CASE-CONTROL ANALYSIS IN HIGHWAY SAFETY: ACCOUNTING FOR SITES WITH MULTIPLE CRASHES

Frank Gross, PhD, PE
Vanasse Hangen Brustlin, Inc.
333 Fayetteville St, Suite 1450
Raleigh, NC 27601
(919) 834-3972
fgross@vhb.com

ABSTRACT

There is an increased interest in the use of epidemiological methods in highway safety analysis. The case-control and cohort methods are commonly used in the epidemiological field to identify risk factors and quantify the risk or odds of disease given certain characteristics and factors related to an individual. This same concept can be applied to highway safety where the entity of interest is a roadway segment or intersection (rather than a person) and the risk factors of interest are the operational and geometric characteristics of a given roadway. One criticism of the use of these methods in highway safety is that they have not accounted for the difference between sites with single and multiple crashes. In the medical field, a disease either occurs or it does not; multiple occurrences are generally not an issue. In the highway safety field, it is necessary to evaluate the safety of a given site while accounting for multiple crashes. Otherwise, the analysis may underestimate the safety effects of a given factor.

This paper explores the use of the case-control method in highway safety and two variations to account for sites with multiple crashes. Specifically, the paper presents two alternative methods for defining cases in a case-control study and compares the results in a case study. The first alternative defines a separate case for each crash in a given study period, thereby increasing the weight of the associated roadway characteristics in the analysis. The second alternative defines entire crash categories as cases (sites with one crash, sites with two crashes, etc) and analyzes each group separately in comparison to sites with no crashes. The results are also compared to a "typical" case-control application, where the cases are simply defined as any entity that experiences at least one crash and controls are those entities without a crash in a given period. In a "typical" case-control design, the attributes associated with single-crash segments are weighted the same as the attributes of segments with multiple crashes.

The results support the hypothesis that the "typical" case-control design may underestimate the safety effects of a given factor compared to methods that account for sites with multiple crashes. Compared to the first alternative case definition (where multiple crash segments represent multiple cases) the results from the "typical" case-control design are less pronounced (i.e., closer to unity). The second alternative (where case definitions are constructed for various crash categories and analyzed separately) provides further evidence that sites with single and multiple crashes should not be grouped together in a case-control analysis. There is clearly a need to differentiate sites with single and multiple crashes in a case-control analysis. While the results

suggest that sites with multiple crashes can be accounted for using a case-control design, further research is needed to determine the optimal method for addressing this issue.

Keywords: transportation, safety, case-control

INTRODUCTION

The preferred method for estimating the effect of a given factor (e.g., treatment or countermeasure) is an experimental study design. Experimental studies are planned in that entities (e.g., sites or individuals) are identified for some treatment and then randomly assigned to either the treatment or control group. It is generally unethical (and also uneconomical) to "randomly" assign treatment in road safety; there are sites that warrant treatment based on historical or expected safety performance. Those sites with the greatest need generally receive treatment. As such, observational studies are more common than experimental studies in road safety research.

There are various types of observational studies, including before-after study designs and cross-sectional study designs. Before-after studies involve a treatment at some point in time and a comparison of the selected performance measure before and after treatment. In cross-sectional studies, there is no before and after period. Instead, the selected performance measure is compared at sites with and without the treatment of interest.

In highway safety, well-designed observational before-after studies are generally preferred to observational cross-sectional studies to estimate the safety effectiveness of a given treatment (Harwood et al., 2000). The empirical Bayes before-after study is considered the current state-of-the-practice (Hauer, 1997). It is a rigorous method in that it can account for regression-to-the-mean, changes in traffic volume, and other temporal factors that may change from the before to the after period. There are, however, several practical limitations that may preclude the use of the empirical Bayes before-after method, including:

1. Confounding factors: several improvements may be implemented simultaneously, making it difficult to isolate the effect of a single countermeasure using a before-after study. Similarly, changes in traffic volume, driver population, vehicle mix, and other factors may occur over the analysis time period in a before-after study.
2. Sample size: it is sometimes difficult to find an adequate sample of sites where the treatment of interest has actually been implemented. Results from a limited sample will have a high level of statistical uncertainty. When there are few or no sites being treated with the countermeasure of interest, a before-after study is difficult to employ.
3. Study period: a before-after study requires a time sequence, where it is necessary to implement a countermeasure and wait for sufficient data in the after period. While data collection can be time consuming for any safety evaluation, waiting several years after implementation is a practical concern in before-after studies.

Given the limitations associated with observational before-after studies, alternative evaluation methods are sometimes needed to estimate the safety effectiveness of a countermeasure. Cross-

sectional study designs are particularly useful for estimating safety effects where there are insufficient "installations" of a countermeasure. For example, there may be few projects where the degree of horizontal curvature is changed from 15 degrees to 10 degrees, yet there are many horizontal curves that are 10 degrees and many others that are 15 degrees. In this case, a before-after design may be undesirable because there are too few projects that change the degree of curvature. Instead, crash data could be collected for the two groups of curves and compared in a cross-sectional design.

Multivariate regression models are typically used to analyze cross-sectional data. Hauer (2010) argues that cross-sectional studies have not proven successful to identify cause and effect in road safety because multivariate regression typically does not produce consistent results between studies. He suggests that an observational epidemiological approach may, however, be a viable method to control for the many sources of variation present in cross-sectional data.

The case-control method is commonly used in epidemiology to identify risk factors and quantify the risk or odds of disease given certain characteristics and factors related to an individual (Woodward, 2005). It has recently been applied in the highway safety field to investigate the safety effects of geometric variables (Gross, 2006; Gross and Jovanis, 2007). The case-control method is also identified as an alternative method for estimating safety effectiveness in *A Guide to Developing Quality Crash Modification Factors* (Gross et al., 2010). One criticism of the case-control method in highway safety is that it typically does not account for the difference between sites with single and multiple crashes. If this potential weakness can be shored-up, this method may hold greater potential in the field of highway safety.

In the medical field, a disease either occurs or it does not; multiple occurrences are generally not an issue. As such, this issue has not been raised or addressed in other fields. A literature does not exist on how to account for multiple occurrences using the case-control method. This research is an attempt to shed light on the subject and provide a foundation for future research on the topic. Specifically, the objective of this study is to investigate alternative ways to account for sites with multiple crashes using the case-control method. The "typical" case-control analysis in highway safety would define cases as those sites that had at least one crash during the study period. This study compares the results from a "typical" case-control analysis to the results from two alternative case-control analyses, which investigate methods to differentiate between sites with multiple crashes using the case-control study design.

CASE-CONTROL STUDIES

Overview of Case-Control Studies

Case-control methods have been used in certain areas of highway safety, but few have focused on the effects of geometric design elements. For example, case-control studies have been applied to investigate the effectiveness of motorcycle-helmet use (Tsai et al., 1995) and the crash risk of hours of service for truck drivers (Jovanis et al., 2005). More recently, the case-control method was employed to estimate safety effectiveness for geometric design elements, including lane and shoulder width (Gross, 2006; Gross and Jovanis, 2007).

