

Study Design and Analysis in Molecular Epidemiology (I)

Andrew Rundle, DrPH

Associate Professor of Epidemiology
Columbia University
Mailman School of Public Health



Class Outline

Study Design and Analysis in Molecular Epidemiology (1)

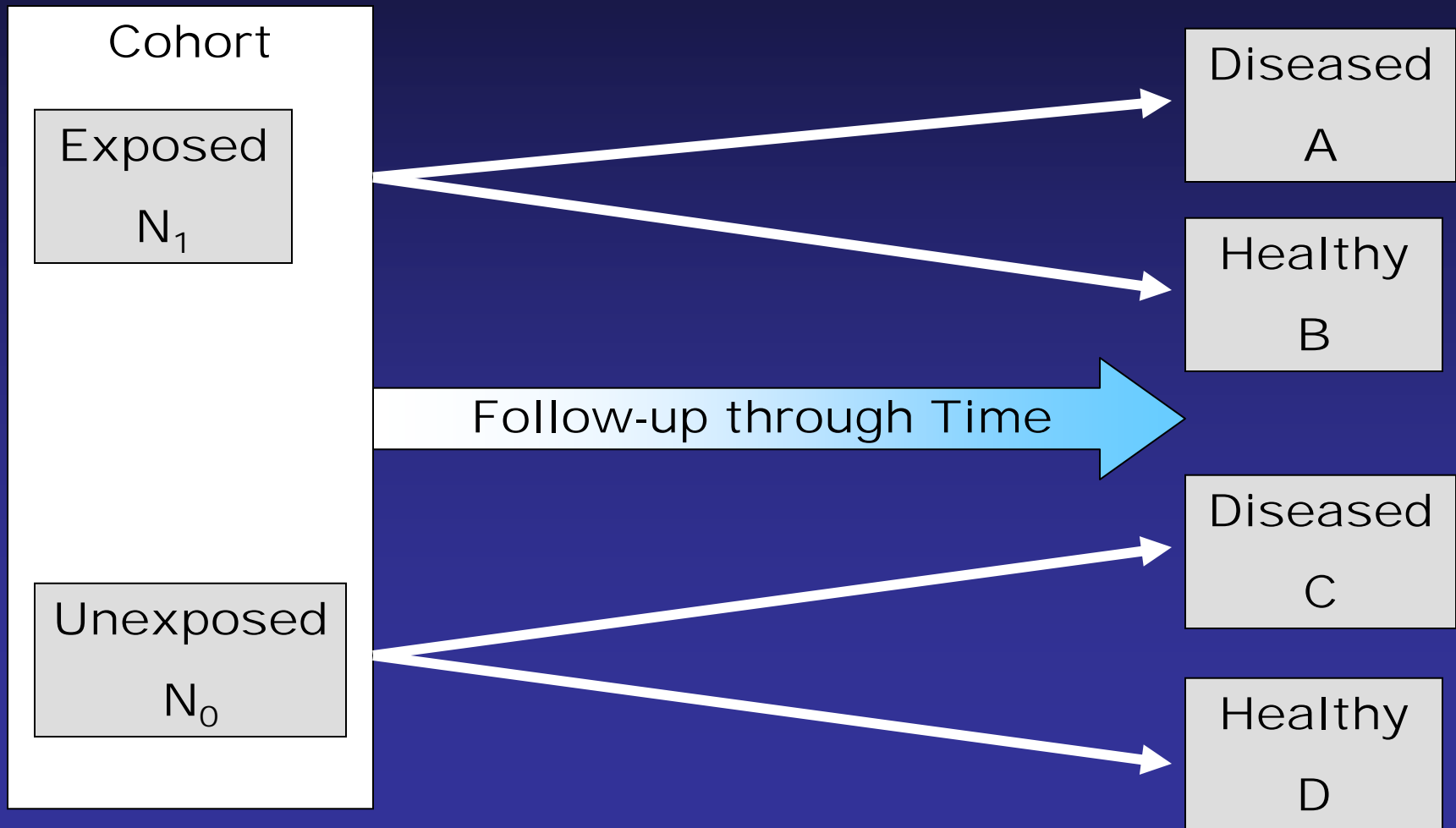
- Cohort study design
 - Measures of effect that can be calculated from the design
- Case-control design
 - Case-control studies conducted within an existing cohort (nested case-control studies)
 - Relationship between OR and RR and IRR
 - Case-control studies conducted in population and hospital settings

Study Design and Analysis in Molecular Epidemiology (2)

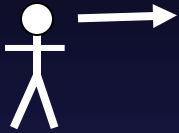
- Nested case-control and case-cohort designs
- Issues in using biomarkers in nested case-control and case-cohort designs.

The Cohort Study

The cohort study is the primary conceptual study design of observational epidemiology.



Prospective Cohort Study



Cohort Follow-up

Examples: EPIC, UK-BioBank, HEALS

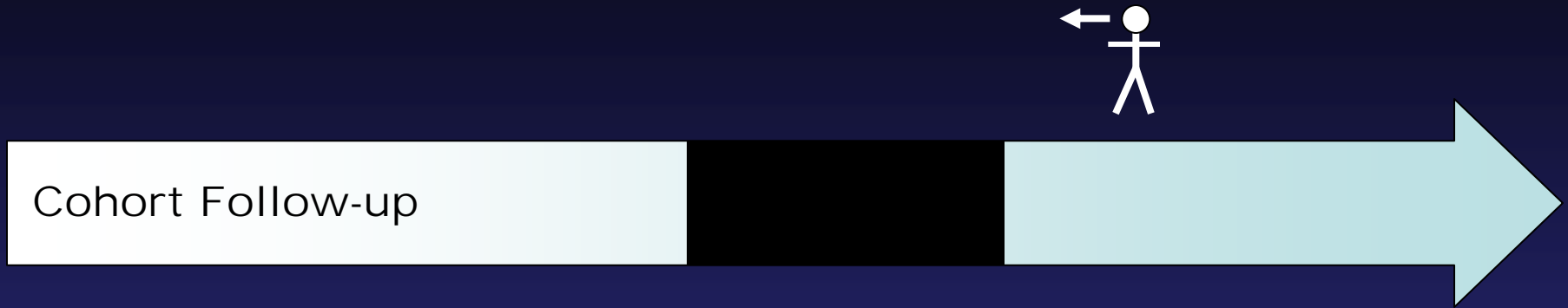
Advantages:

Full control over sampling, data collection, bio-specimen handling and storage.

