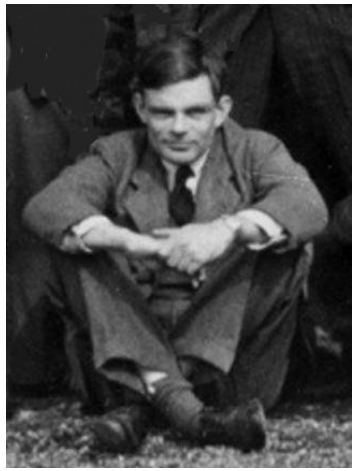
# Algoritmi: il soffio dell'intelligenza

Roberto Raschetti

Alan Turing

**Finalità**: non imbarcarsi nella teoria matematica degli algoritmi, ma sottolineare l'importanza di una adeguata documentazione

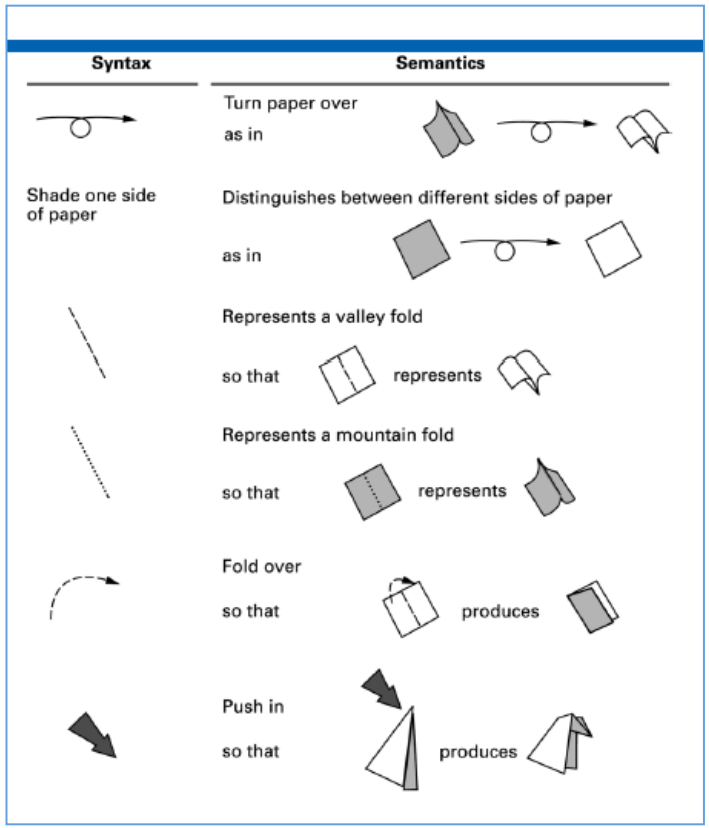
*Il termine **algoritmo** ha origine nel Medio Oriente. Essa proviene dall'ultima parte del nome dello studioso persiano Abu Jµafar Mo-hammed Ibn Musa Al-Khowarizmi, il cui testo di aritmetica (825 d.C. circa) esercitò una profonda influenza nei secoli successivi.*

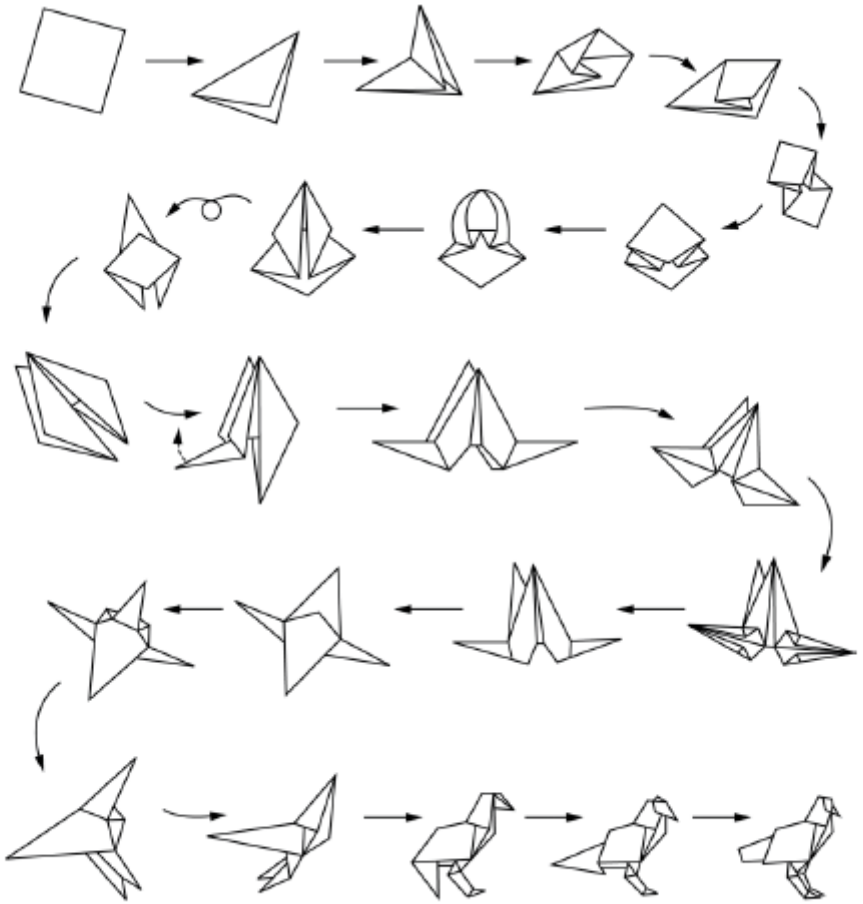
# Una definizione generale

Un algoritmo è una sequenza di azioni che consente di pervenire alla soluzione di un problema mediante una sequenza finita di operazioni, completamente e univocamente determinate.


Un procedimento risolutivo che riceve dei dati in ingresso (input) esegue una qualche elaborazione e restituisce il risultato della trasformazione (output).

