PharmaSUG 2025 - Paper RW-110

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

Jennifer Dusendang, MPH and Yuval Koren, MSc, Graticule Inc.

ABSTRACT

Linking data from disparate sources for scientific studies is highly valuable when using real-world data. A common example is linking electronic medical record (EMR) data to medical claims data or data from specialized providers outside the EMR system. Although creating privacy-preserving record linkage (PPRL) tokens is part of the linkage process, additional methods and considerations are necessary to produce a reliable and usable linked data set for scientific research.

Linkage rates provide an upper bound of data set patient coverage and usability for analyses. However, these are typically basic calculations of how often PPRL tokens match between two data sets, without regard to availability of key data elements, study period overlap, or issues with duplicate or low-specificity tokens. In particular, not identifying an appropriate timeframe in which linked data is applicable for a study can lead to unexpected decreases in sample size and limited study feasibility. Additionally, the study cohort that results from linking likely has different characteristics than the original, unlinked cohort.

To produce a linked data set that is appropriate for scientific studies, researchers and programmers should consider expected overlap of base populations within the data sets, reduction of linkage rates due to lack of data during relevant study periods and using stable patient characteristics to handle false-positive linked patients.

Although these concepts are applicable across programming languages, examples use SQL, Python, or R. This content is applicable for all skill levels.

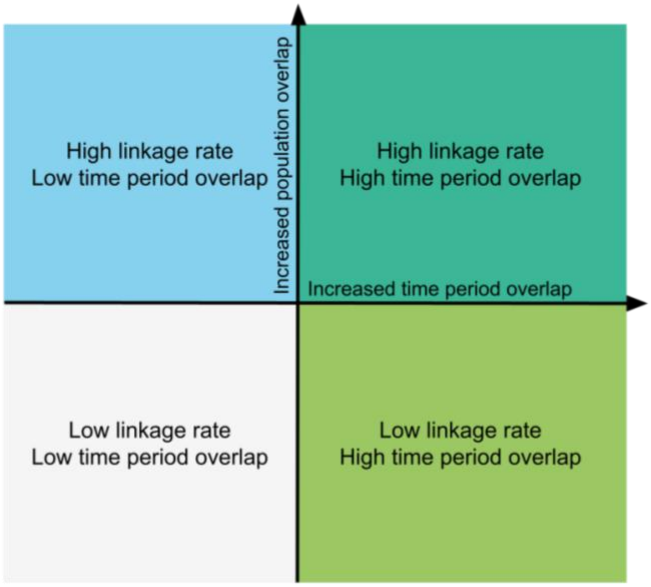
INTRODUCTION

In real-world studies, linked data sources (DS) are highly valuable because they can join disparate sources of data to derive novel conclusions. To produce linked data that is useful for scientific research, there are many details that must be considered, which have implications for sample size, accurate patient linkages, and usability for analyses.

Privacy-preserving record linkage (PPRL) allows data providers with access to protected health information (PHI) to create de-identified tokens, which are combinations of patient characteristics (e.g., name, date of birth, address, etc.) to share with other users of the data without sharing actual PHI.[1-3] Tokens can be matched across DS – therefore linking patients between DS.

Linkage proportions are often used to indicate the breadth of coverage of linked data and how usable data is for analysis. However, linkage proportions are simple calculations of how often a token from a primary data source (DS1) matches a token in a secondary data source (DS2) without regard to data availability, continuity, time period overlap, relevant study periods, or issues with duplicate or non-specific tokens (Figure 1).

Figure 1. Axes of Data Similarity Between Data Sources: Patient Populations and Time Period Overlap

