Measures of Disease-Exposure Association

- Relative risk RR
- Odds ratio OR
- The OR as an approximation to the RR
- *Symmetry* of roles of disease and exposure in the OR
- Excess Risk
- Regression-based estimates

Measures of Disease-Exposure Association

Purposes of epidemiology:

- **quantification** of the occurrence of a disease [*descriptive* studies]
- **strength of the association*** between exposure and the onset of the event [*analytical* studies]

Estimate measures of *disease-exposure* association (or **measures of effect**).

Disease frequency in the **exposed group** is compared with the frequency of disease in the **group of those not exposed**, making use of the appropriate measure of occurrence.

This comparison can occur in two ways: in **absolute** terms and in **relative** terms.

***note that we are not using the term *causal* effect!**

In a *general* sense, each disease could be *associated* with one (or more…) factors.

In a *quantitative* sense, we can start by comparing **the occurrence** of a disease in two [or more..] groups that differ by *one certain* feature [***univariable*** analysis].

- absolute scale: **difference** between two prevalences, two risks (Cum Inc) or two incidence rates
- relative scale: **ratio** of two prevalences, two risks (Cum Inc) or two incidence rates
- *attributable risk*: ***proportion of cases*** *attributable to exposure in a population*

Does a mother's marital status **affect** the risk of a baby's death in the first year? ***To what extent*** ? And what about birthweight?

Relative risk

The Relative Risk for an outcome D associated with a *binary* risk factor E*, denoted by RR, is defined as follows:

$$RR = \frac{P(D|E)}{P(D|\bar{E})}$$

	D	Not D	Tot
E	a	b	a+b
Not E	c	d	c+d
Tot	a+c	b+d	N

*we will take into account **uncertainty** around these estimates in particular related to **sample size**

*Could be extended to **continuous** risk factors through regression

Some simple implications immediately follow:

RR
- <1 : **lower** risk or probability of *D* when exposed than when unexposed
- =1 : ***null*** value equivalent to saying that *D* and *E* are independent $P(D|E) = P(D|\bar{E})$
- >1 : **greater** risk or probability of *D* when exposed than when unexposed

The Relative Risk is the basis of a *multiplicative model for risk* :

$$Risk_{Exposed} = Risk_{unexposed} * RR$$

Baseline Risk

If you smoke cigarettes, your lifetime risk of lung cancer increases tenfold, i.e., the Relative Risk for lung cancer associated with cigarette smoking is 10.

Restrictions on the range:

$$0 < RR \leq \frac{1}{P(D|\bar{E})}$$

For instance, if $P(D|\bar{E}) = 1/3$ (30%) then RR $\leq$ 3 since $P(D|E) \leq 1$

This *restriction* could become an issue with *common* diseases

RR is ***not symmetric*** in the role of the two factors D and E.

The Relative Risk for E associated with D is a different measure of association:

$$\frac{P(D|E)}{P(D|\bar{E})} \neq \frac{P(E|D)}{P(E|\bar{D})}$$

Relative Effect

Infant Mortality	Mother's Marital Status	
	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

The Relative Risk for infant mortality in the U.S., associated with a mother being unmarried at the time of birth, is:

$$RR = \frac{16712}{1213854} : \frac{18784}{2897205} = 2.12$$

the risk of an infant death with an unmarried mother is **double** the risk w.r.t. mother is married.

The RR for infant mortality in the U.S., associated with a low-birthweight infant, is:

	Birthweight		
Infant Mortality	Low Birthweight	Normal Birthweight	Total
Death	21,054	14,442	35,496
Live at 1 year	271,269	3,804,294	4,075,563
Total	292,323	3,818,736	4,111,059

$$RR = \frac{21054}{292323} : \frac{14442}{3818736} = 19.0$$

Much greater association of birthweight on infant mortality than we saw for a mother's marital status.

Odds Ratio

An alternative quantity that is used in health research is the ***odds*** of D as given by: $\frac{P(D)}{P(\bar{D})}$

The odds gives the likelihood of D occurring relative to it not occurring: "how likely am I to win?" as compared to "how likely am I to lose?"

An even odds event D (odds of D are 1) is equivalent to P(D)=1/2, that is, the **same chance** of winning as losing.

