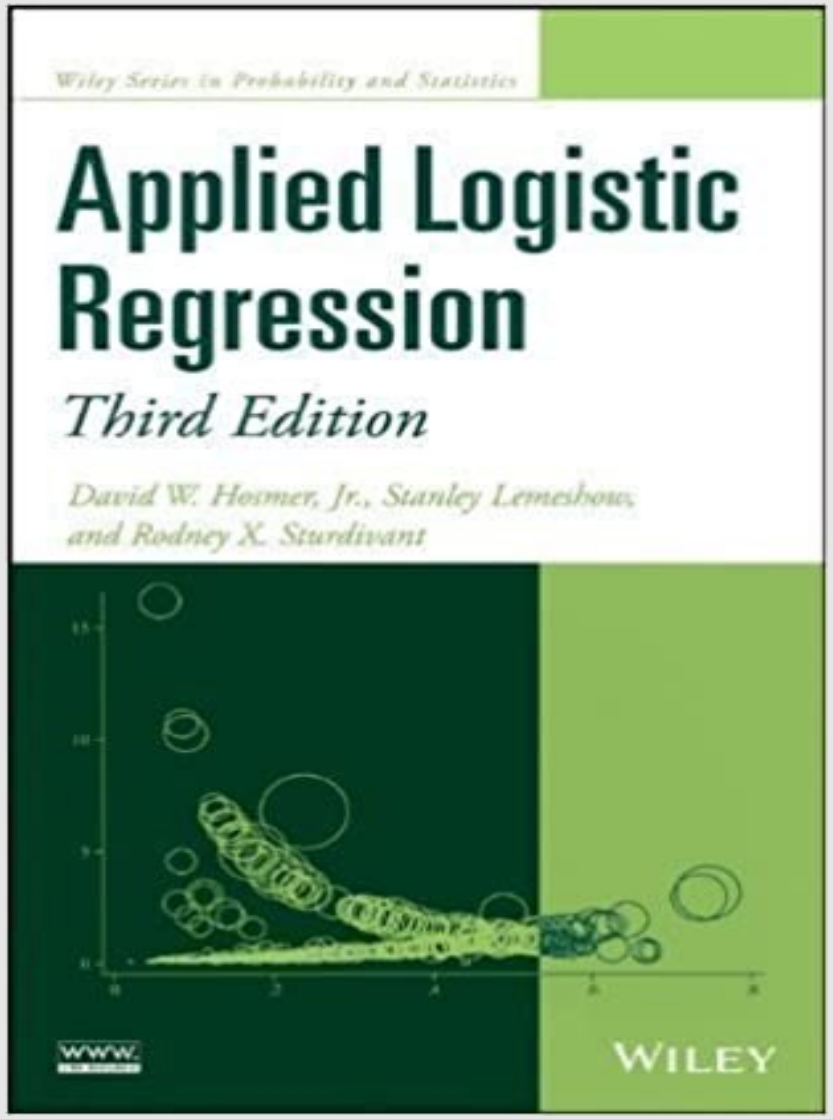
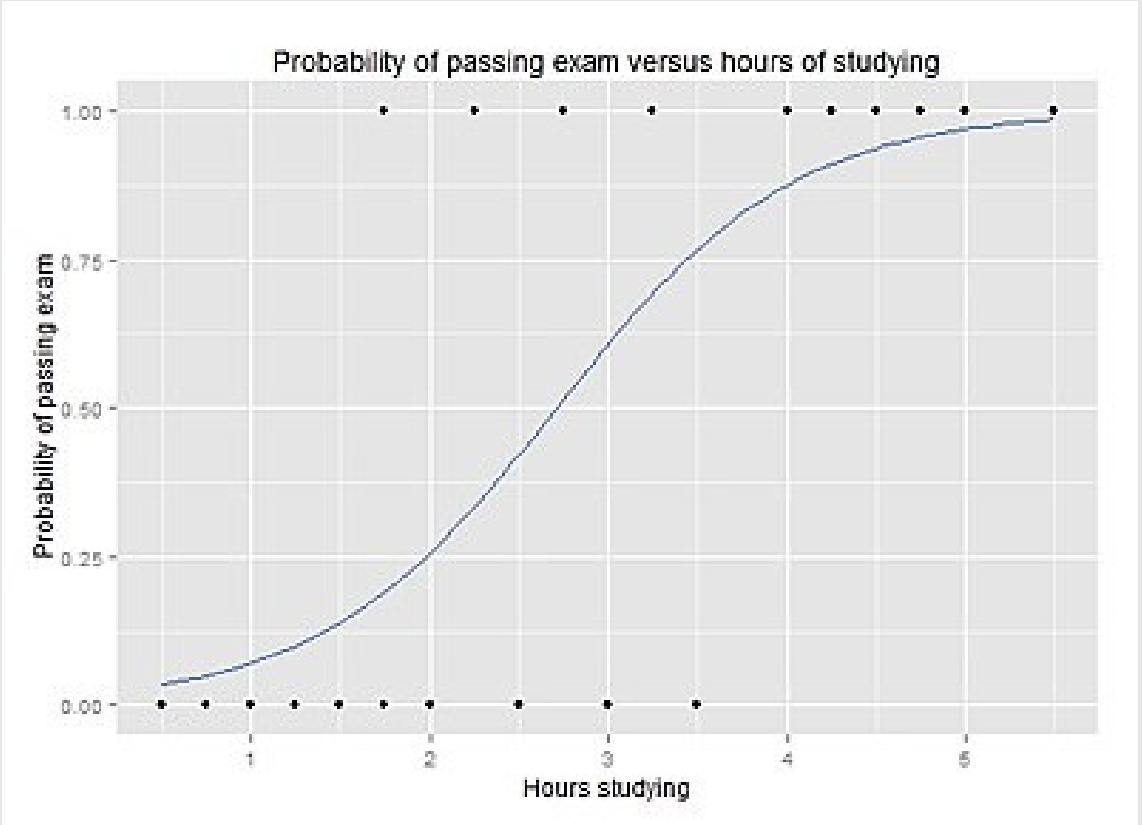
Logistic regression





