Kort om statistik och epidemiologi

- Statistik
  - Mått
  - Urval
  - Konfidensintervall
  - Hypotestest
  - p-värde, Typ II fel etc.
  - Fördelningar

- Epidemiologi
  - Studietyper
  - Riskmått
  - Bias
  - Confounding
  - Exempel (fall-kontroll)

Statistik (sammanfattande mått)

- **Median (”mittenvärdet” Bra att använda om extrema värden förekommer i data materialet)**
- **Typvärde (”vanligaste värdet”)**
- **Medelvärde (används ofta)**
- **Geometriskt medelvärde (bra att använda för att beräkna doser av läkemedel etc.)**

Statistik (spridningsmått)

- Variationsvidd (eng. Range):
  - skillnaden mellan högsta och lägsta värdet.
- Percentiler, kvartiler
- Varians och standardavvikelse.

Varför?
Därför att det mesta (allt) varierar!

Statistik
Varians och standardavvikelse

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	1	- 1	1
	2	0	0
	3	1	1
Summa	6	0	2

$$\bar{X} = \frac{\sum X}{n} = \frac{6}{3} = 2$$

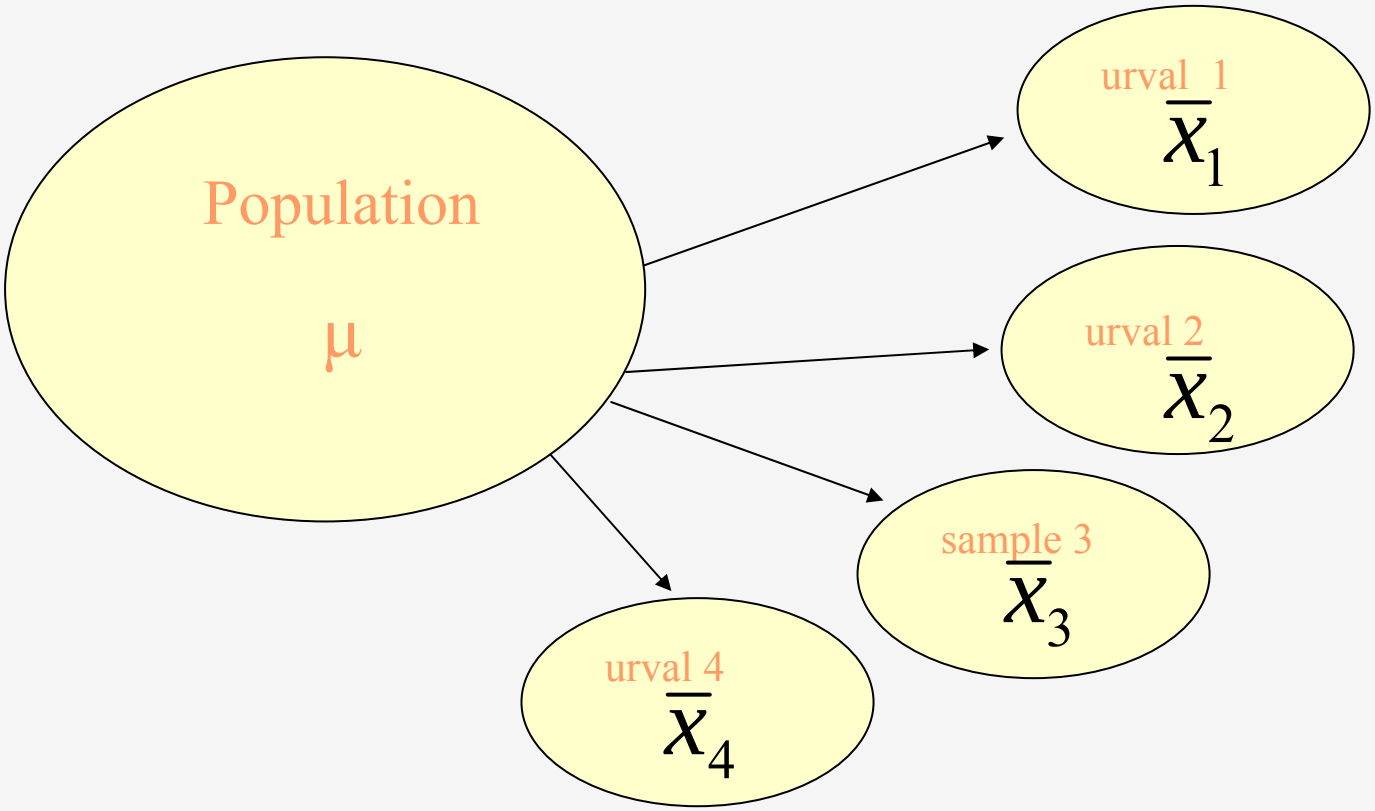
$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{2}{3 - 1} = 1$$

