L'INTEGRAZIONE DI ARCHIVI ELETTRONICI

PER L'EPIDEMIOLOGIA E LA SANITA' PUBBLICA: FINALITA' E METODI

17-18 maggio 2007, Istituto Superiore di Sanità, Roma

PERFORMANCE DELLE PROCEDURE DI LINKAGE TRA ARCHIVI PER L'EPIDEMIOLOGIA

Gruppo di lavoro. C. Fornari,[1] M. Demaria,[2] F. Madotto,[1] P. Pepe,[3] M. Raciti,[3] A. Romanelli,[3] V. Tancioni,[4] P. Trerotoli,[5] GC. Cesana,[1] G. Corrao[1]

[1] Università degli Studi di Milano–Bicocca. [2] Epidemiologia Ambientale – ARPA Regione Piemonte.
[3] CNR, Istituto di Fisiologia Clinica, Pisa. [4] LAZIOSANITA', Agenzia di Sanità Pubblica, Regione Lazio.
[5] Università degli Studi di Bari.

Con la partecipazione di sismec

Background

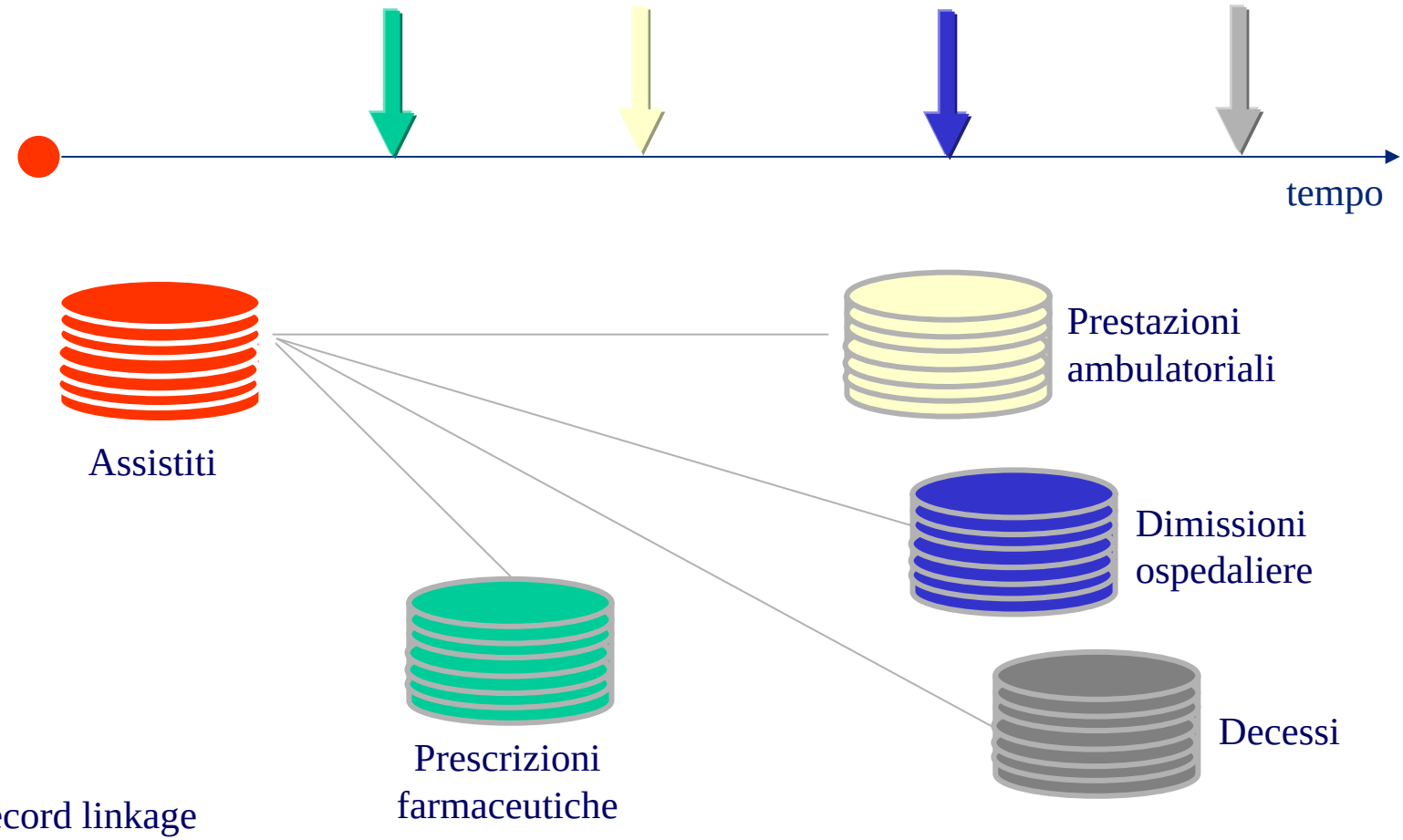
1. Record linkage

"Each person in the world creates a **book of life**. This book starts with birth and ends with death. **Record linkage** is the name of the process of assembling the pages of this book into a volume".[1]

[1] Dunn H.L. Record Linkage. *Am J Publ Health* 1946;**36**:1412–6.

