False Discovery estimation in Record Linkage

Kayané Robach

Amsterdam UMC
Universitair Medische Centra

EPIDEMIOLOGY AND
DATA SCIENCE

Bigstatistics

May 19, 2025

Contents

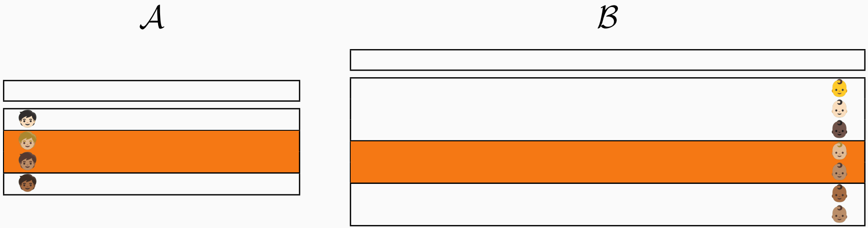
Problem

Record Linkage: a motivational example

The Netherlands Perinatal Registry gathers about 96% of all deliveries

We could study the risk of pre-term birth using characteristics of the mothers and data from past deliveries

Data are at the scope of the babies, family portraits need to be assembled



Record Linkage: a motivational example

Make use of 'partially identifying variables' *postal code, birth date*

Combine data sources to recover the siblings:
linked data common to $\mathcal{A}$ and $\mathcal{B}$

The true linkage structure is latent

$\mathcal{A}$

zipcode	delivery date	pre-term
1012GL	28-06-2021	yes
1112XJ	13-04-2019	no
8043VD	14-10-2015	yes
3572TC	03-08-2008	yes

$\mathcal{B}$

Age	ART	zipcode	delivery date	pre-term	past delivery
25	yes	1012GL	02-04-2022	no	
45	yes		21-01-2020	no	
51	no	8043VD	03-09-2009	yes	29-05-1995
45	no	1112XJ	12-01-2020	yes	13-04-2019
33	no	8011PK	15-04-2018	no	14-10-2015
22	yes	3572TC	27-08-2019	no	
29	no	3522BB	18-01-2013	yes	09-05-2010

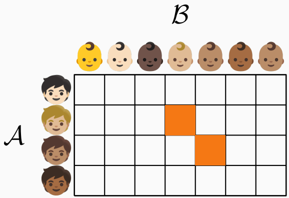
Record Linkage: a motivational example

Make use of 'partially identifying variables' *postal code, birth date*

Combine data sources to recover the siblings:
linked data common to $\mathcal{A}$ and $\mathcal{B}$

The true linkage structure is latent

$\Delta :=$



Record Linkage, and then what?

Analyses are made on linked data without specifying the linkage process nor the expected reliability of this linkage

Record Linkage, and then what?

Analyses are made on linked data without specifying the linkage process nor the expected reliability of this linkage

- RL applied on Perined data
  - to study mother/children dynamics
    $\rightarrow$ RL to combine Perined with external source(s)
  - to study pre-term birth, post-term birth, stillbirth risks
    $\rightarrow$ RL to link the siblings

How to evaluate a method?

Sensitivity / Specificity are often used in the literature to evaluate RL methods (on sets for which we know the true linkage structure)

How to evaluate a method?

Sensitivity / Specificity are often used in the literature to evaluate RL methods (on sets for which we know the true linkage structure)

- False Negative Rate (FNR $= 1 -$ sensitivity) captures the missed links
  - Missed links are the hardest pairs to detect
    $\rightarrow$ registration errors (missing values or mistakes)
    $\rightarrow$ changes over time (moving)
    $\rightarrow$ processes that we try to estimate within the RL model

How to evaluate a method?

Sensitivity / Specificity are often used in the literature to evaluate RL methods (on sets for which we know the true linkage structure)

- False Negative Rate (FNR $= 1 -$ sensitivity) captures the missed links
  - Missed links are the hardest pairs to detect
    $\rightarrow$ registration errors (missing values or mistakes)
    $\rightarrow$ changes over time (moving)
    $\rightarrow$ processes that we try to estimate within the RL model

- False Discovery Rate (FDR $= 1 -$ specificity) captures the falsely linked pairs
  - What about falsely linked pairs?

