Analysis of Crash Rates and Surrogate Events

Unified Approach

Tim J. Gordon, Lidia P. Kostyniuk, Paul E. Green, Michelle A. Barnes, Daniel Blower, Adam D. Blankespoor, and Scott E. Bogard

A preliminary study was done into the use and validation of crash surrogates, which are obtained from naturalistic driving studies for the detailed analysis of risk factors. The approach is based on a unified statistical analysis of crash data and surrogate events that uses a spatial referencing system and a common measure of exposure. Statistical methods based on a bivariate response and Bayesian update models were adapted to the joint analysis of crashes and surrogates. The study specifically addresses road-departure crashes involving a single vehicle. It is proposed that suitable surrogates be based on underlying continuous measures of disturbance in the driver's lateral control of the vehicle. Naturalistic driving data from a field operational test conducted in southeastern Michigan were spatially joined with highway data and crash data from the same area, and a set of candidate crash surrogates was tested. Analysis results indicated that simple lateral lane position did not provide a satisfactory surrogate, whereas estimated time to road departure was found to show the correct statistical dependencies, consistent with the crash data. The approach developed in the study provided a way to assess crash risk in a common framework and also to validate or invalidate candidate surrogates. When applied to data from the future SHRP 2 naturalistic driving study, the increased statistical power resulting from the much larger data set will provide more definitive conclusions about surrogate validity and factors influencing overall crash risk.

For decades, crashes have been studied as discrete events focusing on the circumstances of the crash. This type of analysis has been used to identify the characteristics of highway features associated with higher crash experience (*1–3*), but other factors, such as traffic volumes, driver characteristics, land use, and environmental conditions, were also needed to explain or describe crash events (*4–6*). Furthermore, cross-sectional analyses of crash events did not address circumstances leading to a crash.

Crashes are rare events, and there are conditions in which a crash, although likely, does not occur. Thus, a crash can be considered a high-probability event given a set of conditions. However, because crashes are rare, it is difficult to accumulate enough of them for timely study. Accordingly, it is desirable to identify crash surrogates, which are related to crash risk but which occur more frequently than crashes. Traffic conflicts have been used as crash surrogates (*7*), but the relationship between conflicts and crashes is not yet fully understood. Advances in vehicle instrumentation technology have made it possible to collect longitudinal naturalistic data about the vehicle, driver, and roadway, accumulating information about events preceding crashes if they occur, or near crashes and other events that could be associated with crash risk. Such data provide the opportunity to identify crash surrogates. For example, a recent study conducted by researchers at Virginia Polytechnic Institute and State University looked at near crashes in naturalistic driving data (NDD) as crash surrogates (*8*).

This paper summarizes work conducted as part of SHRP 2 to provide a validated quantitative link between detailed measurements of naturalistic driving behavior, road-departure crashes, and road segment characteristics (*9*). Because no such link had existed, this study developed appropriate analysis methods that could associate exposure-based crash risk with quantitative metrics (crash surrogates) available from NDD. When applied to results of the future SHRP 2 naturalistic driving study, the methods should provide quantitative relationships between driving and crash risk, provide validated surrogates for these types of crashes, and develop new understanding of risk factors that can be used to improve highway safety. The present work is exploratory and uses existing driving data from Southeast Michigan to develop initial statistical models and formulate appropriate metrics.

The crash problem addressed in this research is that of road-departure crashes of single vehicles, and the study is based on the idea that the underlying mechanisms leading to such crashes are the same as those that create variations in normal driving—especially those involving "disturbed" lane-keeping control. Disturbed control is any interruption or delay in the process of perception, recognition, judgment or decision, or action. Specifically, the road-departure crash problem was formulated according to the following set of hypotheses:

- Single-vehicle road-departure crashes occur only under conditions of disturbed control.
- NDD contain measurable episodes of disturbed control.
- Crash surrogates exist and are based on a combination of objective measures of disturbed control, highway geometric factors, and off-highway factors.
- Crash surrogates can be related to actual crashes.

Transportation Research Institute, University of Michigan, 2901 Baxter Road, Ann Arbor, MI 48109-2150. Corresponding author: L. P. Kostyniuk, lidakost@umich.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2237, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 1–9.
DOI: 10.3141/2237-01

DATA

Data for this study were developed from an NDD resource obtained in a field operation test (FOT) conducted in southeast Michigan, referenced in a common spatial framework with highway information and crash data from the same region.

