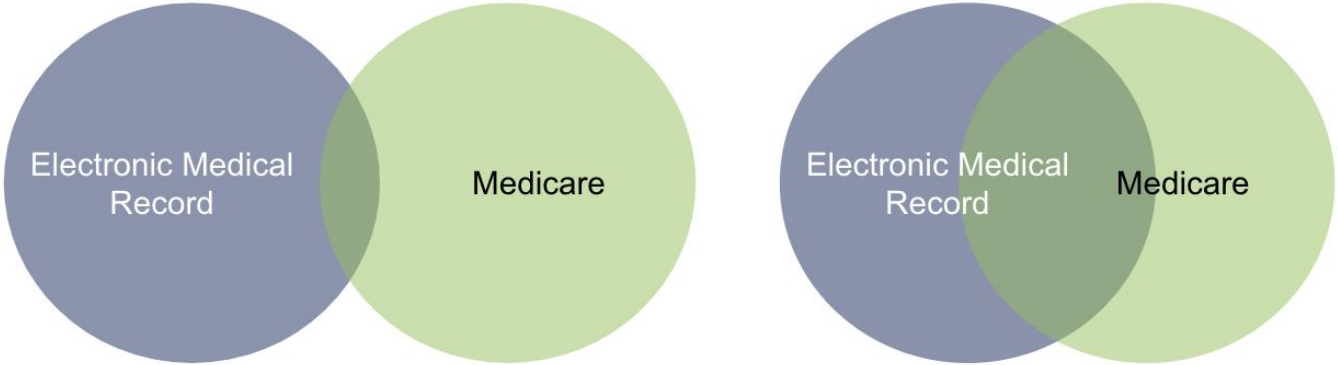
Types of Linkage Studies



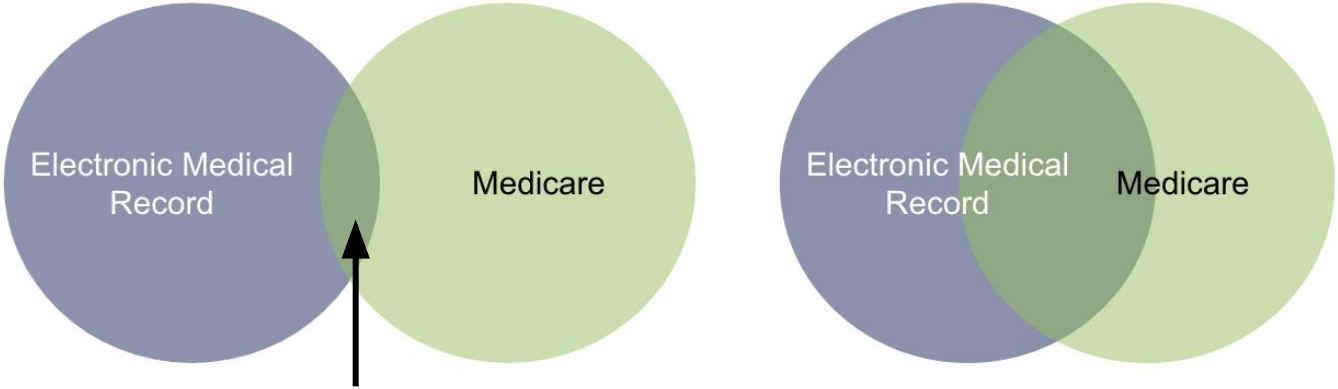
Same data elements to fill in gaps

Multiple Dimensions of Attrition

Base Population Overlap

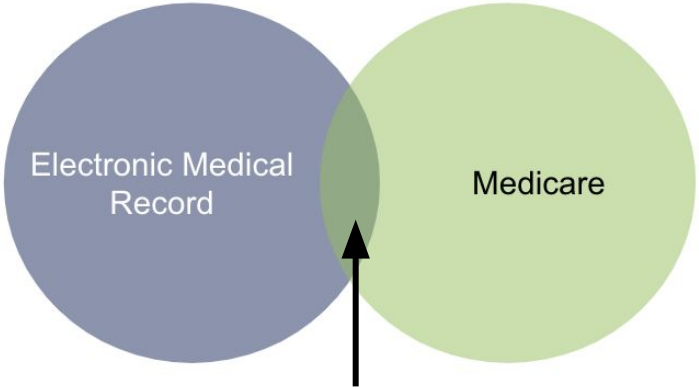


Base Population Overlap

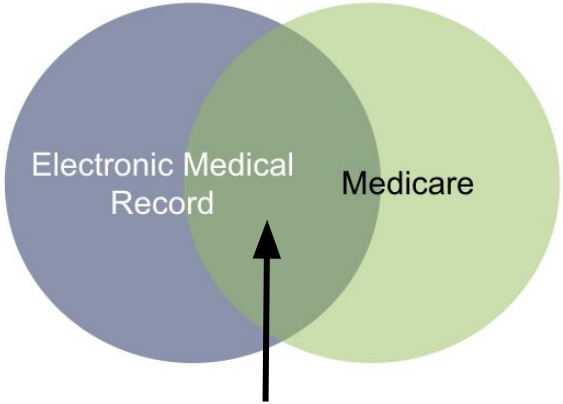


EMR cohort of younger patients:
Small linked cohort

Base Population Overlap

