

Study Design and Analysis in Molecular Epidemiology (II)

Andrew Rundle, Dr.P.H.

Department of Epidemiology
Mailman School of Public Health
Columbia University

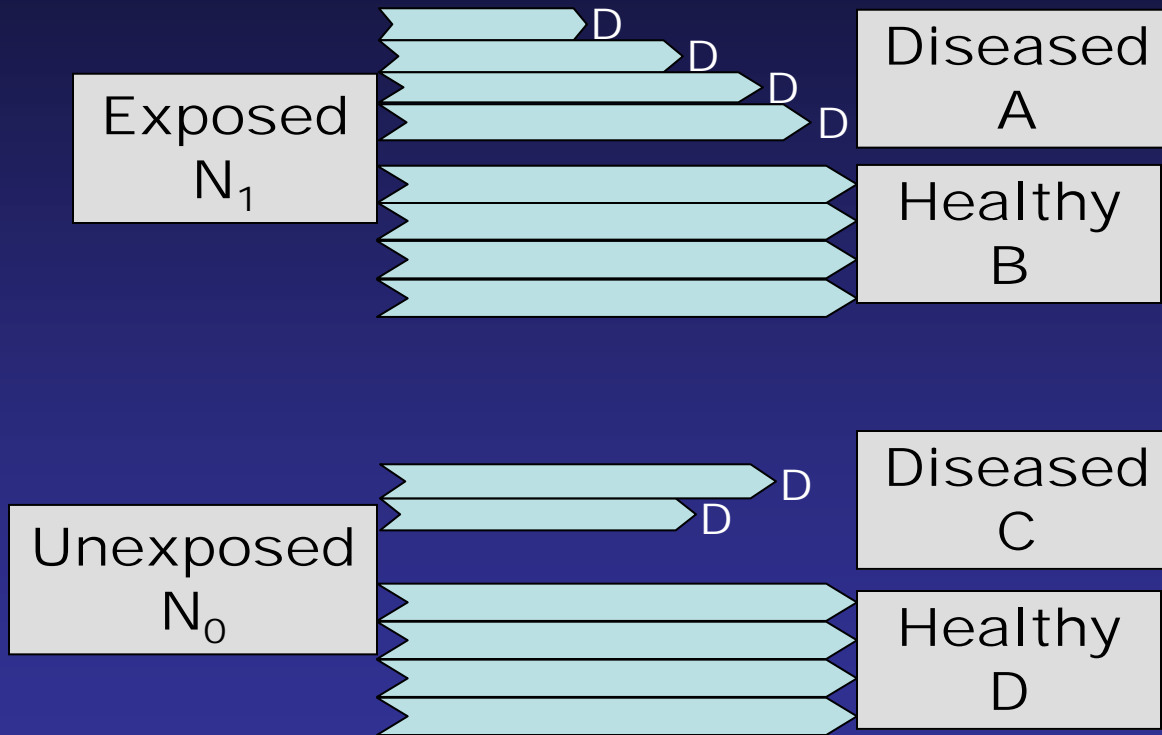


Using Biomarkers in Large Cohorts

There are now several large cohort studies with stored blood samples that can be used for molecular epidemiologic studies (i.e. EPIC, PHS, Nurses Health, UK-Biobank, HEALS). But it is inefficient to assay samples from all of the members of the cohort.