Naturalistic Driving Data

NDD come from the *Road Departure Crash Warning System Field Operational Test*, which tested a combined lane-departure and curve speed warning system (*10*). The FOT exposed a fleet of 11 Nissan Altima cars equipped with a road-departure crash warning system to 10 months of naturalistic driving by 78 drivers, randomly selected and recruited from Michigan driver licensing records. A total of 9,582 trips covered 133,290 km (82,773 mi) in the NDD.

Data gathered included more than 400 data signals: video of the forward driving scene and driver's face; differential Global Positioning System time and position; lane tracking including boundary type (solid, dashed), forward and side radar returns, distance to lane edge, and available maneuvering distance; vehicle velocity, yaw, pitch, and roll; and data on lights and windshield wipers. The data file (raw and derived) is approximately 250 GB and is stored on a Sequel server database.

Highway Data

The Enhanced Highway Performance Maintenance System (HPMS) for Michigan in 2005 is the main source of highway data. HPMS contains administrative and extent-of-system information on all public roads. It also contains descriptive information in a mix of universe and sample data for the arterial and collector functional systems, and areawide summary information for urbanized, small urban, and rural areas. The road system is divided into individual segments with unique identifiers in a linear referencing system, which allows them to be located spatially and joined to geospatial databases. Information on road type and rural or urban designation is available for all segments, and number of through lanes is available for all segments except those on minor collectors and local streets. Information on curves, grades, shoulder and median types and widths, and traffic information [annual average daily traffic (AADT), speed limits, peak capacity] are available only for a sample of segments.

Crash Data

Michigan police-reported crash data for 2001 to 2005 were used to identify road-departure crashes for analysis. Virtually all crashes in Michigan are geolocated with latitude and longitude coordinates.

Analysis Data

Geographic information system (GIS) software tools were used to spatially join the NDD to the highway and crash data from eight counties in Southeastern Michigan (*11*). A spatial base map of Michigan provided the key layer for the GIS, on which a digital map for all public roads in Michigan was overlaid. Only HPMS segments that were also in the NDD were included. Because traffic volumes were needed for exposure, HPMS segments without AADT information were excluded. Examination of the segments without volume information showed them to be minor collectors and local streets.

Information on road type, rural or urban designation, and types and widths of shoulders (when available) was taken directly from HPMS data. Otherwise, shoulder variables were assigned the median value obtained from the HPMS sample road segments of the same functional class, with the same number of through lanes, in the same county, reflecting traffic engineering practices of the locality.

HPMS segments are not directional, that is, the data are for both directions of travel if the road is two-way. Because directions of curves and road departures are relevant in analyses of road departures, directional road segments were developed on the basis of the descriptors of the HPMS segment and direction of travel. (Although it is possible that the close spatial proximity of the opposite sides of the same segment of road could cause unmodeled correlations, the two travel directions were considered to be correlated only via the coincidence of the explanatory highway variables used.) Thus the basic unit of analysis is the directional road segment for which traffic volume information was provided and that had been traversed at least once in the FOT. The analysis database contained 9,526 directional road segments.

Collecting information on horizontal curvature for the directional segments was a challenge. Horizontal curves in sample HPMS segments are classified into six ranges of curvature and by the total length of curves in each range. There was no meaningful way to summarize curve information for a segment and no credible way to impute curve information for segments not in the HPMS sample. It was possible to obtain the degree of curvature and length of horizontal curves from the vehicles' path and yaw rate, but the procedure was labor intensive and did not solve the problem of what to do if there was more than one curve in the segment. For the present study, the yaw rates in the NDD for the segment were used to define a variable indicating the presence or absence of at least one curve in the segment. It was clear from this experience that a different way of defining road segments (perhaps with one curve per segment) for future studies is needed.

The crash data were joined to the directional segments. Of the 71,308 road-departure crashes in the crash data file, 21,340 were on the directional segments in the analysis database. Of these crashes, 7,562 departed the road to the right, 4,372 to the left; in 9,406 cases, the direction of the departure was unknown.

Exposure measures for each directional segment were developed. Exposure for crashes was based on the volume of vehicles entering the segment in 5 years and on segment length. Exposure for surrogate events was based on the number of traversals of instrumented vehicles and segment length. Roadside information could not be coded into the analysis database, which precluded including roadside features in the analyses.

STATISTICAL METHODS

Various methodologies have been used to investigate associations between crashes and risk factors. Depending on the application under investigation, Poisson, negative binomial, random effects, and hierarchical Bayesian data models have been used to analyze data collected from historical crash databases. The response variable under these models typically is the number of crashes that can be cross classified into contingency tables according to explanatory factors hypothesized to be associated with the response variable. Because

the goal for this study was to develop exposure-based risk measures to determine if certain surrogates were good surrogates for crashes, inclusion of the relationship between surrogates and risk factors was desired. Surrogate measures of crashes would come from NDD, which rarely provide sufficient data from actual crashes, and data for crash outcomes would be derived from historical crash databases. Therefore, it is necessary to consider attributes of crash data and NDD simultaneously to provide a link between crashes and crash surrogates in a unifying framework that accounts for and possibly adjusts for inherent differences in the types of variables available from the two sources of data.

