

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

Jenny Dusendang & Yuval Koren, Graticule Inc.

PharmaSUG 2025

Introduction & Background

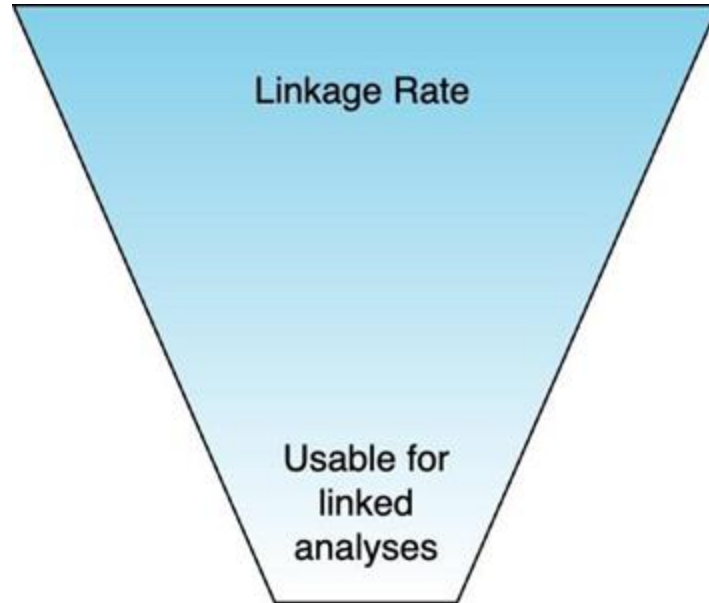


Background

- 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



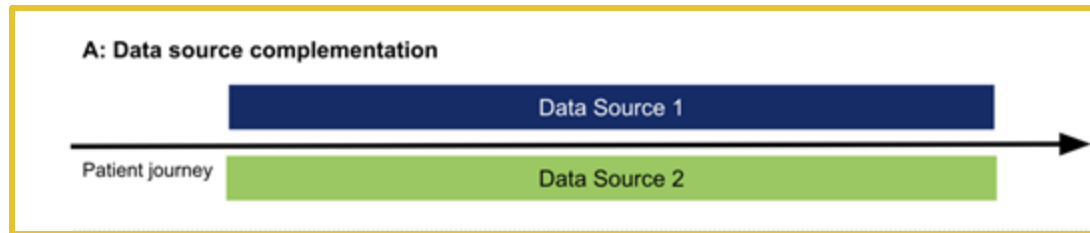
Top of Attrition Funnel: Linkage Rate