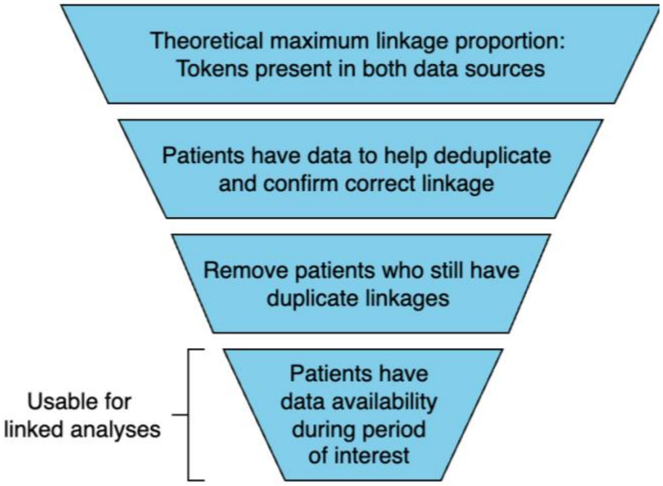


Time period overlap describes the concurrence of data between DS. This can be defined study-agnostically (e.g., both DS have data available between 2020 and 2022) or in a study-specific manner (e.g., patients in the cohort have data available at specific time periods that are relevant for analyses).

IMPACT OF LINKAGE ON STUDY SAMPLE SIZE AND FEASIBILITY

Each of the considerations discussed in this paper can significantly decrease the linked sample size of a study (Figure 2). Particularly if a study starts with a small population or involves subgroups or stratifications, the resulting linked sample size may be too small to perform the study. Although determining the magnitude of sample size reduction prior to conducting a study may not be feasible due to data not being readily available, it is important to understand that the initial linkage proportion of a DS is likely the maximum possible and that any steps to confirm the correct patients are linked at time periods relevant for a study will only lead to decreases in linkage proportions.

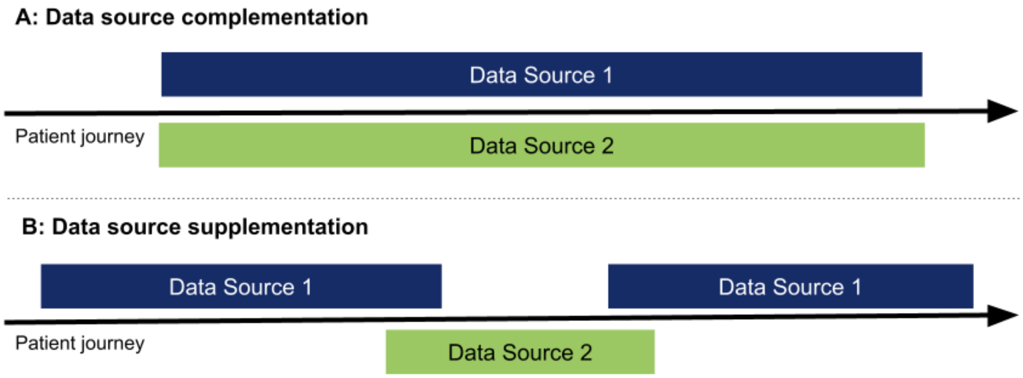
Figure 2. Hypothetical Attrition Diagram for Analyses of Linked Patients


Theoretical maximum linkage proportion: Tokens present in both data sources
Patients have data to help deduplicate and confirm correct linkage
Remove patients who still have duplicate linkages
Patients have data availability during period of interest
Usable for linked analyses

TYPES OF LINKAGE STUDIES

Linked healthcare data may be utilized in studies either to complement a base DS with other data elements (e.g., linking EMR data with claims data to describe the relationship between medical care and costs) (Figure 3A) or to supplement a base DS with the same data elements (e.g., linking two EMR systems to capture more complete patient journeys) (Figure 3B). In both scenarios, the considerations for linkage are similar although baseline expectations for linkage proportions and linked cohort size will vary.

Figure 3. Two Types of Linkage Studies: Data Source Complementation & Data Source Supplementation