Disadvantage:

Hugely expensive, and a long time period between initiation and any results.

Retrospective Cohort Study



Examples: Child Development and Health Study

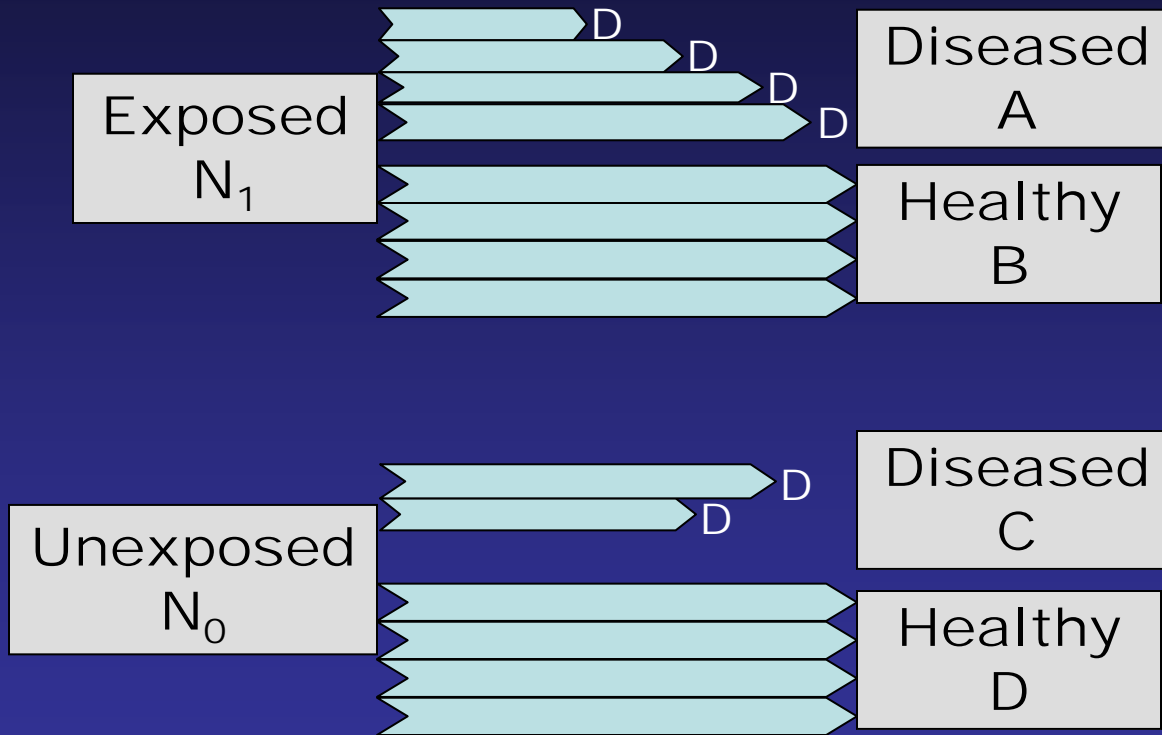
Advantages:

Can be implemented and produce results very fast, relatively inexpensive.

Disadvantage:

Relies on finding a suitable historic cohort, no control over data collection, sample collection or handling.