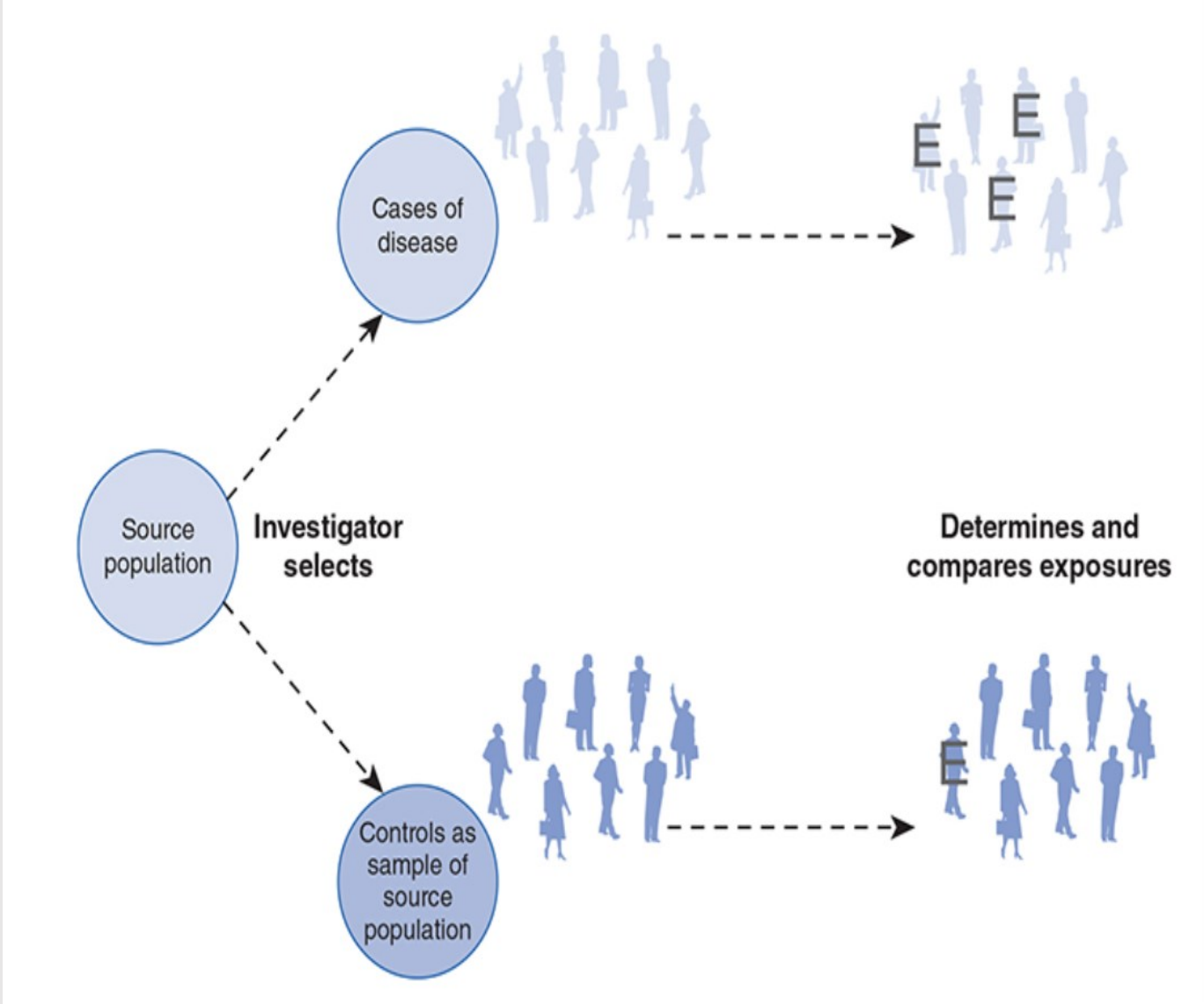
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Now we focus on a **regression model particularly useful** to evaluate data obtained from a **case-control** study design.

The purpose of this design is to assess the **magnitude of the association** between an exposure and a specific disease or health-related event.

This is the most **cost-effective** study design and is recommended when the **incidence** of the disease or condition of interest is **rare** or has a **long latency.**

The aim could be both on **explaining** effects, or making *predictions*, for example ***prediction of diagnosis*** could be a target.



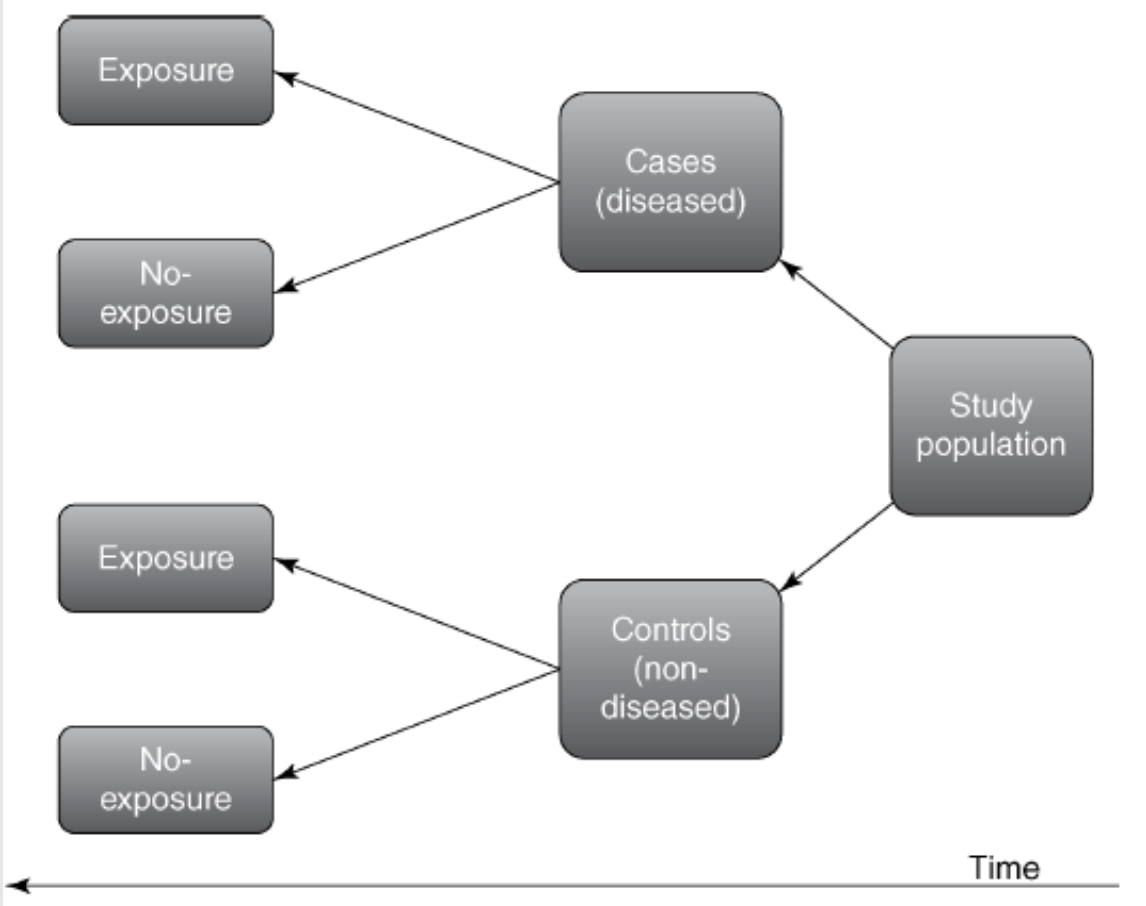
Specific Objectives (mostly related to **explanatory/causal** purposes):

- Define the concepts of ***crude*** and ***adjusted*** odds ratios (ORs) to assess the magnitude of the association between an exposure and a specific disease

- Assess ***confounding*** and ***effect modification [interaction]***

Remember: in a standard case-control study design, it is not possible to estimate **disease incidence** in those who are exposed and those who are unexposed, since participants are selected according to the disease status, not on the basis of their exposure status.

However, it is possible to calculate the odds of exposure among cases and controls, and then the exposure/disease odds ratio.

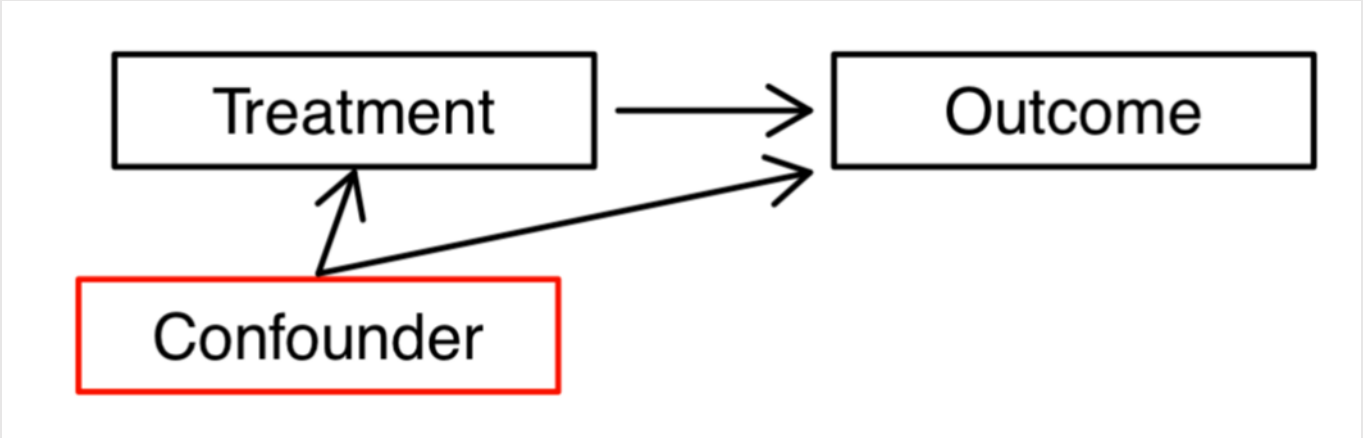


Confounding Assessment (explanatory setting)

To evaluate the effect of potential **confounding variables** in a case-control study, it is necessary to compare the estimate of the ***crude*** odds ratio $\widehat{OR}_{crude}$ with the estimate of the ***adjusted*** odds ratio $\widehat{OR}_{adjusted}$

In general, if the $\widehat{OR}_{crude}$ is *similar* to $\widehat{OR}_{adjusted}$ we can think that the confounding variables have no effect in the **magnitude** of the association of interest.