Options: Nested Case-Control Studies and Case-Cohort Studies

Standard Cohort Analyses



Relative Risk

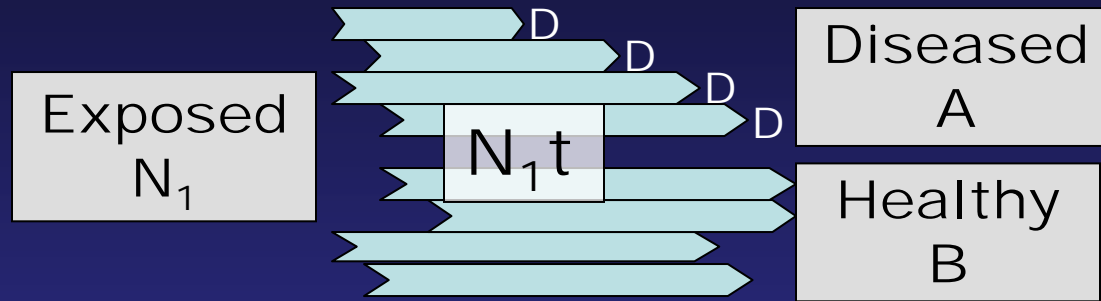
$$I_1 = A/N_1$$

$$I_0 = C/N_0$$

$$RR = I_1/I_0$$

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

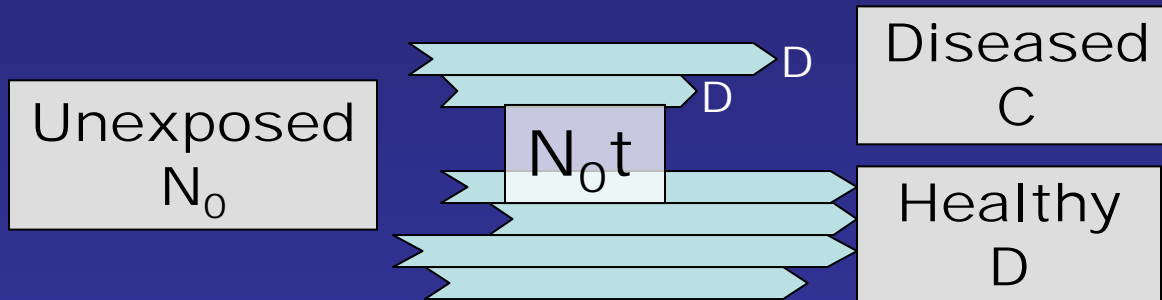
Standard Cohort Analyses



Incidence Rate Ratio

$$IR_1 = A/N_1t$$

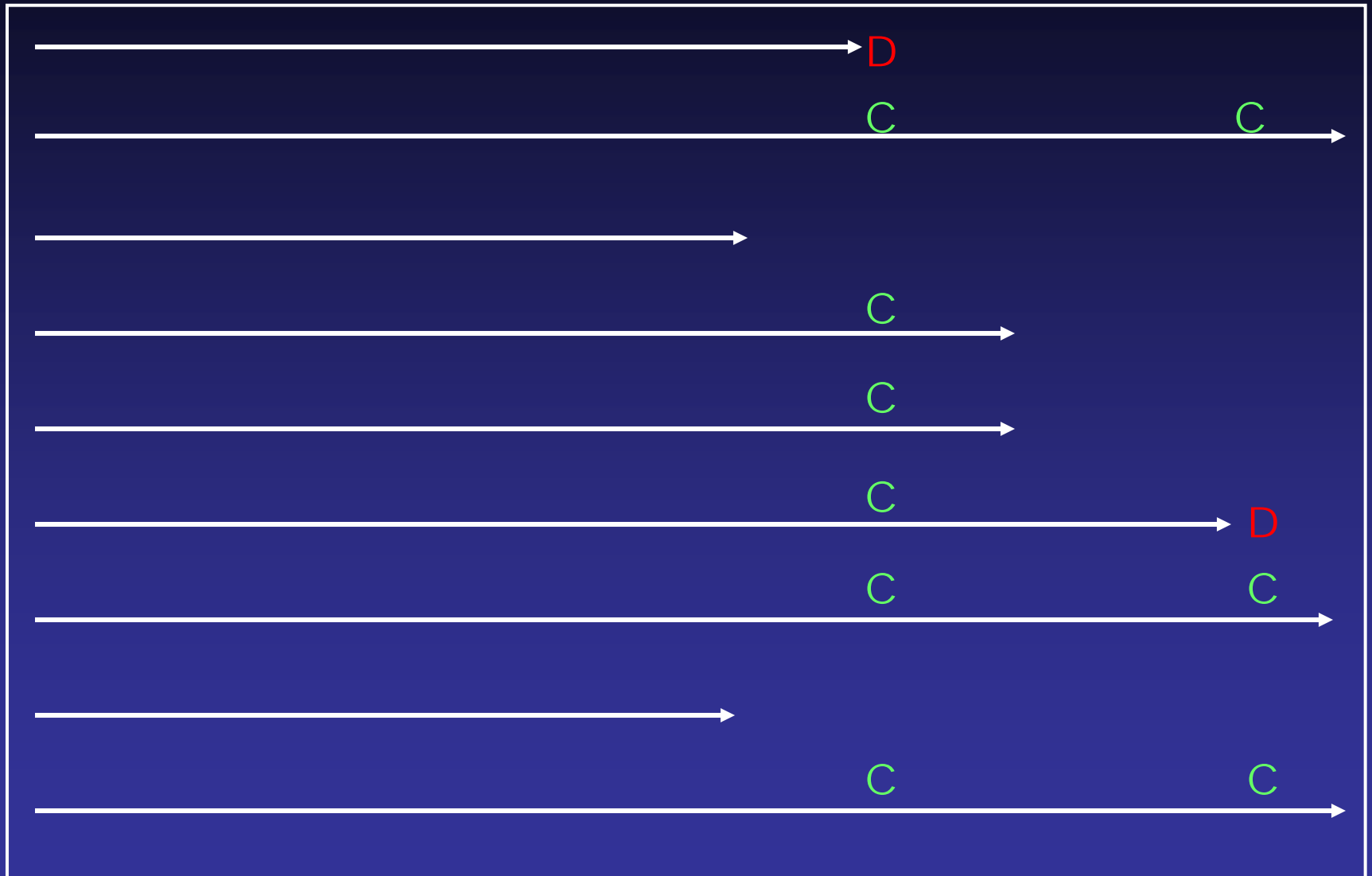
$$IR_0 = C/N_0t$$



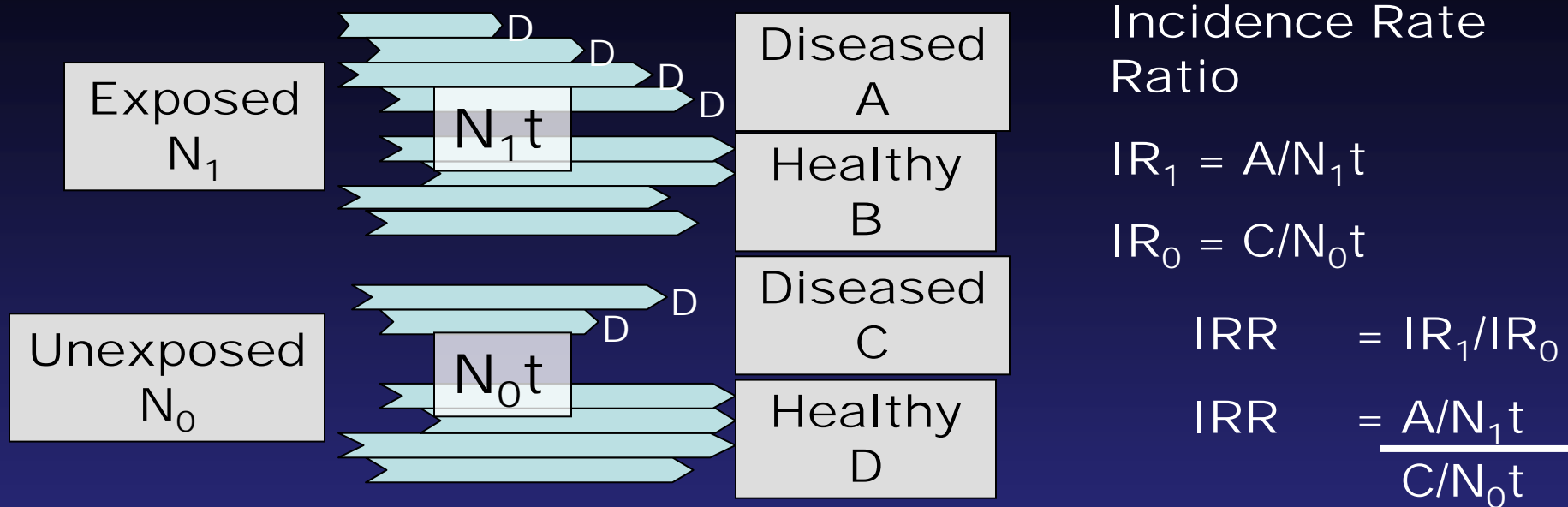
$$IRR = IR_1/IR_0$$

$$IRR = \frac{A/N_1t}{C/N_0t}$$

Nested Case-Control Study



Nested Case-Control Study



In a nested case-control study all of the cases and a matched sample, r , of the exposed and unexposed person time are used

	Cases	Controls
Exposed	A	$N_1t (r)$
Unexposed	C	$N_0t (r)$

$$\text{Cross product} = \frac{AN_0t (r)}{CN_1t (r)} = \frac{A/N_1t}{C/N_0t} = IRR$$

Matching in Nested Case-Control Studies

All nested case-control studies match controls to cases on length of follow-up in cohort.

