Near Crashes as Crash Surrogate for Naturalistic Driving Studies

Feng Guo, Sheila G. Klauer, Jonathan M. Hankey, and Thomas A. Dingus

Naturalistic driving is an innovative method for investigating driver behavior and traffic safety. However, as the number of crashes observed in naturalistic driving studies is typically small, crash surrogates are needed. A study evaluated the use of near crashes as a surrogate measure for assessment of the safety impact of driver behaviors and other risk factors. Two metrics, the precision and bias of risk estimation, were used to assess whether near crashes could be combined with crashes. The principles and exact conditions for improved precision and unbiased estimation were proposed and applied to data from the 100-Car Naturalistic Driving Study. The analyses indicated that a positive relationship exists between the frequencies of contributing factors for crashes and for near crashes. The study also indicated that analyses based on combined crash and near-crash data consistently underestimate the risk of contributing factors compared to use of crash data alone. At the same time, the precision of the estimation will increase. This consistent pattern allows investigators to identify true high-risk behaviors while qualitatively assessing potential bias. In summary, the study concluded that the use of near crashes as a crash surrogate provides definite benefit when naturalistic studies are not large enough to generate sufficient numbers of crashes for statistical analysis.

Traffic crashes are considered as one of the primary measures of traffic safety. However, crash data are not universally available or useful when crash data cannot be collected directly, for example, for proposed new transportation facilities, or when crash data are too sparse given that traffic accidents are rare events. In these situations, there are not enough observations for reaching statistically significant conclusions.

For these reasons, alternative measures of safety have been proposed, and crash surrogates represent an indirect measure of safety. The most widely used surrogate measure technique, the traffic conflict technique, was developed to evaluate vehicle safety (*1, 2*). The traffic conflict technique attempts to measure safety by using cameras and other sensors at various roadway locations to assess how drivers negotiate the roadway. The reliability and validity of the traffic conflict technique has been a major concern, and a number of studies have tried to address this issue (*3–6*). Some empirical studies found that there were clear relationships between traffic conflicts and crashes (*7, 8*). Despite the concerns about these issues, traffic conflict techniques have been used in various studies to evaluate safety (*9–12*).

Surrogate measures are especially useful for evaluating the performance of new roadway designs. These surrogate measures are typically conducted in computer simulation models. Gettman and Head (*13*) and Gettman et al. (*14*) found a relatively weak relationship between traffic conflicts and crash rate.

In recent years, advanced approaches have been used, including the extreme value method (*15*), the counterfactual approach (*16*), and an automated road safety analysis that uses a probabilistic framework (*17*).

Naturalistic driving is similar to the traffic conflict technique, except that the instrumentation is located in the vehicle (vehicle based) rather than roadway location or intersection based. The naturalistic driving method is an innovative method for investigating driver behavior and traffic safety. The power of naturalistic data is in the precise vehicle kinematic data (i.e., acceleration, velocity, and position collected at 10 Hz) and driver behavior and performance (as viewed by using continuous video). These continuously recorded high-resolution data for crashes, near crashes, and normal driving conditions allow for more sensitive analyses than does use of other crash data.

Although the cost per vehicle year of data collected is rapidly declining, the traditionally high expense of conducting these studies limits the number of participants used during data collection. To help alleviate the limitation of a small number of crashes, researchers have proposed the use of near crashes in combination with crash events (*18*). The near crash may have several analytical benefits for such analyses. First, a near crash is an event that should be avoided since, by definition, a successful, last-second evasive maneuver is needed to avoid a crash. Second, near crashes can provide unique insight into the elements and factors associated with successful crash-avoidance maneuvers for comparison to unsuccessful maneuvers or crash circumstances. Third, and the focus of this paper, near crashes, since they (by definition) have many of the same elements as a crash, may provide useful insight into the risk associated with driver behavior and environmental factors in combination with crashes. This third benefit, if it can be validated, can provide a powerful tool for analyzing naturalistic driving data since near crashes occur at a rate of roughly 10 to 15 times more frequently than crashes. Thus, there is a need to better understand the relationship between crashes and near crashes as well as the impact of use of crash surrogate measures during assessment of crash risk.

100-CAR NATURALISTIC DRIVING DATA

The 100-car study is a large-scale naturalistic driving study (*18*). The study instrumented 100 cars and recorded continuous driving data for 1 year. The study collected approximately 2 million vehicle miles

F. Guo, Virginia Tech Transportation Institute and Department of Statistics; S. G. Klauer and J. M. Hankey, Center for Automotive Safety Research, Virginia Tech Transportation Institute; and T. A. Dingus, Virginia Tech Transportation Institute, Virginia Polytechnic Institute and State University, 3500 Transportation Research Plaza, Blacksburg, VA 24061. Corresponding author: F. Guo, feng.guo@vt.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2147, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 66–74.
DOI: 10.3141/2147-09

and 43,000 h of driving data. The detailed information provides an unprecedented opportunity for evaluating the driver behaviors and factors that significantly affect traffic safety.

