

Data Linkage Methods and Challenges

ISCTM 13th Annual Scientific Meeting – February 22, 2017

Andrew Kress
CEO
HealthVerity, Inc.



Practical Considerations > Accessing RWD

Specific applications regarding the utility of RWD covered elsewhere...

1. To link to RWD, need a method for locating and linking accurately to the relevant records

AND- need to do this in a privacy-protecting, HIPAA de-identified manner

2. Need to decide which data to use, keeping in mind research limitations imposed by sample selection. There are two main use cases:
 - Starting with a target dataset (e.g. a registry), and looking to extend it to include additional data
 - Doing a study solely in 3rd party data sources

Practical Considerations >

Poor Patient Data Hygiene in Source Data

Let's say you just wanted to use a single database as a target for RWD acquisition.

- Master patient entry duplicate rates within a single entity range from below 8% to significantly higher
- Match rates using various approaches vary based on number of factors - Kaiser Permanente reported a match rate >90% within each EMR instance; rate fell to around 50-60% when sharing between regions using separate instances.¹

The broader the number of sources, the less control on data entry standards and quality.

- Different sources permit or require different inputs, e.g. SS# is often not available

And, for privacy preservation, this may need to be done de-identified (often driven by rules of target data source)

1. PATIENT IDENTIFICATION AND MATCHING FINAL REPORT, Office of the National Coordinator for Health Information Technology, 2014

De-Identification of PHI

Protected health information includes many common identifiers (e.g., name, address, birth date, Social Security Number), plus potential “quasi-identifiers” which may permit a recipient of the data to determine the identity of the corresponding subject.

Most de-identification methods rely on some form of

- Hashing algorithms to transform/conceal PHI fields
- “Expert Determination” of results (vs “Safe Harbor”)

“Expert determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and documents the methods and results of the analysis”

Challenges of PHI Data Linkage

First Name	Last Name	Age	3-digit Zip	Phone Number
William	Jones	45	190	555-1234
Bill	Jones	?	086	555-7234

Nicknames & Synonyms

Names and places may be written or abbreviated in different ways.

Common Values

Not all values carry the same amount of information. Common names or values are less identifying, and can lead to confusion.

Missing Values

Some fields may be missing or not recorded by a site. Matching with different amounts of evidence is not easy.

Changing Data

About 12% of families move each year. A million women get married and change their name. Tracking patients across changes is very hard.

Typos

Small errors – from typos, OCR errors, or phonetic similarity – are ubiquitous, and can affect up to 5% of records at some sites.

Deterministic Approaches

Deterministic approaches try to avoid the problem and just push through

- All fields are encrypted with a simple hash function
- Number of fields kept small – Name, DOB, gender, zip code
- Values truncated to avoid errors and changes – first initial, 3-digit zip
- Exact matches, or simple rules of thumb

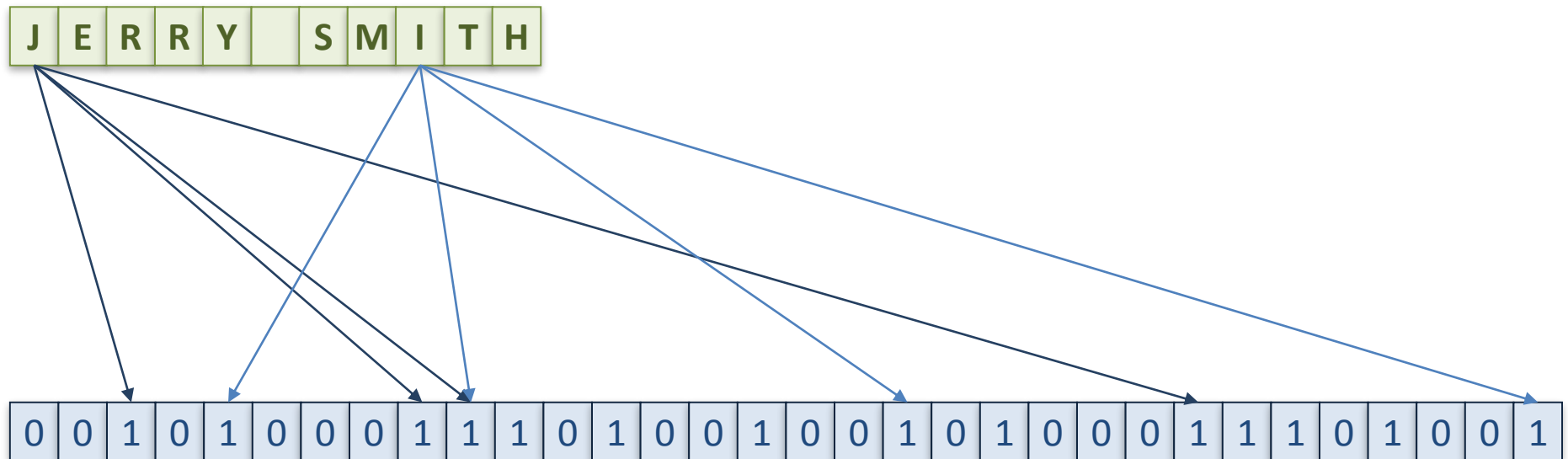
Results are predictably weak – and get worse with more data sets