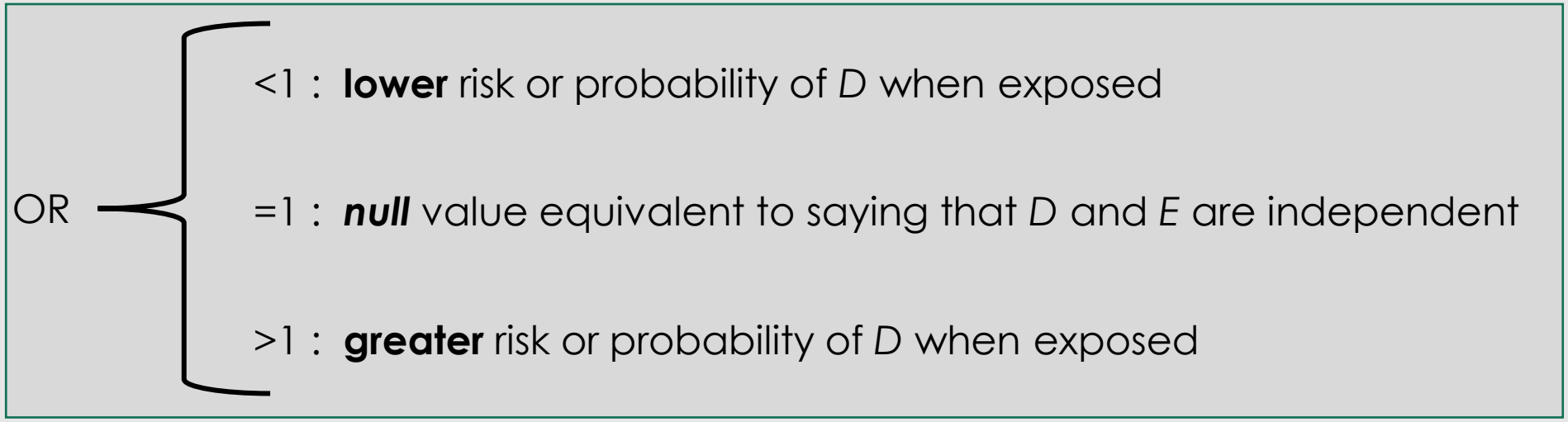
$$Odds = \frac{Probability\ of\ event}{1 - Probability\ of\ event}$$

The **Odds Ratio** measures association by **comparing the odds** of D in the exposed and unexposed.

The Odds Ratio for D associated with E is defined by:

$$OR = \frac{P(D|E)}{P(\bar{D}|E)} : \frac{P(D|\bar{E})}{P(\bar{D}|\bar{E})} \qquad OR = \frac{P_{exp}/(1 - P_{exp})}{P_{unexp}/(1 - P_{unexp})}$$

Relative Effect



The Odds Ratio is also the basis of a *multiplicative model* for the risk of D.

Like RR, OR > 0, but unlike RR, OR *has no upper limit* whatever the *baseline* risk $P(D|\bar{E})$ is.

Thus, the OR can be effectively used as a measure for association even when $P(D|\bar{E})$ is large.

Infant Mortality	Mother's Marital Status: Unmarried	Mother's Marital Status: Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

The OR for infant mortality associated with an unmarried mother is:

$$OR = \left[\frac{16712}{1213854} : \frac{1197142}{1213854}\right] : \left[\frac{18784}{2897205} : \frac{2878421}{2897205}\right] = 2.14$$

Associated with low birthweight, the OR is: $$OR = \left[\frac{21054}{292323} : \frac{271269}{292323}\right] : \left[\frac{14442}{3818736} : \frac{3804294}{3818736}\right] = 20.4$$

Infant Mortality	Birthweight: Low Birthweight	Birthweight: Normal Birthweight	Total
Death	21,054	14,442	35,496
Live at 1 year	271,269	3,804,294	4,075,563
Total	292,323	3,818,736	4,111,059

The odds ratio as an *approximation* to the relative risk

If the risk of disease is **low** - that is, the disease is **rare** - in both exposed and unexposed, $P(\bar{D}|E)$ and $P(\bar{D}|\bar{E})$ are both close to 1 and the OR and the RR are *approximately* equal:

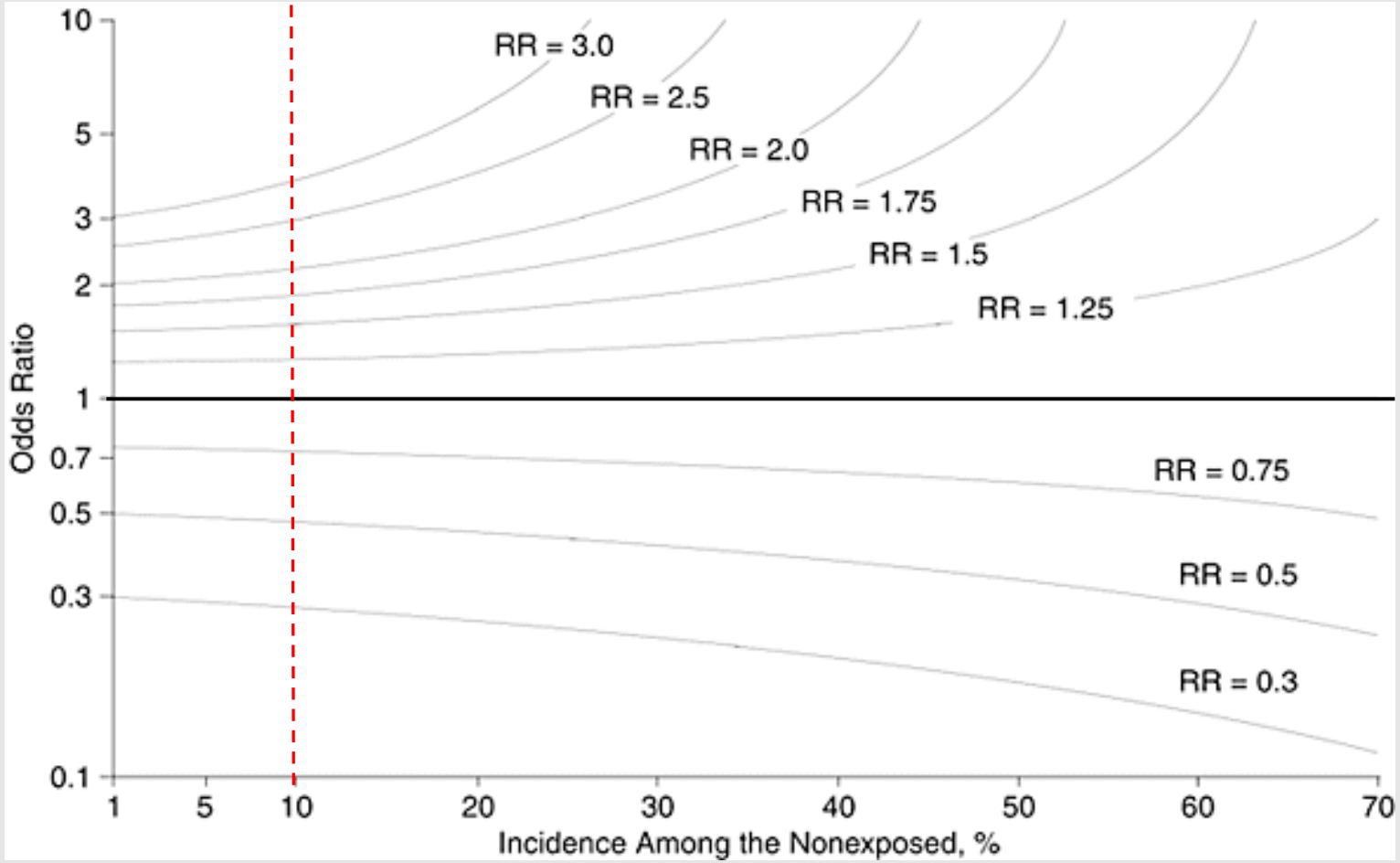
$$P(\bar{D}|E) \approx P(\bar{D}|\bar{E}) \approx 1$$

$$OR \approx \frac{P(D|E)}{P(D|\bar{E})} = RR$$

Generally, OR is *similar* to the RR when the *sum of the risks* - in the exposed and unexposed - is < 0.1

https://jamanetwork.com/journals/jama/fullarticle/188182

The relationship between relative risk (RR) and odds ratio (OR) by *baseline* incidence of the outcome:



When the ***incidence*** of an outcome is low (<10%), the odds ratio is *close* to the relative risk.

The more frequent the outcome becomes, the more the odds ratio will **overestimate** the relative risk when it is more than 1 or **underestimate** the relative risk when it is less than 1.

https://www.youtube.com/watch?v=76W4Wymv2Ec

Symmetry of disease and exposure in the odds ratio

The Odds Ratio is confusing when first encountered, particularly in contrast to the simplicity of the Relative Risk. Why is the Odds Ratio then used so often*? A fundamental reason is that the Odds Ratio is ***symmetric*** in the roles of D and E.

Reversing the roles of D and E makes **no difference** in Odds Ratio : this is the **key** to estimating association between an exposure and disease in **case-control studies** [**block 2 !!**].

$$OR = \frac{P(D|E)}{P(\bar{D}|E)} : \frac{P(D|\bar{E})}{P(\bar{D}|\bar{E})} = \frac{P(D\&E)/P(E)}{P(\bar{D}\&E)/P(E)} : \frac{P(D\&\bar{E})/P(\bar{E})}{P(\bar{D}\&\bar{E})/P(\bar{E})}$$

$$= \frac{P(D\&E)}{P(\bar{D}\&E)} : \frac{P(D\&\bar{E})}{P(\bar{D}\&\bar{E})} = \frac{P(D\&E)}{P(D\&\bar{E})} : \frac{P(\bar{D}\&E)}{P(\bar{D}\&\bar{E})}$$

$$= \frac{P(D\&E)/P(D)}{P(D\&\bar{E})/P(D)} : \frac{P(\bar{D}\&E)/P(\bar{D})}{P(\bar{D}\&\bar{E})/P(\bar{D})}$$

$$= \frac{P(E|D)}{P(\bar{E}|D)} : \frac{P(E|\bar{D})}{P(\bar{E}|\bar{D})}$$

*Also for the popularity of the logistic regression model [**block 3 !!**]…

Relative vs Absolute Effects…

Hours of inactivity slumped in front of a television **sharply raises the risk** of **dying from a blood clot in the lungs**, say scientists…

https://www.ahajournals.org/doi/10.1161/circulationaha.116.023671

RESEARCH ARTICLE | Originally Published 26 July 2016 |

Watching Television and Risk of Mortality From Pulmonary Embolism Among Japanese Men and Women: The JACC Study (Japan Collaborative Cohort)