# Un esempio: Origami



# Un esempio: Il nodo papillon

Tappa n°1: posizionate le due estremità del papillon in modo asimmetrico, una più bassa dell'altra.

Tappa n°2: attorno al collo, incrociate l'estremità più lunga su quella più corta.

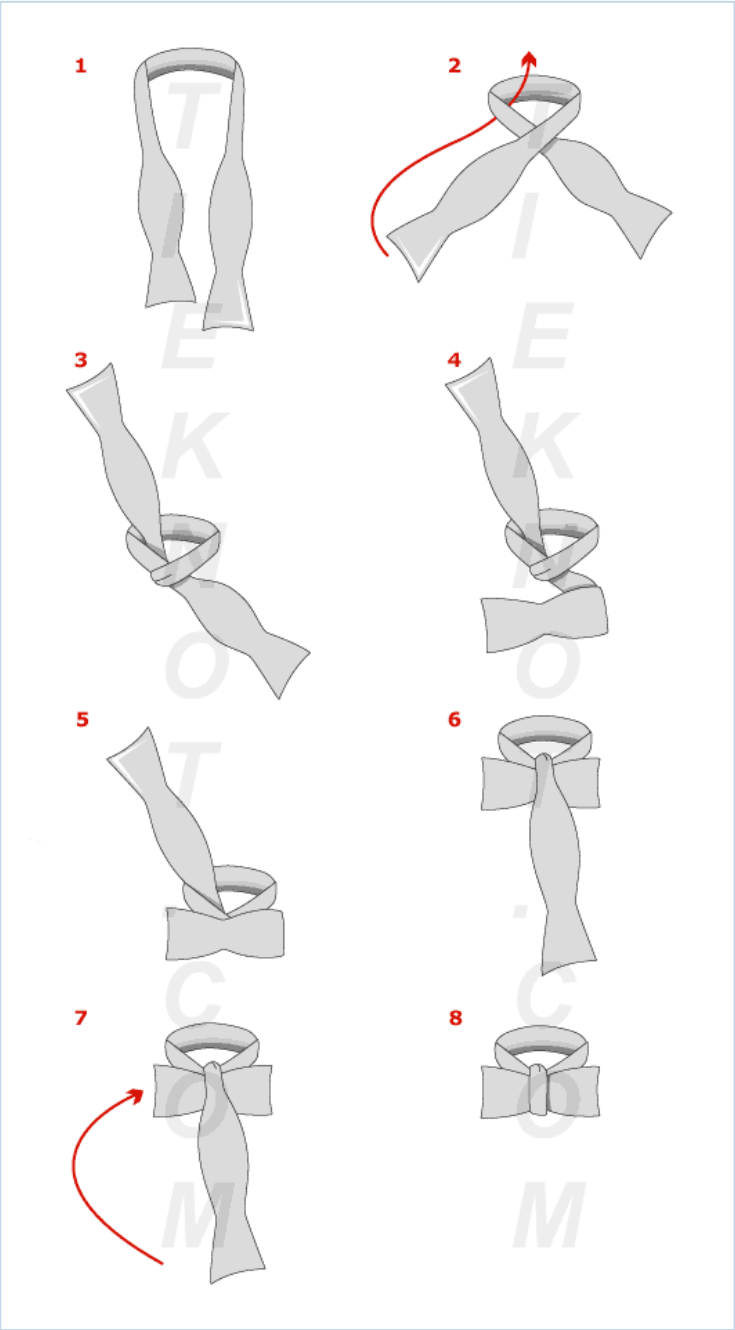
Tappa n°3: fate scivolare l'estremità più lunga verso l'alto, sotto il papillon.

Tappe n°4/5: formate le due ali del papillon, piegando orizzontalmente l'estremità più corta.

Tappa n°6: fate scendere l'estremità più lunga davanti al nodo in formazione.

Tappa n°7: nascondete poi l'estremità più lunga sotto l'estremità piegata.

Tappa n°8: aggiustate il papillon tirando sulle due ali.



# Proprietà di un algoritmo

L' algoritmo deve essere:

**Finito**, costituito cioè da un numero limitato di passi e l'esecuzione deve avere termine dopo un tempo finito (*terminazione*);

**Definito** (non-ambiguo), i passi costituenti devono essere interpretabili in modo diretto e univoco dall'*esecutore*, sia esso umano o artificiale;

**Effettivo** l'esecuzione deve portare ad un risultato univoco ;

**Eseguibile**, cioè la sua esecuzione deve essere possibile con gli strumenti di cui si dispone;

**Deterministico**, ad ogni passo deve essere definita una ed una sola operazione successiva.

# Rappresentazione di algoritmi

Lo stesso algoritmo può essere rappresentato in vari modi (diagrammi, testo, ecc) e  a diversi livelli di astrazione.

Ogni rappresentazione si deve basare su un insieme di operazioni primitive ben definite, comprensibili all'esecutore.

# Linguaggi di descrizione di algoritmi

Le operazioni primitive sono:

- La sequenza
- La selezione condizionale
- La iterazione

Due tipici esempi di linguaggi semiformali sono:

- schemi di flusso (diagrammi di flusso, flow-chart)
- pseudo-codice

# Schemi di flusso e pseudo-codici

## Schemi di flusso (diagrammi di flusso, flow-chart)

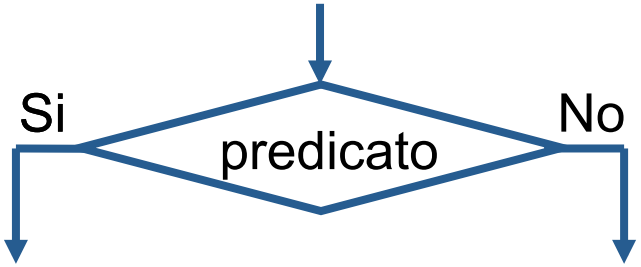
Sono un formalismo descrittivo che fa uso di elementi grafici per indicare il flusso di controllo e per indicare i tipi di operazioni.
Semplici da usare e facilmente leggibili, in particolare per algoritmi di limitata complessità e soprattutto nelle fasi di bozza iniziale della descrizione, o anche come documentazione esplicativa di un programma.

