Data Analysis after Record Linkage: Sources of Error, Consequences, and Possible Solutions

Martin Slawski

jointly with Brady West[⋆] and Emanuel Ben-David[†]

George Mason University

[⋆]University of Michigan [†]CSRM, U.S. Census Bureau

May 12, 2022

AAPOR Annual Conference

Session Title: "Are you willing to share your life? Data Linkage & Consent"

Themes of this Presentation

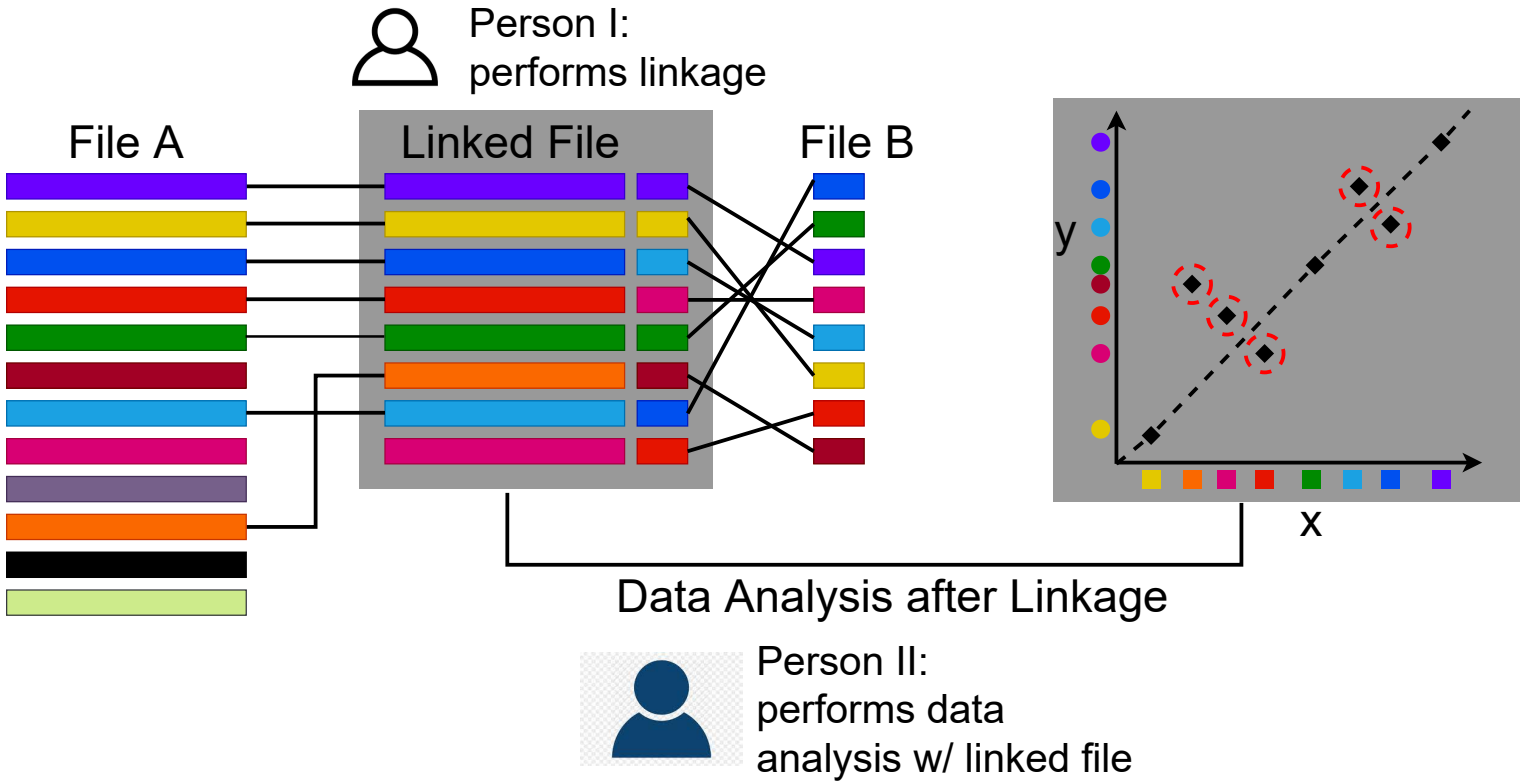
This presentation is NOT (primarily) about

- Obtaining consent to link,
- Collecting sensitive attributes from survey respondents.

