National Cancer Institute

Maximizing Internal and External Validity in Epidemiology Studies

American Society of Human Genetics
October 24, 2007

Robert N. Hoover, M.D., Sc.D.
Director
Epidemiology and Biostatistics Program
October 24, 2007

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Validity (Internal Validity)

- Both patients and comparison groups are representative of the same "study-base" (population)

- If selective factors enter into the choice of patients, the same factors should enter into the selection of the comparison group

External Validity (Representativeness)

- The study-base for your cases and controls is similar in all relevant aspects to the population to which you wish to extrapolate the results.

Study Base (Internal Validity) Questions

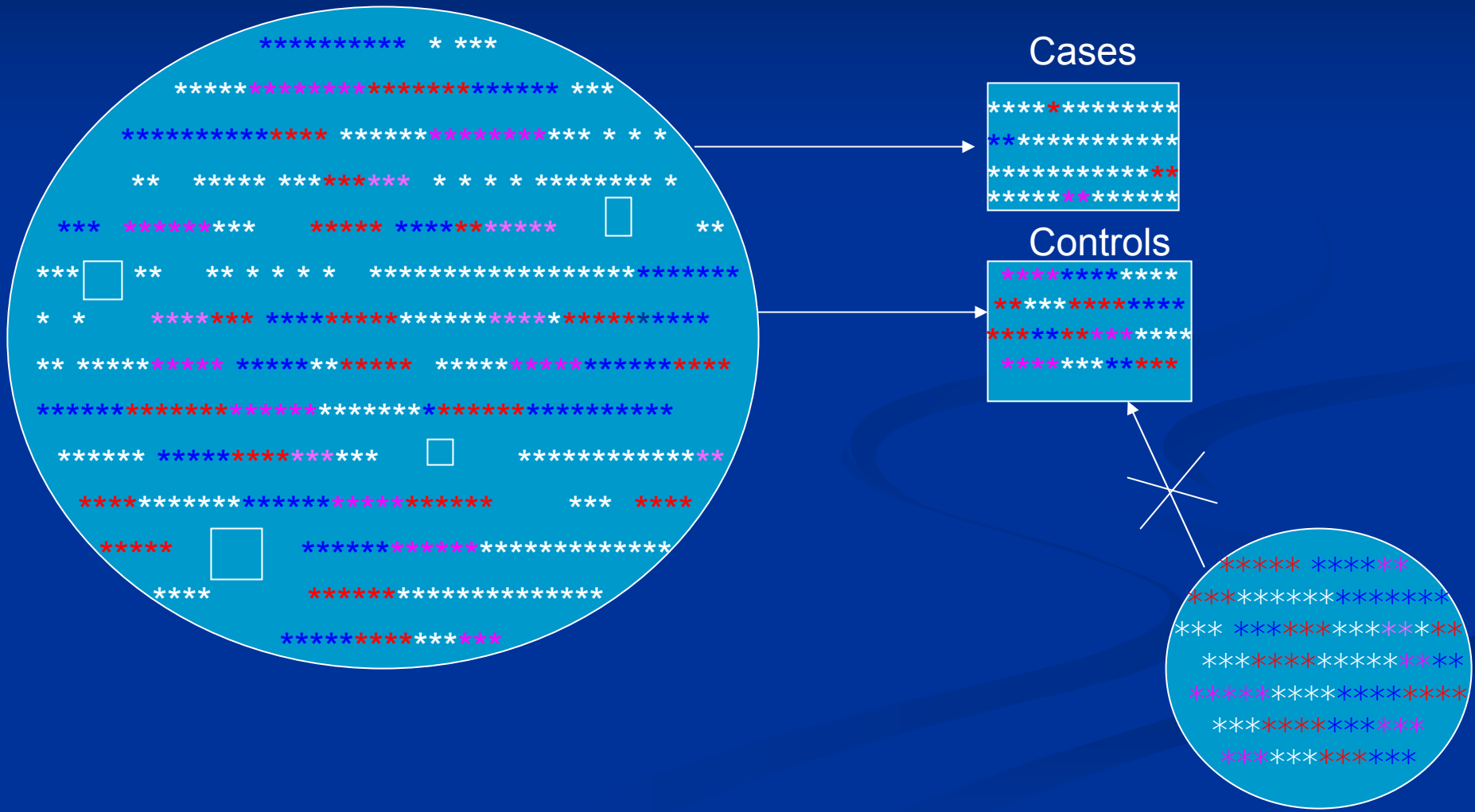
- If one of your controls had contracted the disease, would he/she be in your case group?
- If one of your cases had not been a case would he/she have been equally likely to be chosen as a control as any of the controls in your study?