Background

1. Record linkage



Background

- Linkage deterministico
- Linkage semi-deterministico
- Linkage probabilistico

Background

1. Record linkage

2. Strategie di record linkage

2.1. Linkage deterministico

Archivio A

nome e cognome	genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GNN	CRR	M	23/10/1953	020001

Archivio B

nome e cognome	genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GVN	CRR	.	23/10/1953	020001

Regola decisionale: due record appartengono alla stessa unità se, e solo se, tutti i campi della chiave di linkage coincidono.

Background

1. Record linkage

2. Strategie di record linkage

2.1. Linkage deterministico

Archivio A

nome e cognome		genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GNN	CRR	M	23/10/1953	020001

Archivio B

nome e cognome		genere	data di nascita	comune di nascita
CRL	FRN	F	15/01/1967	010330
GVN	CRR	.	23/10/1953	020001

?

La capacità di riconoscere due record come appartenenti alla stessa unità dipende dalla qualità degli archivi.

Background

1. Record linkage

2. Strategie di record linkage

2.1. Linkage semi- deterministico

2.2. Linkage semi-deterministico

Archivio A

nome e cognome	genere	data di nascita	comune di nascita
	FRN	15/01/1967	010330
	CRR	23/10/1953	020001

Archivio B

nome e cognome	genere	data di nascita	comune di nascita
	FRN	15/01/1967	010330
	CRR	23/10/1953	020001

Razionale: alcuni campi vengono esclusi (perché meno informativi o più soggetti ad errori?).

Regola decisionale: due record appartengono alla stessa unità se, e solo se, tutti i campi della chiave di linkage "ridotta" coincidono.

Background

1. Record linkage

2. Strategie di record linkage

2.1. Linkage semi- deterministico

2.2. Linkage semi-deterministico

Archivio A

nome e cognome		genere		data di nascita	comune di nascita
	FRN			15/01/1967	010330
	CRR			23/10/1953	020001

Archivio B

nome e cognome		genere		data di nascita	comune di nascita
	FRN			15/01/1967	010330
	CRR			23/10/1953	020001

La decisione di escludere alcuni campi è arbitraria (non veicolata da alcun processo basato su modelli probabilistici).

Background

1. Record linkage

2. Strategie di record linkage

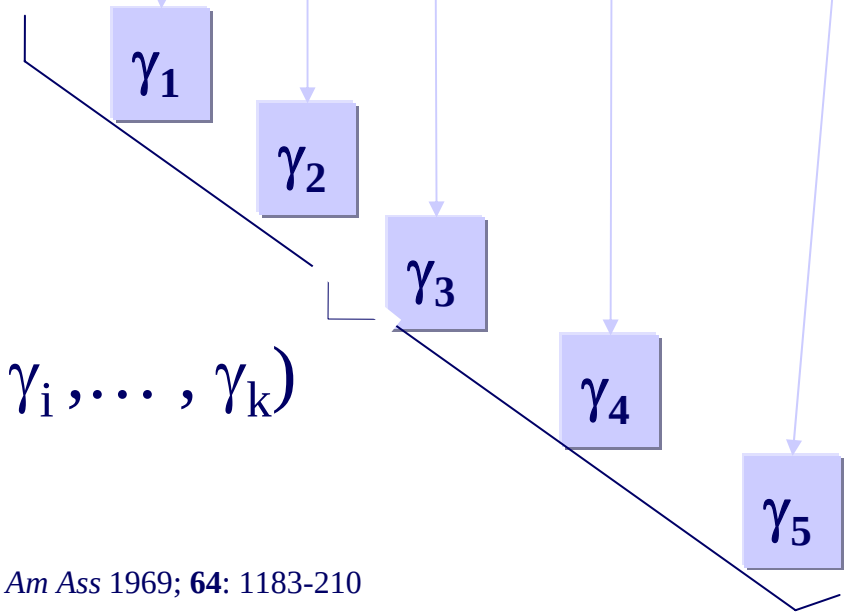
2.1. Linkage deterministico

2.2. Linkage semi- deterministico

2.3. Linkage probabilistico

Razionale:

	nome e cognome		genere	data di nascita	comune di nascita
Archivio A	CRL	FRN	F	15/01/1967	010330
Archivio B	CCL	FRN	.	15/01/1967	010330



$$\mathbf{w} = f(\gamma_1, \gamma_2, \ldots, \gamma_i, \ldots, \gamma_k)$$

Fellegi IP, Sunter AB. A theory for record linkage. *J Am Ass* 1969; **64**: 1183-210

