



Quantifying the Impact of Selection Bias in Cohort Studies using Monte Carlo Simulations

C Pizzi^{1,2}, L Richiardi ¹, B De Stavola², F Merletti ¹

¹*Università di Torino*

²*London School of Hygiene & Tropical Medicine*

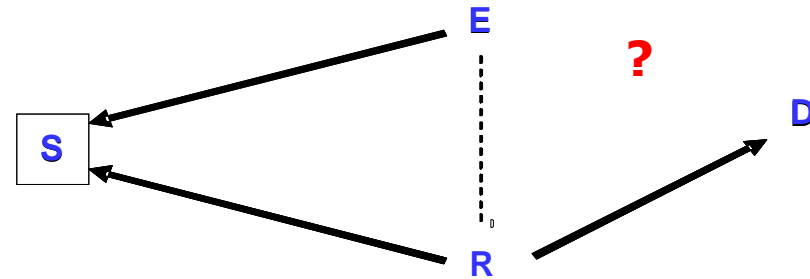
Pavia, 18 Settembre 2009 - SISMEC

Introduzione

- Gli studi osservazionali sono spesso condotti su popolazioni selezionate in base a criteri pre-specificati: questa restrizione puo' comportare problemi di **generalizzabilità** dei risultati
- Se il processo di selezione del campione dipende sia dall'esposizione di interesse sia da un fattore di rischio per la malattia, la restrizione può introdurre **problemi di validità interna delle stime**

Il setting

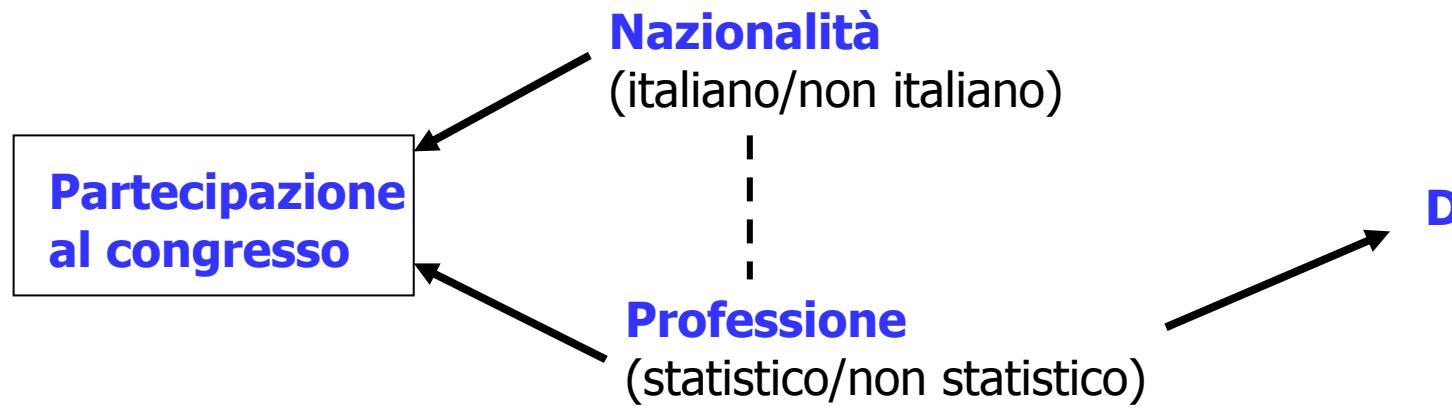
E: Esposizione binaria
D: Outcome binario
R: Fattore di rischio binario per D
S: Indicatore di inclusione nel campione



Parametro d'interesse: associazione E-D

- **Condizionare su $S=1$** crea una associazione spuria tra E ed R
- Se R è non noto o non misurato il backdoor path E—R—D non è bloccato e la stima dell'effetto E-D sarà distorta
- Il bias dipende da due componenti: l'associazione E-R indotta dalla restrizione e l'associazione vera tra R e D → **R confondente**
- La quantificazione del bias e l'effettivo impatto sulla validità degli studi di coorte sono stati poco discussi in letteratura

