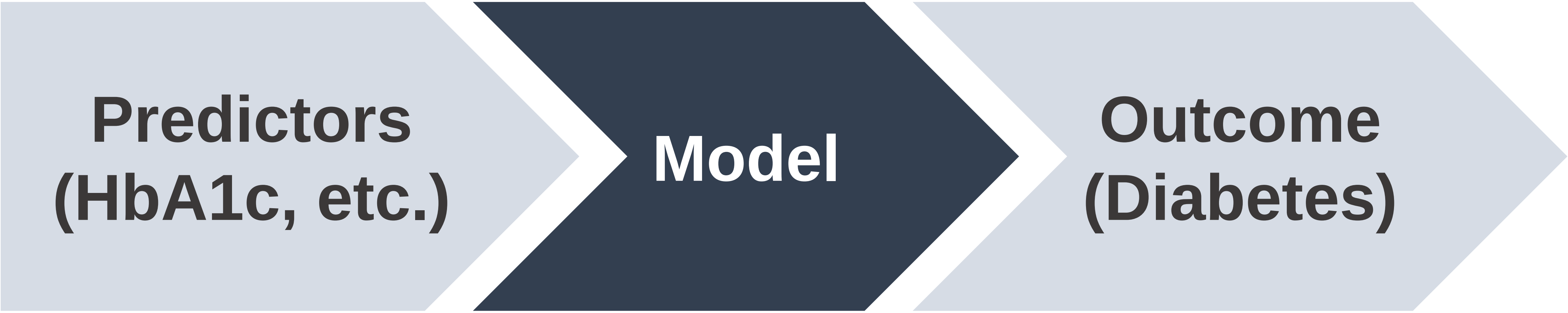
Validation with a non-representative gold standard: methods for administrative data linkage

Evelyn Lauren[1,2], Dickman Gareta[3,4,5], Khumbo Shumba[1], Kobus Herbst[3], Dorina Onoya[1], Jacob Bor[1,3,6,7]

[1]Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[3]Africa Health Research Institute, Somkhele, South Africa
[4]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
5University of Bern, Graduate School for Health Sciences, Bern, Switzerland
[6]Department of Global Health, Boston University School of Public Health, Boston, MA, USA
[7]Boston University School of Public Health, Department of Epidemiology, Boston, United States

Validation: a core problem in epidemiology

In epidemiology, we often want to make predictions or inferences on things we don’t see using data that we do see.



Predictors (HbA1c, etc.)

Model

Outcome (Diabetes)

We can evaluate the model performance (Sen, Spec, PPV, NPV) by comparing it against a gold standard.

Validation: a core problem in record linkage

Records in administrative databases often contain data entry errors

National ID	Record no.	First name	Last name	DOB	Place of residence
N/A	1001	Peter	Parker	10/08/2001	*N/A*
N/A	1002	Pedro	Packer	01/08/2001	New York

Do these records belong to the same individual?

Validation: a core problem in record linkage

Records in administrative databases often contain data entry errors

National ID	Record no.	First name	Last name	DOB	Place of residence
N/A	1001	Peter	Parker	10/08/2001	*N/A*
N/A	1002	Pedro	Packer	01/08/2001	New York

Do these records belong to the same individual?

Which records belong to which individuals?

Validation: a core problem in record linkage

Records in administrative databases often contain data entry errors

National ID	Record no.	First name	Last name	DOB	Place of residence
N/A	1001	Peter	Parker	10/08/2001	*N/A*
N/A	1002	Pedro	Packer	01/08/2001	New York

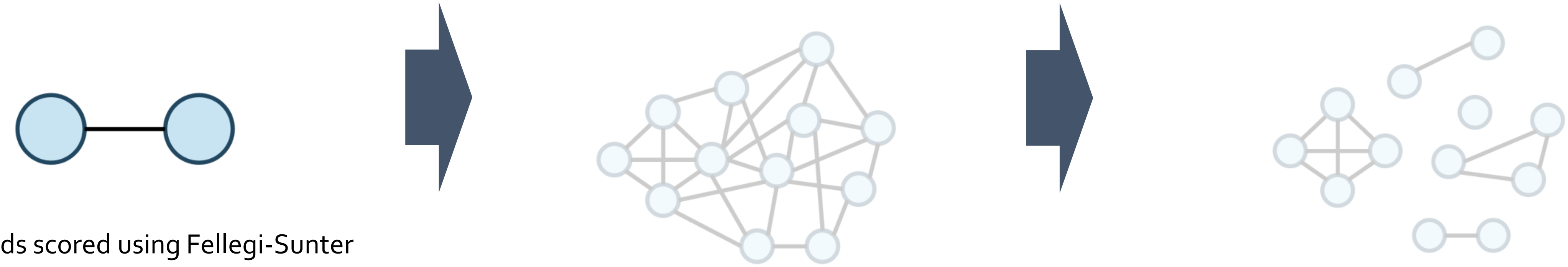
Do these records belong to the same individual?