- How did you end up with the cases you have?

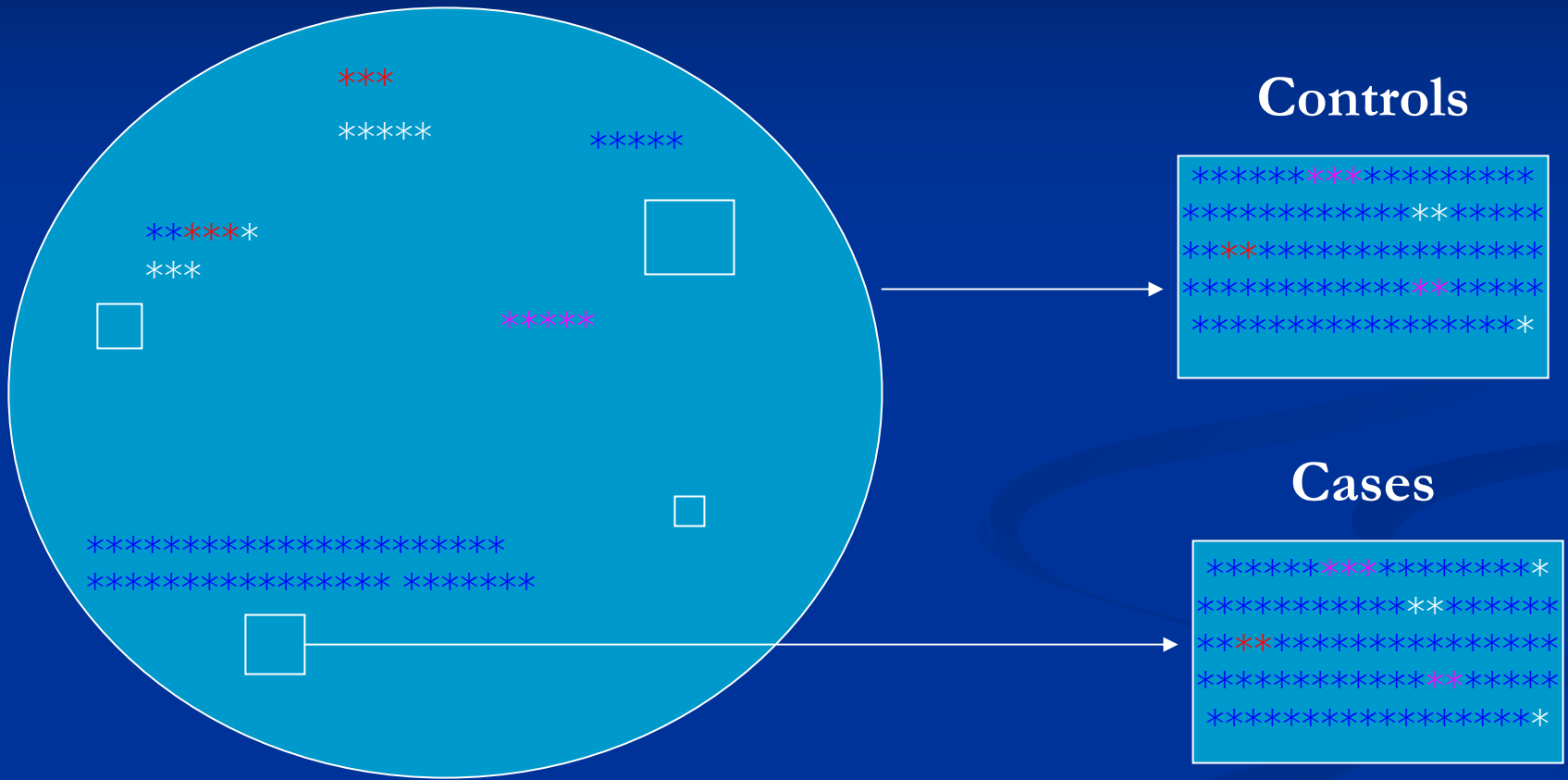
Common Starting Places for A Study Base

- General Population
- Special Population
  - HMO
  - Cohort
- Hospital/Clinic
- Neighborhood