Otherwise, it is recommended to determine if the $\widehat{OR}_{crude}$ over-estimates or under-estimates the association using as a reference the $\widehat{OR}_{adjusted}$



The multivariable logistic regression model allows us to estimate the OR (*crude* and *adjusted*) to assess the magnitude of the association between the exposure of interest and the disease under study taking into account **multiple** confounders **with different scales** of measurements.

LR estimates* the probability of disease in the exposed and unexposed groups as follows:

$$OR = \frac{P_{exposed}/(1 - P_{exposed})}{P_{unexposed}/(1 - P_{unexposed})}$$

$$p_i = \frac{1}{1 + e^{-\left(\beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}\right)}}$$

β_E coefficient of the exposure

β_j coefficient of the j-th *confounding*** variable

p_i

probability of having the disease for the subject *i*

E_i

exposure for the subject *i* (quantitative or categorical). If dichotomous, defined by an indicator variable(dummy variable)1=present and 0=absent.

C_{ij}

*Confounding*** variable for subject *i*, for *j=1,….m*

*estimate β by *maximising the likelihood*, i.e. probabilities to observe the data in hand get maximal.

**or independent predictor

The probability of observing a control (non diseased person) through the LR model is:

$$1 - p_i = 1 - \frac{1}{1 + e^{-\left(\beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}\right)}}$$

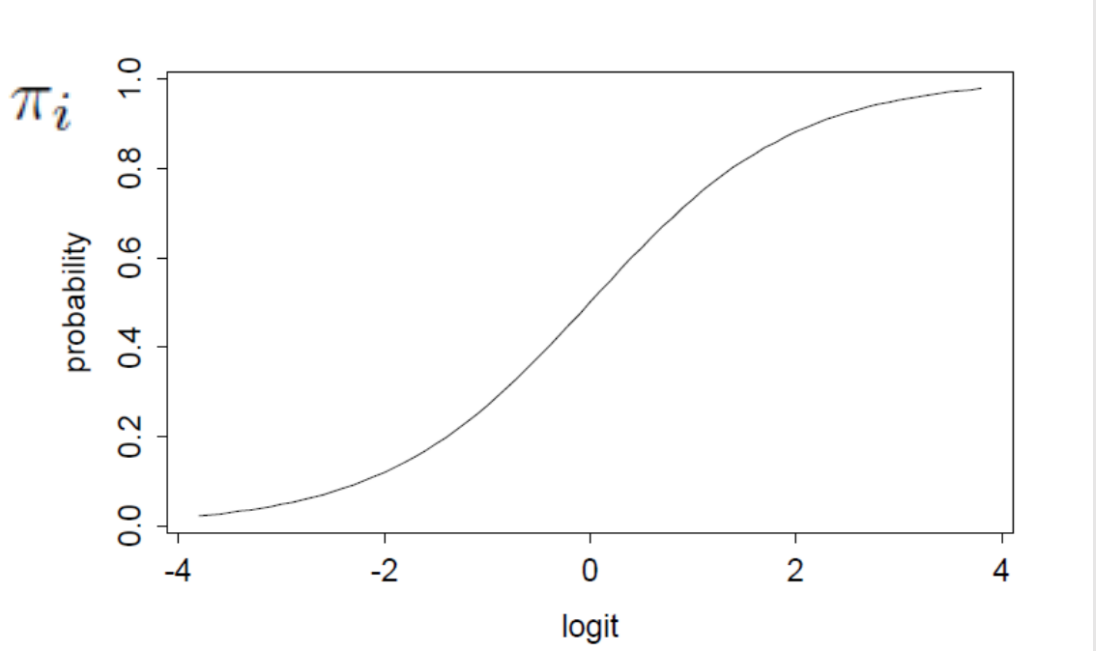
As a result, the odds of disease can be defined by:

$$\frac{p_i}{1 - p_i} = e^{\left(\beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}\right)}$$

On a logarithmic scale, the odds of disease would be:

$$log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}$$

Logit function [**link** function]



$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

On the logit scale we come back to a **linear** model

OR 'crude' (univariable analysis):

To obtain the OR 'crude' in the LR model (i.e. when **only one** predictor is in the model, the exposure) assuming that the exposure factor is a dichotomous variable (0,1) the odds would be:

$$Odds_1 = \frac{p_1}{1-p_1} = e^{\beta_0+\beta_E}$$

$$Odds_0 = \frac{p_0}{1-p_0} = e^{\beta_0}$$

$$OR_{crude} = \frac{Odds_1}{Odds_0} = e^{\beta_E}$$

Role of stem cell renewal factor BMI-1 in primary and metastatic melanoma: binary covariates

	n	Univariate OR	*p*-value
$p16^{ink4a}$ low *vs.* high	35/29	3.0 (1.0–8.6)	0.04
BMI-1 high *vs.* Low	41/23	4.5 (1.3–15.6)	0.02
$p16^{ink4a}$ low/BMI-1 high *vs.* others	22/42	3.2 (1.4–7.3)	0.005

Y=presence of metastasis

Interpretation of the LR coefficients [binary covariates]

	CASE (Y=1)	CONTROL (Y=0)
E (X=1)	P(Y | X=1)	1-P(Y | X=1)
Not E (X=0)	P(Y | X=0)	1-P(Y | X=0)

$$\frac{\frac{exp(\alpha+\beta)}{1+exp(\alpha+\beta)} * \frac{1}{1+exp(\alpha)}}{\frac{exp(\alpha)}{1+exp(\alpha)} * \frac{1}{1+exp(\alpha+\beta)}} = \frac{exp(\alpha+\beta)}{exp(\alpha)} = exp(\beta)$$

	CASE (Y=1)	CONTROL (Y=0)
E (X=1)	$\frac{exp(\alpha+\beta)}{1+exp(\alpha+\beta)}$	$\frac{1}{1+exp(\alpha+\beta)}$
Not E (X=0)	$\frac{exp(\alpha)}{1+exp(\alpha)}$	$\frac{1}{1+exp(\alpha)}$

Here we denote with α the intercept of the model

The intercept α in the model is the log-odds of disease (i.e. to be a case) in the unexposed.

Interpretation of the LR coefficients [continuous exposure/covariates]

In linear regression: If x changes by one unit, the mean of y is expected to change by β_1 units.

Relation between p(x)=P(y=1 | x) and x is linear in logits:

$$g(x) = log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Thus: change in x by one unit ➡ change in logit of p(x) by β_1 units

odds ratio = exp(β_1) is a measure for an increase in risk (in odds) when x changes by one unit.

logit-increase when x changes by k units: $log(OR) = \left(\beta_0 + \beta_1(x+k)\right) - (\beta_0 + \beta_1 x)$

OR for change of x by k units: $exp(k\beta_1) = exp(\beta_1)^k = OR^k$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Example: a study on prostate cancer

Increasing levels of phosphatase are related to the presence of nodal metastases?