Toru Shirakawa, MD, Hiroyasu Iso, MD, PhD, MPH, Kazumasa Yamagishi, MD, PhD, Hiroshi Yatsuya, MD, PhD, Naohito Tanabe, MD, PhD, Satoyo Ikehara, PhD, Shigekazu Ukawa, PhD, and Akiko Tamakoshi, MD, PhD | AUTHOR INFO & AFFILIATIONS

Circulation • Volume 134, Number 4 • https://doi.org/10.1161/CIRCULATIONAHA.116.023671

8.2 cases each 100.000 p-years
vs
2.8 cases each 100.000 p-years

= 5.4 *extra* cases each 100.000 p-years …

Table. HRs for Mortality From Pulmonary Embolism According to Hours Spent Watching Television

	Time Spent Watching Television, h/d			
	<2.5	2.5–4.9	≥5.0	Increment by 2 h (95% CI), *P* Value
Cases, n/person-y	19/678 199	27/562 449	13/157 922	
Mortality rate per 100 000 person-y	2.8	4.8	8.2	
Age- and sex-adjusted HR	1.0	1.6 (0.9–2.8)	2.4 (1.2–4.9)	1.3 (1.0–1.8), 0.06
Multivariable HR*	1.0	1.7 (0.9–3.0)	2.5 (1.2–5.3)	1.4 (1.0–1.8), 0.04

CI indicates confidence interval; and HR, hazard ratio.

*Adjusted for age, sex, body mass index, history of hypertension, history of diabetes mellitus, smoking status, perceived mental stress, educational level, walking activity, and sports activity.

you can also say that
(if this effect is ***causal…***) *a subject* would have to watch 5+ hours a day for nearly 19.000 years to expect one event to happen …

Excess risk

To convey an **absolute** measure of the impact of exposure on risk, the Excess Risk, denoted by ER, could be estimated:

$$ER = P(D|E) - P(D|\bar{E})$$

The Excess Risk uses the same basic components as the Relative Risk (and the Odds Ratio), but looks at the **absolute**, rather than relative, difference in risk levels.

The Relative Risk for lung cancer associated with cigarette smoking is about **6** times as great as the Relative Risk for CHD due to smoking.

On the other hand, the Excess Risk for CHD **is larger** since it is the **most common** disease.

Therefore, from a health policy or public health point of view, cigarette intervention programs may be more important in terms of their **impact** on CHD.

Absolute effect

ER
- <0 a greater risk for D when unexposed than when exposed
- =0 independence of D and E
- >0 a greater risk for D when exposed than when unexposed

$$-1 \leq \mathrm{ER} \leq 1$$

Excess Risk* is the basis of an ***additive*** *model* for risk:

$$Risk_{Exposed} = Risk_{unexposed} + ER$$

Interpretation of the Excess Risk ??

*Note that in epidemiological studies ER is computed as a difference between *cumulative incidence* or *incidence rates* and expressed in terms of *#events* in a *population* over a *period of time*

$P(D) = P(D|\bar{E})$ All **not exposed**: number of cases $\# cases = \boldsymbol{N} * \boldsymbol{P}(\boldsymbol{D}|\bar{\boldsymbol{E}})$

$\# cases = \boldsymbol{N} * \boldsymbol{P}(\boldsymbol{D})$

$P(D) = P(D|E)$ All **exposed**: number of cases $\# cases = \boldsymbol{N} * \boldsymbol{P}(\boldsymbol{D}|\boldsymbol{E})$

Excess Risk = the "excess" number of cases when population members ***are all exposed*** as compared to them ***all being unexposed***.

Example: association between appendectomy and infections

Cumulative Incidence (CI) with appendectomy = 5.3% = 53/1000

Cumulative Incidence (CI) without appendectomy = 1.3% = 13/1000

Excess Risk (ER) = 40/1000= 4/100 (0.04%)

Interpretation:
Subjects who had appendectomy had 4 ***additional cases*** of infection per 100 people compared to subjects who did not have appendectomy.

There were 4 *excess infections* per 100 subjects in the group that had appendectomies, compared to the group without appendectomy.

Absolute effect

Excess Risk for infant mortality in the U.S. in 1991 associated with the mother's marital status:

	Mother's Marital Status	
Infant Mortality	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

$$ER = \frac{16712}{1213854} - \frac{18784}{2897205} = 0.0073$$

Excess Risk for infant mortality associated with low birthweight:

	Birthweight		
Infant Mortality	Low Birthweight	Normal Birthweight	Total
Death	21,054	14,442	35,496
Live at 1 year	271,269	3,804,294	4,075,563
Total	292,323	3,818,736	4,111,059

$$ER = \frac{21054}{292323} - \frac{14442}{3818736} = 0.0682$$

Low birthweight is *more influential* than marital status on both the absolute and relative comparative scales.

We would expect the infant mortality **to increase by 7%** if all births exhibited low birthweight as compared to all those being of normal birthweight (w.r.t **0.7%** in case of marital status).

So…relative or absolute risk measures ? Better both!

Relative measures lose information on risk levels, so you can find relative risks *relatively low* associated with *very high absolute* differences, and viceversa.