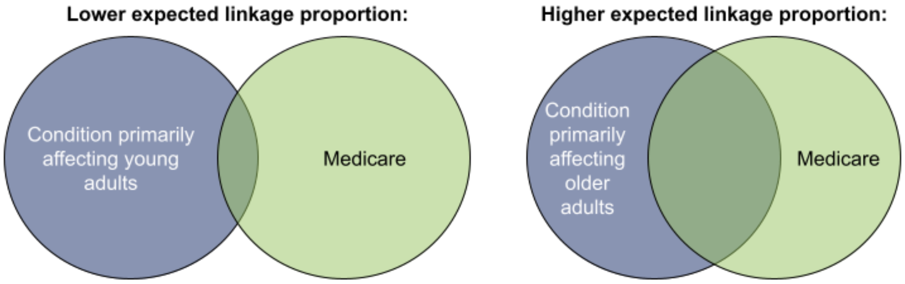


In the first example, the hope is that linkage proportions are high because analyses may not be feasible for patients who are not linked and therefore do not have complementary data for analysis. In addition, spans of direct time period overlap may be necessary to match specific periods of time needed for each study element. In the second example, linkage proportions will likely be low as linkage is only acting as a supplement in the analysis. We may not require or even expect overlapping time periods between sources as DS2 is filling in gaps where we don't see care in DS1.

BASE POPULATION OVERLAP

A seemingly obvious but potentially overlooked concept in linking DS is that there should be reasonable assurance that some patients occur in both sources. DS contain base populations that may vary in their geographic distribution, patient demographics, medical conditions, etc. For example, a study of a condition that occurs primarily in young adults will have low linkage proportions to a claims DS that primarily contains Medicare data (Figure 4). Separately, a study of a hospital system in California will have low linkage proportions to a hospital system in Washington.

Figure 4. Base Population Overlap & Expected Linkage Proportions


Lower expected linkage proportion:
Condition primarily affecting young adults
Medicare
Higher expected linkage proportion:
Condition primarily affecting older adults
Medicare

Even when patients come from the same base population and reasonable overlap can be expected, the cohort of patients who are successfully linked will likely differ in their characteristics from patients in the original unlinked cohort. For example, patients with a specific condition in an EMR may have different overall characteristics compared to the subset of those patients who are linked to a certain insurance provider. The cohort being analyzed for the linked analysis is no longer patients with the condition in EMR, but patients with the condition in EMR who also have specific insurance coverage. The characteristics of linked patients should be analyzed in addition to those of the original cohort if analyses are to be performed in both populations.

HIGH-CONFIDENCE PATIENT LINKAGE VERSUS SAMPLE SIZE PRESERVATION

Primary considerations for linking DS using PPRL tokens include:

- Which PPRL tokens are available (or can be produced) in both DS?
- How many patients have these tokens (or elements of tokens) available in each DS?
- How accurate are these tokens in identifying patients?

Studies will have different priorities for linkage accuracy and sample size, and the number of patients who share the same token increases with study sample size.[4] In studies with large cohorts, removing more patients due to potential false-positive linkages may be acceptable because the remaining sample size may be large enough for analyses. For studies with small cohorts, programmers may choose to prioritize retaining as many patients as possible, so the resulting linked DS has a large enough sample size to perform analyses. Additionally, programmers may prioritize ensuring high-confidence in linked patients if the risk of having false-positive linked patients is deemed detrimental to study results.

TOKEN PRECISION

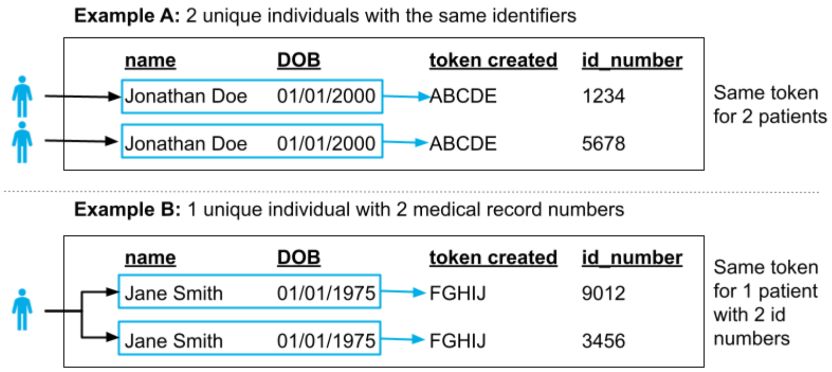
When designing the details of a linkage algorithm, there may be trade-offs between confidence in linkage and retaining sufficient sample sizes. PPRL tokens with less-specific identifiers (e.g., first name, gender) may be available in each DS for a higher proportion of patients than more personal and specific tokens, like those that use social security numbers. Linkage using more specific tokens minimizes false-positive and duplicate linkages, but this should be considered in the context of the proportion of patients with those tokens available.