- For each data source:
- **1 in 26** missed patient matches – rises to **1 in 8** when including patients that moved (FN)
- **1 in 200** patients matched incorrectly (FP)

Creates longitudinal data sets that are sparse, noisy, and heavily biased

For Comparison: Using Bloom Filters for Matching

A **Bloom Filter** is just a different way of hashing a name to random sequence of 0s and 1s

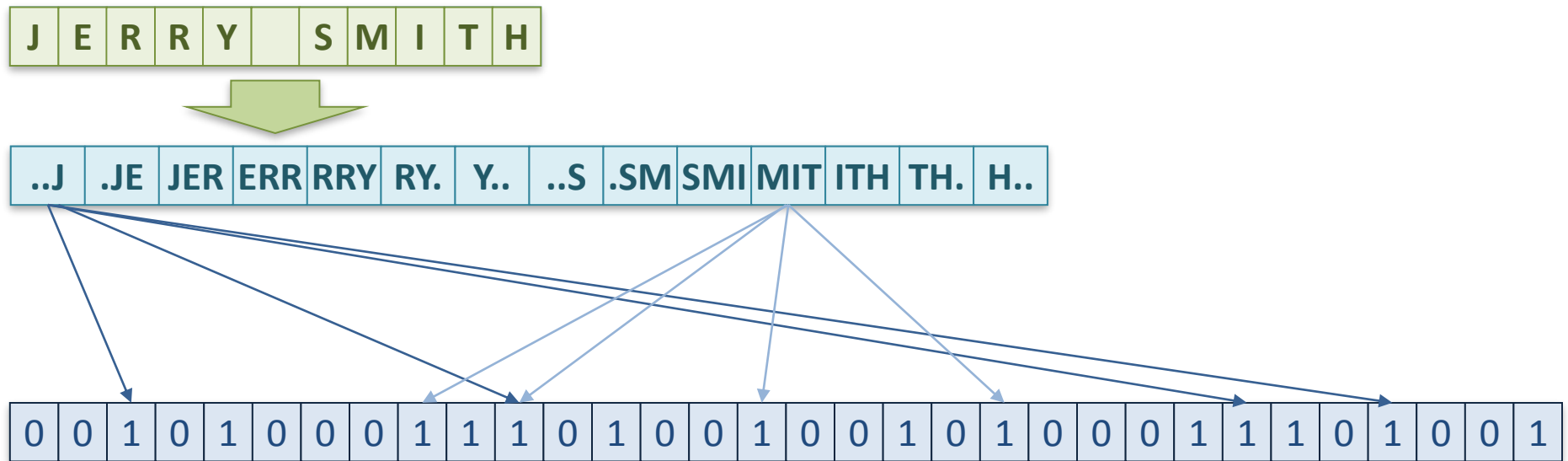


Simple Implementation:

- Each letter in the name maps to 4 different bits in the sequence, turning them to a 1.
- Which bits are used is determined by cryptographic key- the encryption is still safe
- Changing one letter means that no more than 8 bits are different.
- Similar names have similar bit patterns – small errors can still be matched with confidence.