What about false discoveries?

The (only) proposed estimate of the error made when linking records relies on the RL model (sensitive to the difficulty of the task)

What about false discoveries?

The (only) proposed estimate of the error made when linking records relies on the RL model (sensitive to the difficulty of the task)

- RL methods target:
  - Δ: indicator matrix defined by cartesian product of sets A and B, 1 for a link, 0 for a non-link

What about false discoveries?

The (only) proposed estimate of the error made when linking records relies on the RL model (sensitive to the difficulty of the task)

- RL methods target:
  - Δ: indicator matrix defined by cartesian product of sets A and B, 1 for a link, 0 for a non-link
- Probabilistic RL models provide:
  - $\widehat{\Delta}(\xi)$: matrix of linkage probabilities from the RL model (taking $\xi > 0.5$ determines coherent linked pairs)

What about false discoveries?

The (only) proposed estimate of the error made when linking records relies on the RL model (sensitive to the difficulty of the task)

- RL methods target:
  - Δ: indicator matrix defined by cartesian product of sets A and B, 1 for a link, 0 for a non-link
- Probabilistic RL models provide:
  - $\widehat{\Delta}(\xi)$: matrix of linkage probabilities from the RL model (taking $\xi > 0.5$ determines coherent linked pairs)

$$\mathsf{FDR} = \frac{\mathit{FP}}{\mathit{FP} + \mathit{TP}} = 1 - \frac{\mathit{TP}}{\text{linked records}}$$

$$\mathbb{P}\widehat{\mathsf{FDR}}(\xi) = 1 - \frac{\sum_{i,j} \hat{\Delta}_{i,j} \cdot \mathbb{1}\{\hat{\Delta}_{i,j} > \xi\}}{\sum_{i,j} \mathbb{1}\{\hat{\Delta}_{i,j} > \xi\}}$$

False discovery estimation

What would we like?

Difficulty of the RL task:

- low discriminative power of the Partially Identifying Variables (PIVs)
- registration errors / information changes between data collections
- complex distributions of the PIVs
- dependencies among the PIVs

What would we like?

Difficulty of the RL task:

- low discriminative power of the Partially Identifying Variables (PIVs)
- registration errors / information changes between data collections
- complex distributions of the PIVs
- dependencies among the PIVs

Tendency to overestimate linkage probabilities in practice

What would we like?

Difficulty of the RL task:

- low discriminative power of the Partially Identifying Variables (PIVs)
- registration errors / information changes between data collections
- complex distributions of the PIVs
- dependencies among the PIVs

Tendency to overestimate linkage probabilities in practice
(i.e. underestimate the $\mathbb{P}\widehat{\text{FDR}}$)

$\mathbb{P}\widehat{\text{FDR}}$ is seldom used and often not available from the implemented RL algorithms

What would we like?

Difficulty of the RL task:

- low discriminative power of the Partially Identifying Variables (PIVs)
- registration errors / information changes between data collections
- complex distributions of the PIVs
- dependencies among the PIVs

Tendency to overestimate linkage probabilities in practice
(i.e. underestimate the $\mathbb{P}\widehat{\text{FDR}}$)

$\mathbb{P}\widehat{\text{FDR}}$ is seldom used and often not available from the implemented RL algorithms

We want an estimation procedure that is independent of the RL model

Estimation procedure: a new $\widehat{\mathsf{FDR}}$

`Input`

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\left\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\right\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

Estimation procedure: a new $\widehat{\text{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\left\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\right\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$

Estimation procedure: a new $\widehat{\text{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\left\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\right\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$
$\quad\quad \widehat{\Delta}(\xi) \leftarrow \left\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [\xi, 1]\right\}$

Estimation procedure: a new $\widehat{\text{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$

$\quad \widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [\xi, 1]\big\}$

$\quad FP_{\text{synth}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{S}\text{ynth}\}$