We probabilistically assign patient IDs to records

1. Obtain similarity score based on demographics

2. Multiple records link to the same individual

3. Identify records belonging to underlying, but unobserved, individuals



*records scored using Fellegi-Sunter (1969) and Jaro-Winkler (1995)

Validation: a core problem in record linkage

Records in administrative databases often contain data entry errors

National ID	Record no.	First name	Last name	DOB	Place of residence
N/A	1001	Peter	Parker	10/08/2001	*N/A*
N/A	1002	Pedro	Packer	01/08/2001	New York

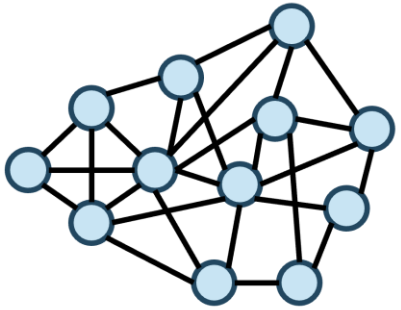
Do these records belong to the same individual?

We probabilistically assign patient IDs to records

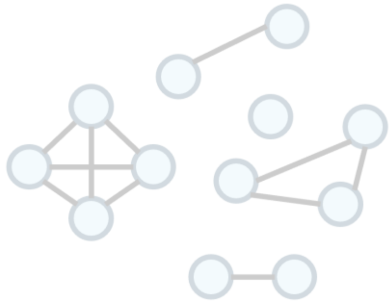
1. Obtain similarity score based on demographics

*records scored using Fellegi-Sunter (1969) and Jaro-Winkler (1995)

2. Multiple records link to the same individual



3. Identify records belonging to underlying, but unobserved, individuals



Validation: a core problem in record linkage

Records in administrative databases often contain data entry errors

National ID	Record no.	First name	Last name	DOB	Place of residence
N/A	1001	Peter	Parker	10/08/2001	*N/A*
N/A	1002	Pedro	Packer	01/08/2001	New York

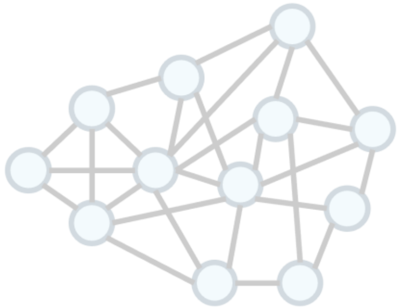
Do these records belong to the same individual?

We probabilistically assign patient IDs to records

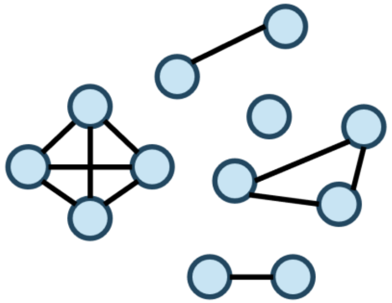
1. Obtain similarity score based on demographics

*records scored using Fellegi-Sunter (1969) and Jaro-Winkler (1995)