Instead of two separate models, one for crashes and one for a surrogate, a model based on the method of seemingly unrelated regressions (SUR) was used; proposed by Zellner in 1962 (*12*), it allows both models to be fit in a unifying framework. A log relative risk (RR) difference was derived to serve as the exposure-based risk measure to determine if there were significant differences in RR for crash and surrogate events. The hypothesis is that these two measures may be similar for crashes and their surrogates. If the log RR difference is zero, the conclusion is that the surrogate is a good candidate for crash risk. The validation of the surrogate is based on the comparison of the relative risks, that is, on the ratios of the rates of crashes and surrogates, not on the rates themselves.

SUR Model

SUR is developed in a normal theory framework and incorporates a correlation structure between crashes and crash surrogates. It allows formal tests of hypotheses to determine whether the risks associated with explanatory factors, or, more important, subsets of explanatory factors, are the same or different for crashes and crash surrogates.

The model takes the form

$$Y_1 = X_1\beta_1 + \epsilon_1$$

and

$$Y_2 = X_2\beta_2 + \epsilon_2 \tag{1}$$

where subscript 1 refers to the crash model and subscript 2 refers to the surrogate model. The equations resemble ordinary regression equations where Y_1 and Y_2 are the response variables, X_1 and X_2 are data matrices of explanatory variables, β_1 and β_2 are regression parameters, and ϵ_1 and ϵ_2 are error terms with normal distributions. In the SUR framework, the crash data are stacked on top of the surrogate data to form a system of equations:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{2}$$

The matrices of explanatory variables are not required to be the same, for either variables or dimension. Therefore, variables collected from NDD can be different from those collected from crash data. Because crash data are stacked on top of surrogate data, the system of equations satisfies a linear model of the form

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \Sigma) \tag{3}$$

where

$$\mathrm{Var}(\epsilon) = \Sigma = \begin{bmatrix} \sigma_{11}I & \sigma_{12}I \\ \sigma_{21}I & \sigma_{22}I \end{bmatrix} \tag{4}$$

and I are identity matrices. Suppose Y_1 has dimension $N_1 \times 1$ and Y_2 has dimension $N_2 \times 1$ so that Y has length $N_1 + N_2 = N$. Then, the matrix Σ has dimension $N \times N$. Since this model satisfies the properties of a linear model with a defined covariance matrix, the parameters can be estimated by weighted least squares (WLS).

$$\hat{\beta} = \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}Y \tag{5}$$

The parameters σ_{11} and σ_{22} represent the variances in the crash and surrogate regressions, respectively. The parameter $\sigma_{12} = \sigma_{21}$ is the covariance between Y_1 and Y_2. These parameters are estimated by fitting separate independent regressions for the crash data and the surrogate data and by using the usual residual sum of squares for the variances and the sum of the residual cross-product terms for the covariance.

The utility of this unifying framework is that tests of hypotheses of the form $H_0 : \beta_1 = \beta_2$ can be conducted with the usual F-test in a regression setting. This hypothesis tests whether crash model parameters equal surrogate model parameters. It is possible to test whether only certain crash model parameters equal certain surrogate model parameters. This point is important to the application of this framework to the simultaneous modeling of crashes and surrogates, because in many cases only a subset of the variables will be common to both.

Poisson Log-Linear Models Estimated with WLS

The Poisson log-linear model is the standard model for the analysis of rates. However, this model has limited use in practice because for a Poisson random variable, the mean is restricted to equal the variance. This has caused researchers to consider more flexible models, such as negative binomial, generalized linear mixed models, or Bayesian models.

It is well known that WLS can be used to estimate maximum likelihood parameters in Poisson log-linear models (*13*). Therefore, the SUR framework can be used to estimate parameters in a log-linear model since parameters in an SUR model can be estimated with WLS. The WLS solution depends on asymptotic theory, so the only restriction is that the data are not too sparse. Use of normal theory on which the SUR model is based to estimate parameters in a log-linear model requires a square-root transformation of the data (*14*). In particular, the dependent and independent variables are

$$Y' = \sqrt{Y}\log(Y) \qquad X' = \sqrt{Y}X \tag{6}$$

WLS with covariance matrix Σ is used to regress the variable Y' on X'. This model does not suffer from the restrictions of the Poisson model. That is, in addition to the mean parameter, the normal model has two parameters for variance and one parameter for the covariance. Therefore, it can handle the extravariability or overdispersion often encountered in observational studies that the standard Poisson model cannot.