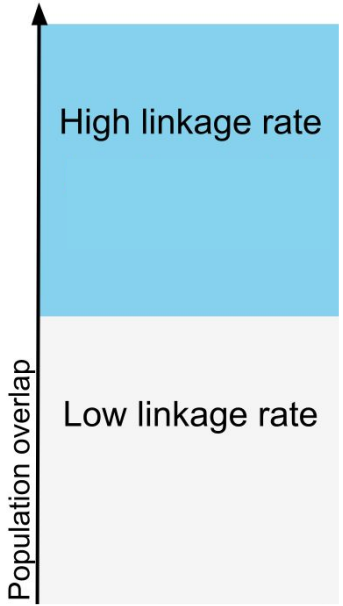


EMR cohort of younger patients:
Small linked cohort



EMR cohort of older patients:
Larger linked cohort

Population Overlap



Data Continuity Periods

Enrollment spans:

Data has periods defined during which we can expect data

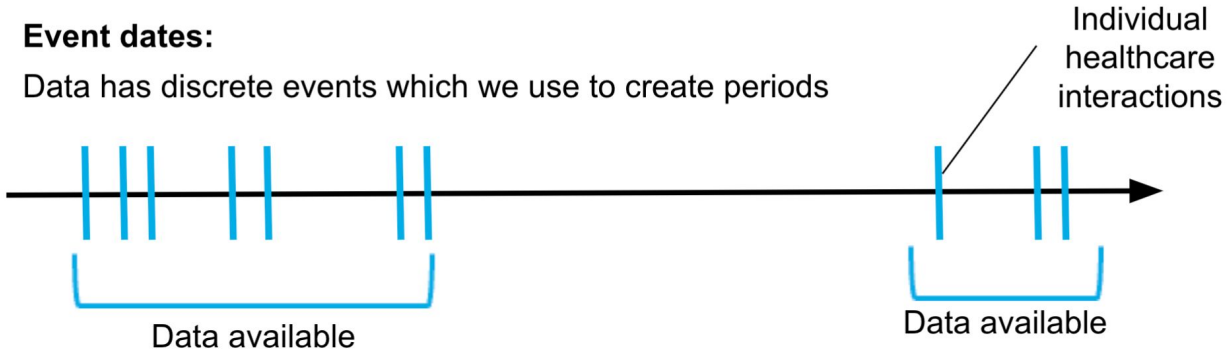
Data Continuity Periods

Enrollment spans:

Data has periods defined during which we can expect data

Event dates:

Data has discrete events which we use to create periods



Time Period Overlap

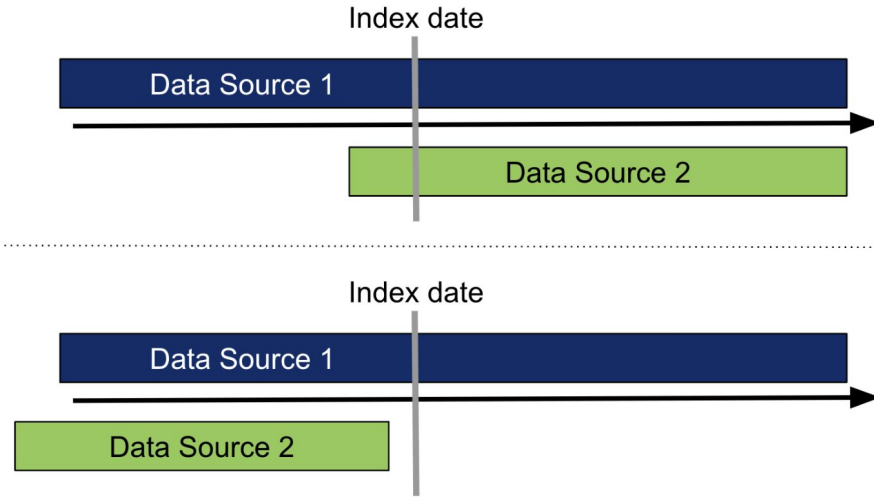
Example study requirements:

1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2

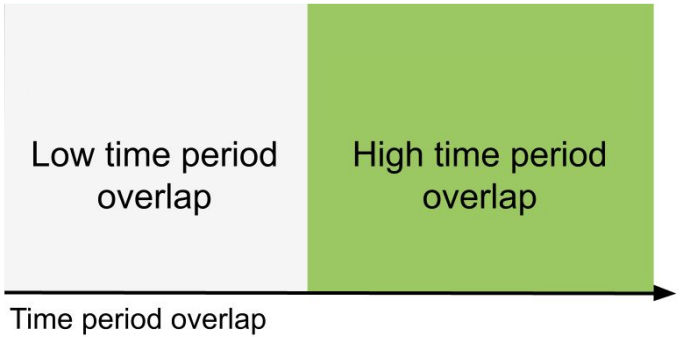
Time Period Overlap

Example study requirements:

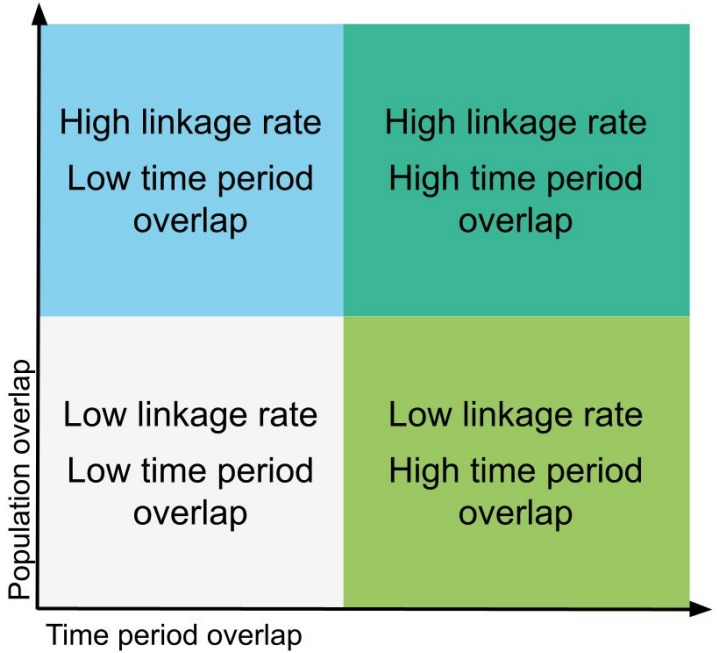
1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2



Time Period Overlap



Population and Time Period Overlap



Real-World Tokenization Trade-Offs

Linkage Considerations

- Which PPRL tokens are available in both data sources?