The data-reduction framework for the event database was developed to identify various driving-behavior and environmental characteristics for safety-relevant conflicts. The operational definitions for the pertinent severity levels are as follows:

- Crash. Any contact with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated. This includes other vehicles, roadside barriers, objects on or off the roadway, pedestrians, cyclists, and animals.
- Near crash. Any circumstance that requires a rapid, evasive maneuver by the participant vehicle, or any other vehicle, pedestrian, cyclist, or animal, to avoid a crash. A rapid, evasive maneuver is defined as steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle's capabilities.

The vehicle kinematic data were used to detect safety-relevant events, including lateral acceleration, longitudinal acceleration, critical incident button, forward time to collision (TTC), rear TTC, and yaw rate (*18*). The reduced events include 69 crashes and 761 near crashes. The 69 crashes occurred for 41 drivers, and 15 drivers had more than one crash.

The exposure variables recorded for each of these safety-relevant events were selected on the basis of previous instrumented-vehicle studies (*19, 20*), national crash databases (General Estimates System and Fatality Analysis Reporting System), and questions on the Virginia state police accident reports. For each safety event, data reductionists watched 6-s epochs for each event and reduced and recorded information regarding the nature of the event, driving behavior before the event, state of the driver, environment, and so forth. A total of four areas of data reduction were recorded for each event: vehicle variables, event variables, environmental variables, and driver state variables.

The analysis of the naturalistic study follows a case-based approach in which the exposure variables (i.e., potential contributing factors) for the crashes were compared with the exposure variables for the normal driving conditions to evaluate the impact of each factor. Guo and Hankey showed that the use of a case-cohort design provides an efficient means for evaluating risk factors by using relatively smaller sample sizes than are used in typical crash database analyses (*21*).

The case-control baseline database used in this study contains approximately 17,000 6-s segments in which the vehicle maintained a velocity of greater than 5 mph (referred to as an epoch). Kinematic triggers on driving performance data were not used to select these baseline epochs. Rather, these epochs were selected at random throughout the 12- to 13-month data collection period per vehicle. The number of baseline samples for each vehicle was proportional to the total driving hours of the vehicle.

METHODOLOGY

Evaluation Risk

The primary metric in evaluating the impact of a factor on traffic safety is the odds ratio (OR). The odds of event occurrence are defined as the probability of event occurrence divided by the probability of nonoccurrence. A contributing factor is considered to have a significant effect on safety if the odds of presence for safety events are much higher than the odds of presence in baseline events. Therefore, the ratio of the odds for safety events to the odds for baseline is a measure of the safety impact of the contributing factor. The odds ratio is defined as

$$\mathrm{OR} = \frac{\mathrm{odds}_{\mathrm{event}}}{\mathrm{odds}_{\mathrm{baseline}}} = \frac{\dfrac{p_{\mathrm{event}}}{\left(1 - p_{\mathrm{event}}\right)}}{\dfrac{p_{\mathrm{baseline}}}{\left(1 - p_{\mathrm{baseline}}\right)}}$$

where p_{event} and p_{baseline} are the probability of the presence of a contributing factor for incidents and baseline. An odds ratio with a value of 1.0 indicates a factor that has no elevated risk in the event situation above normal, baseline driving. An odds ratio of greater than 1.0 indicates that this activity increases one's relative risk of a crash or near crash by the value of the odds ratio. For example, if the odds ratio for drowsiness is 6.0, then this indicates that the odds of a driver being involved in a crash or near crash while driving drowsy is 6 times higher than if the driver is alert.

Principle of Surrogate Measures

The primary reason to use near crashes as a surrogate measure for crashes in naturalistic driving studies is that the number of crashes observed is not large enough for use in evaluating the risks involved for specific contributing factors. Surrogate measures are closely related to Heinrich's triangle (*22*). The tenet of Heinrich's triangle is that less-severe accidents happen more frequently than do severe accidents, and the frequency of severe injuries can be reduced by reducing the frequency of minor injuries. Following this philosophy, two principles are proposed for using near crashes as a surrogate measure for crashes:

- The causal mechanisms for surrogates (near crashes) and crashes are the same or similar.
- There is a strong association between the frequency of surrogate measures and crashes under different settings.

Causal Mechanism

A key requirement for using near crashes as a surrogate measure is that they possess the same causal mechanism as crashes. (The only difference between a crash and an appropriate near-crash surrogate is the severity of the safety outputs.) For example, suppose there are two potential calamities in a working environment: (*a*) a hand injury caused by operating a machine and (*b*) a fatality caused by objects falling from a roof. Although the frequency of hand injuries is higher than that of falling-object fatalities, the former is not a good surrogate measure of the latter because the two calamities are based on different causal mechanisms. In other words, making safety improvements to reduce the number of hand injuries will not reduce the risk of fatalities. In the context of naturalistic studies, the contributing factors for near crashes and crashes should be similar or identical (e.g., distractions can lead to both near crashes and crashes), and their differences should be merely of severity. Only then can near crashes be used to evaluate factors that affect traffic safety.

In naturalistic studies, a near crash is identified through the combination of several factors, including vehicle kinematic measures and visual evaluation for the severity of events. A near crash contains more information and is potentially a better surrogate measure than those based on a single metric, such as TTC. Furthermore, many factors associated with near crashes share similar characteristics with crashes. For example, the same kinematic triggers were used to detect both crashes and near crashes, and this paper indicates similar contributing factors for both crashes and near crashes. Thus it is expected that the causal mechanism for near crashes would be similar for crashes.