Background

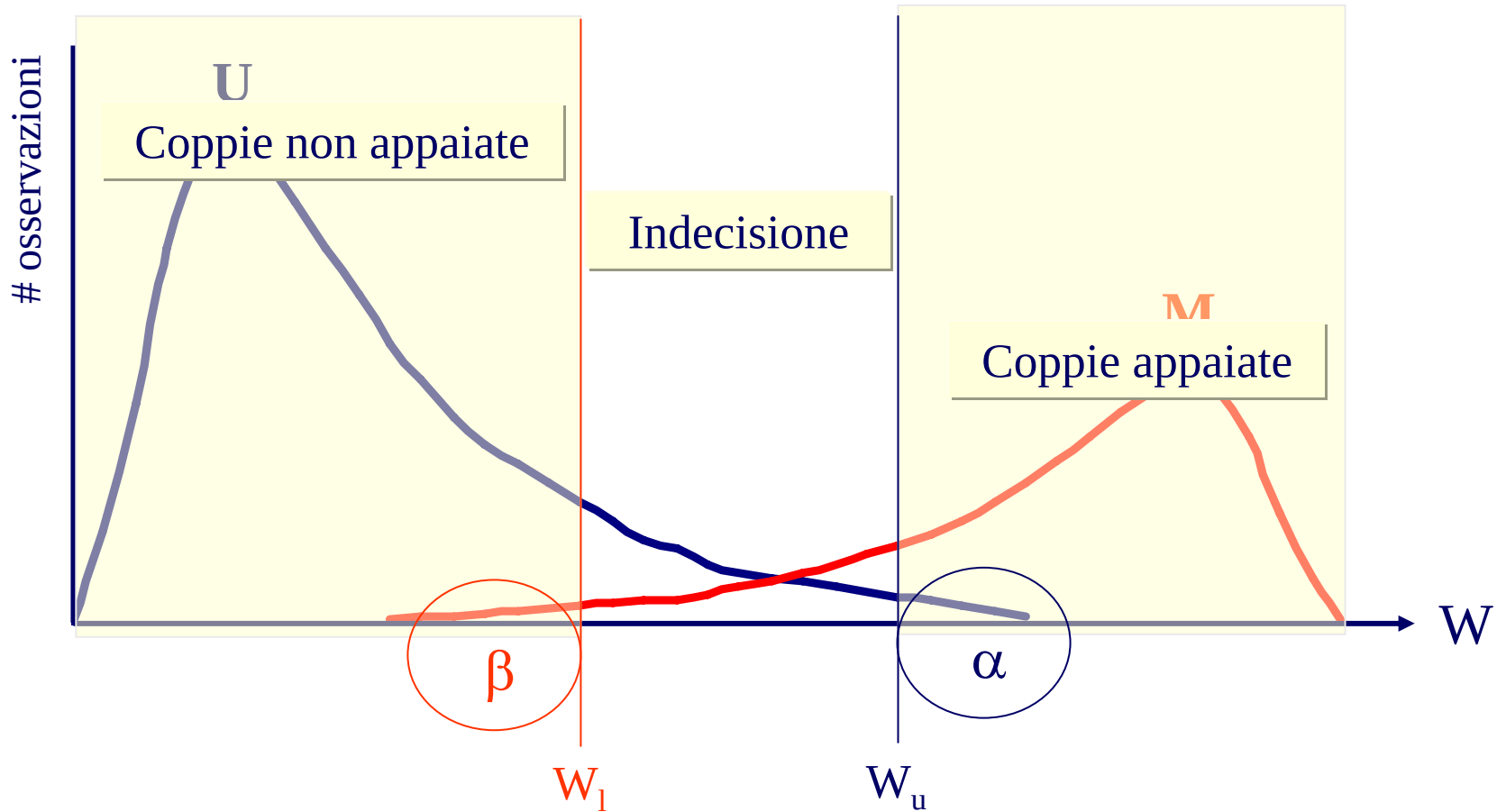
1. Record linkage

2. Strategie di record linkage

2.1. Linkage deterministico

2.2. Linkage semi- deterministico

2.3. Linkage probabilistico



Obiettivo

Valutazione dell'impatto delle strategie di record linkage sulla validità delle misure epidemiologiche generate da un servizio di "epidemiologia e sanità pubblica" locale o regionale.

Metodi

1. Centri partecipanti

Piemonte

TORINO. Epidemiologia Ambientale - ARPA Piemonte -

Pisa

PISA. CNR - Istituto di Fisiologia Clinica-

Roma

ROMA. LAZIOSANITA' - Agenzia si sanità pubblica – Regione Lazio.

Puglia

BARI. Università degli Studi di Bari

Metodi

2. Popolazioni coperte

	Numerosità popolazione coperta	Ricoveri per IMA*
Piemonte	4,313,038	8,668
Pisa	91,212	204
Roma	2,834,419	10,002
Puglia	3,801,120	5,409

* Infarto Miocardico Acuto (selezione AIE-SISMEC): ICD-9: 410 diagnosi principale; 427.1 - 427.41 - 427.42 - 427.5 - 428.1 - 429.5 - 429.6 - 429.71 - 429.79 - 429.81 - 518.4 - 780.2 - 785.51 - 414.10 - 423.0 diagnosi principale, se accompagnati da 410 in almeno una delle diagnosi secondarie)