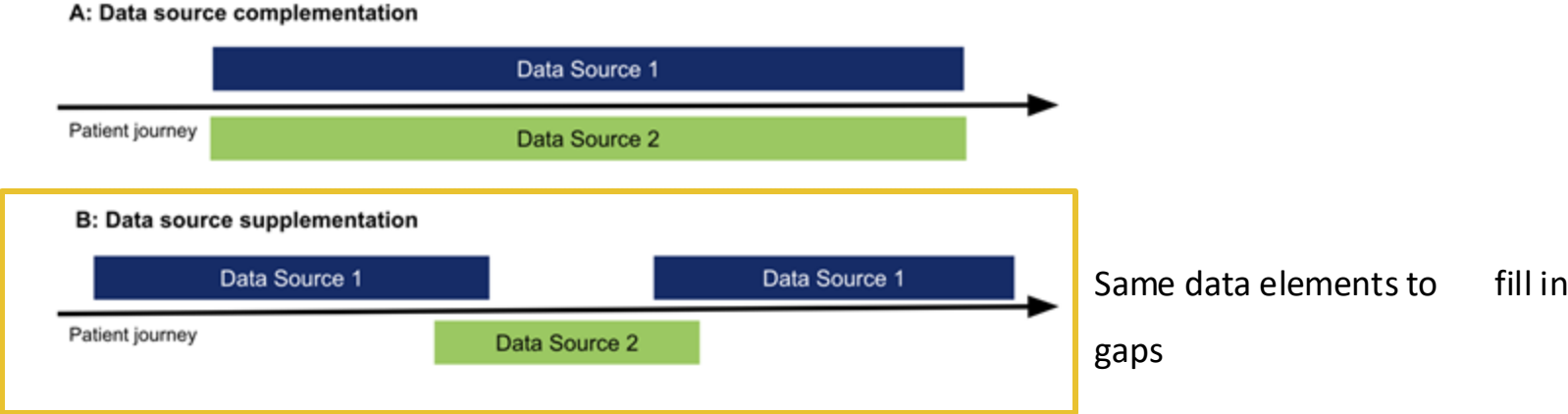
Types of Linkage Studies



Types of Linkage Studies



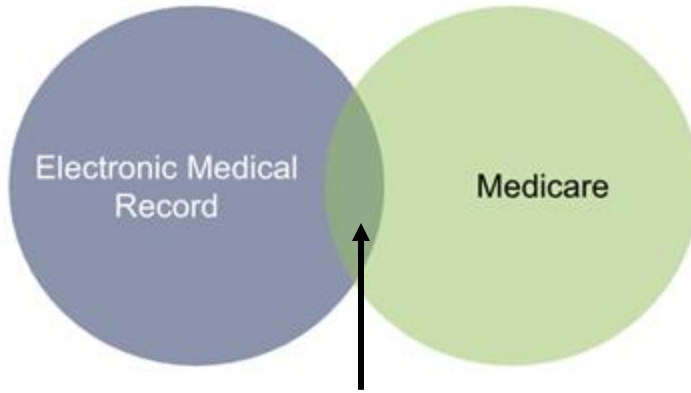
Types of Linkage Studies



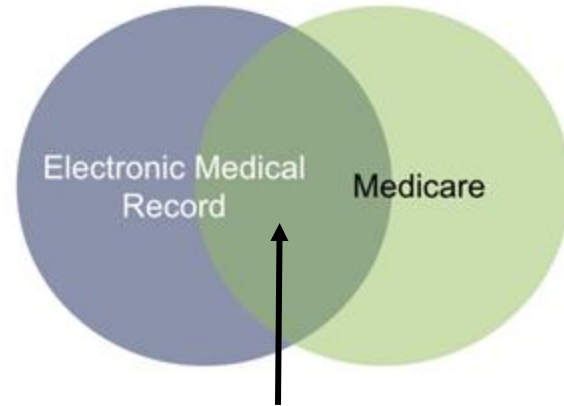
Multiple Dimensions of Attrition



Base Population Overlap



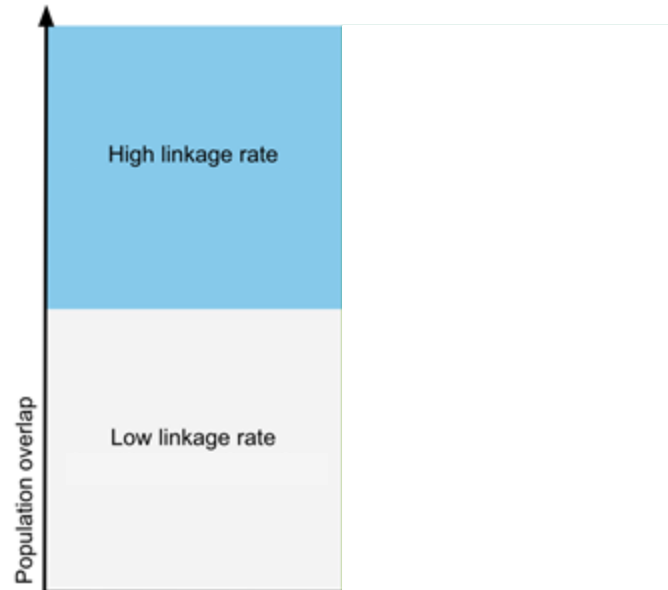
EMR cohort of younger patients:
Small linked cohort