Case-control studies are based on cross-sectional data, but they should not be confused with cross-sectional studies in general. For cross-sectional studies, samples are generally selected based on the presence and absence of a specific characteristic (e.g., lighting) or based on a specific roadway or intersection type, ignoring whether there was a crash there or not. Case-control studies select sites based on outcome status (e.g., crash or no crash) and then determine the prior treatment (or risk factor) status within each outcome group (e.g., presence of lighting).

The most important step in a case-control study is defining the cases and controls. Ambiguous or broad definitions for cases and controls may lead to misclassification and will likely produce unclear results. The case definition (or variations of the case definition) may be an option to explore the difference in risk among sites with a single crash and sites with multiple crashes.

Statistical Analysis of Case-Control Studies

Case-control studies assess whether exposure to a potential treatment is disproportionately distributed between the cases and controls, thereby indicating the likelihood of an actual benefit from the treatment. The safety effect is expressed as the odds ratio between two levels of a variable. For example, it may be found that the odds of a crash occurring on horizontal curves with a degree of curvature greater than 15 degrees is 1.5 times the odds of a crash occurring on curves less than 15 degrees. The odds ratio is a direct estimate of the safety effectiveness.

The case-control method can be used to estimate the safety effect of binary variables (e.g. median barrier, roadway lighting, or guardrail) or multi-level variables such as lane width (e.g. 9, 10, 11 and 12 foot lanes). Multiple logistic regression techniques (or conditional logistic regression in the case of a matched case-control study), are commonly used to clarify these relationships because they are able to examine the risk or benefit associated with one factor while controlling for other factors.

The ratio of controls to cases may vary and often depends on the availability of time, budget, and potential sites. Increasing the number of controls will increase the power of the study, especially when there are relatively few cases. Power is defined as the probability that the test will reject a false null hypothesis. In a matched design, controls are sampled randomly and matched to each case based on similar values of the potential confounding variable. Matching provides a balanced design and automatically adjusts the estimates for the potential confounding effects of variables included in the matching scheme.

The conditional probability of an outcome associated with the unmatched variables $x_1,\ldots,x_p$ for each member of the j^{th} matched set is given by Equation 1 (Schlesselman, 1982).

$$\Pr(Y) = 1/\{1 + \exp[-(\alpha_j + \sum_{i=1}^{p} \beta_i x_i)]\} \quad (1)$$

Where:

Y = the outcome (1 = case and 0 = control).
α_j = the effect of matching variables for each matched set.
β_i = estimated coefficients for explanatory variables.
x_i = unmatched explanatory variables included in the model.

Estimates of the coefficients for the explanatory variables are obtained by maximizing the likelihood expression in Equation 2.

$$L(\beta_i) = -\sum_{j=1}^{n} \ln\left[1 + \sum_{k=1}^{c} \exp\left\{\sum_{i=1}^{p} \beta_k \left(x_{jki} - x_{j0i}\right)\right\}\right] \quad (2)$$

Where:

$L(\beta_i)$ = likelihood estimate of coefficient i.
n = number of cases.
c = number of controls matched to each of n cases.
x_i = unmatched explanatory variables.
x_{j0i} = value of x_i for a case in the j^{th} matched set.
x_{jki} = value of x_i for the k^{th} matched control in the j^{th} matched set.

Strengths and Limitations of Case-Control Studies

The case-control method is useful for studying rare events (such as crashes) because the number of cases and controls is predetermined. Another advantage of the case-control design is that multiple treatments may be investigated in relation to a single outcome using the same sample (i.e., a single sample may be used to investigate any variables that are not included in the selection or matching criteria for cases and controls). While case-control studies may be used to explore multiple treatments, they can only investigate one outcome per sample. The sampling is conducted separately within the case and control populations based on outcome status and different outcomes produce different samples. As such, it is necessary to draw separate samples from the database to investigate multiple target crash types (e.g., total crashes, run-off-road crashes, etc).

The case-control method cannot demonstrate causality because there is no time sequence of events in the analysis. Instead, the odds ratio indicates the increased/decreased likelihood of a crash occurring when a treatment (e.g., roadway characteristic) is present. In general, it also does not recognize differences between locations with many crashes or a single crash. This is a loss of potentially important information and thus, the true increase in risk could be underestimated.

EMPIRICAL SETTING

Method

A matched case-control design is set-up to evaluate the effects of shoulder width and lane width on roadway segment crashes. A sample is selected from the population of all rural, two-lane,

undivided roadway segments in Pennsylvania. This study population is used to eliminate the variability between rural and urban segments, multi-lane segments, and those segments with and without a median. The study period includes five years of data from 1997 to 2001. Each year of the study period is analyzed separately as opposed to aggregating the five years of data before selecting cases and controls. If the five years of data were aggregated before case selection, it is likely that many more segments would experience at least one crash in this period, leaving relatively few controls for comparison.

Case Definition

This study investigates specific crash types that tend to be influenced by lane and shoulder width. These "related" crash types are head-on, run-off-road, opposite direction sideswipe, and same direction sideswipe crashes as presented in the Highway Safety Manual (AASHTO, 2010). The case definition is the primary focus of this investigation, and three different case definitions are explored to investigate alternative methods for evaluating sites with multiple crashes. The three case definitions are as follows:

1. Typical case definition: Cases are defined as segments that experience at least one "related" crash during a particular year of the study period, regardless of the number of crashes (i.e., there is no differentiation of sites with one crash and sites with multiple crashes). For example, a segment with three "related" crashes in the year 1999 would be defined as a single case, as would a segment with one "related" crash in 1998.
2. Multiple crashes = multiple cases: Cases are defined as segments that experience at least one "related" crash during a particular year of the study period, but each "related" crash represents an individual case. For example, a segment with three "related" crashes in the year 1999 would represent three cases and each case would be defined with similar attributes (e.g., roadway geometry and traffic volume). A segment with one "related" crash in 1998 would represent a single case. In this way, the characteristics of sites with multiple crashes will be more prevalent in the database, thereby increasing the odds associated with those specific features.
3. Each crash category represents a separate case definition: Cases are defined by the number of "related" crashes reported on a particular segment in a given year. This method redefines the case definition in each of a series of analyses. Cases are first defined as sites with a single crash in a given year and controls are those sites without a crash in the same year. Next, the case definition is modified to include sites with two crashes in a given year and controls remain the same (i.e., sites without a crash in the same year). This step is repeated as many times as necessary to adjust the case definition so that it includes sites with multiple crashes. The final category could lump sites that exceed a certain threshold to overcome what could become relatively small sample sizes. For example, a database may contain 50 sites with three crashes, 35 sites with four crashes, and 15 sites with five crashes. Each category individually may be too small for a meaningful analysis; however, the case definition for the last category could include "sites with three or more crashes", which would result in an aggregate category of 100 sites.