Additionally, multiple tokens can be used to link patients, or tokens can be used via a waterfall method where additional tokens are only used when the primary token linkage is unsuccessful. For example, programmers may first attempt linkage using a highly specific token that contains social security number but is only available for a subset of the population and only use less specific tokens for patients who do not have the primary token.[2] Using more complex algorithms may lead to higher linked sample sizes but may be more difficult to implement and document.

WITHIN DATA SOURCE DUPLICATE TOKENS

Even within a single DS, multiple patients may share tokens. This can occur when tokens are not specific (e.g., for some token algorithms, patients with the same name and date of birth (DOB) will have the same token) (Figure 5, Example A) or because the same patient has two identifiers (e.g., medical record number, ID number) in the DS (Figure 5, Example B). If it is expected that individual patients could have multiple identifiers in the DS and tokens are specific enough or additional data or demographic information can be used to confirm that multiple identifiers belong to the same patient, programmers may consider merging multiple identifiers into one record before performing linkage to DS2.

Figure 5. Within Data Source Duplicate Tokens



Example A: 2 unique individuals with the same identifiers

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

Same token for 2 patients

Example B: 1 unique individual with 2 medical record numbers

name	DOB	token created	id_number
Jane Smith	01/01/1975	FGHIJ	9012
Jane Smith	01/01/1975	FGHIJ	3456

Same token for 1 patient with 2 id numbers

It is likely not necessary to de-duplicate patients with the same token in an entire data source. Distinct patients who share the same token are unlikely to both have the same condition of interest and meet all inclusion and exclusion criteria for the study. Programmers can likely assess duplicate tokens and potential false-positive matches after the unlinked cohort is created, although this will depend on tolerance for false-positives and prevalence of the condition being studied.

If two patients in the study cohort share tokens, programmers can consider whether investigation into additional data points would help to distinguish patients or if all patients who share tokens should be removed from linked analyses. The challenge with retaining patients who share tokens is that when linked to DS2, additional steps need to be taken to determine which patient in DS1 the linked data belongs to.

In Program 1, we join a cohort to a list of tokens on the patient identifier (ID) variable. We then count how many distinct IDs exist for each token and filter to patients in the cohort who have a count of 1, indicating that they have a token, but do not share that token with anyone else.

Program 1. Removing Patients with the Same Token (SQL code)

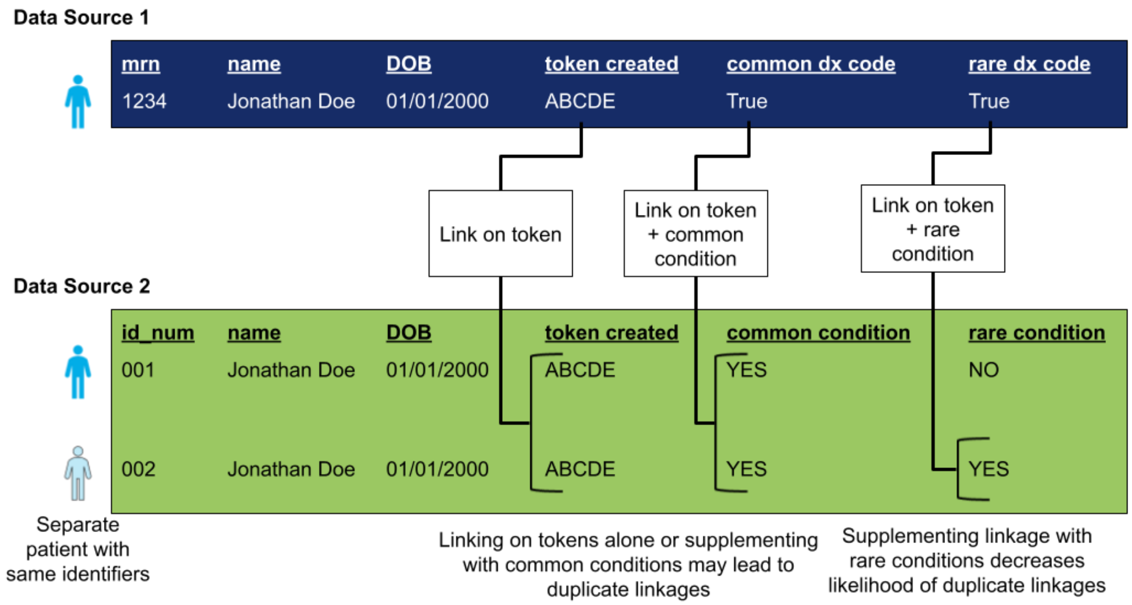
```
SELECT cohort.*,
    CASE
        WHEN token_dataset.token IS NULL THEN 0
        ELSE COUNT(DISTINCT cohort.ID) OVER (PARTITION BY token_dataset.token)
        END AS count_id_per_token
FROM cohort
LEFT JOIN token_dataset
    ON cohort.ID = token_dataset.ID
WHERE count_id_per_token = 1
```