## Pseudo-codici

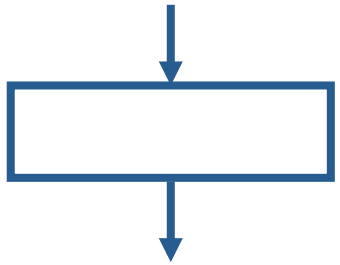
La rappresentazione con pseudo-codice è completamente testuale.
I costrutti di controllo sono descritti con la forma e le parole chiave simili a quelle dei linguaggi di programmazione, mentre le operazioni possono essere descritte in modo informale e sintetico.

# Diagrammi di flusso

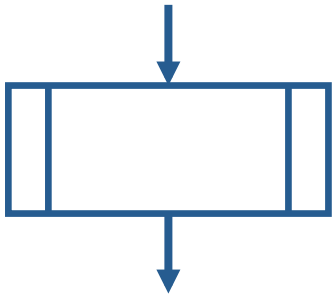
Start/Fine




Selezione condizionale



Elaborazione



Sottoprocedura

Nodo connettore

# Pseudocodifica

Un possibile formalismo per la descrizione di un algoritmo utilizza un sottoinsieme del linguaggio naturale. Utilizza termini del tipo:

INIZIO/FINE

LEGGI (ACQUISISCI)/SCRIVI

SE …ALLORA….ALTRIMENTI

MENTRE…ESEGUI

# Progettazione top-down

Quando il problema è articolato può non essere immediato trovare subito i passi elementari per la sua soluzione.

Una tecnica che semplifica questo compito è la tecnica per *raffinamenti successivi*.
Il problema viene suddiviso in sottoproblemi da risolvere.

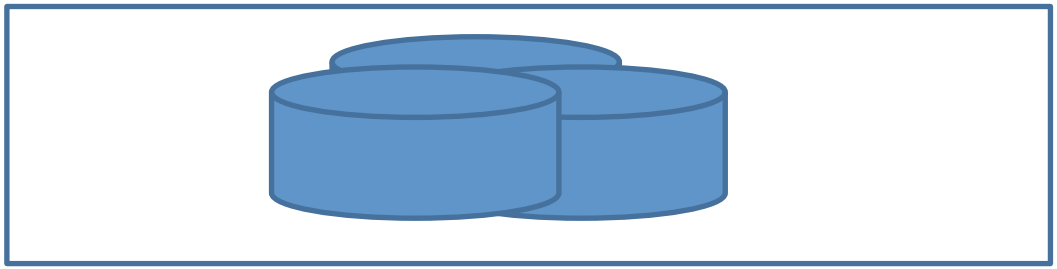
Tale procedimento di scomposizione è detta tecnica top-down.

Pianificazione

**Criteri**

**Algoritmo**

**Fonti di dati sanitari correnti**



Selezione

```
SELECT ...
    FROM …
    WHERE ….
    OR (….AND ….)
ORDER BY ….
```

**Data set selezionato**

Incidenza
Prevalenza
Rischi relativi

Stima

# Un cenno alle misure epidemiologiche

# Prevalenza

Proporzione di individui nella popolazione in studio che, in un certo istante o in un certo intervallo di tempo, possiede la caratteristica di interesse.

**Prevalenza puntuale**

$$P = \frac{\text{N di individui con la caratteristica in un dato istante}}{\text{Popolazione totale}}$$

**Prevalenza di periodo**

$$P = \frac{\text{N di individui con la caratteristica in un dato periodo di tempo}}{\text{Popolazione totale}}$$

# Incidenza Cumulativa (Rischio - Risk )

Probabilità di sviluppare l'evento in studio in un definito intervallo di tempo.

Si calcola come la proporzione di individui, inizialmente privi dell'evento, che sviluppano l'evento durante il periodo di osservazione

$$IC_{t1\text{-}t0} = \frac{\text{N° di nuovi casi}_{\,t1\text{-}t0}}{\text{Popolazione priva dell'evento a t0}}$$

# Densità di incidenza (tasso - rate)

Numero di eventi che si verificano nel corso del periodo di osservazione sulla massa a rischio (tempo-persona) costituita dalla somma dei periodi individuali di tempo a rischio

$$DI = \frac{\text{N° di nuovi casi}}{\text{Tempo persona}}$$

Descrive quanto rapidamente avverranno i nuovi eventi nella popolazione

Si esprime in unità di tempo

$0 \leq \text{Rate} \leq \infty$

# Effetto relativo – rapporto fra incidenze

$$RR = \frac{I_t}{I_{nt}}$$

**Rischio relativo**

**Rapporto di tassi**

**Rate ratio**

**Rapporto di rischi**

**Risk ratio**

Equivale a misurare l'incremento (o il decremento) di incidenza negli esposti in multipli dell'incidenza nei non esposti.

Il valore di un effetto relativo dipende dal valore dell'incidenza di riferimento.