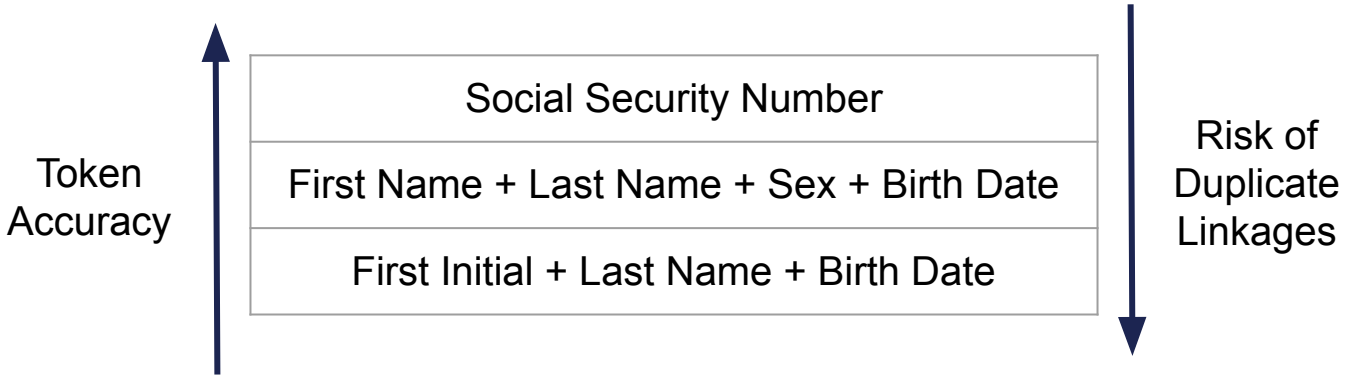
Linkage Considerations

- Which PPRL tokens are available in both data sources?
- How many patients have these tokens available in each data source?

Linkage Considerations

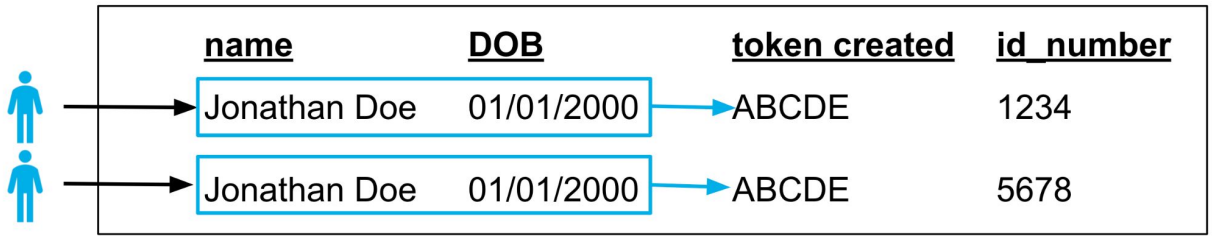
- Which PPRL tokens are available in both data sources?
- How many patients have these tokens available in each data source?
- How accurate are these tokens in identifying patients?

