Estimating Crash Risk Using Naturalistic Driving Study Data

Feng Guo, Ph.D.
Virginia Tech Transportation Institute
Department of Statistics
Virginia Tech

September 1\textsuperscript{st} 2010
The Second International Naturalistic Driving Symposium
Estimating Crash Risk: Outline

• Overview

• Case-cohort and case-crossover approaches: results and lessons learned

• Near-crashes as crash surrogates for risk assessment purpose
# Naturalistic Driving versus Crash Database

<table>
<thead>
<tr>
<th></th>
<th>Crash Database</th>
<th>Naturalistic driving data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample population</strong></td>
<td>All drivers: limited selection bias</td>
<td>Participants only: selection bias</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>Police report: small proportion of actual accidents</td>
<td>Data collection system: all safety events</td>
</tr>
<tr>
<td><strong>Information source</strong></td>
<td>Driver/witness statements, retrospective: information bias</td>
<td>High resolution video and instrument recording</td>
</tr>
<tr>
<td><strong>Driver behavior</strong></td>
<td>Limited/unreliable information</td>
<td>Accurate/detailed information through data reduction</td>
</tr>
</tbody>
</table>
Risk Assessment

• Presence of a factor at crash ≠ Risk

• Comparing exposure status for safety events and for normal driving conditions.

• Naturalistic driving data provides detailed and accurate exposure information
Modeling Crash Likelihood Framework

Data collection

Identify Safety Events

Baseline sampling

Data reduction (event/baseline)

Statistical Modeling

Study Design
Modeling 100-Car: Case-Cohort Approach

- Sample short (6 second) epochs from the videos
- Sampling Scheme: Random sampling stratified by vehicle travel time
- Independent of crash/near-crash
Case-Crossover

Exposure information collected

Sample exposure immediate before crashes

Sample exposure for time interval some period before crash

Control exposure  ▲ Case Exposure  ● Crash
Matched Factors

• Driver ID
• Day of week (weekday versus weekend)
• Time of day (± 2 hours)
• Same GPS Location (± 100 Meters OR match to relation to junction)
• Must occur prior to crash/near-crash occurrence.

• Goal: 15 baselines for every crash/near-crash event.
Differences between the Case-Crossover and Case-Control
Individual Variation: Good Driver, Bad Driver

Teen driver crash/near crash rate

Number of safety events by subject (100-Car)
Case-Cohort: Generalized Mixed Effect Model

Model specification

\[ y_i = \begin{cases} 
1 & \text{Crash} \\
0 & \text{No Crash} 
\end{cases} \]

\[ y_i \sim \text{Binomial}(1, p_i) \]

\[ \log\left( \frac{p_{ij}}{1 - p_{ij}} \right) = X_{ij}\beta + Z_{ij}\alpha_i \]

where \( p_i \) is the probability of crash for \( i \)th observation

\( X_{1i} \) is the 1st covariate for event \( i \);

\( \beta \)'s are the regression parameters

\( \alpha_i \) is the driver specific random random effect
Case-Crossover: Conditional Logistic Regression

Let $p_{ij}$ be the probability of crash/near-crash for $j$th observation in $i$th matched set.

Define

$$Y_{ij} = \begin{cases} 
1 & \text{if the $j$th observation in $i$th matched set is a crash/near-crash.} \\
0 & \text{if the $j$th observation in $i$th matched set is a baseline.}
\end{cases}$$

The matched sampling mechanism leads to:

$$\sum_j Y_{ij} = 1$$

$$\text{logit}(p_{ij} | \sum_j Y_{ij} = 1) = \beta \ast \text{drowsy}_{ij},$$

In this model $\exp(\beta)$ is the estimated OR.
Case-Cohort: Crash Risk

- Drowsiness increases crash risk by 6 times
- Complex secondary task increase the risk by 3 times
- Crashes are more likely to happened in roadway junction areas (6-fold increase)
- Crashes are 5 times more likely to happened on wet, snowy, or muddy road surface.
Comparing Secondary Task Engagement OR for Case-Crossover versus Case-Control