# Esempi di algoritmi che operano sulle fonti di dati correnti

| Categoria diagnostica | Fonti |
|---|---|
| diabete | SDO, PF, ET |
| cardiopatia ischemica | CM, SDO, PF, ET |
| IMA | CM, SDO |
| ictus acuto | CM, SDO |
| asma | CM, SDO, PF, ET |
| BPCO | CM, SDO |
| MPCO | CM, SDO, PF, ET |

CM: cause di morte; *causes of death*
SDO: schede di dimissione ospedaliera; *hospital discharges*
PF: prescrizioni faramaceutiche; *drug prescriptions*
ET: esenzione ticket; *health-tax exemption*

# Un esempio: stima delle amputazioni nelle persone con diabete

Schede di Dimissione Ospedaliera (SDO) italiane dal 2001 al 2007

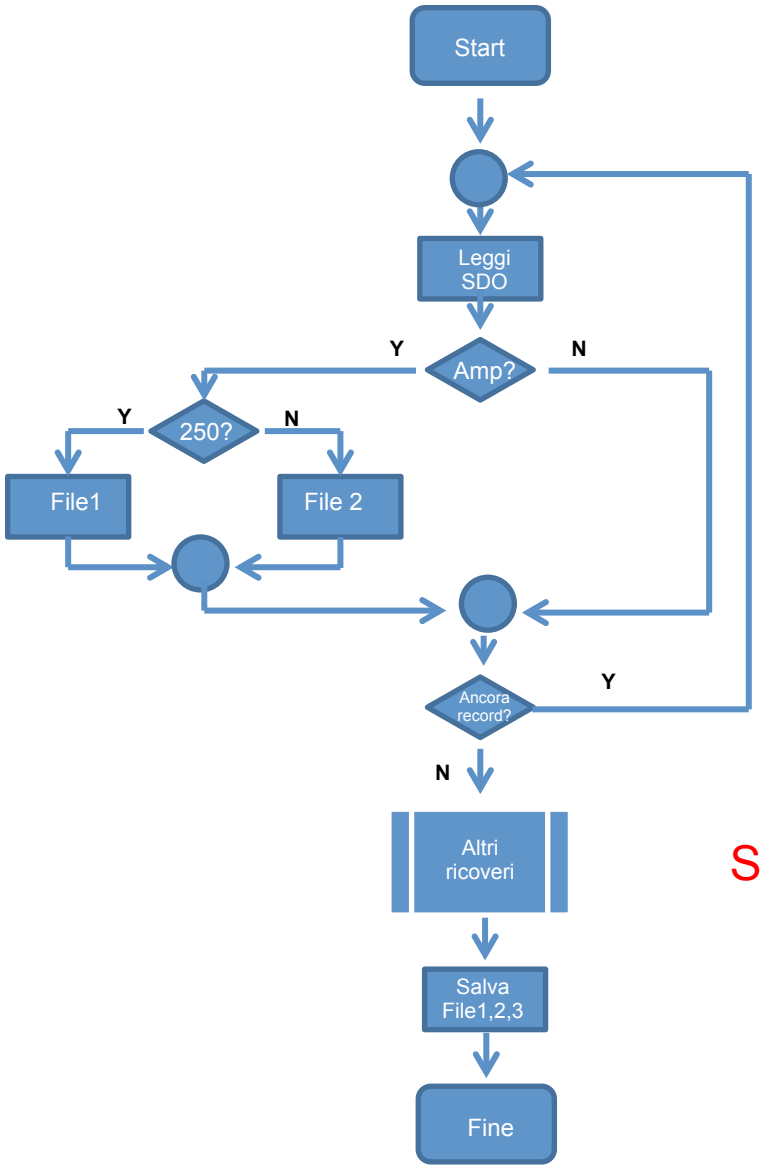
**Persone con diabete**: codice ICD9 250 in diagnosi principale o secondaria.
Sono state escluse le pazienti con diabete gestazionale (ICD9 648.80-648.84).

**Intervento di amputazione** agli arti inferiori: codice 841 in intervento principale o secondario, con esclusione delle amputazioni per traumatismi e neoplasie.

Amputazioni minori: dita del piede o piede (ICD9 84.11-84.12)

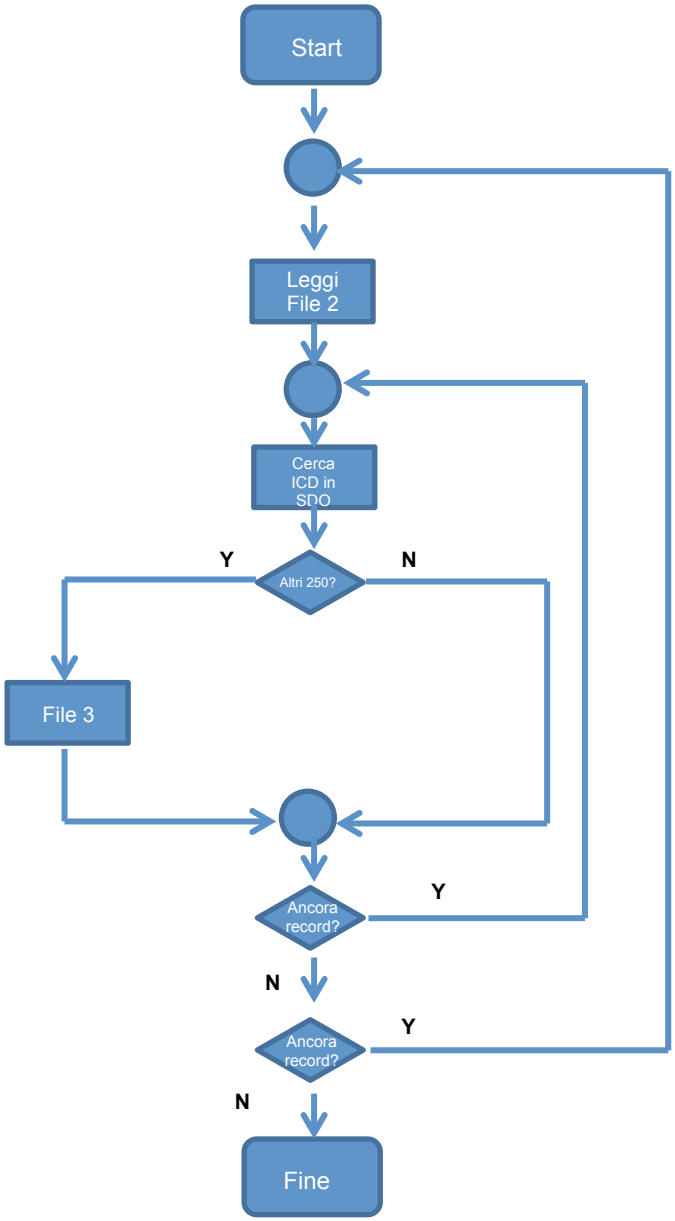
Amputazioni maggiori: sopra il livello del piede (ICD9 84.13-84.19)

**Amputazione in persona con diabete**:
intervento di amputazione all'arto inferiore di paziente ricoverato con diagnosi di diabete nello stesso anno ( cod 250 nel ricovero dell'intervento o in altro ricovero nello stesso anno).