- Families (Siblings, Spouses)

Population Study-Base



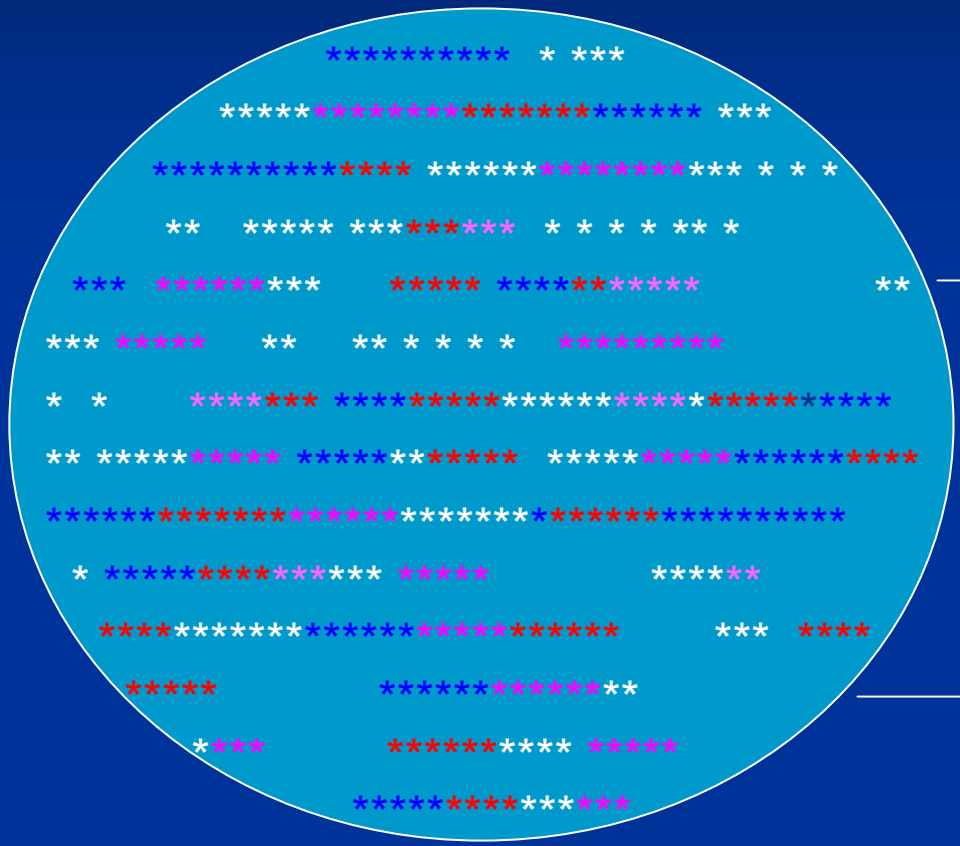
Hospital/Clinic Study-Base



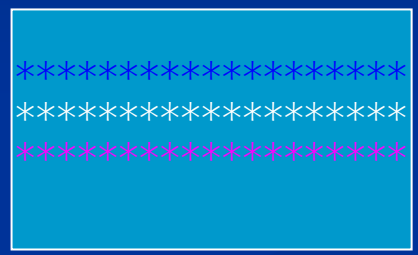
General Population-Based Study of Lymphoma

