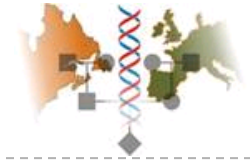




Montreal Spring School of Population Genomics and Genetic Epidemiology

Review of Epidemiology

Marie-Hélène Roy-Gagnon, PhD
May 31, 2010



Learning Objectives

- ▶ Definition and objectives of epidemiology
- ▶ Research process in epidemiology
 - ▶ Case definition
 - ▶ Measuring disease occurrence
 - ▶ Study designs
 - ▶ Measures of risk and association
- ▶ Research process in genetic epidemiology
- ▶ Role of epidemiological methods in genetic research
- ▶ Why population genetics concepts are important to genetic epidemiology research

What is epidemiology?



- ▶ The study of the distribution and determinants of health-related states or events in specified human populations and the application of this study to the control of health problems
 - ▶ Examples: disease, death, hospital admission, cholesterol levels, car accidents, treatment compliance, ...
 - ▶ The underlying hypothesis is that the distribution of states/events is not random in human population but is instead influenced by individual characteristics
 - ▶ These characteristics can be genetic, environmental, social, cultural, and can interact with each other
 - ▶ These characteristics are the basis for interventions

Objectives of Epidemiology



1. Identify the cause of a disease and its risk factors
2. Determine the extent of disease found in communities and populations
3. Study the natural history and prognosis of disease
4. Evaluate new preventive and therapeutic measures
5. Provide foundation for developing public policy and regulatory decisions

Fields of Epidemiology



- ▶ Infectious disease
- ▶ Chronic disease
- ▶ Environmental
- ▶ Genetic
- ▶ Clinical
- ▶ Social
- ▶ Aging
- ▶ Pharmaco
- ▶ ...

Epidemiological research

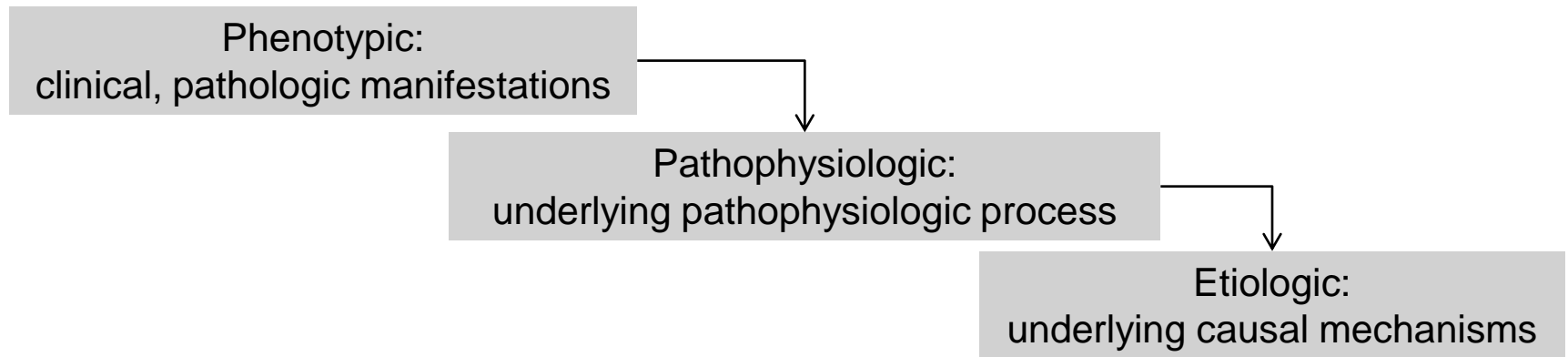


1. Case definition, disease classification
(in genetic epidemiology phenotype definition)
2. Determine research questions and underlying hypotheses within a conceptual framework
 - ▶ Descriptive data and previous literature
3. Select an appropriate study design
4. Plan data collection, including sample size calculation and ethics approval
5. Data management and quality assurance
6. Data analysis
7. Inference

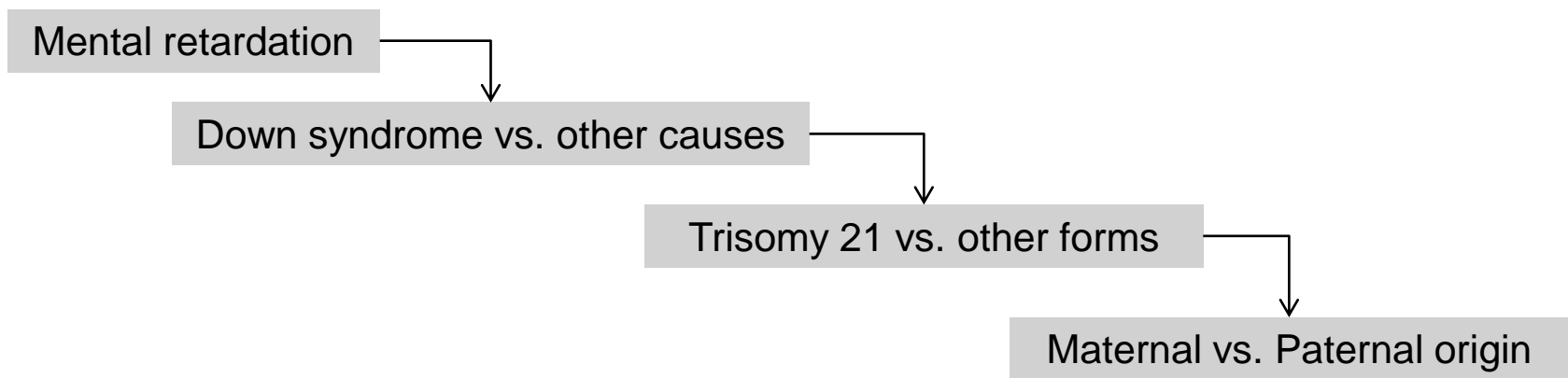
Disease classification



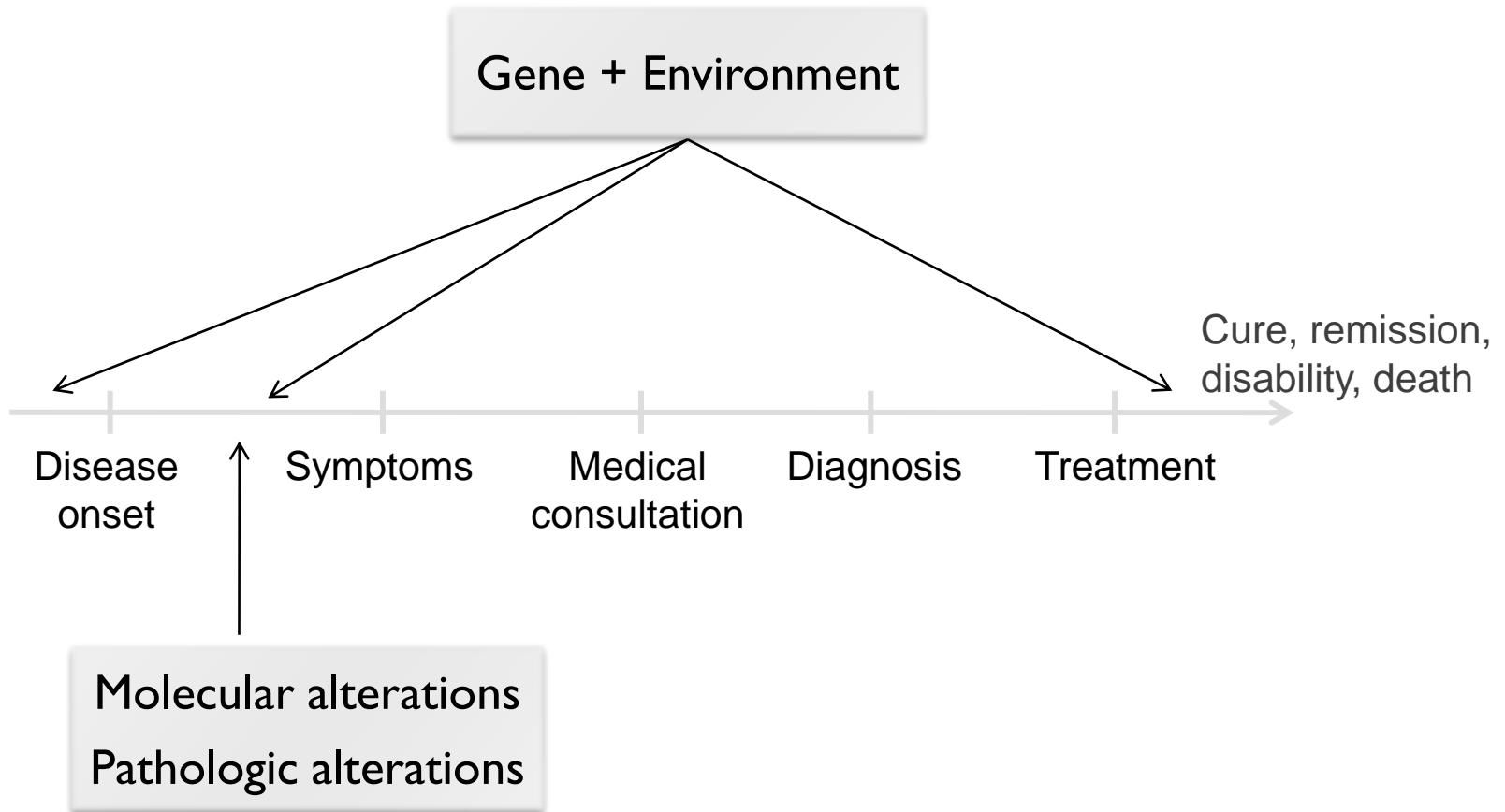
- ▶ Disease classification goes from general to specific:



- ▶ Example:



