



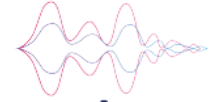
**Ministero
dell'Università
e della Ricerca**



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



PNC
Piano nazionale per gli investimenti
complementari al PNRR



Anthem
Advanced Technologies For Human-centred Medicine

Cleaning e analisi dei dati: il punto di vista statistico

Bergamo, 21 Novembre 2024

Stefania Galimberti

Bicocca Bioinformatics Biostatistics and Bioimaging B4 Centre

Dipartimento di Medicina e Chirurgia

Università di Milano-Bicocca



**Fondazione IRCCS
San Gerardo dei Tintori**

Uno dei mantra degli statistici

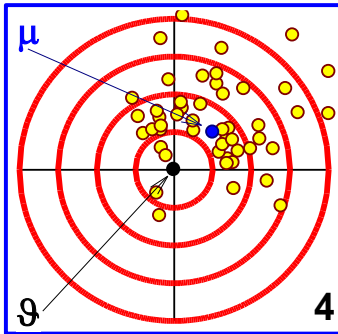


Rilevanza, integrità e qualità del dato

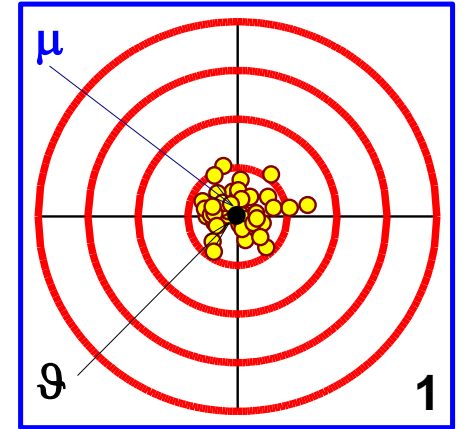
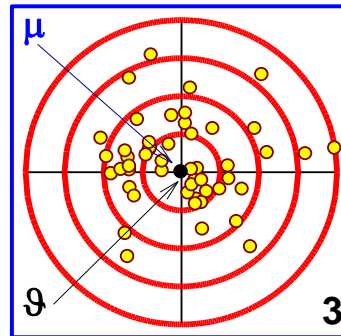
I rischi

