Cohort and Case-control Approaches

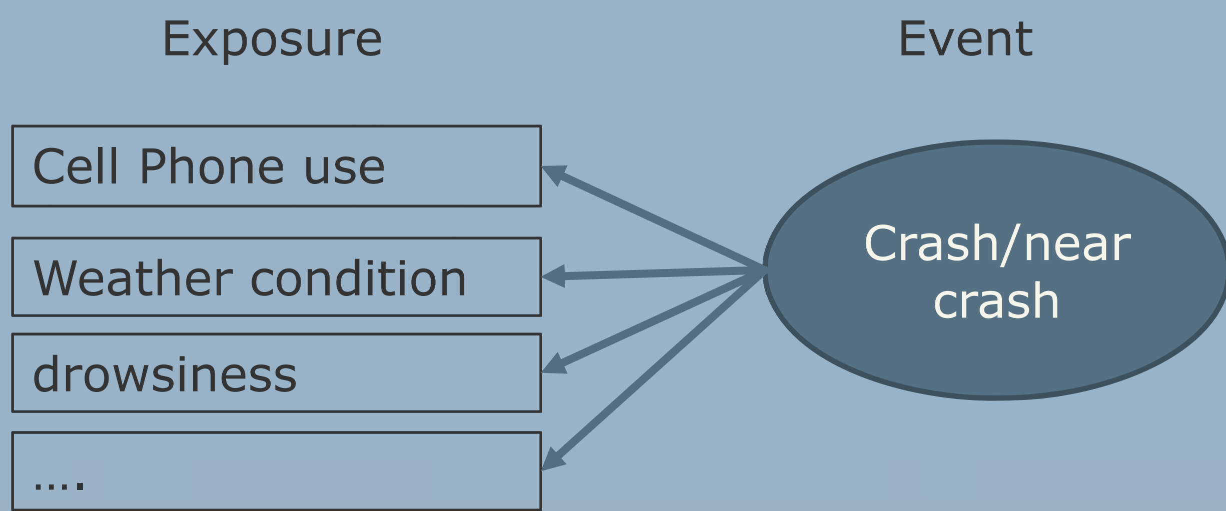
Feng Guo Ph.D.
Department of Statistics
VTTI, CASR
Virginia Tech

First Human Factors Symposium: Naturalistic Driving Methods & Analyses
August 26th 2008

Driving Transportation with Technology

Virginia Tech Transportation Institute

Framework



Hypothetical examples:

Example 1

In a naturalistic study, it is found that in 95 out of 100 crashes observed, the driver was listening to music. Can we conclude that listening to music contributes to crashes?

Evaluating risk factors

Example 2

If it is found in 10 crashes, the driver fallen in sleep for more than 6 seconds. Can we conclude that drowsiness/fatigue contributes to crashes?

Have to compare with "Normal" (Baseline) conditions!

- 95% of the times people are listening to music when driving : listening to music is unlikely a risky behavior.

- Essentially nobody sleep when driving: Sleeping during driving is dangerous.