BETWEEN DATA SOURCE DUPLICATE TOKENS

Duplicate linkages can also occur when a single token in DS1 links to multiple identifiers in DS2. Again, this could be an issue with a single patient having multiple IDs in DS2 or could be the result of non-specific tokens, therefore leading to a false-positive linkage. Similar steps should be taken as above in ‘Within data source duplicates’ to de-duplicate records in DS2.

If a single token in DS1 links to multiple IDs in DS2, we can reduce the number of duplicates by ensuring that the ID in DS2 aligns with data available in DS1. For example, a cohort of patients with a condition could be linked on tokens and indication of the condition in DS2 (Figure 6). This does not remove the possibility that both patients with the same token could have the same condition, however, it does reduce the number of patients excluded for having duplicate linkages. Using the confirmatory method above is also dependent on the data available in both DS and the frequency of the indicator used to de-duplicate. A prevalent condition such as high blood pressure may not be as successful in de-duplicating linkages as a rarer condition that would be less likely to occur in both duplicate patients.

Figure 6. Supplementing Tokens with Other Data to Decrease Duplicate Linked Patients



Data Source 1

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

Data Source 2

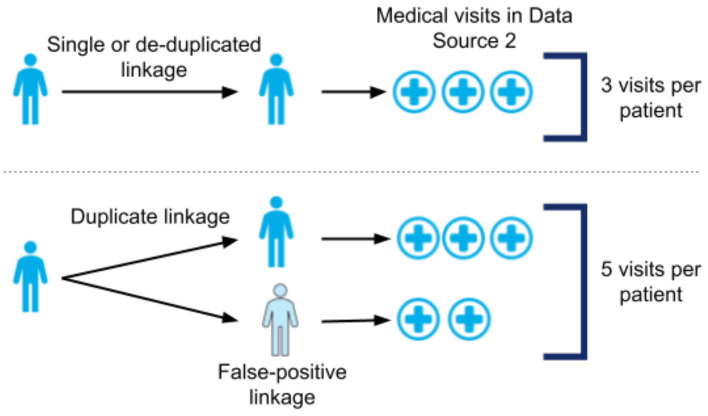
id_num	name	DOB	token created	common condition	rare condition
001	Jonathan Doe	01/01/2000	ABCDE	YES	NO
002	Jonathan Doe	01/01/2000	ABCDE	YES	YES

Link on token
Link on token + common condition
Link on token + rare condition
Separate patient with same identifiers
Linking on tokens alone or supplementing with common conditions may lead to duplicate linkages
Supplementing linkage with rare conditions decreases likelihood of duplicate linkages

RISKS OF INCLUDING DUPLICATE TOKENS OR FALSE-POSITIVE LINKAGES

If de-duplication is not performed and both identifiers from DS2 remain in analyses, there is a risk of overestimating events in DS2. For example, if the goal of linking DS1 to DS2 is to estimate visit counts per patient, keeping multiple linked patient IDs in DS2 linked to a single DS1 ID can lead to visits from multiple patients being summed as visits of a single patient – potentially doubling the visits per patient (Figure 7).

