

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)

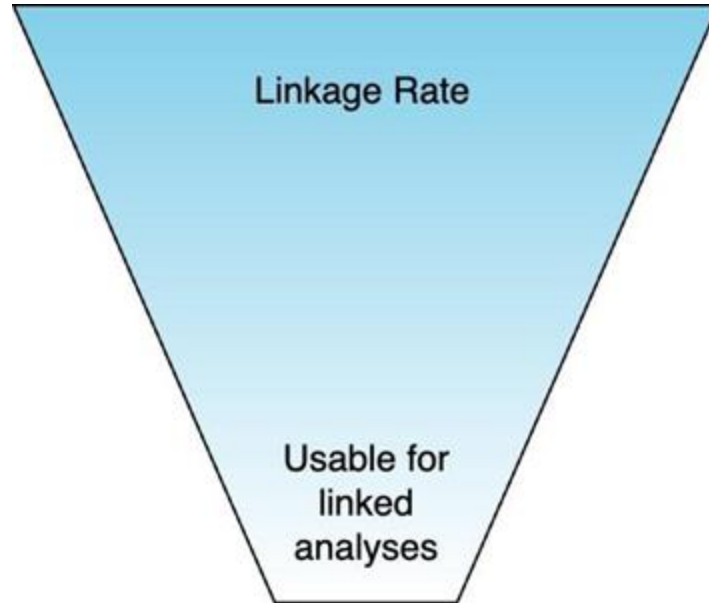


Background

- 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

A: Data source complementation



B: Data source supplementation



Types of Linkage Studies

A: Data source complementation



Different data elements not available in original data

B: Data source supplementation



Types of Linkage Studies

A: Data source complementation



B: Data source supplementation