Relative and absolute risk measures between incidence rates (per 100.000 pyrs) of disease in smokers and non-smokers:

	Smokers	Not Smokers	RR	ER*
Lung cancer	48.33	4.49	**10.8**	43.84
Cardiovascular disease	294.67	169.54	1.7	**125.13**

For this reason, it is important to estimate (especially in public health studies) also the absolute differences between risks / rates (note that this is possible *in some types of studies* but not in others, block 2).

*ER here is expressed as a difference between incidence rates: an «excess» of cases each 100.000 pyears

For **uncommon** events such as clinically problematic rare adverse events, relative measures will tend to *exaggerate* differences. For **common** events such as therapeutic response, relative measures may *minimize* differences.

Knowledge of the **baseline rates** of the outcome of interest can help understand situations when the absolute difference is very small but the relative effect is very large.

Another possibility is to compute the so-called ***attributable risk*** measures that *combine* some of the advantages of both absolute and relative measures.

Estimating Odds Ratios via Logistic Regression

The logistic regression model allows us to estimate the OR to assess the magnitude of the association between a specific factor and the disease under study taking into account **multiple** covariates **with possibly 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

probability of having the disease for the subject *i*

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

Factors «X» in whatever scale of measure: binary, numerical, categorical..

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

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

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

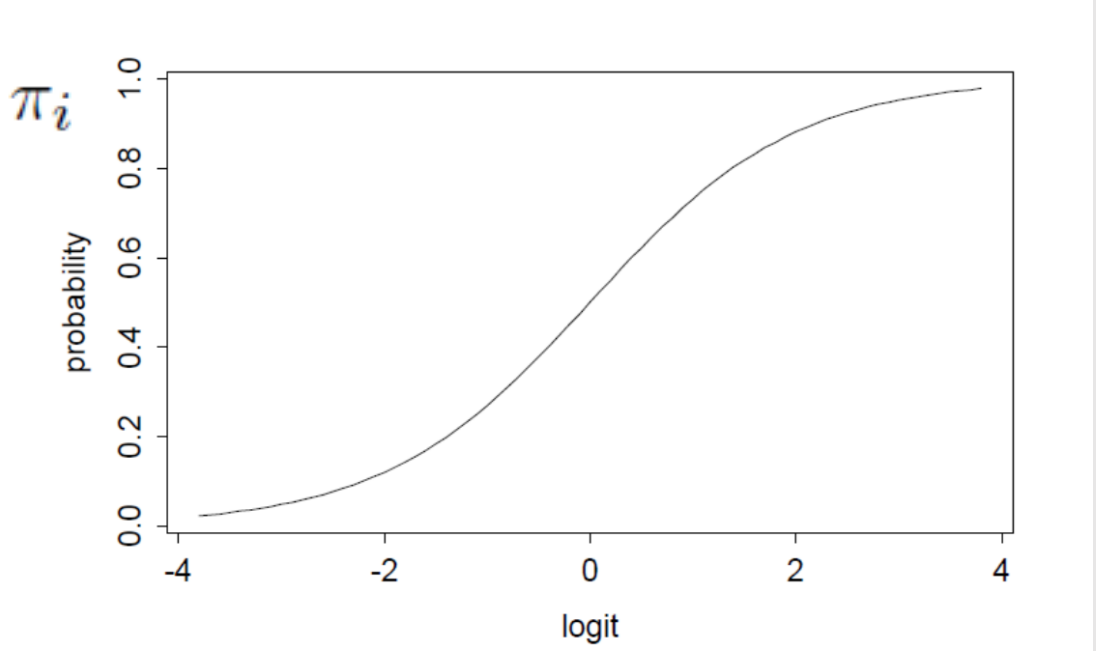
As a result, the odds of disease can be defined by:

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

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

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

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 '<u>crude</u>' (univariable analysis):

Assuming that the only variable of interest is a binary one (E):

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

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

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

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 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: $log(OR) = \big(\beta_0 + \beta_1(x+k)\big) - (\beta_0 + \beta_1 x)$

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

Example: a study on prostate cancer

Increasing levels of phosphatase are associated 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 Relative Risks via Poisson Regression

The Poisson regression model estimates the incidence of an event ***under different conditions***.

The Poisson model ***estimate a relationship****/[makes a prediction]* between the **expected** number of cases and the covariates included in the model.

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 can be written as:

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

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

X_{ij} *j-th covariates*

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** here 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 are **assuming** that μ is linked to the **exponential** of a linear function of the candidate associated factors; so the changes in the incidence resulting from the combined effects of factors are multiplicative.

[incidence of events]

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

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

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 we have a binary variable E:

$$I_1 = \frac{\mu_1}{P_1} = e^{\beta_0+\beta_E}$$ Incidence in the exposed

$$I_0 = \frac{\mu_0}{P_0} = e^{\beta_0}$$ Incidence in the unexposed

$$\Rightarrow \quad RR = \frac{I_1}{I_0} = e^{\beta_E}$$

When we have a continuous variable X:

Let's say we want to estimate the impact of air temperature on the expected number of hospital admissions for respiratory issues.