Estimation procedure: a new $\widehat{\text{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\, i \in \mathcal{A},\, j \in \tilde{\mathcal{B}},\, p \in [0, 1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$

$\widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\, i \in \mathcal{A},\, j \in \tilde{\mathcal{B}},\, p \in [\xi, 1]\big\}$

$FP_{\text{synth}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\, j \in \mathcal{S}\text{ynth}\}$

$N_{\text{real linked}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\, j \in \mathcal{B}\}$

Estimation procedure: a new $\widehat{\mathsf{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$

$\widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [\xi, 1]\big\}$

$FP_{\text{synth}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{S}\text{ynth}\}$

$N_{\text{real linked}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{B}\}$

$$\frac{FP_{\text{synth}}(\xi)}{N_{\text{synth}}} \approx \frac{FP(\xi)}{N_{\mathcal{B}}}$$

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0,1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

`For` $\xi \in (0.5, 1)$

$\widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [\xi, 1]\big\}$

$FP_{\text{synth}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{S}\text{ynth}\}$

$N_{\text{real linked}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{B}\}$

$\dfrac{FP_{\text{synth}}(\xi)}{N_{\text{synth}}} \approx \dfrac{FP(\xi)}{N_{\mathcal{B}}},\ N_{\text{real linked}}(\xi) = FP(\xi) + TP(\xi)$

Estimation procedure: a new $\widehat{\text{FDR}}$

Input RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\, i \in \mathcal{A},\, j \in \tilde{\mathcal{B}},\, p \in [0, 1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

For $\xi \in (0.5, 1)$

$$\widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\, i \in \mathcal{A},\, j \in \tilde{\mathcal{B}},\, p \in [\xi, 1]\big\}$$

$$FP_{\text{synth}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\, j \in \mathcal{S}\text{ynth}\}$$

$$N_{\text{real linked}}(\xi) \leftarrow \sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\, j \in \mathcal{B}\}$$

$$\widehat{\text{FDR}}(\xi) \leftarrow \frac{FP_{\text{synth}}(\xi) \cdot N_{\mathcal{B}} / N_{\text{synth}}}{N_{\text{real linked}}(\xi)}$$

Estimation procedure: a new $\widehat{\text{FDR}}$

`Input` RL algorithm, synthesiser, file $\mathcal{A}$, file $\mathcal{B}$, $N_{\mathcal{A}} \leq N_{\mathcal{B}}$

\# Synthesise $N_{\text{synth}} = 0.10 \times N_{\mathcal{B}}$ records based on file $\mathcal{B}$
$\mathcal{S}\text{ynth} \leftarrow \text{synthesiser}(N_{\text{synth}}, \mathcal{B})$

$\tilde{\mathcal{B}} \leftarrow \text{concat}(\mathcal{B}, \mathcal{S}\text{ynth})$

\# Run RL between file $\mathcal{A}$ and augmented file $\tilde{\mathcal{B}}$
$\big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [0, 1]\big\} \leftarrow \text{RL}(\mathcal{A}, \tilde{\mathcal{B}})$

`For` $\xi \in (0.5, 1)$

$$\widehat{\Delta}(\xi) \leftarrow \big\{(i, j, p),\ i \in \mathcal{A},\ j \in \tilde{\mathcal{B}},\ p \in [\xi, 1]\big\}$$

$$FP_{\text{synth}}(\xi) \leftarrow \textstyle\sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{S}\text{ynth}\}$$

$$N_{\text{real linked}}(\xi) \leftarrow \textstyle\sum_{\ell \in \widehat{\Delta}(\xi)} \mathbb{1}\{\ell := (i, j, p) \in \widehat{\Delta}(\xi),\ j \in \mathcal{B}\}$$

$$\widehat{\text{FDR}}(\xi) \leftarrow \frac{FP_{\text{synth}}(\xi) \cdot N_{\mathcal{B}} / N_{\text{synth}}}{N_{\text{real linked}}(\xi)}$$