Standard Cohort Analyses: Calculation of the Relative Risk



Relative Risk

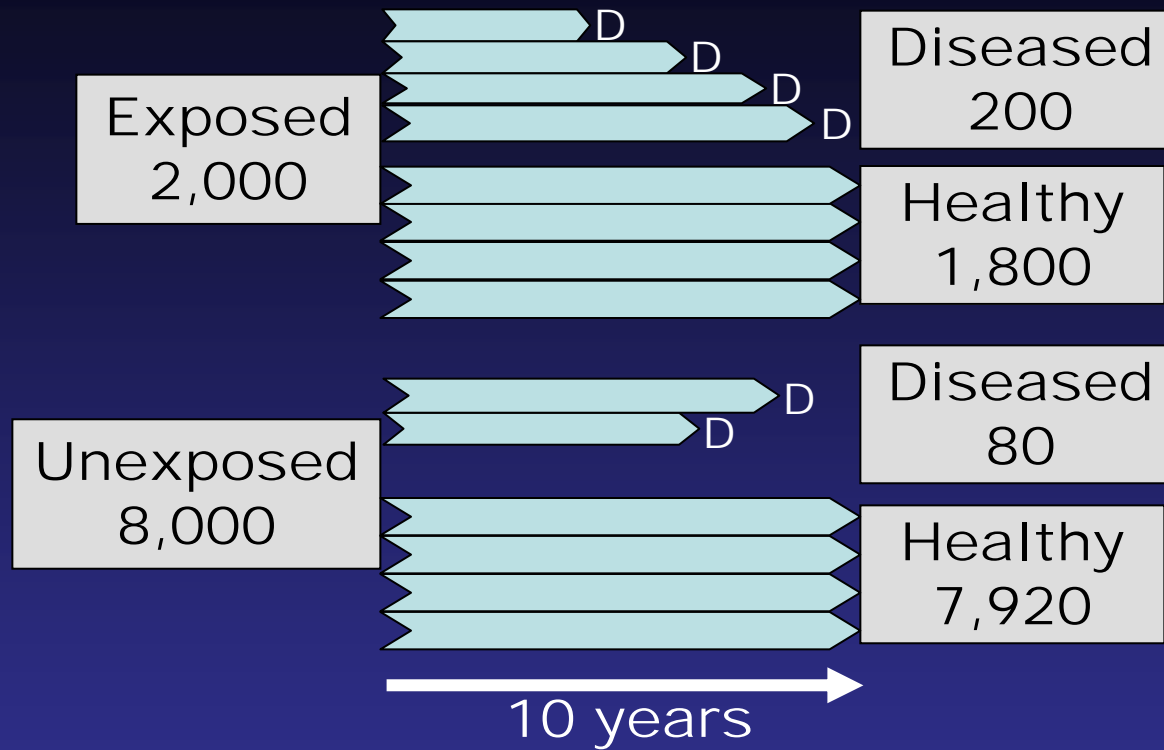
$$I_1 = A/N_1$$

$$I_0 = C/N_0$$

$$RR = I_1/I_0$$

$$RR = \frac{A/N_1}{C/N_0}$$

Calculation of the Relative Risk



Relative Risk

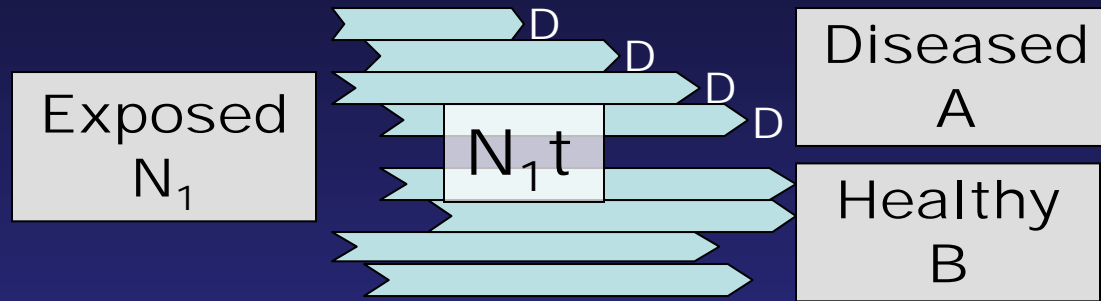
$$I_1 = 200 / 2,000$$

$$I_0 = 80 / 8,000$$

$$RR = I_1 / I_0 = 200 / 2,000 / 80 / 8,000 = 10$$

Standard Cohort Analyses

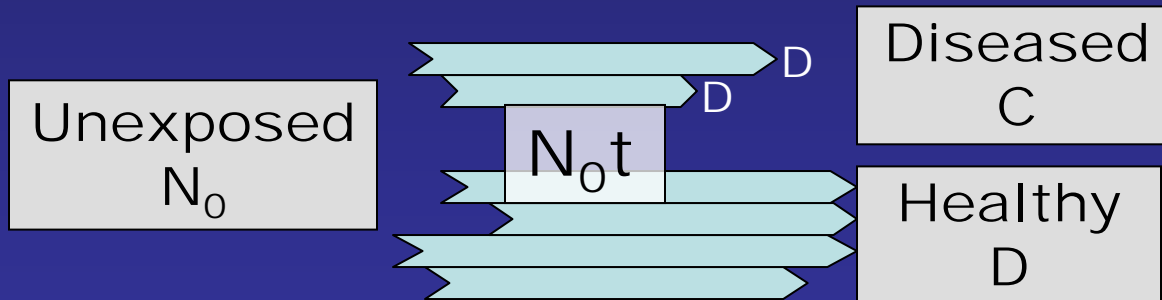
Calculation of the Incidence Rate Ratio



Incidence Rate Ratio

$$IR_1 = A/N_1 t$$

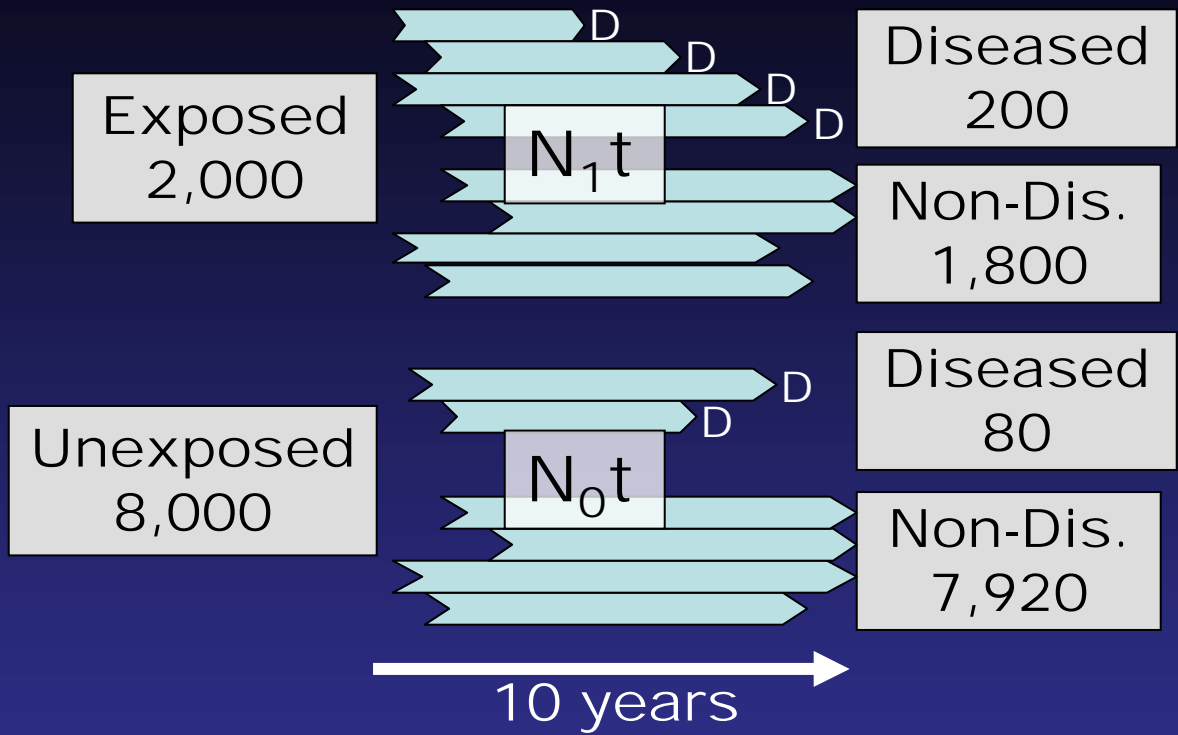
$$IR_0 = C/N_0 t$$



$$IRR = IR_1/IR_0$$

$$IRR = \frac{A/N_1 t}{C/N_0 t}$$

Calculation of the Incidence Rate Ratio



Mean follow up in diseased = 6 yrs

Mean follow up in non-diseased = 8 yrs

$$N_1t = 1,200 + 14,400$$

$$N_0t = 480 + 63,360$$

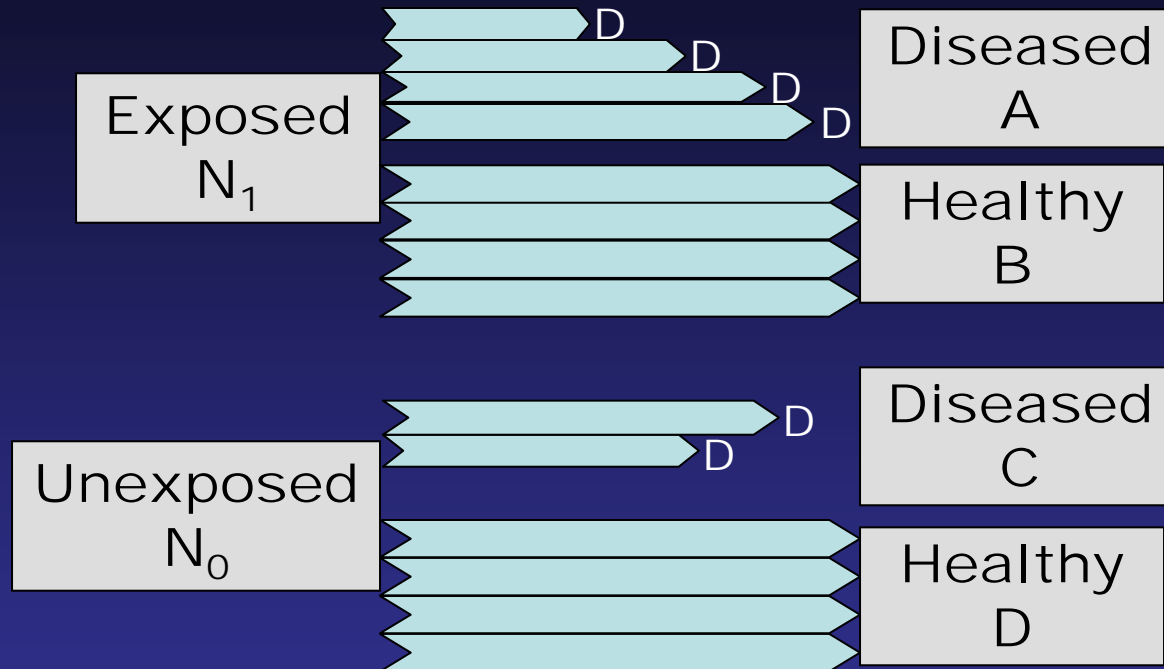
Incidence Rate Ratio