2. Multiple records link to the same individual



3. Identify records belonging to underlying, but unobserved, individuals



Finding gold standard data can be challenging

Resource/cost-intensive

Ethical concerns

Confidentiality concerns

Noise/bias

Non-representative data

Obtaining a gold standard in record linkage

1. **Manual review**

   (+) Can be done on a representative sample

   (–) Expensive, not scalable, reviewer bias*

*may lead to bias in estimated SEN and PPV

Obtaining a gold standard in record linkage

1. **Manual review**

   (+) Can obtain a representative sample

   (–) Expensive, not scalable, subject to reviewer bias*

2. **Known ground truth for a subset of records**

   (+) Cheap, scalable

   (–) Non-representative sample*

*may lead to bias in estimated SEN and PPV

Obtaining a gold standard in record linkage

1. **Manual review**

   (+) Can obtain a representative sample

   (–) Expensive, not scalable, subject to reviewer bias*

2. **Known ground truth for a subset of records**

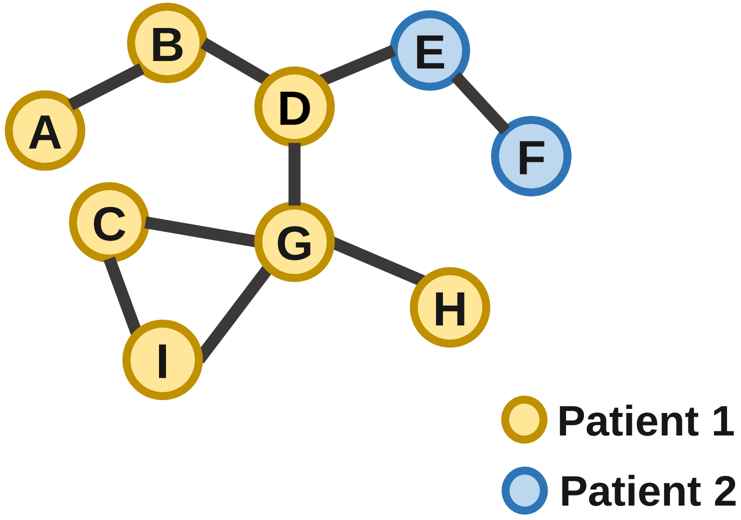
   (+) Cheap, scalable

   (–) Non-representative sample*

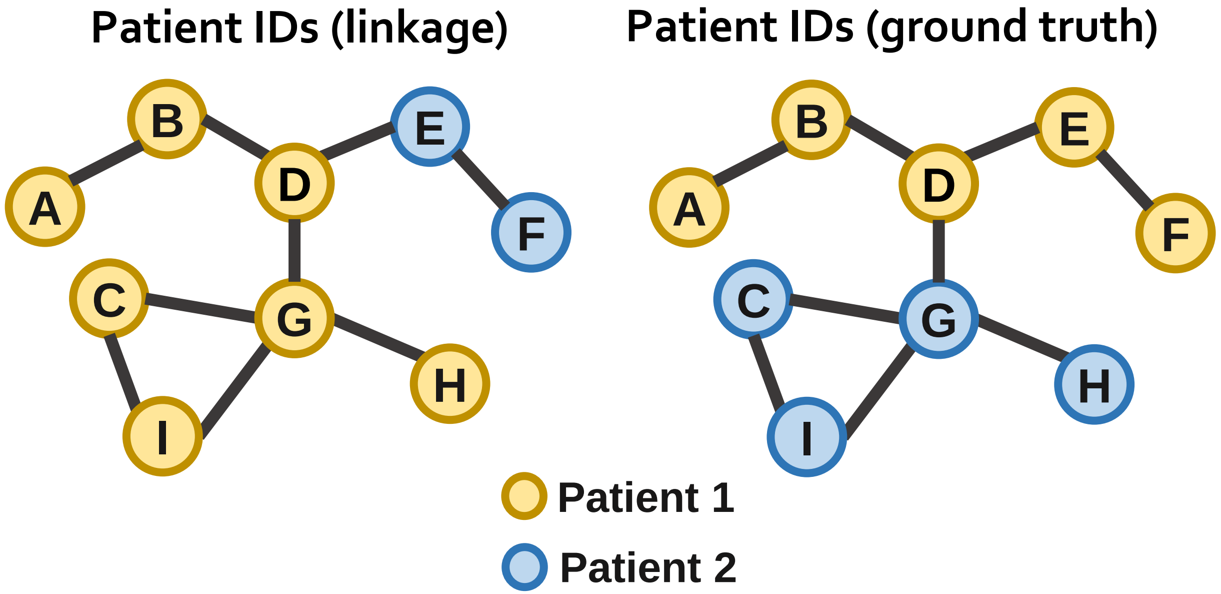
*leads to bias in estimated SEN and PPV

Assessing linkage performance

Patient IDs (linkage)