Obiettivo di ogni ricerca è ottenere risultati **accurati e precisi**

Evitare
errori
sistematici



Minimizzare
errori casuali



non si possono trarre conclusioni affidabili da dati inadeguati

I Principi

- ✓ **Accuratezza**

I dati devono riflettere correttamente i fenomeni che rappresentano

- ✓ **Completezza**

I dati devono includere tutte le informazioni necessarie per il loro utilizzo

- ✓ **Aggiornamento**

I dati devono essere aggiornati in modo opportuno

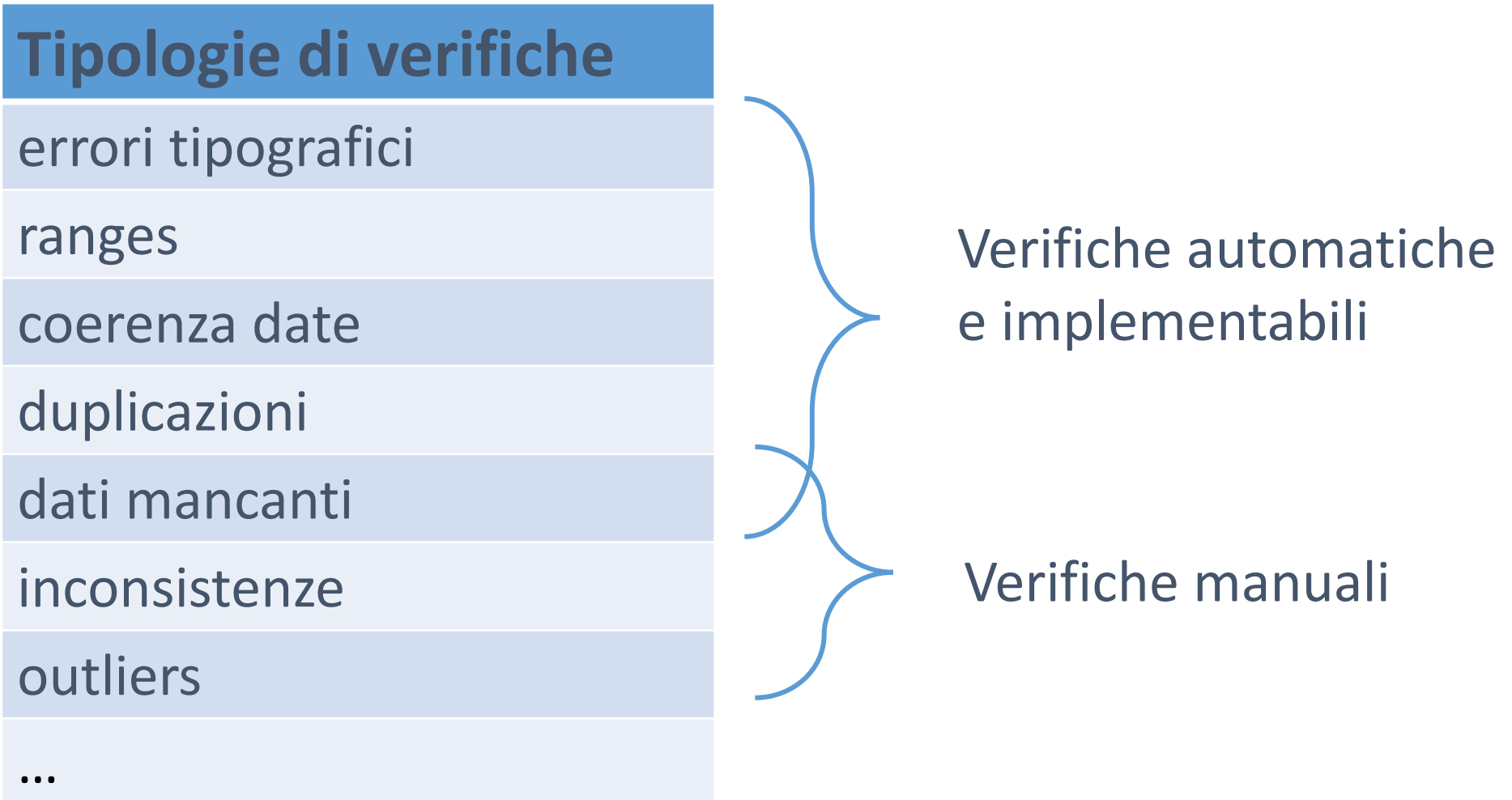
- ✓ **Affidabilità**

I dati devono essere consistenti

- ✓ **Coerenza**

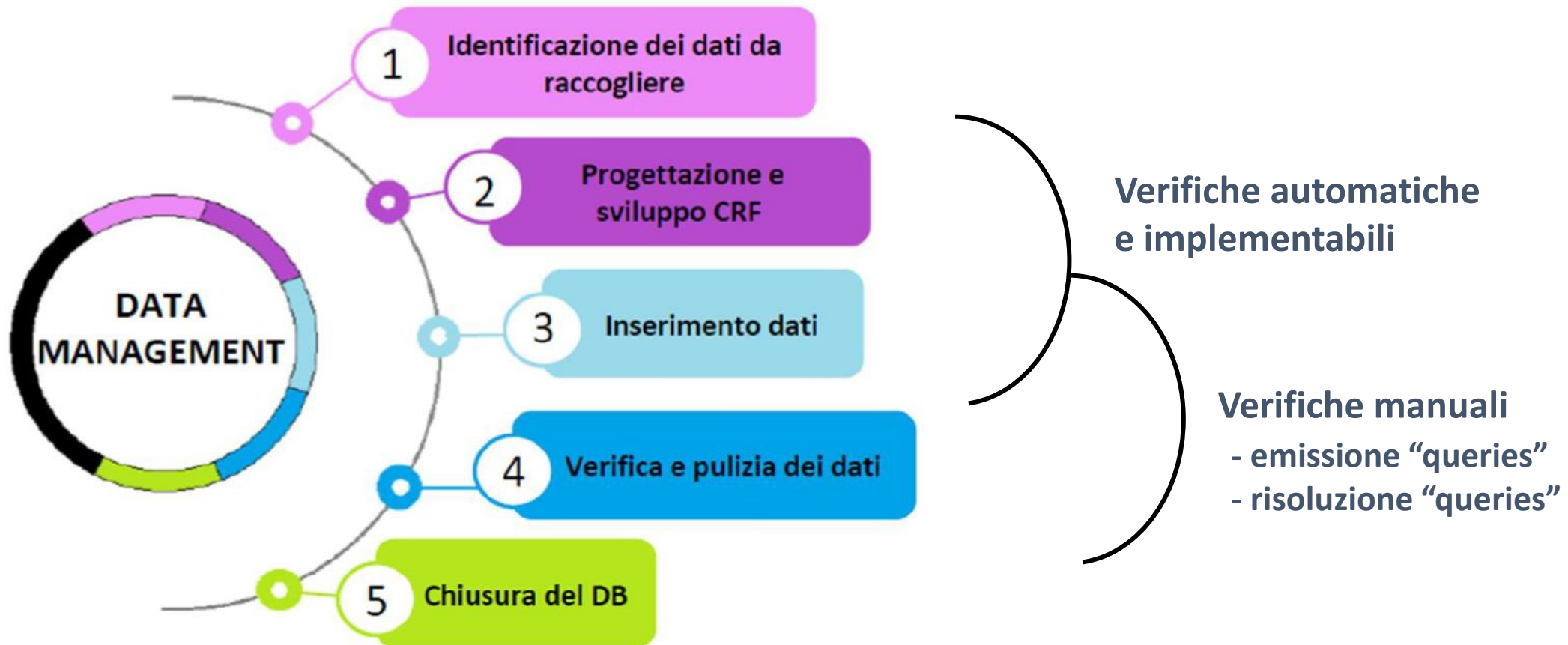
I dati devono essere tra loro compatibili e non presentare contraddizioni