<table>
<thead>
<tr>
<th>Distraction</th>
<th>Case Crossover Odds Ratio</th>
<th>95% Odds Ratio Confidence Limits</th>
<th>Case-Control Odds Ratios</th>
<th>95% Odds Ratio Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>0.8</td>
<td>0.62</td>
<td>1.05</td>
<td>1.2</td>
</tr>
<tr>
<td>Moderate</td>
<td>1.3</td>
<td>1.00</td>
<td>1.70</td>
<td>2.1</td>
</tr>
<tr>
<td>Complex</td>
<td>2.1</td>
<td>1.19</td>
<td>3.58</td>
<td>3.1</td>
</tr>
</tbody>
</table>
Crash Risk Increase Monotonically with Total Eyes Off Forward Roadway

Odds Ratio Point Estimate/Confidence Intervals

Total TEOR (seconds)
Percent TEOR (out of 15 seconds)
Comparison of Case-Cohort and Case-Crossover

• Case-Cohort: Greater generalizability to not only driver behavior but also environmental and roadway risk assessment.

• Case-Cohort: Simpler to conduct and less resource intensive.

• Case-Crossover: Greater precision as potential confounding factors are controlled through baseline sampling.
Can Near-Crashes Serve as Crash Surrogates?

• They are different by definition!

• It depends on the purpose the study.
  – This analysis focuses on the impacts for risk assessment purpose.
Crash Surrogate

- Less-severe events happen more frequently than severe events
- Severe events can be reduced by reducing less severe events

1. The causal mechanism for surrogates (near-crashes) and crashes are the same or similar.
2. There is a strong association between the frequency of surrogate measures and crashes under different settings.
# Driver Reaction on Crash and Near-crash

The image presents a flowchart outlining the relationship between different factors and maneuvers in the context of crash and near-crash events. The flowchart contains the following elements:

- **Pre-Incident Maneuver**
- **Precipitating Factor**
- **Evasive Maneuver**
- **Contributing Factors**

## All Conflict Types

<table>
<thead>
<tr>
<th>i</th>
<th>Crash</th>
<th>Near-Crash</th>
<th>Reaction</th>
<th>No reaction</th>
<th>Perc. Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>45</td>
<td>723</td>
<td>Reaction</td>
<td>5</td>
<td>36%</td>
</tr>
<tr>
<td>No-Reaction</td>
<td>23</td>
<td>37</td>
<td>No reaction</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Perc. Reaction</td>
<td>66%</td>
<td>95%</td>
<td>Perc. Reaction</td>
<td>36%</td>
<td>100%</td>
</tr>
</tbody>
</table>

## Conflict with Leading Vehicle

- **Reaction**: 5 Crash, 377 Near-Crash
- **No reaction**: 9 Crash, 0 Near-Crash
- **Perc. Reaction**: 36% Crash, 100% Near-Crash
# Frequency Relationship

<table>
<thead>
<tr>
<th>Factors</th>
<th>Constant Crash to Near-Crash Ratio</th>
<th>Measure for Association</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>Significant</td>
</tr>
<tr>
<td>Gender</td>
<td>0.26</td>
<td>NO</td>
</tr>
<tr>
<td>Age Group</td>
<td>0.23</td>
<td>NO</td>
</tr>
<tr>
<td>Level of Service (LOS)</td>
<td>&lt;0.001</td>
<td>YES</td>
</tr>
<tr>
<td>Lighting Conditions</td>
<td>0.414</td>
<td>NO</td>
</tr>
<tr>
<td>Road Alignment</td>
<td>0.02</td>
<td>YES</td>
</tr>
<tr>
<td>Road Surface Condition</td>
<td>0.02</td>
<td>YES</td>
</tr>
<tr>
<td>Weather</td>
<td>0.32</td>
<td>NO</td>
</tr>
</tbody>
</table>

\[ y_i \sim \text{Poisson}(\lambda_i) \]

\[ \log(\lambda_i) = \beta_0 + \beta_1 x_i \]

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.31</td>
</tr>
<tr>
<td>Near-Crash</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Sensitivity Analysis

The graph shows the odds ratio (95% CI) for different factors: Distraction, Lighting, Road Surface, Weather, and Drowsiness. The blue line represents 'Crash Only' while the red line represents 'Crash & Near-Crash'.
Surrogate Measure: Summary

• Using crashes plus near-crashes will lead to a conservative but more precise result in risk assessment.

• For smaller studies with an insufficient number of observed crashes, there is a definite benefit to using near-crashes as a crash surrogate.

• Caution should be used when interpreting the results of risk evaluation.