Interview + DNA	Cases	Controls
Participation Rate	68%	47%
Response Rate	53%	40%

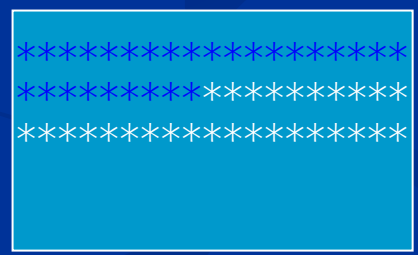
Population Study-Base With Response Bias



Cases



Controls



Maternal Urinary Mercury & Neural-Tube Defects in Mexican-Americans

	Cases	Study Controls	Natl Sample
% ≥ 5.62mg/L	28%	17%	5%
Odds Ratio		1.8	7.4

Endometrial Cancer & Estrogen Use

	Other Cancer Controls	D & C Controls
Relative Risk	10.8	1.7
P-value	0.005	0.12

External Validity (Representativeness)

- The study-base for your cases and controls is similar in all relevant aspects to the population to which you wish to extrapolate the results.

"Relevant Aspects"

Factor	Measure Affected
Underlying Biology	Relative Risk Absolute Risk Attributable Risk
Prevalence of Risk Factors That Modify The Effect	Same as above
Prevalence of The Study Risk Factor	Absolute Risk Attributable Risk

Underlying Biology/Presence of Modifiers

BrCa1 in high-risk families → 85% lifetime risk

BrCa1 in population sample → 50% lifetime risk

Heavy Smoking/Heavy Alcohol Use and Esophageal Cancer

	Socioeconomic Status		
	High	Medium	Low
Relative Risk	34	95	421

Prevalence of Risk Factors

Estrogen Therapy and Uterine Cancer

	California	Minnesota
Relative Risk	5	5
% Exposed	50	3
Etiologic Fraction (Attributable Risk)	67%	11%

Selection of Cases (Issues of External Validity)

- Self Report – validated or not
- Clinically Defined (Hospital, Registry, Physician)