$$s(std.avvikelse) = \sqrt{s^2(\text{var}ians)}$$

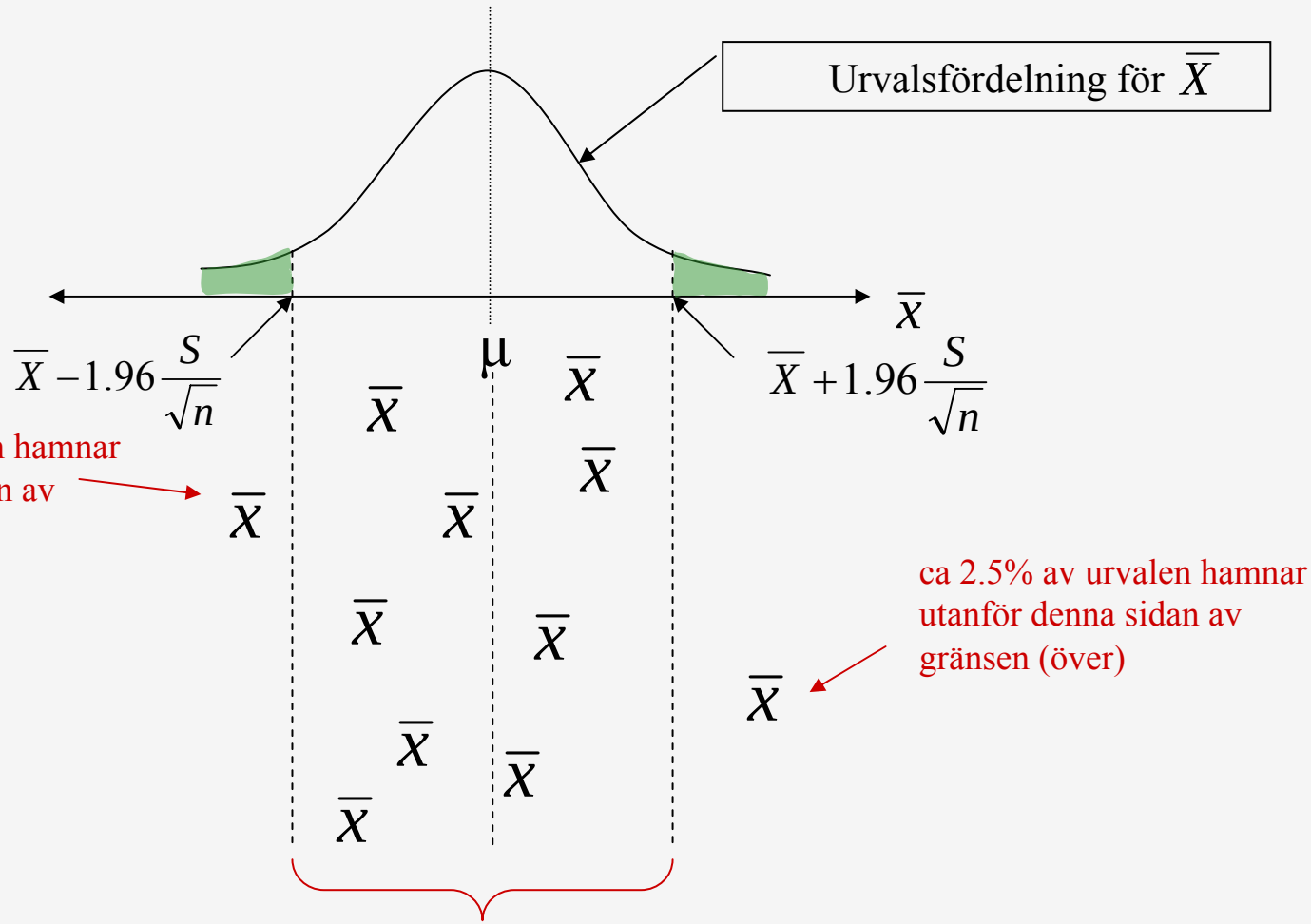
Statistik (urval)

Varför statistik?

Allt Varierar, dyrt ev. omöjligt att mäta allt eller alla



Statistik (konfidensintervall)



Statistik (konfidensintervall)

Ett konfidensintervall ger ett mått på precisionen av skattningen.

Tolkning av ett 95% confidence interval:

”med 95% sannolikhet finns det okända uppskattade populationsvärdet inom dessa gränser”

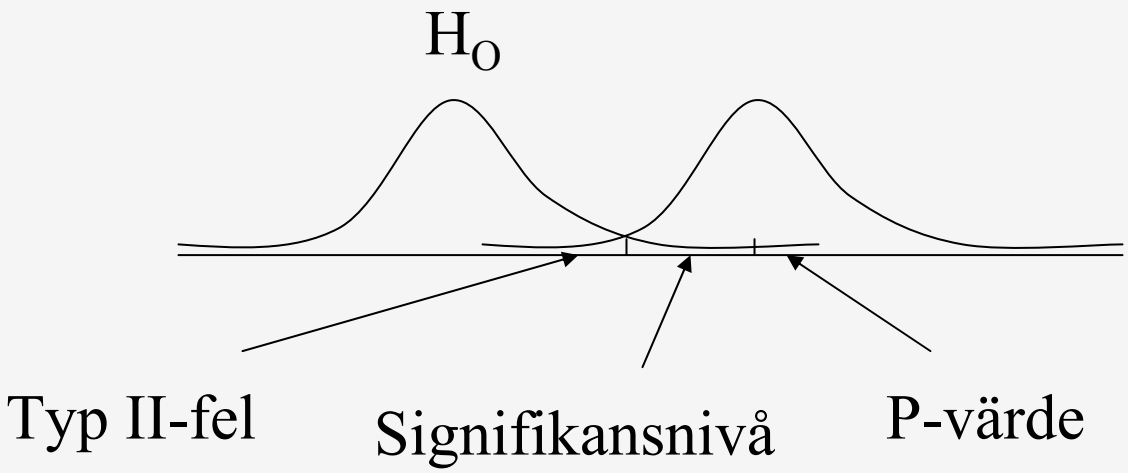
Statistik (Hypotestestning)

Hur sannolikt är det att slumpen förklarar skillnaden mellan det observerade värdet och värdet som specificerats i hypotesen?

- **Nollhypotes (ofta ingen skillnad)** H_0
- **Alternativ Hypotes**

Statistik
Hypotestestning (P-värde mm)

- **P-värdet anger hur stor sannolikheten är att vi observerar ett extremare värde förutsatt att H_O är sann.**



Statistik Hypotestest (p-värde vs. Konfidensintervall)

- **Konfidensintervall anger magnitud**
- **Konfidensintervall anger signifikansnivå indirekt**
- **Konfidensintervall mer kliniskt tillämpbart.**
- **P-värde kan beräknas ur konfidensintervall och medelvärde ej tvärtom.**

Vad är Bäst?

Statistik (Centrala principer kort!)

- **Litet p-värde kan förklaras av:**
  - **Många individer i studien (n)**
  - **Stora skillnader mellan grupper eller mättillfällen.**
  - **Liten variation, spridning (varians)**

Statistik (fördelningar)

- **Det finns många olika sannolikhetsfördelningar.**
  - **Normalfördelning**
  - **t-fördelningar**
  - **χ^2 (Chi-två fördelningar)**
  - **Binomialfördelningen (proportioner)**
  - **Poissonfördelningen (Incidens)**

- **Alla fördelningar kan approximeras med normalfördelningen förutsatt att man har tillräckligt många observationer (centrala gränsvärdessatsen)**

Statistik Centrala principer kort!