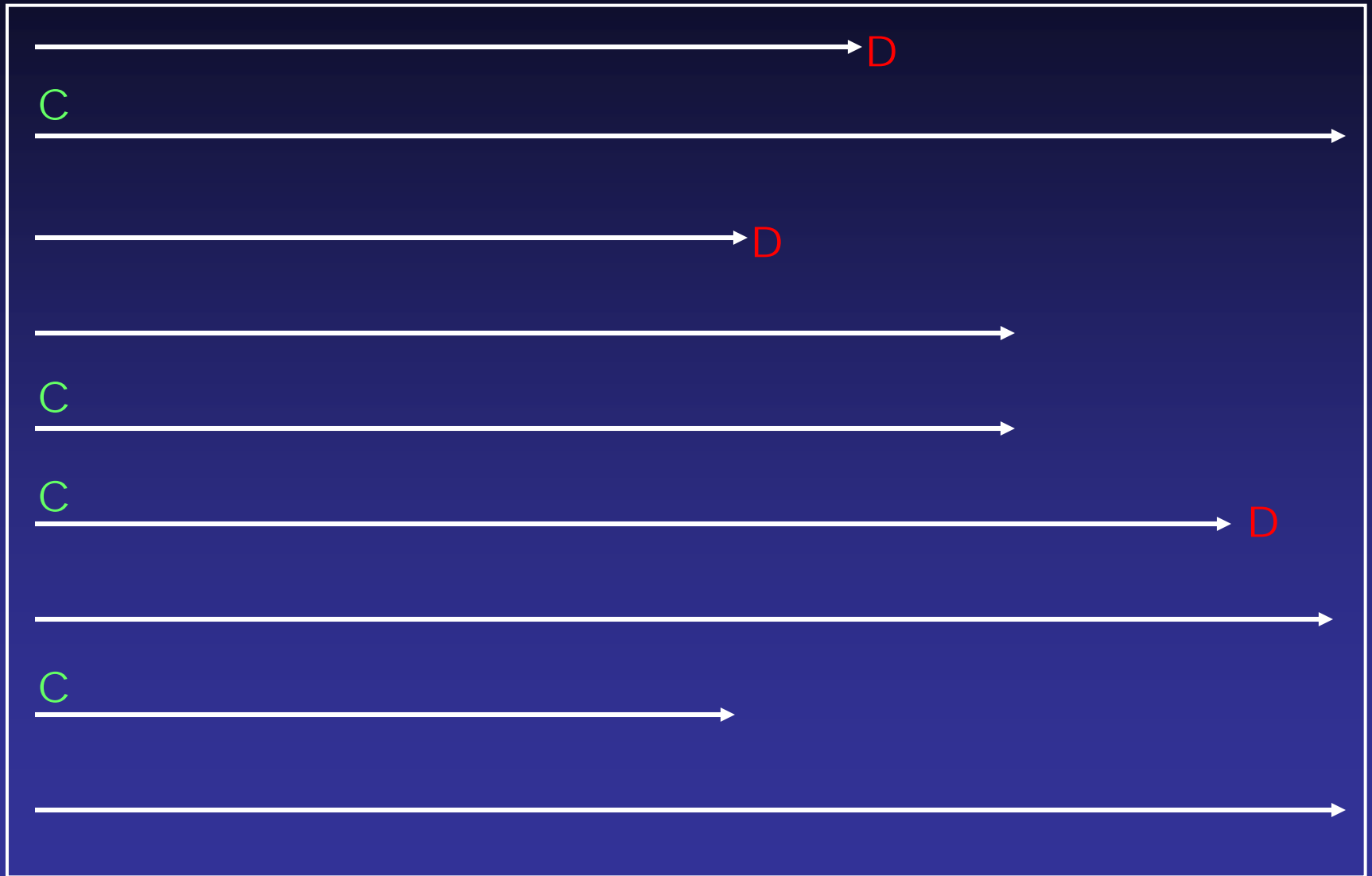
In addition it is typical to match on age and gender. Additional more complex matching is also common:

EPIC/GEN-AIR also matches on country and smoking status

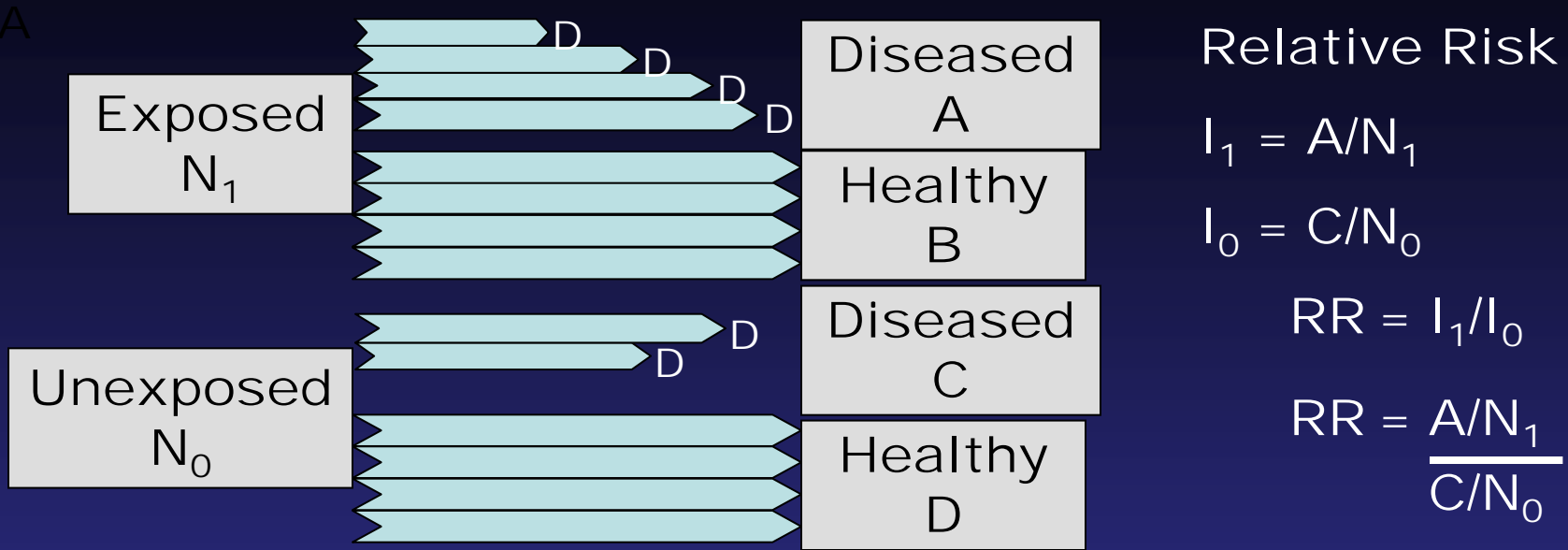
Tang et al., study in PHS also matched on smoking status and among current smokers matched on cigs/day