Trigrams for Bloom Filters

We use **Trigrams** – 3 letter combinations – instead of single letters to preserve ordering



Benefits of Trigrams:

- Relative ordering is preserved – the fact that **M** is between an **S** and an **I** is important
- Absolute ordering is not critical – a missing letter does not affect all of the later letters
- Trigrams are more secure than single letters – frequency-based attacks are useless

Missing or Changing Data: Probabilistic Matching

- Probabilistic Matching deals with missing or changing data
- Each field contributes a probability in the observed data
 - ◎ There's a probability that the values match – based on moving, frequency of value, etc.
 - ◎ There's a different non-zero probability that the value has changed
 - ◎ Those different values can be combined – using Bayesian probability – to find the default probability to use when a value is missing



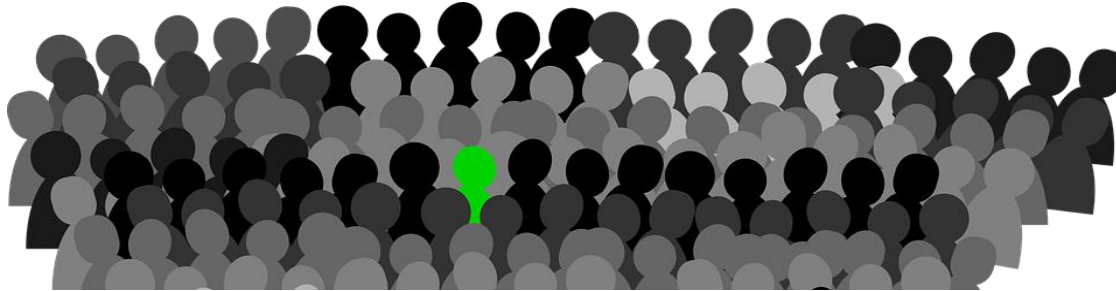
$\text{Prob}(\text{zip code}_{\text{new}} = \text{Jacksonville} \mid \text{zip code}_{\text{old}} = \text{Atlanta}) = 0.2\%$

$\text{Prob}(\text{zip code}_{\text{new}} = \text{Atlanta} \mid \text{zip code}_{\text{old}} = \text{Atlanta}) = 85\%$

$\text{Prob}(\text{zip code}_{\text{new}} = \text{???} \mid \text{zip code}_{\text{old}} = \text{Atlanta}) = 2.1\%$

- Each field contributes a probability based on whether it matches or not
- Now we can use all of the extra fields and link patients that have moved

Common Values: Frequency Normalization



Certain names and places are much more common than others

- ⦿ Consider James Smith, a 28 year-old living in Manhattan
- ⦿ There are approximately 34 of them – same name, same age, same 3-digit zip
- ⦿ There is roughly an 80% chance that two of them share the exact same birthday
- ⦿ Correctly matching records to the right James Smith could be error prone

Other names are unique - little or no more info required

- ⦿ There is only one person named Justin Bieber in the US
- ⦿ Matching the name means very few additional fields needed to assign records

Normalizing for frequency removes significant bias from the matching

- ⦿ Names, Locations, Ages, etc.
- ⦿ Correcting for the underlying distribution to understand – and correct – disparities in the error rates for different populations

Comparison of Methods

Tested Probabilistic Matching vs. Deterministic Matching

- ⦿ Public database of 1M Florida residents
- ⦿ Individuals matched based on Name, DOB, Gender, and 3-digit zip
- ⦿ Representative noise was synthesized
- ⦿ Zip code changes are based on actual relocations over a one year period

Probabilistic Matching provides a significant advantage

- ⦿ Moving and typos were the largest source of errors for Deterministic methods
- ⦿ Most of the false negatives for probabilistic matching came from individuals with identical names and ages – the matching deferred linking instead of risking an error

Full potential of Probabilistic Matching is still untapped

- ⦿ Significant number of informative fields were not used to provide a fair comparison
- ⦿ All of the fields in this test data was 100% populated – missing data was not an issue

Results on 1M Florida Residents

	False Positives	False Negatives	FP %	FN %
Deterministic	8291	117686	0.829	11.769
Probabilistic	12	3863	0.001	0.386

Examples of Available Toolkits for Linkage

Table 1. A systematic comparison of existing generic and privacy preserving record linkage software tools.