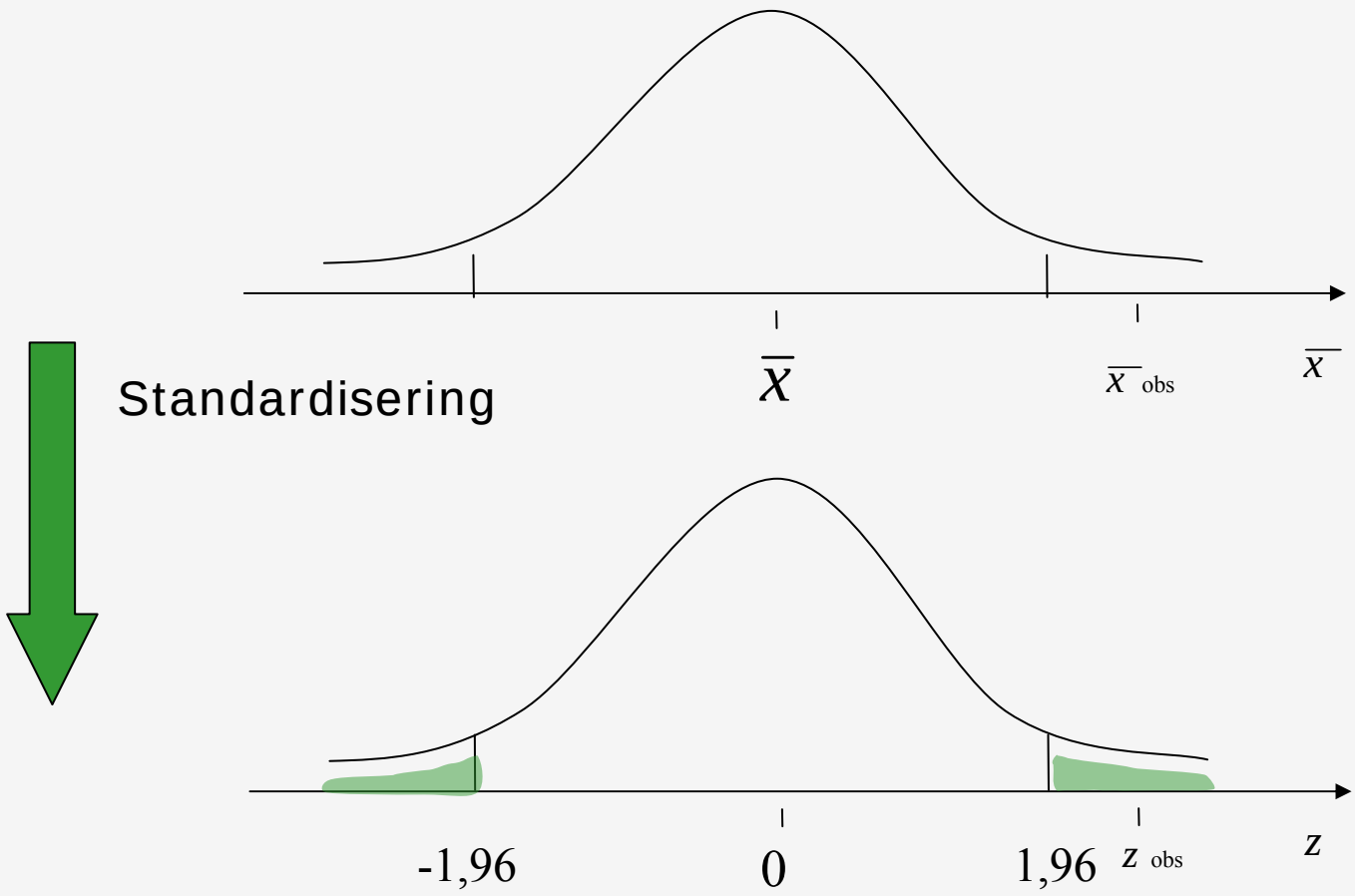
- **För att slippa tusen tabeller så standardiserar man:**

$$st.värde\,(z,t) = \frac{Medelvärde}{\sqrt{\text{varians}}}$$

- **På följande sätt får man konfidensintervall:**

$$Medelvärde \pm tabellvärde \times \sqrt{\text{varians}}$$

Statistik (Standardisering)



Statistik (t- fördelningstabell)

Konfidensint. Bredd	0,9	0,95	0,99
Frihetsgrad (n-1) etc.			
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03
6	1,94	2,45	3,71
7	1,89	2,36	3,50
8	1,86	2,31	3,36
9	1,83	2,26	3,25
10	1,81	2,23	3,17

Epidemiologi

- **Epidemiologi ~ Läran om sjukdomars utbredning och uppkomst**
- **Epidemiologiska studier är sk. Observationsbaserade studier.**
- **Stora Randomiserade studier anses vara bättre än Observationsbaserade studier.**

