CAUSAL RELATIONSHIP IN EPIDEMOLOGY

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics Dept. of Diagnostics & Public Health University of Verona

A) Descriptive relationship

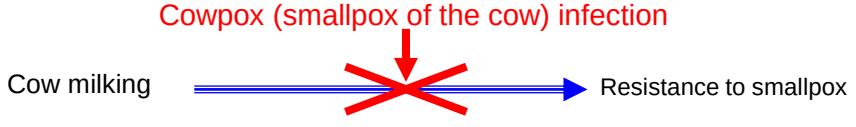
determinant —(No hypothesis of causal dependence)→ Parameter of occurrence

For example:

Yellow fingers → Lung cancer

Cow milking → Resistance to smallpox

However if we control for smoking habits by separately studying smokers and non-smokers



Descriptive relationship allows to identify groups at high risk

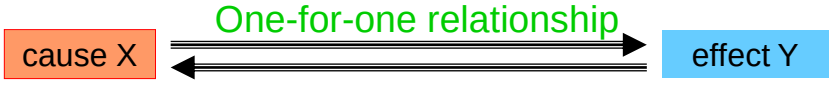
In epidemiology an empirical relationship between a determinant and a measure of occurrence is considered causal, when it still holds after controlling for all possible confounders.

B) Causal relationship

EXPERIMENT	PLANNED OBSERVATION
Researchers actively modify the course of events	Researchers just observe the course of events, without attempting to modify it
Only **positive perturbations** can be applied: 1) Preventive interventions, such as adding fluorine to tap water, or iodine to salt 2) Therapeutic measures (early thrombolysis in myocardial infarction, segmental vs total mastectomy) 3) Rehabilitation interventions	Also **etiologic factors** with **deleterious** health effects can be studied: 1) wrong lifestyle (smoking, excessive alcohol intake) 2) environmental situation (Chernobyl)
RANDOMIZATION	**SELF-SELECTION**
Participants are randomly assigned to different treatments ↓ Other risk factors (potential confounders) are balanced among groups	Potential confounders are not eliminated. For instance, it could be hypothesized that: Unknown genes → Craving for smoking Unknown genes → Increased risk of lung cancer

Classical (deterministic) interpretation of causality

X is a cause of Y if, in a perfectly stable system, any change in X produces a change in Y.



Specificity of the cause (necessary and sufficient): X is the only cause of Y

Specificity of the effect: Y is the only effect of X

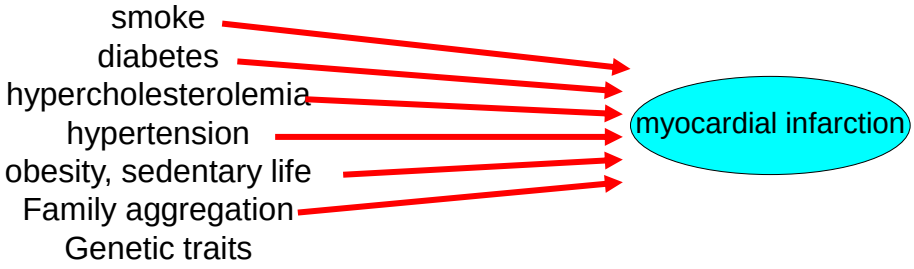
In the Nineteenth century Koch attempted to apply the deterministic interpretation of causation to the study of infectious diseases.

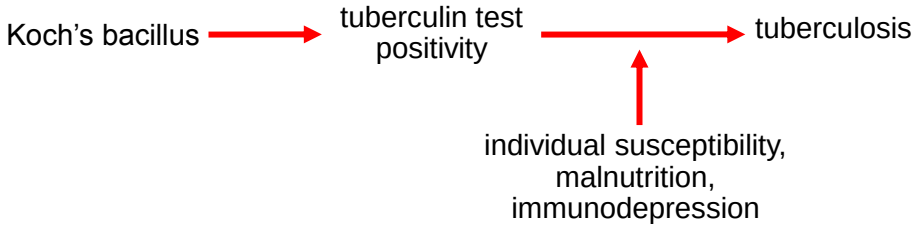
Koch's postulates (adapted)