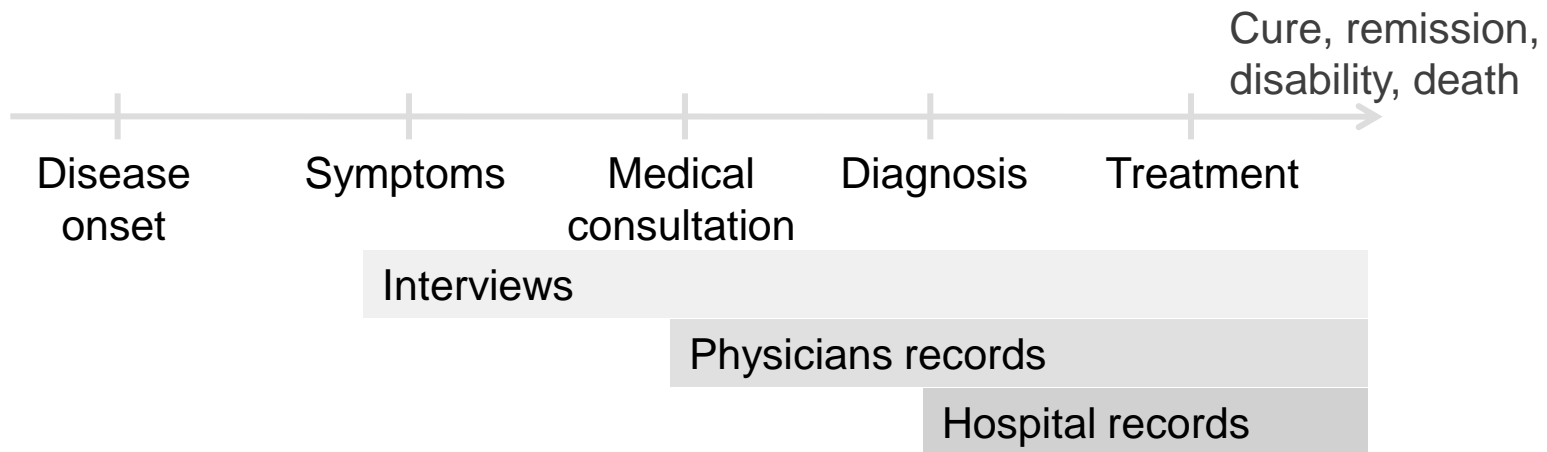
Disease natural history



Measuring disease occurrence



- ▶ The source of the data used to measure disease occurrence influences what we measure:

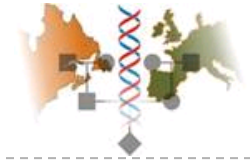


- Disease occurrence is measured in terms of rates or proportions
- Rates indicate how fast a disease is occurring in a population
- Proportions indicate what fraction of the population is affected

Measuring disease occurrence



- ▶ Incidence (measure of risk)
 - ▶ Incident case = new case
 - ▶ Cumulative incidence = $\frac{\text{Number of new cases during a given period of time}}{\text{Total number of persons at risk during the same period}}$
- ▶ The denominator must include at-risk persons only
 - ▶ Example : for uterine cancer incidence we would only include women who did not have their uterus removed
 - ▶ Incidence rate = $\frac{\text{Number of new cases during a given period of time}}{\text{Total person - time during this period}}$
 - ▶ Total person-time is the sum of the time periods contributed by each person observed for all or part of the time period



Measuring disease occurrence

▶ Prevalence

- ▶ All disease cases at a give time (prevalent cases)

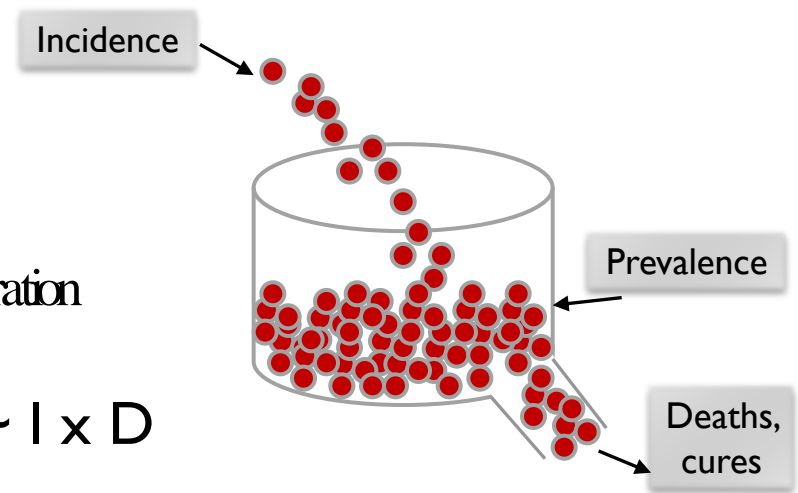
▶
$$\text{Prevalence} = \frac{\text{Number of cases in the population at a given time}}{\text{Number of persons in the population at that time}}$$

- ▶ Measures the burden of disease in a population

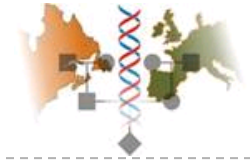
- ▶ When all factors are constant (under equilibrium):

$$\frac{\text{Prevalence}}{1 - \text{Prevalence}} = \text{Incidence} \times \text{Disease duration}$$

- ▶ If prevalence is low ($P < 10\%$), $P \sim I \times D$

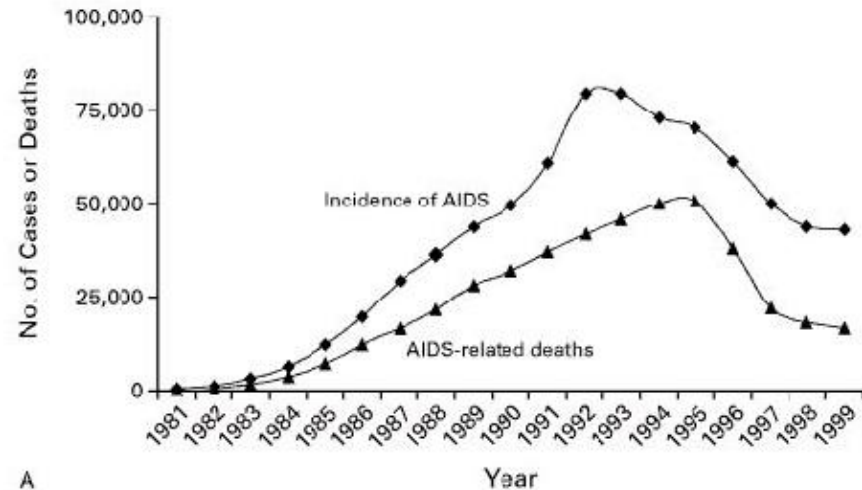


Measuring disease occurrence

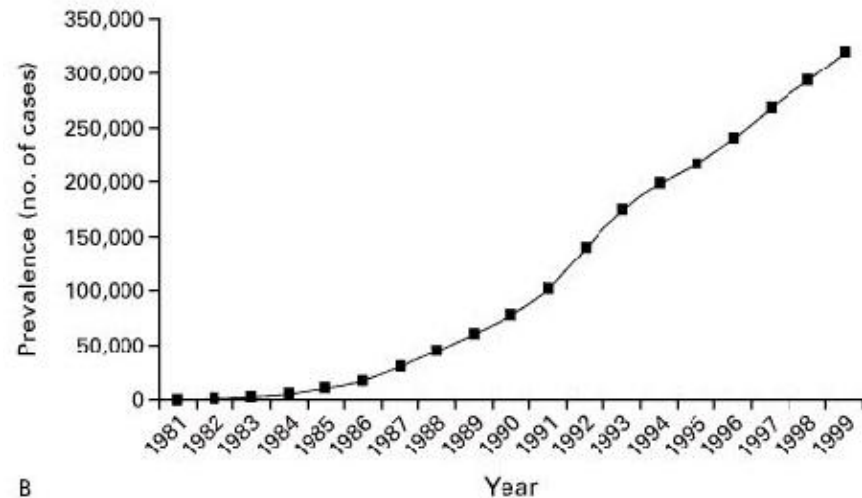


Example from Sepkowitz,
KA (2001) *NEJM* 344:1764-72

Figure 1. U.S. Trends in New AIDS Cases (Incidence) and AIDS-Related Deaths (Panel A), People Alive with AIDS (Prevalence, Panel B), and Federal Spending for AIDS Care, Prevention, and Research (Panel C), 1981 to 1999.



A



B

Measuring disease occurrence



▶ Mortality

- ▶ Death is an incident event
- ▶ Measured by mortality rates similar to incidence rates
- ▶ Mortality rates measure the intensity of deaths in a given period of time and is calculated using person-time units
- ▶ The proportion or probability of death is calculated similarly to the cumulative incidence
- ▶ Mortality is calculated for all causes, per disease or group of diseases

Epidemiological research

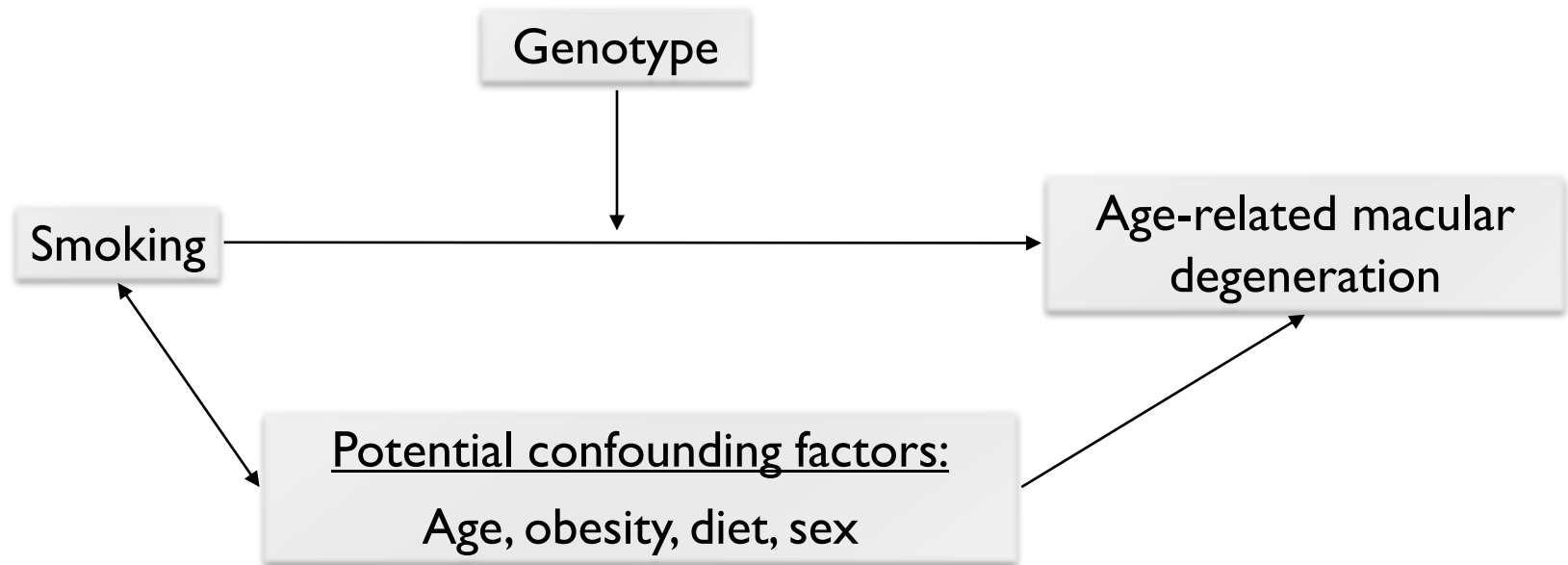


1. Case definition, disease classification
(in genetic epidemiology phenotype definition)
2. Determine research questions and underlying hypotheses within a conceptual framework
 - ▶ Descriptive data and previous literature
3. Select an appropriate study design
4. Plan data collection, including sample size calculation and ethics approval
5. Data management and quality assurance
6. Data analysis
7. Inference