Accuracy vs Sample Size Tradeoffs

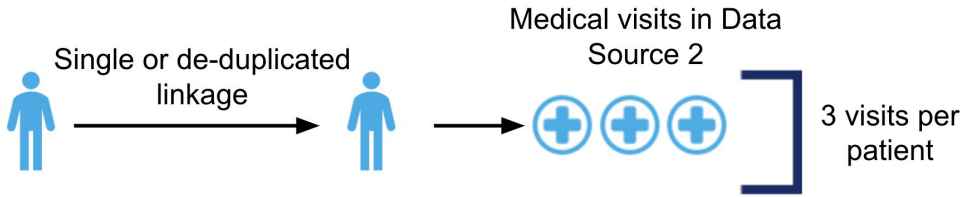


Tokenization Yields Duplicates

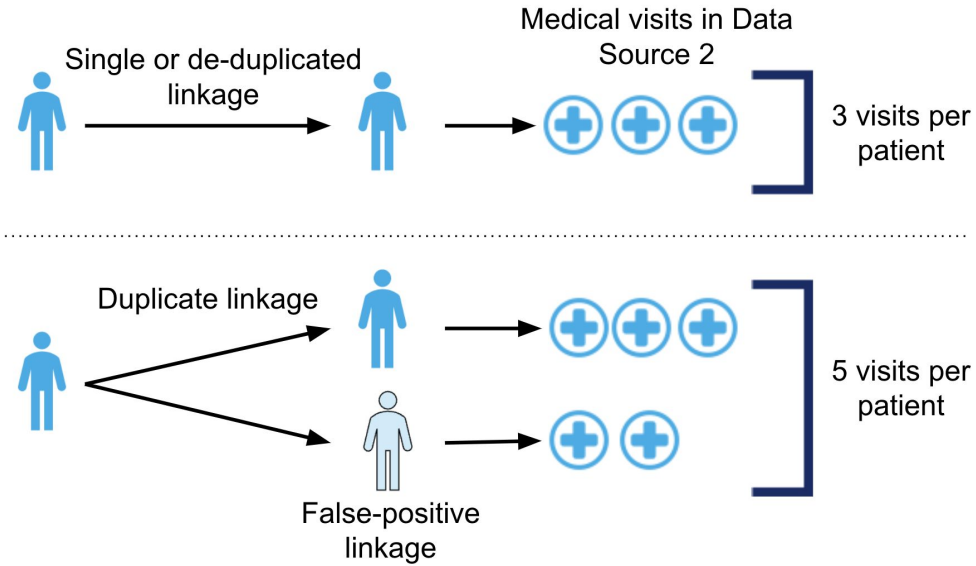
Example A: 2 unique individuals with the same identifiers