Bayesian SUR for Log-Linear Models

Now that the model is set up in the context of a normal theory linear model, the extension to a Bayesian model is straightforward. Methods for Bayesian data analysis of normal regression models are well developed (*15*). A likelihood function for the data and a prior distribution for the parameters must be specified. The likelihood function and prior distributions are as follows:

Likelihood:

$$Y' \mid \mu, \Sigma \sim N(\mu, \Sigma) \qquad \Sigma \text{ fixed}$$

Prior 1:

$$\mu_i \mid \beta, \tau \sim N(\lambda_i, \tau) \qquad i = 1, \ldots, N \qquad \lambda_i = \beta_0 x'_{i0} + \beta_1 x'_{i1} + \cdots + \beta_p x'_{ip}$$

where N is the number of observations.

Prior 2:

$$\beta_j \sim N(0, 10^6) \qquad j = 0, \ldots, p \qquad 1/\tau \sim \text{Gamma}(0.001, 0.001)$$

where P is the number of explanatory variables.

In the likelihood, the matrix Σ is assumed fixed and contains two parameters for variance, σ_{11} and σ_{22}, and one parameter, σ_{12}, for covariance. These parameters are estimated with the residual sum of squares and residual sum of cross-product terms from ordinary independent regression models fit to crash and surrogate data, respectively. The regression model equation is incorporated in the first prior as the mean of a normal distribution and is designated by λ, which is a linear combination of the regression parameters β and the explanatory variables X. The second prior is proper and takes a standard noninformative prior. Use of proper priors ensures propriety of posterior distributions.

Estimation proceeds with Markov chain Monte Carlo simulation, which is used to generate random variables from the posterior distributions of the parameters μ, β, and τ. Because calculation of posterior distributions directly is not possible in closed form, the output generated from Markov chain Monte Carlo simulation is used to estimate characteristics of posterior distributions. These Markov chains are designed to converge in distribution to the desired posterior distributions. To ensure convergence, Markov chains are run with 60,000 iterations, and the first 30,000 are discarded for burn-in.

The Bayesian model has an important advantage over the classical model. Because the regression model is specified in the prior, the posterior estimate of μ will tend to be a weighted average of the data Y' and the regression estimate λ. The weights depend on the estimates of variance, namely, Σ and τ. Therefore, if the regression model displays lack of fit, indicated by large τ, the posterior estimate will be smoothed toward the data. Accordingly, in the Bayesian SUR model, interest focuses on the posterior estimates of μ and not on the regression estimates λ. The estimates of RR produced by the Bayesian model that are the focus of this analysis depend on μ. Since the Bayesian model produces estimates that are a weighted average of the data and the regression model, in the case of lack of fit, the Bayesian model smoothes estimates toward the data. This was an important property in the models fit in this study. In a classical model, RR would be estimated by the regression equation for λ alone.

Since the SUR model is estimated on a transformed scale to normality, it is necessary to transform back to make inference about the RRs. The RR is simply a ratio of rates comparing one combination of explanatory variables in the numerator with another combination in the denominator. Running the Markov chain will produce samples generated from the posterior distributions of μ. The transformation of the dependent variable is

$$Y' = \sqrt{Y}\log(Y) \tag{7}$$

Therefore, the simulated values should be transformed by the formula

$$\frac{\mu}{\sqrt{Y}} - \log(\text{exposure}) \tag{8}$$

to calculate a posterior sample for the log rates. Then log RRs can be formed by taking differences of log rates based on combinations of certain explanatory variables. The log RR is used because the sampling distribution of the RR on the log scale is close to normal. A main hypothesis of interest is whether the difference between the crash log RR and a surrogate log RR is 0 when certain explanatory variables are controlled for.

CANDIDATE SURROGATES

The surrogate is intended to capture aspects of crash mechanisms in the form of disturbed vehicle control by the driver. For road-departure crashes this clearly relates to lateral (steering) control, suggesting such variables as lane deviation and steering correction. Also, driver assistance systems on equipped vehicles are designed to give alerts when an apparently high-risk scenario occurs; two such alerts are available as candidate surrogates in the NDD. The list of candidate surrogates formulated and initially considered included time to lane crossing, time to road edge crossing, peak lateral deviation in lane, peak projected lateral deviation, coherency between tracking error and vehicle or driver response, steering rate less than small threshold for at least 4 s, driver looking away for 2 s or more, and alerts for imminent lane departures and speeds too fast for an approaching curve. Data availability, complexity of data processing, and resource constraints contributed to the selection of the following three candidate surrogates for analysis:

- LDEV (lateral deviation). Vehicle lateral deviation from the center of the lane exceeds a threshold based on an overall frequency distribution obtained from the driving data;
- LDW (lane-departure warning). Onboard lane departure system used in the driving study gave an alert to the driver; and
- TTEC (time to edge crossing). Estimated time to departing the paved surface, based on lane position and shoulder width, is less than a certain threshold.