Epidemiological research



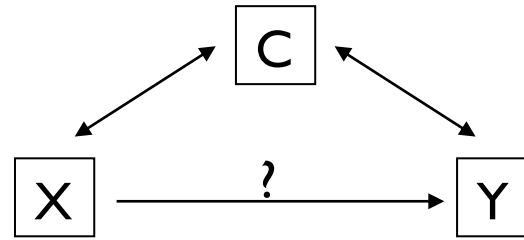
- ▶ Example of conceptual model:



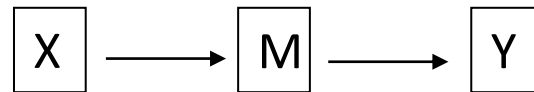
Epidemiological research



- ▶ Confounding factor:



- ▶ C is not in the causal pathway from X to Y
A mediator (intervening variable) is not a confounder:

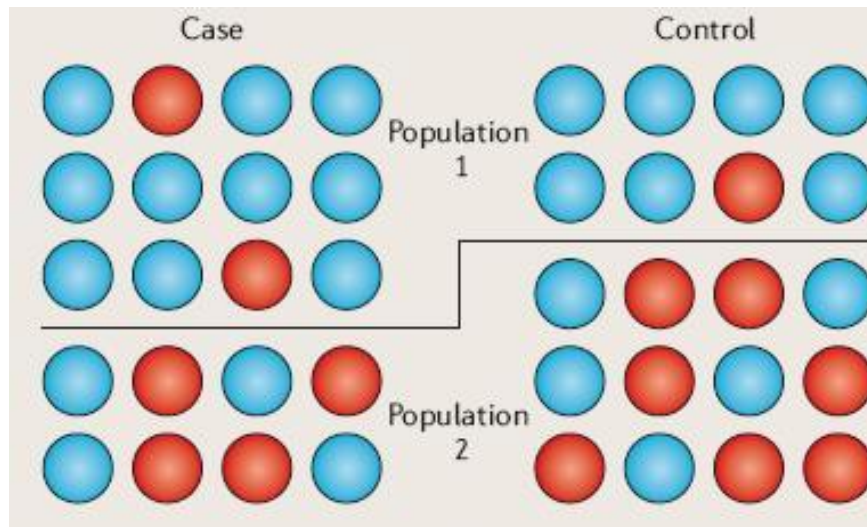
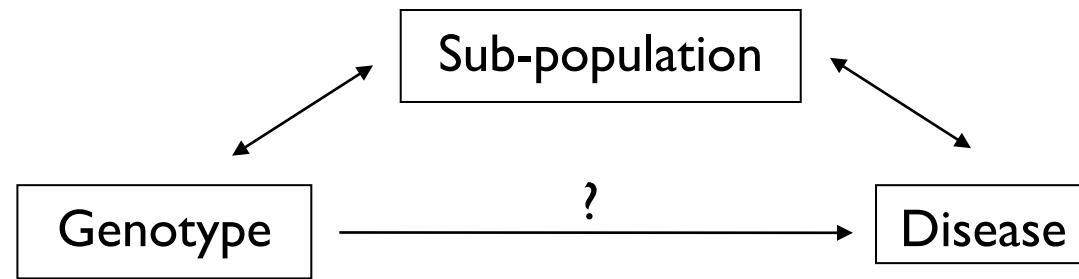


- ▶ Interaction = modification of the effect of a risk factor
 - ▶ Different level (quantitative interaction) or direction (qualitative interaction) of an association according to the levels or categories of the effect modifier

Epidemiological research



- ▶ Example of confounding factor: population structure (stratification)



Balding *Nat Rev Genet* (2006)

Epidemiological research



- ▶ Example of mediation (Simanek et al., 2009)



Epidemiological research



- ▶ Example of interaction between a genetic factor and an environmental factor (Caspi et al. 2002)

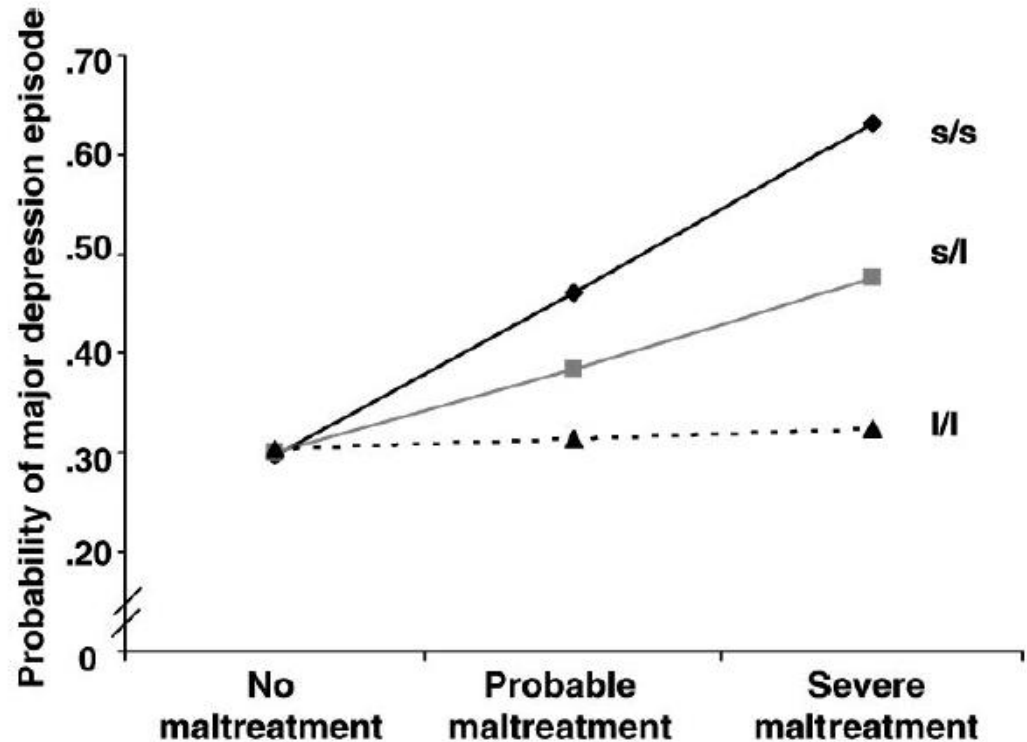


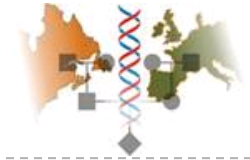
Fig. 2. Results of regression analysis estimating the association between childhood maltreatment (between the ages of 3 and 11 years) and adult depression (ages 18 to 26), as a function of 5-HTT genotype. Among the 147 s/s ho-

Epidemiological research



- ▶ Controlling for confounding:
 - ▶ A priori by matching or randomizing
 - ▶ A posteriori by stratifying or by including confounding factors in multiple regression models
- ▶ Only possible to control for known potential confounders
 - ▶ A limitation of all observational studies
- ▶ Residual confounding can still be present
 - ▶ Examples:
 - ▶ Diet is a confounder but was not measured in the study
 - ▶ The age categories used during data collection are too wide and do not allow proper adjustment

Epidemiological research



- ▶ Example of adjustment for population stratification bias (Price et al. *Nat Genet* 2006)

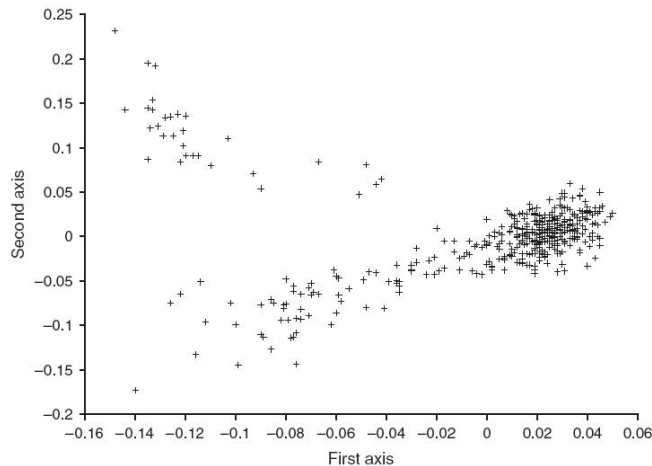


Figure 2 The top two axes of hypothesize that the first axis and southeast Europe, with a European ancestry (first axis separates two southeast Euro

Table 2 SNPs outside chromosome 2 that are spuriously associated to the lactase persistence phenotype

SNP	χ^2	Genomic control	EIGENSTRAT
rs10511418	45.11 (0.0000022)	31.55 (0.0023)	11.57 (1.00)
rs2493880	26.12 (0.037)	18.27 (0.89)	8.17 (1.00)
rs4306808	26.04 (0.039)	18.21 (0.90)	8.83 (1.00)
rs2243133	25.60 (0.049)	17.90 (0.93)	5.88 (1.00)

Epidemiological research



▶ Inferring causality

1. **Temporality**
2. **Strength of association**
3. **Dose-response relationship**
4. **Replication of the findings**
5. **Biological plausibility**
6. **Consideration of alternative explanations**
7. Cessation of exposure
8. **Consistency with other knowledge**
9. Specificity of the association