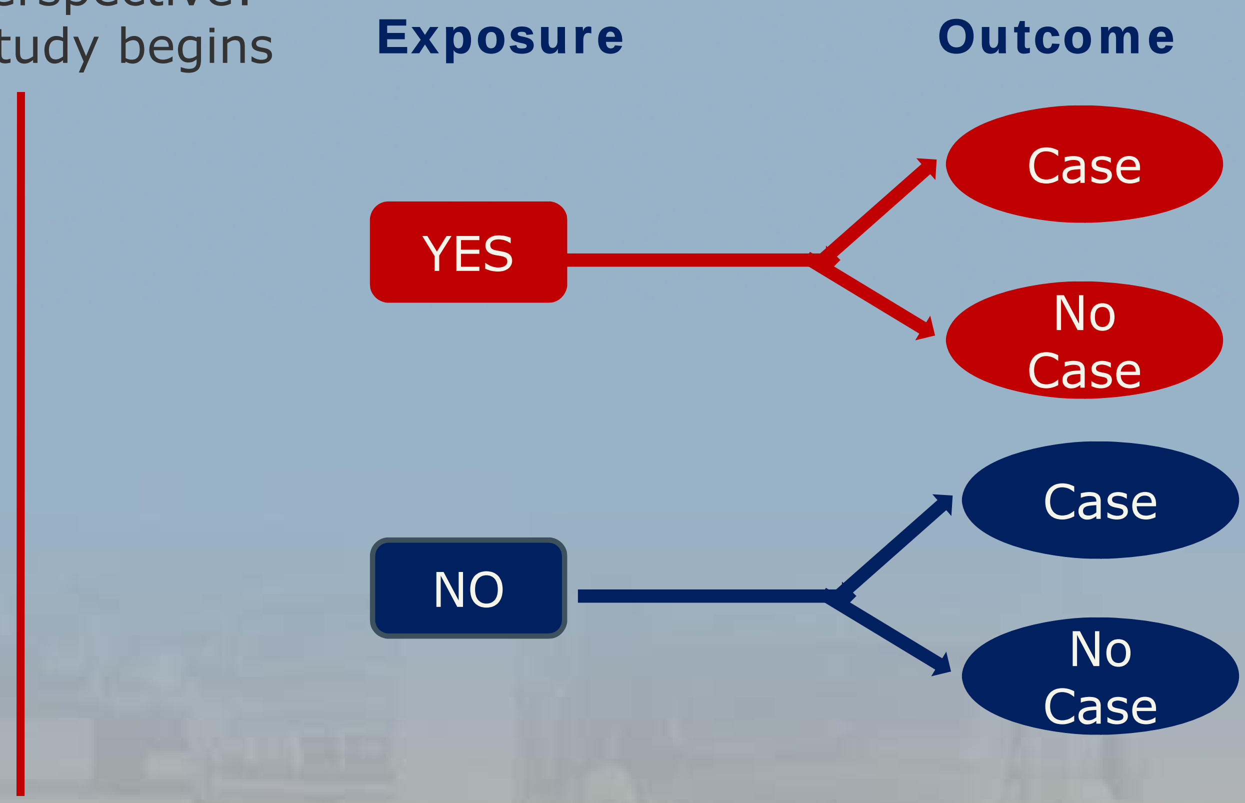
How to get exposure information on normal and event situations?: study design

- Cohort
- Case-control
- Case-cohort
- Case-crossover

- **Major issue: how to reduce bias.**
- **Analysis / modeling is directly related to study design!**