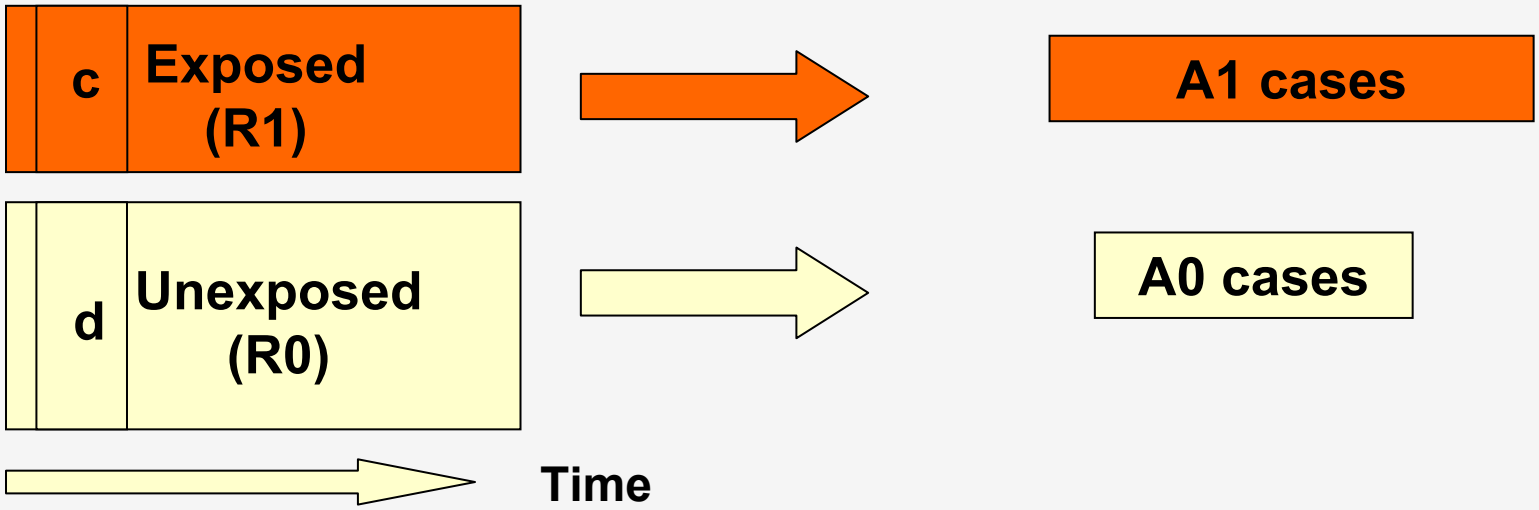
Epidemiologi (studier)

Common observational studies

- **Cohort**
  - **Good for single exposures.**
  - **Time consuming and expensive.**
  - **Often prospective**
  - **Esimate RR**
- **Case-control**
  - **Good for many exposures.**
  - **More efficient than Cohort studies.**
  - **-Retrospective (Implictaions on causal relationship)**
  - **Estimate OR (RR)**

- **Tvärsnittsstudier**
- **Ekologiska studier (grupper)**

Epidemiologi (studier och mått)

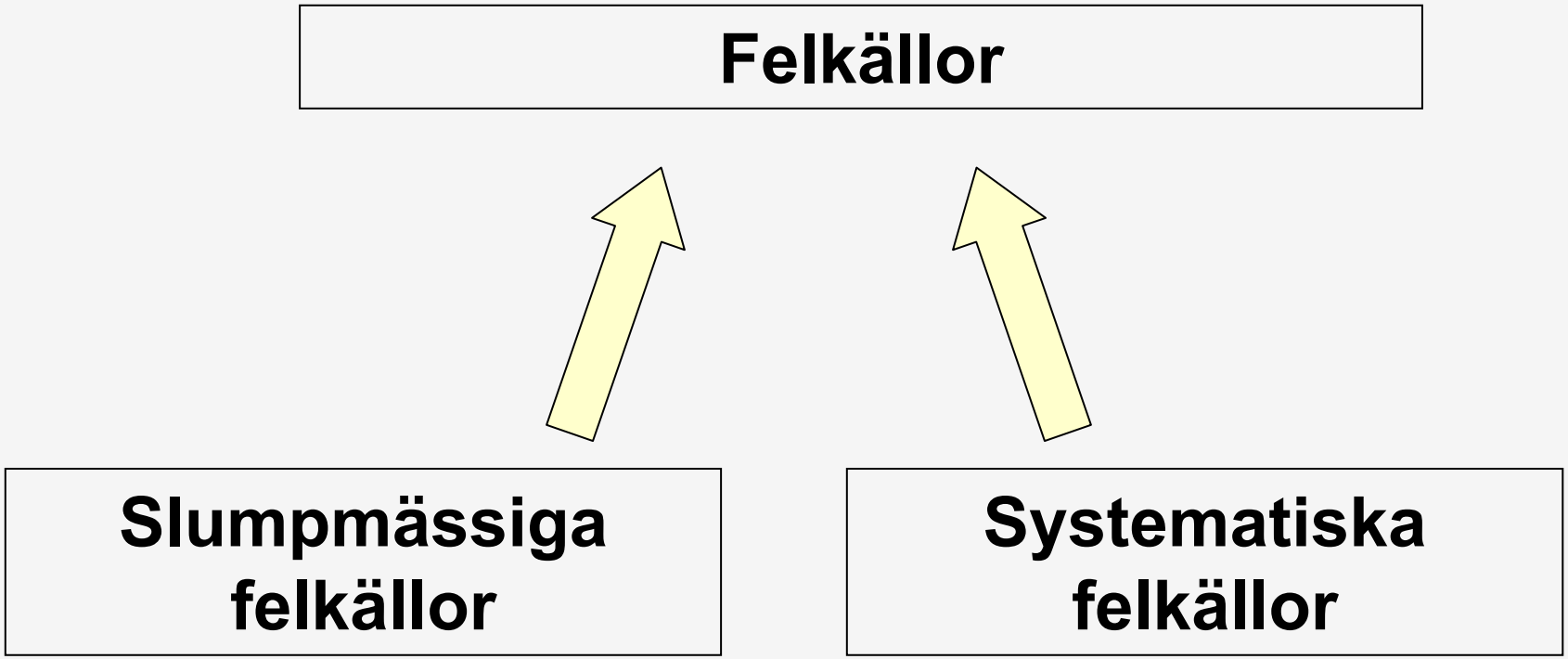


RR=A1/A0/R1/R0 (RR>1 → increased risk for disease when exposed)