EMR cohort of older patients:
Larger linked cohort



Population Overlap



Data Continuity Periods

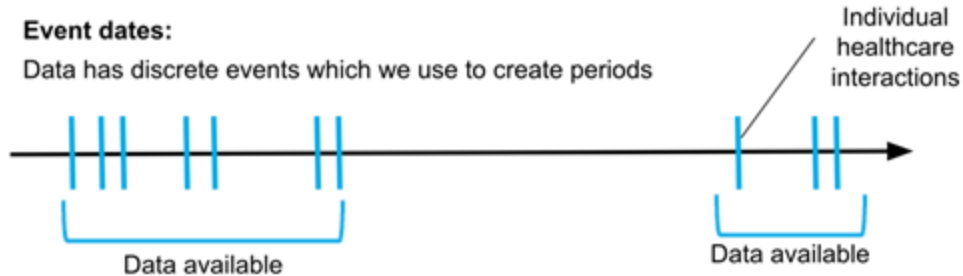
Enrollment spans:

Data has periods defined during which we can expect data



Event dates:

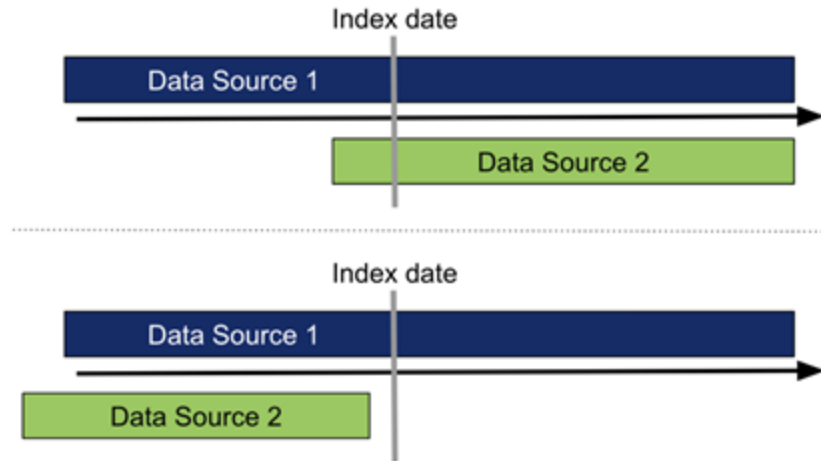
Data has discrete events which we use to create periods



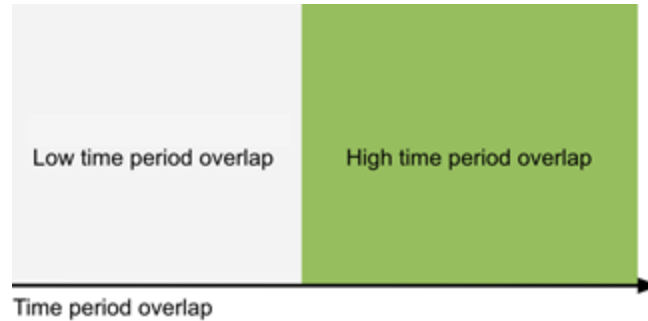
Time Period Overlap

Example study requirements:

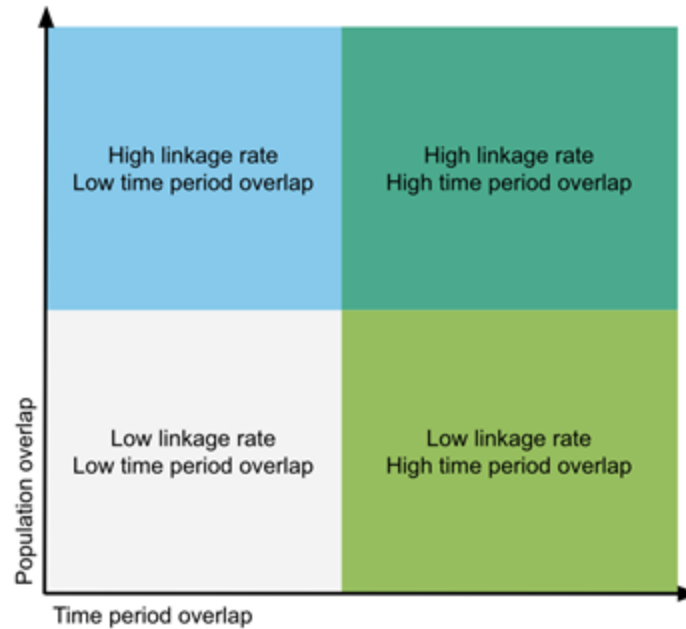
1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2