Also possibility of counter matching

Case-Cohort Study



Case-Cohort Study

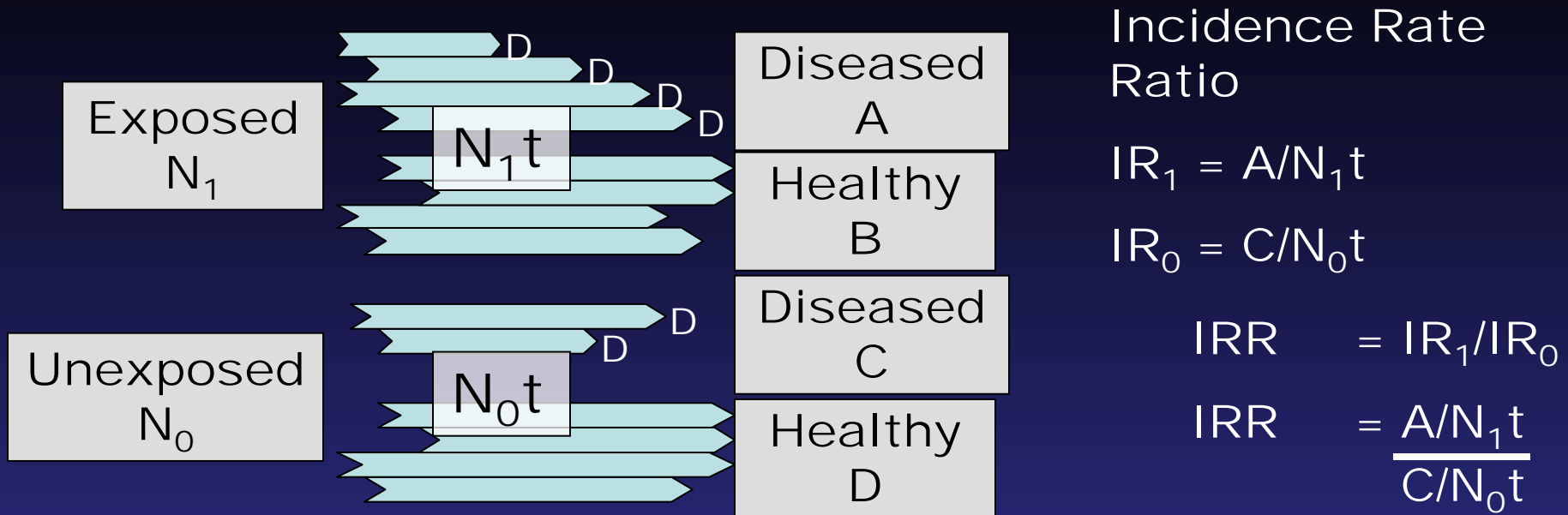


In a case-cohort study in a cohort with complete follow-up, all of the cases and a sample, r , of the cohort are used

	Cases	Cohort
Exposed	A	$N_1 (r)$
Unexposed	C	$N_0 (r)$

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

Case-Cohort Study



In a case-cohort study in a cohort that has variable follow-up, all of the cases are used and a sample, r , of the cohort is used to estimate the follow-up experience of the cohort

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

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

Nested Case Control Studies vs. Case-Cohort Studies

Nested case-control studies match on length of follow-up and can match on other factors as well

- allows for calculation of the IRR
- allows for efficient control for confounders

Case-cohort studies involve no matching and use a random sample of the cohort (sub-cohort) at baseline as the referent group

- the sub-cohort can be used as a referent group for future case series
- prevalence of exposure can be estimated and external comparisons can be made
- time scale is not fixed by design

Concerns with the Nested Case-Control Design

Due to the intricate matching,

- the control series is not intuitively understood and is difficult to work with
- controls are not representative of the cohort population
- the control series has few other uses, so the investment in biomarker analyses cannot be leveraged for other research

Growing interest in case-cohort analyses

Biomarkers Cause Logistical Problems with Executing a Case-Cohort Study

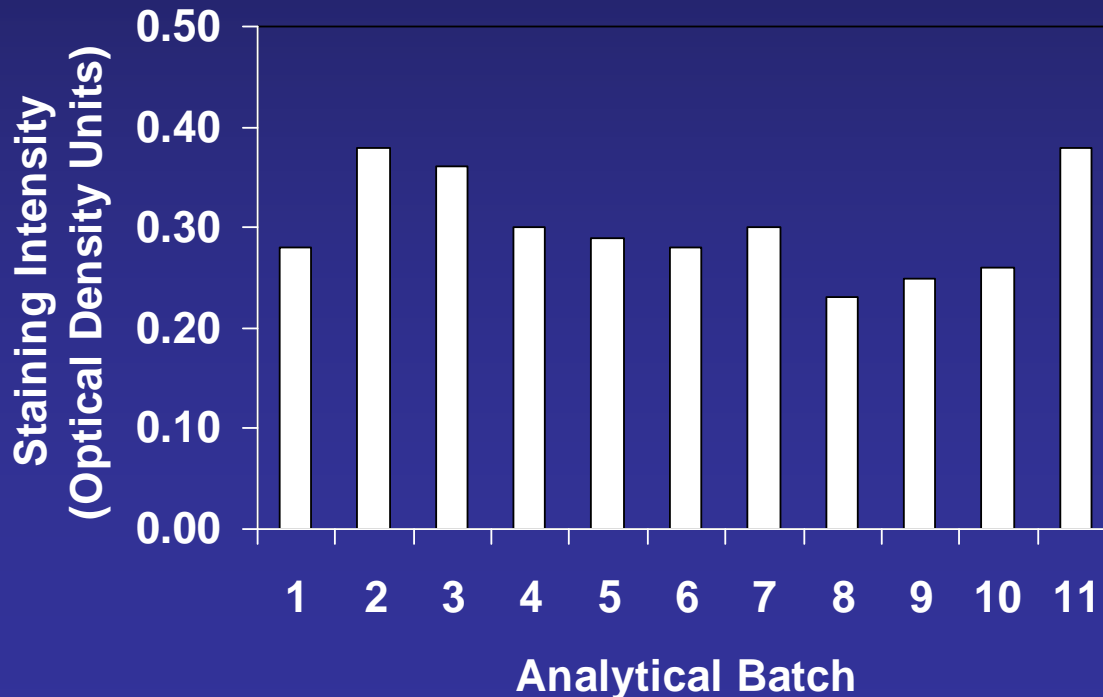
The case-cohort study relies on the assumption that exposure can be equally well measured in the sub-cohort as in the cases, and subsequent case series.