Sotto procedura

# Sotto procedura
# Altri ricoveri



# Stima dell'incidenza dell'infarto miocardico acuto basata su dati sanitari correnti mediante un algoritmo comune in differenti aree italiane

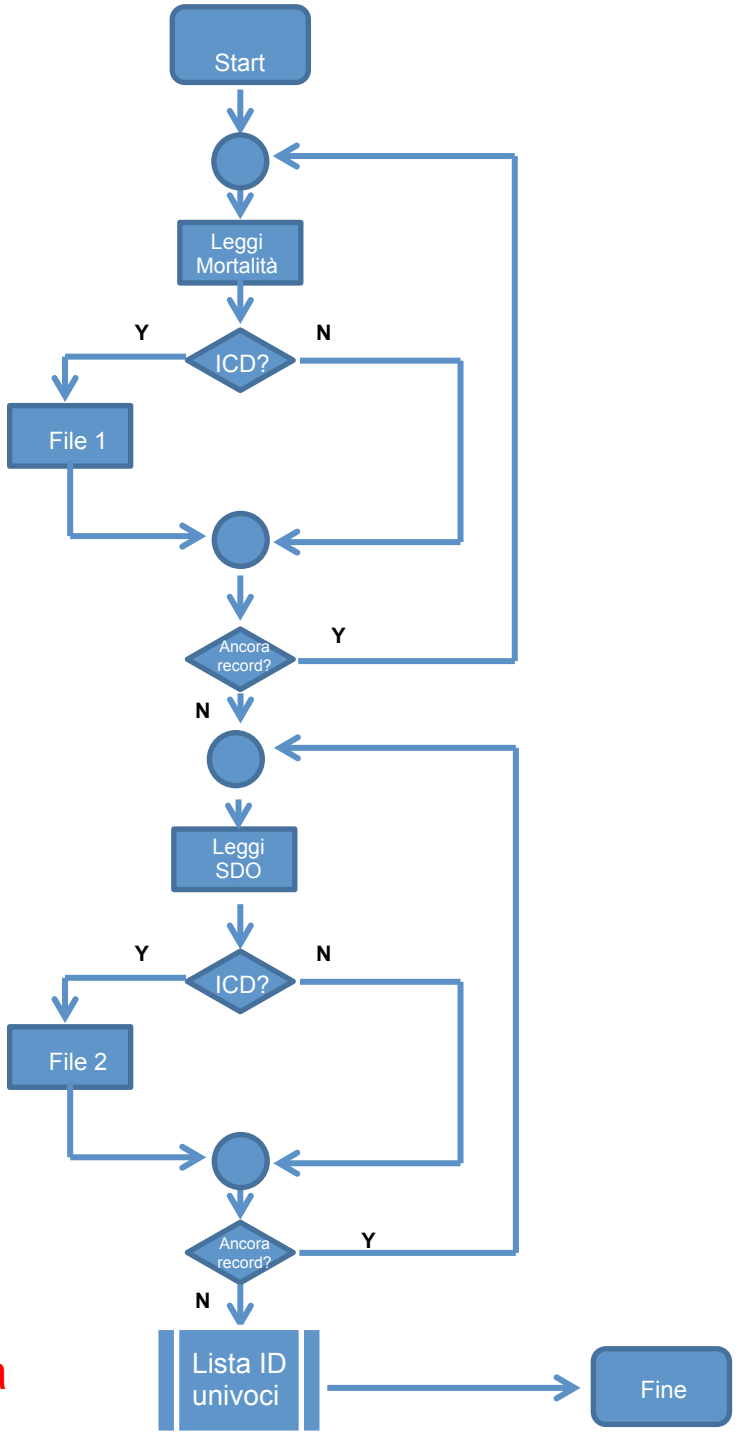
Alessandro Barchielli,[1] Daniela Balzi,[1] Antonella Bruni,[2] Cristina Canova,[3] Giulia Cesaroni,[4] Roberto Gnavi,[5] Roberta Picariello,[5] Andrea Inio,[6] Mariangela Protti,[7] Anna Romanelli,[7] Roberta Tessari,[3,6] Maria Angela Vigotti,[8] Lorenzo Simonato[3]

| Fonte | Criteri di selezione casistica | Criteri per la definizione di incidenza |
|---|---|---|
| mortalità | decesso per infarto miocardico acuto in diagnosi di morte principale | assenza di altri ricoveri con diagnosi di dimissione principale o secondaria con i codici ICD9-CM 410* o 412* (infarto miocardico pregresso), nei 5 anni precedenti alla data di ammissione o di morte |
| schede di dimissione ospedaliera | ricovero ordinario per infarto miocardico acuto (ICD9-CM1: 410*) in diagnosi di dimissione principale o secondaria, se associata ad alcuni specifici codici in diagnosi principale^ | |

ICD9-CM: Classificazione internazionale delle malattie – 9ª revisione – con modifiche cliniche; *International classification of diseases – 9th revision – clinical modifications*