Instead, our research considers data analysis **after** record linkage (**RL**) when

- linkage is associated with substantial uncertainty given the lack of unique or "high-quality" identifiers,
- data analysis and data linkage are performed separately by different individuals; the data analyst(s) may not know how data were linked (**secondary analysis**),
- Errors in linkage (**mismatches** and **missed matches**) affect quality of downstream analyses.

Overview



- "Linkage": "Probabilistic Linkage" associated with **uncertainty** about matching records
- Person II (data analyst) may know nothing about how files were linked (**secondary analysis** setting).

Probabilistic RL

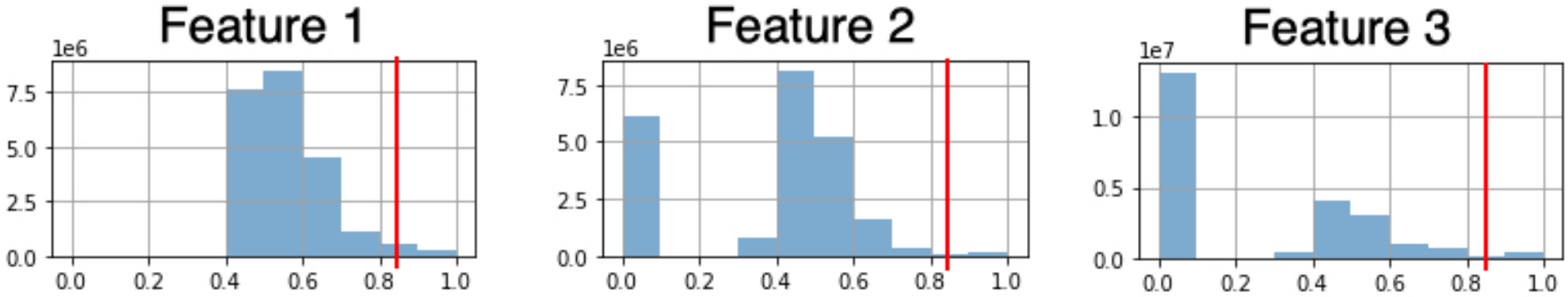
Which records belong to the same individual?

f.name	m.name	l.name	m.o.b	lives in
Emanuel	Hyatt	Bendavid	Mar	New York, NY
Emmanuel	Ben	David	Dec	Washington, DC
Emanuel	NA	Ben-Dawid	Nov	Stanford, CA
Emanuel	NA	Ben-David	Mar	Ashland, OR
E.	NA	Ben-Davit	Nov	San Diego, CA

Common practice in (probabilistic) RL is to consider pairs of records (a, b) with a from File A and b from File B and assign a score $\gamma(a, b)$.

Probabilistic RL (c'ted)

Example: Scores based on three features (Last, First, Middle) names.



The resulting scores will be aggregated and converted (thresholded) into a binary decision: **match** or **non-match**.

Potential errors:

- False matches **(mismatches, mismatch error)**,
- False non-matches **(missed matches)**.

Principled choice of the threshold frequently challenging given lack of training data (e.g., based on clerical review).

Balacing linkage and mismatch rates

Option 1 – Confining data analysis to "safe correct matches":

- Linkage rate $\downarrow$
- Danger of Selection Bias $\uparrow$
- Statistical Power $\downarrow$