Metodi

3. Strategie di record linkage

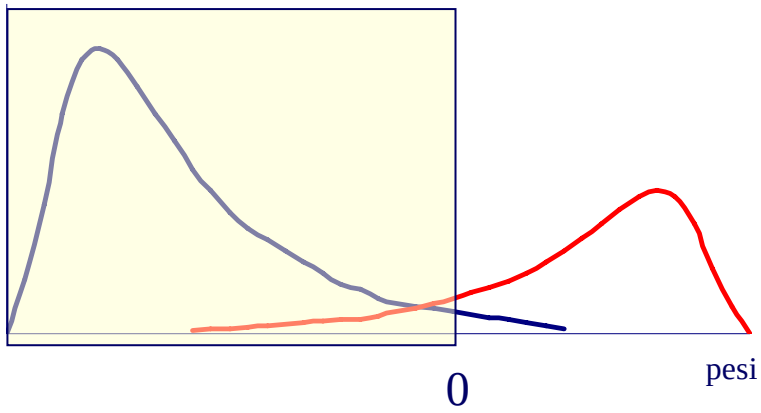
	Chiave di linkage	Procedura di linkage
Piemonte	Codice fiscale	Semi-deterministico con 30 passi
Pisa	Nome, cognome, data di nascita, comune di nascita	Semi-deterministico con 3 passi
Roma	Nome, cognome, sesso, data di nascita, comune di nascita	Semi-deterministico con 7 passi
Puglia	Nome, cognome, data di nascita, comune e provincia di residenza	Deterministico

Metodi

4. Linkage probabilistico

Chiave di linkage		Passo 1
Nome		Nome
Cognome		Cognome
Sesso		Sesso
Giorno nascita	blocco →	Giorno nascita
Mese di nascita		Mese di nascita
Anno di nascita		
Comune di nascita		
Comune di residenza		Comune di residenza

Non appaiati



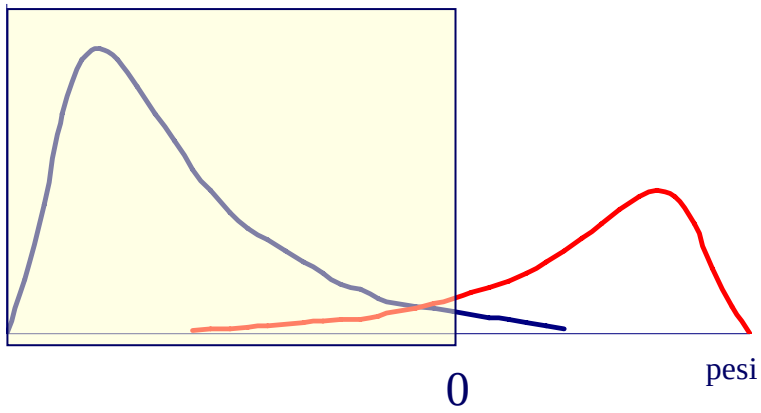
Coppie appaiabili (1)

Metodi

4. Linkage probabilistico

Chiave di linkage		Passo 2
Nome Cognome Sesso Giorno nascita Mese di nascita Anno di nascita Comune di nascita Comune di residenza	blocco	Nome Cognome Sesso Anno di nascita Comune di nascita