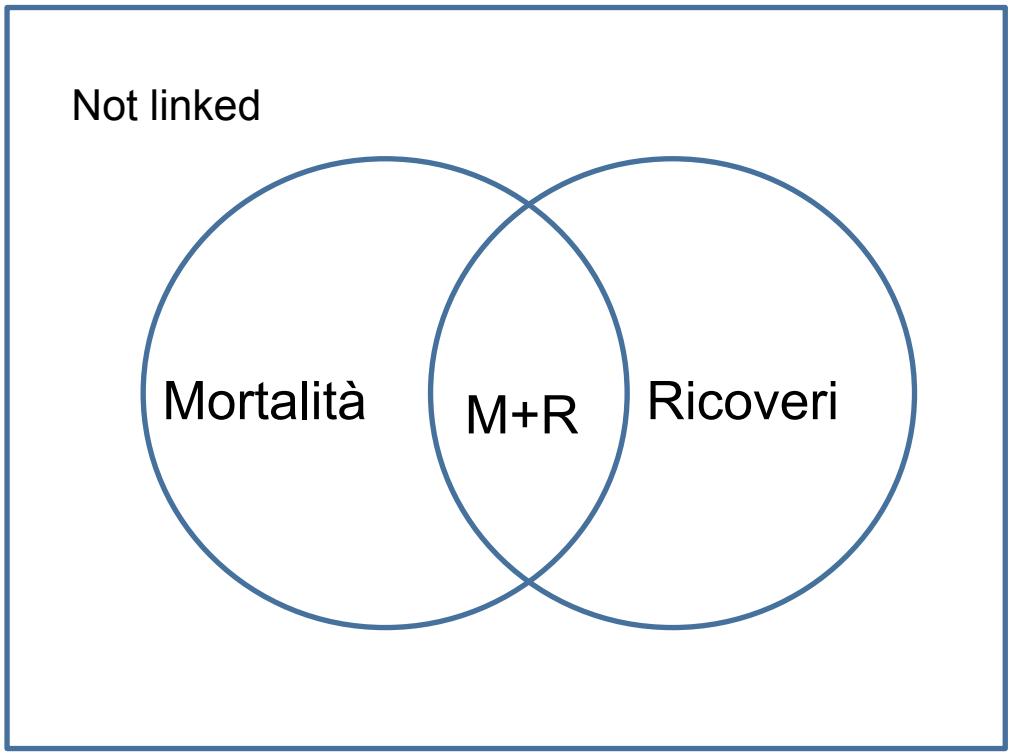
^ codici in diagnosi di dimissione principale: 427.1, 427.41, 427.42, 427.5, 428.1, 429.5, 429.6, 429.71, 429.79, 429.81, 518.4, 780.2, 785.51, 414.10, 423.0; *codes in principal diagnosis: 427.1, 427.41, 427.42, 427.5, 428.1, 429.5, 429.6, 429.71, 429.79, 429.81, 518.4, 780.2, 785.51, 414.10, 423.0*

*Tabella 1. Fonti dei dati e criteri di definizione di caso incidente di infarto miocardico acuto.*