Epidemiological research



Necessary
and sufficient

Factor A

Necessary
but not
sufficient

Factor A + Factor B

Sufficient
but not
necessary

Factor C ou Factor D

Neither
sufficient nor
necessary

Factor A + Factor B
ou
Factor C + Factor D
ou
Factor E + Factor F

Disease

Epidemiological research



- ▶ Consideration of alternative explanations:
The observed association can be due to:
 - ▶ Chance (type I error)
 - ▶ Biases
 - ▶ Confounding
- ▶ Uncertainty is quantified using statistical methods
 - ▶ Examples: standard error and confidence interval around an estimate, *p-value* of a hypothesis test
- ▶ Sample size must be large enough
- ▶ A bias is any systematic error in the *design*, execution or analysis of a study, which leads to inaccurate estimates of the effects of risk factors on disease

Epidemiological research

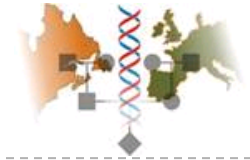


- ▶ Types of bias: selection and information bias
 - ▶ Selection: study participants are not selected according to the same criteria
 - ▶ Example: Controls are younger than cases
 - ▶ Information bias:
 - ▶ Leads to misclassification of disease or exposure
 - ▶ For example, interview bias occurs when an interviewer's style of questioning or interpretation of answers differ according to study group (i.e., case/control, exposed/unexposed)
 - ▶ Or in the case of recall bias, affected individuals recall events differently than unaffected
 - Cases may remember events that they feel may have caused their disease
 - ▶ Non-differential misclassification: true association is diluted
 - ▶ Differential misclassification: false positive or negative

Study Designs

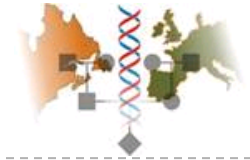


- ▶ Population comparisons (ecological design)
 - ▶ Examples from genetic epidemiology:
 - ▶ Migrant studies
 - ▶ Admixture studies
- ▶ Design more specific to genetic epidemiology:
Family studies
 - ▶ Fixed sets of relatives
 - ▶ Twin studies, Adoption studies
 - ▶ Arbitrary pedigrees



Study Designs

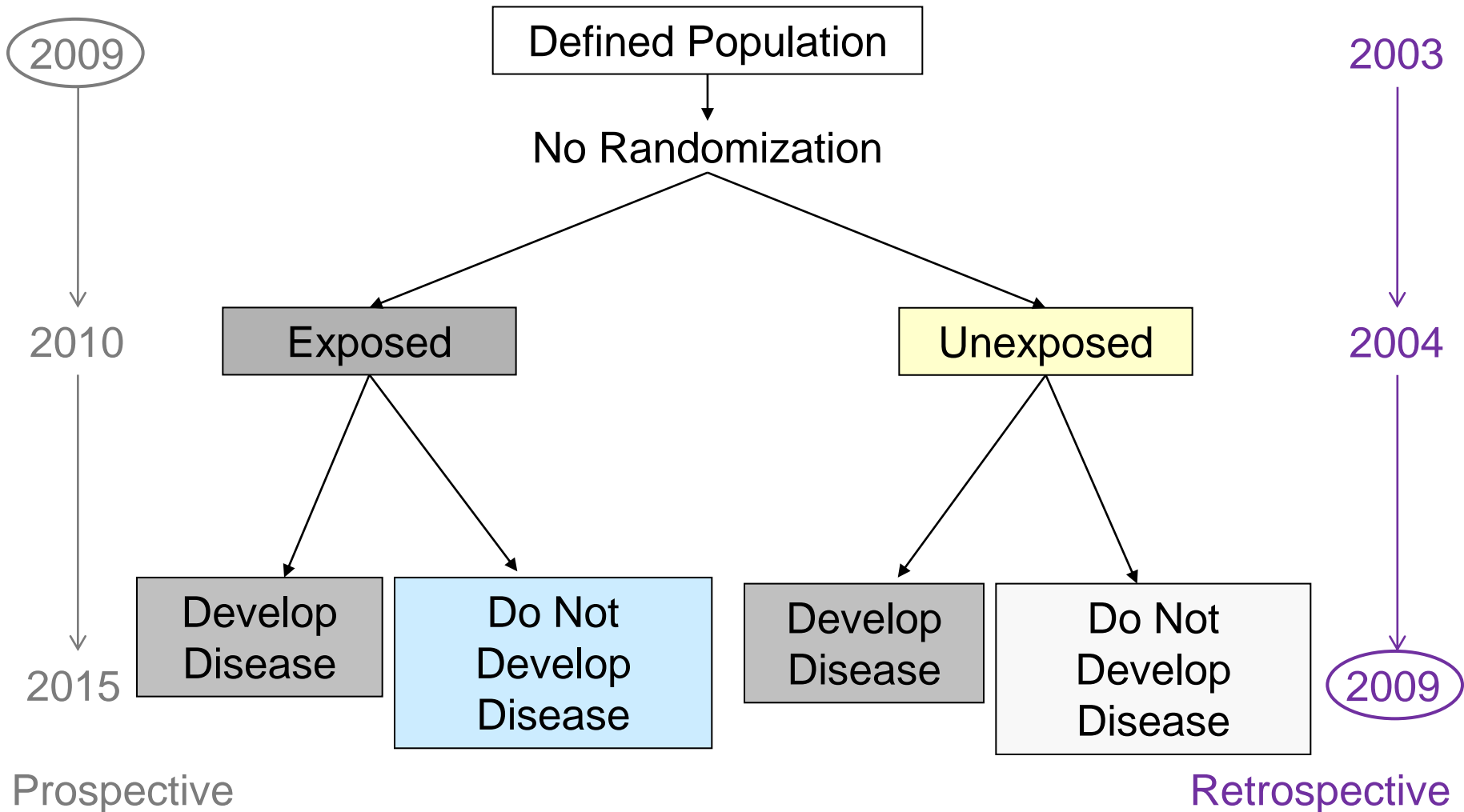
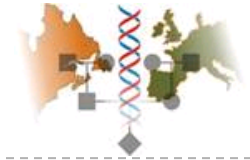
- ▶ Cohort study
 - ▶ Prospective (longitudinal)
 - ▶ Retrospective
 - ▶ Case-control study
 - ▶ Cross-sectional study
- } Observational studies
- ▶ Randomized clinical trial → Experimental study
 - ▶ Randomization: random assignment of study subjects to different groups, for example to different treatments
 - ▶ In genetic epidemiology, Mendelian transmission constitutes a natural randomization that can be exploited



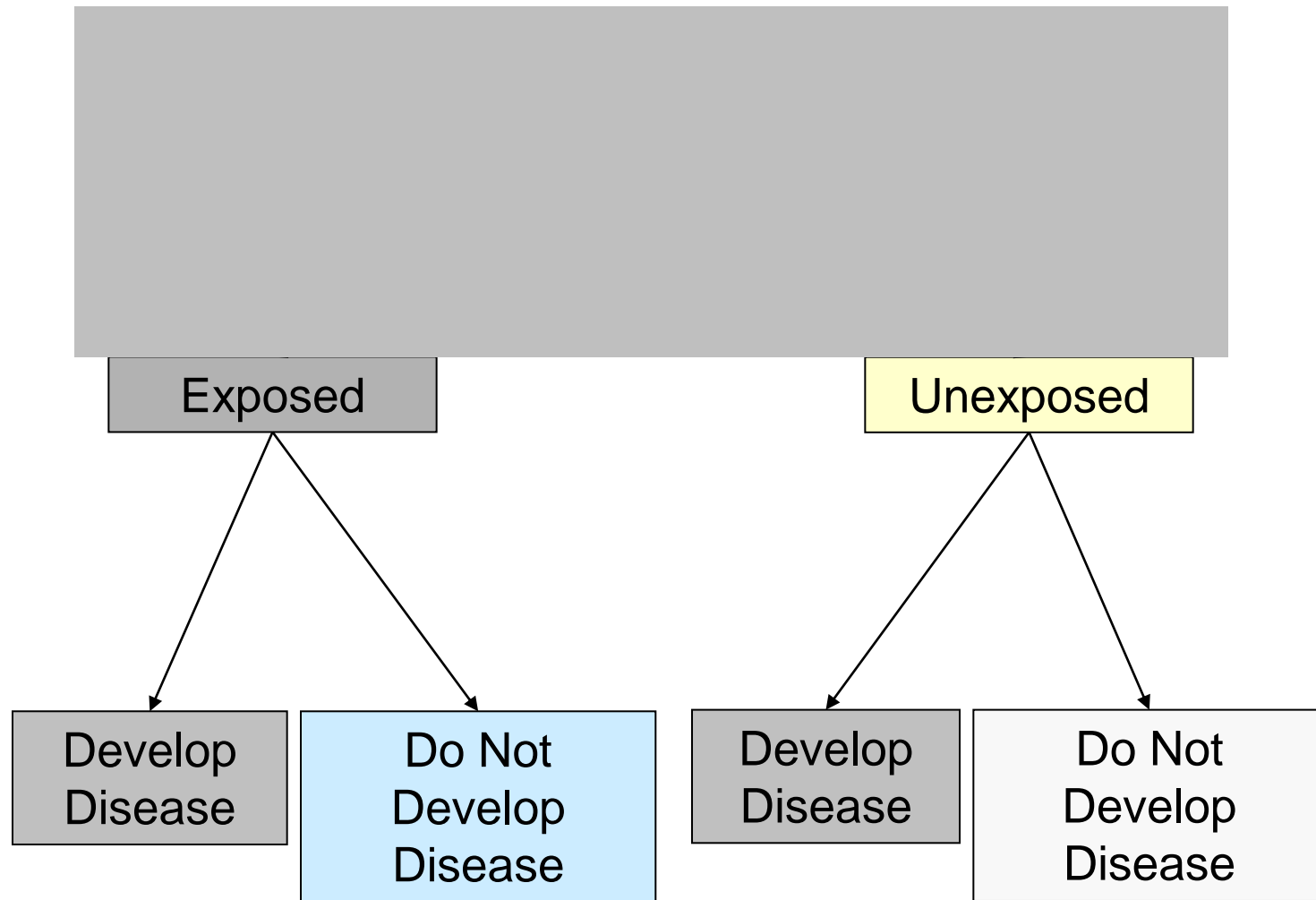
Study designs

- ▶ **Experimental studies**
 - ▶ The conditions of the study are controlled by the researcher
 - ▶ The study participants are assigned to the experimental or control group in a random fashion (gold standard)
- ▶ **Observational studies**
 - ▶ Observational study of changes or differences in factors in relation to a trait of interest (like asthma or blood pressure) without any intervention from the researcher
 - ▶ Observational descriptive study : hypothesis-generating
 - ▶ Observational analytic study: test hypotheses
 - ▶ Advantages: better feasibility, more natural situation
 - ▶ Disadvantages : little control of study conditions, can be susceptible to bias, unique study