Tool	PRL	Free	Open Source	Extensible	Communication ^a	Support
<i>LinkKing</i> ²⁰	No	No ^c	Yes	Limited ^e	Manual	Desktop ^b
<i>LinkPlus</i> ¹⁹	No	Yes	No	No	Manual	Desktop ^b
<i>FEBRL</i> ²⁵	No ^d	Yes	Yes	Limited ^e	Manual	Desktop ^b
<i>FRIL</i> ²⁴ + <i>LinkIt</i> ²⁸	Yes	Yes	Yes	Limited ^e	Manual	Desktop ^b
<i>MTB</i> ²⁷	Yes	Yes	No ^f	Limited ^e	Manual	Desktop ^b
<i>OpenEMPI</i> ¹⁴	No	Yes	Yes	Yes ^g	Yes ^h	Enterprise
<i>OpenMRS</i> ²²	No	Yes	Yes	Yes ^g	Yes ^h	Enterprise
<i>RECLINK</i> ²⁷	No	No ^c	Yes	Limited ^e	Manual	Desktop ^b
<i>SOEMPI</i>	Yes	Yes	Yes	Yes ^g	Yes ⁱ	Enterprise

^a Communication with other entities.

^b Requires RDP or VNC to view GUI of the server.

^c The script is free in itself, but requires additional SAS or Stata license to run.

^d Proposed, but not implemented.

^e Requires specific programmer knowledge.

^f BloomEncode and SafeLink sourcecode is available only for research projects.

^g SOA software, designed for extensibility, requires programmer knowledge.

^h With standard HCO actors, but not in a complex record linkage protocol.

ⁱ With other SOEMPI instances for record linkage.

Source: Toth C, Durham E, Kantarcioglu M, Xue Y, Malin B. SOEMPI: A Secure Open Enterprise Master Patient Index Software Toolkit for Private Record Linkage. *AMIA Annual Symposium Proceedings*. 2014;2014:1105-1114.

Now that we've figured out how to link the patient....we can link to their real-world data, right?

Often, linkage to RWD driven by access or convenience. Based on the methodology and data sources you are using:

Extensive longitudinal data can introduce selection bias

Available data sources skew which patients are available

Not all patients are observable for the same duration

Source selection is equally biasing

Frequency and quality of data varies with different data sources

As longitudinal data increases, so does the risk for spurious correlation

Best practices for big data need to be maintained

Results should be tested against an independent validation set

As more information is collected, and is more complete, potential for previously unanticipated identification risk grows

What other information or data sets can this be cross linked to?

- How many patients have seen these exact four doctors?
- Can enhanced demographics or location patterns leak identity or household information?

Understanding the implicit research limitations and potential for bias imposed due to sample selection

Source Type	Benefit	Representativeness	Activity Capture
EMR	Deeper clinical data (in particular if access to notes) All encounters with provider	Inaccurate measure of drug exposure May overweight certain populations or standards of care	Patients lost if seeing providers outside the sample
Medical Claims (health plan)	Complete record of paid claims for patient	Population bias May reflect plan controls	Patient lost pre- or post-enrollment
Medical Claims (other sources)	Representative across all payers, patient types	Incomplete view of patient activity	Patients lost if seeing providers outside the sample
Pharmacy (health plan, PBM)	Complete record of paid Rx's for patient	No cash Doesn't mirror medical benefits Population bias May reflect plan controls	Patient lost pre- or post-enrollment
Pharmacy (other sources)	Representative across all payers, patient types, Cash Rx's	Incomplete view of patient activity	Patients lost if seeing providers outside the sample

Lastly, Re-Identification Risk

Re-identification risk is a factor if the data you are linking to has been de-identified pursuant to HIPAA, as is common in many epidemiological databases.

Broad-scale RWD linkages potentially leak identifiable information about the patient:

- Medical claims place of service codes (claims have service delivery address)

12	Home
13	Assisted Living Facility
14	Group Home

- Dx codes, e.g.

E80*-E84*	Vehicle accident
-----------	------------------

- Demographics

(C) Age-related information more specific than one year and greater than 85 years old
(D) Race and Ethnicity
(E) Marriage status
(F) Language preference
(G) Death status within a month of the event

However, some of this may be useful or necessary, e.g. indicators of mortality

Given the caveats- why do it?

Relatively low-cost access to large-scale RWD:

- Billions of annual data points on Rxs, medical events, lab results, EMR records already generated as a byproduct of healthcare treatment and payment.
- Growing access to new sources like NLP from notes, device and genomics data.
- A lot of willing data supplier participants as long as use case is acceptable and patient privacy can be protected.

Thank You.

Andrew Kress
akress@healthverity.com
215.582.2008
100 North 20th Street
Suite 203
Philadelphia, PA 19103