Quantitatively verifying the similarity of the causal factors for near crashes and crashes is challenging. First, the causal mechanisms for crashes often are different and thus probably are different for near crashes as well. Second, the causal mechanisms for crashes are complex. Rarely does a single factor cause a crash, but the combination of several factors will. Therefore, it is not appropriate to state that one particular factor causes an accident. In epidemiology, the sufficient risk set (which is defined as the set of all necessary conditions for an accident to happen) is used. A similar concept, the cut set, can be found in the fault tree analysis (*23*). The sufficient risk set or cut set in most cases is unique for each event. It is challenging to find all elements of the sufficient risk set for a particular crash by using frequency data.

The random nature of the crash and the near crash implies that the presence of a high-risk factor will not necessarily lead to a crash or near crash. Instead, the probability of a crash or near crash will increase. For example, the presence of drowsiness will not lead to a crash definitively, but a crash or near crash is more likely to happen in the presence of drowsiness. The same causal mechanism can be understood as the factor that will lead to an elevated probability of a crash will also lead to a high probability of a near crash. Because the probability can be measured by relative frequency, the frequency of exposure factors for crashes and near crashes implies a similar causal mechanism from this perspective. This paper presents a thorough examination of the contributing factors for crashes and near crashes, and this indirectly verifies the causal mechanism principle.

Frequency Relationship

The frequency of a surrogate measure should be strongly associated with the frequency of crashes. That is, if a certain number of surrogates are observed, a good estimate of the number of crashes that will occur should be obtainable. This directly corresponds to the motivation for the use of surrogate measures: instead of modeling or evaluating the relatively small number of crashes observed, one can use the relatively larger number of surrogates to get an improved assessment of traffic safety. A strong association will ensure that an analysis that uses surrogate measures will not depart significantly from the results of an analysis that uses crashes only. In this paper, an empirical study of the frequency relationship between crashes and noncrashes was conducted by using the 100-car database.

Role of Constant Crash-to-Near-Crash Ratio

The relationship between frequencies of crashes and near crashes is the key for assessing crash risk. For the case-based analysis approach (*21, 24*), a stronger assumption for the association between frequencies of crashes and near crashes is a constant crash-to-near-crash ratio condition. If the constant crash-to-near-crash ratio assumption holds, the follow results are true:

- The odds ratios for using crashes alone and combining crashes and near crashes will be identical (unbiasness) and
- The combined analysis will provide more accurate estimation (smaller variance and tighter confidence intervals).

This conclusion applies for a general situation, as shown by the following derivation. The contingency tables of safety outcomes versus a risk factor with two exposure levels for crash alone and crash and near crash combined are shown in the following table:

Safety Event	*Exposure*	*Nonexposure*
Crash only	A	B
Crash-only baseline	C	D
Crash and near crash combined	$(1+\lambda)A$	$(1+\lambda)B$
Crash and near crash baseline	C	D

The notation A through D represents the number of observation in each category, for example, A is the number of crashes for the exposure group. Assume a surrogate measure was used and the constant near-crash-to-crash ratio is λ. That is,

$$\frac{A_{\text{surrogate}}}{A} = \frac{B_{\text{surrogate}}}{B} = \lambda$$

The total observations for crash and near crash combined are $A + A_{\text{surrogate}} = (1+\lambda)A$ for exposure, and $B + B_{\text{surrogate}} = (1+\lambda)B$ for nonexposure, as shown in the table.

The odds ratio for the risk factor with crash alone is

$$\text{OR} = \frac{A * D}{B * C}$$

The variance of the estimation is given by

$$\text{Var}\left(\log\left(\text{OR}\right)\right) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

The 95% confidence interval is

$$\left(\text{OR} \cdot \exp\left(-z\sqrt{v}\right) \qquad \text{OR} \cdot \exp\left(z\sqrt{v}\right)\right)$$

where z is the $100(1-\alpha/2)$ percentile of the standard normal distribution. For the 95% confidence interval, $z = 1.96$. The effect of a risk factor is statistically significant if the corresponding confidence interval does not include the neutral value 1. The length of the confidence interval is

$$\text{OR}\left(\exp\left(z\sqrt{v}\right) - \exp\left(-z\sqrt{v}\right)\right)$$

The odds ratio of combined data using both crashes and surrogates, denoted by OR′, is equal to that of using crashes alone, as shown in the following simple derivation:

$$\text{OR}' = \frac{A(1+\lambda) * D}{B(1+\lambda) * C} = \frac{A * D}{B * C} = \text{OR}$$

The variance of the estimation using crashes and surrogates combined, denoted by v', is smaller than that of using crash alone:

$$v' = \mathrm{Var}(\log(\mathrm{OR})) = \frac{1}{A(1+\lambda)} + \frac{1}{B(1+\lambda)} + \frac{1}{C} + \frac{1}{D}$$

$$< \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} = v \text{ for all } \lambda > 0$$

Thus the corresponding length of the confidence interval is smaller:

$$\mathrm{OR} \cdot \left\{\exp\left(z\sqrt{v'}\right) - \exp\left(-z\sqrt{v'}\right)\right\} < \mathrm{OR} \cdot \left\{\exp\left(z\sqrt{v}\right) - \exp\left(-z\sqrt{v}\right)\right\}$$

These resulting three equations indicate that a constant ratio will lead to identical point estimations for the odds ratios but shorter confidence intervals (i.e., a more precise estimation) given a positive ratio $\lambda > 0$. Furthermore, the reduction in length of confidence interval depends on factor λ; a larger λ will lead to a more precise estimation.