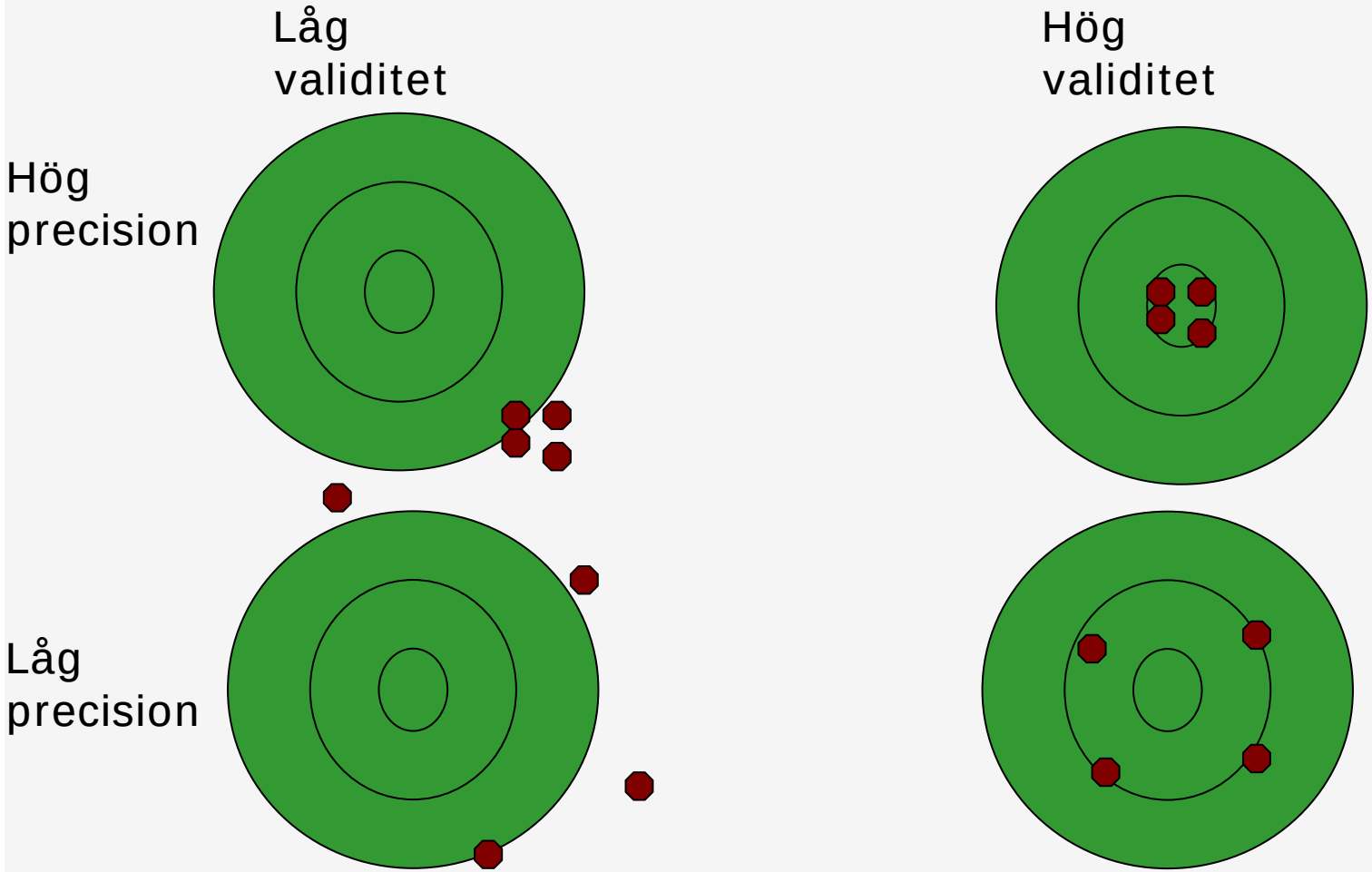
R1 and R0 can be proportions during a specified time interval or time counted in person years.

OR=A1/A0/c/d (= RR if c/d = R1/R0). OR is only based on proprotions.

Epidemiologi (slumpmässiga och systematiska felkällor)



Epidemiologi
Bias I (felkällor)



Epidemiologi
-Bias II (Systematiska felkällor)

Cohort

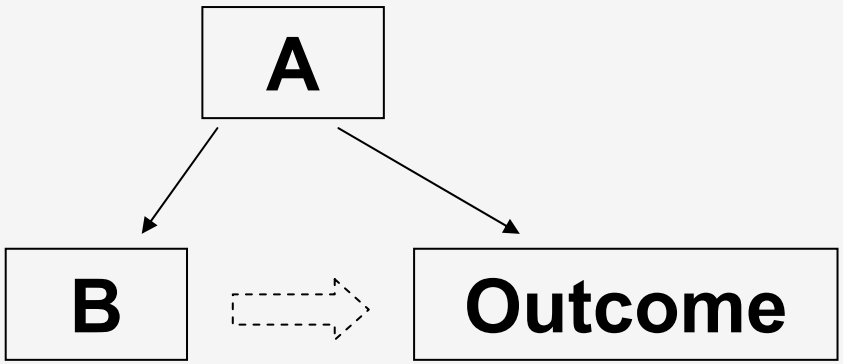
- ”Selection”
  - Loss to follow up
- Felklassificering
  - Differential
  - Non differential
- Confounding

Fall-kontroll

- Selection
  - Kontrollgrupp
- Felklassificering
  - Differential (recall)
  - Non differential
- Confounding

Problems with case-control studies

- Temporality (is the exposure preceding the outcome or not)
  - Time dependent questions and time of onset of disease.
- Bias (recall)
  - Incident cases.
- Confounding (spurious associations)
  - Adjusting (and matching)



Epidemiologi
Fall-kontroll

Odds Ratio

	Exposed	Unexposed
Case	a	b
Control	c	d

OR = a/b/c/d, often estimated by using Logistic regression ($OR=e^{\beta 1*exposure}$)

The EIRA study

- **EIRA= Epidemiological Investigation of Rheumatoid Arthritis (RA)**
  - **Population based Case-control study.**
  - **Incident cases (at present more than 2000 RA cases)**
  - **Randomly choosen Controls matched to cases on age, sex and living area.**
- **Cases and controls asked to fill in an extensive questionnaire regarding life style, exposures, diseases, education etc.**
- **Provide blood sample for genetic and serological analysis.**
- **Participating rate: 96% for cases and 82% for controls**
- **Aim: Investigate genetic and environmental risk factors for RA.**

Real world example (Rheumatoid Arthritis (RA))

- **Established Risk factors for RA (RA with certain antibodies called ACPA):**

  - Shared epitope alleles (SE alleles). Alleles in the HLA-DRB1 region. (These alleles enhance affinity to citrullinated peptides)