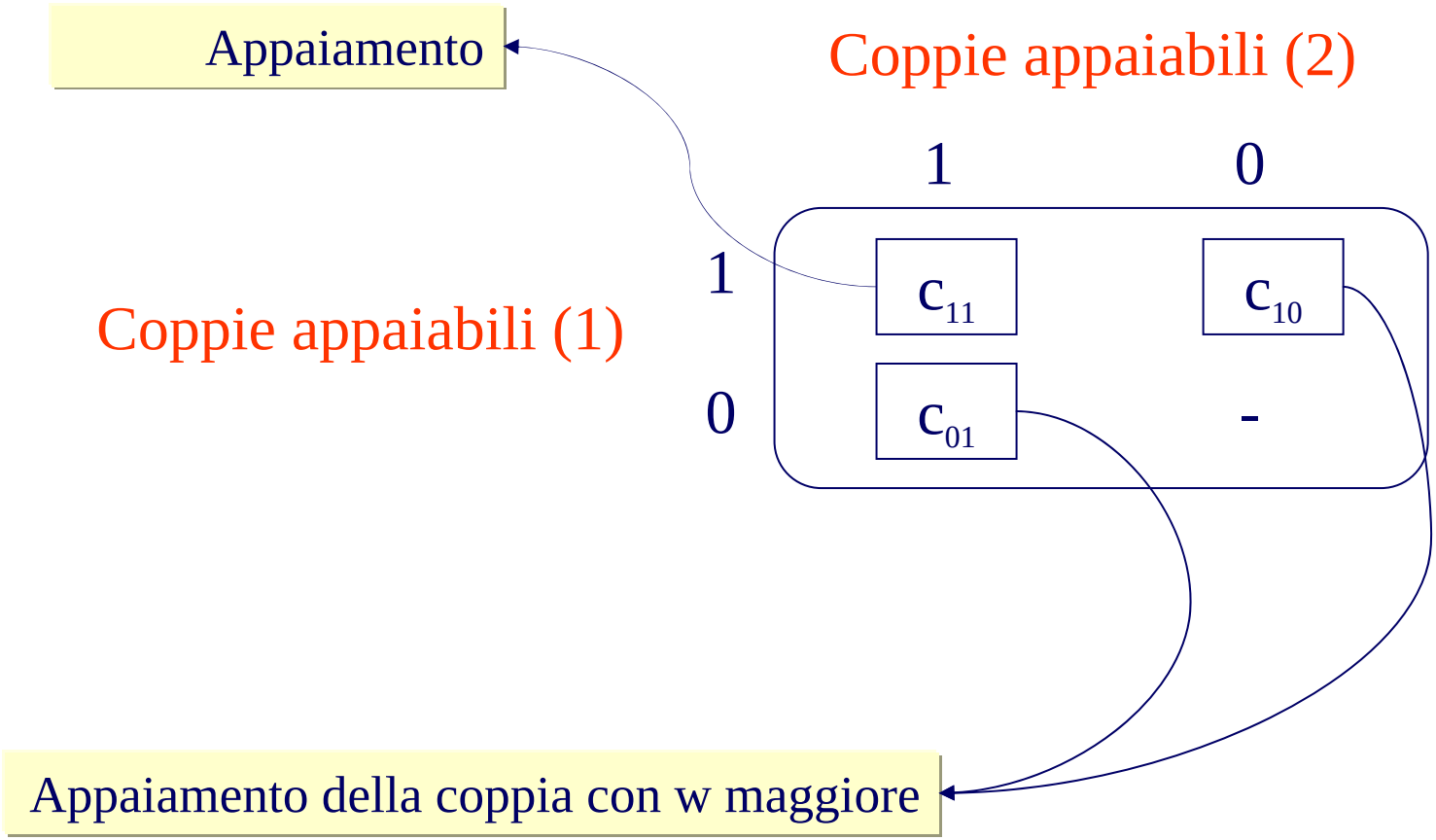
Non appaiati



Coppie appaiabili (2)

Metodi

4. Linkage probabilistico



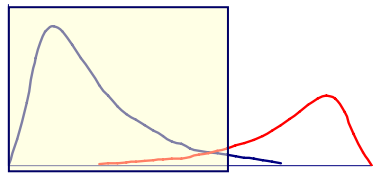
Metodi

4. Linkage probabilistico

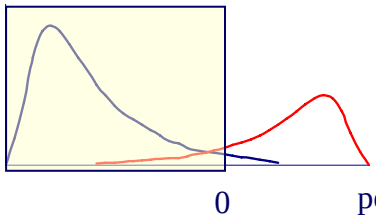
Quali soglie?