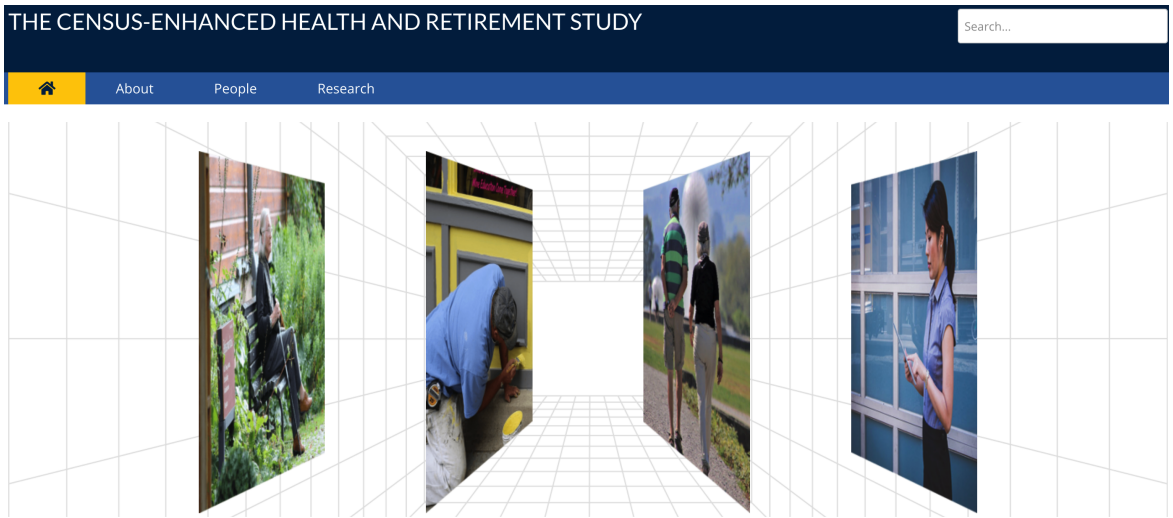
Option 2 – Maximizing #matches:

- Danger of Selection Bias $\downarrow$
- Danger of Data Contamination $\uparrow$

A suitable approach involves a balance between the two extremes. **Post-hoc adjustment** for data contamination may allow for more aggressive selection of potential matches.

Example: Linkage of the HRS and the Census BR

In a recent study, Abowd et al. (2022) consider linkage of the Health & Retirement Study (HRS) and the U.S. Census Business register (BR) to enrich information from the HRS.



Abowd et al. (2022) re-analyze the relationship between establishment size and hourly wages after inferring the establishment size from the BR.

Example: Linkage of the HRS and the Census BR

Linkage is considerably hampered by the fact that only 70% of HRS respondents consent to SSA linkage, and hence lack an Employer Identification Number (EIN) – linkage based on establishment addresses may yield of thousands of potential matches in the BR.

Despite considerable sophistication in terms of RL strategy, the subsets of linked and unlinked respondents are still markedly different:

	linked (92%)	**unlinked** (8%)
∅Age	57.6	56.9
%White	68	57
%Black	22	24
%Hispanic	14	26
%Native born	87	69
$Earnings	43.2k	33.3k

Source: based on from a recent presentation by Dhiren Patki in the ISR record linkage series U Michigan.

Illustration: Accounting for mismatch error

Research question: how long does it take for nurses to receive a regular nurse license after being on a temporary license?

astName	FirstName	MiddleName	CredentialType	Status	BirthYear	CEDueDate	FirstIssueDate	LastIssueDate	Expiration
aldonado	Ashley	Elizabeth	Medical Assistant Certification	ACTIVE	1991.0	NaN	20141107.0	20211001.0	20231(
Strange	Danielle	Nicole	Registered Nurse Temporary Practice Permit	SUPERSEDED	1981.0	NaN	20191212.0	20191212.0	20200
Mahar	Sean	Patrick	Medical Assistant Registration	ACTIVE	1994.0	NaN	20210512.0	20211001.0	20231(
Namru	Lobsang	Tsering	Registered Nurse Temporary Practice Permit	EXPIRED	1968.0	NaN	20210401.0	20210401.0	20211(
...	...	...	...	...	...	...	...	...	
Wood	Kyle	Alan	Registered Nurse License	ACTIVE	1995.0	20230917.0	20190119.0	20210715.0	20220