Il setting



- **Condizionare su $S=1$** crea una associazione spuria tra E ed R ^(2,3)
- Se R è non noto o non misurato il backdoor path E—R—D non è bloccato e la stima dell'effetto E-D sarà distorta
- Il bias dipende da due componenti: l'associazione E-R indotta dalla restrizione e l'associazione vera tra R e D → **R confondente**
- La quantificazione del bias e l'effettivo impatto sulla validità degli studi di coorte sono stati poco discussi in letteratura

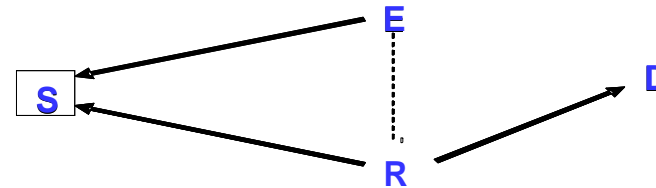
Obiettivo

Quantificare il bias indotto dal processo di selezione del campione nella stima dell'associazione tra esposizione e malattia negli studi di coorte



Simulazioni di Monte Carlo

Simulazioni – Generazione dati(1)



- **Background variables:**

E e **R** marginalmente indipendenti con prevalenze P_E e P_R

- **Selection model:**

$$\text{logit}(\mathbf{S}=1) = a_S + \beta_{SE} E + \beta_{SR} R + \beta_{inter} E * R$$

$P_S \rightarrow$ prevalenza baseline; $a_S \rightarrow \text{logit}(P_S)$

β_{SE} e $\beta_{SR} \rightarrow$ log-OR per l'effetto di E e R

$\beta_{inter} \rightarrow$ log-OR per l'interazione di E e R

- **Outcome model:**

$$\log \lambda = \log \lambda_0 + \beta_{DE} E + \beta_{DR} R$$

D \rightarrow modello esponenziale (tasso di occorrenza dell'evento, λ , costante nel tempo)