Selection of Controls

In each variation of the case definition, controls are defined similarly as those segments with no reported crashes in a given year. Controls could be defined as those segments with no "related" crashes in the same year as a case segment; however, there is the potential that crashes could be miscoded by type. To partly overcome the issue of miscoding, controls were always defined as segments that did not experience any crashes in the same year as a case segment.

Control segments are randomly selected, at a ratio of 1:1, from the same population as each case segment. In this case, the entire population of rural, two-lane, undivided roadway segments in Pennsylvania are identified. The data are then binned by year and coded as cases and controls based on the outcome status. Cases and controls are separated into two datasets for each given year and a random number generator is used to match cases and controls from the same year. The cases and controls are selected without regard to additional geometric or traffic characteristics, differentiated only by outcome status (crash or no crash) during a particular year.

Confounding Variables

A confounding factor is a variable that completely or partially accounts for the apparent association between an outcome and a treatment. Specifically, a confounder is a variable that is a risk factor for the outcome under study, and is associated with, but not a consequence of, the risk factor in question (Collett, 2003). Traffic volume and segment length are both significant predictors of crash frequency and may also be associated with design characteristics. If a specific design characteristic (e.g. lane or shoulder width) is suspected to be a risk factor of crashes then the effects of traffic volume and segment length must be separated before the true effects of the variable of interest may be known (Persaud et al., 1999 and Hauer et al., 2004). This holds for many variables and emphasizes the importance of controlling for outside effects (i.e. effects from sources other than the variable of interest).

Adjustment for potential confounders is applied during the selection of cases and controls (i.e., matching) as well as during the model estimation process. The risk factors of interest in this particular study include lane width and paved shoulder width. Potential confounders included in the matching scheme include area type, number of lanes, median type, and year as these variables were used to define cases and controls. Potential confounders included in the analysis include AADT, segment length, posted speed limit, additional (unpaved) shoulder width, and PennDOT district. PennDOT district identifies the general location of a segment within the state and was included to help account for regional influences such as topography, weather, maintenance practices, driver populations, and crash reporting.

Data

Geometric, traffic and crash data were obtained for rural, two-lane, undivided highway segments in Pennsylvania from 1997 – 2001. The data were obtained in two parts (1) a crash inventory database extracted from the Pennsylvania Crash Reporting System and (2) a roadway inventory file. Crash data were available for each year of the study period; however, only one geometric

file was available for the five-year period. The crash data were merged with the geometric data using the three unique identifying features (i.e. county, route number and segment number) and separated by year.

The crash inventory data include all reportable crashes for mid-block locations (i.e. non-intersection crashes). Reportable crashes are defined as those in which at least one vehicle is towed from the scene. This dataset does not contain crashes occurring at or near intersections and any data from "phantom" or "hit-and-run" crashes are excluded. The dataset includes state roads only and does not include turnpike crashes.

Table 1 indicates the total number of case-control pairs for each of the alternative case definition schemes. Comparison of the case definition schemes illustrates the effects of the case definition on sample size. It is understood that the number of available cases increases as the case definition is changed to represent each crash as a separate case. For example, a segment with three crashes would be represented by a single case under the first case definition, but would be represented by three individual cases under the second case definition. Under the third alternative case definition, sample sizes decrease as the number of crashes increases. In this case, separate case definitions are used to represent segments with one, two, and three or more crashes. Generally, there will be more segments with one crash in a given year than segments with two or three crashes in a given year.

Table 1 Sample Sizes for Alternative Case Definitions

Case Definition Scheme		Case-Control Pairs
1. Typical Case Definition: Cases are defined as segments that experience at least one "related" crash during a particular year of the study period, regardless of the number of crashes.		27,523
2. Multiple Crashes = Multiple Cases: Cases are defined as segments that experience at least one "related" crash during a particular year of the study period, and each "related" crash represents an individual case.		36,206
3. Separate Case Definition for Each Crash Category: Cases are defined by the number of "related" crashes reported on a particular segment in a given year.	Segments with 1 crash in a given year.	21,205
	Segments with 2 crashes in a given year.	4,765
	Segments with 3+ crashes in a given year.	1,553

Descriptive statistics are provided in Table 2 for the entire population of cases and controls (note that a random sample of controls was drawn from the population for each specific case definition). Traffic volume and segment length are included in the analysis as continuous variables. Segment length is approximately normally distributed with a mean of approximately 0.5 miles and standard deviation of approximately 0.13 miles. Traffic volume (AADT) is not normally distributed and several transformations were tested in an attempt to normalize the variable. A cube root transformation was selected ($AADT^{1/3}$), which has a mean and standard deviation of 14.0 and 4.2, respectively. Posted speed limit ranges between 15 mph and 55 mph

with a mean and standard deviation of 48 mph and 7.6 mph, respectively. Lane width ranges from 6 to 33 feet with a mean of approximately 11 feet. Note that the very narrow and very wide lane widths may be errors in the data, but a review of video logs for a sample of these sites indicated some very narrow and very wide pavement widths. It was decided to leave these segments in the data because they could not be confirmed as errors. Paved shoulder width ranges from 0 to 15 feet with a mean of 2.8 feet. Additional (unpaved) shoulder width ranges from 0 to 13 feet with a mean of 1.3 feet.

Table 2 Descriptive Statistics for Study Population

Variable	Mean	Standard Deviation	Minimum	Maximum
Segment Length (ft)	2529	677	23	7793
AADT (vehicles/day)	3488	3098	95	25844
$AADT^{1/3}$	14.0	4.2	4.6	29.6
Speed Limit (mi/h)	48	7.6	15	55
Lane Width (ft)	11.1	1.7	6	33
Paved Shoulder Width (ft)	2.8	2.4	0	15
Additional Shoulder Width (ft)	1.3	1.9	0	13

RESULTS

Table 3 compares the results from the typical case definition with the first alternative. As a reminder, the typical case definition defines cases as those segments that experience at least one crash in a given year, but does not differentiate segments with a single crash from segments with multiple crashes. The alternative case definition creates a separate case for each crash in a given year.

Examining the results in Table 3, it is apparent that the two methods produce relatively consistent results. The odds ratio increases as AADT increases, which is consistent with previous research (Schoppert, 1992) and the safety models presented in the Highway Safety Manual (AASHTO, 2010). The odds ratio is approximately 1.0 for segment length. This is also consistent with previous research that shows crash risk to increase linearly with segment length (i.e., the odds of a crash is proportional to the segment length). The effect of speed limit is somewhat counterintuitive in that the odds ratio decreases as speed increases. While this is not an intuitive result, similar effects have been shown in other research (Solomon, 1964; Milton and Mannering, 1998). The odds ratio generally decreases as lane width and paved shoulder width increase. The effects of lane and paved shoulder width are less consistent at the extremes and these results are further explained in previous work (Gross, 2006). It should be noted that paved shoulder widths of zero feet and two feet to six feet are most prevalent in the dataset. For lane width, the most prevalent widths are 10, 11, and 12 feet. The moderate sample sizes for lane widths less than ten feet and greater than thirteen feet are the result of combining several small samples in these ranges.