1) The microorganism (virus, bacterium) must be found in abundance in all organisms suffering from the disease (necessary cause)
2) The microorganism should not be found in healthy organisms (or in organisms suffering from different diseases) (specificity of the effect)
3) The cultured microorganism should cause disease when introduced into a healthy organism (sufficient cause)

For example: rabies virus

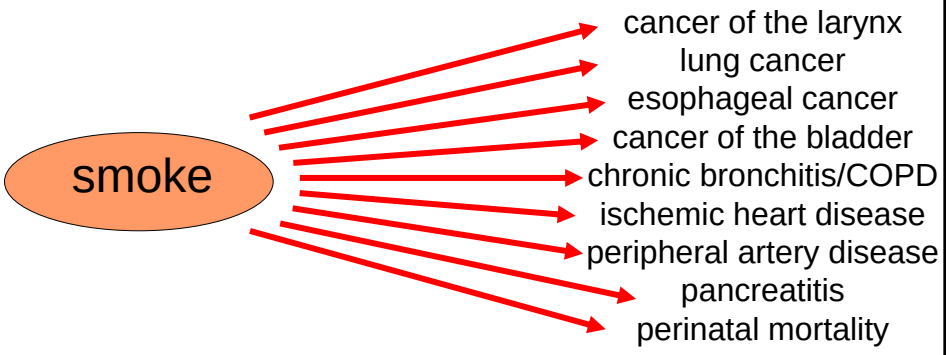
Multifactorial etiology: Diseases usually have several causes





Multiplicity of effects:

many risk factors have several harmful effects



necessary cause

	diseased	healthy
exposed	a	b
unexposed	---	d

unexposed are all healthy

infectious diseases (TBC, influenza)

sufficient cause

	diseased	healthy
exposed	a	---
unexposed	c	d

all exposed are sick

post-traumatic pneumothorax, traumas

necessary and sufficient cause

	diseased	healthy
exposed	a	---
unexposed	---	d

all exposed are sick

all unexposed are healthy

genetic diseases (Down's syndrome), rabies

probabilistic model of cause

	sick	healthy
exposed	a	b
unexposed	c	d

The disease is more frequent in exposed than in unexposed

p(dis./exposed) > p(dis./unexposed)

chronic-degenerative diseases

probabilistic model of cause

- Present knowledge of chronic-degenerative diseases suggests that cause-effect relationships involved are much weaker than deterministic relationships
- Causes (risk factors) involved are neither necessary nor sufficient

Smoke → lung cancer

hypercholesterolemia → myocardial infarction

Probabilistic interpretation of cause

A risk factor is an exposure that changes in a regular and predictable way the risk (probability) of disease

Example: the increase in lung cancer incidence in women is predicted by cumulative exposure to cigarette smoking (pack-years)

HILL'S CRITERIA FOR CAUSATION

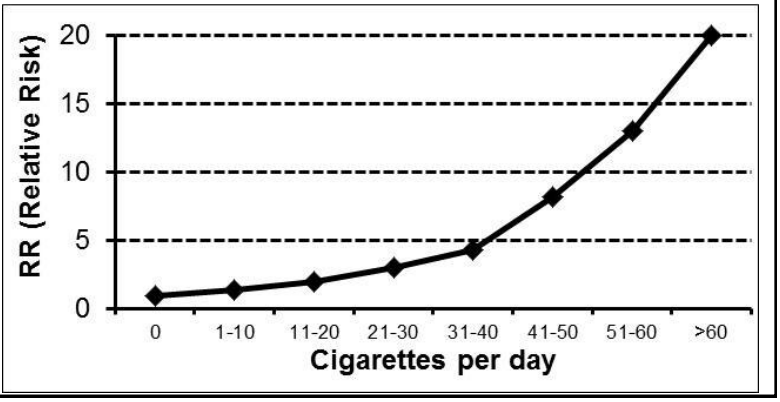
1) STRENGTH OF ASSOCIATION (effect size): small effects likely represent random fluctuations, while large effects are more likely to reflect a cause-effect association.
   For instance, the association between cigarette smoking and lung cancer (Relative Risk =14) is stronger than the association between cigarette smoking and myocardial infarction (RR = 1.62).