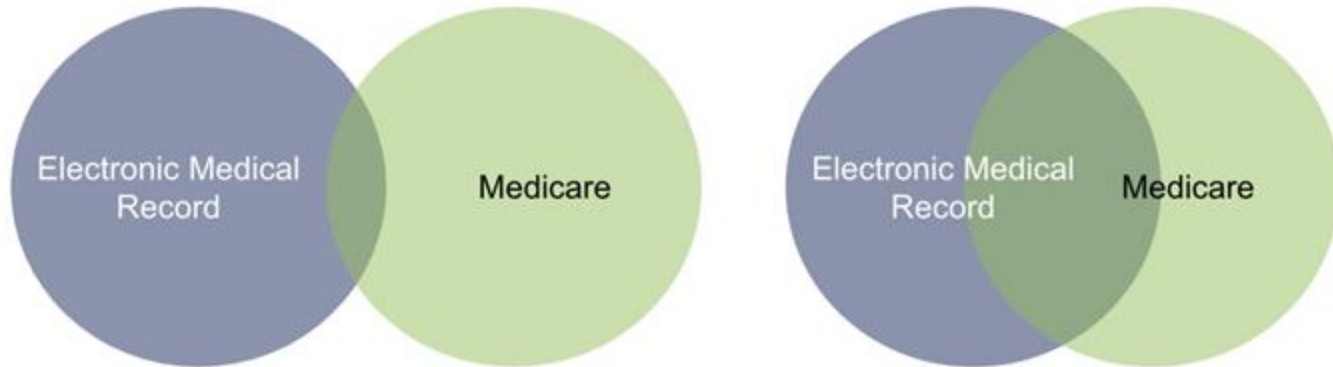
Same data
elements to
fill
in gaps



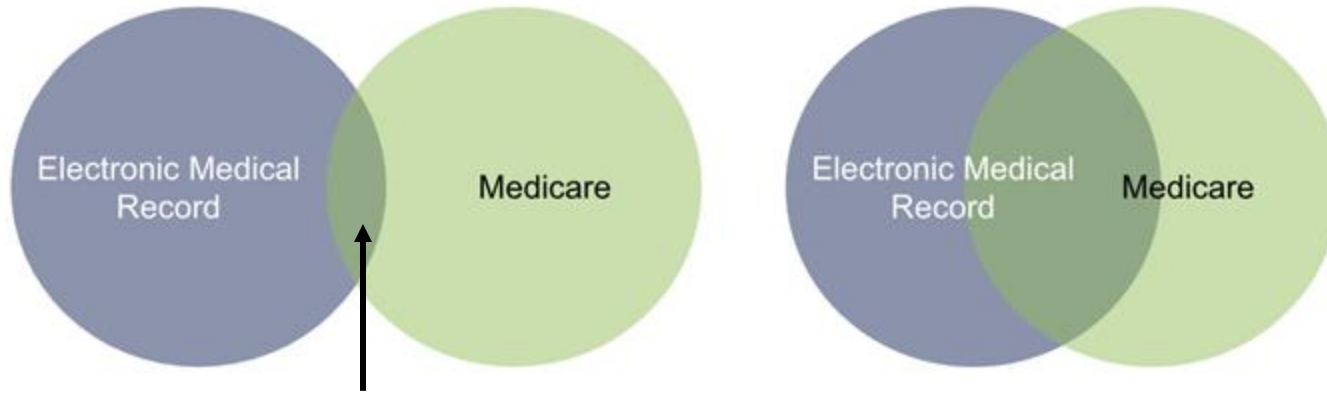
Multiple Dimensions of Attrition



Base Population Overlap



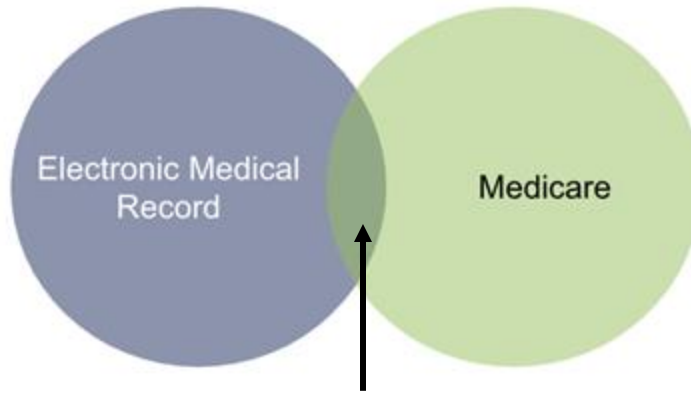
Base Population Overlap



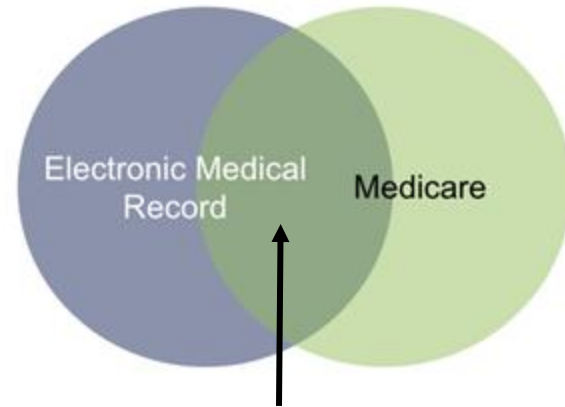
EMR cohort of younger patients:
Small linked cohort



Base Population Overlap



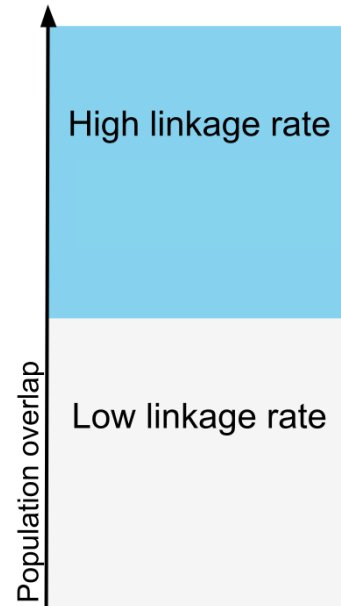
EMR cohort of younger patients:
Small linked cohort