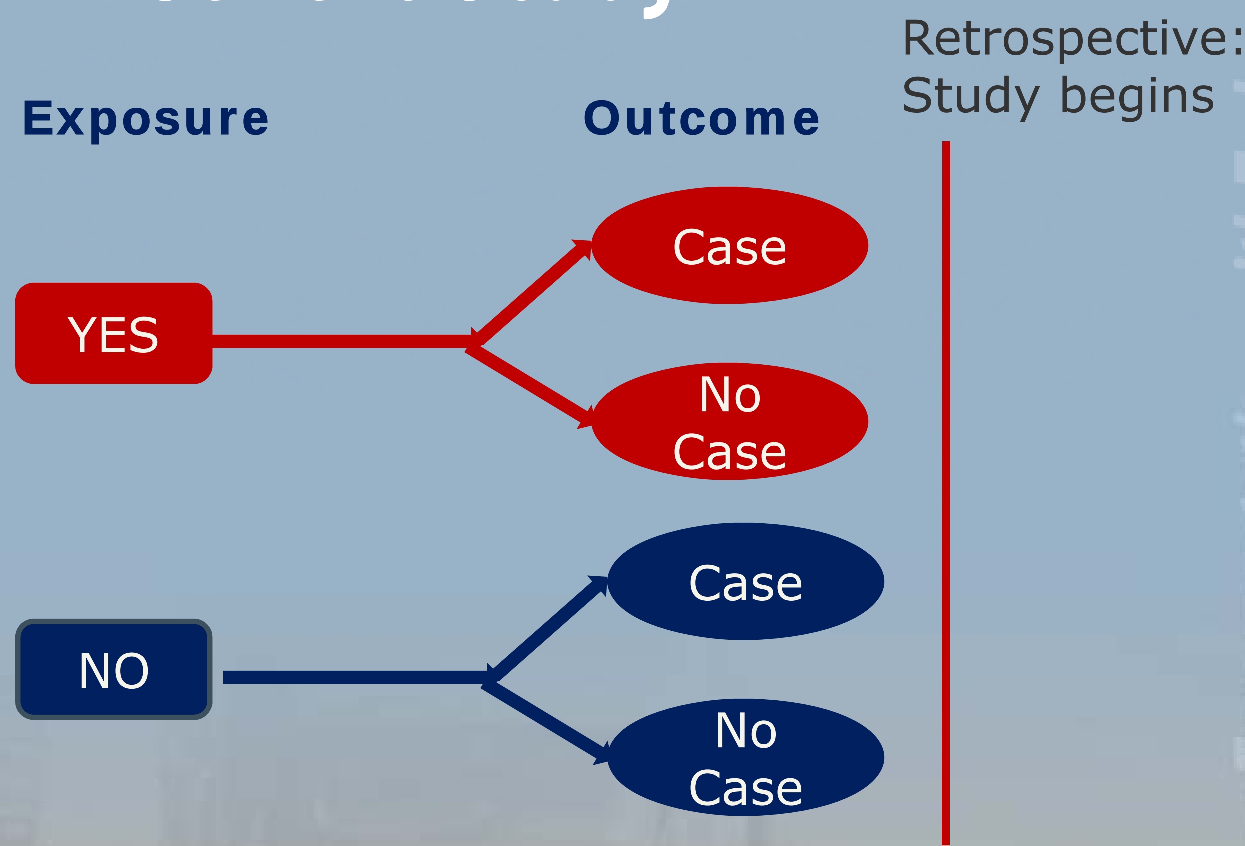
Cohort Study

Perspective:
Study begins



Forward direction: from exposure to case

Cohort Study



Forward direction: from exposure to case

Cohort Study

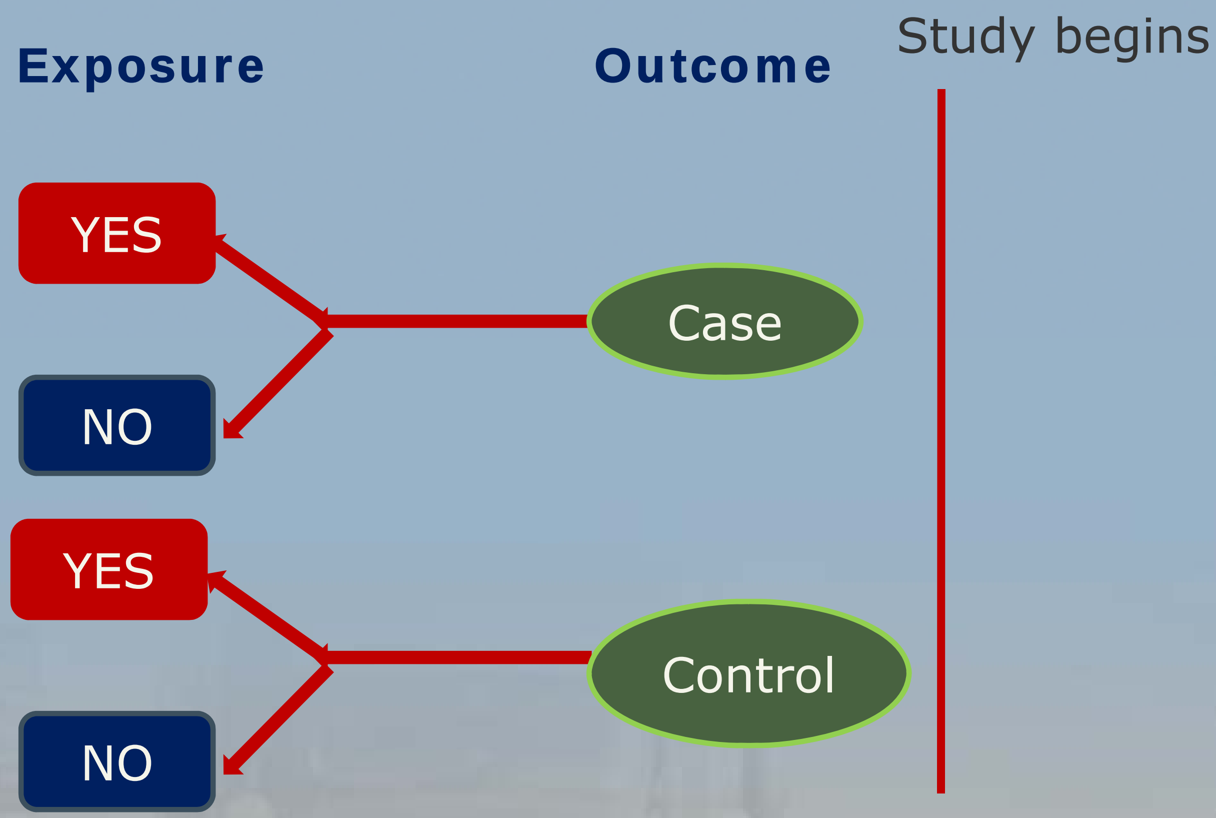
Pros:

- Least prone to bias
  - Relative to other observational study designs
- Can address several diseases in same study
- Retrospective can be relatively low cost and quick
  - Frequently used in occupational studies

Cons:

- Loss to follow-up is potential source of bias
- Prospective cohort study
  - Quite costly and time consuming
  - May not find enough cases if disease is rare

Case-control Study



Backward direction: from outcome to exposure
Backward timing: study begins after outcome

Case-Control Study

Pros:

- Less expensive and time-consuming
- Optimal for rare diseases
  - Subjects selected based on disease status
- Allows several exposures to be evaluated
  - Multiple etiologic factors for a single disease

Case-Control Study

Cons:

- More susceptible to selection bias (than cohort studies)
  - Presence or absence of exposure may influence selection of disease and non-disease groups
- More susceptible to information bias
  - Observer bias
  - Recall bias

- Does not allow direct estimation of risk
  - Not possible to calculate rate of development of disease given exposure status
- Does not allow several diseases to be evaluated
- Generally not feasible for rare exposures

Hybrid Design

- Mixture of cohort, case-control, crossover, and cross-sectional design

- Case-cohort
- Case-crossover

Case-Cohort Study

Study begins

Exposure | **Outcome**

YES

NO

YES

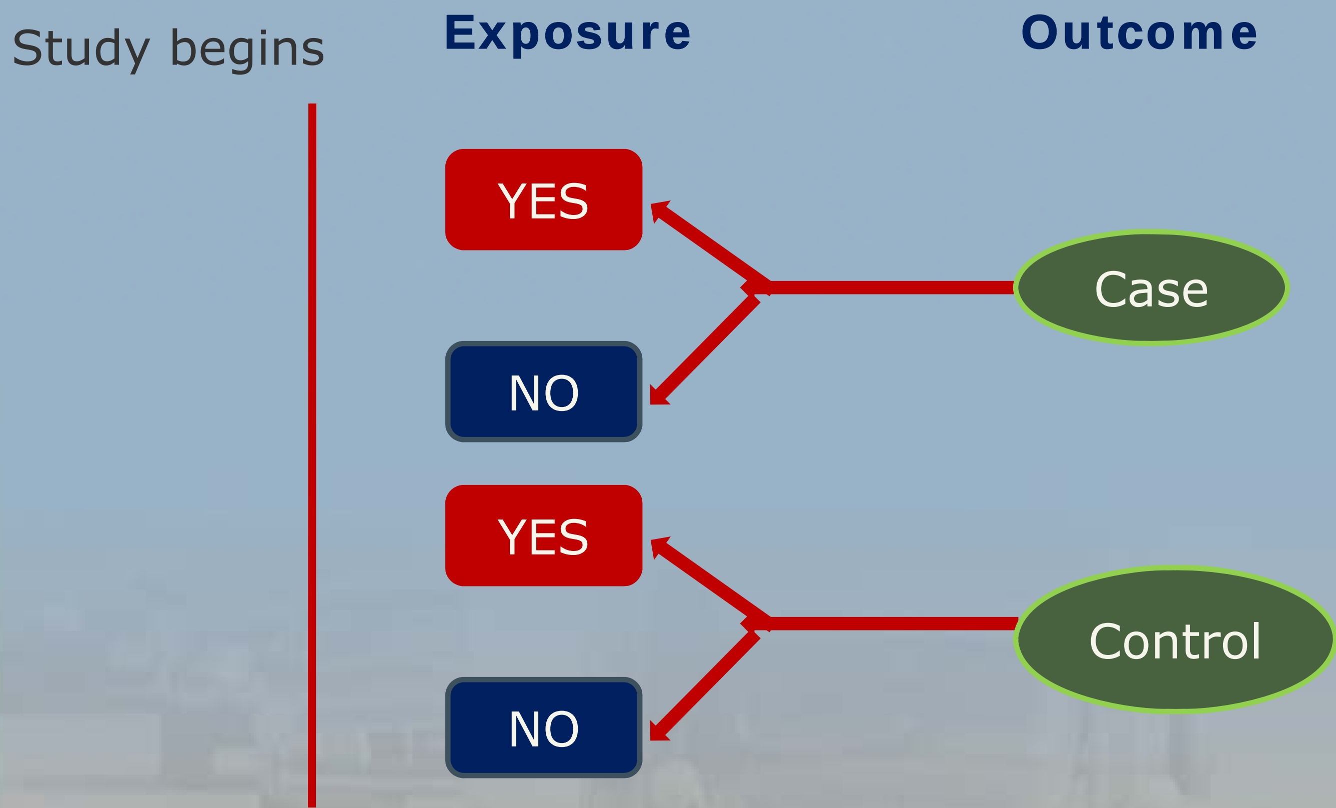
NO

Case

Control

Backward direction: from outcome to exposure
Forward timing: study begins BEFORE outcome