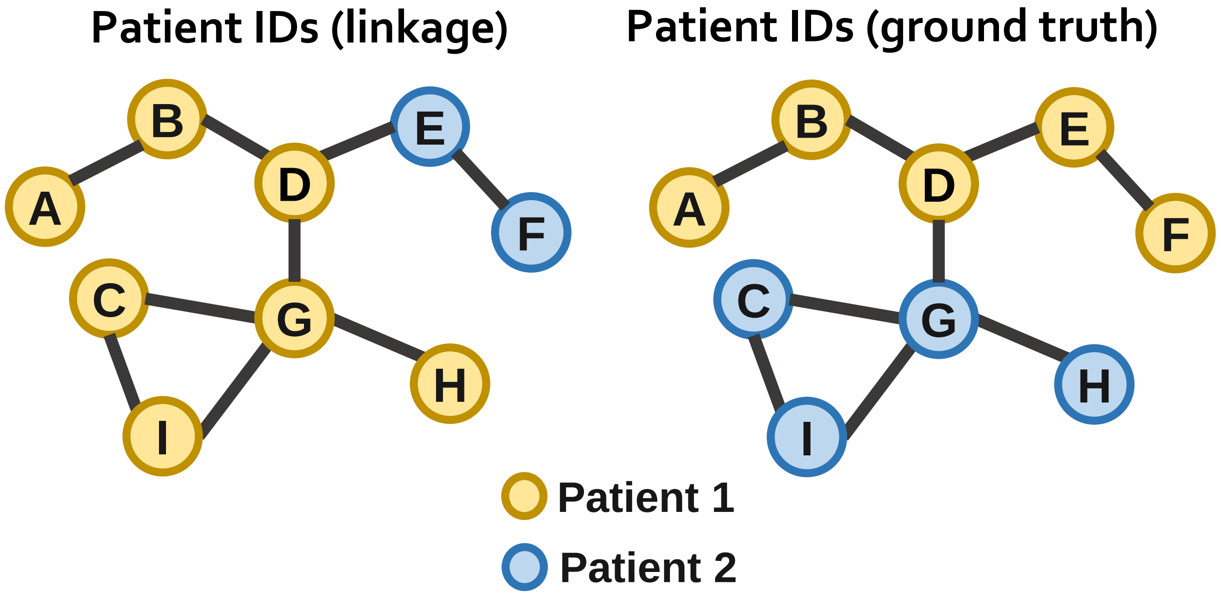
Assessing linkage performance



Patient IDs (linkage)

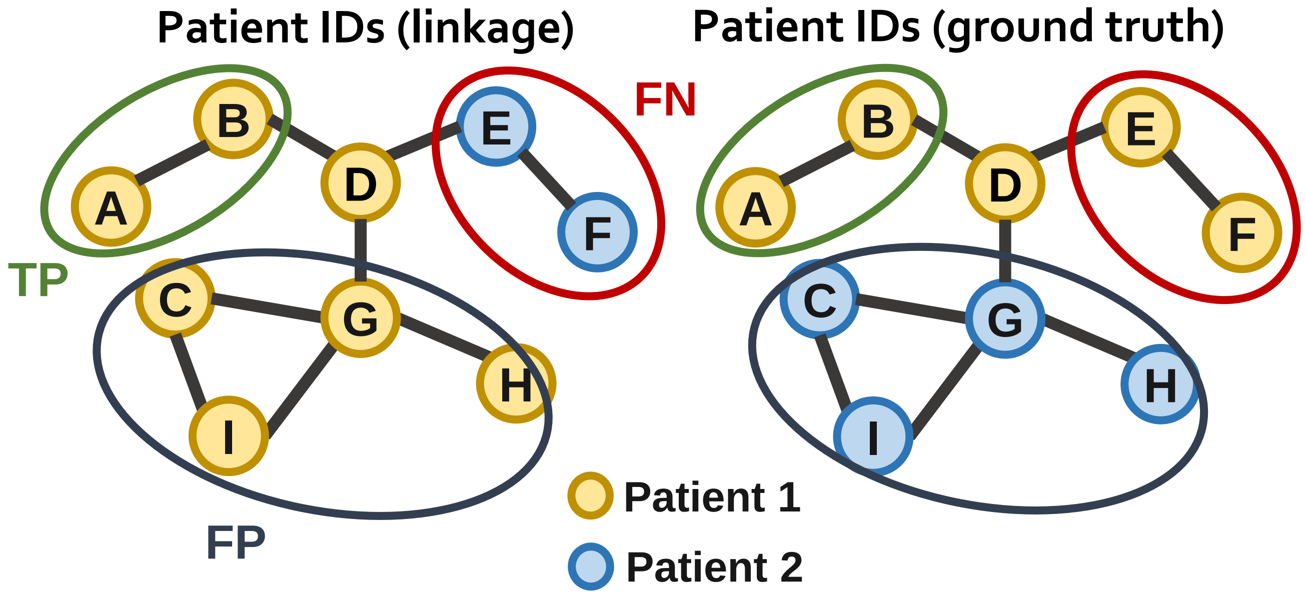
Patient IDs (ground truth)

Assessing linkage performance



		Ground truth	
		Match	**Non-match**
Linkage	**Match**	True positives (TP)	False positives (FP)
	Non-match	False negatives (FN)	True negatives (TN)