$$IR_1 = 200 / 15,600$$

$$IR_0 = 80 / 63,840$$

$$IRR = I_1 / I_0 = (200 / 15,600) / (80 / 63,840) = 10.23$$

Standard Cohort Analyses: Calculation of the Exposure Odds Ratio



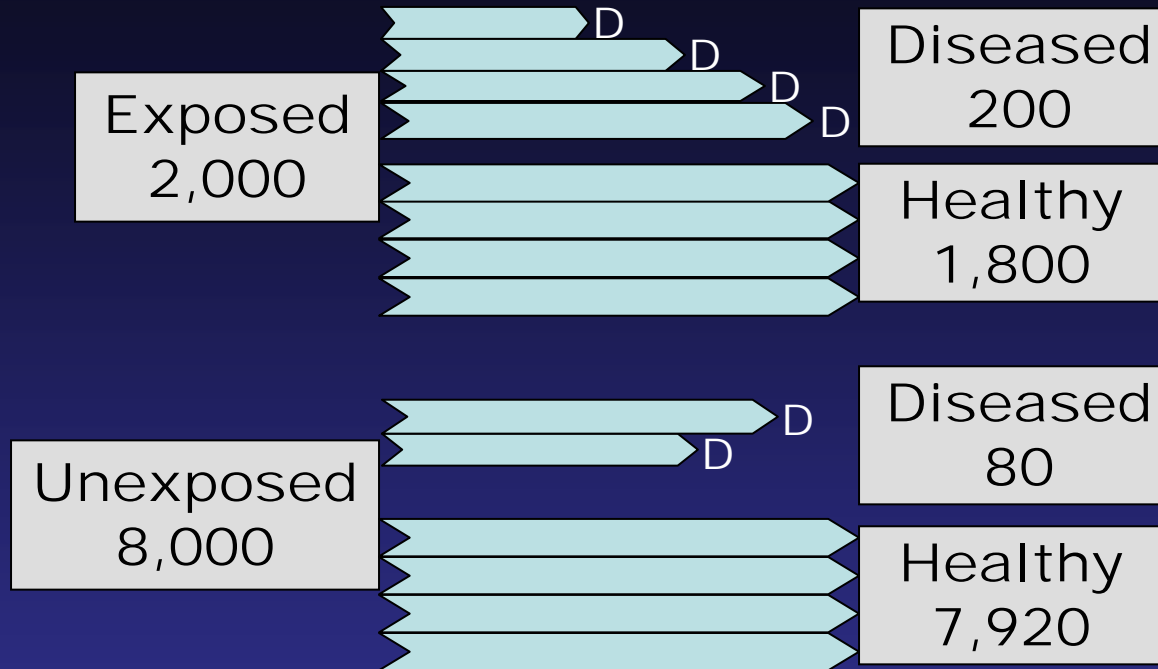
Exposure Odds Ratio

$$\text{Odds}_1 \text{ (odds of D in } N_1) = A/B$$

$$\text{Odds}_0 \text{ (odds of D in } N_0) = C/D$$

$$\text{Exposure OR} = \text{Odds}_1 / \text{Odds}_0 = A/B / C/D = AD/BC$$

Calculation of the Exposure Odds Ratio



Exposure Odds Ratio

$$\text{Odds}_1 (\text{odds of D in } N_1) = 200 / 1,800$$

$$\text{Odds}_0 (\text{odds of D in } N_0) = 80 / 7,920$$

$$\text{Exposure Odds Ratio} = 200 / 1,800 / 80 / 7,920 = 11$$

Relationships Between the RR, IRR and Exposure OR

- The RR is traditionally considered to be the primary measure of interest.
- When disease is rare and follow-up is complete $IRR \sim RR$ and $EOR \sim RR$.