Le Azioni



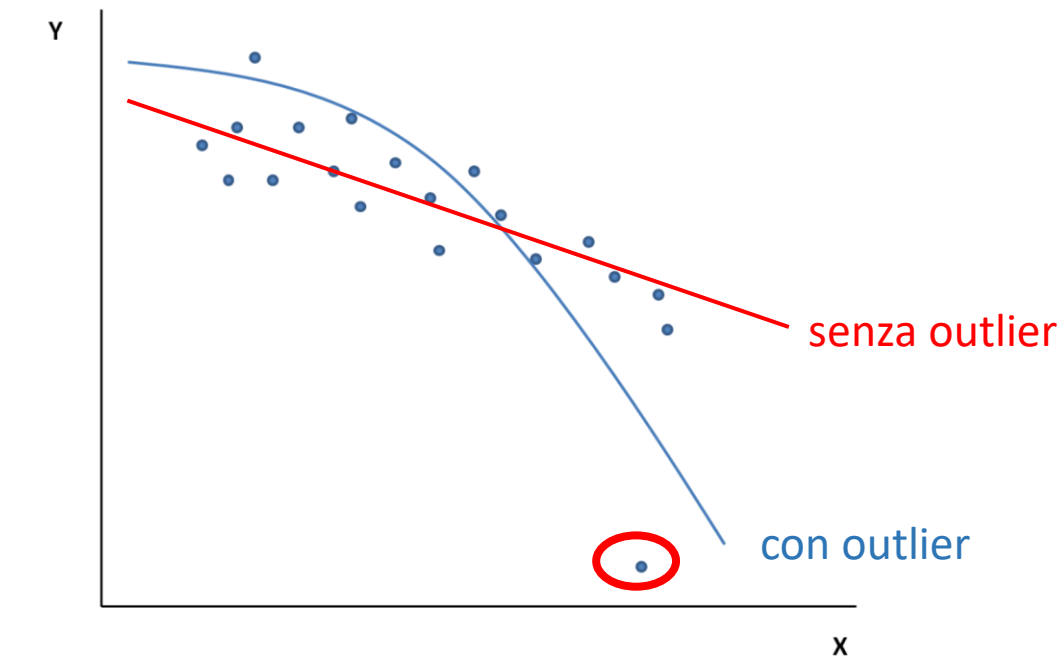
Il data management

Combinazione di più processi che hanno l'obiettivo di garantire che i dati siano acquisiti, controllati, convalidati, archiviati e protetti in modo standardizzato.



Outliers

Gli outliers devono essere controllati e non cancellati

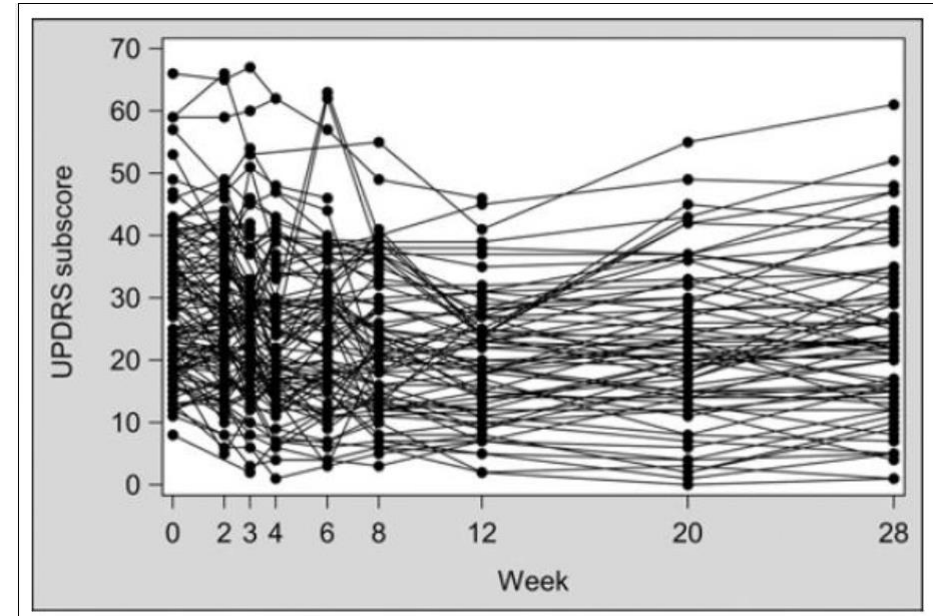


Dati mancanti

Esempio: studio clinico nel Parkinson

- Farmaco sperimentale vs placebo
 - 9 visite (visit 0=baseline) in 28 settimane
 - Endpoint primario: sub-score dell'Unified Parkinson's Disease Rating Scale (UPDRS) a 28 settimane
- valori alti: prognosi peggiore