Assessing linkage performance



		Ground truth	
		Match	Non-match
Linkage	Match	True positives (TP)	False positives (FP)
	Non-match	False negatives (FN)	True negatives (TN)

$$\mathbf{SEN} = \mathbf{share\ of\ true\ links\ that\ are\ linked}$$

$$= \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} = \frac{\mathbf{2}}{\mathbf{2} + \mathbf{2}} = \mathbf{0.5}$$

$$\mathbf{PPV} = \mathbf{share\ of\ links\ that\ are\ true\ links}$$

$$= \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} = \frac{\mathbf{2}}{\mathbf{2} + \mathbf{4}} = \frac{\mathbf{2}}{\mathbf{6}} = \mathbf{0.333}$$

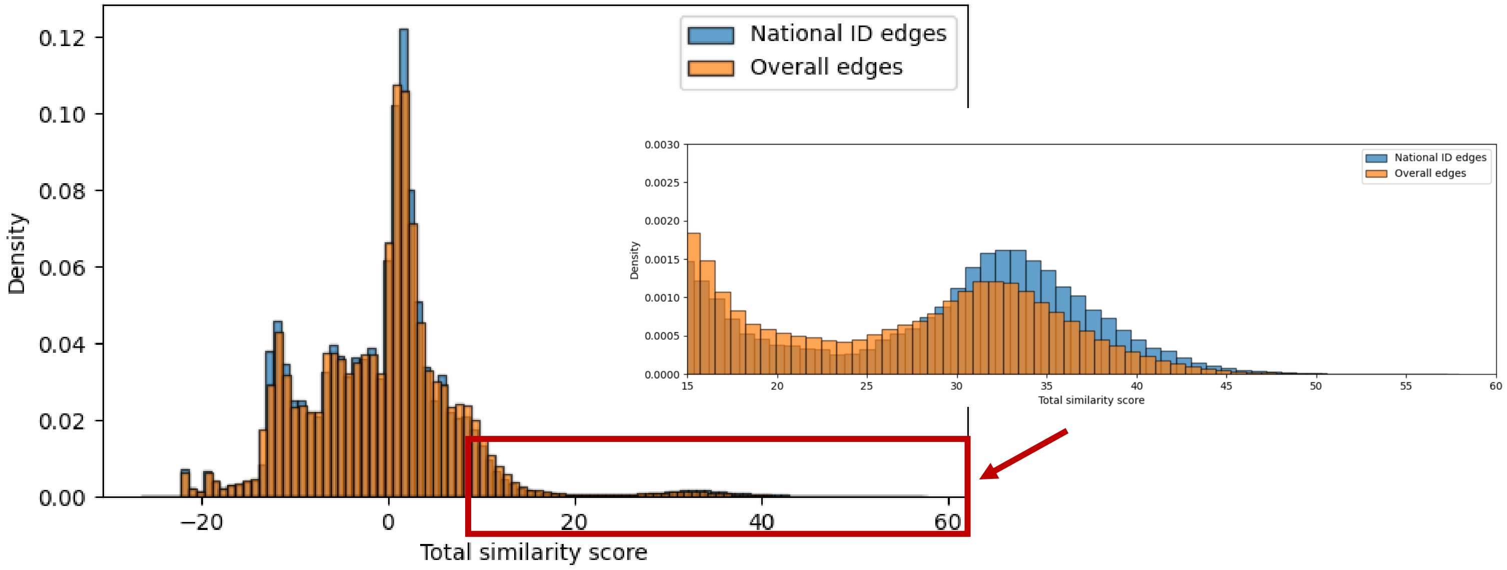
Problem:

We don’t know the ground truth for every observation

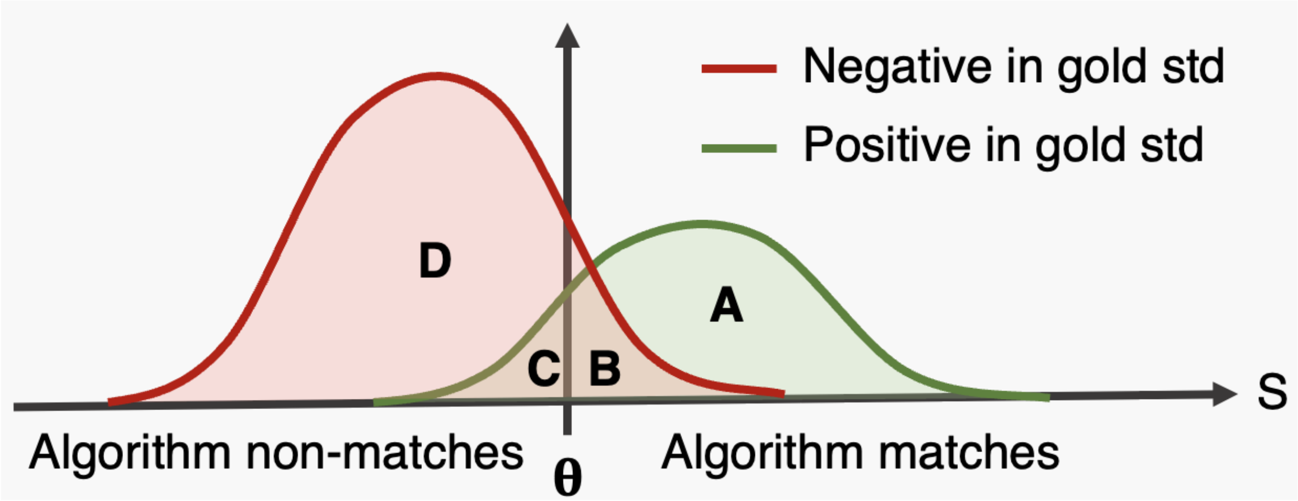
Bias when ground truth is non-representative