ANALYSIS

The relationship between crashes and near crashes is thoroughly evaluated in this section. The analysis follows two principles for crash surrogates: whether the causal mechanism differs substantially and whether there is a strong relation between the frequencies of crashes and near crashes. The analyses include three major parts: the sequential factor analysis, frequency analysis, and sensitivity analysis.

Sequential Factor Analysis

For each crash and near crash, a sequence of factors was recorded to capture not only the vehicle trajectory but also the relevant driver behavior and driver response to the situation. This sequence of behaviors and factors is as shown in Figure 1. The preincident maneuver captures the driver's action just before the beginning of the event. The precipitating factor is the action by the 100-car study driver or another driver in the near vicinity that begins the sequence of factors. The evasive maneuver is the action, typically by the 100-car study driver, that is performed either successfully (resulting in a near crash) or unsuccessfully (resulting in a crash). Contributing factors include particular driver behaviors (e.g., driver drowsiness), environmental states (e.g., icy road conditions), or mechanical problems with the vehicle (e.g., flat tire).

Evasive Maneuver

The evasive maneuver is defined as the participant driver's reaction in response to the precipitating factor. This is independent of maneuvers associated with the resulting crash or near crash. The analysis was conducted for all events and for single-vehicle, lead-vehicle, and following-vehicle conflicts separately. The types of conflicts were defined on the basis of the relationship between the 100-car vehicle and the vehicles in close vicinity. The definitions are as follows:

- Conflict with a lead vehicle: Interaction with a vehicle in front of the subject vehicle (traveling in the same direction as the subject vehicle or stopped);
- Conflict with a following vehicle: Interaction with a vehicle behind the subject vehicle (traveling in the same direction as the subject vehicle); and
- Single-vehicle conflict: Any nonmotor vehicle conflict occurring on or off the roadway not described in other categories.

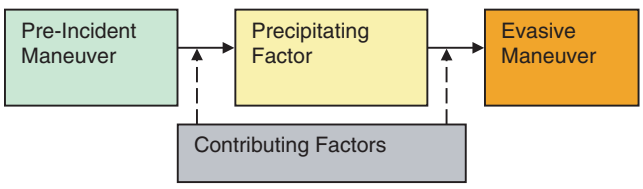

FIGURE 1 Sequential factors of crash or near crash.

Table 1 lists the reactions of drivers before each crash and near crash. It can be seen that 45/(23 + 45) = 66% drivers reacted before a crash. In contrast, in near crashes, about 723/(37 + 723) = 95% drivers reacted. The statistical chi-square test for equal percentage is highly significant, supporting the observed unequal percentages. The lead-vehicle conflict contains the most obvious patterns. There were 377 near crashes and five crashes in which the driver reacted. In contrast, there are nine crashes in which the drivers failed to react. The message is clear: in the conflict with a leading vehicle, failure to react will dramatically increase the risk of crash (in this case, all nine events with no reaction led to crashes). The results from following-vehicle and single-vehicle conflicts are not conclusive. The statistical chi-square test for equal probability is not significant, in part because of the relatively small sample size.

The significant difference in driver reaction for crashes and near crashes implies that driver response is critical in distinguishing between these two types of events. However, this difference shall not be considered as evidence against the identical causal mechanism. The causal mechanism in this study is considered as the risk factors that trigger the safety events, not a driver's last response to avoid a crash. A crash and a near crash can have the same causal mechanism but a different safety outcome because of the evasive maneuver.

Number of Contributing Factors

The analysis for number of contributing factors is motivated by the hypothesis that crashes can be caused by the combination of several

TABLE 1 Driver Reaction to Crash and Near Crash

Event Type	Driver Reaction	Crash	Near Crash	*p*-Value
All events	Reaction	45 (66%)	723	≤.0001
	No reaction	23	37	
			95%	
Lead-vehicle conflict	Reaction	5 (36%)	377	<.0001
	No reaction	9	0	
			100%	
Following-vehicle conflict	Reaction	5 (42%)	49	.0558
	No reaction	7	21	
			70%	
Single-vehicle conflict	Reaction	20 (91%)	47	.1789
	No reaction	2	1	
			98%	

factors. Thus the number of contributing factors represents the level of driving burden. If crashes are associated with a higher driving burden than near crashes, the number of factors for crashes should be higher than that for near crashes. Table 2 displays the frequencies with which the following six factors occurred for the three main conflict types and the aggregate of all conflict types:

- Distraction,
- Surface conditions,
- Traffic density,
- Lighting,
- Weather, and
- Visual obstruction.

The statistical hypothesis being tested is that the mean number of contributing factors is equivalent for crashes and near crashes. A standard *t*-test was used to test the difference. The mean number of contributing factors for each level is given in the table. As can be seen, the only statistically significant result is for conflict with lead vehicle, for which the mean number of contributing factors for crashes (2.93) is higher than that for near crashes (2.27). This implies that for this type of conflict, the presence of more contributing factors will more likely result in crashes.