2) CONSISTENCY (reproducibility): Findings should be replicated by different groups in different places with different samples.
   For instance, the association between alcohol intake and esophageal cancer should be found in Europe as well as in the Far East.

4) TEMPORALITY: the exposure precedes the effect (*post hoc, propter hoc*).

For instance, a treatment with estrogens can be causally linked to thrombophlebitis only if started before the onset of thrombophlebitis itself.

5) BIOLOGICAL GRADIENT (dose-response relation): Greater exposure should generally lead to greater incidence of the disease.



6) PLAUSIBILITY: the new cause-effect relationship should be in line with current scientific knowledge.

However, the relation between Zodiac signs an myocardial infarction, found in a study, has no scientific explanation.

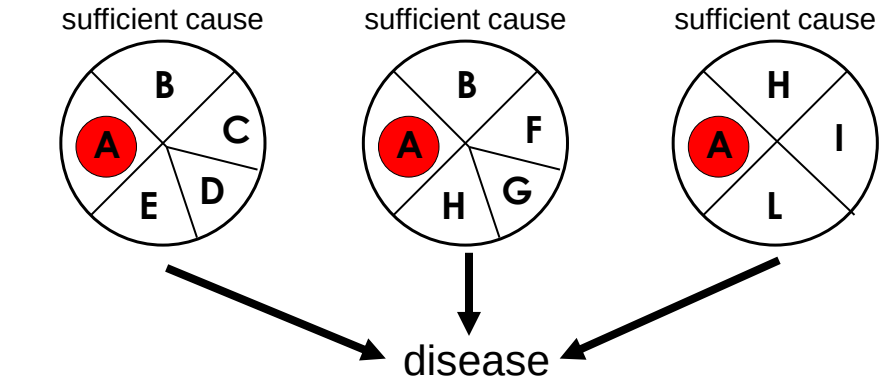
8) EXPERIMENT: the cause-effect association, found in an observational study, should be confirmed by an experimental study.

For instance, the association between smoking and lung cancer has been confirmed by experiments on animal models [Hutt, Carcinogenesis 2005]

Rothman's causal pie model

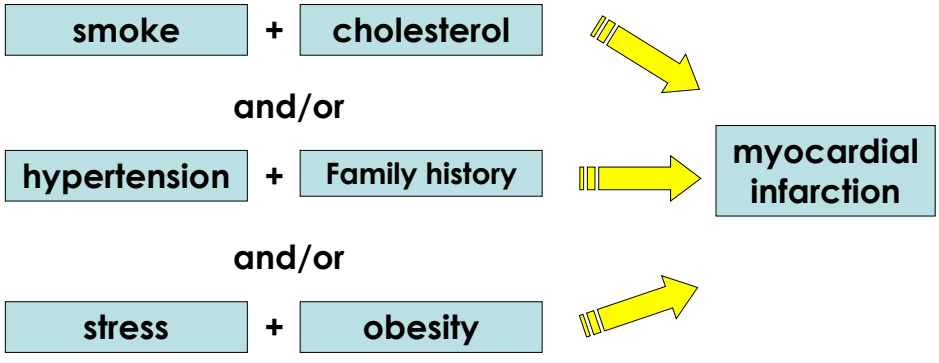
Multifactorial etiology = a disease is due to several causes. A set of causes occurring together, just sufficient to initiate the disease process, is called a **causal complex.** It is depicted as a pie with several slices, each representing a single component cause.

The same disease can be elicited by different causal complexes.



A = necessary cause

Multiplicity of causes (and effects)



Rothman's model is intrinsically deterministic. It turns probabilistic as we do not know all the risk factors involved in causal complexes.

Eighty percent of cancer cases are due to environmental causes,

Ninety percent of cancer causes is due to genetic causes.

The sum (80% + 90% = 170%) is greater than 100%.

This paradox can be explained by applying Rothman's causal pie model:

environmental causes are present in 80% of causal complexes; genetic causes are present in 90% of causal complexes.