That is:

$$N_{1t} / N_{0t} \sim N_1 / N_0$$

$$B / D \sim N_1 / N_0$$

- When disease is common $IRR > RR$ and $EOR \gg RR$.

Measures of Incidence

- **Cumulative Incidence:** Disease & non-disease status is known only at the end of follow-up, exact data of diagnosis is unknown.
 - Relative risk and exposure odds ratio can be calculated.
- **Incidence Density:** Exact date of diagnosis is known and time to event can be calculated.
 - Relative risk, incidence rate ratio and exposure odds ratio can be calculated.

Cumulative Incidence Assessment of Common Outcomes

- Often in molecular epidemiology we study common outcomes where we don't know the exact date the outcome occurred (cumulative incidence).
 - e.g. studies on adenomas of the colon or Barrett's esophagus or studies where the presence of a biomarker is the outcome.
- Time to disease and IRR cannot be calculated and the OR overestimates the RR

Biomarker Based Follow-up of Vinyl Chloride Workers.

Retrospective cohort study of vinyl chloride workers with blood samples collected at the end of follow-up. Blood samples assayed for the presence of mutant P53 protein.

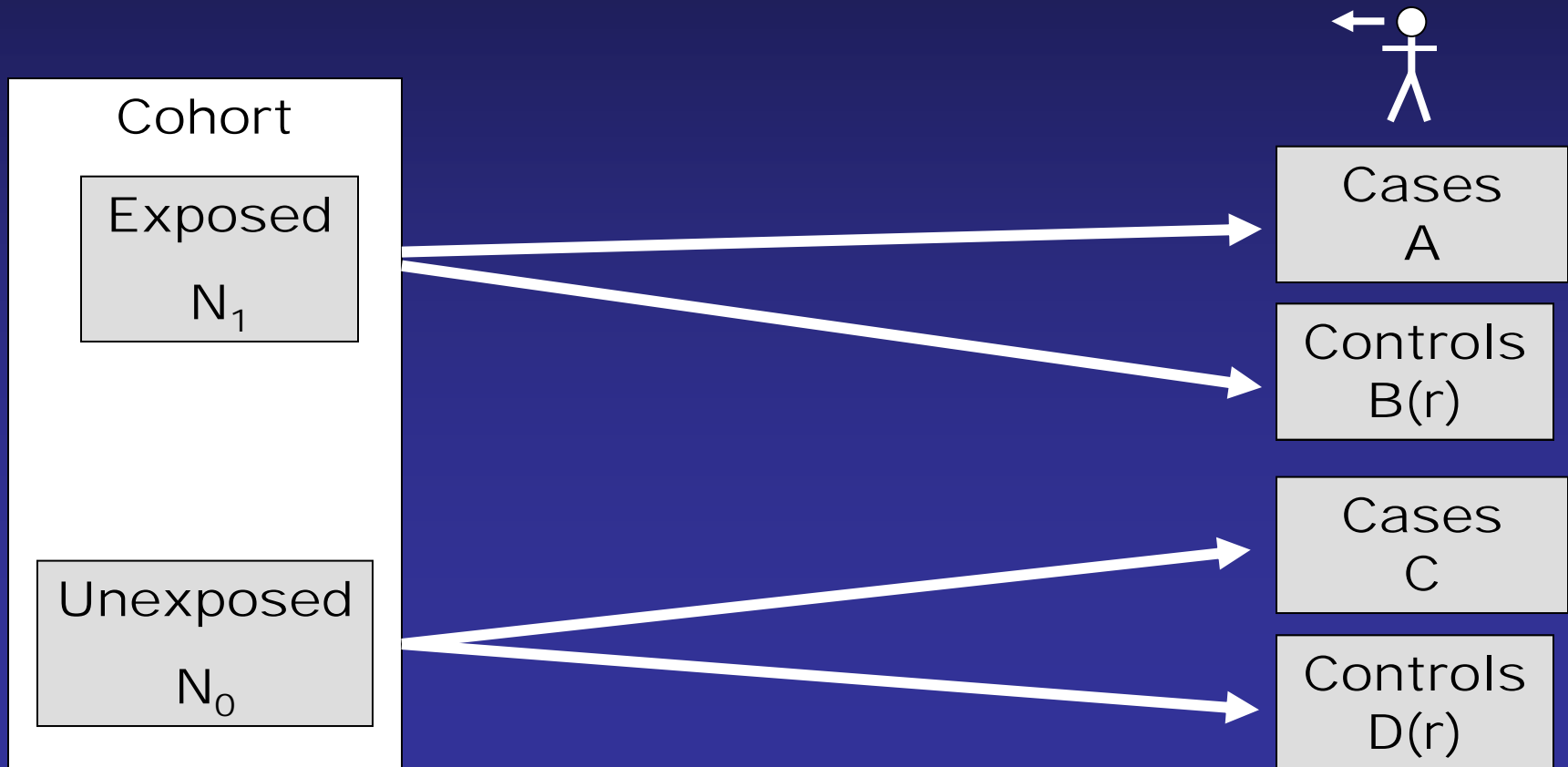
Exposure ¹	P53+	P53-	OR	RR	% over est.
<500	16	38	1	1	-
501-2,500	21	41	1.22	1.15	6%
2,501-5,000	24	27	2.11	1.59	33%
>5,000	30	28	2.54	1.75	45%

1 PPM-Years of VC exposure

[Smith et al., 1998]

Case-Control Studies

- A case-control study is understood as a cohort study done in reverse.
- Not to be confused with a retrospective cohort, here we select on disease status, not exposure.



Case-Control Studies

- Derive their validity from their ability to estimate results that otherwise would have been generated by a cohort study.
- All case-control studies must be understood to occur within a cohort study.
- Cases and controls are sampled from an underlying cohort.
 - The cohort may exist, that is, there is already a fully enrolled cohort.
 - The cohort may be theoretical, that is, the cohort is not enumerated. The cohort that produced the cases is imagined and controls are sampled from it.

Case-Control Studies

Purpose of Controls:

To estimate the prevalence of exposure and covariates in the source population from which the cases are derived.