Tutti i dati disponibili

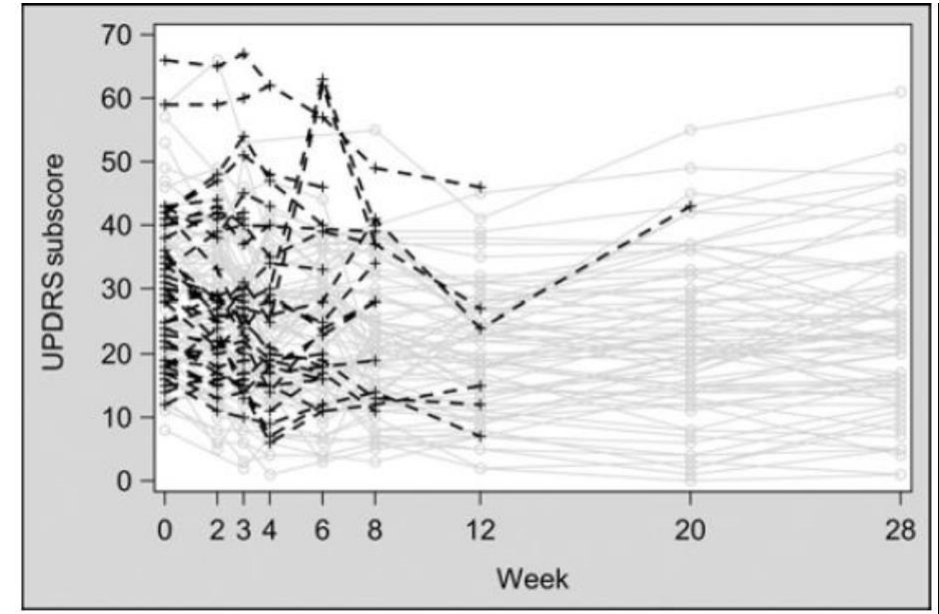


Dati mancanti

Perchè non posso usare i dati completi?

- Farmaco sperimentale vs placebo
 - 9 visite (visit 0=baseline) in 28 settimane
 - Endpoint primario: sub-score dell'Unified Parkinson's Disease Rating Scale (UPDRS) a 28 settimane
- valori alti: prognosi peggiore

In evidenza i soggetti che interrompono lo studio

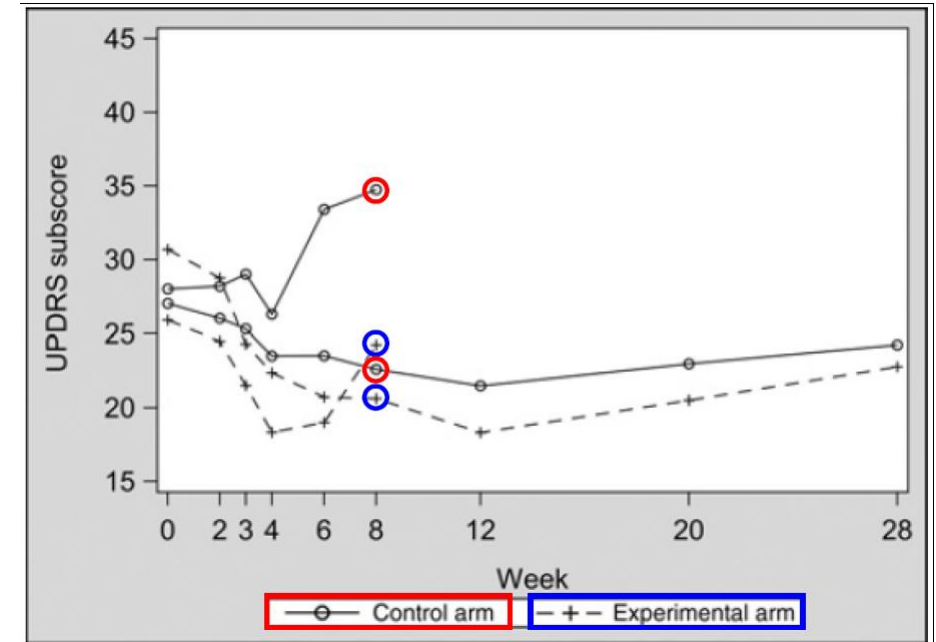


Dati mancanti

Perchè non posso usare i dati completi?

I pazienti che completano lo studio nel gruppo di controllo hanno una traiettoria più favorevole di quelli che interrompono.

Le informazioni parziali dei pazienti che non completano lo studio sono inutilizzate.