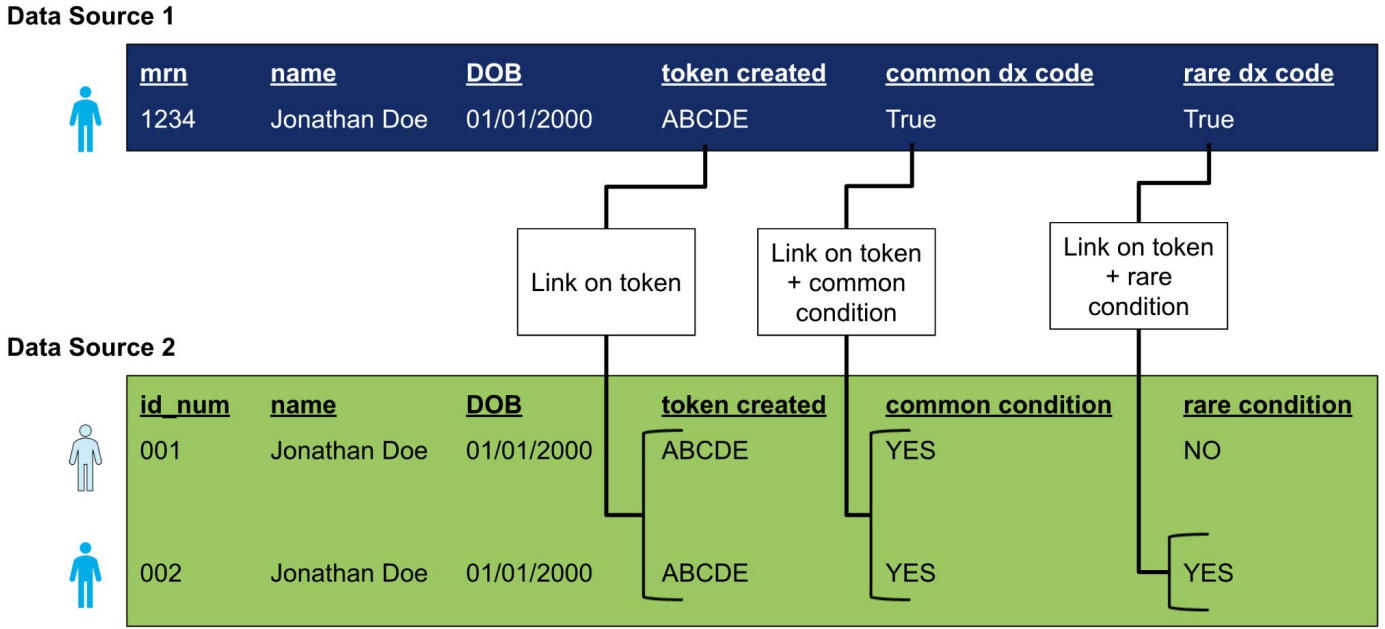
Duplicates Can Impact Results



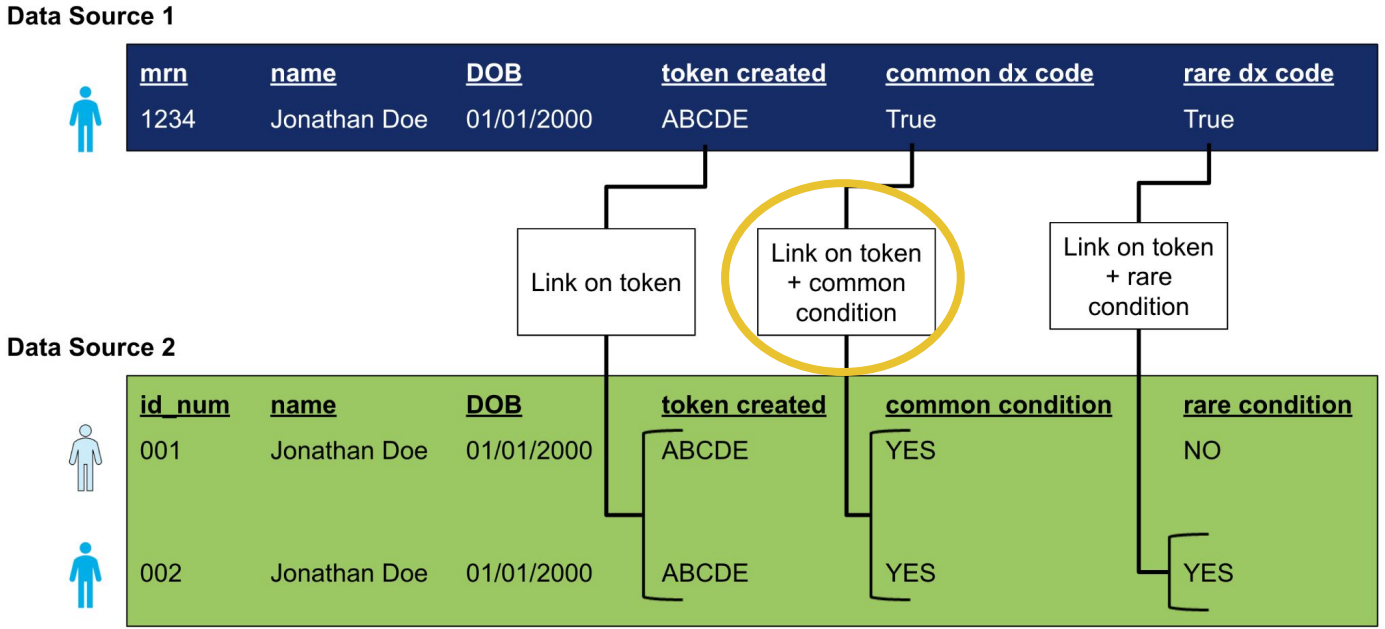
Duplicates Can Impact Results



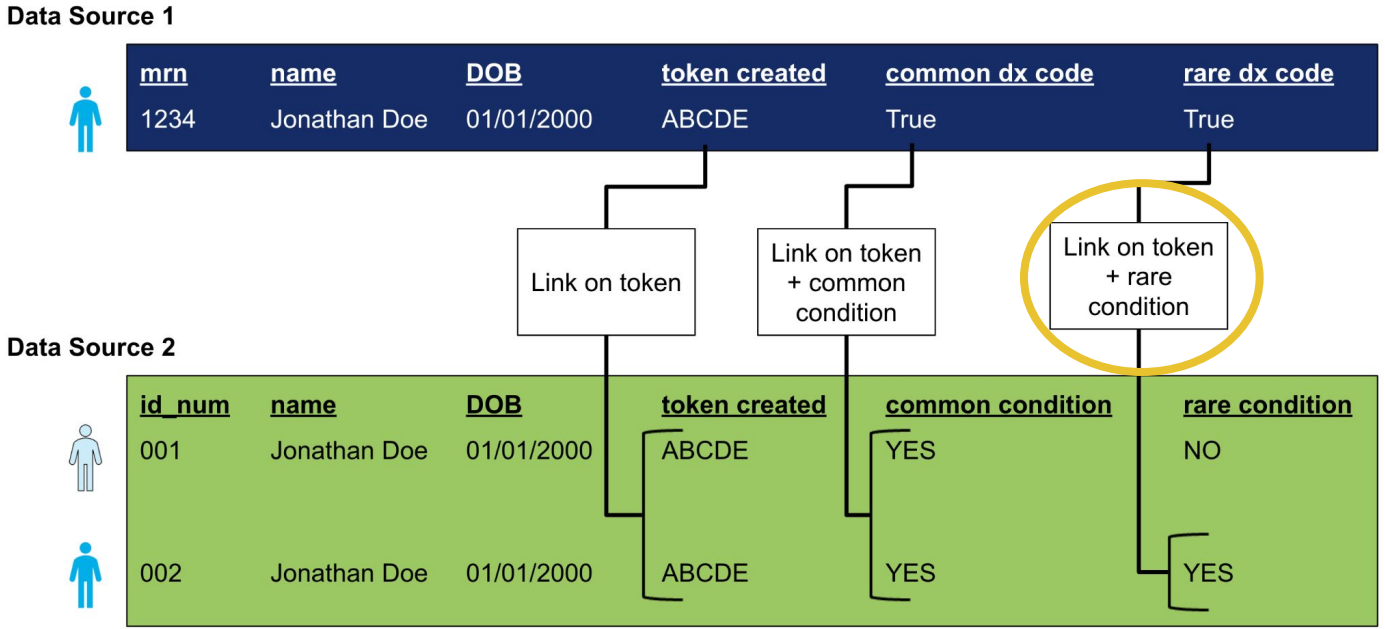
Deduplication Strategies



Deduplication Strategies