EFFECT MODIFIER

EFFECT MODIFIER

QUANTITATIVE INTERACTION = the effect of a risk factor gets stronger or weaker in different levels of the other factor.

For instance, the carcinogenic effect of alcohol changes as a function of genetic variants of aldehyde dehydrogenase, which detoxifies acetaldehyde, a genotoxic metabolite of alcohol. People who are heterozygous for the inactive enzyme, are at higher risk for esophageal cancer when drinkers [Lewis 2005; Yokoyama 2005].

QUALITATIVE INTERACTION = a factor has opposite effects (increase vs decrease) in different levels of the other factor.

For instance, acetylcholine, when administered to an isolated artery, causes vasodilation if the endothelium is intact, vasoconstriction is the endothelium has been removed

CONFOUNDING

Cohort study — WHOLE SAMPE

INFARCTION

		Yes	No	
Alcohol	Yes	100	900	1000
	No	20	980	1000
		120	1880	2000

P(infarction) =
P(infarction/alcohol) =
P(infarction/teetotaler) =

RR (Relative Risk)=
OR (Odds Ratio) =

SMOKERS

INFARCTION

		Sì	No	
Alcohol	Yes	99	801	900
	No	11	89	100
		110	890	1000

P(infarction) =
P(infarction/alcohol) =
P(infarction/teetotaler) =

RR (Relative Risk)=
OR (Oddds Ratio) =

NON-SMOKERS

INFARCTION

		Yes	No	
Alcohol	Yes	1	99	100
	No	9	891	900
		10	990	1000

P(infarction) =
P(infarction/alcohol) =
P(infarction/teetotaler) =

RR (Relative Risk)=
OR (Oddds Ratio) =

Cohort study — WHOLE SAMPE

INFARCTION

		Yes	No	
Alcohol	Yes	100	900	1000
	No	20	980	1000
		120	1880	2000

P(infarction) = 120/2000 = 6%
P(infarction/alcohol) = 100/1000 = 10%
P(infarction/teetotaler) = 20/1000 = 2%

RR (Relative Risk)= 0.10 / 0.02 = 5
OR (Odds Ratio) = (100*980) / (20*900) = 5.44

SMOKERS

INFARCTION

		Sì	No	
Alcohol	Yes	99	801	900
	No	11	89	100
		110	890	1000

P(infarction) = 110/1000 = 11%
P(infarction/alcohol) = 99/900 = 11%
P(infarction/teetotaler) = 110/1000 = 11%

RR (Relative Risk)= 0.11/0.11 = 1
OR (Odds Ratio) = (99*89) / (11*801) = 1

NON-SMOKERS

INFARCTION

		Yes	No	
Alcohol	Yes	1	99	100
	No	9	891	900
		10	990	1000

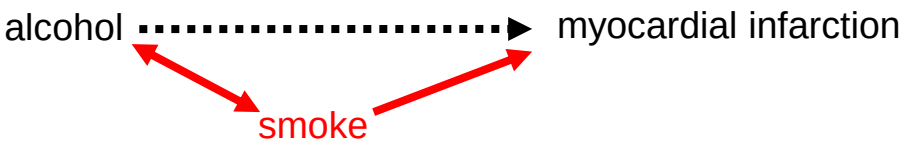
P(infarction) = 10/1000 = 1%
P(infarction/alcohol) = 1/100 = 1%
P(infarction/teetotaler) = 9/900 = 1%

RR (Relative Risk)= 0.01/0.01 = 1
OR (Odds Ratio) = (1*891) / (9*99) = 1

INTERPRETATION: The risk of myocardial infarction apparently increases five-folds in drinkers.

alcohol ⟶ myocardial infarction

Actually heavy drinkers also tend to smoke more than non-drinkers; for this reason, they present a higher incidence of myocardial infarction.



In epidemiological terms smoke is a confounder of the relation between alcohol and myocardial infarction.

A variable can be considered a confounder if:

1. It is independently associated with the outcome (i.e. it is a risk factor).
2. It is also associated with the exposure under study in the source population. In other words, the confounder must differently distributed in different levels of the potential risk factor under study.
3. It is not an intermediate step in the causal pathway between the exposure and the outcome under study.