Frequency Relationship Between Crashes and Near Crashes

Regression Analysis by Driver

The relationship of driver involvement in crashes versus near crashes was analyzed for each individual driver; 234 drivers were used in the analysis. A strong relationship would indicate that the number of crashes a driver would be involved in is predictable based on the driver's frequency of involvement in near crashes. As the number of crashes is a count variable, a Poisson regression model was fitted by using the number of near crashes as covariates. The model setup is as follows:

$$y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

where

y_i = number of crashes,
λ_i = expected number of crashes,
x_i = number of near crashes for driver i, and
β_0, β_1 = regression parameters.

The model fitting parameters are $\widehat{\beta_0} = -2.31$ [standard error (SE) $= (0.25, p\text{-value} < .001]$ and $\widehat{\beta_1} = 0.21$ (SE = 0.25, p-value < .001). The coefficient for near crashes is highly significant, and for every additional near crash a driver is involved in, the frequency of crash involvement will increase by a factor of exp(0.21) = 1.23, which confirms the positive relationship between the frequency of crash and near crash.

Frequency Relationship

The number of crashes versus near crashes for six environmental factors—weather condition, road surface, age group, traffic condition, lighting condition, and road alignment—was examined, and the results are shown in Figure 2. The present analysis generally indicates a positive correlation between the number of crashes and the number of near crashes. Categories with more near crashes tended also to have more crashes. The constant crash-to-near-crash ratio is a stronger condition and was satisfied only when a straight trend line crossed the origin of axis. The analyses indicated that the ratio depended on the context of the observed scenario. For example, the apparent curve patterns found for traffic conditions by the level of service indicated that the crash-to-near-crash ratios change as a function of traffic density. Also, there is a trend that the crash-to-near-crash ratio for poor surface conditions is higher than the ratio for normal driving conditions. A possible reason is that in poor driving conditions, the tolerance for driver mistakes or other risks is relatively low, which will lead to crashes that can be avoided under normal driving conditions. Although the changes in these ratios are easily explained from an engineering point of view, they do suggest that the crash-to-near-crash ratios are not stable. A direct consequence of the nonconstant ratio is that the risk estimation that uses near crash will be biased according to the result in the section on the role of constant crash-to-near-crash ratio.

TABLE 2 Contributing Factors for All Event Types

	All Events		Lead-Vehicle Conflict		Following-Vehicle Conflict		Single-Vehicle Conflict	
No. of Factors	Crash	Near Crash	Crash	Near Crash	Crash	Near Crash	Crash	Near Crash
0	3	25	0	5	0	3	1	3
1	21	183	1	76	2	15	13	18
2	25	311	6	163	7	29	6	22
3	11	161	4	88	2	15	3	2
4	6	62	2	36	1	7	1	4
5	1	14	1	6	0	1	0	0
6	1	4	1	2	0	0	0	0
Mean	2.04	2.14	2.93	2.27	2.17	2.16	1.58	1.71