`Output` Sets $\big\{(i, j, p),\ i \in \mathcal{A},\ j \in \mathcal{B},\ p \in [\xi, 1]\big\}_{\xi \in (0.5, 1)}$ of real linked records and corresponding $\widehat{\text{FDR}}(\xi)$

Estimation formula

The proposal is upper bounded by 1 if

$$\frac{FP_{\text{synth}}(\xi)}{N_{\text{synth}}} < \frac{N_{\text{real linked}}(\xi)}{N_B} \tag{2}$$

Estimation formula

The proposal is upper bounded by 1 if

$$\frac{FP_{\text{synth}}(\xi)}{N_{\text{synth}}} < \frac{N_{\text{real linked}}(\xi)}{N_B} \tag{2}$$

The synthetic data can only be involved in the linkage as TN or FP and as a consequence we can assume

$$\frac{\mathbb{E}[FP_{\text{synth}}(\xi)]}{N_{\mathcal{A}} N_{\text{synth}}} = \frac{FP(\xi)}{N_{\mathcal{A}} N_{\mathcal{B}}} \tag{3}$$

which ensures that the estimate is unbiased

Estimation formula

The proposal is upper bounded by 1 if

$$\frac{FP_{\text{synth}}(\xi)}{N_{\text{synth}}} < \frac{N_{\text{real linked}}(\xi)}{N_B} \tag{2}$$

The synthetic data can only be involved in the linkage as TN or FP and as a consequence we can assume

$$\frac{\mathbb{E}[FP_{\text{synth}}(\xi)]}{N_{\mathcal{A}} N_{\text{synth}}} = \frac{FP(\xi)}{N_{\mathcal{A}} N_{\mathcal{B}}} \tag{3}$$

which ensures that the estimate is unbiased

Equation (3) $\implies$ eq. (2) so we can at least get rid of biased estimates identified thanks to eq. (2) being unfulfilled

We cannot check eq. (3) in unlabelled real-life RL applications

Data synthesisers

The recent developments made in density estimation and data synthesis provide `Python` and `R` packages

We studied the estimation procedure with 2 recent methods:

- synthpop: sequential modelling using classification and regression trees on the conditional distribution of the data
- arf: adversarial random forest $\rightarrow$ generative modelling: learn by classifying data into real or synthetic

FDR estimation on real data applications

It works on large data sets

For the labelled SHIW application (16445 and 14917 records, 6430 in common)

BRL	*synthpop*	*FastLink*	*synthpop*	*FlexRL*	*synthpop*
FDR	0.23 (0.006)	FDR	0.70 (0.003)	FDR	0.42 (0.001)
$\widehat{\textbf{FDR}}$ bias	-0.035 (0.001)	$\widehat{\textbf{FDR}}$ bias	-0.062 (0.002)	$\widehat{\textbf{FDR}}$ bias	0.009 (0.008)
		prob$\widehat{\text{FDR}}$ bias	-0.57 (0.001)	prob$\widehat{\text{FDR}}$ bias	-0.10 (0.003)
condition 3	$2e^{-07}$ (e^{-07})	condition 3	e^{-06} (e^{-06})	condition 3	e^{-05} (e^{-06})

The % bias relative to the true FDR on "large" applications lays around 15% on average over the different RL methods

It works on large data sets

For the labelled NLTCS application (20484 and 9532 records, 7612 in common)

BRL	*synthpop*	*FastLink*	*synthpop*	*FlexRL*	*synthpop*
FDR	0.23 (0.006)	FDR	0.70 (0.003)	FDR	0.42 (0.001)
$\widehat{\mathbf{FDR}}$ bias	-0.035 (0.001)	$\widehat{\mathbf{FDR}}$ bias	-0.062 (0.002)	$\widehat{\mathbf{FDR}}$ bias	0.009 (0.008)
		prob$\widehat{\text{FDR}}$ bias	-0.57 (0.001)	prob$\widehat{\text{FDR}}$ bias	-0.10 (0.003)
condition 3	$2e^{-07}$ (e^{-07})	condition 3	e^{-06} (e^{-06})	condition 3	e^{-05} (e^{-06})