Yet three issues with biomarkers make this assumption questionable.

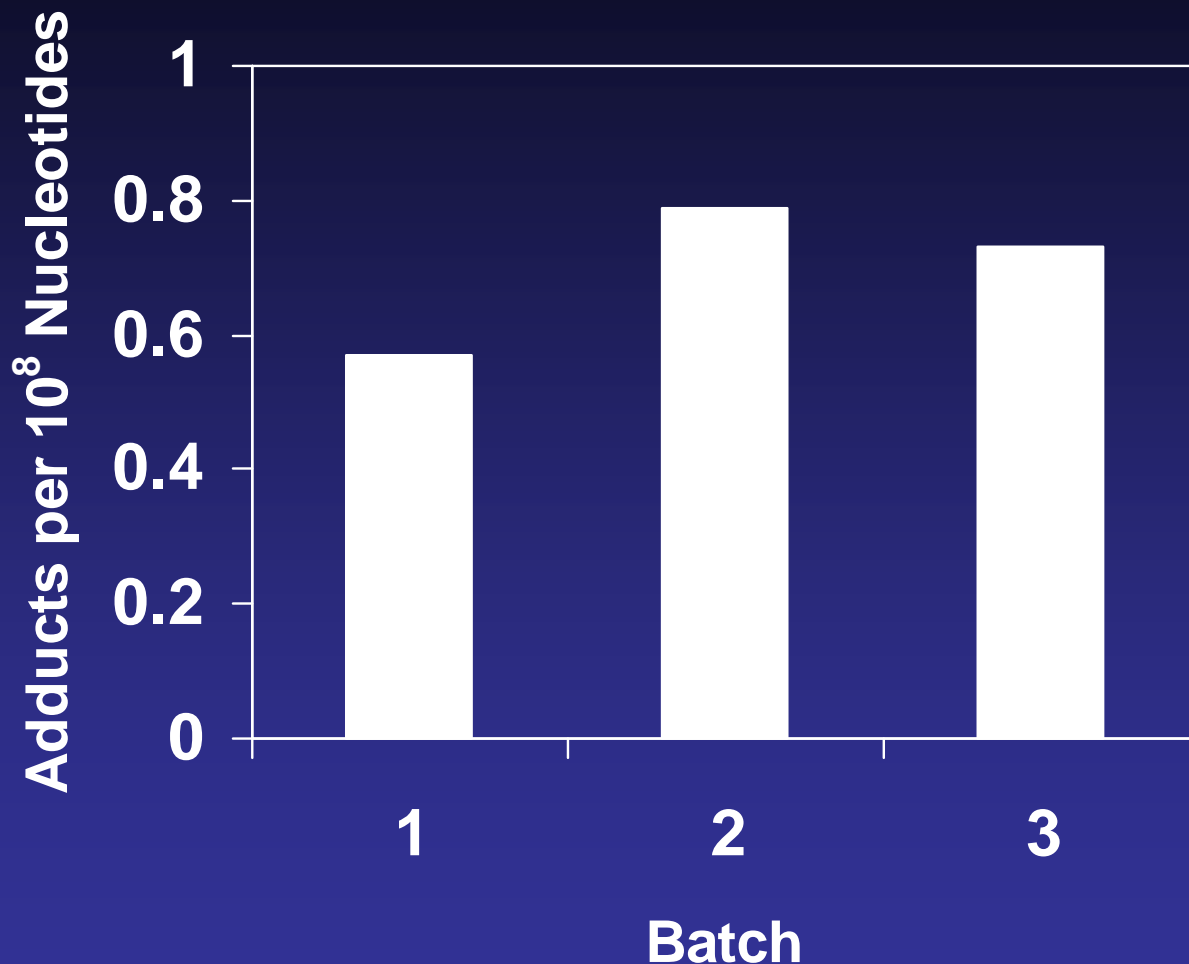
- Batch effects
- Storage effects
- Freeze-thaw cycles

Batch Effects

There are technological and staffing limits to how many samples can be analyzed in one go. So samples are run in batches or groups. The hope is that circumstances of the analyses do not vary by batch, that placing a sample in one particular batch versus another batch does not influence the analytical result. It is clear that for many biomarkers this is not true, there are substantial batch effects.



Batch Effects In EPIC/GEN-AIR



$P < 0.01$ for difference between batch 1 and 3 after control for gender, smoking (never vs. ex-smoker), country and EPIC center.

Storage Effects

Biological samples are typically stored at -70°C or lower. However, not all biomarker targets are stable at this temperature.