Study Designs: Cohort Study



Study Designs: Cohort Study



Study Designs: Cohort Study



▶ Advantages :

- ▶ The temporality can be verified such that the occurrence of the risk factor is before the onset of the disease
- ▶ Many diseases or traits can be studied at the same time
- ▶ Efficient for rare risk factors
- ▶ Less susceptible to bias than other observational studies

▶ Disadvantages:

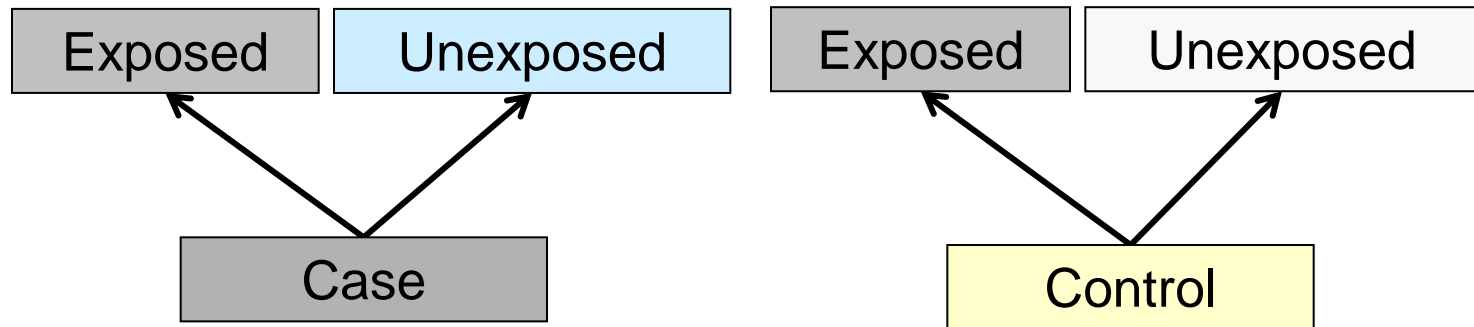
- ▶ Not good for rare diseases
- ▶ Long observation time can be required
- ▶ Can be very costly
- ▶ Risk factors can change during the study: can be difficult to keep track, measure and take into account of these changes

Study Designs: Cohort Study



- ▶ Possible biases:
 - ▶ Interviewer bias
 - ▶ Solution = masked (or blinded) data collection
 - ▶ Information bias if the quantity and or quality of information collected is different across exposure groups
 - ▶ Especially a problem for retrospective chart reviews
 - ▶ Bias due to non-response or losses to follow-up
 - ▶ Non-response and losses to follow-up must be non-differential (i.e., similar across exposure groups) to avoid bias
 - ▶ Analytic bias
 - ▶ Confounding
 - ▶ Can statistically adjust or match for known, measured confounders

Study Designs: Case-Control Study



- ▶ Must take great caution when selecting the cases and controls
 - ▶ Ideally select a group of incidence cases
 - ▶ Which controls to select?
 - ▶ Matching of controls?

Study Designs: Case-Control Study



▶ Case selection:

- ▶ Case definition and eligibility criteria based on our hypotheses and on the population to which we want to generalize
- ▶ Choose incident cases of the disease unless the duration (survival) of the disease is independent of the exposure
- ▶ Sources : hospitals, registers, schools, surveillance data, ...

▶ Selection of controls:

- ▶ Must reflect the probability of exposure in cases under the assumption of no association
 - ▶ Same eligibility criteria as the cases
 - ▶ Exclude conditions related to the exposure or similar to the disease
 - ▶ Sources : hospitals, neighbourhood, spouses, parents and friends, colleagues of the case, representative population controls

Study Designs: Case-Control Study



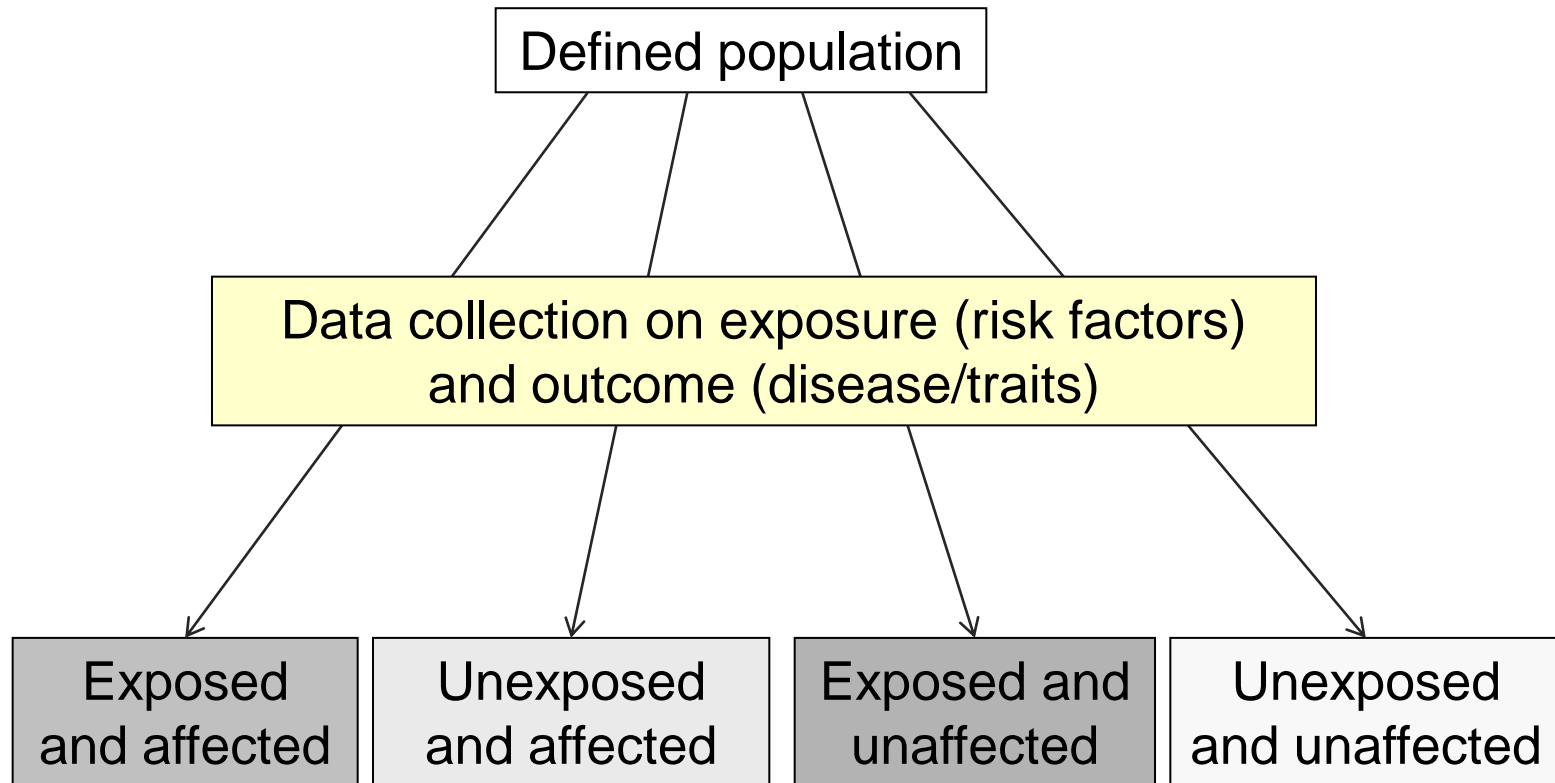
- ▶ Advantages :
 - ▶ Good for rare diseases or diseases with a long latency
 - ▶ Relatively rapid and less costly
 - ▶ Can study many exposures
- ▶ Disadvantages:
 - ▶ Not good for rare exposures
 - ▶ Difficult to establish temporality
 - ▶ Studying the mechanism of action of a risk factor is almost impossible
 - ▶ Especially subject to selection bias and recall bias
 - ▶ Impossible to directly calculate incidence rates
 - ▶ If it is a population study (i.e. subjects are representative of the general population), they can be derived

Study Designs: Case-Control Study



- ▶ Possible biases:
 - ▶ Selection bias
 - ▶ Selection of cases and controls is not independent from exposure
 - ▶ Includes: surveillance bias, non-response bias, survival bias
 - ▶ Information bias occurs when the amount and/or quality of information obtained differ according to case/control status
 - ▶ Example: recall bias
 - ▶ Analytical bias
 - ▶ Confounding
 - ▶ Can statistically adjust or match for known, measured confounders
- ▶ Case-control studies can be nested within cohort studies

Study Designs: Cross-sectional



Study Designs: Cross-sectional

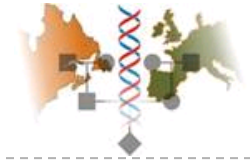


▶ Advantages:

- ▶ Good for studying characteristics stable over time (ex.: genetic factors)
- ▶ Easier to carry out
- ▶ Can generate hypotheses to guide future studies
- ▶ Can measure the prevalence of the exposure and the disease

▶ Disadvantages:

- ▶ Cannot establish temporality of exposure and outcome
- ▶ Not good for diseases that are rare, have short duration or poor survival
- ▶ Impossible to distinguish if exposure is associated with initiation or duration of the disease (incidence-prevalence bias)



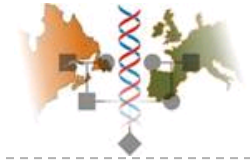
Measures of Risk

- ▶ Absolute risk = incidence
- ▶ Relative risk (RR) is the ratio of two incidence rates
- ▶
$$\text{Relative Risk} = \frac{\text{Risk in exposed}}{\text{Risk in unexposed}}$$
- ▶ Cohort study:

	Affected	Unaffected	Total
Exposed	a	b	$a + b$
Unexposed	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

$$\text{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

} = 1, no association
} > 1, positive association
} < 1, negative association



Measures of Risk

- ▶ Example of a calculation of the RR in a study of incident cases of age-related macular degeneration

	Affected	Unaffected	Total
Exposed (genotype CT or CC)	132	884	1016
Not Exposed (genotype TT)	53	636	669
Total	185	1520	1705

$$RR = \frac{\frac{132}{1016}}{\frac{53}{669}} = 1.64$$

Measures of Risk



▶ Odds of disease given exposure = $\frac{P(\text{affected} | \text{exposed})}{1 - P(\text{affected} | \text{exposed})}$

▶ Odds ratio (OR) for disease:
$$\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

▶ Odds of exposure given disease = $\frac{P(\text{exposed} | \text{affected})}{1 - P(\text{exposed} | \text{affected})}$

Measures of Risk

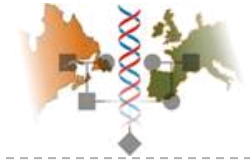


▶ Case-control study:

	Affected	Unaffected	Total
Exposed	a	b	$a + b$
Unexposed	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Odds ratio for exposure:

$$\text{OR} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$$



Measures of Risk

- ▶ Example of a calculation of the OR in a case-control study of age-related macular degeneration

	Affected	Unaffected	Total
Exposed (genotype CT or CC)	132	884	1016
Unexposed (genotype TT)	53	636	669
Total	185	1520	1705

$$\text{OR} = (132 / 53) / (884 / 636) = 1.79$$

$$\text{OR} = (132 / 884) / (53 / 636) = 1.79$$



Measures of Risk

- ▶ Advantages of odds ratios (OR):
 - ▶ Odds ratios can be estimated from all study designs (cohort, case-control, cross-sectional)
 - ▶ Is obtained directly from a logistic regression model
 - ▶ Similar interpretation to a RR except that we speak of odds and not of risk
 - ▶ Approximates the RR when the disease is rare
- ▶ Disadvantages:
 - ▶ Interpretation less easy to understand than RR
 - ▶ Exaggerates the RR when disease is not rare

Measures of Risk



- Matched case-control study:

Case	Control	
	Exposed	Unexposed
Exposed	<i>a</i>	<i>b</i>
Unexposed	<i>c</i>	<i>d</i>

$$OR = \frac{b}{c}$$

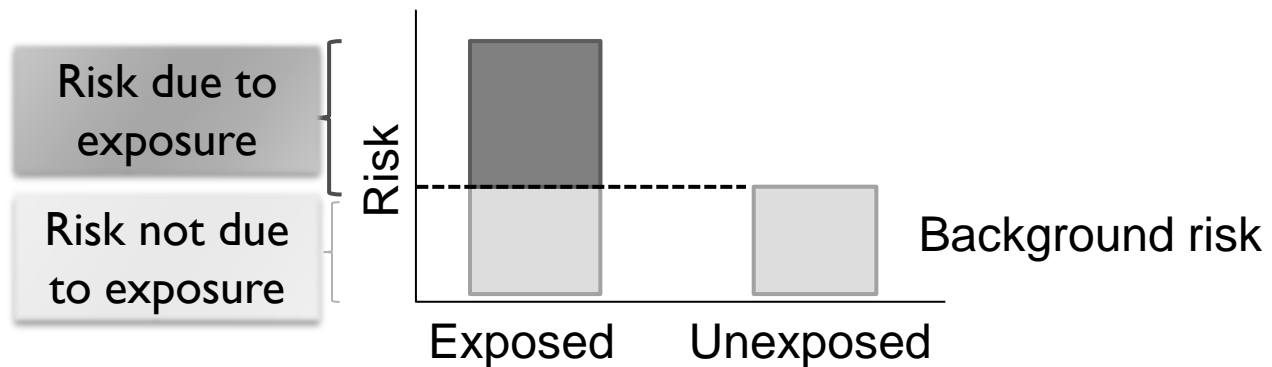
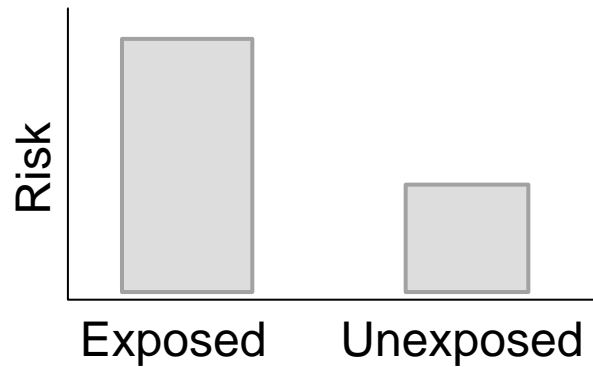
- *b* = Pairs in which the case is exposed, control is not
- *c* = Pairs in which the control is exposed, case is not
- The two other cells (*a* and *d*) tell us nothing about the risk of the case compared to the control
- Obtained with conditional logistic regression

Measures of Risk



- ▶ **Attributable risk**
 - ▶ AR (attributable risk)
 - ▶ AR% (attributable risk percentage)
 - ▶ PAR (population attributable risk)
 - ▶ PAR% (population attributable risk percentage)
- ▶ **Useful for:**
 - ▶ Evaluation of the impact of the risk factor
 - ▶ Measure of the impact of eliminating the risk factor in the population
 - ▶ Perspective of benefit of disease prevention

Measure of Risk



Measures of Risk



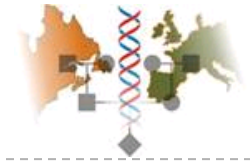
Cohort study:

$$AR = \text{Incidence in exposed} - \text{Incidence in unexposed}$$

$$AR\% = \frac{\text{Incidence in exposed} - \text{Incidence in unexposed}}{\text{Incidence in exposed}}$$

$$PAR = \text{Incidence in the population} - \text{Incidence in unexposed}$$

$$PAR\% = \frac{\text{Incidence in the population} - \text{Incidence in unexposed}}{\text{Incidence in the population}}$$



Measures of Risk

Levin's formula for Population Attributable Risk %*

$$= \frac{p_e \times (RR - 1)}{p_e \times (RR - 1) + 1} \times 100\%$$

$$p_e \times (RR - 1) + 1$$

p_e = exposure prevalence in the population

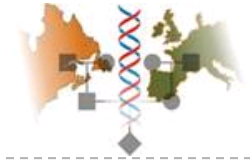
*Should not be used if using an adjusted RR

$$= p_d \left(\frac{RR - 1}{RR} \right)$$

p_d = proportion of cases exposed to risk factor

See Rockhill et al. Use and misuse of population attributable fractions, Am J Public Health, 1998

Measures of Risk



▶ Example :

	Ankylosing spondylitis	COPD
Incidence	0.002	0.05
Genetic Factor	HLA-B27	Alpha-1 antitrypsin deficiency
Mode	Dominant	Recessive
Allele Frequency	0.036	0.02
GRR	100	20
PAR%	87.5	0.8

GRR = Genotype relative risk

$$\text{PAR}\% = \frac{\text{Frequency of risk genotype} \times (\text{GRR} - 1)}{\text{Frequency of risk genotype} \times (\text{GRR} - 1) + 1}$$

Genetic Epidemiology



- ▶ Study of genetic factors and their interaction with environmental factors as they relate to disease distribution in human populations, with the ultimate goal of controlling and preventing diseases
- ▶ Related to genetics (genomics), epidemiology, population genetics, public health, and biomedical sciences
- ▶ Uses tools from statistical genetics and bioinformatics

Genetic Epidemiology



- ▶ Why study genetic factors?
 - ▶ Understand biological processes leading to disease
 - ▶ Diagnostic and prognostic
 - ▶ Prevention
 - ▶ Screening and genetic counselling
 - ▶ Treatment
 - ▶ Develop new treatment
 - ▶ Personalized medicine (pharmacogenetics)
 - ▶ Strengthen/Confirm epidemiological inference for modifiable environmental factors (Mendelian randomization)

Genetic Epidemiology Process



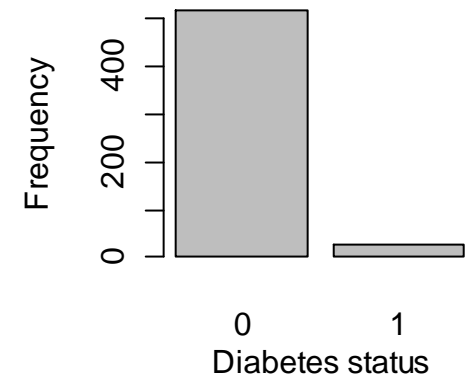
- ▶ **First step: Carefully define and measure phenotype**
- ▶ A phenotype is an observable trait that must be measurable
 - ▶ Examples: hair color, presence/absence of a condition, physiological and biological measurements (blood pressure, serum levels of different factors, ...)
- ▶ Quality of phenotype measurement is essential to study success
 - ▶ Validity, Reliability/Reproducibility
- ▶ Relationship between genotype and disease-related phenotypes can be simple or very complex

Genetic Epidemiology Process



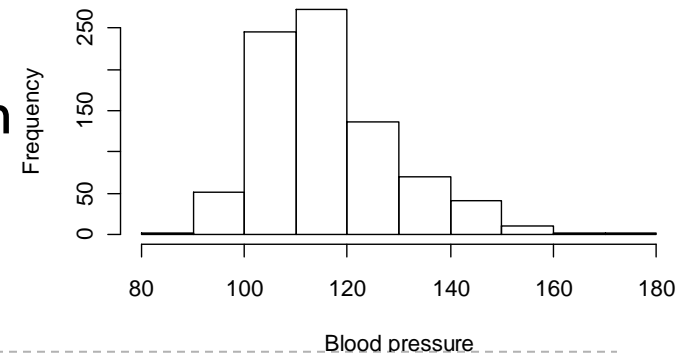
▶ Qualitative phenotype

- ▶ Trait has a discrete distribution in the population
- ▶ Examples:
 - ▶ Dichotomous: Affected/Unaffected
 - ▶ Ordinal: Severity scale Mild/Moderate/Severe
 - ▶ Nominal: Eye color Blue/Green/Brown