Time Period Overlap



Population and Time Period Overlap



Real-World Tokenization Trade-Offs

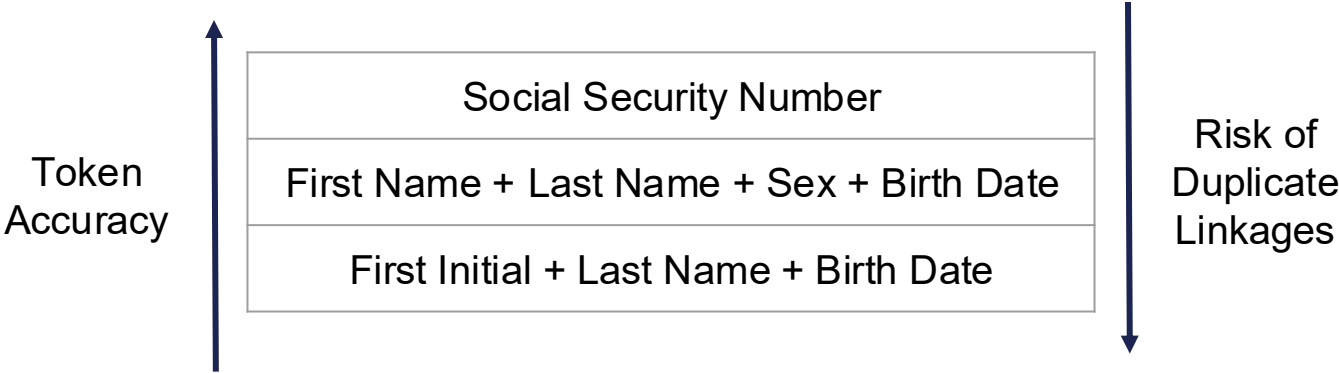


Linkage Considerations

- 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



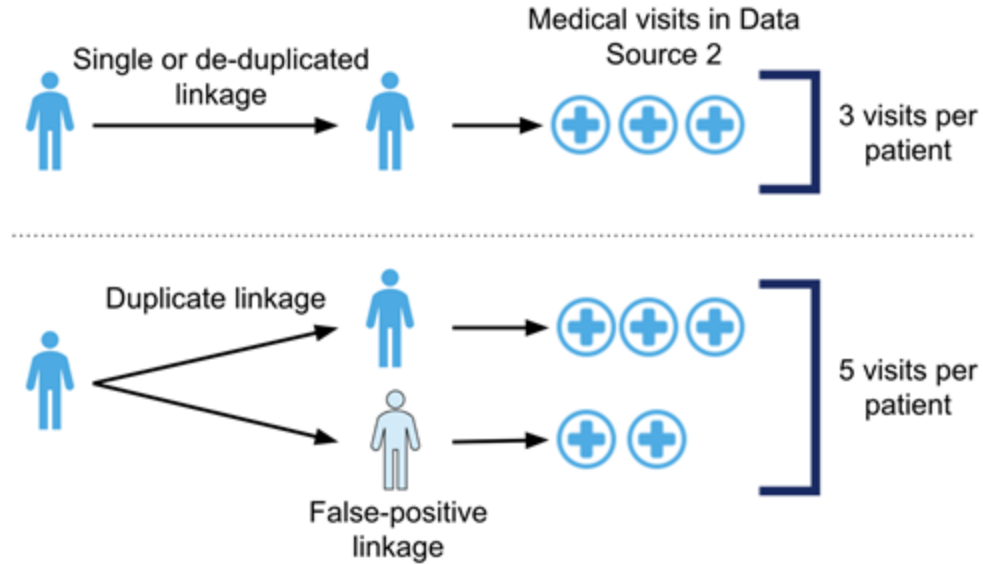
Tokenization Yields Duplicates

Example A: 2 unique individuals with the same identifiers

<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>id_number</u>
Jonathan Doe	01/01/2000	ABCDE	1234
Jonathan Doe	01/01/2000	ABCDE	5678



Duplicates Can Impact Results



Deduplication Strategies

Data Source 1





<u>mrn</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common dx code</u>	<u>rare dx code</u>
1234	Jonathan Doe	01/01/2000	ABCDE	True	True

Link on token

Link on token
+ common
condition

Link on token
+ rare
condition

Data Source 2



<u>id_num</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common condition</u>	<u>rare condition</u>
001	Jonathan Doe	01/01/2000	ABCDE	YES	NO
002	Jonathan Doe	01/01/2000	ABCDE	YES	YES



Deduplication Strategies

Data Source 1





<u>mrn</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common dx code</u>	<u>rare dx code</u>
1234	Jonathan Doe	01/01/2000	ABCDE	True	True

Link on token

Link on token
+ common
condition

Link on token
+ rare
condition

Data Source 2

<u>id_num</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common condition</u>	<u>rare condition</u>
001	Jonathan Doe	01/01/2000	ABCDE	YES	NO
002	Jonathan Doe	01/01/2000	ABCDE	YES	YES



Deduplication Strategies

Data Source 1





<u>mrn</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common dx code</u>	<u>rare dx code</u>
1234	Jonathan Doe	01/01/2000	ABCDE	True	True

Link on token

Link on token
+ common
condition

Link on token
+ rare
condition

Data Source 2



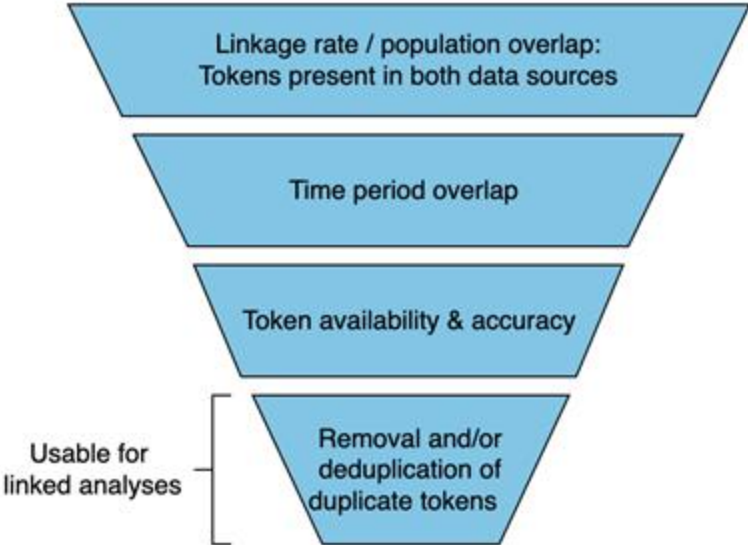
<u>id_num</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common condition</u>	<u>rare condition</u>
001	Jonathan Doe	01/01/2000	ABCDE	YES	NO
002	Jonathan Doe	01/01/2000	ABCDE	YES	YES



Conclusions



Sample Size and Feasibility



Questions?



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



jdusendang@graticule.life



[linkedin.com/in/jennifer-dusendang](https://www.linkedin.com/in/jennifer-dusendang)