Modalità opportuna di follow-up

Esempio:

- studio che ha reclutato 25 nuove diagnosi di una certa malattia con esordio tra:
1/8/1989 – 31/7/1990
- fine dello studio: 31 Luglio 1991 (data limite analisi)

Due possibili modalità di follow-up:

- 1) Follow-up disponibile alla data limite + notifica delle morti appena possibile (\Rightarrow “*bad news come first*”)
- 2) Follow-up uniformemente aggiornato alla data limite

Modalità opportuna di follow-up

| Paziente n° | Data ingresso in studio | Tempo di osservazione e stato del paziente | | | |
|----------------|----------------------------|--|-------|------------------------|-------|
| | | Dati disponibili al 31/07/91 (a) | | Dati aggiornati (b) | |
| 1 | 4/08/89 | 18/12/90 | vivo | 18/07/91 | vivo |
| 2 | 5/08/89 | 15/04/91 | morto | 15/04/91 | morto |
| 3 | 11/08/89 | 3/01/91 | vivo | 25/08/91 | vivo |
| 4 | 25/08/89 | 19/09/89 | morto | 19/09/89 | morto |
| 5 | 8/09/89 | 25/01/91 | vivo | 20/07/91 | vivo |
| 6 | 15/09/89 | 5/02/91 | morto | 5/02/91 | morto |
| 7 | 29/09/89 | 18/04/91 | vivo | 8/08/91 | vivo |
| 8 | 12/10/89 | 25/03/91 | vivo | 10/08/91 | vivo |
| 9 | 30/10/89 | 15/04/91 | vivo | 31/07/91 | vivo |
| 10 | 5/11/89 | 10/05/91 | vivo | 18/08/91 | vivo |
| 11 | 11/12/89 | 1/04/91 | morto | 1/04/91 | morto |
| 12 | 27/12/89 | 31/01/91 | vivo | 31/01/91 | perso |
| 13 | 30/12/89 | 18/07/91 | vivo | 18/07/91 | vivo |
| 14 | 21/01/90 | 18/03/91 | vivo | 30/07/91 | vivo |
| 15 | 30/01/90 | 18/07/90 | vivo | 20/07/91 | vivo |
| 16 | 22/02/90 | 3/01/91 | vivo | 25/08/91 | vivo |
| 17 | 7/03/90 | 29/12/90 | vivo | 29/07/91 | vivo |
| 18 | 7/04/90 | 16/07/91 | morto | 16/07/91 | morto |

Modalità opportuna di follow-up

| Tempo di osservazione e stato del paziente | | | | | |
|--|-------------------------|----------------------------------|-------|---------------------|-------|
| Paziente n° | Data ingresso in studio | Dati disponibili al 31/07/91 (a) | | Dati aggiornati (b) | |
| 1 | 4/08/89 | 18/12/90 | vivo | 18/07/91 | vivo |
| 2 | 5/08/89 | 15/04/91 | morto | 15/04/91 | morto |
| 3 | 11/08/89 | 3/01/91 | vivo | 25/08/91 | vivo |
| 4 | 25/08/89 | 19/09/89 | morto | 19/09/89 | morto |
| 5 | 8/09/89 | 25/01/91 | vivo | 20/07/91 | vivo |
| 6 | 15/09/89 | 5/02/91 | morto | 5/02/91 | morto |
| 7 | 29/09/89 | 18/04/91 | vivo | 8/08/91 | vivo |
| 8 | 12/10/89 | 25/03/91 | vivo | 10/08/91 | vivo |
| 9 | 30/10/89 | 15/04/91 | vivo | 31/07/91 | vivo |
| 10 | 5/11/89 | 10/05/91 | vivo | 18/08/91 | vivo |
| 11 | 11/12/89 | 1/04/91 | morto | 1/04/91 | morto |
| 12 | 27/12/89 | 31/01/91 | vivo | 31/01/91 | perso |
| 13 | 30/12/89 | 18/07/91 | vivo | 18/07/91 | vivo |
| 14 | 21/01/90 | 18/03/91 | vivo | 30/07/91 | vivo |
| 15 | 30/01/90 | 18/07/90 | vivo | 20/07/91 | vivo |
| 16 | 22/02/90 | 3/01/91 | vivo | 25/08/91 | vivo |
| 17 | 7/03/90 | 29/12/90 | vivo | 29/07/91 | vivo |
| 18 | 7/04/90 | 16/07/91 | morto | 16/07/91 | morto |