Figure 7. Duplicate Linkages Can Impact Magnitude of Results

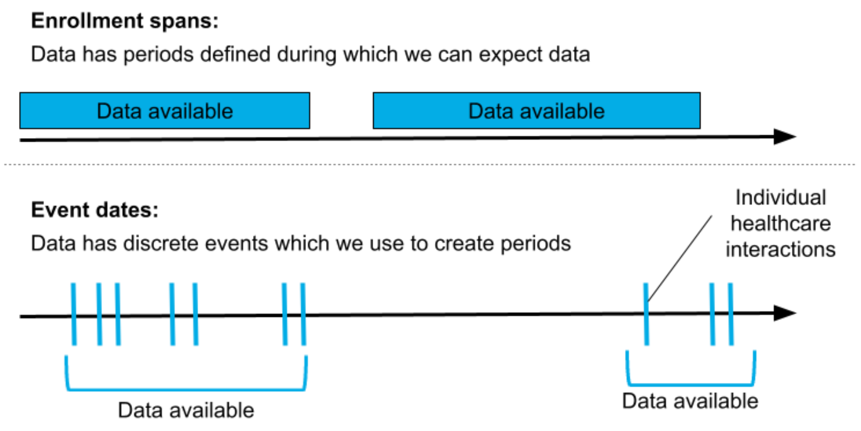


IMPORTANCE OF DATA TEMPORALITY IN LINKED DATA

DATA CONTINUITY PERIODS

To determine when data is available for each patient, we must consider how to define data continuity periods. Some DS include variables indicating periods of data continuity (e.g., enrollment periods), while other DS only contain discrete events such as visits or encounters (Figure 8). In the latter case, creating a definition of data continuity periods is useful to define patient-level periods of overlap. For example, we may expect at least 1 event (e.g., visit, lab test, etc.) to occur every 365 days, and if a patient does not meet this criterion, they are considered to not have data during at that time. These criteria will vary depending on the frequency of care we expect patients to receive given the DS and condition of interest.

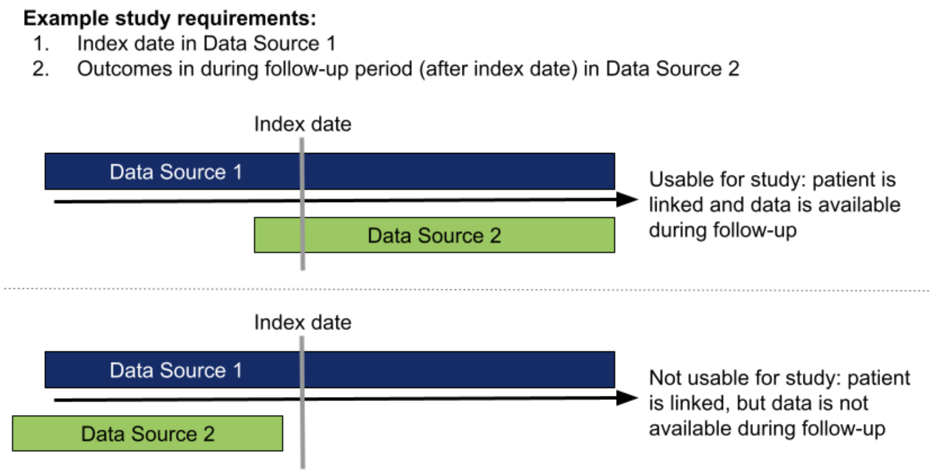
Figure 8. Determining Data Continuity with Enrollment Spans or Event Dates


Enrollment spans:
Data has periods defined during which we can expect data
Data available
Data available
Event dates:
Data has discrete events which we use to create periods
Individual healthcare interactions
Data available
Data available

TIME PERIOD OVERLAP

A critical consideration for linked study feasibility is determining the periods of time period overlap that are required in each DS for each patient. Patients may be linked, but if data is not available during the study period that is relevant for the research questions, then the linked patients are unusable for the analysis. For example, if it is important to have data from DS1 during a baseline period prior to the index date, and data from DS2 during the follow-up period, a patient that only has data in DS2 prior to the index date will be unusable for analyses, even though they are linked (Figure 9).