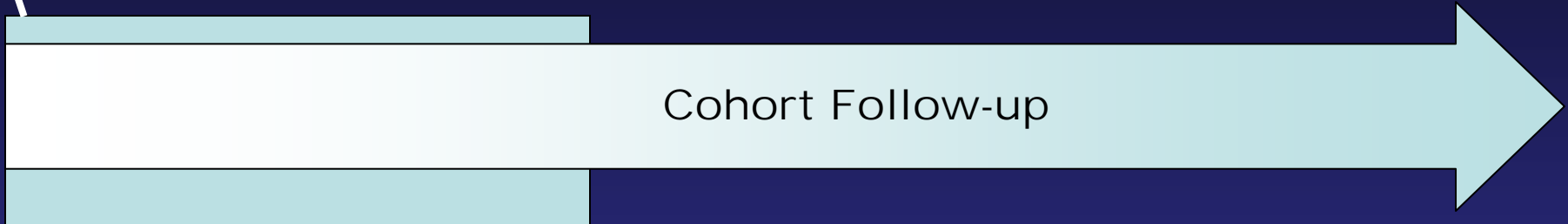
- Evidence that antioxidant micronutrients in serum, cotinine and B[a]P-DNA adducts are stable
- Evidence that serum cholesterol, free PSA, serum sex hormones, salivary Ab, and IH targets in tissue sections are not stable.

Freeze-Thaw Cycles

As biological samples freeze and thaw the pH and ionic balance of the liquid phase of the sample can be very different from the natural condition of the sample. Changes in pH and ionic balance can degrade biomarker targets.

- There is evidence that lipoprotein (a), antibodies, endogenous antioxidants, saliva cortisol, EGFR and DNA quality degrade during freeze-thaw cycles.

A Case-Cohort Study Conducted Prospectively



Period in which the work of a case-cohort study is performed.

- Selection of sub-cohort
- Identification of cases
- Lab analyses

But when do you start analyzing study samples?

A Case-Cohort Study Conducted Prospectively, cont

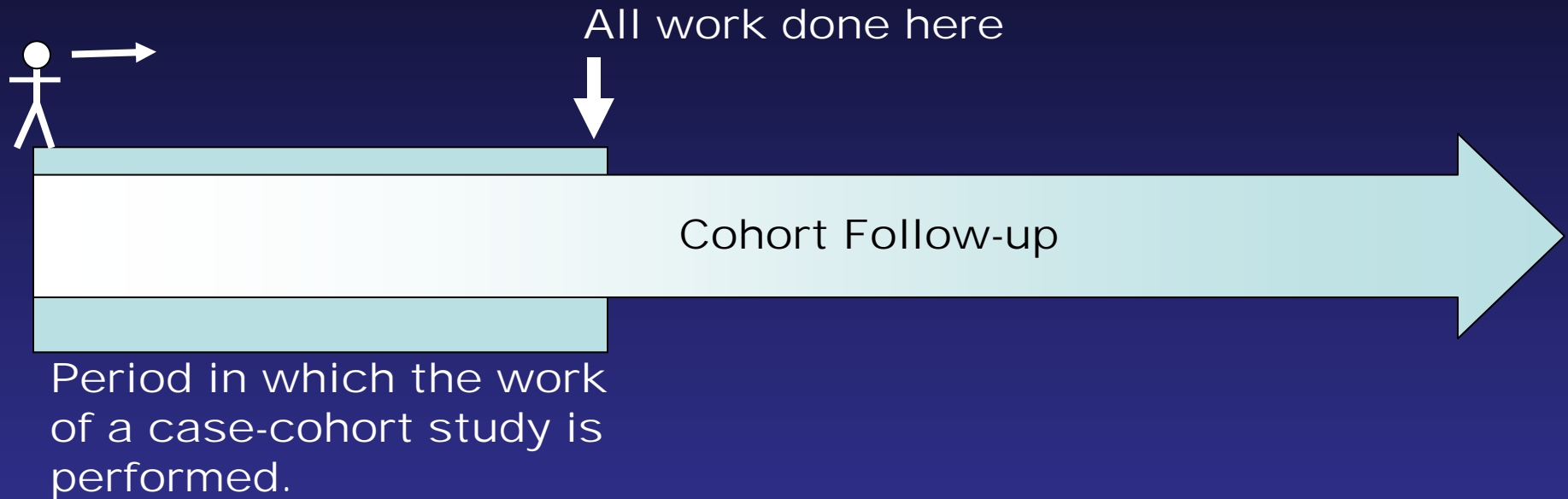
Incentives to analyze the sub-cohort samples right away.

- Spreads the work out over the study period, reduces pressure on the lab.
- Allows for cross-sectional analyses of determinants of biomarkers to begin.
- Can perform cross-sectional analyses of prevalent cases and sub-cohort.
- What else are you going to do with your time?

But most cases will accrue towards the end of the work period, and will not be subjects included in the sub-cohort.

- Case and control samples analyzed in different batches.
- Cases will tend to have longer storage durations.

A Case-Cohort Study Conducted Prospectively



- Selection of sub-cohort
- Identification of cases
- Lab analyses

A Case-Cohort Study Conducted Retrospectively



Cohort Follow-up

Period in which the work of a case-cohort study is performed.

- Retro. selection of sub-cohort
- Retro. identification of cases
- Lab analyses

A Case-Cohort Study Conducted Retrospectively, cont

- Since sub-cohort and cases identified retrospectively biological samples can be randomized to batches, removing bias.
- Depending on how long it took to assemble the cohort, storage duration could vary by several years among subjects.
- Certain samples may already have undergone freeze-thaw cycles.

Subsequent Case Series

An often cited strength of the case-cohort design is that the sub-cohort can be re-used as a referent group for future case series.