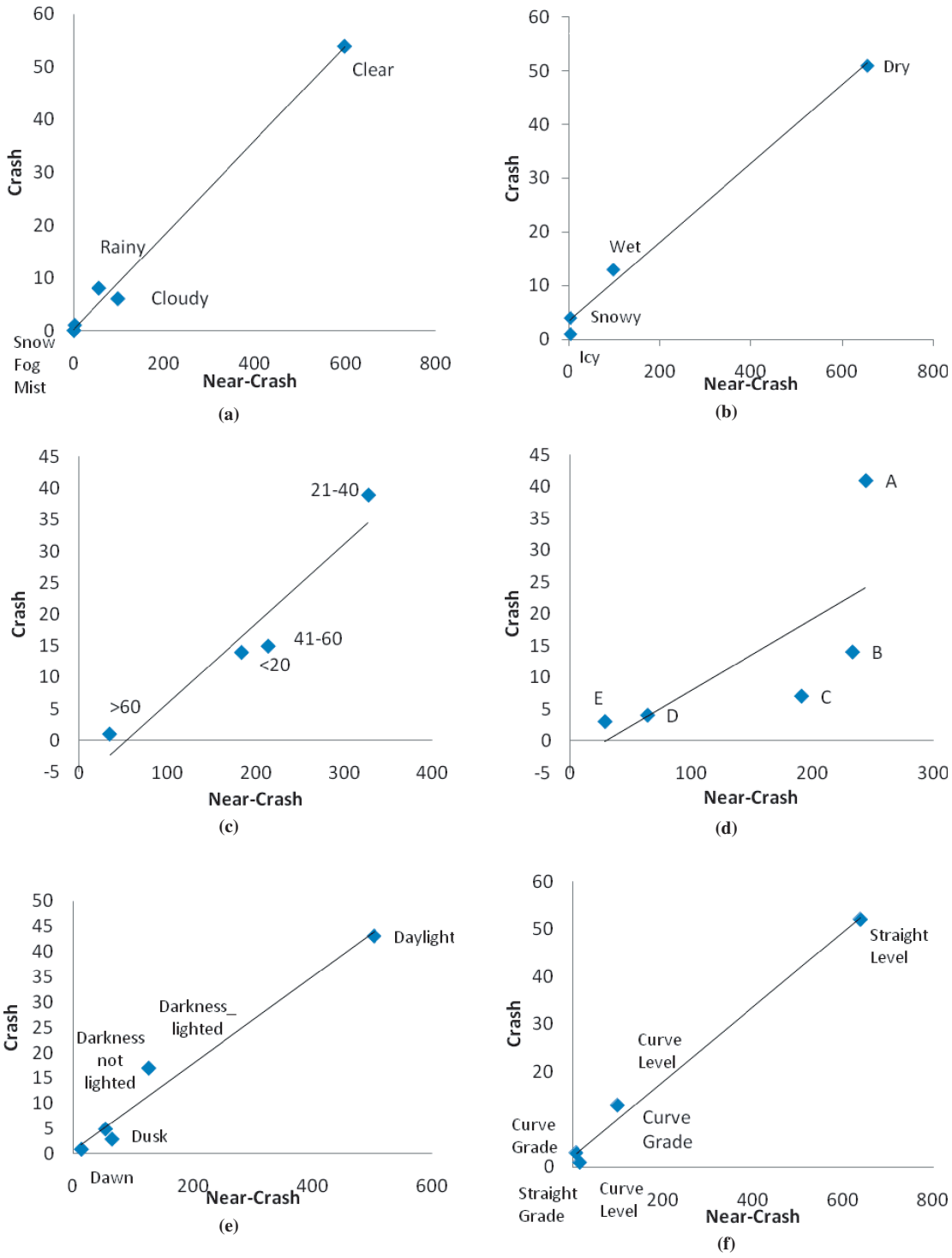

FIGURE 2 Frequency relationship between crash and near crash for (*a*) weather conditions, (*b*) road surface, (*c*) age (in years), (*d*) traffic conditions (level of service), (*e*) lighting conditions, and (*f*) road alignment.

Sensitivity Analysis

The bias and precision of the odds ratio estimate are the most critical criteria in quantitative evaluation of risk factors. Precision is directly related to the sample size and, thus, improved precision can be obtained by combining surrogates with crashes into any analysis. The degree of bias, introduced by combining the surrogate measure with crashes, is thus the key to assessing a proper crash surrogate and is directly related to the validity of these analyses.

It is reasonable to assume that risk estimation (OR) based on crashes alone will lead to the correct or unbiased estimation. As proved before, the unbiased estimation can be achieved only for

TABLE 3 Sensitivity Analysis

	Safety Events	Odds Ratio	*p*-Value	95% Confidence Limits	
Distraction	Crash only	5.58	<.001	2.23	13.98
	Crash & near crash	4.29	<.001	3.11	5.91
Lighting condition	Crash only	1.40	.173	0.86	2.28
	Crash & near crash	1.21	.01	1.04	1.4
Surface condition	Crash only	3.11	<.001	1.81	5.33
	Crash & near crash	1.56	<.001	1.28	1.9
Weather condition	Crash only	1.79	.097	0.89	3.63
	Crash & near crash	1.12	.364	0.87	1.44
Drowsiness	Crash only	7.12	<.001	3.94	12.87
	Crash & near crash	4.32	<.001	3.48	5.36

a constant crash-to-near-crash ratio, which is rarely true for real data, as indicated by the frequency relationship analysis. It is expected that a certain level of bias will be introduced by use of near crashes as surrogates. The primary question is whether the benefits of the combined analysis outweigh whatever bias may exist. The analyses presented here thus focus on evaluating the magnitude and direction of the potential bias, and this is conducted through a sensitivity analysis.

The sensitivity analysis is commonly used to evaluate the impact of invalid model assumptions on risk evaluation. In the context of this study, the focus is the magnitude and direction of bias caused by an invalid assumption, that is, no constant ratio. The sensitivity analysis compares the risk estimated by using crashes alone with the risk estimated by using crashes and near crashes combined. This comparison indicates the impact (i.e., bias introduced) of using near crashes as a surrogate measure. The data used in this analysis were the 69 crashes, the 761 near crashes, and 17,344 randomly sampled baseline data.

Five potential risk factors were evaluated. The results of the sensitivity analysis are given in Table 3 and illustrated in Figure 3. As can be seen, the sensitivity analysis produced very consistent patterns: the point estimation for odds ratios using combined data