Modalità opportuna di follow-up

| Paziente n° | Data ingresso in studio | Tempo di osservazione e stato del paziente | | | |
|----------------|----------------------------|--|-------|------------------------|-------|
| | | Dati disponibili al 31/07/91 (a) | | Dati aggiornati (b) | |
| 1 | 4/08/89 | 18/12/90 | vivo | 18/07/91 | vivo |
| 2 | 5/08/89 | 15/04/91 | morto | 15/04/91 | morto |
| 3 | 11/08/89 | 3/01/91 | vivo | 25/08/91 | vivo |
| 4 | 25/08/89 | 19/09/89 | morto | 19/09/89 | morto |
| 5 | 8/09/89 | 25/01/91 | vivo | 20/07/91 | vivo |
| 6 | 15/09/89 | 5/02/91 | morto | 5/02/91 | morto |
| 7 | 29/09/89 | 18/04/91 | vivo | 8/08/91 | vivo |
| 8 | 12/10/89 | 25/03/91 | vivo | 10/08/91 | vivo |
| 9 | 30/10/89 | 15/04/91 | vivo | 31/07/91 | vivo |
| 10 | 5/11/89 | 10/05/91 | vivo | 18/08/91 | vivo |
| 11 | 11/12/89 | 1/04/91 | morto | 1/04/91 | morto |
| 12 | 27/12/89 | 31/01/91 | vivo | 31/01/91 | perso |
| 13 | 30/12/89 | 18/07/91 | vivo | 18/07/91 | vivo |
| 14 | 21/01/90 | 18/03/91 | vivo | 30/07/91 | vivo |
| 15 | 30/01/90 | 18/07/90 | vivo | 20/07/91 | vivo |
| 16 | 22/02/90 | 3/01/91 | vivo | 25/08/91 | vivo |
| 17 | 7/03/90 | 29/12/90 | vivo | 29/07/91 | vivo |
| 18 | 7/04/90 | 16/07/91 | morto | 16/07/91 | morto |