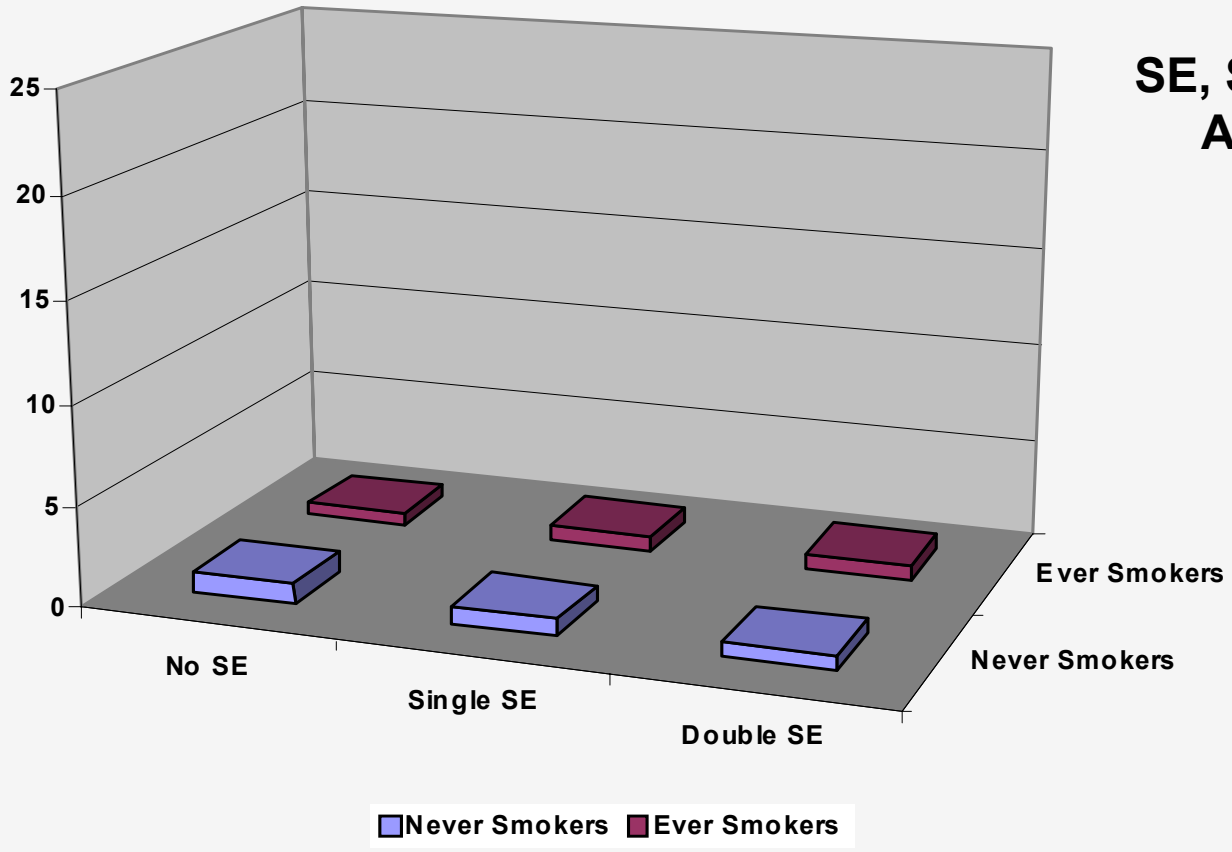
  - Smoking (causes citrullination of peptides)

Real world example (Rheumatoid Arthritis (RA))

- **SE allele and risk of RA**
  - **ACPA+ RA: RR= 5.8 (95% CI: 4.7 – 7.0)**
  - **ACPA- RA: RR= 1.2 (95% CI: 0.9 – 1.4)**
- **Smoking and risk of RA**
  - **ACPA+ RA: RR= 1.8 (95% CI: 1.5 – 2.1)**
  - **ACPA- RA: RR= 0.8 (95% CI: 0.8 – 1.2)**

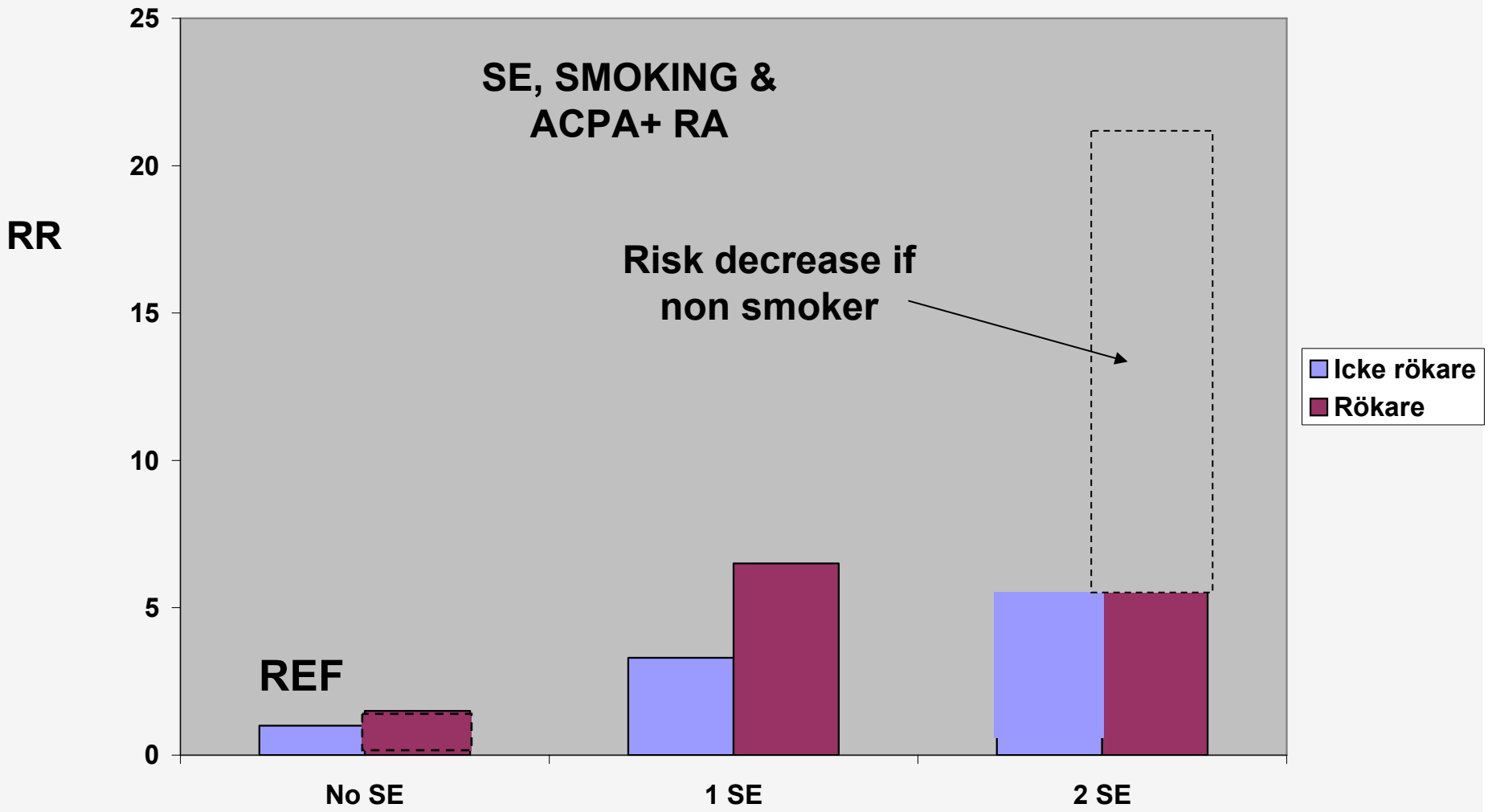
SE and smoking combined?