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9919	0.6033	1.64	0.1001
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058

Interpretation of the intercept is quite theoretical: log-odds of disease when log2(Phosphatase)=0, i.e when Phosphatase =1

The normal range for serum Phosphatase level is 20 to 140 IU/L

OR when phosphatase changes by a factor of 2:

$$exp(2.4198) = 11.24$$

OR for a change by a factor of 1.5:

log(1.5,base=2)=0.585

$$1.5 = 2^{0.585}$$

$$OR = 11.24^{0.585} = 4.1$$

Interpretation of the LR coefficients [categorical covariates]

For categorical or ordinal X one has to introduce binary ***dummy variables***, with a category as **reference**.

X categorical with 3 levels (A, B, C):

$$\log(Odds) = \beta_0 + \beta_B(X = B) + \beta_C(X = C)$$

With three parameters: β_0, β_B and β_C, and X = A (reference)

Then:

If X=A: $\boldsymbol{log(Odds) = \beta_0}$
If X=B: $\log(Odds) = \beta_0 + \beta_B$
If X=C: $\log(Odds) = \beta_0 + \beta_C$

If the reference coding is changed (X=C is reference) a new model is formulated:

$$\log(Odds) = \beta_{0,new} + \beta_{A,new}(X = A) + \beta_{B,new}(X = B)$$

Where: X = C (reference) and:

If X=A: $\log(Odds) = \beta_{0,new} + \beta_{A,new}$
If X=B: $\log(Odds) = \beta_{0,new} + \beta_{B,new}$
If X=C: $\boldsymbol{log(Odds) = \beta_{0,new}}$

The *new* model parameters and the *old* model parameters are related:

$\beta_{0,new} = \beta_0 + \beta_C$ and $\beta_{A,new} = -\beta_c$

we have: $\beta_{0,new} + \beta_{A,new} = \beta_0$ which is: $(\beta_0 + \beta_C) + \beta_{A,new} = \beta_0$

$\beta_{B,new} = \beta_B - \beta_C$ and we have: $\beta_{0,new} + \beta_{B,new} = \beta_0 + \beta_B$ which is: $(\beta_0 + \beta_C) + \beta_{B,new} = \beta_0 + \beta_B$

```
glm(formula = chd ~ wt4, family = binomial(link = logit), data = wcgs)

Coefficients:
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  -2.9392   0.1622      -18.117   < 2e-16 ***
wt4(155,170] 0.4261    0.2028      2.101     0.035628 *
wt4(170,182] 0.9029    0.2055      4.393     1.12e-05 ***
wt4(182,320] 0.6843    0.2036      3.361     0.000777 ***
---
Signif. codes:  0 '***'
```

Each class is compared to the reference class:

(155 – 170] pounds vs <= 155 pounds -> OR 1.53

(170 – 182] pounds vs <= 155 pounds -> OR 2.47

(182 – 320] pounds vs <= 155 pounds -> OR 1.98

Relationship between CHD (coronary heart disease) and body weight.

Body weight in 4 groups:

<= 155 pounds [reference]
(155 – 170] pounds
(170 – 182] pounds
(182 – 320] pounds

Estimating the 'adjusted' OR:

To obtain the adjusted OR in the presence of potential *J confounding* variables C in the LR model:

Assuming here that the exposure factor E is a dichotomous variable (0,1):

$$Odds_1 = \frac{p_1}{1-p_1} = e^{\beta_0+\beta_E+\sum_{j=1}^{J}\beta_j C_j}$$

$$Odds_0 = \frac{p_0}{1-p_0} = e^{\beta_0+\sum_{j=1}^{J}\beta_j C'_j}$$

$$\Rightarrow \quad OR_{adjusted} = \frac{Odds_1}{Odds_0} = e^{\beta_E+\sum_{j=1}^{J}\beta_j\left(C_j-C'_j\right)}$$

The adjusted OR is obtained under the assumption $C_j = C'_j$: $OR_{adjusted} = e^{\beta_E}$

To ensure the *validity* of the OR adjusted, it is necessary to check that there is no **interaction effect** between the confounding variables and the exposure factor

$$0.9 < \frac{\widehat{OR}_{adjusted}}{\widehat{OR}_{crude}} < 1.1$$

⇒ No confounding effect (data-driven…)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Interaction again… [epi jargon: effect modification]

To correctly estimate the adjusted OR, it could be necessary to verify if there are **interactions** between the exposure and potential confounding variables. If any significant interaction terms are found, the OR **will depend** also on the coefficient associated with the interaction term.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_E E + \beta_C C + \gamma(E * C)$$

Simplest case (E binary exposure):

$$\log(Odds_1) = \beta_0 + \beta_E + \beta_C C + \gamma C$$

$$\log(Odds_0) = \beta_0 + \beta_C C'$$

$$OR_{adjusted} = \frac{Odds_1}{Odds_0} = e^{\beta_E + \beta_C(C-C') + \gamma C}$$

Usually, a first model is fitted **with** the interaction terms and compared to a second model **without** it by means of a ***likelihood ratio test***

Depending on the possible combinations of C and C' we obtain **different values** for the adjusted OR

If we assume C=C': $OR_{adjusted} = e^{\beta_E + \gamma C}$

Example of logistic regression as a diagnostic model

Renal artery stenosis is a rare cause of hypertension.

The reference standard for **diagnosing** renal artery stenosis, renal angiography, is **invasive** and **costly**.

Aim: develop a *prediction rule* for renal artery stenosis from clinical characteristics.

The rule might then be used **to select patients** for renal angiography.

Logistic regression analysis performed with data from **477** hypertensive patients who underwent renal angiography. A simplified **prediction rule** was derived from the regression model for use in clinical practice.

Age, sex, atherosclerotic vascular disease, recent onset of hypertension, smoking history, body mass index, presence of an abdominal bruit, serum creatinine concentration, and serum cholesterol level were selected as predictors.

Diagnostic accuracy of the regression model was similar to that of renal scintigraphy. The conclusion was that this clinical prediction model **can help to pre-select patients** for renal angiography in an efficient manner by reducing the number of angiographic procedures without the risk of missing many renal artery stenosis.

Krijnen et al., A clinical prediction rule for renal artery stenosis. Annals of Internal Medicine(1998)

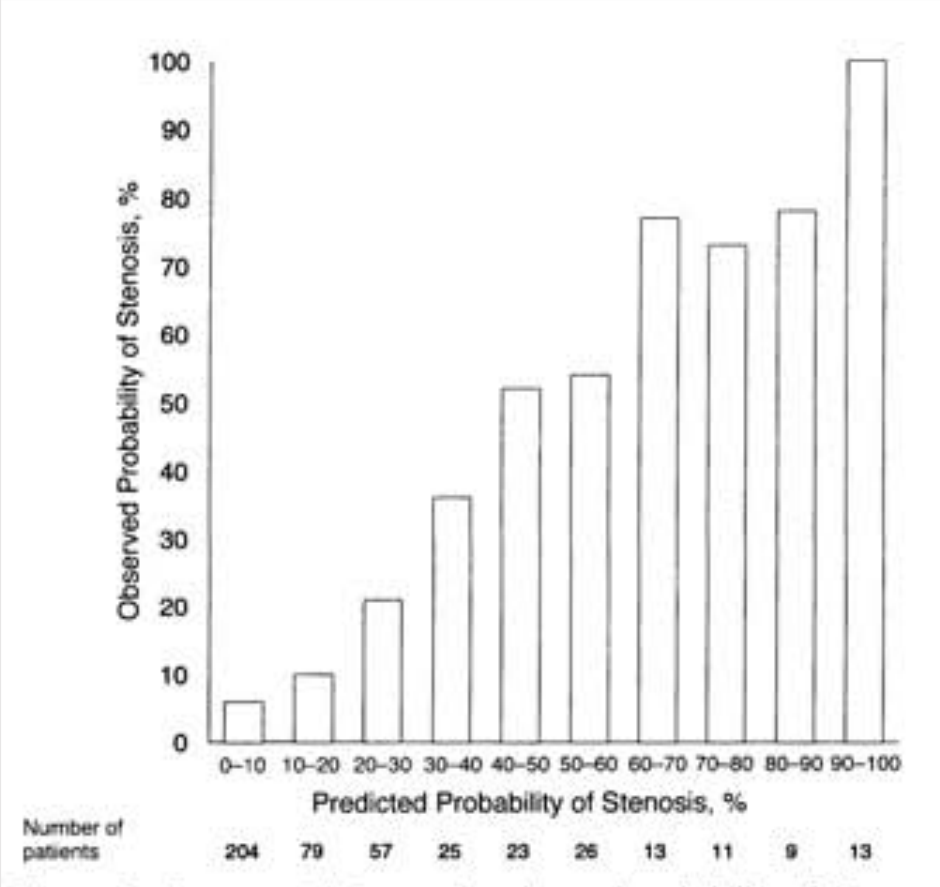
Derivation of Scores in the Prediction Rule