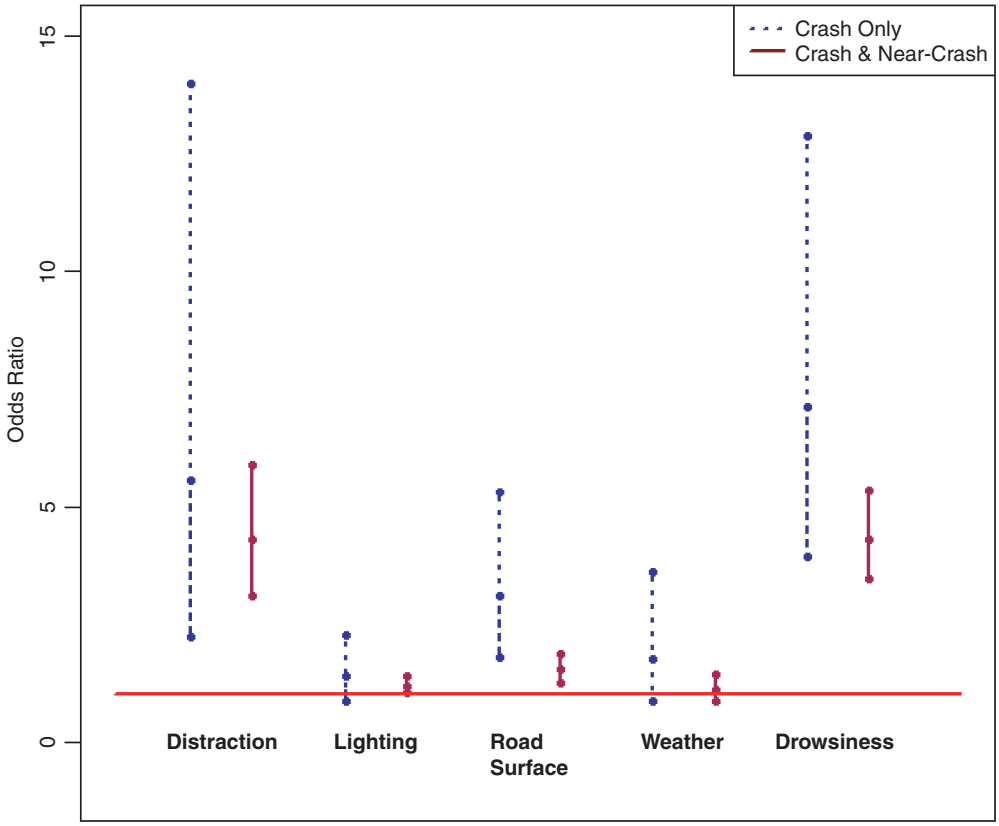

FIGURE 3 Risk output for sensitivity analysis.

was always smaller than for using crash data alone. The precision of the estimator, as measured by the length of the confidence intervals, is always better than that of using crashes alone. The consistency of the results has a significant implication: use of surrogate measures tends to provide conservative risk estimates, yet with statistically significant test results. Therefore, a significant risk factor identified by using near-crash surrogates will be at least as dangerous as the analysis indicated with crashes alone. The estimated odds ratio can be considered as a lower bound of the mean of "true" odds ratios by using crashes alone (if there are sufficient data). This suggests that assessing the risk of various contributing factors by using near crashes as surrogates will provide conservative results as compared with calculating risk by using crashes alone.

SUMMARY AND DISCUSSION

Naturalistic driving studies can significantly improve understanding of traffic safety from several perspectives that cannot be sufficiently evaluated by using other methods. Although the naturalistic crash data sample size is small, from a crash database perspective there is very detailed information about the driver, vehicle, traffic, and environment; in addition, there is precise information regarding the timing of factors (e.g., distraction). This detailed information surrounding naturalistic data has the potential to provide great insights into improving driver safety, and it demands that novel approaches to data analysis and modeling be developed. The definition of a near crash combines various kinematic and environmental factors; thus, near crashes provide more information than the single kinematic-measure-based surrogates. Use of near crashes as a surrogate metric for crashes can both increase the sample size and increase the amount of information collected from the data. This analysis focused on the validity of the approach of combining the surrogates with actual crashes.

There is no debate that crashes and near crashes are two different types of events. Not only is this true by operational definition, but several results in this report demonstrate that the two cannot be identical. However, this does not eliminate the use of near crashes as valid surrogates for crashes. Two principles, namely, the identical causal mechanism and the constant crash-to-near-crash ratio, were proposed and discussed.

Three analyses were conducted to assess these two principles:

- Sequential factor analysis,
- Frequency relationship of the presence of behaviors between crashes and near crashes, and
- Sensitivity analysis.

The following are the primary results from each of these analyses:

- No significant differences were found in the number of contributing factors present for both crashes and near crashes, which suggests that for both types of events, the driver is clearly in a complex situation.
- The crash-to-near-crash ratio analysis indicates that there is a positive relationship between crashes and near crashes. However, as a much stronger condition, the crash-to-near-crash ratio is highly scenario dependent and not constant in general.
- The sensitivity analysis indicated that when combined crash and near-crash data are used, the precision of the estimation will increase as measured by the shorter length of the confidence interval surrounding the point estimate. This is expected because of the increased sample size. An interesting result relates to the pattern in the bias of odds ratio: the odds ratios from combined analyses are consistently smaller (thus more conservative) than when crashes alone are used. This result has significant implications in the analysis of naturalistic data. The combined analysis will provide a conservative odds ratio point estimate but may indicate a statistically significant effect. Therefore, the risk factors identified through combined crash and near-crash data appear to provide a more conservative estimate but provide enough data with which to identify significant effects, that is, the truly risky behaviors. The other analyses provide support that these data can be combined to provide better estimates of the true effect (although it is likely conservative).

The relationship between crashes and near crashes is complex, and there are no simple or absolute criteria with which to accurately predict crashes by using near crashes or vice versa. This study provides a general procedure for evaluating the appropriateness of use of near crashes, as well as other safety outcomes, as a surrogate for crashes. The empirical study that used 100-car data indicates two main conclusions: (*a*) use of near crashes will result in conservative risk estimates, and (*b*) use of near crashes as surrogates can significantly improve the precision of the estimation. This trade-off between bias and precision can be found in many statistical estimation problems. For small-scale studies with low numbers of crashes, use of near crashes as a surrogate measure is informative and will point transportation researchers in the right direction when assessing crash and near-crash risk.