One notable difference is that the estimated coefficients from the typical case definition are generally closer to unity than the estimated coefficients from the alternative case definition. In other words, when cases are defined as any crash-related site (regardless of number of crashes), the estimated safety effect is closer to 1.0 relative to the alternative case definition (i.e., when

cases are defined by each crash, not by a single site). This result suggests that there is a risk of underestimating effects if sites with multiple crashes are not accounted for in the case-control analysis.

Figures 1 and 2 further illustrate the disparity between the results from the two case definitions. Figure 1 compares the results for lane width and Figure 2 compares the results for paved shoulder width. In Figures 1 and 2, the dashed line represents the typical case definition, while the solid line represents the alternative case definition where each crash is used as a separate case. The dashed line is generally closer to 1.0 than the solid line, indicating that the typical case definition may underestimate the effects of lane and paved shoulder width.

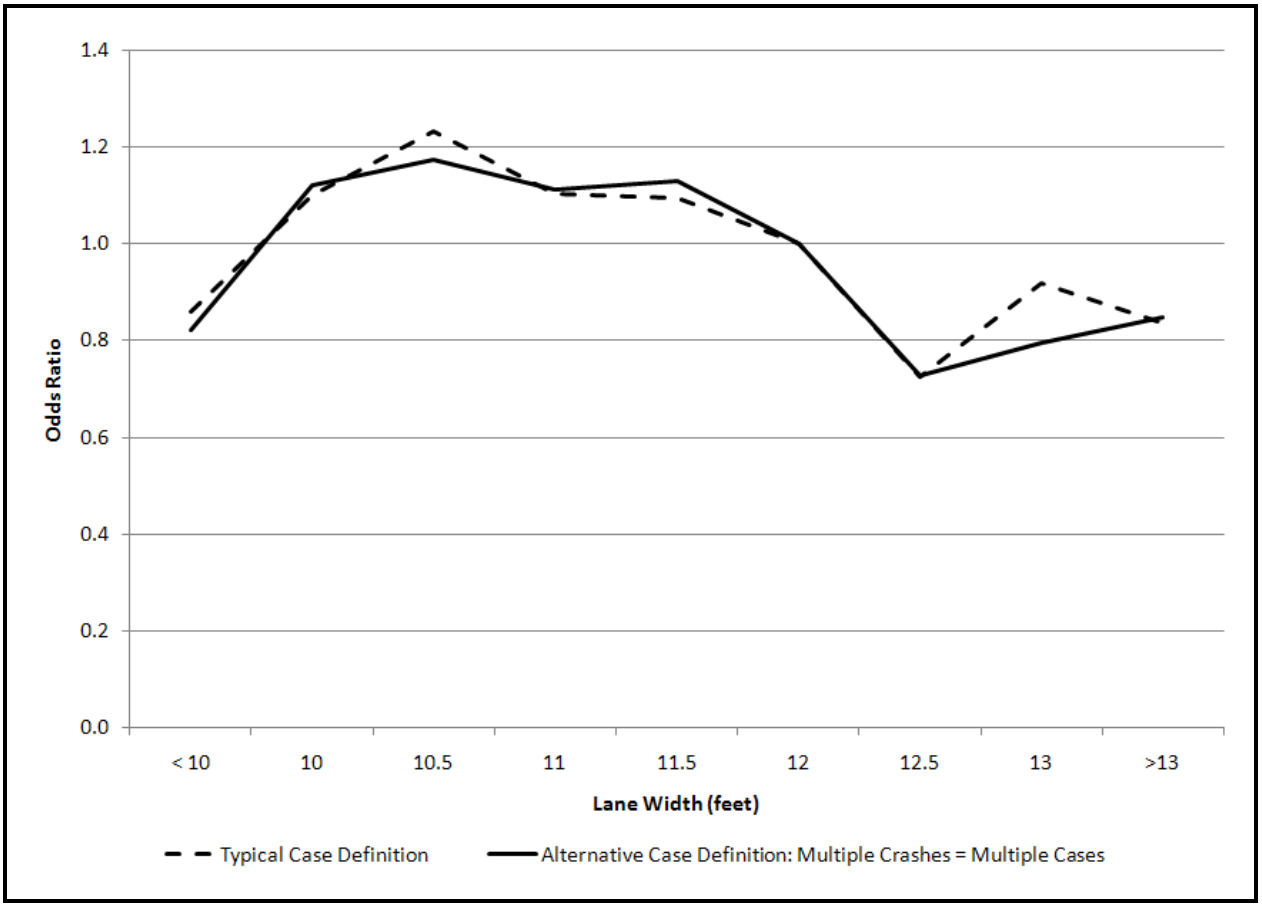

Figure 1 Comparison of Results for Typical and Alternative Case Definition: Lane Width