Subset of records with known ground truth (NATL ID) may be:

- More accurately recorded
- Higher representation in:
  - very low scores (certain non-matches)
  - Very high scores (certain matches)



Correcting for bias due to sample non-representativeness



A: True positives
B: False positives
C: False negatives
D: True negatives

We can express Sen and PPV using distributions from the linkage

- The distribution of similarity scores, $f(S)$
- The probability of being a true match conditional on the similarity score S, $P(match|S)$

$$SEN = \frac{\int_{\theta}^{\infty} f(S) \cdot P(match|S)\, dS}{\int_{-\infty}^{\infty} f(S) \cdot P(match|S)\, dS} \qquad PPV = \frac{\int_{\theta}^{\infty} f(S) \cdot P(match|S)\, dS}{\int_{\theta}^{\infty} f(S)\ dS}$$

Correcting for bias due to sample non-representativeness

We observe $\boldsymbol{P(match|S, ID)}$ instead of $\boldsymbol{P(match|S)}$

Assumption: IDs are missing at random (MAR) conditional on S, $P(match|S, ID) = P(match|S)$

- S encapsulates all similarity information between any record pair

$$SEN = \frac{\int_{\theta}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID) \ ds}{\int_{-\infty}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID) \, ds}$$

$$\text{where } w(S) = \frac{f(S)}{f(S|ID)}$$

$$PPV = \frac{\int_{\theta}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID) \ ds}{\int_{\theta}^{\infty} f(S|ID) \cdot w(S) \ ds}$$

Correcting for bias due to sample non-representativeness

We observe $\boldsymbol{P(match|S, ID)}$ instead of $\boldsymbol{P(match|S)}$

Assumption: IDs are missing at random (MAR) conditional on S, $P(match|S, ID) = P(match|S)$

- S encapsulates all similarity information between any record pair

$$SEN = \frac{\int_{\theta}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID)\ ds}{\int_{-\infty}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID)\, ds}$$

$$\text{where } w(S) = \frac{f(S)}{f(S|ID)}$$

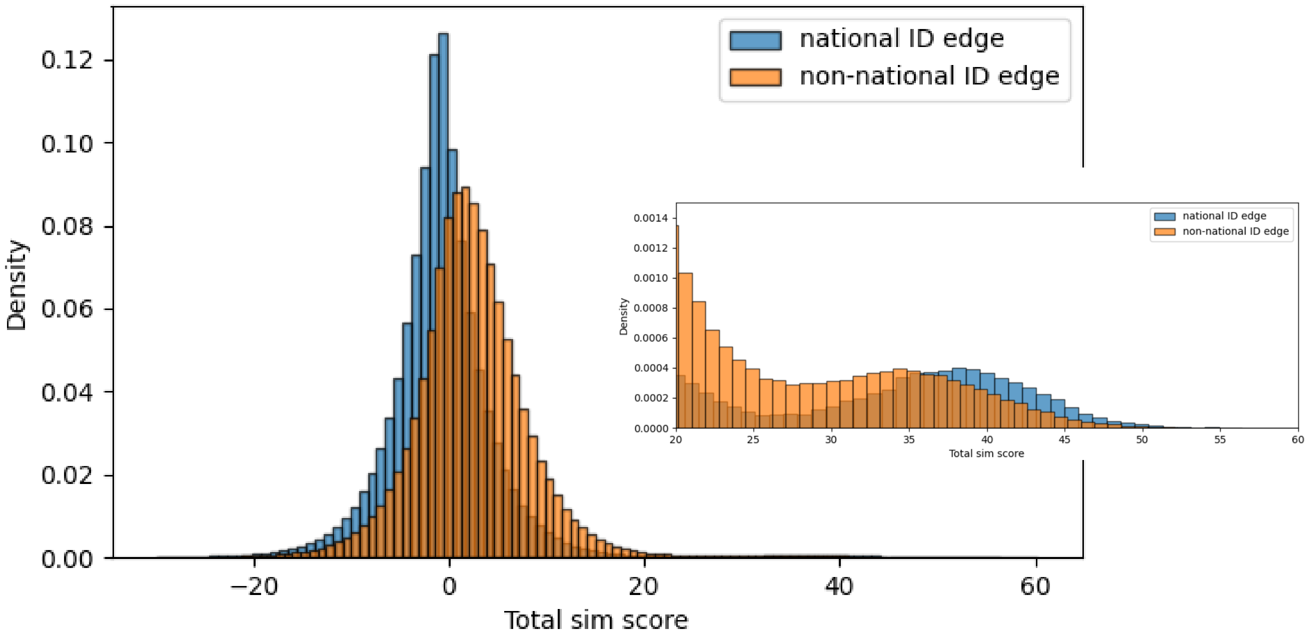
$$PPV = \frac{\int_{\theta}^{\infty} f(S|ID) \cdot w(S) \cdot P(match|S, ID)\ ds}{\int_{\theta}^{\infty} f(S|ID) \cdot w(S)\ ds}$$