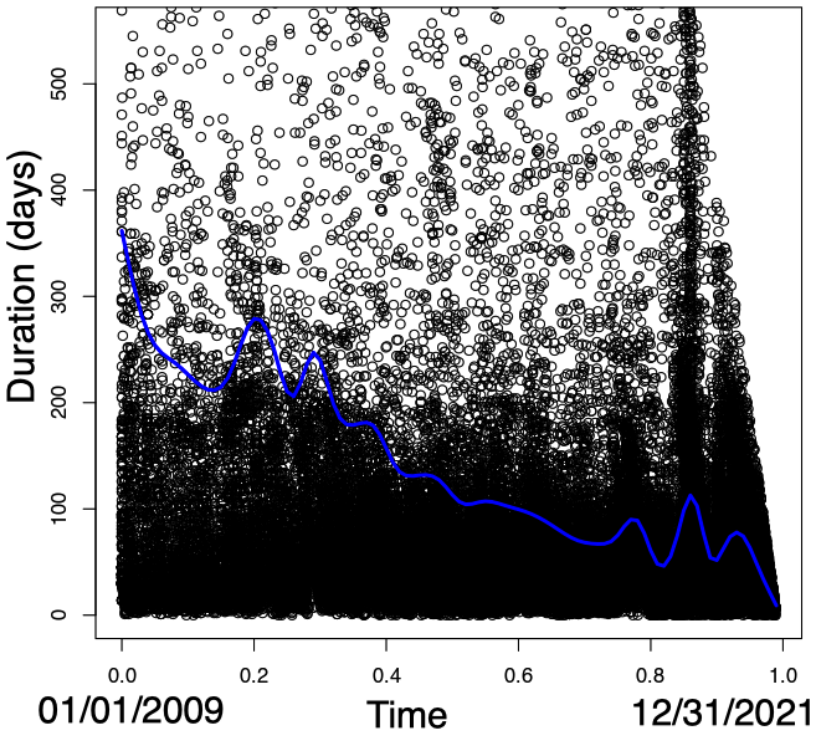
- File A: contains all records with temporary practice permit
- File B: contains all records with regular nurse license
- Link File A on B: first block on DOB, last initial, then use probabilistic linkage based on names.