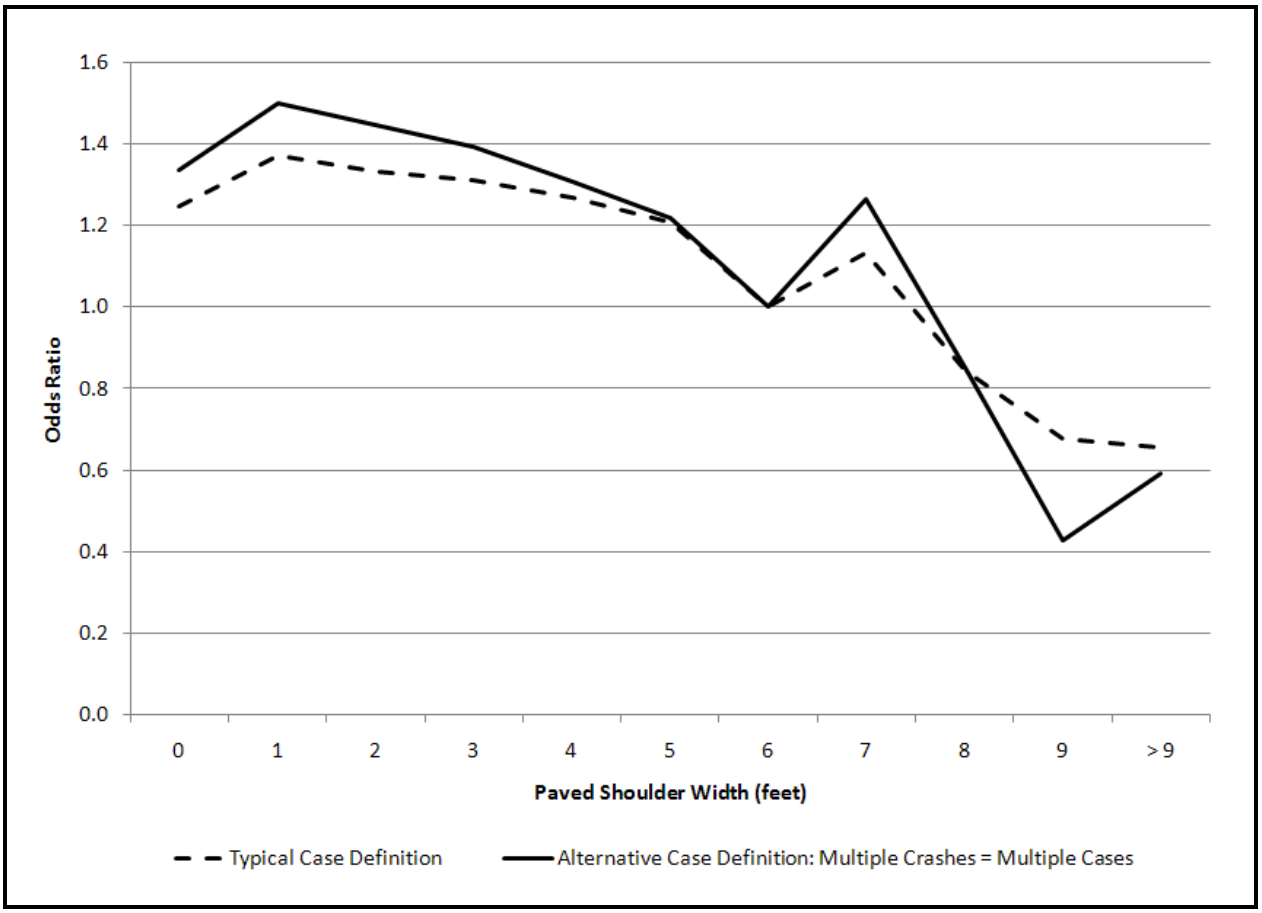

Figure 2 Comparison of Results for Typical and Alternative Case Definition: Shoulder Width

Table 3 Comparison of Conditional Logistic Regression Results for Case Definition #1 and #2

	Case Definition #1			Case Definition #2		
Variable	Coefficient	Standard Error	Sample Size	Coefficient	Standard Error	Sample Size
$AADT^{1/3}$	1.17	0.00	55,046	1.19	0.00	72,412
Segment Length	1.00	0.00	55,046	1.00	0.00	72,412
Speed Indicator (1=50mph or greater)	0.92	0.02	55,046	0.88	0.02	72,412
Lane Width < 10 ft	0.86	0.04	3,735	0.82	0.04	4,758
Lane Width = 10 ft	1.10	0.04	14,645	1.12	0.03	19,181
Lane Width = 10.5 ft	1.23	0.07	2,017	1.17	0.06	2,673
Lane Width = 11 ft	1.10	0.03	20,338	1.11	0.03	26,994
Lane Width = 11.5 ft	1.09	0.10	605	1.13	0.10	828
Lane Width = 12 ft	1.00	NA	10,574	1.00	NA	13,828
Lane Width = 12.5 ft	0.73	0.12	202	0.73	0.10	274
Lane Width = 13 ft	0.92	0.11	426	0.79	0.08	597
Lane Width > 13 ft	0.84	0.04	2,504	0.85	0.04	3,279
Shoulder Width = 0	1.25	0.07	15,411	1.34	0.06	19,875
Shoulder Width = 1	1.37	0.11	1,258	1.50	0.11	1,661
Shoulder Width = 2	1.33	0.07	7,180	1.45	0.07	9,653
Shoulder Width = 3	1.31	0.06	9,324	1.39	0.06	12,247
Shoulder Width = 4	1.27	0.06	11,997	1.31	0.05	15,950
Shoulder Width = 5	1.21	0.07	3,057	1.22	0.06	4,009
Shoulder Width = 6	1.00	NA	3,168	1.00	NA	4,133
Shoulder Width = 7	1.13	0.13	517	1.26	0.12	699
Shoulder Width = 8	0.84	0.05	2,274	0.85	0.05	3,047
Shoulder Width = 9	0.68	0.18	78	0.43	0.10	113
Shoulder Width > 9	0.65	0.06	782	0.59	0.05	1,025
Additional Shoulder Width = 0	1.00	NA	32,663	1.00	NA	43,310
Additional Shoulder Width = 1	0.97	0.05	2,408	1.00	0.05	3,186
Additional Shoulder Width = 2	1.03	0.03	8,341	1.02	0.03	10,972
Additional Shoulder Width = 3	0.97	0.04	3,535	0.96	0.04	4,480
Additional Shoulder Width = 4	0.95	0.04	5,275	0.93	0.04	6,825
Additional Shoulder Width = 5	0.81	0.07	656	0.80	0.07	851
Additional Shoulder Width = 6	0.82	0.06	1,183	0.87	0.06	1,472
Additional Shoulder Width = 7	0.50	0.11	111	0.49	0.10	136
Additional Shoulder Width = 8	0.76	0.07	606	0.81	0.07	779
Additional Shoulder Width = 9	0.67	0.22	47	0.46	0.12	70
Additional Shoulder Width > 9	0.69	0.11	221	0.57	0.08	283
District 1	1.00	NA	6,027	1.00	NA	7,587
District 2	0.98	0.04	8,017	0.99	0.04	10,346
District 3	1.05	0.04	6,938	1.05	0.04	8,780
District 4	1.32	0.06	5,067	1.31	0.05	6,655
District 5	1.62	0.08	4,033	1.94	0.08	5,813
District 6	1.29	0.10	1,247	1.39	0.09	1,884
District 8	1.17	0.05	7,688	1.20	0.04	10,426
District 9	1.00	0.04	5,640	1.06	0.04	7,371
District 10	1.17	0.05	5,098	1.21	0.05	6,638
District 11	1.00	0.07	1,360	1.04	0.06	1,779
District 12	1.19	0.06	3,931	1.26	0.05	5,133

Table 4 presents the results of the second alternative case definition (case definition scheme #3 from Table 1). In this scenario, separate case definitions (and analyses) are constructed for various crash categories. The first case definition only looks at segments with one crash in a

given year and compares them to segments with no crashes in the same year. This is also done for segments with two crashes and segments with three or more crashes. In this way, risk factors such as lane and shoulder width can be assessed for each crash category. If the results are consistent across all categories then it may be reasonable to define cases as those segments with at least one crash in a given year as is done in the "typical" case-control analysis.