- All
- Advanced

- Incident
- Prevalent

Prevalent Conditions

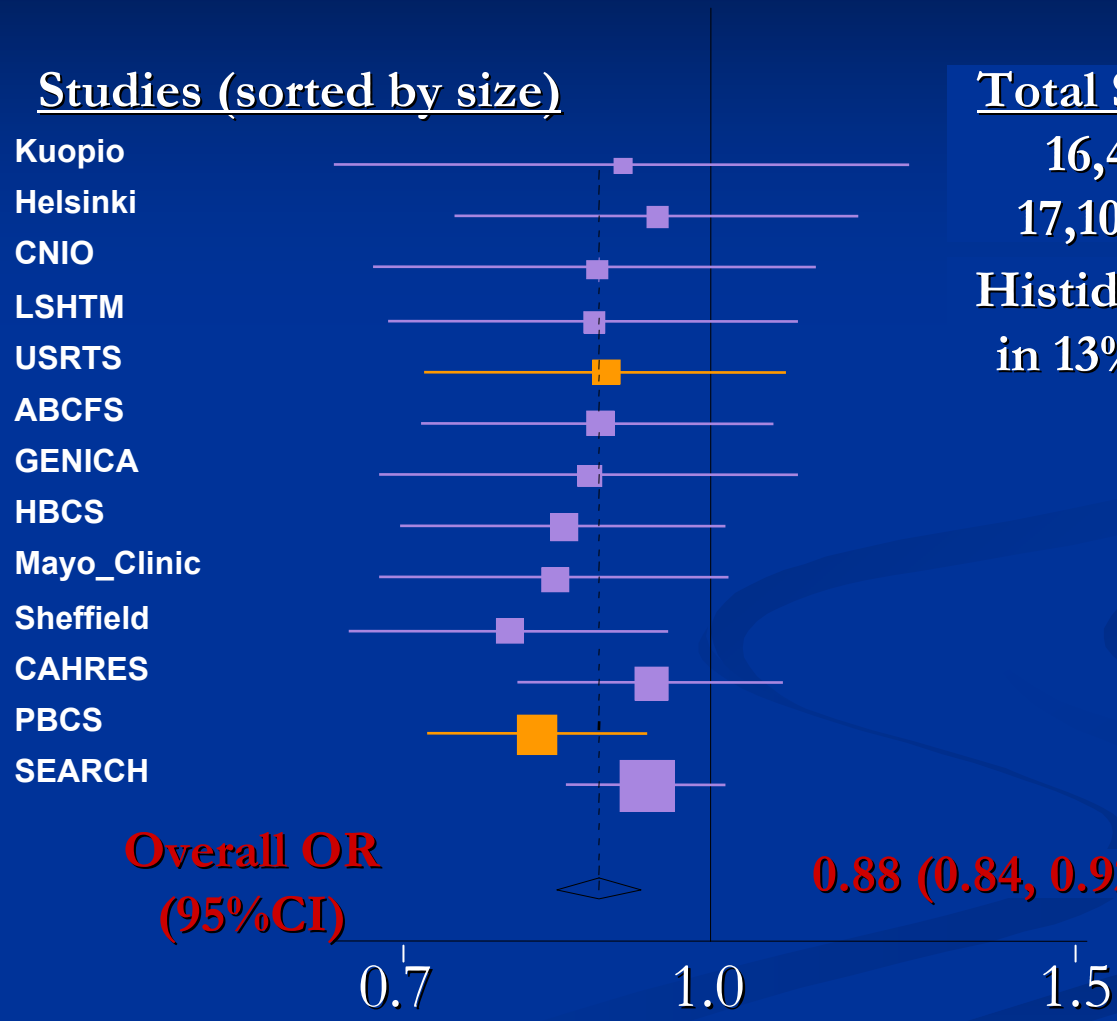
Prevalence = Incidence x Duration

Prevalent samples → weighted with long-term survivors

Selection of Controls

Cautionary Note Re: Extreme Phenotype And Violation of Study-Base Principle

Caspase 8 *(CASP8)* D302H Variant Decreases Breast Cancer Risk



Total Sample Size
16,423 cases
17,109 controls

Histidine (H) allele in 13% of controls

Nat Genet 2007;39:352-8

Lung Cancer Risk and CYP2D6*

	Study 1	Study 2	Study 3
Relative Risk	15. 6 (4.8 – 55.9)	6.1 (2.2 – 17.1)	0.6 (0.3 – 1.2)
Epidemiologic Quality	Low	Intermediate	High
(% participation)	(?)	(26%)	(80%)

* Risk of homozygous extensive metabolizers compared to homozygous poor metabolizers.

Study-Base Flaws in GWAS

False Positives

Better Designs in Replications

False Negatives

No GXE

Likely not great

GXE

Could be Substantial

Summary

- Internal Validity is paramount to achieve case and control series that are comparable in order to confidently uncover causal risk factors
- The key to achieving this comparability is to define the "study-base" for which the cases are all, or a representative sample of all, of those affected in the study-base, and to draw a representative sample of unaffected from this same study-base as controls.

Summary (con't)

- External Validity relates not to the validity of the association noted, but to how it may be extrapolated to other circumstances.
- The greater the difference between the study-base for the study and the populations to which the results are extrapolated, the more problematic the extrapolation.

Conclusion

- Genetic risk factors are likely to be less affected by potential selection biases introduced by study-base flaws than environmental and life-style risk factor.
- There will, however, be some effect, and for genetic risk factors operating through GXE interaction, the effects could be substantial.

Selection Bias in Genetic Studies

- Genetic variants associated with selection factors
- Genes associated with risk factors for the disease that are biased by sampling
  (e.g. obesity, smoking, alcohol)
- Genetic effect from altered GXE for Biased Exposures