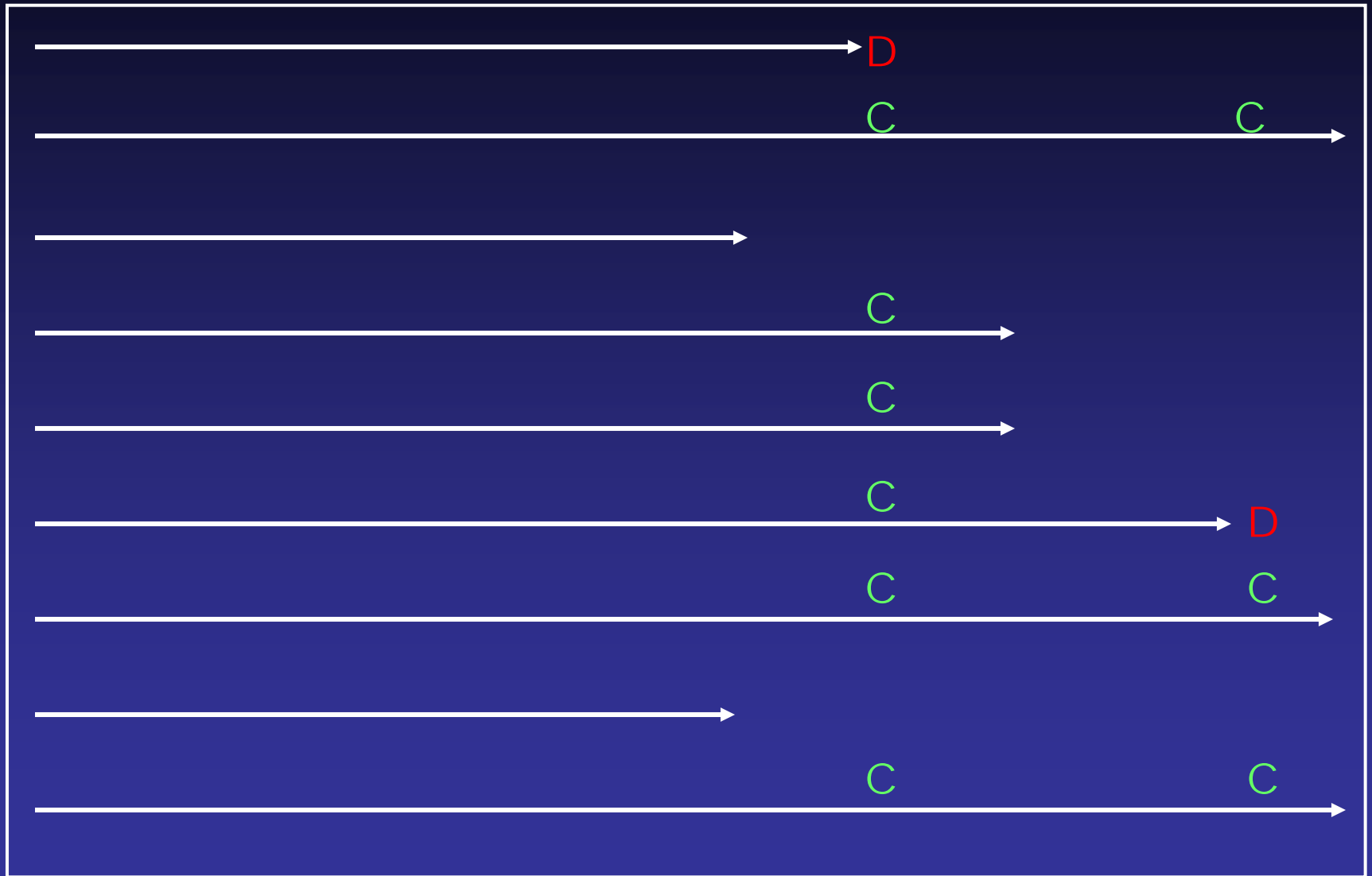
But for case-cohort analyses 2 all the case samples will be analyzed years after the samples from the sub-cohort were analyzed. Assuring batch effects and differences in storage duration.

Flexibility in Time Scale of Analyses

In a Case-Cohort study the time scale used in the analyses is not fixed by the design. Could use duration of follow-up, age, time since some event of interest, or any other scale.

- Changing the time scale will alter which subjects are included in risk sets and the calculation of person years
- Duration of follow-up may be appropriate
- Age may be appropriate for genetic risks
- Time since onset of exposure might be more appropriate for other exposures

Nested Case-Control Study

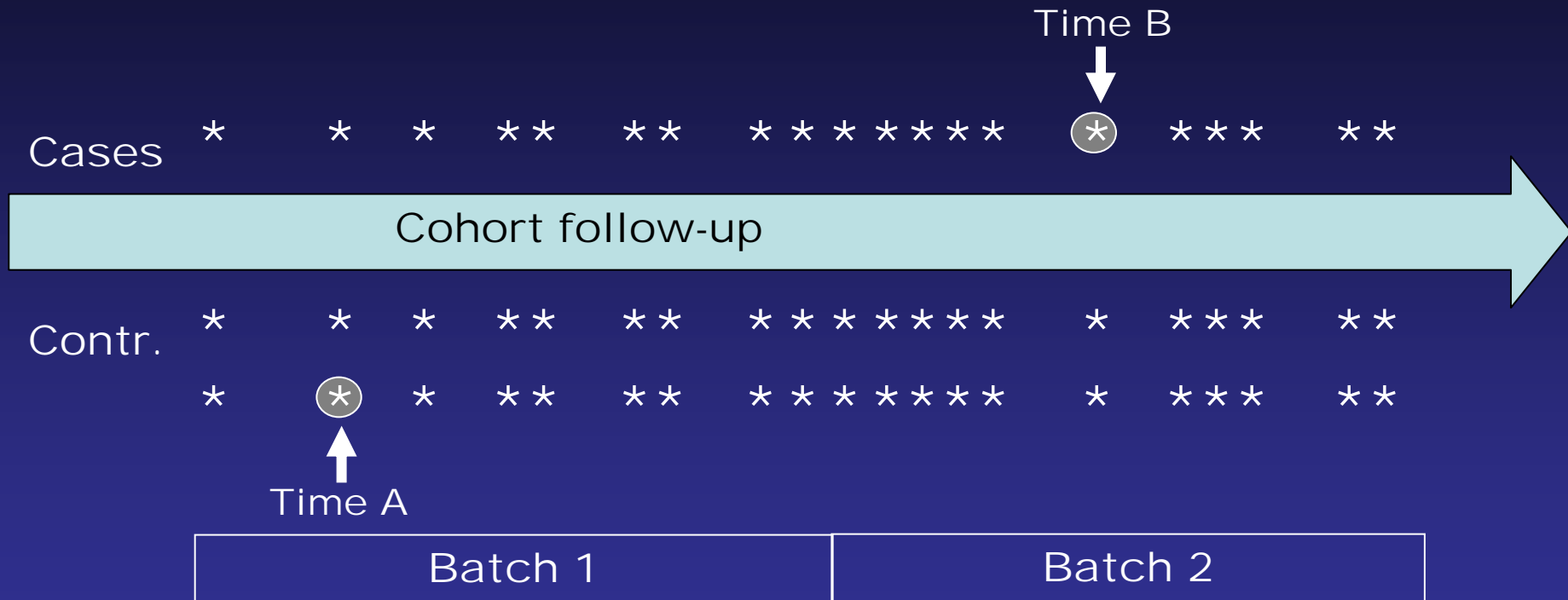


Nested Case-Control Studies

- Matching on length of follow-up, typically means that the cases and controls are matched on sample storage duration.
- Because cases and controls are identified simultaneously, samples can easily be matched on batch.
- It is also possible to match on number of freeze thaw cycles.