The results from Table 4 are relatively consistent with the results presented in Table 3. The coefficients for AADT, segment length, and posted speed limit indicate that the magnitude and direction of the effect are consistent with previous results. The odds ratio for AADT is greater than one for all three case categories indicating that AADT is a risk factor (i.e., odds of a crash segment increases as AADT increases). Note also that the odds ratio for AADT increases as the number of crashes per segment increases from one to two to three or more crashes. This indicates that AADT is a greater risk factor for segments with three or more crashes than it is for segments with one crash. While the general trend for lane and paved shoulder width are similar to the results from Table 3, it is obvious that the odds ratio varies among the three case categories for any given lane or paved shoulder width.

The differences among the three case categories are further illustrated in Figures 3 and 4 for lane width and paved shoulder width, respectively. It is apparent that lane width and paved shoulder width are less of a risk factor for segments with one crash compared to segments with two or three or more crashes. Specifically, the trend lines for segments with one crash are much closer to unity than the trend lines for segments with two crashes or three or more crashes. For lane width, a similar comparison can be made between segments with two crashes and segments with three or more crashes. The same conclusion can be made that lane width is less of a risk factor (i.e., the trend line is less steep and closer to unity) for segments with two crashes compared to segments with three or more crashes. For paved shoulder width, the odds ratio for segments with two crashes is generally greater than the odds ratio for segments with one crash and the odds ratio for segments with three or more crashes is consistently lower than the other two case categories.

It is apparent that the odds ratios for lane and paved shoulder width are not consistent among the three case categories, indicating that it may not appropriate to combine all three categories together, as is the case in the "typical" case-control analysis. However, the second alternative case definition did not produce completely reasonable results. It was expected that the results for the three case categories (1-crash segments, 2-crash segments, and 3-crash segments) would follow some logical order in terms of the respective coefficients. It is clear from Figure 6 that this is not the case. There are several possible explanations, but there is clearly a need for further investigation. Specifically, the second alternative case definition could be further explored, using alternative modeling techniques. In this study, the logistic regression model was applied to analyze the data. The logistic regression model is appropriate when outcomes are 0 or 1. While the cases and controls can be represented by a 0/1 indicator, it may be more appropriate to use a count-based model to compare the various case categories (1, 2, and 3+ crashes).

Table 4 Conditional Logistic Regression Results for Case Definition #3

	Cases = 1 Crash			Cases = 2 Crashes			Cases = 3+ Crashes		
Variable	Coeff.	S.E.	S.S.	Coeff.	S.E.	S.S.	Coeff.	S.E.	S.S.
$AADT^{1/3}$	1.15	0.00	42,410	1.26	0.01	9,530	1.33	0.03	3,106
Segment Length	1.00	0.00	42,410	1.00	0.00	9,530	1.00	0.00	3,106
Speed Indicator (1=50mph or greater)	0.93	0.02	42,410	0.79	0.05	9,530	0.79	0.10	3,106
Lane Width < 10 ft	0.90	0.05	2,945	0.73	0.11	562	0.60	0.19	182
Lane Width = 10 ft	1.08	0.04	11,290	1.08	0.09	2,467	1.34	0.26	822
Lane Width = 10.5 ft	1.12	0.07	1,599	1.16	0.17	358	1.78	0.61	98
Lane Width = 11 ft	1.06	0.03	15,487	1.07	0.08	3,660	1.15	0.19	1,201
Lane Width = 11.5 ft	1.03	0.11	467	1.09	0.29	106	1.10	0.58	40
Lane Width = 12 ft	1.00	NA	8,156	1.00	NA	1,826	1.00	NA	564
Lane Width = 12.5 ft	0.87	0.16	160	0.61	0.24	38	0.16	0.17	16
Lane Width = 13 ft	0.82	0.10	333	0.78	0.22	83	0.43	0.29	31
Lane Width > 13 ft	0.84	0.05	1,973	0.61	0.09	430	0.75	0.23	152
Shoulder Width = 0	1.19	0.07	12,070	1.67	0.24	2,492	0.83	0.25	779
Shoulder Width = 1	1.30	0.12	955	1.75	0.38	221	0.83	0.37	85
Shoulder Width = 2	1.25	0.07	5,473	1.87	0.26	1,237	1.43	0.41	452
Shoulder Width = 3	1.24	0.07	7,150	1.61	0.21	1,679	1.38	0.38	529
Shoulder Width = 4	1.23	0.06	9,171	1.46	0.18	2,144	1.21	0.32	715
Shoulder Width = 5	1.18	0.08	2,382	1.46	0.22	539	0.80	0.26	154
Shoulder Width = 6	1.00	NA	2,403	1.00	NA	560	1.00	NA	173
Shoulder Width = 7	1.17	0.15	347	3.37	1.07	93	0.52	0.28	37
Shoulder Width = 8	0.87	0.06	1,747	0.79	0.13	426	0.37	0.13	134
Shoulder Width = 9	0.66	0.19	59	0.92	0.75	11	0.06	0.09	7
Shoulder Width > 9	0.66	0.06	653	0.60	0.15	128	0.16	0.08	41
Additional Shoulder Width = 0	1.00	NA	25,034	1.00	NA	5,774	1.00	NA	1,957
Additional Shoulder Width = 1	1.00	0.06	1,805	1.03	0.13	457	0.87	0.25	147
Additional Shoulder Width = 2	0.98	0.04	6,492	1.01	0.09	1,440	1.63	0.32	432
Additional Shoulder Width = 3	1.01	0.05	2,721	0.90	0.11	573	1.10	0.29	189
Additional Shoulder Width = 4	0.93	0.04	4,156	0.96	0.11	858	0.87	0.22	248
Additional Shoulder Width = 5	0.83	0.08	512	1.03	0.28	92	0.86	0.43	35
Additional Shoulder Width = 6	0.95	0.08	877	0.78	0.15	179	1.98	0.90	52
Additional Shoulder Width = 7	0.60	0.14	91	0.26	0.17	15	0.03	0.04	6
Additional Shoulder Width = 8	0.78	0.08	490	0.76	0.21	103	1.48	1.07	25
Additional Shoulder Width = 9	0.75	0.24	46	0.55	0.54	5	0.34	0.46	4
Additional Shoulder Width > 9	0.71	0.12	186	0.27	0.14	34	0.51	0.49	11
District 1	1.00	NA	4,727	1.00	NA	979	1.00	NA	242
District 2	0.98	0.04	6,233	1.00	0.11	1,299	0.89	0.22	425
District 3	0.99	0.04	5,502	1.15	0.13	1,099	1.03	0.27	325
District 4	1.24	0.06	3,950	1.39	0.17	850	1.50	0.40	279
District 5	1.36	0.07	2,836	2.28	0.29	837	5.74	1.65	354
District 6	1.02	0.09	882	1.15	0.22	256	4.40	1.83	134
District 8	1.05	0.05	5,912	1.25	0.13	1,376	1.89	0.46	490
District 9	0.89	0.04	4,326	1.06	0.13	999	1.36	0.37	322
District 10	1.07	0.05	3,962	1.24	0.15	890	1.36	0.37	258
District 11	1.00	0.07	1,072	0.88	0.16	240	1.62	0.64	67
District 12	1.13	0.06	3,008	1.16	0.14	705	1.69	0.49	210