Real world example (Rheumatoid Arthritis (RA))



SE, SMOKING & ACPA- RA

Gen-Miljö interaktion



RR for developing ACPA+ RA

	SE allele zygousity		
Smoking status	RR (95% CI) (No SE allele, 0)	RR (95% CI) (Heterozygous, A)	RR (95% CI) (Homozygous, AA)
Never smokers, 0	1.0 (Ref)	3.3 (1.8 – 5.9)	5.4 (2.7 – 10.8)
Ever smokers, B	1.5 (0.8 – 2.6)	6.5 (3.8 – 11.4)	21.0 (11.0 – 40.2)

Definition

Statistical level

- **Multiplicative interaction logistic scale**

$$\log it(P(Y = y, SMK, SE, SMK * SE)) = \alpha + \beta_{SMK} \times SMK + \beta_{SE} \times SE + \beta_{SMK,SE} \times SMK \times SE + \varepsilon$$

$$OR_{SE,SMK} > OR_{SE} \times OR_{SMK} \Rightarrow OR_{SE,SMK} > e^{\beta_A * SMK} \times e^{\beta_B * SE}$$

- **Additative model**

$$OR_{SMK,SE} > OR_{SMK} + OR_{SE} \Rightarrow OR_{SMK,SE} > e^{\beta_{SMK} * SMK} + e^{\beta_{SE} * SE}$$

Gen-Miljö interaktion

Interaction between SE alleles, smoking ACPA+ RA (RERI and AP calculation)

- **RERI (Relative Excess Risk due to Interaction) regarding ACPA+ RA, Smoking (B) and hetero-(A) or Homo-(AA),zygous SE allele.**

$$RERI_{AB} = RR_{AB} - RR_A - RR_B + 1 = 6.5 - 3.3 - 1.5 + 1 = 2.7$$

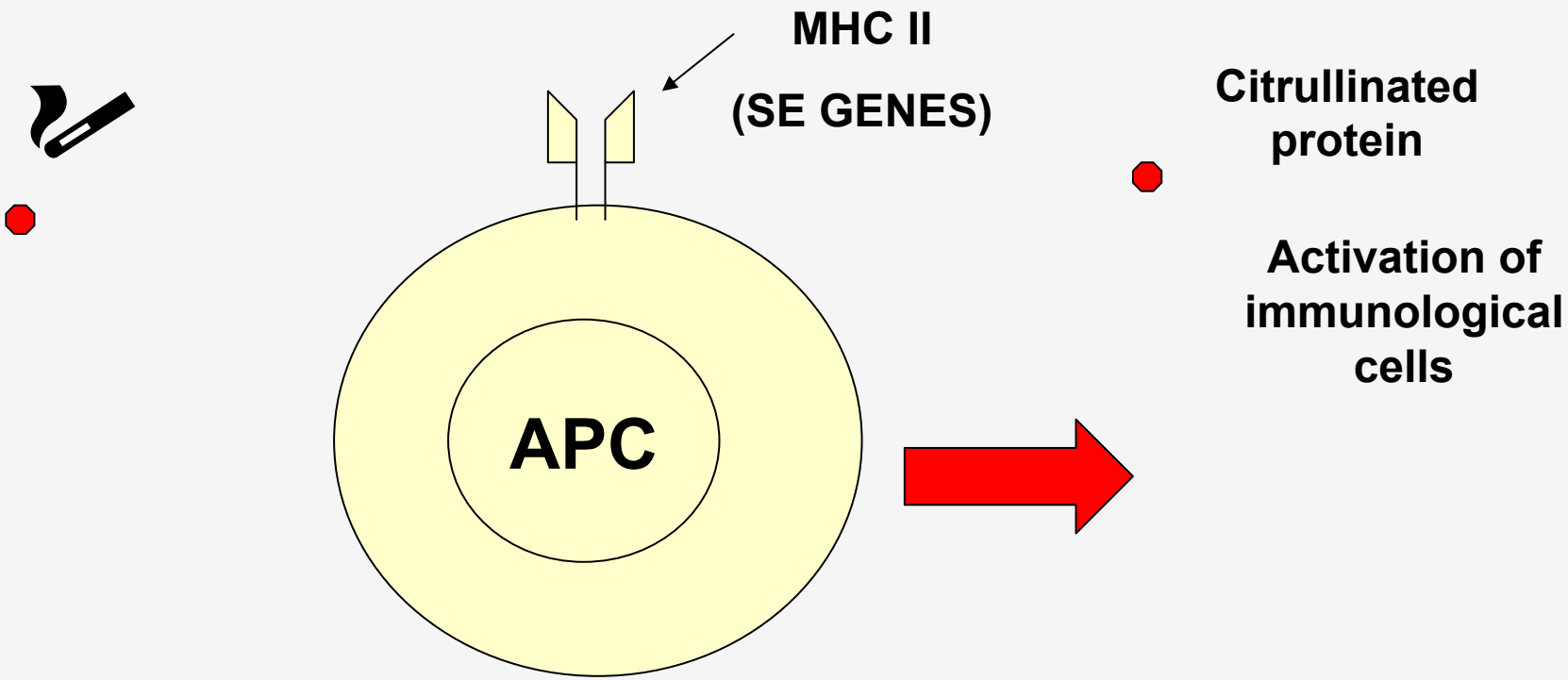
$$RERI_{AAB} = RR_{AAB} - RR_{AA} - RR_B + 1 = 21 - 5.4 - 1.5 + 1 = 15.1$$