The % bias relative to the true FDR on "large" applications lays around 15% on average over the different RL methods

It works on medium size data sets

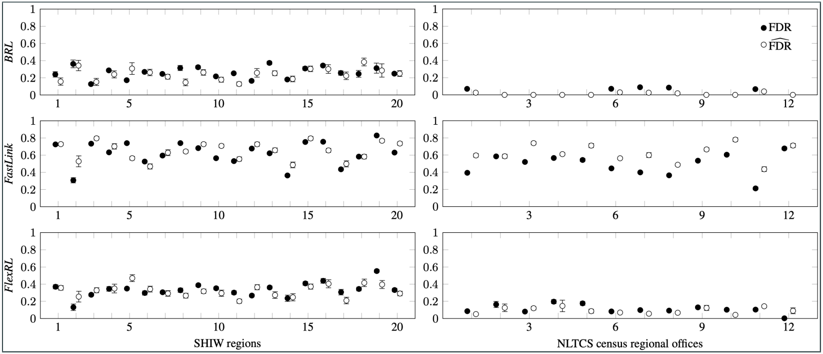
For the labelled SHIW application, North, Centre, South subsets (between 6700 and 3000 records, approximately 2000 in common)

BRL	*synthpop*	*arf*	*synthpop*	*arf*	*synthpop*	*arf*
FDR	0.073 (0.021)	0.076 (0.028)	0.222 (0.016)	0.218 (0.017)	0.188 (0.11)	0.187 (0.012)
$\widehat{\mathbf{FDR}}$ bias	0.093 (0.031)	0.019 (0.070)	-0.001 (0.012)	-0.049 (0.008)	-0.001 (0.024)	-0.077 (0.022)
condition 3	e^{-07} (e^{-07})	e^{-07} (e^{-07})	e^{-06} (e^{-07})	e^{-06} (e^{-06})	e^{-07} (e^{-07})	e^{-07} (e^{-07})
FastLink	*synthpop*	*arf*	*synthpop*	*arf*	*synthpop*	*arf*
FDR	0.791 (0.003)	0.789 (0.004)	0.727 (0.006)	0.727 (0.004)	0.741 (0.004)	0.745 (0.005)
$\widehat{\mathbf{FDR}}$ bias	-0.043 (0.012)	-0.009 (0.012)	0.002 (0.008)	-0.014 (0.004)	0.011 (0.008)	0.010 (0.011)
prob$\widehat{\text{FDR}}$ bias	-0.591 (0.001)	-0.589 (0.001)	-0.494 (0.001)	-0.495 (0.000)	-0.555 (0.001)	-0.555 (0.005)
condition 3	e^{-05} (e^{-06})	e^{-06} (e^{-06})	e^{-05} (e^{-05})	e^{-06} (e^{-06})	e^{-06} (e^{-06})	e^{-06} (e^{-06})
FlexRL	*synthpop*	*arf*	*synthpop*	*arf*	*synthpop*	*arf*
FDR	0.386 (0.019)	0.371 (0.013)	0.405 (0.017)	0.413 (0.019)	0.410 (0.014)	0.402 (0.016)
$\widehat{\mathbf{FDR}}$ bias	0.009 (0.019)	0.025 (0.031)	-0.023 (0.018)	-0.077 (0.007)	-0.028 (0.016)	-0.052 (0.029)
prob$\widehat{\text{FDR}}$ bias	-0.052 (0.005)	-0.031 (0.003)	-0.082 (0.002)	-0.086 (0.002)	-0.098 (0.002)	-0.090 (0.016)
condition 3	e^{-06} (e^{-07})	e^{-07} (e^{-07})	e^{-06} (e^{-06})	e^{-06} (e^{-06})	e^{-06} (e^{-06})	e^{-06} (e^{-06})

The % bias relative to the true FDR on "medium" applications lays around 10% on average over the different RL methods

It works on small data sets

For the labelled SHIW and NLTCS applications, regional subsets (between 150 and 2500 records)