The multivariable logistic regression model can be written as:

predicted probability of stenosis = $1/1 + e^{-(LP)}$,

where linear predictor $LP = -7.859 + 0.059 \times \text{age} + 0.033 \times (75 - \text{age}) \times \text{ever smoked} - 0.996 \times \text{sex} + 0.585 \times \text{atherosclerotic vascular disease} + 0.642 \times \text{recent onset} - 1.027 \times \text{obesity} + 1.693 \times \text{abdominal bruit} + 0.502 \times \text{hypercholesterolemia} + 0.032 \times \text{serum creatinine concentration}$.

The regression coefficients were multiplied by a shrinkage factor of 0.88, which was derived from bootstrapping procedures. Shrinkage of the regression coefficients aims to improve calibration of predictions in future patients:

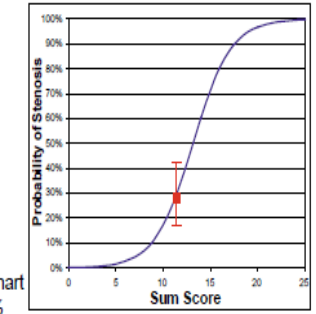


The area under the ROC curve was 0.84 on the full data set and 0.82 after a boostrap procedure

Prediction rule for renal artery stenosis

Predictor		Value	Score
Smoking	former or current =1	1	-
Current age	years	45	4.4
Gender	male = 1	1	0
Atherosclerotic vascular disease*	yes = 1	0	0
Onset of hypertension within 2 years	yes = 1	1	1
Body mass index >= 25 kg/m2	yes = 1	0	2
Presence of abdominal bruit	yes = 1	0	0
Serum creatinine concentration	µmol/L	112	4.1
Serum cholesterol level > 6.5 mmol/L**	yes = 1	0	0
Sumscore			*11*

	Formula	Score chart
Predicted probability of renal artery stenosis	28%	*25%*
Confidence interval	17% -	43%



See figure for graphical illustration

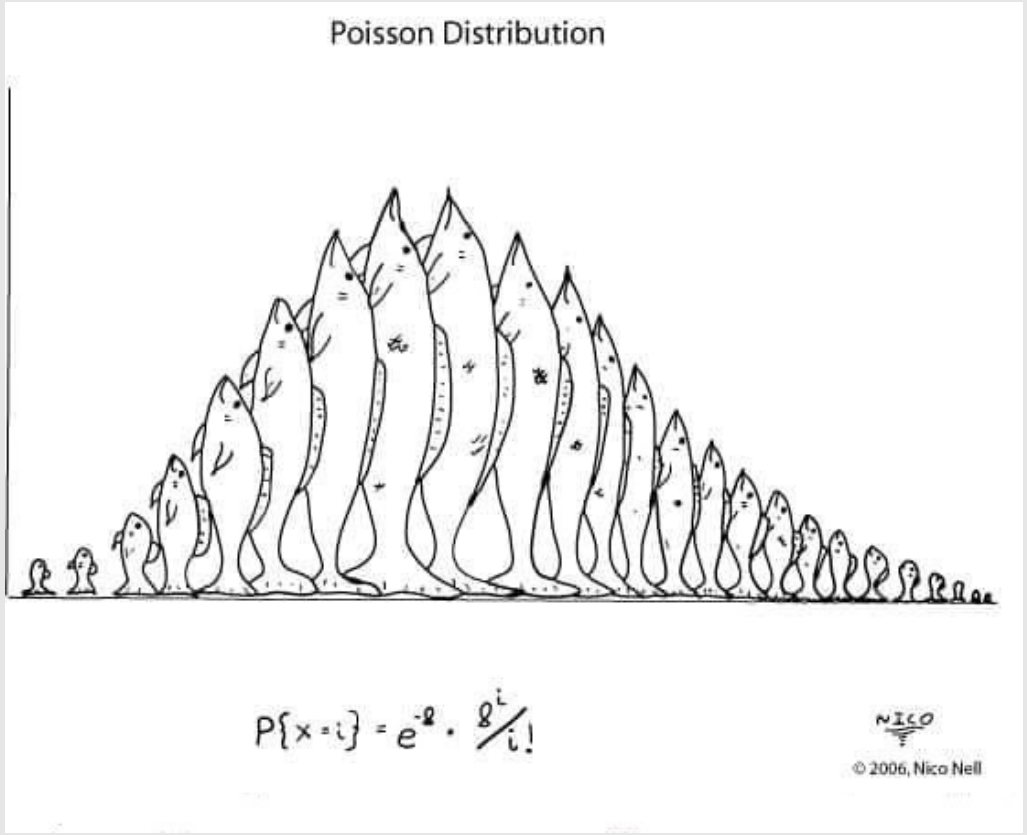
* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery

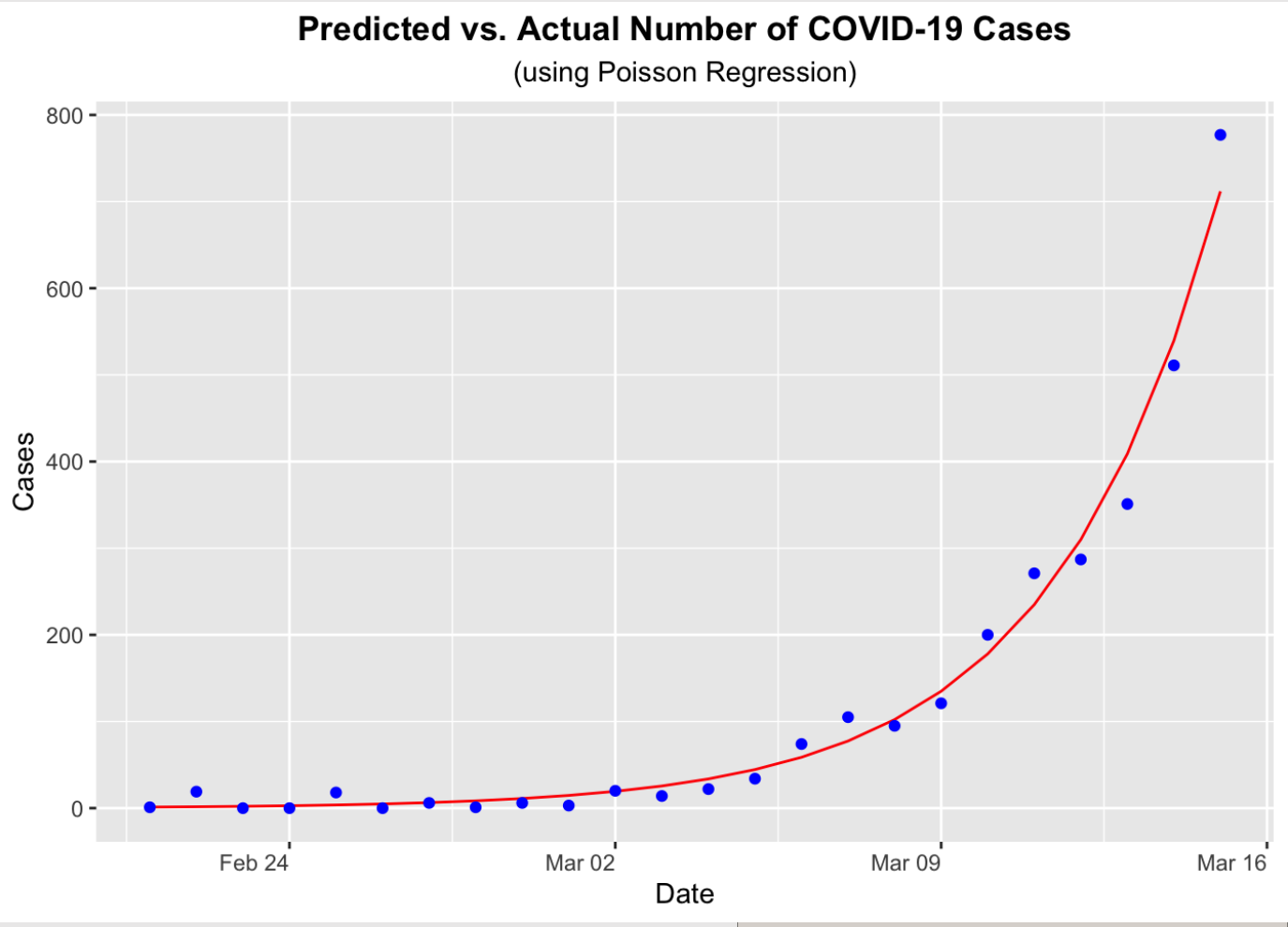
** or cholesterol lowering therapy

45-year-old male with recent onset of hypertension.

The sum score was 11, the estimated probability or renal artery stenosis was 28% [95% confidence interval 17–43%].

Poisson regression

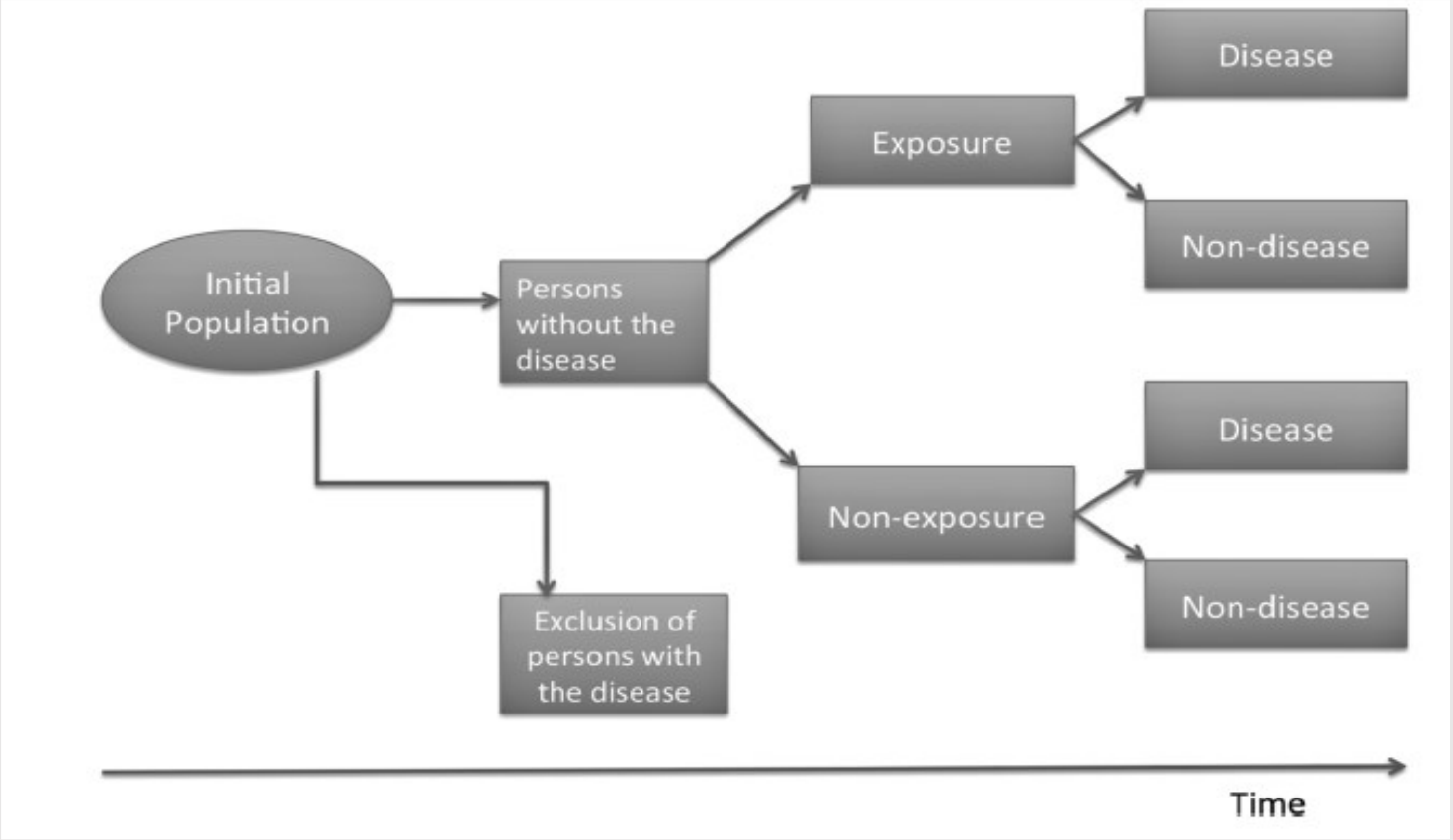




Remind: Population based / Cohort Studies / [RCTs]

In a pop.based/cohort/[RCT] study, a population is selected on the basis of their exposure/treatment and **followed over a period of time** to determine the occurrence of a disease or any other health-related event.

The **incidence** of disease could then compared in exposed and unexposed groups.



UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

In the explanatory setting the main goal is to estimate **disease incidence among exposed** and **unexposed** individuals and then quantifying the **magnitude** of the association between exposure and disease.