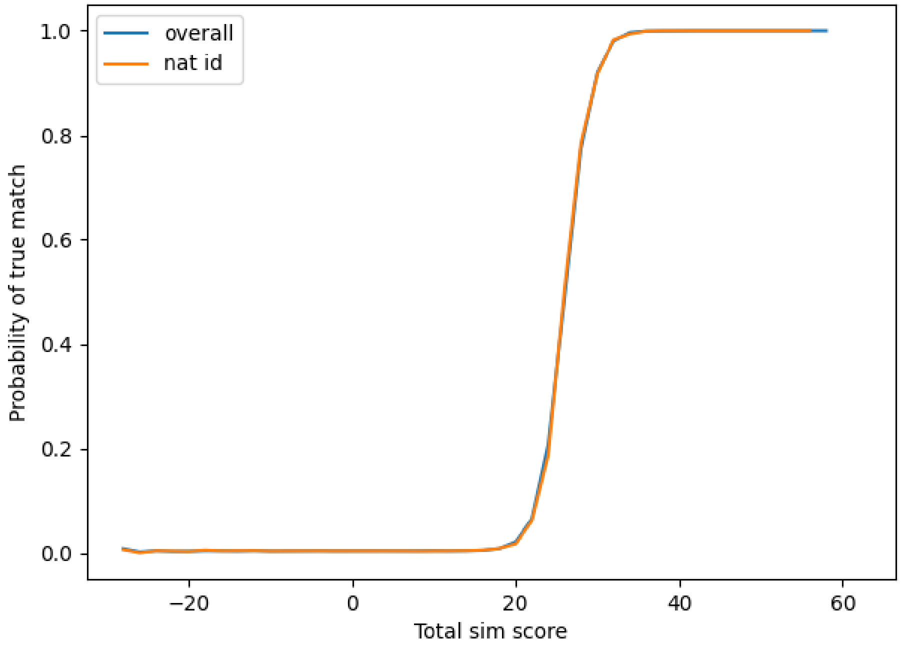
We can estimate PPV and SEN without bias by reweighting the data

Simulation study

Simulation study

- Construct simulated data as follows:
  - Have lower proportion of missing IDs for very low/very high S
  - IDs missing at random conditional on S





Simulation study

	SEN (%)	PPV (%)	SEN difference from ground truth (%)	PPV difference from ground truth (%)
Ground truth data	54.6	99.8	-	-
50% missing national ID, unadjusted	66.1	99.9	+11.5	+0.1
50% missing national ID, bias-corrected	57.3	99.7	+2.7	-0.1

In simulated data, applying bias-correction approach reduces bias

Application

Application: Africa Health Research Institute (AHRI) clinical and laboratory HIV databases