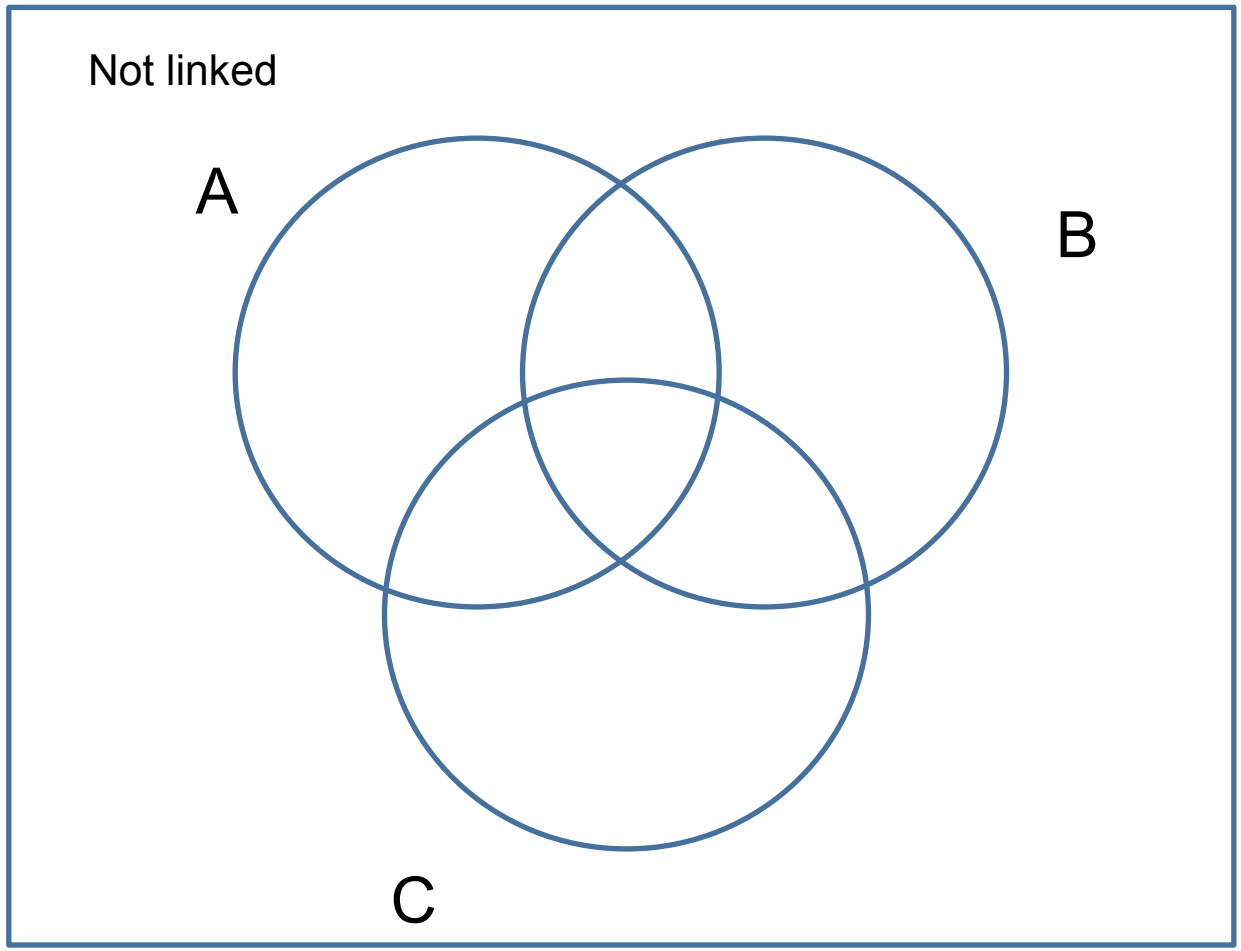



Sotto procedura

# I diagrammi di Venn

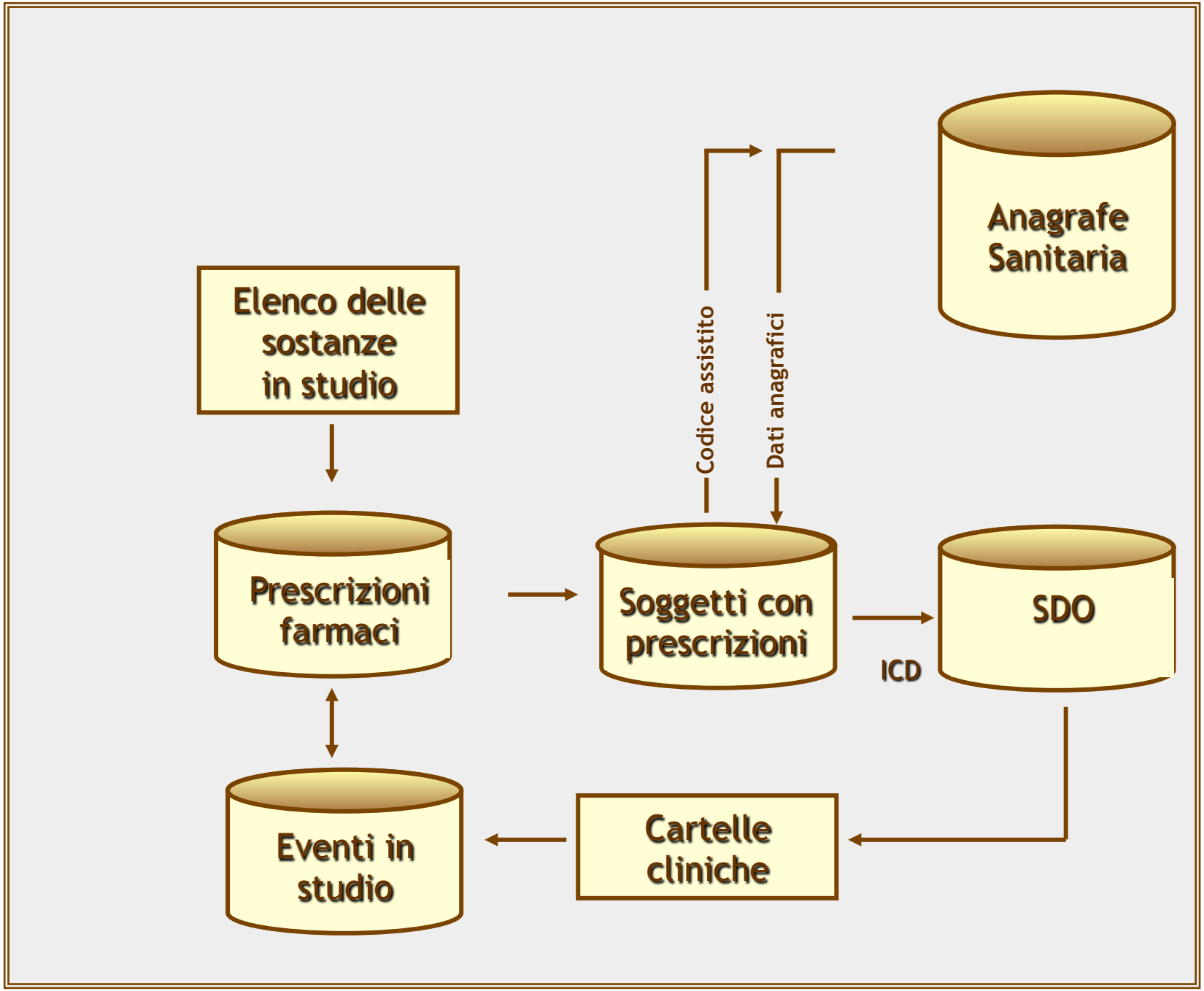


# I diagrammi di Venn (3 fonti)




$2^K - 1$ partizioni

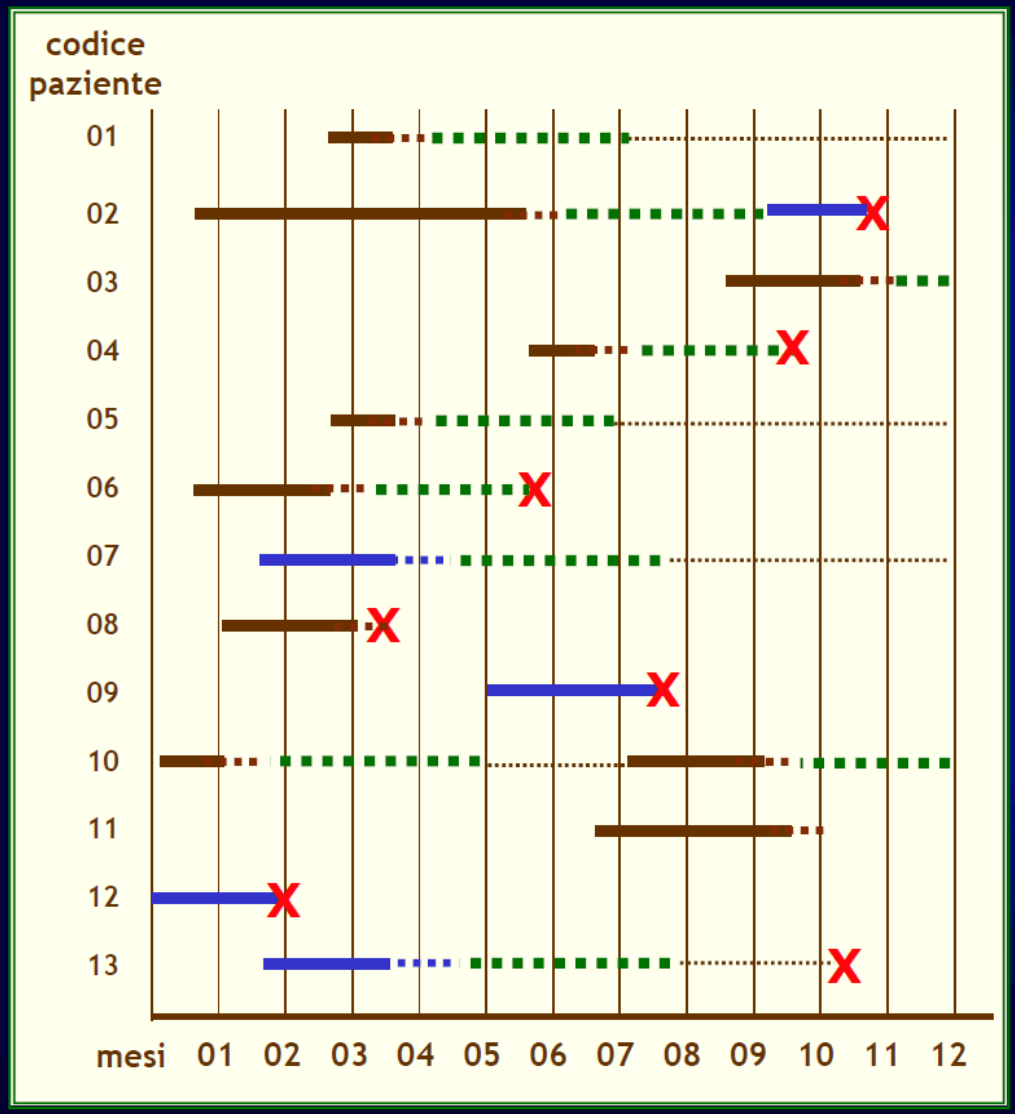
# Applicazioni nella Farmacoepidemiologia


Anagrafe Sanitaria
Elenco delle sostanze in studio
Codice assistito
Dati anagrafici
Prescrizioni farmaci
Soggetti con prescrizioni
SDO
ICD
Eventi in studio
Cartelle cliniche


Esposizione ai trattamenti in studio

Periodi di non trattamento

# Matrici temporali di esposizione




**farmaco A**
**farmaco B**

uso corrente

uso recente

non uso

X evento


# Tecniche di cattura - ricattura

# L'idea della cattura-ricattura


Si estrae dalla popolazione un campione di $n_1$ soggetti che vengono poi "marchiati" per consentirne una successiva identificazione e reintrodotti nella popolazione. Dopo un periodo di tempo, tale da consentire ai soggetti marchiati e non di mischiarsi, si estrae un campione di $n_2$ individui di cui $m$ risultano marchiati. Ipotizzando che la proporzione di marchiati nel secondo campione sia una stima ragionevole proporzione non nota nella popolazione si possono eguagliare i due termini ed ottenere una stima di $N$:

$$\frac{m}{n_2} = \frac{n_1}{N}$$

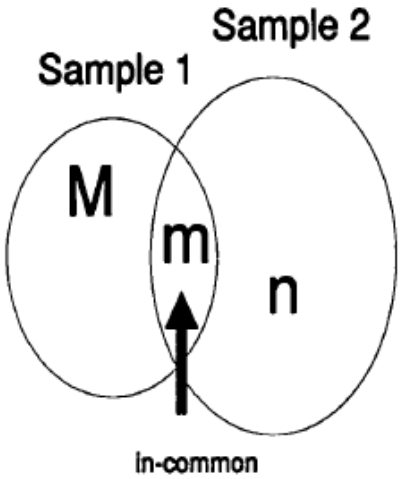
# Counting Diabetes in the Next Millennium

## Application of capture-recapture technology

RONALD E. LAPORTE, PHD
DANIEL MCCARTY, MS