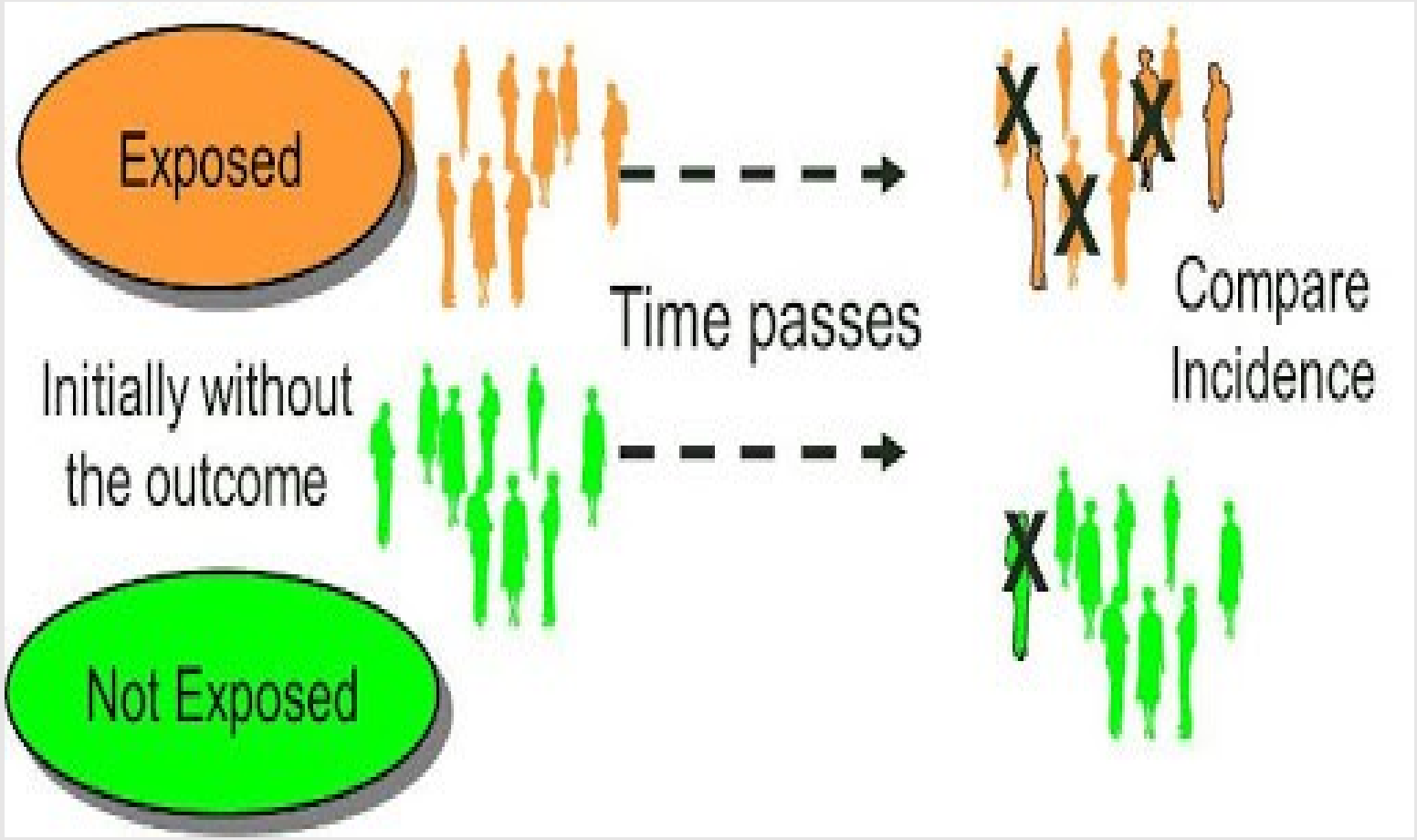
In the prognosis/prediction setting the main goal is to estimate **the probability to develop the outcome in a certain time interval, according to subject's characteristics.**

Advantages of the pop.based/cohort study/[RCTs]:

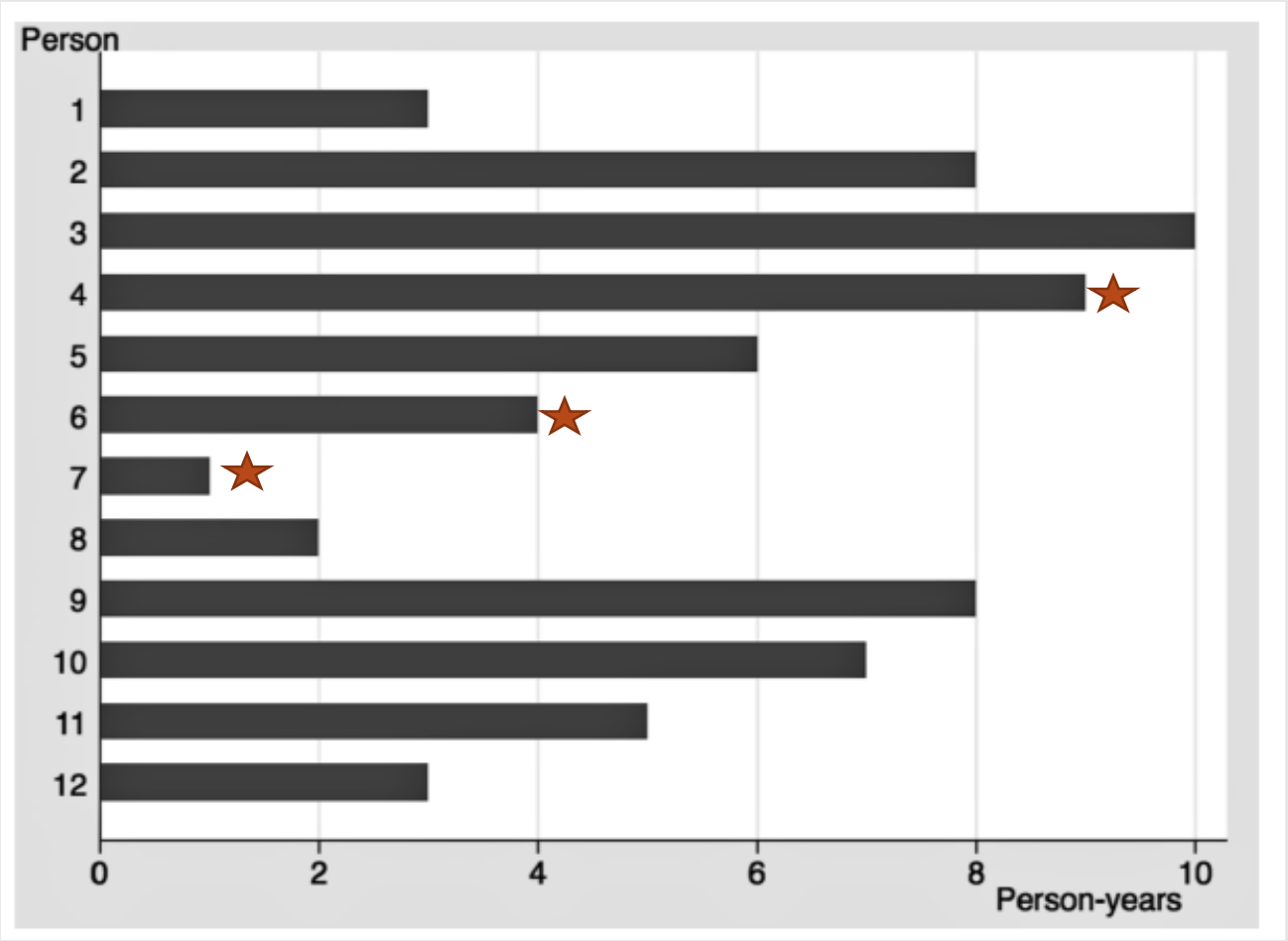
- **Temporal sequence** between the exposure under study and the disease or any other health-related event can be established
- Determination of disease incidence in each exposure group and investigation of the effect of the exposure of interest on possible ***multiple* outcomes** (not all a-priori defined)

Disadvantages

- Susceptibility to (differential) **loss of subjects** during follow up, which may introduce **selection bias** and thus affect the internal validity of the study
- **Comparability** of subjects who remain in the study and those who are lost, **by exposure status**, must be determined in order to assess potential **confounders/bias**
- **Expensive** (time/costs) [if **primary** data collection]
- Not suitable for **rare** diseases



Recap: person-years & Incidence rate



Person-years computation using a hypothetical cohort of 12 subjects observed for a period of 10 years.

3 developed the disease (4,6, and 7).