ACKNOWLEDGMENTS

This study was supported by a NHTSA grant. The authors thank Eric Traube and Mike Perel of NHTSA and John Farbry of Science Applications International Corporation for their advice and support throughout this project.

REFERENCES

1. Perkins, S. R., and J. I. Harris. Traffic Conflict Characteristics—Accident Potential at Intersections. In *Highway Research Record 225,* HRB, National Research Council, Washington, D.C., 1968, pp. 35–43.
2. Parker, M. R., Jr., and C. V. Zegeer. *Traffic Conflict Techniques for Safety and Operations: Observers Manual.* FHWA, U.S. Department of Transportation, 1989.
3. Williams, M. J. Validity of the Traffic Conflicts Technique. *Accident Analysis and Prevention,* Vol. 13, No. 2, 1981, pp. 133–145.
4. Hauer, E. Traffic Conflicts and Exposure. *Accident Analysis and Prevention,* Vol. 14, No. 5, 1982, pp. 359–364.
5. Migletz, D. J., W. D. Glauz, and K. M. Bauer. *Relationships Between Traffic Conflicts and Accidents.* FHWA-RD-84-042. FHWA, U.S. Department of Transportation, 1985.
6. Hauer, E., and P. Garder. Research into the Validity of the Traffic Conflicts Technique. *Accident Analysis and Prevention,* Vol. 18, No. 6, 1986, pp. 471–481.
7. Glauz, W. D., K. M. Bauer, and D. J. Migletz. Expected Traffic Conflict Rates and Their Use in Predicting Accidents. In *Transportation Research Record 1026,* TRB, National Research Council, Washington, D.C., 1985, pp. 1–12.

8. Sayed, T., and S. Zein. Traffic Conflict Standards for Intersections. *Transportation Planning and Technology,* Vol. 22, 1999, pp. 309–323.
9. Rao, V. T., and V. R. Rengaraju. Modeling Conflicts of Heterogeneous Traffic at Urban Uncontrolled Intersections. *Journal of Transportation Engineering,* Vol. 124, No. 1, 1998, pp. 23–34.
10. Retting, R. A., and M. A. Greene. Influence of Traffic Signal Timing on Red-Light Running and Potential Vehicle Conflicts at Urban Intersections. In *Transportation Research Record 1595,* TRB, National Research Council, Washington, D.C., 1997, pp. 1–7.
11. Tiwari, G., D. Mohan, and J. Fazio. Conflict Analysis for Prediction of Fatal Crash Locations in Mixed Traffic Streams. *Accident Analysis and Prevention,* Vol. 30, No. 2, 1998, pp. 207–215.
12. Fazio, J., J. Holden, and N. M. Rouphail. Use of Freeway Conflict Rates as an Alternative to Crash Rates in Weaving Section Safety Analyses. In *Transportation Research Record 1401,* TRB, National Research Council, Washington, D.C., 1993, pp. 61–69.
13. Gettman, D., and L. Head. Surrogate Safety Measures from Traffic Simulation Models. FHWA-RD-03-050. FHWA, U.S. Department of Transportation, 2003.
14. Gettman, D., L. Pu, T. Sayed, and S. Shelby. Surrogate Safety Assessment Model and Validation. FHWA, U.S. Department of Transportation, 2008.
15. Songchitruksa, P., and A. Tarko. The Extreme Value Theory Approach to Safety Estimation. *Accident Analysis and Prevention,* Vol. 28, 2006, pp. 811–822.
16. Davis, G., J. Hourdos, and H. Xiong. Outline of a Causal Theory of Traffic Conflicts and Collisions. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C., 2008.
17. Saunier, N., and T. A. Sayed. Probabilistic Framework for the Automated Analysis of Exposure to Road Collision. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2083,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 96–104.
18. Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. *The 100-Car Naturalistic Driving Study: Phase II—Results of the 100-Car Field Experiment.* DOT HS 810 593. NHTSA, 2006.
19. Hanowski, R. J., W. W. Wierwille, S. A. Garness, and T. A. Dingus. *The Impact of Local/Short Haul Operations on Driver Fatigue.* NHTSA, 2000.
20. Dingus, T., V. Neale, S. Garness, R. Hanowski, A. Keisler, S. Lee, M. Perez, G. Robinson, S. Belz, J. Casali, E. Pace-Schott, R. Stickgold, and J. Hobson. *Impact of Sleeper Berth Usage on Driver Fatigue.* Federal Motor Carrier Safety Administration, Washington, D.C., 2001.
21. Guo, F., and J. Hankey. *Modeling 100-Car Safety Events: A Case-Based Approach for Analyzing Naturalistic Driving Data.* National Surface Transportation Safety Center for Excellence, Blacksburg, Va., 2009.
22. Heinrich, H. W. *Industrial Accidents Prevention.* McGraw-Hill, New York, 1959.
23. Roland, H. E., and B. Moriarty. *System Safety Engineering and Management,* 2nd ed. John Wiley and Sons, New York, 1990.
24. Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. *The Impact of Driver Inattention On Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data.* DOT HS 810 594. NHTSA, 2006.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.