Case-Cohort Study



Backward direction: from outcome to exposure
Forward timing: study begins BEFORE outcome

Case-Cohort Study

- Several diseases can be studied
  - In contrast to case-control study
- Less costly and more efficient than cohort study
  - Smaller number of non-cases
- More prone to measurement error than cohort study
  - Exposure status determined after cases and control
  - Unless exposure status at initial cohort enrollment
- Can be more expensive and time-consuming than case-control study
  - Requires identifying original cohort

Odds Ratio in Different Study Designs

- Case-control Studies: exposure odds ratio
- Cohort studies: risk odds ratio (ROR)

Compare Cohort and Case-control study

FIXED

- Cohort Study

	E+	E-
D+	A	C
D-	B	D
total	N+	N-

P(D+|E+)= (A/N+)
P(D+|E-)= (C/N-)

Risk Ratio (RR)=
P(D+|E+)/ P(D+|E-)

- Case-Control Study

	E+	E-	total
Case	a	c	M1
Control	b	d	M0

P(E+|Case)=a/M1
P(E+|Control)=b/M0

$$Odds(E+|Case) = \frac{P(E+|Case)}{1-P(E+|Case)} = \frac{a/M1}{c/M1} = \frac{a}{c}$$

OR = (a/c) / (b/d)= (ad/bc)

Although conceptually very different, the formulas for Risk OR and Exposure OR are the same: AD/BC

Odd Ratio Approximation of Risk Ratio: Case-Control Studies

In case-control studies, the exposure odds ratio (EOR) approximates the risk ratio when the following 3 conditions are satisfied:

- 1. The rare disease assumption holds
- 2. The choice of controls in the case-control study must be representative of the source population from which the case developed.
- 3. The cases must be incident cases

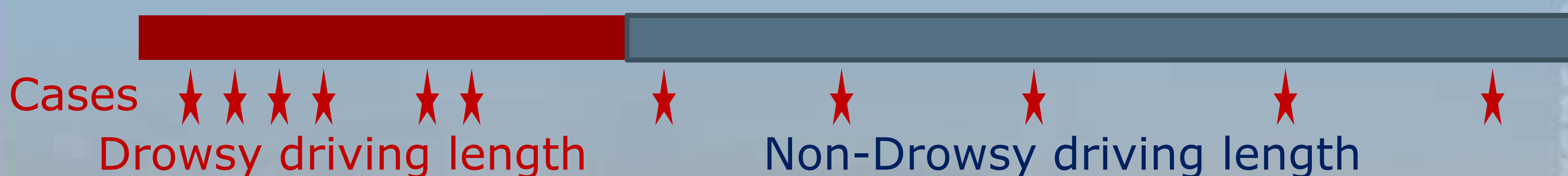
Driving Transportation with Technology

Virginia Tech Transportation Institute

Risk Rate (time variant exposure)

$$\text{Rate1}: \frac{\#\ \text{of Event under drowsiness}}{\text{Miles (time) traveled under drowsiness}}$$

$$\text{Rate2}: \frac{\#\ \text{of Event under NO drowsiness}}{\text{Miles (time) traveled under NO drowsiness}}$$



If Rate1 is significantly greater than Rate2, we considered drowsiness is a risk factor for safety.

Problem: How can we know miles/time traveled under drowsiness?

Odds Ratio Approximation to Rate Ratio

- Cohort Study

	E+	E-
Dis+	A	C
total	PT+	PT-

- IDR = (A/PT+) / (C/PT-)
  = (A/C) / (PT+/PT-)

- Case-Control Study

	E+	E-	total
Case	a	c	M1
Control	b	d	M0

- OR = (a/c) / (b/d)
  ≈ (a/c) / (PT+ / PT-)
  = IDR

Assumptions:

1. M0 subjects are randomly selected via source population
2. Their exposure odds (b/d) similar to that in source population (PT+/PT-).
3. Steady state