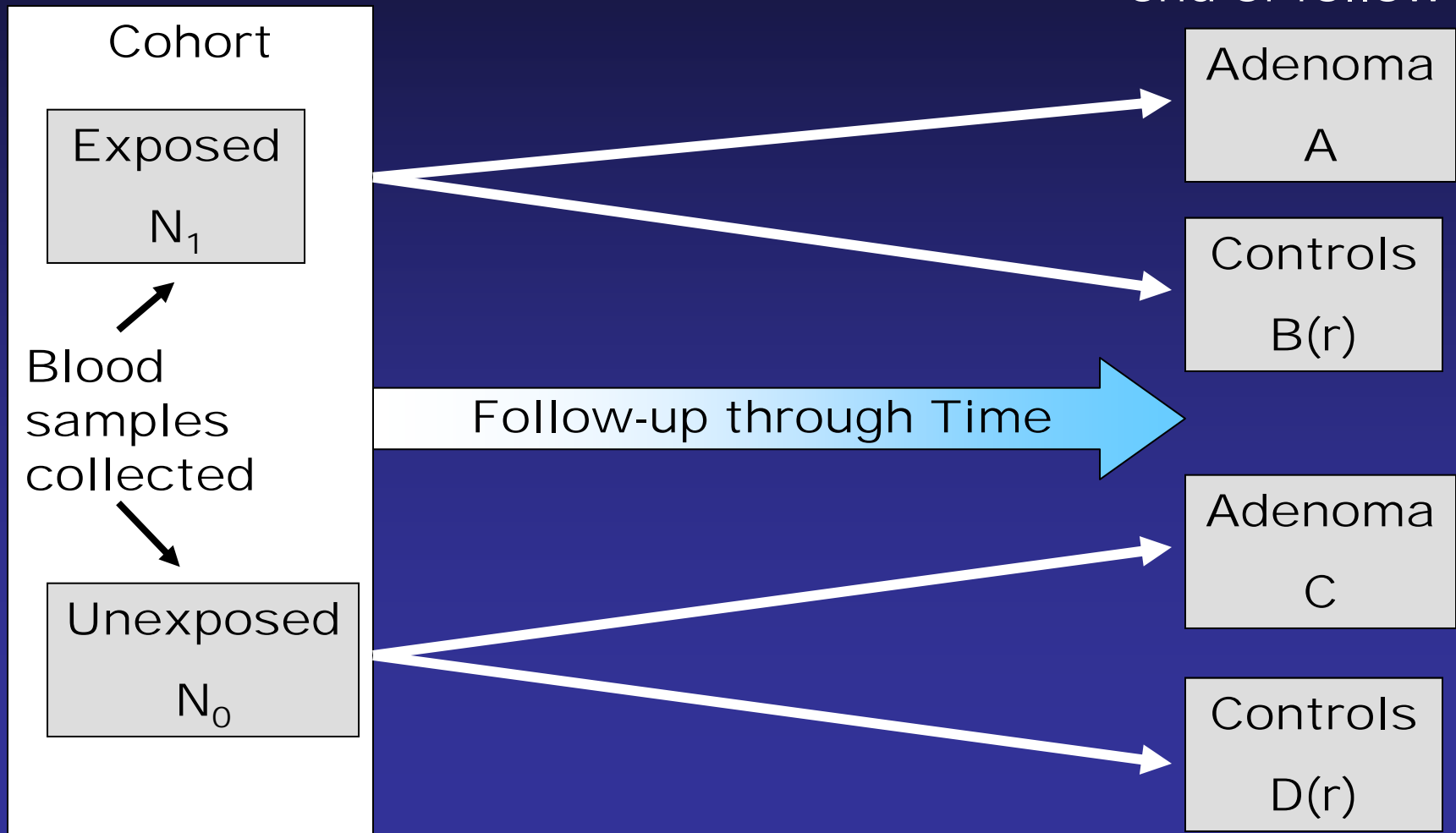
Case-Control Studies

- Nested Case-Control Studies
 - Conducted within a cohort that has been followed up for disease outcomes.
 - Efficiency is gained by collecting exposure data from only a sample of the cohort.
 - The cohort represents the sampling frame for the cases and controls.
- Population or Hospital Based Studies
 - Conducted outside of an existing cohort, in a defined population or clinical setting.
 - The sampling frame for the controls is less clear.