P_i population size of the city (e.g., 500,000 people).

$$\frac{\mu_i}{P_i} = e^{\beta_0+\beta_1 X_i}$$

$$I_1 = e^{\beta_0+\beta_1(X_i+1)} \qquad I_0 = e^{\beta_0+\beta_1 X_i} \quad \Rightarrow \quad RR = \frac{I_1}{I_0} = e^{\beta_1}$$

rate of admissions changes for every 1 degree of increase in temperature.

There are **two** important assumptions for Poisson regression:

- Risk is ***homogeneous*** among person-[times] contributed by different subjects who have the same characteristics of interest (e.g. sex, age-group…) and the same period.

- Asymptotically, or as the sample size becomes larger and larger, the *mean* of the counts is equal to the *variance*.

Note here that the *linear regression model* (assuming constant variance & normal errors) is not appropriate for **count data** for **3** main reasons:

1. the model might lead to the prediction of negative counts

2. the variance of the response may increase with the mean

3. the errors will not be normally distributed

Supplementary materials

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

Attributable risk

An individual may become diseased without being exposed to the risk factor of interest, that is $P(D|\bar{E}) \geq 0$.

Since in that scenario not **all** disease can be due to exposure, it is appealing to ask **how much** of the disease *D* in the population can be explained by the presence of the risk factor *E*.

The **Attributable Risk** is a measure of association designed to provide an answer to this question and is defined as **the fraction of all cases of D in the population (size N) *that can be attributed* to E.**

$$AR = \frac{N * P(D) - N * P(D|\bar{E})}{N * P(D)}$$

$$AR = \frac{P(D) - P(D|\bar{E})}{P(D)}$$

Attributable risk

It could be demonstrated that: $$AR = \frac{P(E)[RR-1]}{1+P(E)[RR-1]}$$

$P(E)$=prevalence of the exposure

Attributable Risk depends on the **strength** of the association between D and E (RR) and the **prevalence** of the risk factor E.

Therefore, it incorporates the advantages of both a relative and an absolute measure of association.

AR
- < 0 exposure to *E* is protective
- = 0 independence of D and E
- > 0 exposure to *E* raises the risk of D

$$-\infty < \mathrm{AR} \leq 1$$

AR can be an arbitrarily large negative number as the disease frequency becomes increasingly smaller and E is protective

The attractiveness of the AR is the insight it promises into the **potential impact** of an intervention program designed to **reduce exposure** to a risk factor E.

However, the assumption that the risk in the unexposed can be applied to individuals who are "changed" from E to not-E *through an intervention program* assumes essentially that the E–D relationship is **causal**.

An additional tacit assumption is that *modification* of an individual's E status does not alter **other risk factors**; in the extreme it is possible that reducing exposure to E may actually increase exposure to other risk factors and thereby make the disease burden greater.

For example, automobile drivers might respond to seat-belt laws by increasing their average speed, under a perception of increased safety, thereby offsetting mortality reductions introduced by higher seat-belt usage.

Both of these concerns - ***causality*** and the ***effect*** of other factors - also apply to the RR and OR !!

[we will discuss the estimation of ***causal effects*** taking into account confounders either by design or using regression approaches, blocks 2/3]

	Mother's Marital Status	
Infant Mortality	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

Attributable risk for marital status:

$$AR = \frac{0.0086 - 0.0065}{0.0086} = 0.25$$

	Birthweight		
Infant Mortality	Low Birthweight	Normal Birthweight	Total
Death	21,054	14,442	35,496
Live at 1 year	271,269	3,804,294	4,075,563
Total	292,323	3,818,736	4,111,059

Attributable risk for low birthweight:

$$AR = \frac{0.0086 - 0.0038}{0.0086} = 0.56$$

Naive interpretation : infant mortality could be reduced by 25% if all mothers were married, or by 56% if we could *eliminate* low birthweight infants.

While it is plausible that a substantial fraction of infant mortality could be prevented by intervention programs designed to eliminate the risk of a low birthweight child, it is not believable that 25% of infant deaths could be eradicated through a program to have single pregnant women marry before they give birth...

This suggests that marital status does not, in fact, **cause** infant mortality; the ***apparent*** association, as captured by either the Relative Risk, Odds Ratio, or Attributable Risk, is likely due to the effect of **other factors** that are related to both marital status and infant mortality.

One drawback in interpreting the AR is that it does not behave as a conventional fraction when **more than one risk** factor is examined.

That is, the AR for **two distinct** exposures **cannot be added** to give the AR for both factors considered simultaneously, even when the exposures are independently distributed.