The % bias relative to the true FDR on "small" applications lays around 20% on average over the different RL methods

A tool for improving inference on linked data

NLTCS longitudinal study: well-being of American elderly

We can link the data and estimate the FDR

NLTCS longitudinal study: well-being of American elderly

We can link the data and estimate the FDR

We can tune the parameters of the RL method to obtain a lower FDR

NLTCS longitudinal study: well-being of American elderly

We can link the data and estimate the FDR

We can tune the parameters of the RL method to obtain a lower FDR

Benchmark	
intercept	0.05 (0.00) $\star$
*FI*82	0.67 (0.02) $\star$
R^2	0.17

FastLink	default	tuned
$\widehat{\text{FDR}}$	0.63	0.35
intercept	-0.05 (0.00) $\star$	0.05 (0.01) $\star$
*FI*82	0.06	0.51 (0.02) $\star$
R^2	0.03	0.11

FlexRL	default
$\widehat{\text{FDR}}$	0.08
intercept	0.05 (0.00) $\star$
*FI*82	0.62 (0.05) $\star$
R^2	0.14

Example above: Linear model to explain the Frailty Index (FI) of 1994 using the one of 1982 on the people we can link (i.e. who survived)

SHIW census: Italian population

We can link the data and estimate the FDR

We can tune the parameters of the RL method to obtain a lower FDR

SHIW census: Italian population

We can link the data and estimate the FDR

We can tune the parameters of the RL method to obtain a lower FDR

DGP		*BRL*	default	*FastLink*	default	tuned	*FlexRL*	default	tuned
		$\widehat{\text{FDR}}$	0.09	$\widehat{\text{FDR}}$	0.75	0.65	$\widehat{\text{FDR}}$	0.39	0.18
intercept	-5	intercept	-4.79 (0.21) $\star$	intercept	-4.72 (0.09) $\star$	-4.81 (0.10) $\star$	intercept	-4.84 (0.19) $\star$	-4.62 (0.40) $\star$
β_1	1	β_1	1.14 (0.24) $\star$	β_1	0.02 (0.09)	0.04 (0.10)	β_1	0.76 (0.18) $\star$	1.12 (0.37) $\star$
β_2	1	β_2	0.98 (0.06) $\star$	β_2	0.07 (0.03) $\star$	0.10 (0.03) $\star$	β_2	0.64 (0.06) $\star$	0.80 (0.14) $\star$
β_3	20	β_3	20.32 (0.52) $\star$	β_3	1.84 (0.22) $\star$	2.27 (0.24) $\star$	β_3	12.24 (0.46) $\star$	14.99 (0.98) $\star$
		R^2	0.98	R^2	0.01	0.01	R^2	0.38	0.58

Example above: (Simulated) Linear model to explain Y in 2020 using X in 2016 on the people we can link

Perinatal registry: Noord-Holland province

BRL	default	*FastLink*	default	*FlexRL*	default
$\widehat{\text{FDR}}$	0.01	$\widehat{\text{FDR}}$	0.01	$\widehat{\text{FDR}}$	0.00
intercept	6.64 (0.83) $\star$	intercept	6.60 (0.83) $\star$	intercept	6.47 (0.94) $\star$
int btw pregnancies	-0.00 (0.05)	int btw pregnancies	-0.01 (0.05)	int btw pregnancies	-0.03 (0.07)
mother age at 1st	-0.04 (0.02) $\star$	mother age at 1st	-0.04 (0.01) $\star$	mother age at 1st	-0.04 (0.02) $\star$
duration pregnancy 1	-0.21 (0.02) $\star$	duration pregnancy 1	-0.21 (0.02) $\star$	duration pregnancy 1	-0.21 (0.02) $\star$
ART pregnancy 1	-0.13 (0.18)	ART pregnancy 1	-0.14 (0.18) $\star$	ART pregnancy 1	-0.32 (0.21) $\star$

We can continue linking these data

... and to do inference on it

Example above: estimation of the pre-term birth risk at the 2nd delivery given characteristics from the 1st delivery and the mother

Thank You!