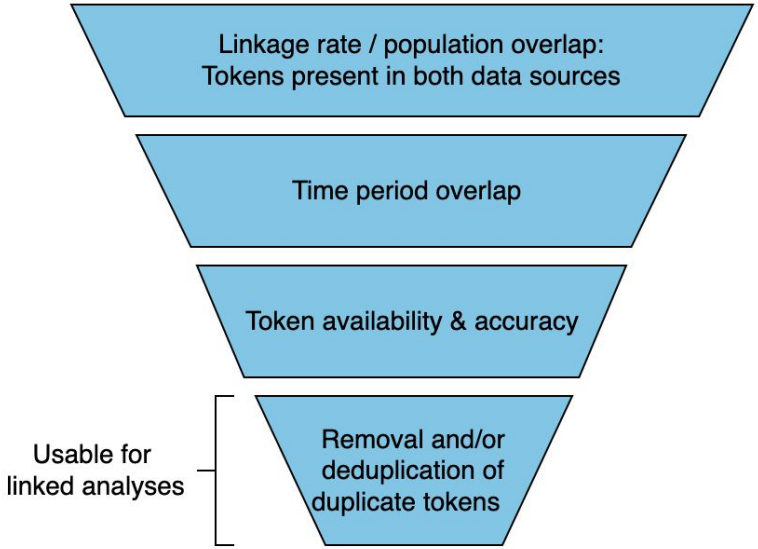


Deduplication Strategies



Conclusions

Sample Size and Feasibility



graticule™

Questions?

Beyond Tokenization: Considerations for Linking Healthcare Data Sets for Scientific Research

ykoren@graticule.life