But, complexities arise when a subject appears multiple times in a data set and matching must be respected in all analyses.

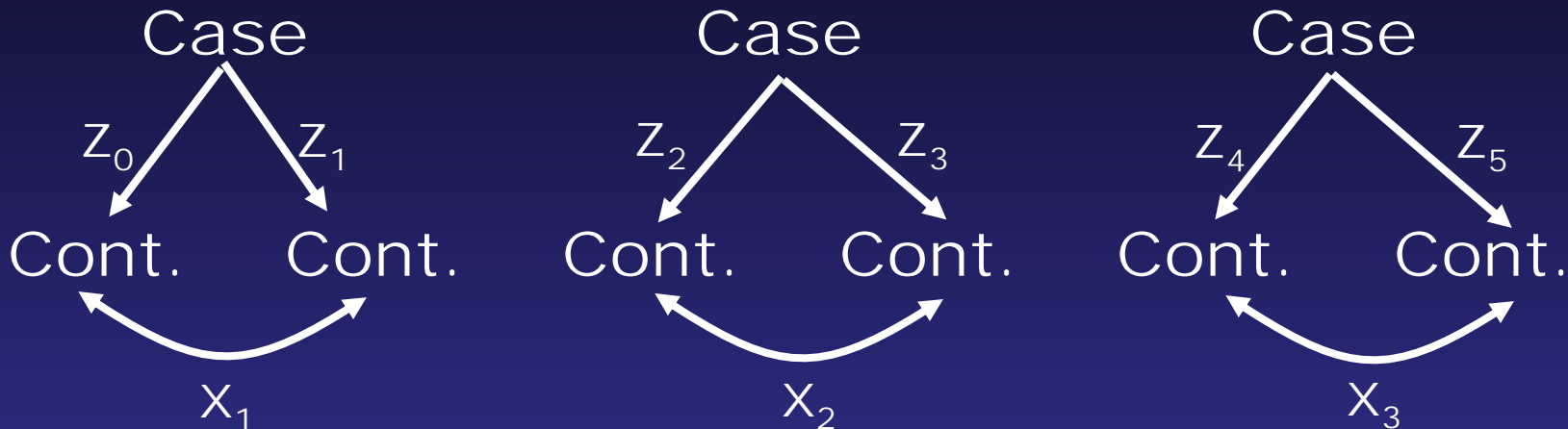
Schematic of a Nested Case-Control Study



A subject is a control at time A and their sample is assayed in batch 1. Then the subject develops disease at time B and is a case, a new aliquot of sample must assayed in batch 2.

Cross-sectional Analyses in Controls

It may be of interest to assess determinants of biomarker levels in controls (Perera, 2002; Sun 2002)



Controls are not independent, they are matched to cases

- will have the same gender and age,
- will have correlated biomarker levels,
- will be similar for all correlates of matching variables,
- in addition they are highly selected.

Xs and Zs are correlation coefficients

Bias and Precision in Cross-Sectional Analyses, the Effect of Matching

Analyses of BMI in EPIC and controls from EPIC/GEN-AIR. Within matched pairs of controls BMI is correlated ($r = 0.17$, $P=0.01$).

	EPIC ¹ Beta	GEN-AIR ¹ Beta, SE	GEN-AIR ² Beta, SE
Gender	-1.06	-0.46, 0.1497	-0.46, 0.1639
Age	0.08	0.03, 0.0082	0.03, 0.0098

1 Linear Regression

2 GEE

Counter Matching

In a nested case-control study only discordant matched pairs contribute to the odds ratio leading to a loss of statistical efficiency.

Counter matching on a correlate/proxy for exposure seeks to maximize the number of pairs discordant on exposure or biomarker, increasing statistical efficiency.

- Matching seeks to increase the efficiency of control for confounding.
- Counter matching seeks to maximize variation in the biomarker within case-control sets.

Counter Matching

Imagine an occupational cohort in which you would like to nest a case-control study to determine whether an exposure related biomarker is associated with disease outcome.

There is data on duration of occupational exposure.

1. Use the distribution in the cases to dichotomize subjects into high and low duration of exposure.
2. Select a control from the exposure category that is counter to the category of the case.
3. Conduct case-control analyses on the biomarker, with a weighting factor for the sampling probability of each subject.

Counter Matching

Case-Control Selection

- Counter matched on duration of exposure.

	Case	Control
Long Duration	S_{case1}	S_{control2}
Short Duration	S_{case2}	S_{control1}

Case-Control Analysis

- Calculate an OR for the biomarker with sampling probability weighting

	Case	Control
Biomarker +	A	B_w
Biomarker -	C	D_w

Counter Matching, cont.

- Produces increased statistical efficiency over sampling that is random with respect to occupational exposure.
- Has been extended to studies of gene-environment interaction.

See for more details:

Cologne J, J. Epidemiol, 2003 and Bernstein JL, Breast Cancer Res. 2004

Studies with Target Tissue Analyses

- All past nested case-control and case-cohort studies using biomarkers have used surrogate tissues (e.g. blood or urine). For non-genotype based biomarkers causal extrapolations are made to target tissue biomarker levels.
- We are conducting a study of prostate cancer using biomarker analyses in target tissue.
 - 5,197 men with available surgical specimens for benign prostate conditions, 1990-2002
 - 800 cases of prostate ca. expected by 2010
 - Analyses of carcinogen-DNA adducts and DNA promoter methylation in benign specimens

Conclusions

For biomarkers not affected by batch, storage, and freeze-thaw cycles (e.g. genotypes) use the case-cohort design.

- Can estimate the risk ratio or the rate ratio.
- The sub-cohort can be used as a referent for multiple case-series.
- Simple random sample allows for valid cross-sectional analyses and external comparisons.

The case-cohort design best leverages the investment in biomarker analyses.

Conclusions

For biomarkers affected by batch, storage or freeze-thaw cycles the nested case-control appears to offer the best approach to control bias.

- Matching allows for efficient control for confounders.
- Counter matching allows for efficient analyses of effect modification.
- Common diseases will have more subjects represented multiple times in the study.
- The controls have few other uses.
- Must respect the matching, much more thought is needed.