GRAZIELLA BRUNO, MD
NAOKO TAJIMA, MD
SHIGEAKI BABA, MD

DIABETES CARE, VOLUME 16, NUMBER 2, FEBRUARY 1993


### *Capture-Recapture*




$$N = \frac{(M+1)(n+1)}{(m+1)} - 1$$

N=Estimate of Number
M=Number in First Sample (those Marked)
n=Number in Second Sample
m=Number of "marked" items in Second Sample

$$Var(N) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2 (m+2)}$$

$$95\%\ CI = \pm 1.96\sqrt{Var(N)}$$

**Appendix**—*Formula for capture-recapture technology.*

# I presupposti

- la popolazione sia chiusa, cioè N sia costante;
- tutti gli individui abbiano la stessa probabilità di far parte del campione;
- la marchiatura non modifichi la probabilità di cattura dei soggetti;
- le fonti devono essere indipendenti.

# Cattura-ricattura con J fonti di dati

Ad ogni soggetto “catturato” almeno una volta viene associato un vettore di risposta:

$$R=[\, r_1, r_2, \ldots .. r_j]$$

Con $r_k = 1$ se il soggetto è stato individuato dalla fonte k,
0 altrimenti

| | $r_1 = 0$ | | $r_1 = 1$ | |
|---|---|---|---|---|
| | $r_2 = 0$ | $r_2 = 1$ | $r_2 = 0$ | $r_2 = 1$ |
| $r_3 = 0$ | $y_{000}\ (?)$ | $y_{010}$ | $y_{100}$ | $y_{110}$ |
| $r_3 = 1$ | $y_{001}$ | $y_{011}$ | $y_{101}$ | $y_{111}$ |

# Modelli di analisi

- Modelli log-lineari (Fienberg, 1972, Cormack, 1989);
- Modello a classi latenti (Cowan e Malec, 1986);
- Modello di Rasch nella versione a classi latenti (Darroch, 1993, Agresti, 1994).