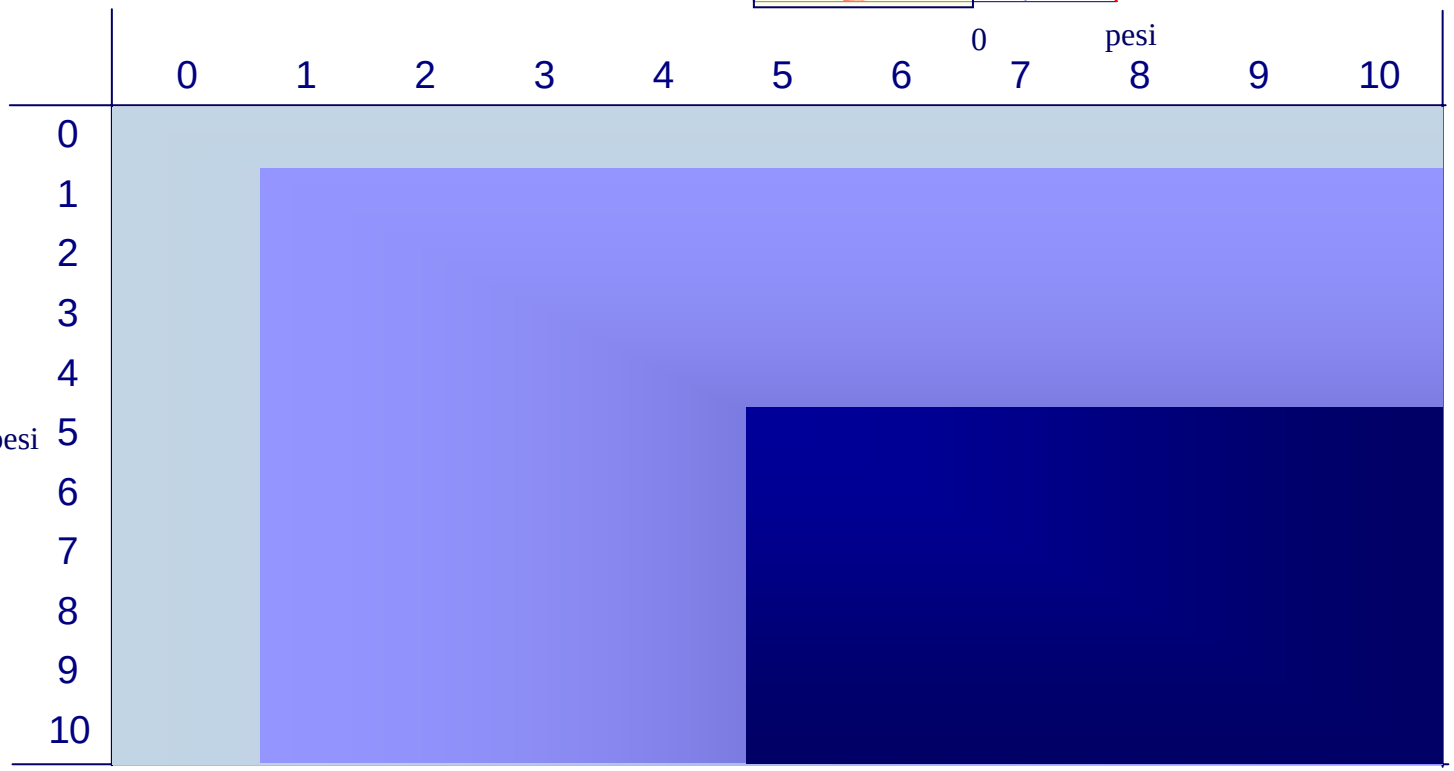
Predittività positiva

Passo 2



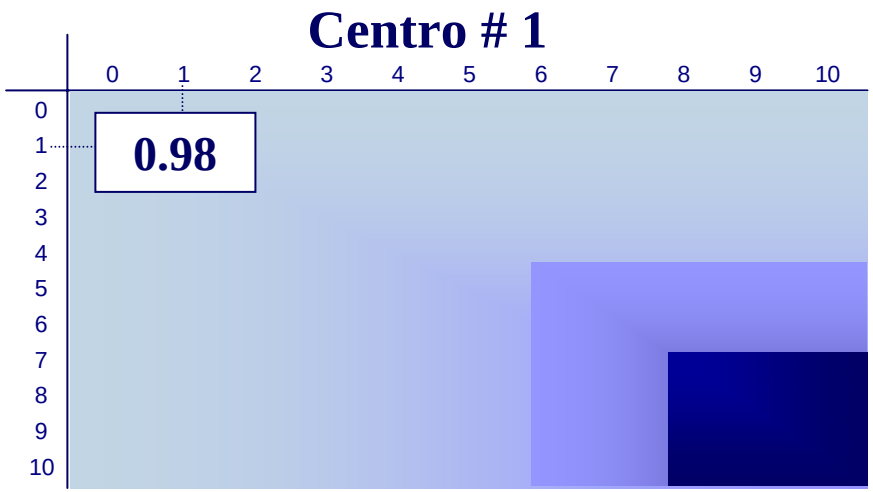
Passo 1

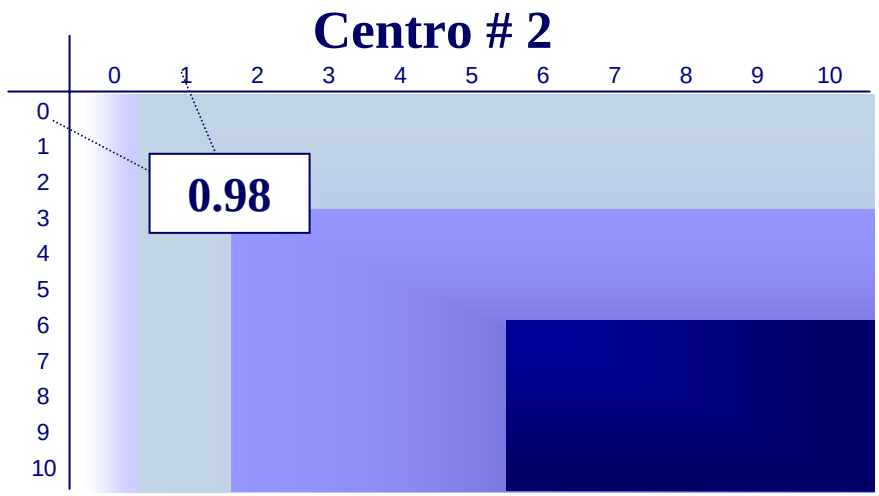


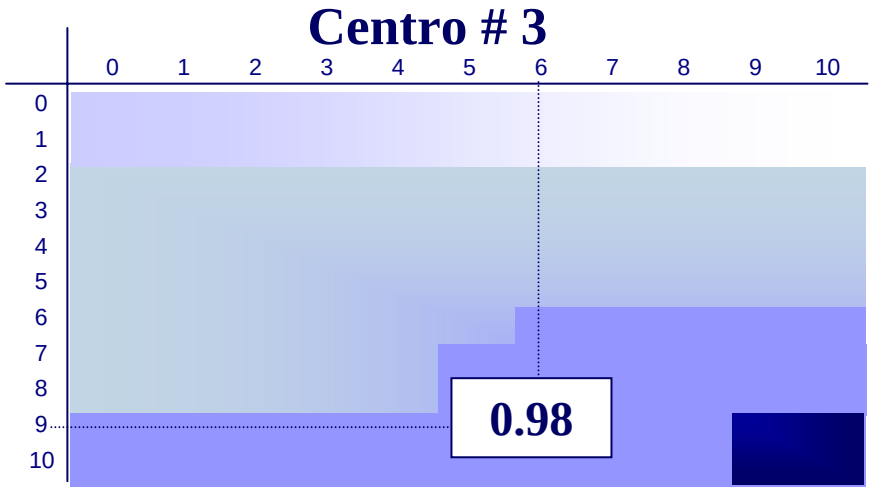


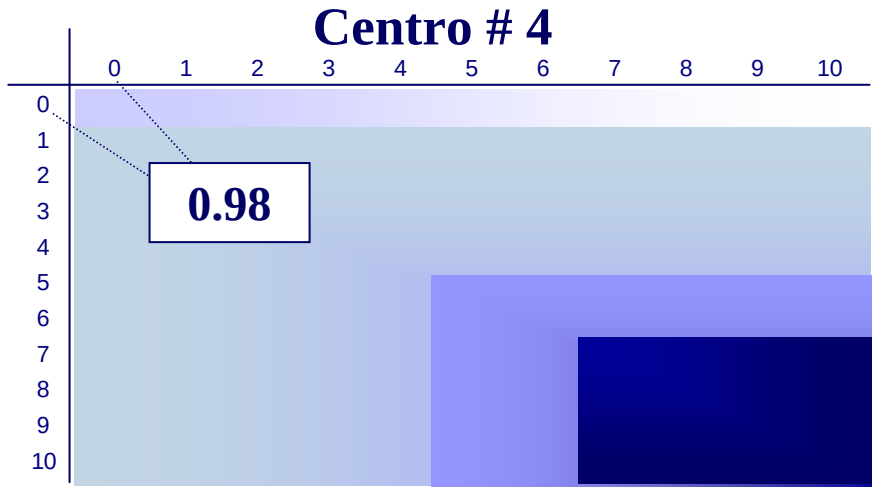
Confronto tra centri

Linkage probabilistico-VPP





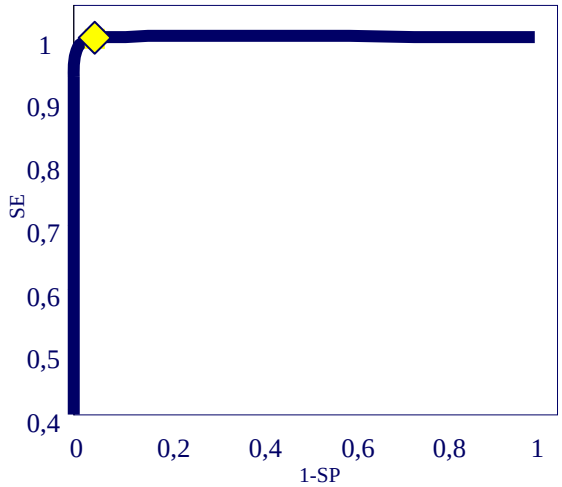




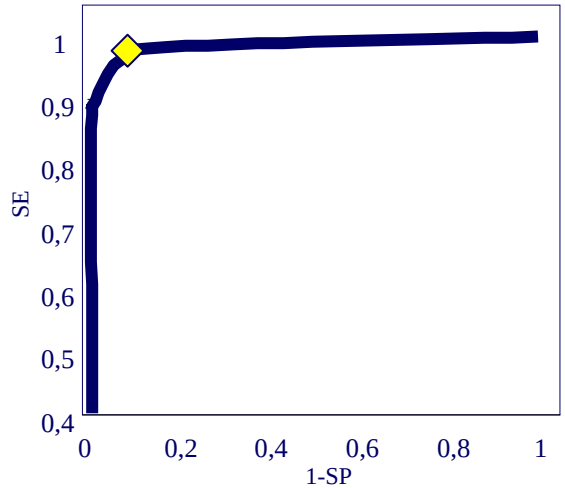
Confronto tra centri

Linkage probabilistico - Curve ROC

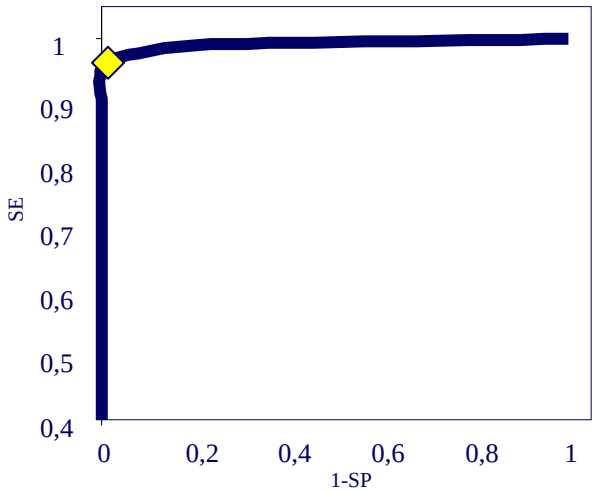
Centro # 1