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

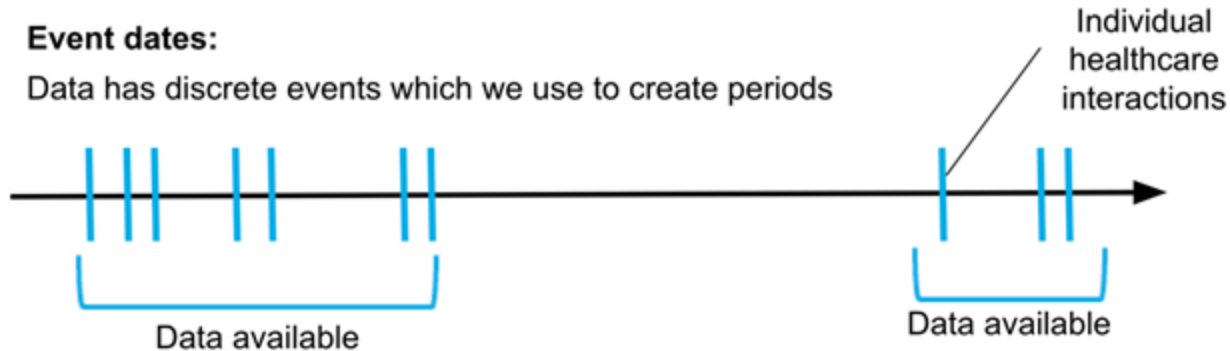
Enrollment spans:

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:

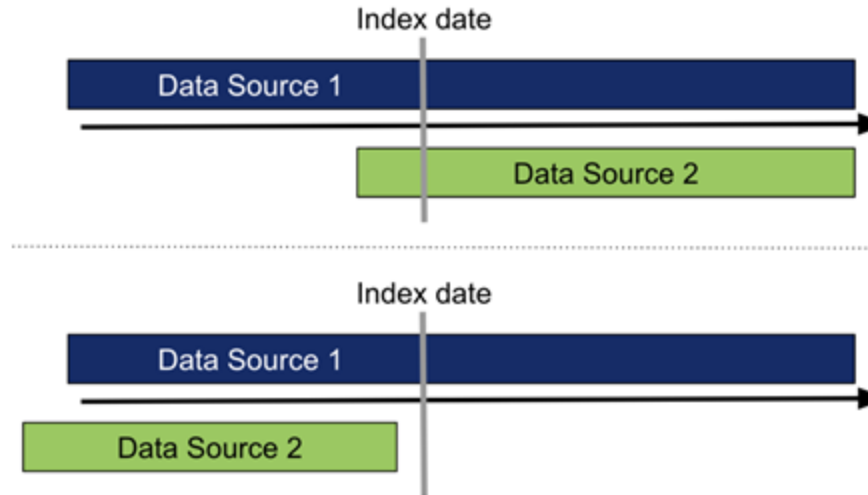
1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2



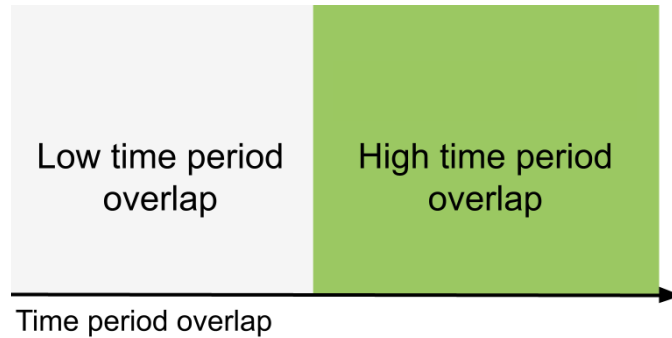
Time Period Overlap

Example study requirements:

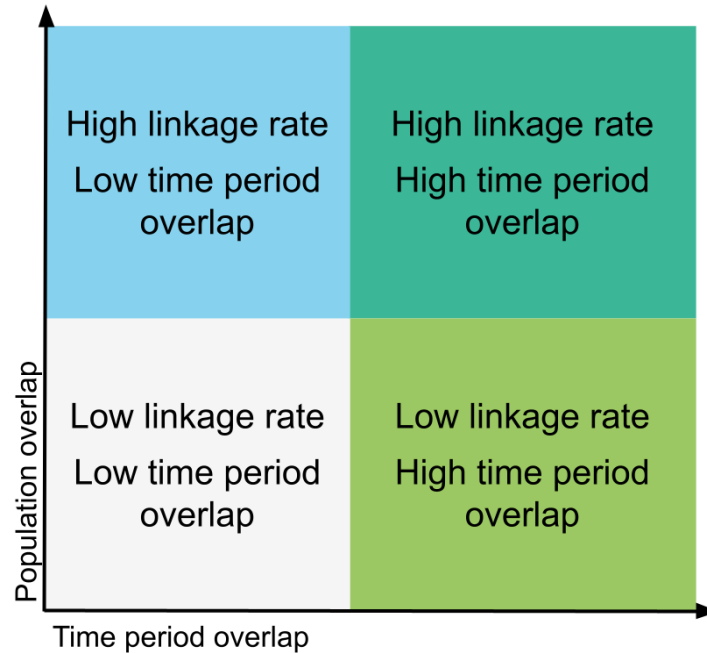
1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2



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?

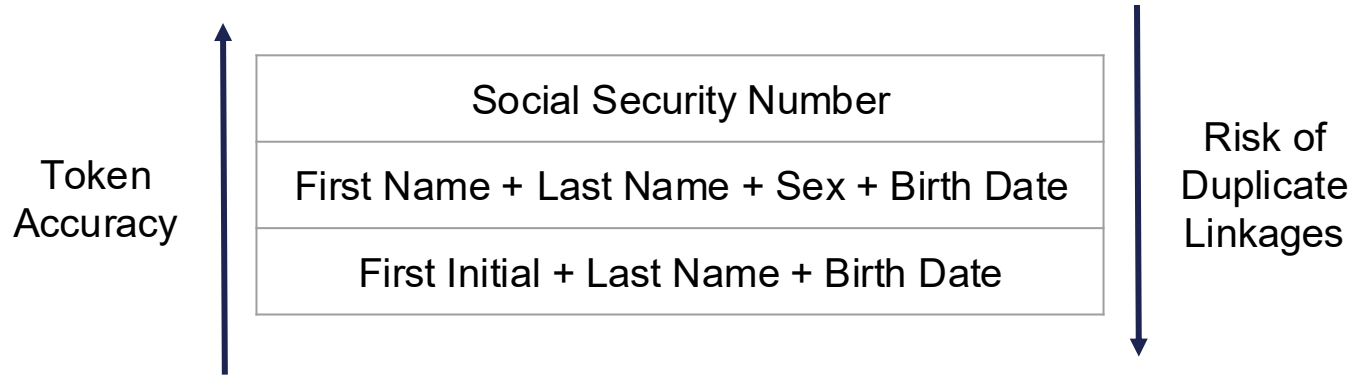


Linkage Considerations

- 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?

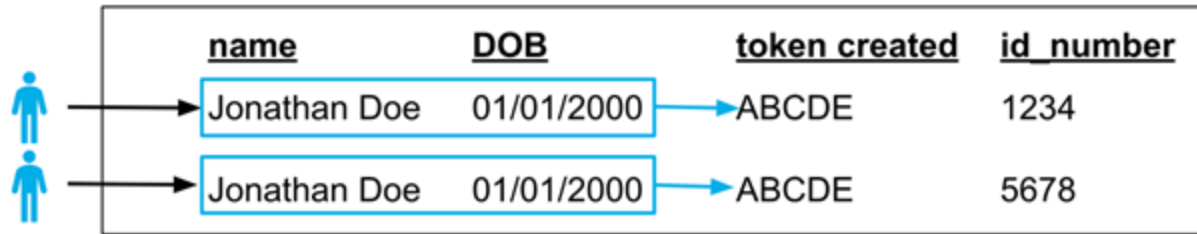


Accuracy vs Sample Size Tradeoffs



Tokenization Yields Duplicates

Example A: 2 unique individuals with the same identifiers



The diagram illustrates a data table with two rows of identical information. On the left, two blue stick figures represent individuals. Arrows point from each figure to a row in a table. The table has four columns: 'name', 'DOB', 'token created', and 'id_number'. The first row contains 'Jonathan Doe', '01/01/2000', 'ABCDE', and '1234'. The second row contains 'Jonathan Doe', '01/01/2000', 'ABCDE', and '5678'. The 'name' and 'DOB' cells in both rows are highlighted with blue borders, and blue arrows point from these cells to the 'token created' column.