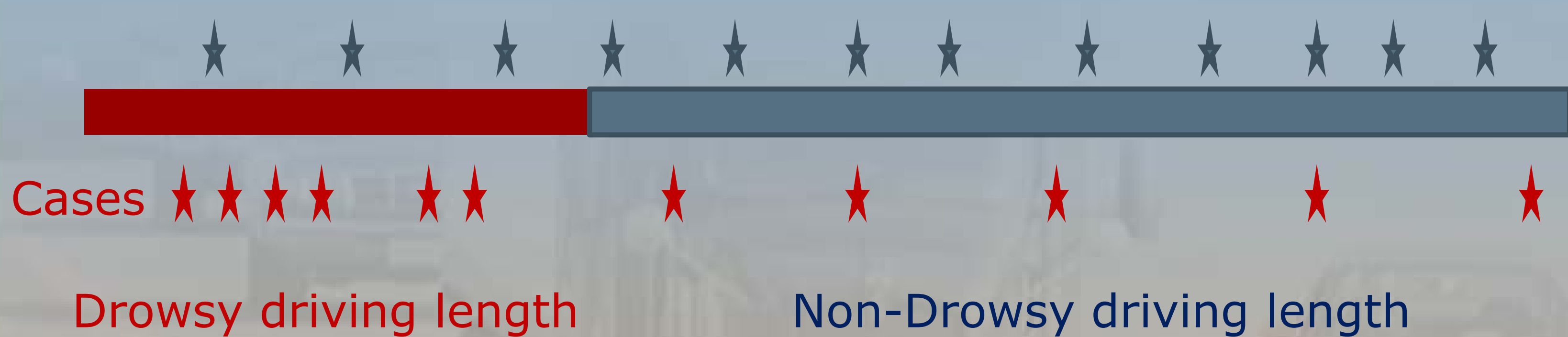
Examples of VTTI research

- Modeling 100 car (STSCE):
  - Random sampling case-cohort design: non-matched design
  - Confounding/interaction factors controlled through modeling
  - Incorporate driver specific correlation through models
- Case-crossover design (NHTSA)
  - Case-crossover sampling: matched design
  - Part of confounding/interaction factors controlled through baseline sampling

Modeling 100 car: Baseline sampling

Principle: ideal control group is representative of the *source population* from which the cases are derived

1. Time variant exposures: risk rate
2. Sampling should reflect the odds ratio to risk rate principles
3. Random sampling stratified by vehicle was adopted



Challenges in analyzing naturalistic study

- Control for confounding and interaction factors.
- Multiple events for same participant: driver specific correlations!

Analysis Options

- Stratified analysis
  - Categorize control variables and form combinations of categories or strata
  - Drawback of running out of numbers when the number of strata is large
- Mathematical modeling
  - Use a mathematical expression for predicting the outcome from the exposure and the control variables
  - Considerations on choice of model and variables to include in initial and final model

Basic Model Setup

- Generalized linear model (GLM) framework
- Baseline Multinomial model
  - Contrast crash, near-crash, and critical incident with base-line separately in a same model
  - The odds ratio is adjusted with respect to other variables in the model

$$y_i \sim Multinomial(1, \mathbf{p})$$

y is a categorical variable corresponding to the events and baseline

$$\log\left(\frac{p_r}{p_0}\right) = \mathbf{X}\ \boldsymbol{\beta}_r$$

Where p_r is the probability of in r^{th} event
p_0 is the probability of baseline
X is the covariates matrix
βr is the vector of parameters for r^{th} event,
it has a direct relationship with odds ratio.

Incorporate driver-specific correlation

Independent assumption for the basic model
One driver have multiple event (baseline)
They should be correlated: good driver, bad driver.