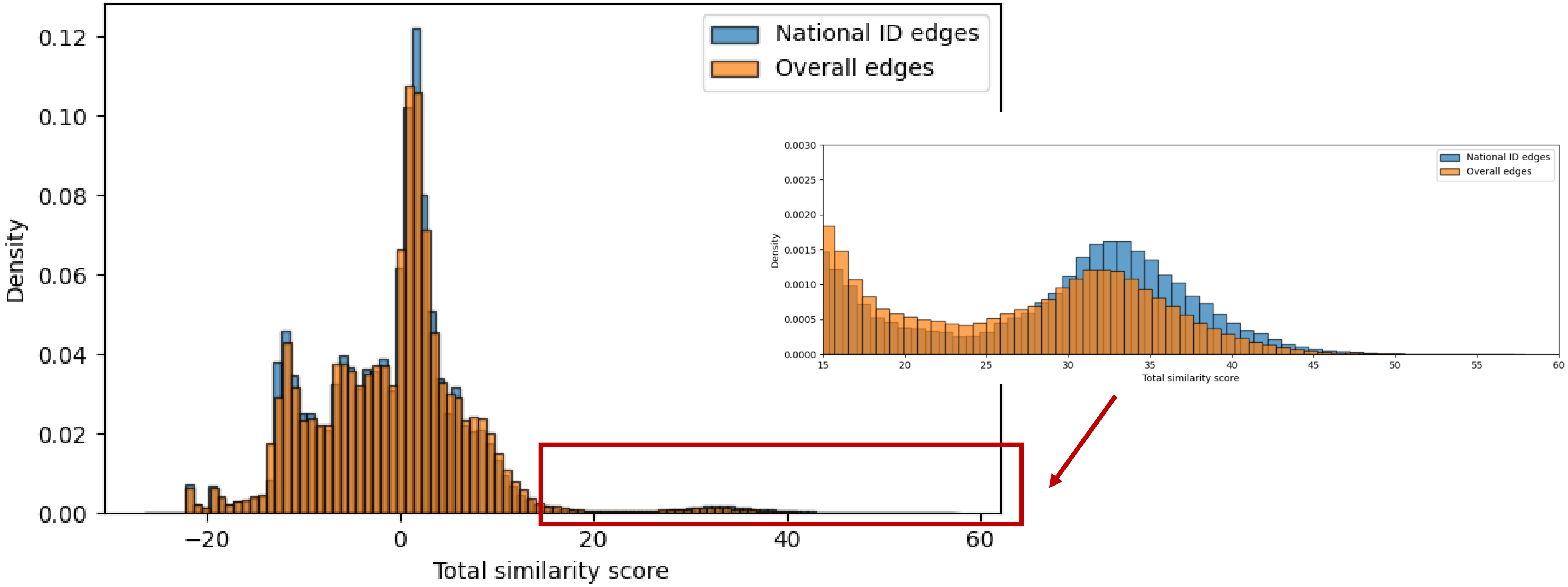
- **Link records from multiple non-deduplicated datasets for HIV care monitoring in South Africa [1, 2]:**
  - Tier.Net, NHLS laboratory database, HDSS, AHRILink

- **National ID information available for:**
  - 71% of TIER
  - 23% of NHLS
  - 40% of HDSS
  - 21% of AHRILink

[1] Bor J, MacLeod W, Oleinik K, Potter J, Brennan AT, Candy S, et al. Building a national HIV cohort from routine laboratory data: Probabilistic record-linkage with graphs. bioRxiv. 2018;

[2] MacLeod WB, Bor J, Candy S, Maskew M, Fox MP, Bulekova K, et al. Cohort profile: the South African National Health Laboratory Service (NHLS) National HIV Cohort. BMJ Open [Internet]. 2022 Oct 1;12(10):e066671. Available from: http://bmjopen.bmj.com/content/12/10/e066671.abstract

Comparison between types of record pairs



Record pairs with national ID are less likely to be missing for very low and very high total similarity scores

Validation performance metrics

Bias correction	SEN (%)	PPV (%)	Undermatch rate (1-SEN)	Overmatch rate (1-PPV)
No bias correction	94.3	96.8	5.7	3.2
With weights correcting for bias due to missing national IDs	91.7	94.8	8.3	5.2

Both SEN and PPV are overestimated if we do not correct for bias

Failure to correct for bias would have led to:

$$\frac{5.7 - 8.3}{5.7} = -46\%$$

46% underestimate in the undermatching error rate

$$\frac{3.2 - 5.2}{3.2} = -63\%$$

63% underestimate in the overmatching error rate

In summary,

- Validation using a non-representative gold standard creates a potential for a cost-effective, easy to implement, and scalable procedure
- Failure to correct for bias will result in incorrect estimation of performance metrics
- Approach can be generalized to any misclassification problem involving a non-representative gold standard

Acknowledgements

NIH National Institute of Allergy and Infectious Diseases

AHRI AFRICA HEALTH RESEARCH INSTITUTE

Health Economics and Epidemiology Research Office
HE²RO
Wits Health Consortium
University of the Witwatersrand

BOSTON UNIVERSITY

NATIONAL HEALTH LABORATORY SERVICE

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

health
Department:
Health
REPUBLIC OF SOUTH AFRICA

Thank you

elauren@bu.edu