$\lambda_0 \rightarrow$ tasso baseline

β_{DE} e $\beta_{DR} \rightarrow$ log-RR per l'effetto di E e R

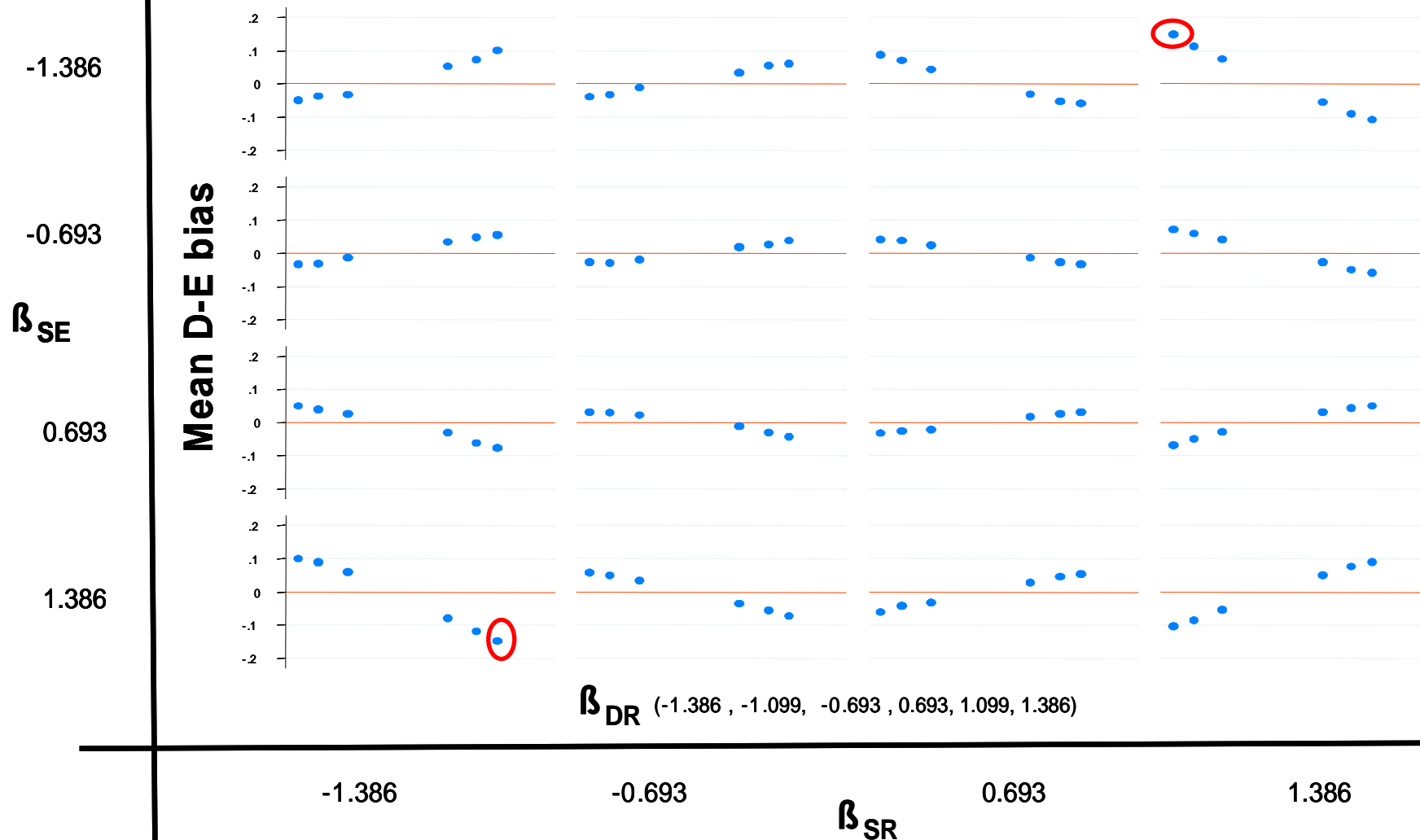
Simulazioni – Generazione dati(2)

⇒ *Parametri :*

$\beta_{SR}, \beta_{SE}, \beta_{DR}$	→ $\pm 0.693, \pm 1.099, \pm 1.386$
β_{inter}	→ 0
β_{DE}	→ 0
P_E	→ 0.5
P_R	→ 0.5
P_S	→ 0.5
λ_0	→ 0.03 eventi/anno
Censura	→ 5 anni
N	→ 5000

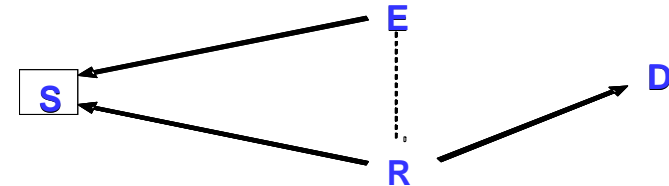
- ❖ 1,000 dataset per ogni combinazione dei parametri considerati
- ❖ Per ogni scenario simulato: media e SD della stima di β_{DE} e del bias
- ❖ β_{DE} stimato tramite modello dei rischi proporzionali di Cox
- ❖ **Bias(β_{DE}): differenza tra il valore vero, 0, e la sua stima.**

Bias nella stima dell'associazione E-D: ($0 - \beta_{DE|S=1}$)



Bias max = ± 0.15 ; Bias max = ± 0.02 quando $|\beta_{SR}|, |\beta_{SE}|$ e $|\beta_{DR}| \leq 0.639$

Scenari alternativi



❖ Interazione:

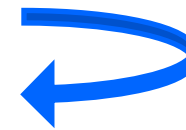
$\beta_{\text{inter}} = \pm 0.693 \rightarrow \beta_{\text{ER|S=1}}$ aumenta drasticamente

Bias max: -0.24; 0.27

❖ Prevalenza di R:

Per un dato $\beta_{\text{ER|S=1}}$, bias max se $P_{\text{R|S=1}} \approx 0.5$

$P_{\text{R}} = 0.1 \rightarrow$ **Bias max** diminuisce a **-0.12; 0.07**



❖ Prevalenza di E ($P_{\text{E}} = 0.25$); Tasso baseline per D ($\lambda_0 = 0.01$ e 0.06); Ampiezza dello studio ($N = 2,500$)



No effetto sulla grandezza del bias

Bias

1. Il bias dipende:

- dalla forza dell'associazione di E e R col processo di selezione del campione (S)
- dalla prevalenza campionaria del fattore di rischio ($P_{R|S=1}$)
- dall' associazione tra R e D

2. Il **bias max** risulta **moderato** (± 0.15) anche per β_{SR} e β_{SE} elevati

3. **Interazione** tra E e R ($OR_{inter} = 2/0.5$) \rightarrow il **bias max** < 0.3 *NB: Equivale, in termini di bias, a scenari con β_{SR} e $\beta_{SE} >>$*

In Pratica...

➤ **Esempio 1:**

Partecipanti reclutati da una sotto-popolazione di soggetti con una elevata prevalenza di esposizione
(***Confronti interni in coorti occupazionali***)

➤ **Esempio 2:**

Coorte creata reclutando da una sotto-popolazione con elevato tasso di partecipazione allo studio o con facilità di reclutamento
(***Nurses' Health Study, NINFEA internet-based***)

	Random sample N = 1,000	Selected samples (S=1)		
		N = 1,000	N = 2,000	N = 4,000
$p(D=1 R=0, E=0)$	0.10	0.10	0.10	0.10
β_{DR}	1.39	1.39	1.39	1.39
$p(S=1 R=0, E=0)$	--	0.50	0.50	0.50
β_{SE}	--	1.39	1.39	1.39
β_{SR}	--	1.39	1.39	1.39
$p(D=1)$	0.20	0.23	0.23	0.23
$p(E=1)$	0.10	0.13	0.13	0.13
$p(R=1)$	0.50	0.61	0.61	0.61
β_{ER}	--	-0.31	-0.31	-0.31
Bias	0	0.10	0.10	0.10
SE (β_{DE})	0.26	0.23	0.16	0.11

Associazione positiva di E e R con S → incremento di esposti ad E e R e di casi
 → aumenta la precisione

Se N aumenta a 4,000 → SE << rispetto al campione casuale

Conclusioni

Condurre studi di coorte su popolazioni selezionate e' spesso inevitabile;

Il bias indotto dalla restrizione e' moderato se non in situazioni estreme e

..la selezione puo' aumentare il potere dello studio.

 conoscere i dettagli del processo di selezione e tenerne conto in fase di analisi

Bibliografia

- 1) Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes JM, et al eds. *Methods in social epidemiology*. San Francisco: Jossey-Bass; 2006.
- 2) Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615-25.
- 3) Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press; 2000.
- 4) Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003 May;14(3): 300-6.

Bias e Precisione

- **Setting:**

D → generato da modello logistico

$\beta_{ER|s=1}$, $P_{E|s=1}$ e $P_{R|s=1}$ → derivati dalle simulazioni

Bias → Calcolato usando equazione algebrica ⁽⁴⁾ :

$$\checkmark \quad \mathbf{Bias}(\beta_{DE}) = Bias_{max}(RR_{ER}, RR_{DR}) = (G+1)^2 / (RR_{ER} + 2G + RR_{DR})$$

con $G = (RR_{ER} * RR_{DR})^{1/2}$

Varianza → Maximum likelihood estimation

- **Scenari a confronto:**

→ i) β_{SR} e $\beta_{SE} = 0$; $N=1,000$

→ ii) β_{SR} e $\beta_{SE} = 1.386$; $N=1,000$ $N=2,000$ $N=4000$

Parametri fissi: $\beta_{DR} = 1.386$; $\beta_{DE} = 0$; $P_D = 0.10$; $P_E = 0.10$; $P_R = 0.50$; $P_S = 0.50$