- **AP (Attributable proportion due to interaction)**

$$AP_{AB} = RERI/RR_{AB} = 2.7/6.5 \approx 0.42$$

$$AP_{AAB} = RERI/RR_{AAB} = 15.1/21 \approx 0.72$$

Genes + Smoking



Gene-Gene interaction in Rheumatoid Arthritis

Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22 and smoking in two subsets of rheumatoid arthritis*

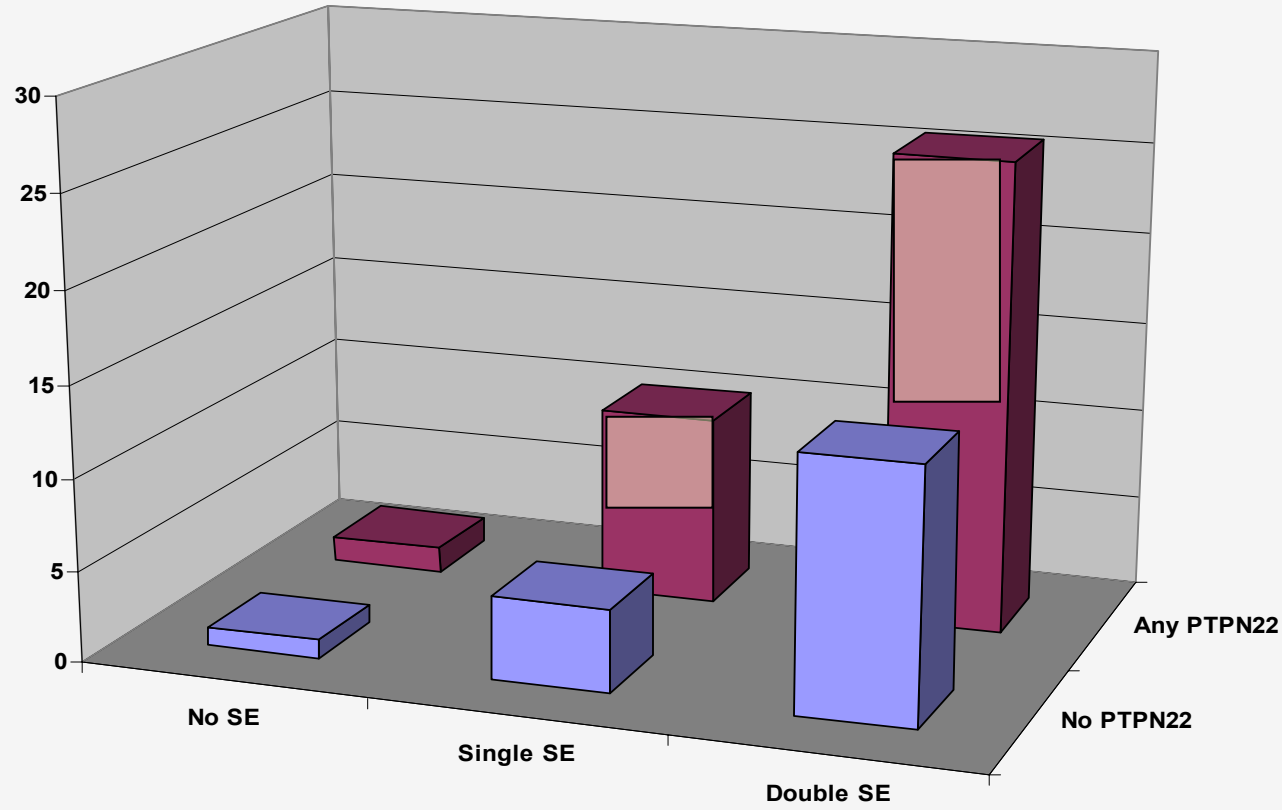
Possible to use the same methods for investigating gene-gene interaction between unlinked loci.

* Källberg H, Padyukov L, Plenge R P, Rönnelid J, Gregersen P K, van der Helm-van Mil A H M, Toes R E M, Huizinga T, Klareskog L, Alfredsson L, EIRA-study group. Am. J. Hum. Gen. 2007

Gene-Gene interaction in Rheumatoid Arthritis

Relative risks for presence of SE alleles, R620W PTPN22 regarding anti-CCP+ RA (Women and Men)

EIRA + NARAC+
LEIDEN EAC



Gene-Gene interaction in Rheumatoid Arthritis

Interaction between HLA-DRB1 SE and R620W PTPN22, in terms of developing anti-CCP+ RA.

	EIRA	NARAC	Leiden EAC	All
Deviation from additivity	p < 0.001	p < 0.001	p = 0.0016	p < 0.001
AP together with 95 % CI	0.5 (0.3 – 0.7)	0.7 (0.5 – 0.9)	0.4 (0.1 – 0.7)	0.5 (0.4 – 0.6)
Deviation from multiplicity	p = 0.06	p = 0.05	p = 0.29	p = 0.025
Deviation from independency of penetrance	p = 0.022	p = 0.035	p = 0.76	p = 0.027

Gene*Environmental and Gene-Gene interactions

Conclusions

- **Smoking and SE alleles are associated with strong interaction regarding risk of developing anti-CCP$^+$ RA**
- **Gene-gene interaction between SE and PTPN22 alleles regarding risk of developing anti-CCP$^+$ RA**

Thank you for your attention!