Note: Coeff. = coefficient, S.E. = standard error, and S.S. = sample size

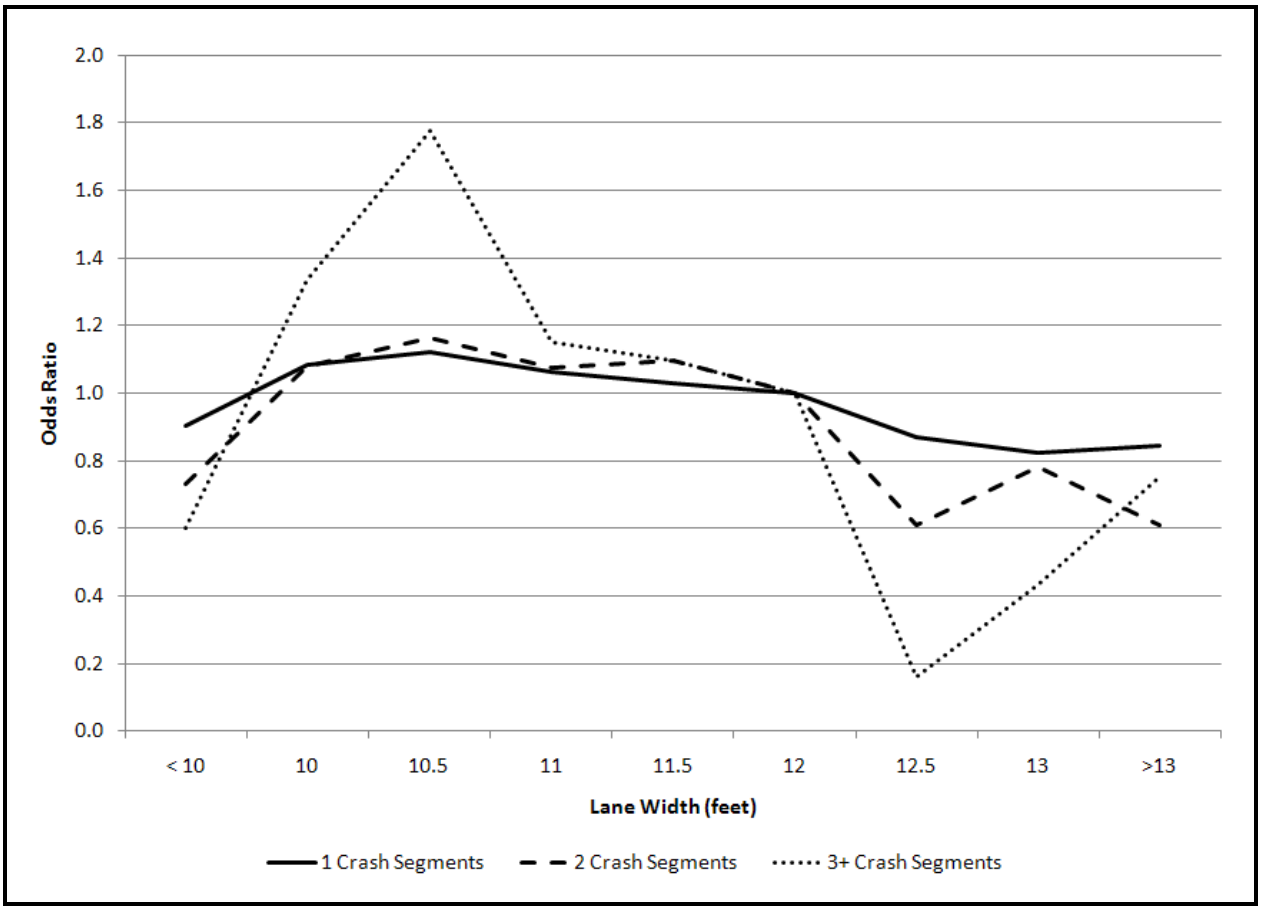

Figure 3 Results for Case Definition #3: Lane Width

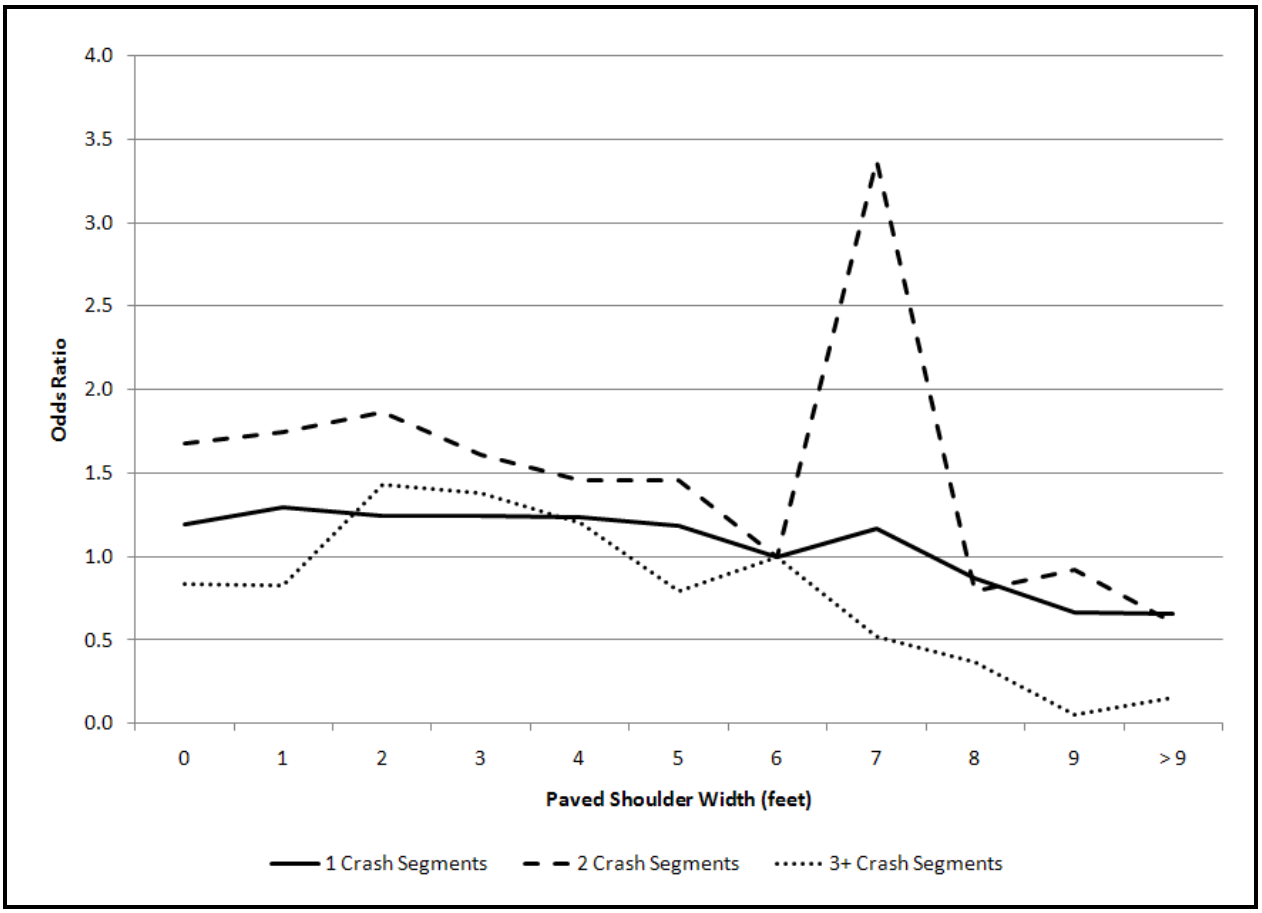

Figure 4 Results for Case Definition #3: Paved Shoulder Width

CONCLUSIONS

Well-designed observational before-after studies are generally the preferred method for estimating the effectiveness of treatments in the highway safety field. Before-after studies are not always feasible, however, due to several practical limitations. As such, alternative evaluation methods are sometimes needed to estimate the safety effectiveness of a countermeasure.

Epidemiological methods have been identified as potential alternatives for estimating the safety effectiveness of treatments in highway safety. Specifically, the case-control method appears to be a potentially viable option for estimating treatment effects. The primary criticism of the case-control method in highway safety is that a "typical" case-control design does not account for sites with multiple crashes in a given period.

This paper employed a case study to explore the use of the case-control method to estimate the safety effects of lane and paved shoulder width. It focused on two alternative approaches for addressing sites with multiple crashes and compared the results to the "typical" case-control design that does not differentiate between sites with single and multiple crashes. In the "typical" case-control design, any roadway segment with one or more crashes in a given time period is

defined as a case. In this way, the attributes associated with single-crash segments are weighted the same as the attributes of segments with multiple crashes.

It was hypothesized that the "typical" case-control design would underestimate the safety effects of a given treatment and the results supported this hypothesis. Specifically, the results from the "typical" case-control design suggested a general decrease in the odds ratio as lane width increases and a similar reduction in the odds ratio as paved shoulder width increases. However, when compared to the first alternative case definition (where multiple crash segments represented multiple cases) the results from the "typical" case-control design were less pronounced (i.e., closer to unity). The second alternative (where case definitions were constructed for various crash categories and analyzed separately) provided further evidence that sites with single and multiple crashes should not be grouped together in the analysis. The second alternative showed how the estimated safety effect of lane width and paved shoulder width followed similar trends for the three case categories, but the magnitude of the effects were substantially different. There is a need to explore the second alternative case definition, using alternative modeling techniques.

This research has confirmed the criticism that case definitions used in the "typical" case-control design may underestimate the odds ratio for associated risk factors, thereby demonstrating the need to differentiate sites with single and multiple crashes. Two alternative methods were presented to account for sites with multiple crashes in the case-control context. While the results suggest that sites with multiple crashes can be accounted for using a case-control design, further research is needed to determine the optimal method for addressing this issue.