Acknowledgment: Abie Flaxman, University of Washington

generously linked

$n = 78,088$

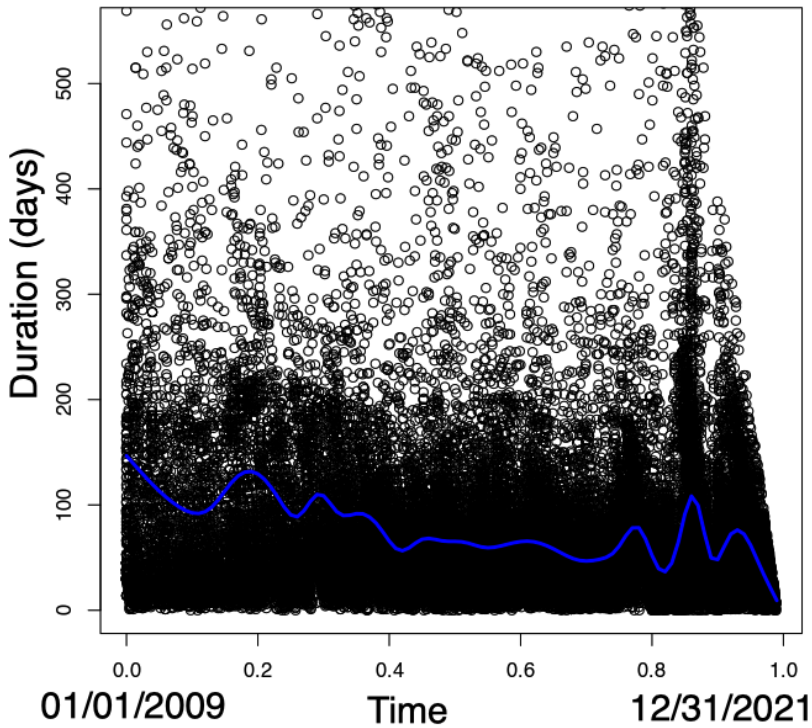
mismatch rate $\geq$7.5%



restrictively linked

$n = 60,842$

mismatch rate $\geq$1.4%



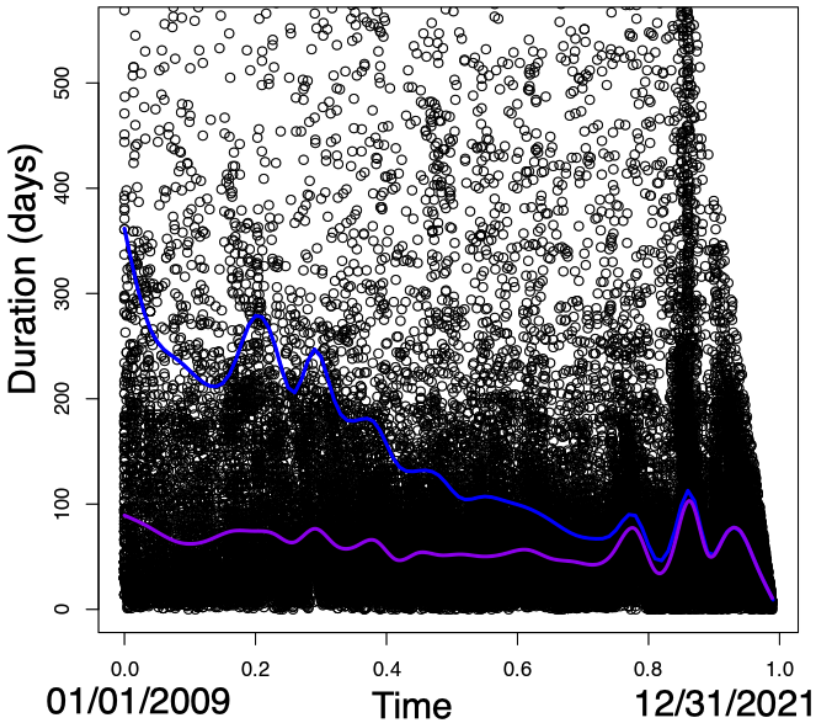
estimated time trend (mean function)

(each after removing apparent mismatches with negative durations)

Corrective, mismatch-aware estimation

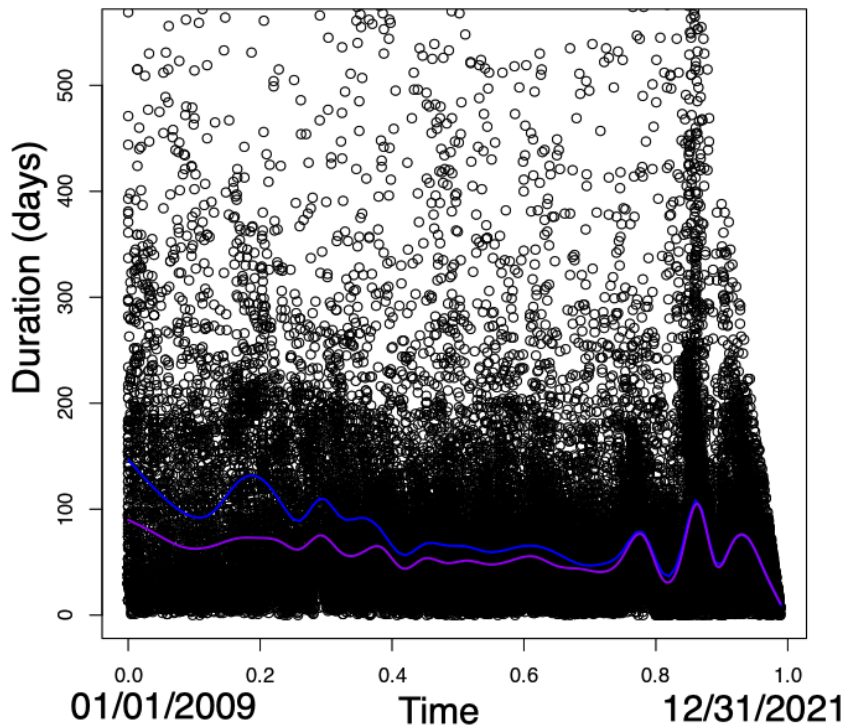
Based on the mixture model proposed in Slawski et al. (2021).

generously linked



Estimated mismatch rate: 7.2%

restrictively linked



Estimated mismatch rate: 4.6%

estimated trend

estimated trend (corrected)

Conclusive remarks and prospective work

- Lack of consent to link and/or lack of access to sensitive variables may affect probabilistic RL
- Advances in statistical methodology & computation are underway to address the resulting challenge, ensuring high linkage rates & quality of post-linkage analysis

As part of an ongoing NSF project*, our team is still actively looking for challenging data linkages for validation purposes.

Thanks for your time & attention !

* The first two authors acknowledge support from NSF grant #2120318.