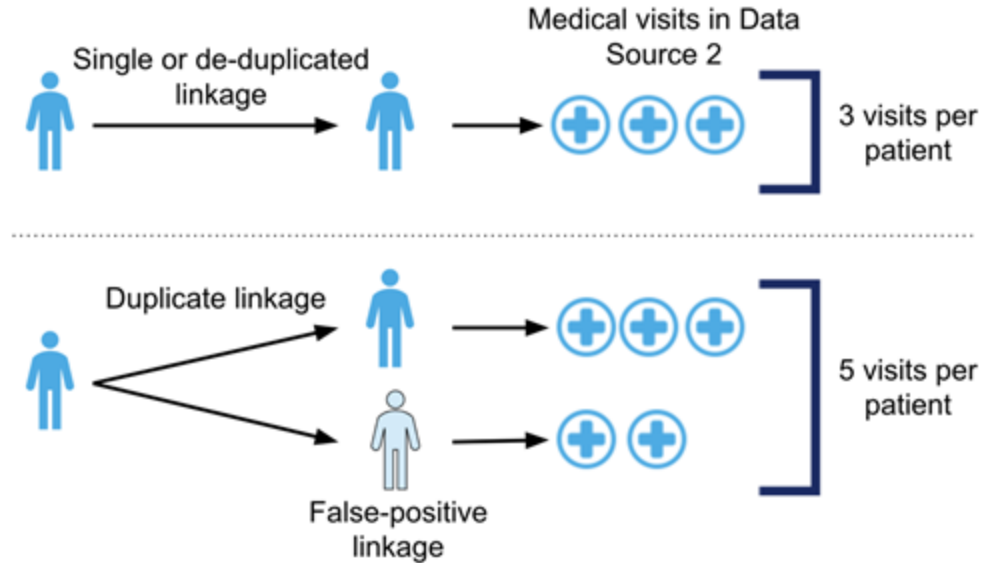
<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>id_number</u>
Jonathan Doe	01/01/2000	ABCDE	1234
Jonathan Doe	01/01/2000	ABCDE	5678



Duplicates Can Impact Results



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



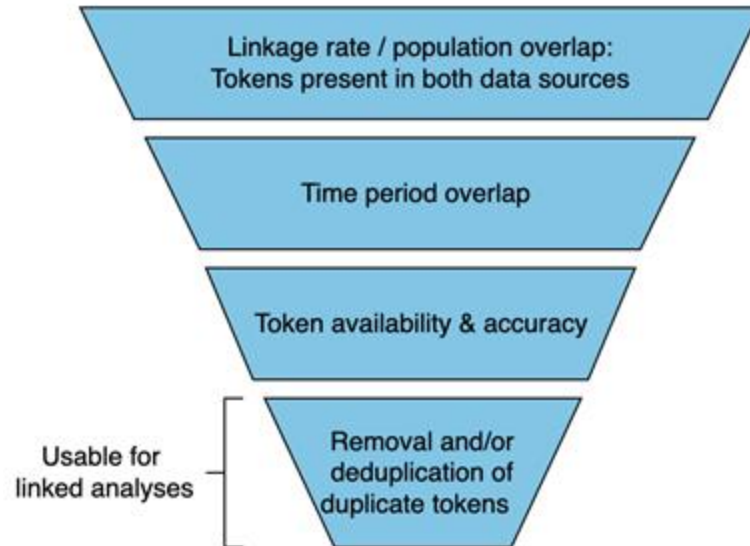
<u>id_num</u>	<u>name</u>	<u>DOB</u>	<u>token created</u>	<u>common condition</u>	<u>rare condition</u>
001	Jonathan Doe	01/01/2000	ABCDE	YES	NO
002	Jonathan Doe	01/01/2000	ABCDE	YES	YES



Conclusions



Sample Size and Feasibility



Questions?



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



ykoren@graticule.life