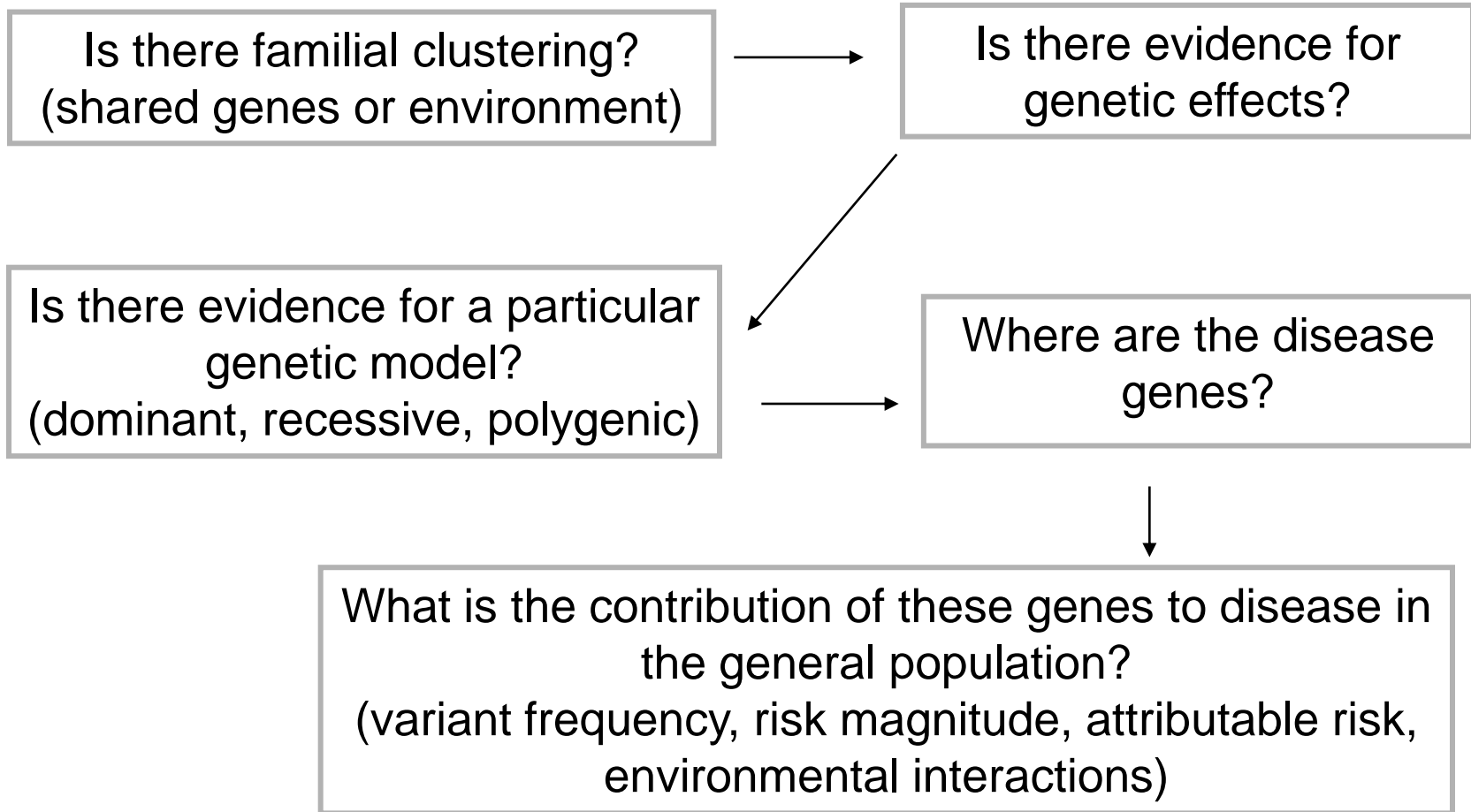


▶ Quantitative phenotype

- ▶ Trait has a continuous distribution in the population
- ▶ Examples: Blood pressure, height, serum levels, test scores,...



Genetic Epidemiology Process



Genetic Epidemiology Process



Familial clustering? – Familial Aggregation studies



Evidence for genetic effects? – Heritability studies



Mode of inheritance model? – Segregation Analysis



Phenotypic
data only

Where are the disease genes? - Disease gene identification

- Genome wide
- Chromosomal regions
- Candidate genes

Linkage analysis (families)

- Model-based
- Model-free

Association studies (families or population samples)

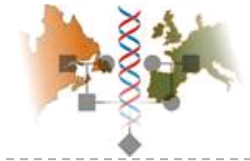
- Indirect
- Direct

Familial Effects?



Population comparisons (ecological design)

- ▶ Disease rates differences among genetically distinct populations provide some evidence for genetic effects (although confounded with environment)
- ▶ Migrant studies: Are migrant more similar to their native or new population? Genetic changes are slower than environmental changes.
- ▶ Admixture studies: Do genetic effects influence risk in mixed offspring?



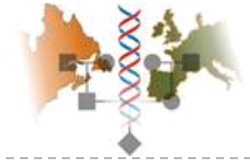
Familial effects?

Familial aggregation studies

- ▶ Case-control with family information design
- ▶ Information on family history:
 - ▶ Any first-degree relatives with positive family history
 - ▶ Number of relatives with positive family history
 - ▶ Relatives under a specific age with positive family history
 - ▶ ...

	Family history	
	Yes	No
Case	a	b
Control	c	d

$$OR = ad/bc > 1 ?$$



Familial effects?

Familial aggregation studies

▶ Family Case-control design

	Affected	Unaffected
Relative of a case	a	b
Relative of a control	c	d

$$OR = ad/bc > 1 ?$$

▶ Familial recurrence risk

$$\lambda_R = \frac{P(\text{affected} / \text{relative is affected})}{P(\text{affected})} = \frac{\text{prevalence in relative of type R}}{\text{prevalence in the population}}$$

Familial effects?

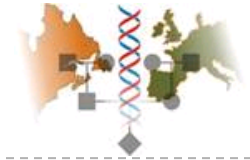


Table 2 FRR for first-degree relatives of cancer probands by site, in decreasing order of prevalence

Data from Utah [Goldgar *et al.* (7)] and Sweden [Dong and Hemminki (8)] are shown.

Site	Utah		Sweden	
	FRR (total)	FRR (early onset)	FRR (Child)	FRR (sibling)
Prostate	2.21	4.08	2.82	9.41
Breast	1.83	3.70	1.86	2.01
Colorectal	2.54	4.53	1.86	4.41
Lung	2.55	2.50	1.68	3.16
Uterine	1.32	1.75	—	—
Melanoma	2.10	6.43	2.50	3.41
Bladder	1.53	5.00	1.53	3.30
Non-Hodgkin's lymphoma	1.68	2.40	1.68	2.37
Brain/CNS	1.97	8.95	1.72	2.37
Cervix	1.73		1.93	2.39
Ovary	2.04		2.94	2.52
Stomach	2.08		1.72	8.82
Lip	2.72			

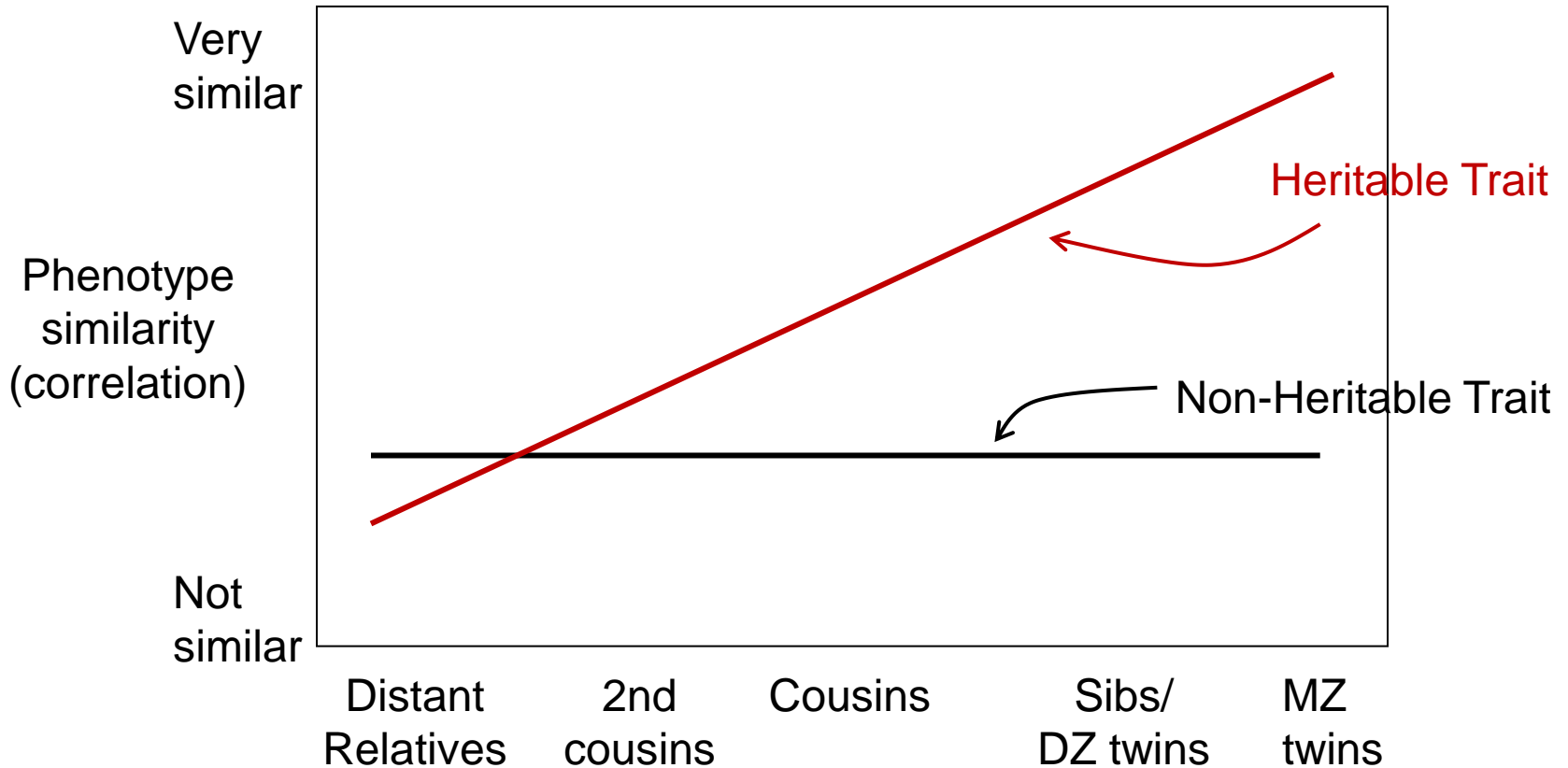
From Risch (2001) *Cancer Epidemiol Biomarkers Prev* 10:733-41

Genetic effects?