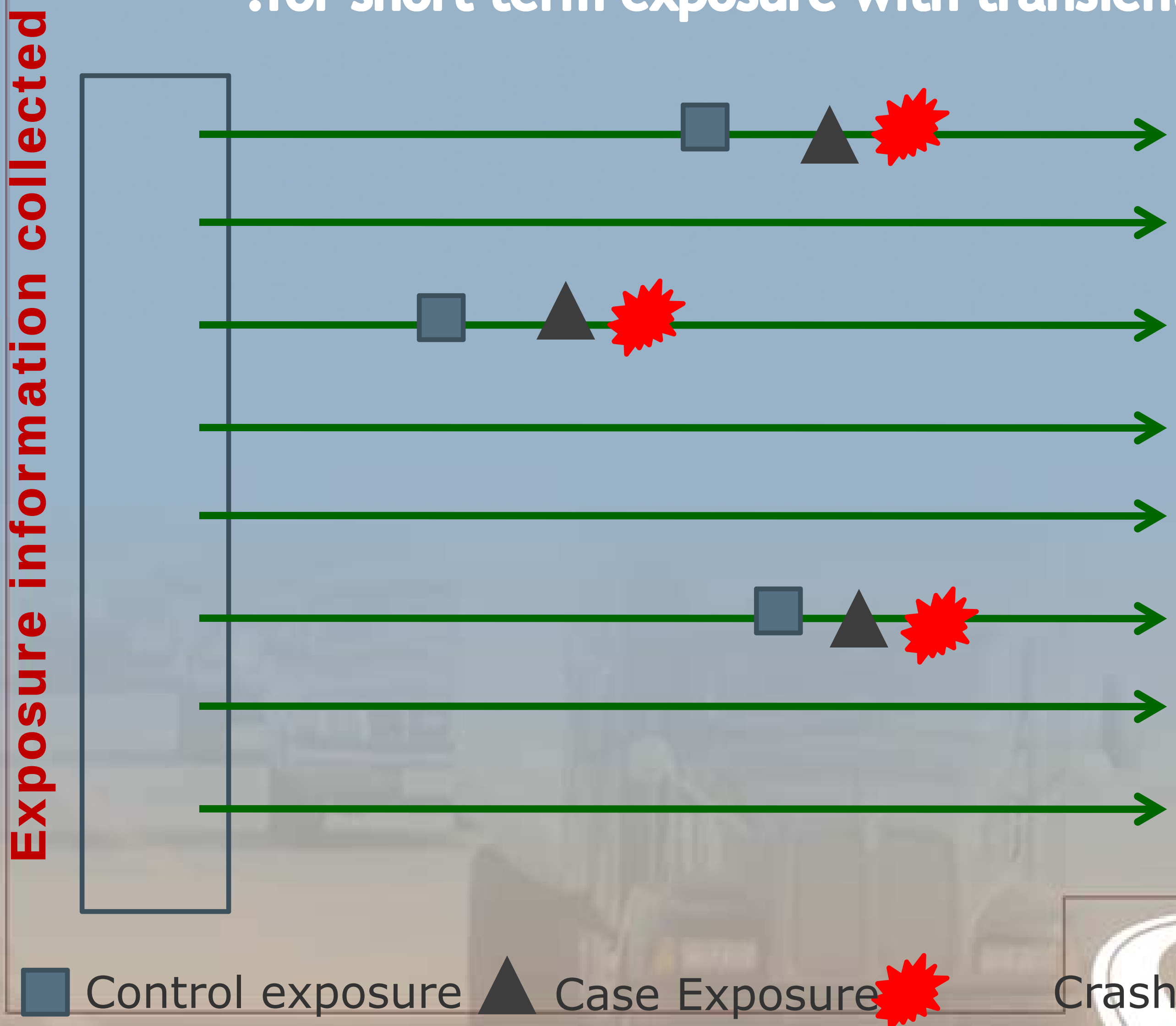
- Random effect model
  - Extension of the basic model

$$\log\left(\frac{p_{ijr}}{p_{ij0}}\right) = \mathbf{X}_{ij}\boldsymbol{\beta}_r + \mathbf{Z}_{ij}\boldsymbol{\alpha}_i$$

  - $\boldsymbol{\alpha}_i$ is the driver specific random effect
- Generalized Estimation Equation (GEE) model
  - Commonly used in longitudinal data analysis
  - Quasi-likelihood based method

Case-Crossover Design

:for short term exposure with transient effect



Sample exposure immediate before crashes

Sample exposure for time interval some period before crash

Case-Crossover compare to study 1

Pros:

1. Less prone to biased
2. More efficient in evaluating the effects of transient exposure factors

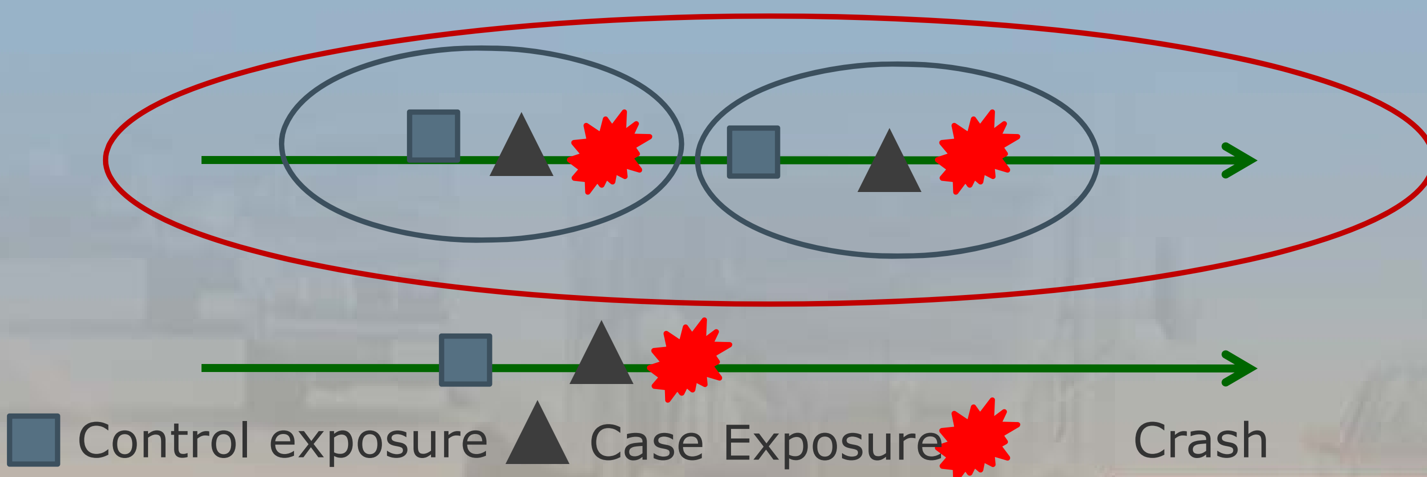
Cons:

1. Cannot be used to evaluate time-invariant effect such as age and gender.
2. Bring another level of correlation into the model

Case-Crossover Analysis

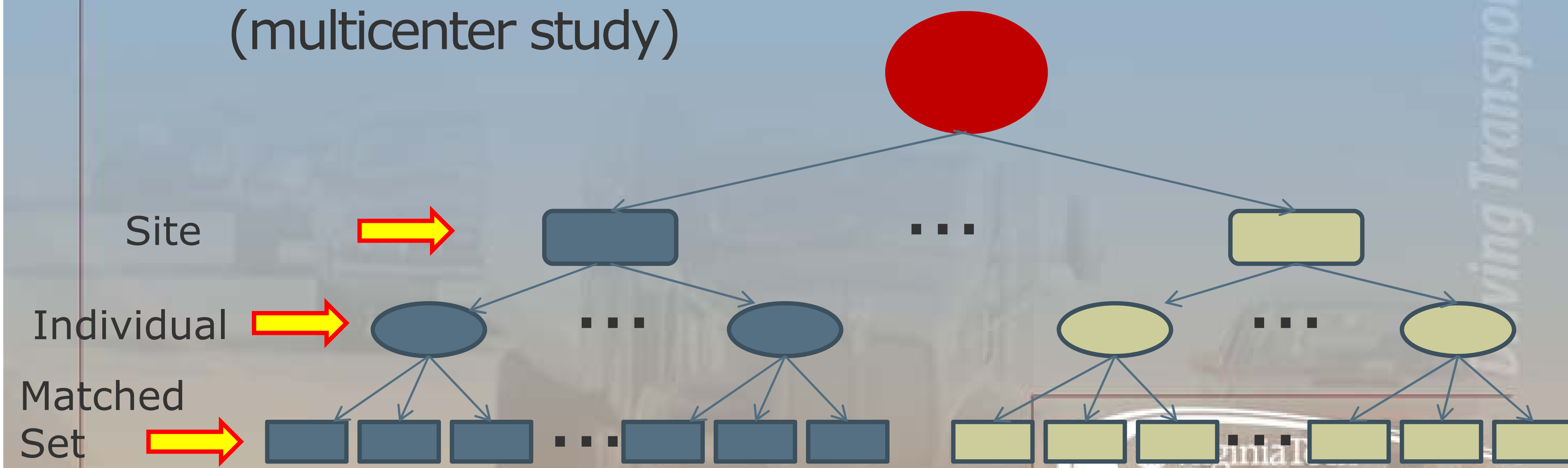
- Matched set correlation

- Driver specific correlation



Case-Crossover Analysis

- Nested random effects model
- Conditional logistic regression model
- Bayesian hierarchical model
  - Fit the context naturally
  - Easy to expend to accommodate more levels (multicenter study)



Bayesian Model

Model setup

$$Y_{ijk} \sim Bernoulli(p_{ijk})$$

$$\text{logit}(p_{ijk}) = \mathbf{X}_{ijk}\boldsymbol{\beta} + \mathbf{Z}_{ijk}\boldsymbol{\alpha}$$

Site i,
individual j,
event k

Prior:

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\beta})$$

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\alpha})$$

Vague: fixed large variance
Informative: prior elicitation
- from previous study
- From expert opinion

Summary

- Appropriate baseline sampling scheme is critical part of analyses.
- Analysis models should reflect the corresponding sampling scheme.
- Considering analysis at the beginning of the study!

Driving Transportation with Technology

Virginia Tech Transportation Institute

- Questions?
- …
- Thanks!