Details on the procedure choices

Different options

- what is the setting we work on?
- what size for the synthetic data set?
- what formula for the FDR estimate?
- which synthesiser?

We investigate these: on 2 real data sets, for 3 RL R packages

The setting

The data synthesis impacts the formula we build for the estimate

Some options may be better than others

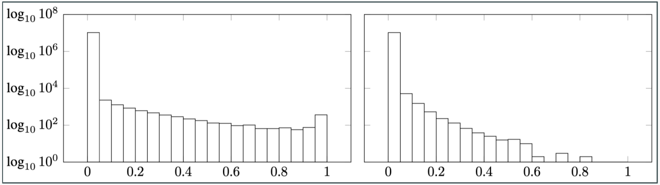
- synthesise data from A that we concatenate to A $\rightarrow$ do RL between augmented A and B
- synthesise data from B that we concatenate to B $\rightarrow$ do RL between A and augmented B
- synthesise data from both $\rightarrow$ do RL between augmented A and augmented B
- synthesise data from both $\rightarrow$ do RL between synthetic A and synthetic B

The setting

synthesise data from both → do RL between synthetic A and synthetic B OR do RL between augmented A and augmented B

NO → too many 'lures'

- the RL algo return too many synthetic pairs
- the RL algo return nothing (task is too noisy)
- the bimodal distr. of linkage probabilities (certainly non-linked vs. certainly linked) disappear



The setting

synthesise data from B that we concatenate to B $\rightarrow$ *do RL between A and augmented B*

YES

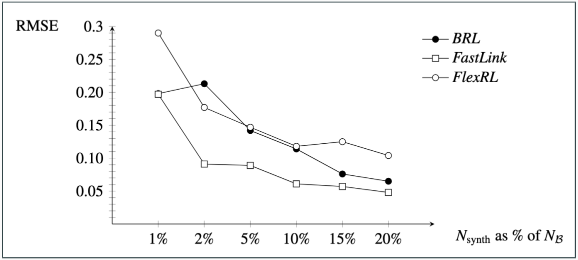
The opposite: *synthesise data from A that we concatenate to A* $\rightarrow$ *do RL between augmented A and B* also works
$\rightarrow$ $N_A \leq N_B$ (clinical data vs. electronic records)
$\rightarrow$ philosophical and practical arguments to eliminate that option

The size

How many synthetic records should we synthesise?

Challenges:

- RL is very slow on large data sets → we do not want to increase the size too much
- RL is less efficient on large data sets (many more potential link to investigate)



Estimation formula

$$\text{FDR} = \frac{FP}{FP+TP} = 1 - \frac{TP}{\text{linked records}}$$

Estimation formula

$$\text{FDR} = \frac{FP}{FP+TP} = 1 - \frac{TP}{\text{linked records}}$$

- estimate FP in the numerator with $\frac{FP_{\text{synth}} \cdot N_B}{N_{\text{synth}}}$
- plug-in $N_{\text{real linked records}}$ in the denominator
- $\frac{FP_{\text{synth}} \cdot N_B / N_{\text{synth}}}{N_{\text{real linked records}}}$

Estimation formula

$$\text{FDR} = \frac{FP}{FP+TP} = 1 - \frac{TP}{\text{linked records}}$$

- estimate FP in the numerator with $\frac{FP_{\text{synth}} \cdot N_B}{N_{\text{synth}}}$
- plug-in $N_{\text{real linked records}}$ in the denominator
- $\frac{FP_{\text{synth}} \cdot N_B / N_{\text{synth}}}{N_{\text{real linked records}}}$

- estimate FP with $\frac{FP_{\text{synth}} \cdot N_B}{N_{\text{synth}}}$
- estimate TP with $N_{\text{linked records}} - \frac{FP_{\text{synth}} \cdot N_B}{N_{\text{synth}}}$
- $\frac{FP_{\text{synth}} \cdot (1+N_B/N_{\text{synth}})}{N_{\text{all linked records}}}$