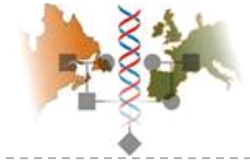


- ▶ Heritability studies (in families or fixed sets of relatives)
- ▶ Heritability can be thought of as the similarity between related individuals that is due to shared genes
- ▶ Familial correlations in phenotype:
If trait is heritable, individuals who share more genes should have higher phenotypic correlations than those who share less

Heritability



Heritability



▶ Example: Breast density in the Amish

Table 3. Heritability estimates (h^2) for breast measures and other breast cancer risk factors

Trait	$h^2 \pm SE$	P	Proportion of total variance explained by	
			Covariates	Genes
Dense area	0.39 ± 0.11	1.8×10^{-5}	0.15	0.33
Percent density	0.35 ± 0.11	1.2×10^{-4}	0.17	0.29
Nondense area	0.71 ± 0.10	1.7×10^{-15}	0.04	0.68
Age at menarche	0.58 ± 0.10	1.9×10^{-12}	<0.01	0.58
Number of live births	0.49 ± 0.11	1.4×10^{-8}	<0.01	0.49
Age at first birth	0.34 ± 0.11	3.4×10^{-4}	<0.01	0.34
Age at natural menopause	0.58 ± 0.17	1.3×10^{-4}	0.02	0.56
Height	0.70 ± 0.10	5.3×10^{-17}	0.06	0.66
Weight	0.58 ± 0.10	4.1×10^{-12}	0.04	0.56
BMI	0.47 ± 0.10	3.6×10^{-8}	0.01	0.46
Waist circumference	0.57 ± 0.11	1.8×10^{-10}	<0.01	0.57

NOTE: Data in second column are after adjustment for age and menopausal status. Dense area, percent density, age at menarche, weight, BMI, and waist circumference were log transformed. Nondense area and age at natural menopause were power transformed (0.3 and 2, respectively). Other variables were not transformed.

Genetic effects?

▶ Twin studies:



MZ: Share 100%
of genome



DZ: Share 50% of
genome on average

- ▶ If genetic effects:
concordance MZ > concordance DZ
correlation MZ > correlation DZ
- ▶ Can be used to estimate heritability

Genetic effects?

- ▶ Adoption studies:
 - ▶ Adoptee, adoptive parents, biological parents, non-related sibs
 - ▶ Comparison of risk or correlation can provide evidence for genetic effects
- ▶ Twins reared apart
 - ▶ Natural design to distinguish between genetic and environmental influences
 - ▶ Small sample sizes

Genetic Epidemiology Process



Familial clustering? – Familial Aggregation studies



Evidence for genetic effects? – Heritability studies



Mode of inheritance model? – Segregation Analysis



Phenotypic
data only

Where are the disease genes? - Disease gene identification

- Genome wide
- Chromosomal regions
- Candidate genes

Linkage analysis (families)

- Model-based
- Model-free

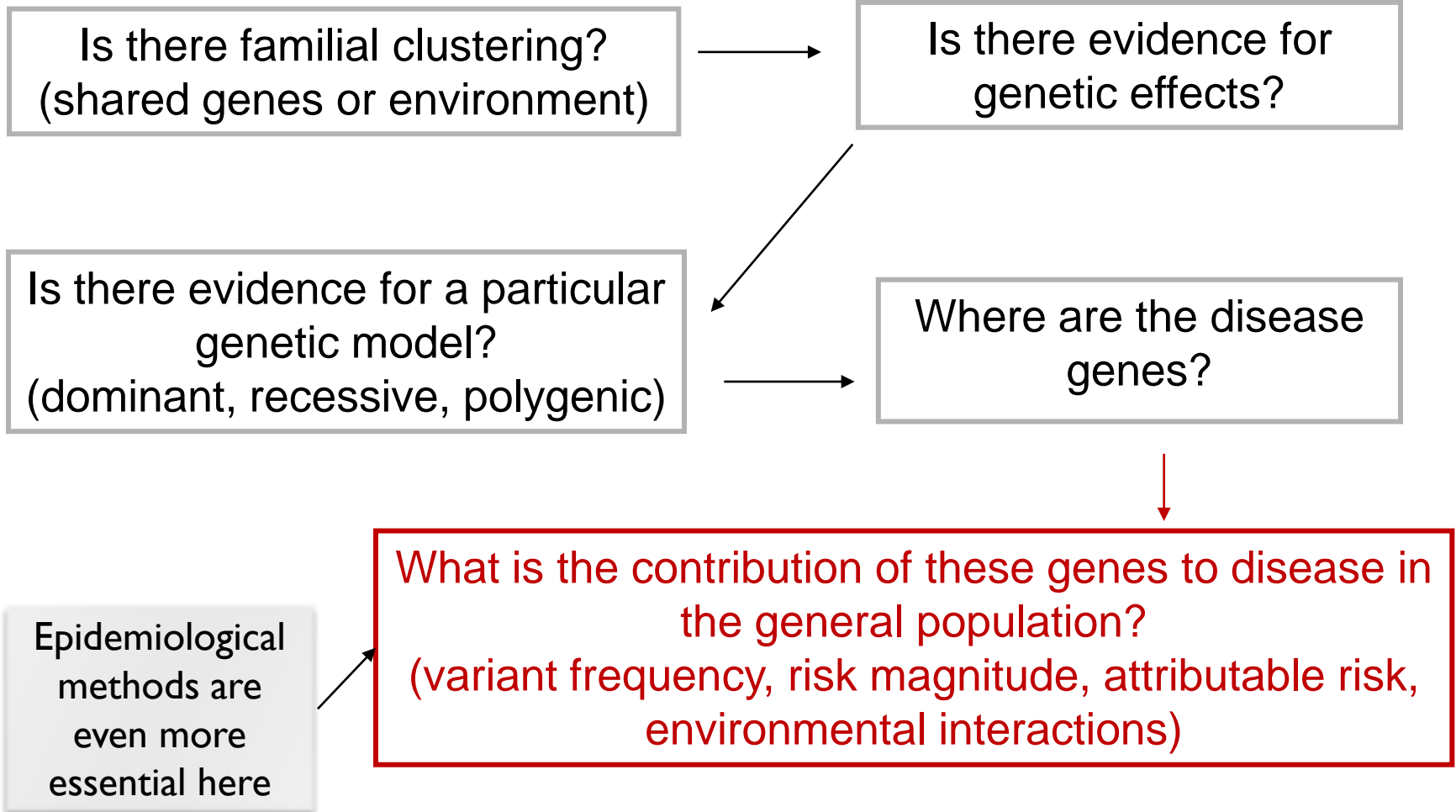
Association studies (families
or population samples)

- Indirect
- Direct

Genetic Epidemiology Process

- ▶ Association studies
 - ▶ Detects the association of genetic markers with disease across families
 - ▶ Exploits linkage disequilibrium
 - ▶ Designs
 - ▶ Population-based (case-control, cohort)
 - ▶ Family-based (e.g., case-parent trios)
 - ▶ May be more appropriate for complex diseases
- ▶ Careful study design is essential and should follow epidemiologic principles

Genetic Epidemiology Process



Population Genetics

- ▶ Studies the origin, transmission and distribution of genetic variation within and between populations
- ▶ Provides essential concepts and information to any fields using genetic variation
- ▶ In genetic epidemiology:
 - ▶ Study design
 - ▶ Appropriate and powerful analytical approach
 - ▶ Adequate interpretation of results

Help from Population Genetics

▶ Linkage:

- ▶ Are models of locus homogeneity plausible?
- ▶ How can we calculate the probability that two relatives share genes identical by descent (IBD)?
- ▶ What are the expected values for IBD sharing between particular relative pairs?

▶ Association:

- ▶ How feasible are LD studies in current populations?
- ▶ What other reasons can explain observed associations?

Help from Population Genetics

- ▶ How is LD best measured?
- ▶ Which factors might influence LD detection in specific populations?
 - ▶ Mutation rates
 - ▶ Recombination rates
 - ▶ Population age
 - ▶ Migration patterns
- ▶ What study designs are most appropriate to exploit LD for disease gene mapping in different populations?

Help from Population Genetics

- ▶ Other reasons for observed association:
 - ▶ Population stratification / subdivision
 - ▶ Recent admixture
 - ▶ Genetic drift
 - ▶ Selection
 - ▶ Assortative mating
 - ▶ Type I error

References

- ▶ Gordis L. *Epidemiology, 4th Edition*. 2009
- ▶ Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology, 3rd Edition*. 2008.
- ▶ Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology*. 1993
- ▶ *Human Genome Epidemiology*, edited by Khoury MJ, Little J, Burke W. 2004
- ▶ Burton PR, Tobin MD, Hopper JL (2005) Key concepts in genetic epidemiology. *Lancet* 366:941-951 (and other articles from this Series on Genetic Epidemiology)
- ▶ Ellsworth DL, Manolio TA (1999) The emerging importance of genetics in epidemiologic research. Part I, II, and III *Ann Epidemiol* vol. 9

Exercise

- ▶ Read the Abstract and Methods of the 2 articles:
 - ▶ Wang et al. *Am J Epidemiol* 2009 169:633-641
 - ▶ Elliott et al. *JAMA* 2009 302:37-48
- ▶ Identify the study design used in each article.
- ▶ Discuss the biases that could potentially occur in each study, particularly considering the genetic effects.
- ▶ Give a few solutions to address these biases.
- ▶ Identify the risk/association measures used in each study and interpret each of them.

Exercise (continued)

- ▶ Could the authors have used the same measures of risk in the 2 studies? Why? Discuss the advantages of each measure.
- ▶ To support which causality inference criteria can Mendelian Randomization be used?
- ▶ The age-related macular degeneration study shows an example of a gene-environment interaction with a genetic factor of strong effect, which is not always the case
 - ▶ What challenges occur when gene-environment interactions are present?
 - ▶ What if a genetic effect cannot be detected in some environmental conditions?