REFERENCES

American Association of State Highway Transportation Officials (AASHTO). *Highway Safety Manual*, 1st Edition, Washington, DC, 2010.

Bahar, G., M. Masliah, C. Mollett, and B. Persaud. Integrated Safety Management Process. *NCHRP Report 501*, Transportation Research Board, National Cooperative Highway Research Program, Washington, DC, 2003.

Collett, D. (2003). *Modelling Binary Data.* Second Edition. New York: Chapman and Hall/CRC.

Gross, F. (2006). A Dissertation in Civil Engineering: Alternative Methods for Estimating Safety Effectiveness on Rural, Two-Lane Highways: Case-Control and Cohort Methods. University Park, PA: The Pennsylvania State University.

Gross, F. and P.P. Jovanis (2007). Estimation of the Safety Effectiveness of Lane and Shoulder Width: The Case-Control Approach. American Society of Civil Engineers, *Journal of Transportation Engineering*, 133(6).

Gross, F., B. Persaud, and C. Lyon. A Guide to Developing Quality Crash Modification Factors. Report No. FHWA-SA-10-032, Federal Highway Administration, Washington, DC, 2010.

Harwood, D.W., F.M. Council, E. Hauer, W.E. Hughes, and A. Vogt. Prediction of the Expected Safety Performance of Two-lane Rural Highways. Report No. FHWA-RD-99-207, Federal Highway Administration, McLean, VA, 2000.

Hauer, E. (1997). *Observational Before–After Studies in Road Safety*. Pergamon Press, Oxford, UK.

Hauer, E. (2004). Statistical Road Safety Modeling. *Transportation Research Record 1897*, Transportation Research Board, National Research Council, Washington, D.C., pp. 81-87.

Hauer, E. (2010). Cause, Effect and Regression in Road Safety: A Case Study. *Accident Analysis and Prevention*, 42(4), 1128−1135.

Jovanis, P.P., S.W. Park, K.Y. Chen, and F. Gross. On the Relationship of Crash Risk and Driver Hours of Service. 2005 International Truck and Bus Safety and Security Symposium, Alexandria, Virginia, November 14–16, 2005.

Milton, J. and F. Mannering (1998). The Relationship Among Highway Geometrics, Traffic-Related Elements and Motor-Vehicle Accident Frequencies. *Transportation*, 25(4), 395-413.

Persaud, B., C. Lyon, and T. Nguyen (1999). Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record 1665*, Transportation Research Board, National Research Council, Washington, D.C., pp. 7-12.

Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

Schoppert, D. (1992). Predicting Traffic Accidents from Roadway Elements of Rural Two-Lane Highways with Gravel Shoulders. *Transportation Research Record 1356*, Transportation Research Board, National Research Council, Washington, D.C., pp. 4 – 26.

Solomon, D. Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle. U.S. Department of Transportation, Federal Highway Administration, 1964.

Tsai, Y.J., J.D. Wang, and W.F. Huang (1995). Case-Control Study of the Effectiveness of Different Types of Helmets for the Prevention of Head Injuries among Motorcycle Riders in Taipei, Taiwan. *American Journal of Epidemiology*, 142(9), 974−981.

Woodward, M. (2005). *Epidemiology: Study Design and Data Analysis.* Second Edition. New York: Chapman and Hall/CRC.