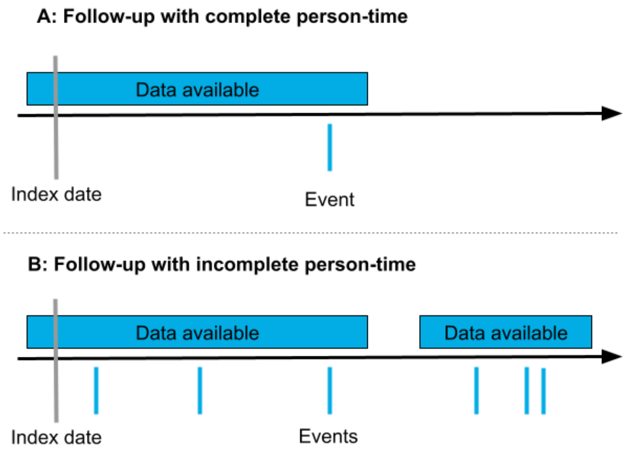
Figure 9. Study Diagram Illustrating Timing of Linked Data


Example study requirements:
1. Index date in Data Source 1
2. Outcomes in during follow-up period (after index date) in Data Source 2
Index date
Data Source 1
Data Source 2
Usable for study: patient is linked and data is available during follow-up
Index date
Data Source 1
Data Source 2
Not usable for study: patient is linked, but data is not available during follow-up

PERSON-TIME CONTINUITY

A study may require that patients have data available in one or more DS during a specific time-point or more flexible criteria may be applied depending on the research question and importance of sample size retention (Figure 10). For example, a study implementing survival analysis may require that patients have data available in a complementary DS at their index date and be followed until event or censoring, enabling complete person-time to be captured during follow-up. Alternatively, data from a complementary DS may only be required at some point in the year after index date to calculate rates of events during enrollment periods during follow-up.

Figure 10. Person-Time Continuity


A: Follow-up with complete person-time
Data available
Index date
Event
B: Follow-up with incomplete person-time
Data available
Data available
Index date
Events

CONCLUSION

Linking real-world DS is valuable for deriving novel cohorts and studying patients across complex healthcare systems. However, many programmers do not consider details of the linking process which are essential to produce usable linked data. These details will vary depending on the study design, DS, and research goals and can dramatically affect the resulting sample size, confidence in patient linkage, and feasibility of analyses.

To produce linked data that are appropriate for scientific studies, programmers should consider expected overlap of base populations, reduction of linkage proportions due to lack of data during relevant study periods, and the use of stable patient characteristics to handle false-positive linked patients.

REFERENCES

1. Pathak A, Serrer L, Zapata D, King R, Mirel LB, Sukalac T, Srinivasan A, Baier P, Bhalla M, David-Ferdon C, Luxenberg S, Gundlapalli, AV. Privacy preserving record linkage for public health action: opportunities and challenges. Journal of the American Medical Informatics Association. 2024;31(11). https://academic.oup.com/jamia/article-abstract/31/11/2605/7720510
2. Frederick National Laboratory for Cancer Research. Evaluating the performance of privacy preserving record linkage systems (PPRLS). 2023. https://surveillance.cancer.gov/reports/TO-P2-PPRLS-Evaluation-Report.pdf
3. Tachinardi U, Grannis SJ, Michael SG, Misquitta L, Dahlin J, Sheikh U, Kho A, Phua J, Rogovin SS, Amor B, Choudhury M, Sparks P, Mannaa A, Ljazouli S, Saltz J, Prior F, Baghal A, Gersing K, Embi PJ. Privacy-preserving record linkage across disparate institutions and datasets to enable a learning health system: The national COVID cohort collaborative (N3C) experience. Learning Health Systems. 2024;8(1). https://onlinelibrary.wiley.com/doi/10.1002/lrh2.10404
4. Huynh S, Liu T, Leshin J, Haskell T. Assessment of the Relationship Between Collision Rate and Sample Size Using a Large US Mortality Dataset. ISPOR Europe. 2021. https://www.ispor.org/docs/default-source/euro2021/posb321stephaniehuynhposter-pdf.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jenny Dusendang
Graticule Inc.
jdusendang@graticule.life