Infant Mortality	$E \,\&\, F$	$E \,\&\, \bar{F}$	$\bar{E} \,\&\, F$	$\bar{E} \,\&\, \bar{F}$	**Tot**
Death	25497	5561	4084	354	35496
Live at 1 yr	1,002,268	1,022,204	1,023,681	1,027,410	4,075,563
Tot	1,027,765	1,027,765	1,027,765	1,027,764	4,111,059

Hypothetical data on two binary exposures, E and F, that might have generated the infant mortality data (the data have been set up so that *E* and *F* are independent)

$P(D|E)$=(25497+5561)/(1027765+1027765)= 0.0151

$P(D|\bar{E})$=(4084+354)/(1027765+1027764)= 0.0022

RR_E=0.0151/0.0022=7 AR_E=0.75

$P(D|F)$=(4084+25497)/(1027765+1027765)=0.0144

$P(D|\bar{F})$=(5561+354)/(1027765+1027764)= 0.0029

RR_F=0.0144/0.0029=5 AR_F=0.67

$P(E) = P(F) = 0.5$

it appears as if infant death is 75% due to E; the other 67% is due to F ….

These two factors are independent and certainly the AR for both combined **cannot be the sum** of the individual ARs since this would greatly exceed 1 …

From another point of view, establishing the AR associated with E to be 0.75 cannot be interpreted as claiming that only 25% of infant mortality remains to be explained in the sense that AR for ***other factors*** will be 0.25 or smaller*.

Di Maso et al., Attributable fraction for multiple risk factors: Methods, interpretations, and examples, Stat Methods Med Res. 2019 https://pubmed.ncbi.nlm.nih.gov/31074326/

Attributable risk in the exposed

The fraction attributable **in the exposed** is the proportion of cases attributable to exposure in the exposed population (i.e. when considering the only population on which exposure *can act*).

$$AR_{Exposed} = \frac{P(D|E) - P(D|\bar{E})}{P(D|E)}$$

It is the risk fraction of those exposed that is attributable to exposure.

$$AR_{Exposed} = \frac{RR - 1}{RR}$$

Note that we lose here the *weight* given by **prevalence** of the exposure in the population

This excess fraction represents the proportion of cases among the exposed that can be attributed to the exposure (*assuming causality*). In other words, it represents the proportion of cases among the exposed that ***could have been prevented*** if they had never been exposed.

Example 5.14. *A total of 34 439 British male doctors were followed up for 40 years and their mortality in relation to smoking habits was assessed (Doll et al., 1994a). Mortality from certain diseases is shown in Table 5.3.*

Underlying cause of death	Never smoked regularly Rate[b] (1)	Current cigarette smoker Rate[b] (2)	Rate ratio (2)/(1)	Rate difference[b] (2)–(1)	Excess fraction (%) $\frac{(2)-(1)}{(2)} \times 100$
Cancer					
All sites	305	656	2.2	351	54
Lung	14	209	14.9	195	93
Oesophagus	4	30	7.5	26	87
Bladder	13	30	2.3	17	57
Respiratory diseases (except cancer)	107	313	2.9	206	66
Vascular diseases	1037	1643	1.6	606	37
All causes	1706	3038	1.8	1332	44

[a] Data from Doll *et al.*, 1994a.
[b] Age-adjusted rates per 100 000 pyrs.

44% of deaths among male British doctors who smoked could be attributed to smoking (*assuming causality*).

The % of deaths that could be attributed to smoking varied by disease.

This % >> for lung cancer (93%) and << for vascular diseases (37%).

However, ***if smokers had never smoked***, the total # of deaths ***prevented*** >> for vascular diseases (606 per 100.000 pyrs) than for lung cancer (195 per 100.000 pyrs)

Therefore [again] here we have a difference between *AR in the exposed* and the absolute measures

Similar measures can be calculated when those exposed have **a lower risk** of developing the disease than those unexposed.

In these circumstances, we would have:

Risk reduction: $P(D|\bar{E}) - P(D|E)$

Prevented fraction: $\dfrac{P(D|\bar{E}) - P(D|E)}{P(D|\bar{E})}$

Example 5.15. *Suppose that a group of oral contraceptive users and a group of never users were followed up in time and their ovarian cancer incidence was measured and compared. The results from this hypothetical study are shown in Table 5.4.*

	Oral contraceptive use	
	Ever	Never
Ovarian cancer cases	29	45
Person-years at risk	345 000	321 429
Rate per 100 000 pyrs	8.4	14.0

Rate ratio = 8.4 per 100 000 pyrs/14.0 per 100 000 pyrs = 0.60
Risk reduction = 14.0 per 100 000 pyrs – 8.4 per 100 000 pyrs = 5.6 per 100 000 pyrs.
Prevented fraction (%) = 100 × (5.6 per 100 000 pyrs / 14.0 per 100 000 pyrs) = 40%.

40% of ovarian cancer cases ***could have been prevented*** among never-users if they had used oral contraceptives

Number Needed to Treat (NNT)

In a **clinical trial** *[mainly]* (block 2) with a binary response, such as dead or alive, there are many ways to quantify the difference between two treatments.

For example, we can use the difference between outcome proportions: $p_{new} - p_{old}$

p_{new} = success rate with the «new» treatment; p_{old}= success rate with the «old» treatment

$$NNT = \frac{1}{p_{new} - p_{old}}$$

NNT is the number of patients to treat with the new drug to achieve a further success (equivalently, to prevent a bad outcome) over the old one.