BIAS

MEASUREMENT ERRORS

Random errors and **systematic errors**.

Random error reduces precision of the estimate (precision = one half of the confidence interval). Random error can be coped with by increasing sample size.

Systematic errors or **biases** are classified as information, selection and confounding biases.

Information bias: for instance, in multicenter studies it is important to centralized the most important laboratory assessment. Otherwise, comparison among lab values collected in different centers can be biased by different laboratory methods.

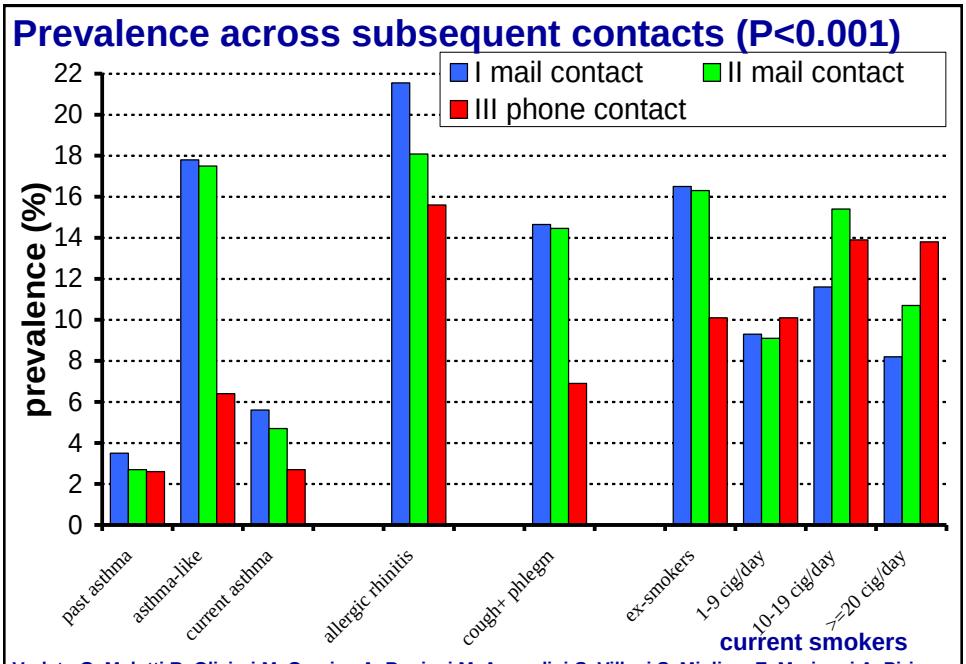
Will-Rogers phenomenon or stage migration: the more lymph nodes are removed in gastric cancer patients, the more metastatic nodes are found [De Manzoni, Verlato et al, Brit J Surg, 2002].

Selection bias: In mailed surveys on respiratory health, asthmatics and ex-smokers tend to be early responders while current smokers tend to be late responders. Hence, if only 50% of the sample respond to the mailed survey, "prevalence rates" of asthma and ex-smokers are over-estimated, while the prevalence of current smokers in under-estimated.

Verlato et al. Asthmatics and ex-smokers respond early, heavy smokers respond late to mailed surveys in Italy. Resp Med 2010.