Cumulative Incidence Case-Control Study in a Cohort

Colonoscopy at baseline

Colonoscopy at end of follow-up



Case-Control Study in a Cohort Cumulative Incidence Type

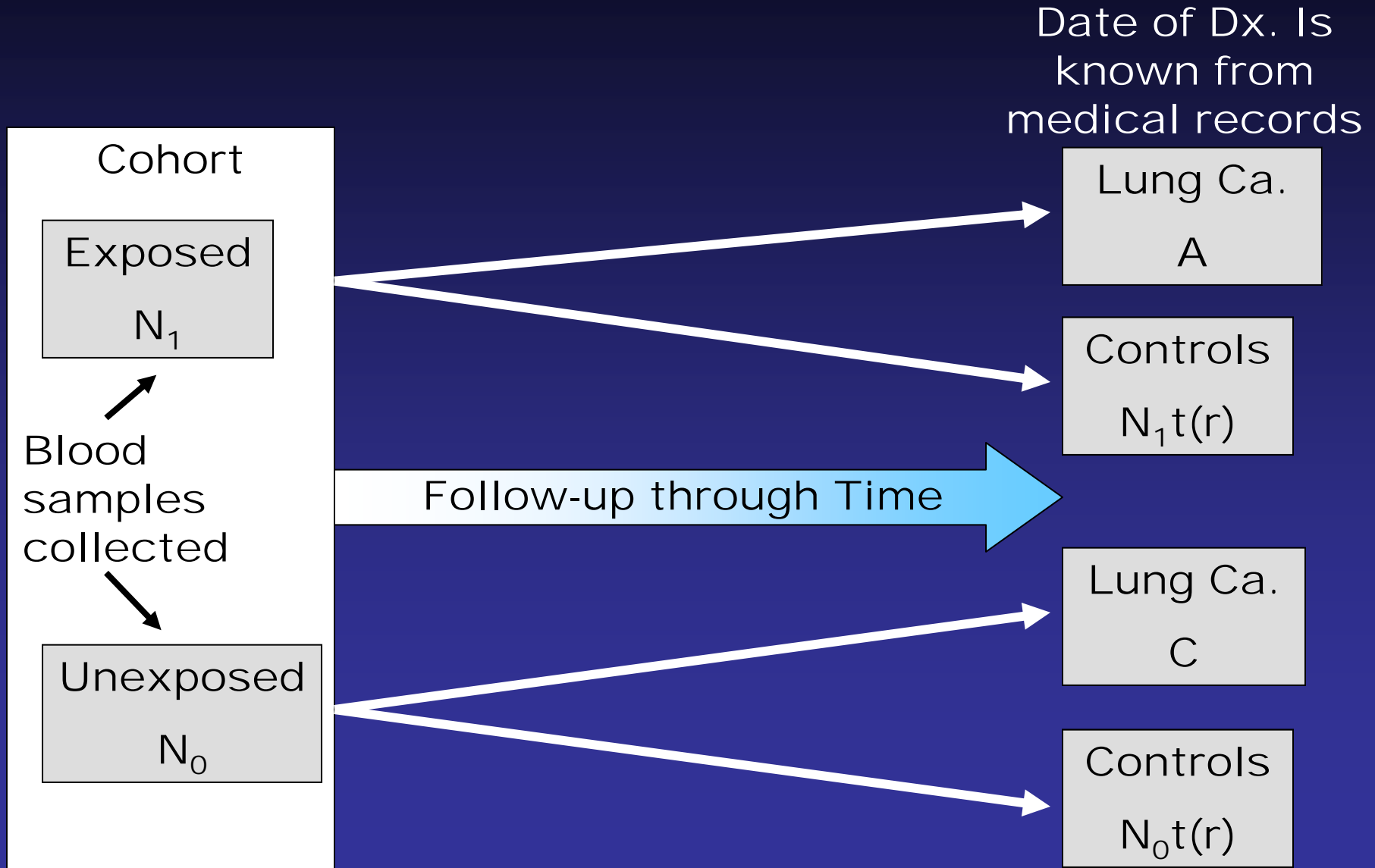


All of the cases and a random sample, r , of the remaining healthy subjects are used.

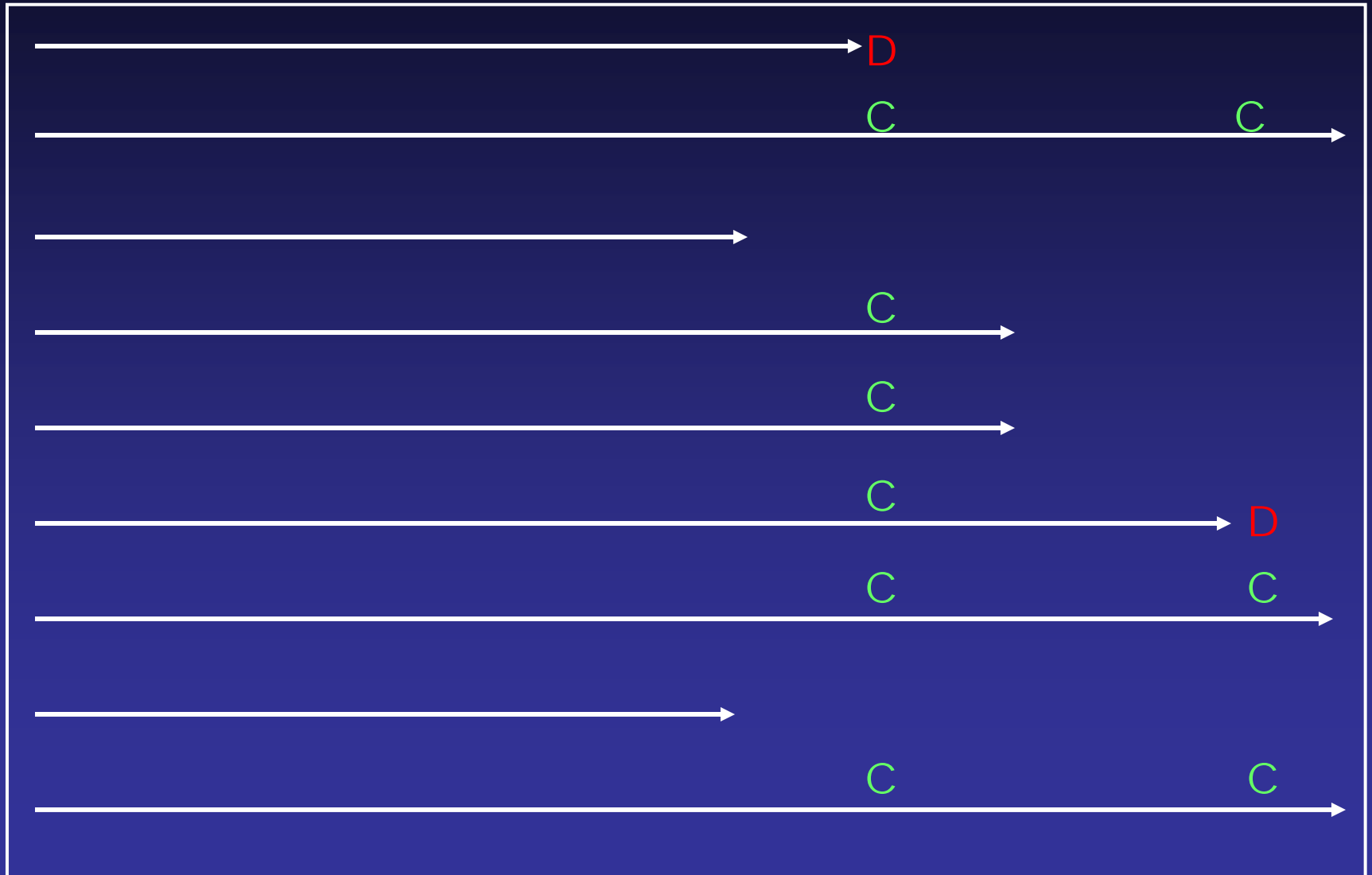
	Cases	Controls
Exposed	A	B (r)
Unexposed	C	D (r)