+ 7 mos

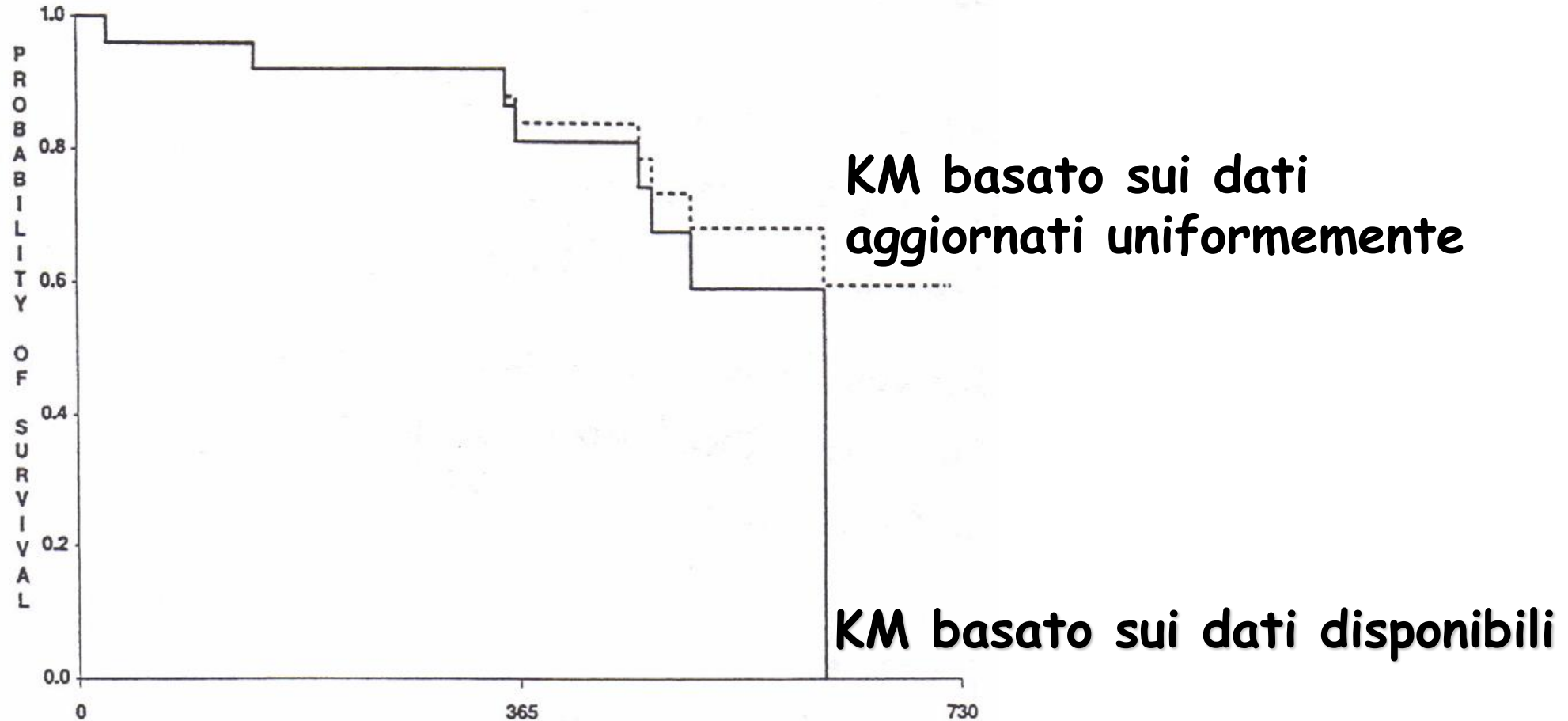
+ 7.7 mos

+ 5.8 mos

+ 0 mos

Modalità opportuna di follow-up

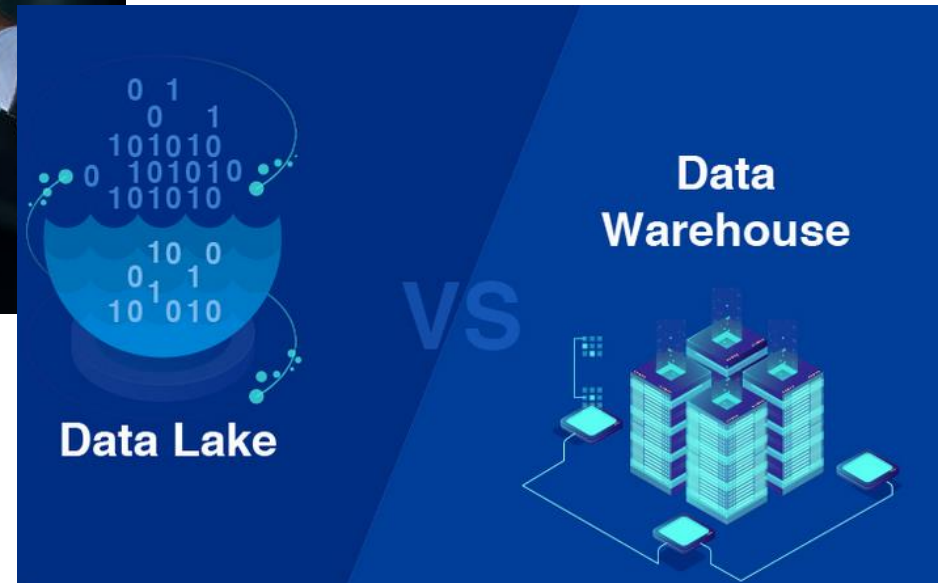
Cautele nell'utilizzo dello stimatore di Kaplan-Meier



I dati disponibili forniscono stime distorte.

I big data

Si moltiplicano i dati, ma il problema della loro qualità rimane



Lavoro di squadra

Solido razionale +

Efficiente disegno +

Rigorosa conduzione +

Dati di qualità +

Appropriata analisi statistica =

CONCLUSIONI SALIENTI

Troppo spesso si pensa che la qualità dei dati sia garantita, ma va curata già dalla fase di pianificazione ed è responsabilità di tutti i componenti del team di ricerca.