$1 \leq NNT < \infty$

- $\frac{1}{1-0}$ New treatment is always good, old treatment never works
- $\frac{1}{0}$ New treatment and old treatment produce same results

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

We could also compute the NNH, number needed to harm, as (taking into account adverse events):

$$NNH = \frac{1}{p.adv._{new} - p.adv._{old}}$$

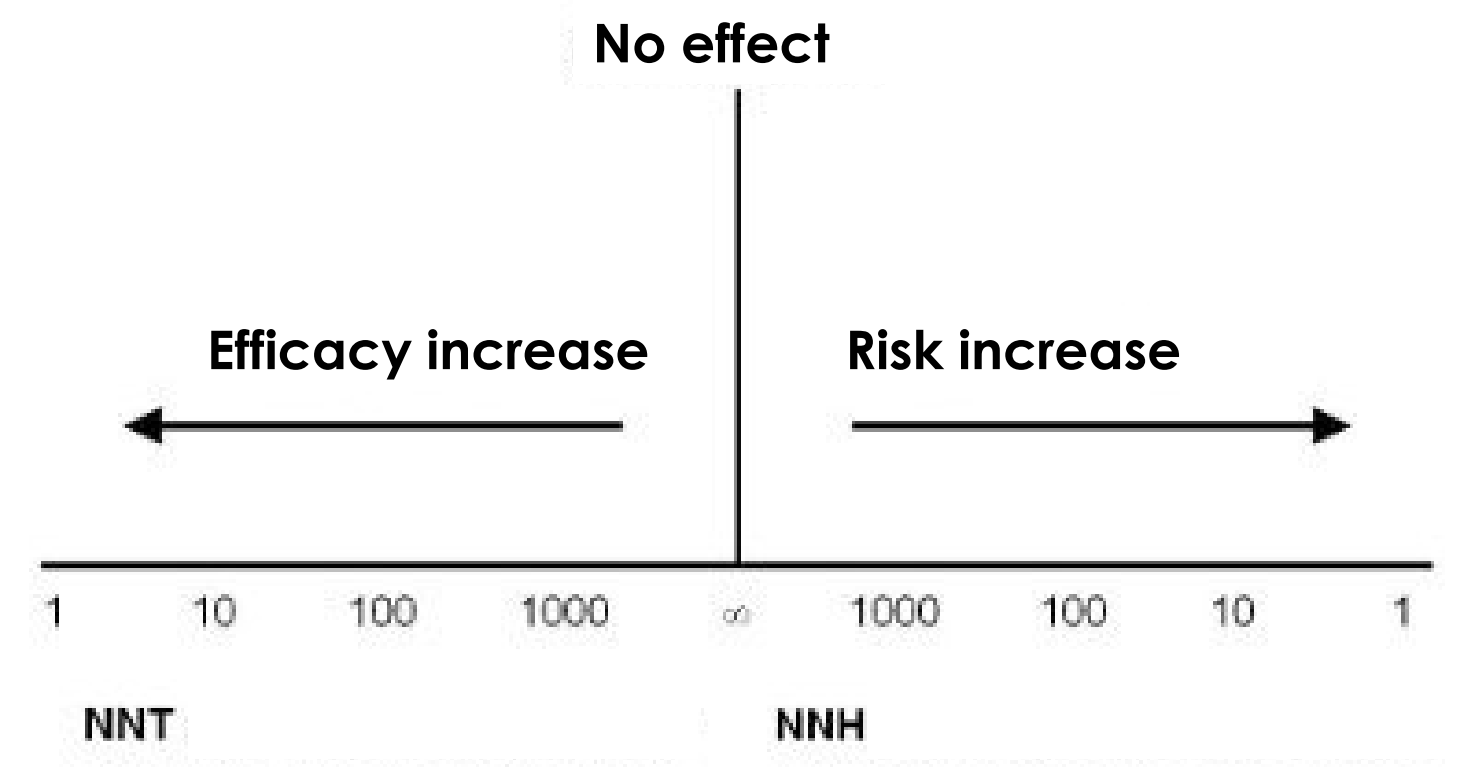
Example:
In a streptomycin study, in which subjects were selected from patients with pulmonary tuberculosis, the proportion of survivors at 6 months was 93% in the treated group vs 73% in the controls.

$$NNT = \frac{1}{p_{new} - p_{old}} = \frac{1}{0.93 - 0.73} = \frac{1}{0.20} = 5$$

The number of patients needed to be treated to prevent one death at 6 months was 5.

It is also possible to calculate **confidence intervals** around NNT/NNH*.

* if the difference between the proportions **is not statistically significant**, i.e. the confidence interval contains zero, there is a problem since "infinite" is a possible value for NNT ...in any case if there is no statistical significance, useless the NNT measure…



As the NNT **decreases**, the efficacy of the treatment increases, so 1 is the ideal NNT: a therapeutic success ***for each patient treated*** (measure of efficacy).

As the NNH **increases**, the probability of adverse events is reduced and the safety of treatment increases, so ***the ideal NNH tends to infinity***, documenting the absence of adverse events (safety measure).