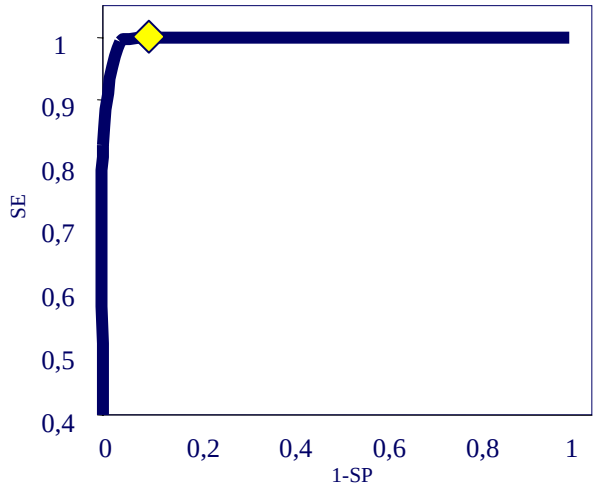
Centro # 2



Centro # 3



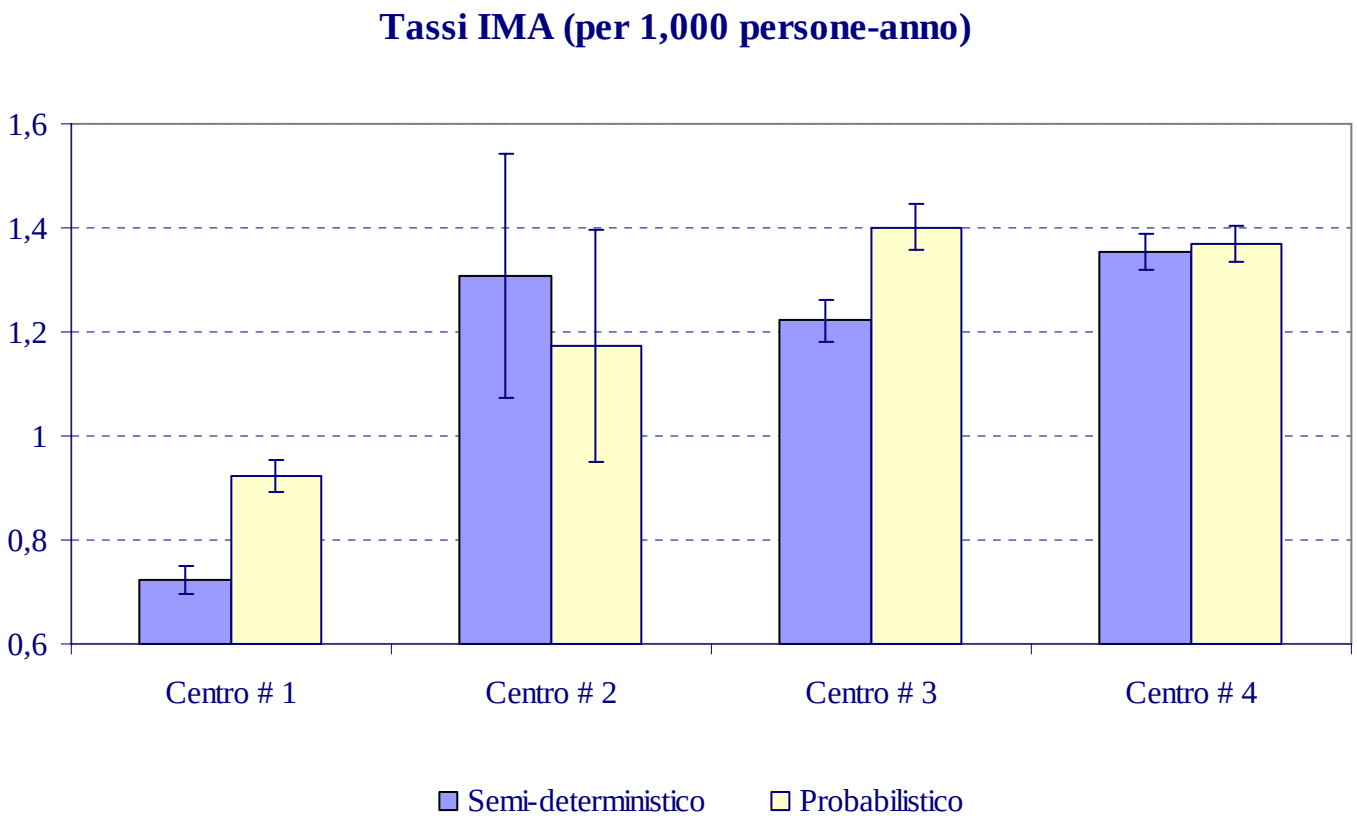
Centro # 4



◇ Soglia prescelta

Confronto tra centri

Impatto strategie di record linkage



Confronto tra centri

Linkage probabilistico – Tempi di esecuzione

	Numerosità popolazione coperta	Ricoveri per IMA	**Tempi macchina**
Piemonte (a)	4,313,038	8,668	2h 30’
Pisa (b)	91,212	204	0h 1’
Roma (c)	2,834,419	10,002	27h 59’
Puglia (d)	3,801,120	5,409	0h 45’

(a) Processore Dual Core 2,8 Ghz, 1GB RAM

(b) Processore 1Ghz, 512 MB RAM

(c) Processore 1Ghz, 512 MB RAM

(d) Processore Dual Core 2 Ghz, 2 GB RAM

Conclusioni

La procedura proposta:

- può migliorare le performance della strategia di linkage
- genera misure epidemiologiche di validità nota e confrontabile
- è sostenibile da un servizio locale o regionale

Problemi aperti

- Privacy
- Implementazione del linkage probabilistico tra archivi completi
- Implementazione del linkage probabilistico tra molteplici archivi
- Duplicati all'interno degli archivi