$$\text{Disease OR} = A / C / B(r) / D(r) = AD / BC = \text{Exposure OR}$$

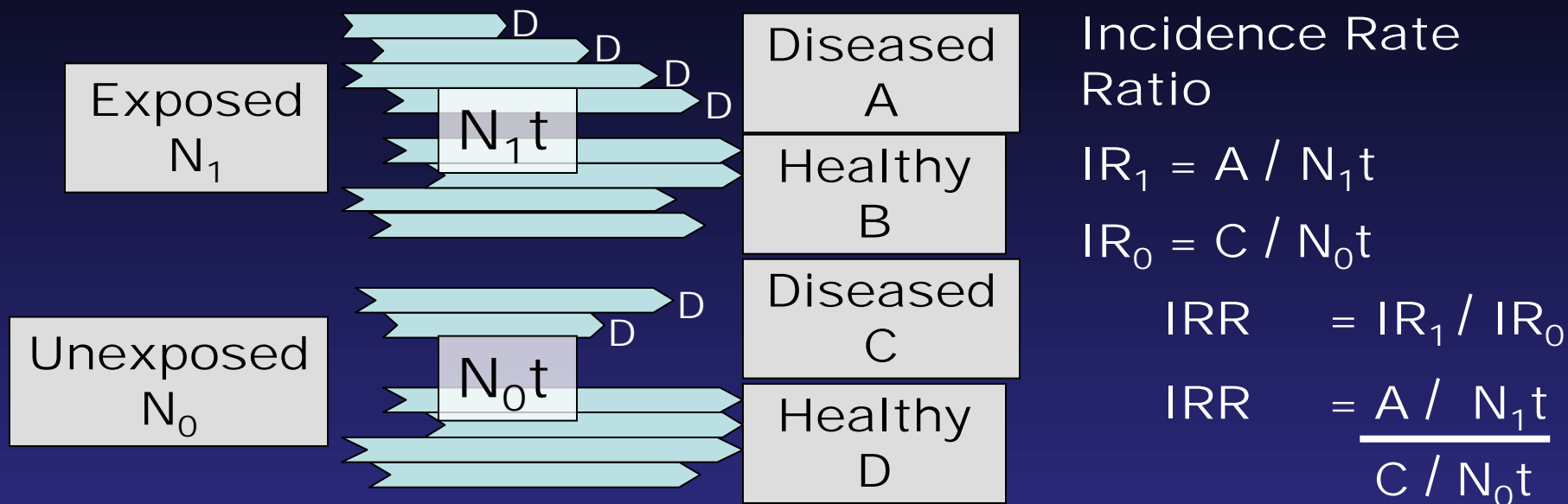
Incidence Density Case-Control Study in a Cohort



Nested Case-Control Study: Incidence Density Sampling



Case-Control Study in a Cohort Incidence Density Type



All of the cases and a random sample, r , of person time are used.

	Cases	Controls
Exposed	A	$N_1 t (r)$
Unexposed	C	$N_0 t (r)$

$$\text{Cross product} = \frac{AN_0 t (r)}{CN_1 t (r)} = \frac{A / N_1 t}{C / N_0 t} = IRR$$

Case-Controls Studies Outside of Existing Cohorts

Population based case-control studies

- Cases are selected from a spatially and temporally defined population.

For example – all cases of prostate cancer occurring in Shanghai from 2008 – 2012.

- Controls are selected from the same spatially and temporally defined population.

For example – a random sample of male residents of Shanghai from 2008 – 2012.

- The conceptual cohort is the population of Shanghai for that 4 year period.

Case-Controls Studies Outside of Existing Cohorts

Hospital based case-control study

- Cases are selected from a hospital(s) over a defined period.

For example: Prostate cancer cases from the Hua Shan Hospital.
- Controls are selected from other services or wards from the same hospital over the same period.

For example: Random sample of men attending the Hua Shan emergency room.
- The conceptual cohort is the population of people, who if they had developed prostate cancer, would be treated at Hua Shan.

Case-Controls Studies Outside of Existing Cohorts

Cumulative incidence sampling.

Cases are recruited until a fixed sample is achieved, and then controls are sampled from among the population that didn't develop prostate cancer during the recruitment period.

For a rare disease $OR \sim RR$ from a cohort study of this population.

Case-Controls Studies Outside of Existing Cohorts

Incidence density sampling

Cases and controls are recruited at the same rate over the recruitment period. It is allowable for a control recruited early in recruitment to later become a case.

OR = IRR from a cohort study of this population.

Case-Controls Studies Outside of Existing Cohorts: Control Selection

- The most difficult aspect of conducting a population or hospital case-control study is control selection.
- Again, the purpose of controls:
 - To estimate the prevalence of exposure and covariates in the population from which the cases arose.
- For a case-control study done within an existing cohort defining that population is easy.
- For a case-control study done outside of an existing cohort defining that population can be difficult.

Case-Controls Studies Outside of Existing Cohorts: Control Selection

Guidelines

- A control is some one, who if they had developed the disease of interest, would have been enrolled as a case in your study.
- For both our Shanghai population study or our Hua Shan hospital study, some one who, if they got prostate cancer, would have gone to Beijing for treatment can't be a control.
- For our Hua Shan hospital study, someone who, if they got prostate cancer, would have gone to Chong Shan hospital can't be a control.

Columbia University Breast Cancer Study

- Cases: Breast cancer patients seen at Columbia's breast service.
- Control group 1: Benign breast disease patients seen at Columbia's breast service.
- Control group 2: Women having gynecological screening whose doctors refer women to Columbia's breast service for breast complaints.

Molecular Epidemiologic Case-Control Studies

- Optimally done nested in a case-control study where bio-samples were collected at baseline enrollment.
- The need to get bio-samples from controls makes molecular epidemiologic studies outside of existing cohorts more difficult.
- Many studies in the literature have very poor control selection.
- Beware the “case-series and some other people study” masquerading as a case-control study.

Case-Series and Some Other People Study

- Study of adducts in breast tissue. Cases enrolled from a hospital in Texas, "controls" were tissue samples from a NIH tissue bank.
- Proteomic study of breast cancer. Cases were from hospitals in Buffalo, NY and Philadelphia PA, "controls" were blood donors from Germany.

Conclusions

- The cohort study is the fundamental study design of observational epidemiology.
- Case-control studies (and all other designs) gain their validity from their ability to estimate the results of a cohort study.
- Case-control studies are understood to occur within the context of cohort study.
 - Imagine the cohort study you would do if you had the resources, then sample cases and controls from within that cohort.