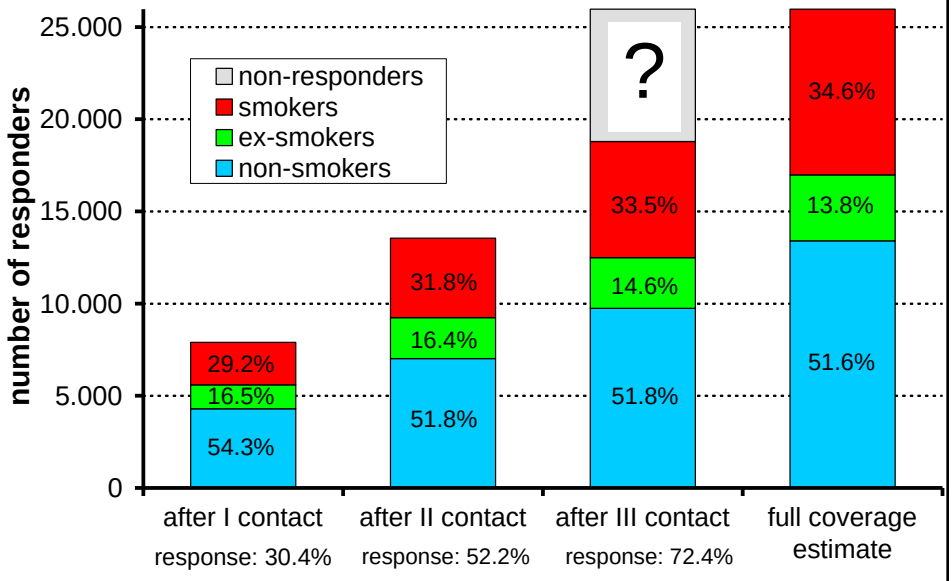
Confounding bias: In the *Verona Diabetes Study*, diabetic women experienced about the same mortality rate as diabetic men (RR = 0.97, 95% CI 0.88-1.07), as if diabetes completely eliminated the "female survival advantage".

However, at baseline diabetic women were older than diabetic men: 68.3±12.2 versus 62.2±13.0 years (mean±standard deviation). Indeed, in multivariable survival analysis, the female survival advantage became evident when controlling for age: RR of women versus men = 0.64, 95% CI 0.58-0.71.



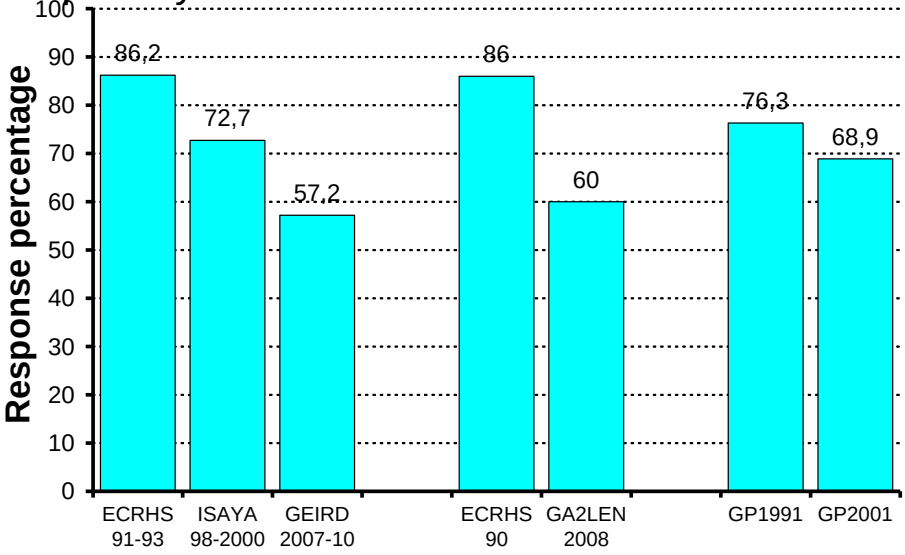
Verlato G, Melotti R, Olivieri M, Corsico A, Bugiani M, Accordini S, Villani S, Migliore E, Marinoni A, Pirina P, Carrozzi L, Bortolami O, Rava M, and de Marco R, for the ISAYA study group (2010) Asthmatics and ex-smokers respond early, heavy smokers respond late to mailed surveys in Italy. *Resp Med*, 104: 172-179

Prevalence of non-smokers, ex-smokers and current smokers after the I, II and III contact. The last column reports estimates for the whole sample, when attributing prevalence recorded in the III contact to hardcore non-responders.



Verlato et al, Respir Med,

Response to mailed epidemiological surveys on respiratory health has decreased in the last decades



Italy (de Marco et al, Eur Respir J 2012)

Sweden (Bjerg et al, PLOS ONE 2011)

England (Frank, Int J Tuberc Lung Dis '07)