Total person-years (66 p-years) : **sum** of the individual time at risk of all subjects.

$$IR = \frac{a}{L} = \frac{\#\ new\ cases}{\sum_{i=1}^{n} \Delta t_i}$$

Δt_i observation time subject *i*

L total accumulated p-years

$$IR = \frac{3}{3 + 8 + \cdots + 3} = \frac{3}{66\ person - years} = 4.54\ x\ 100\ py$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

	# of new cases	p-years	IR
E	a_1	L_1	$IR_1=a_1/L_1$
Not E	a_0	L_0	$IR_0=a_0/L_0$

Incidence Rate Ratio

$$IRR = \frac{IR_1}{IR_0} = \frac{a_1/L_1}{a_0/L_0}$$

Hypothetical data on the association between smoking and cardiovascular disease	Smokers	Nonsmokers
Number of newly diagnosed cases of cardiovascular disease	43	52
Person-years	727	1820

For every 1000 smokers observed in a year, there were 59.1 cases of cardiovascular disease

For every 1000 nonsmokers observed in a year, there were 28.6 new cases of cardiovascular disease

$$\widehat{IRR} = \frac{59.1}{28.6} = 2.06$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Recap: Cumulative Incidence

When the period of observation of each subject in the cohort study is [approximately] constant ($\Delta t_i = t$) (and we observe a *fixed/closed* cohort) the occurrence of an event could be estimated with the cumulative incidence:

$$CI = \frac{\#\ new\ cases}{n}$$

n = population at risk **at the start** of the study

	# of new cases	Total	CI
E	a_1	n_1	$CI_1 = a_1/n_1$
Not E	a_0	n_0	$CI_0 = a_0/n_0$

$$RR = \frac{CI_1}{CI_0} = \frac{a_1/n_1}{a_0/n_0}$$

Complications	Obese	Nonobese	Relative risk
Present	95	77	
Absent	335	463	
Total	430	540	$\hat{RR} = \frac{0.221}{0.142} = 1.55$
Cumulative incidence	$CI_1 = 0.221$	$CI_0 = 0.142$	

Hypothetical study on obesity of mothers and complications during childbirth

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

A regression model for counts/rates

The Poisson regression model estimates the incidence of a disease or health-related event ***under different conditions***.

To determine the incidence, it is necessary to compute the **number of new cases** during the **observation period** and identify the initial conditions of the study, such as the **type of exposure** at baseline and specific values of the potential **confounding** variables.

The Poisson model *establishes a relationship/[makes a prediction]* between the **expected** number of cases and the exposure while **controlling** for potential **confounders**.

Recap:
The Poisson probability distribution can be used when the random variable represents the number of cases (successes) under **3** conditions:

- in a very large number of independent Bernoulli trials [when the constant probability of success is small]
- for a unit of time (e.g., day, month, or year)
- on a unit area (e.g., square meter, square kilometer, or square mile) or volume (e.g., cubic meter or cubic centimeter)

The Poisson regression model establishes a relationship between the **expected number of cases** and the exposure [while controlling for potential confounders/indep.pred.]:

$$\mu_i = P_i e^{\beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}}$$

μ_i **expected value of new cases** in *condition i* : a combination of the values of the covariates. [We assume that the number of new cases is a RV that has a Poisson distribution]

P_i **population** in the *i-th group* of exposure
- Person-time units → Incidence rate
- Population at baseline → Cumulative incidence

E_i *Exposure* variable

C_{ij} *j-th confounding* variable

[or general predictors]

β_0 Intercept of the model. Exp(β_0) : expected incidence of the number of new cases when the exposure and the confounding variables take the value of zero.

We are **assuming** that the response variable is a count of events ***occurring independently*** among different subgroups [number of newly diagnosed cases of kidney cancer at different hospitals every year] and that this random variable follows a Poisson distribution.

We also **assuming** that μ is linked to the **exponential** of a linear function of the candidate predictors; so the changes in the incidence resulting from the combined effects of the exposure and the confounding variables are multiplicative.

[incidence of events]

$$\frac{\mu_i}{P_i} = e^{\beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}}$$

$$ln(\mu_i) = ln(P_i) + \beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}$$

Since the model contains the variable $ln(P_i)$ there is no need to estimate the coefficient for this variable, referred to as an **offset**

When the exposure E is a dichotomous variable coded as 1 (presence) and 0 as absence of the factor, the incidence of events in the two groups is estimated by the following expressions:

$$I_1 = \frac{\mu_1}{P_1} = e^{\beta_0+\beta_E+\sum_{j=1}^{J}\beta_j c_j} \quad \text{exposed}$$

$$I_0 = \frac{\mu_0}{P_0} = e^{\beta_0+\sum_{j=1}^{J}\beta_j c_j} \quad \text{unexposed}$$

$$\Rightarrow \quad RR = \frac{I_1}{I_0} = e^{\beta_E+\sum_{j=1}^{J}\beta_j\left(c_j-c_j'\right)}$$

The adjusted relative risk is obtained under the assumption $C_j = C_j'$: $\widehat{RR}_{adjusted} = \boldsymbol{e^{\beta_E}}$

$$0.9 < \frac{\widehat{RR}_{adjusted}}{\widehat{RR}_{crude}} < 1.1$$

⇒ No confounding effect (data-driven…)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Again, to estimate the adjusted RR, it is necessary to verify that there are **no interactions** between the exposure and potential confounding variables.

If any significant interaction terms are found, the RR will depend on the coefficient associated with the interaction term.

$$I_i = \frac{\mu_i}{P_i} = e^{\beta_0+\beta_E E+\beta_C C+\gamma(E*C)}$$

$$I_1 = \frac{\mu_1}{P_1} = e^{\beta_0+\beta_E+\beta_C C+\gamma C}$$

$$I_0 = \frac{\mu_0}{P_0} = e^{\beta_0+\beta_C C'}$$

$$RR = \frac{I_1}{I_0} = e^{\beta_E+\beta_C(C-C')+\gamma C}$$

If we assume that $C_j = C_j'$: $RR = e^{\beta_E+\gamma C}$

Poisson regression take into account a *crucial* issue not faced by other regression techniques (linear/logistic).

From the data design, different subjects may have **different person-times** of exposure.

Analysing risk factors while **ignoring differences** in person-times is wrong.

Note that logistic regression **completely ignores** this aspect [**difference: cohort vs case-control**]. Observation time **is not accounted for** in the evaluation of the probability of the event.

Supplementary Materials

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Interpreting the output of a Poisson regression model

We are interested in the effect of smoking on death rate, adjusting for age (treated as categorical, 5 levels):

$$ln(Death\ rate) = -7.92 + 0.35 * Smoke + 1.48 * AgeB + 2.63 * AgeC + 3.35 * AgeD + 3.70 * AgeE$$

-7.92 = log death rate for Age category A (reference) and non-Smokers

0.35 = difference in the log death rates for Smokers compared to non-Smokers (at the same age!)

Note that we do not have anything estimated for the offset term, we use it only for the interpretation

$$ln(\mu_i) = ln(P_i) + \beta_0 + \beta_E E_i + \sum_{j=1}^{J} \beta_j C_{ij}$$

[In this dataset we had person-years as time-dimension]

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

$$ln(Death\ rate) = -7.92 + 0.35 * Smoke + 1.48 * AgeB + 2.63 * AgeC + 3.35 * AgeD + 3.70 * AgeE$$

How would we compare the death rate for smokers vs not-smokers ?

$e^{\beta_E} = e^{0.35} = 1.42$ Smokers have 1.42 times the death rate of non-smokers (at the same age!)

How would we calculate the **expected death rate** for a smoker in AgeC ?

$$ln(Death\ rate) = -7.92 + 0.35 * Smoke + 2.63 * AgeC$$

$$ln(Death\ rate) = -4.94$$

$$Death\ rate = e^{-4.94} \qquad Death\ rate = 0.00717\ per\ person - year$$

$$Death\ rate = 715\ per\ 100.000\ person - years$$

$$ln(Death\ rate) = -7.92 + 0.35 * Smoke + 1.48 * AgeB + 2.63 * AgeC + 3.35 * AgeD + 3.70 * AgeE$$

We follow a group of 9783 non-smokers in AgeD for 25 years.

Based on the model we have fit, **how many deaths** would we expect?

$$\frac{\mu_i}{P_i} = \frac{Exp(Deaths)}{9783 * 25} = e^{-7.92+3.35}$$

$$\frac{Exp(Deaths)}{244575} = e^{-4.57}$$

$$Exp(Deaths) = 0.010 * 244575 = 2446$$