LDEV Surrogate

In the LDEV surrogate, the vehicle offset was obtained at a rate of 10 Hz when the subject vehicle was in a lane with a solid right or left boundary, and lane tracking confidence was 70% or higher. The solid lane boundary was used to eliminate drift events into adjacent lanes and to focus on drifts that could lead to a road departure. Only time intervals when the lateral velocity was in the direction of a solid lane boundary and there was no turn signal were used. The vehicle

offset was calculated for the entire driving data set, and the 95th global percentile value of LDEV was obtained to serve as a threshold for identifying lateral deviation events. A lateral deviation event was defined as occurring when a vehicle exceeded the 95th percentile lane offset for a maximum of 10 s. For example, if a vehicle offset greater than the 95th percentile was detected, the LDEV count increased by 1. The next comparison of vehicle offset would occur 10 s later in the vehicle's time history.

LDW Surrogate

In the FOT study that generated the NDD, the LDW surrogate was triggered when the predicted vehicle path was to cross a solid or dashed-line boundary. The vehicle had to be on a nonlocal street, with speed greater than 25 mph, no turn signal, no high steering rate or braking in the past 5 s, and with actual tracking on the boundary to be crossed. The average duration of the LDEV events in the full data set was 0.61 s, and the maximum was 8.9 s. A 10-s delay prevented long events from artificially increasing the event count and excluded only 2.3% of the NDD.

TTEC Surrogate

The TTEC surrogate was based on position and velocity information and was calculated as the quotient of the distance of the vehicle to the outside edge of the road divided by the lateral vehicle velocity. This measure took advantage of an NDD variable, the available lateral maneuvering room, which was derived from side radar reflection, and lane tracking information. The distance to the road edge was determined from the vehicle's position in the lane, the width of the vehicle, and the available maneuvering room. Only periods of driving were considered when the lateral velocity points toward the right solid lane boundary with lateral velocity to the right. To be included, a driving period had to have tracking confidence of 70% or better and no turn signal. A TTEC event was defined as an instance when the vehicle's time to edge crossing was less than the 5th percentile global time to edge crossing value. At that time, the comparison of TTEC against the 5th percentile value was suppressed for 10 s to prevent the long events from inflating the surrogate counts.

SUR MODEL APPLICATION AND RESULTS

Bayesian SUR models were applied to right road-departure crashes and three candidate surrogates. Only surrogate events relative to the right boundary were considered. Three separate models, one for each crash and surrogate pair, were developed. The number of explanatory variables for SUR model application was limited by the data and consisted of four variables reported in the literature to be associated with road-departure crashes, area type, road type, horizontal curvature, and shoulder width. The categorical models are crashes, surrogate events, and exposure measures aggregated into the 24 combinations of the four variables in the models [curve (two), freeway (two), area (two), right shoulder (three)] so that there are $2 \times 2 \times 2 \times 3 = 24$ independent observations. Of the 24 possible cells (combinations of the explanatory variables), only 16 are used as data for the models, six were necessarily empty (meaning the specific combinations were not found in the data, for example, rural freeways with shoulder width 0 to 3 ft on curved and on tangent sections), and two cells had very low values for traversals and crashes, which were dropped from the analysis. The number of traversals for the cells in the analysis ranged from 57 to more than 28,000, and the number of crashes in the cells ranged from 52 to 1,879.

The exposure for crashes in each case was based on the 5-year traffic volume and segment length, and the exposure for each of the surrogates was based on number of traversals in segment and segment length. The same set of explanatory variables was used in each model.

- Curve (1 = yes, 2 = no),
- Freeway (1 = yes, 2 = no),
- Area (1 = rural, 2 = urban), and
- Right shoulder (1 = 0 to 3 ft, 2 = 3+ ft to 8 ft, 3 = 8+ ft).

The real focus of this analysis is the log RR differences used to determine if the RRs of crashes and surrogate events are the same under specified conditions. The regression parameters in the crash and surrogate equations are of secondary concern and are shown to give an indication of the effects of the four variables on crashes and surrogate measures.

Table 1 shows posterior estimates from the regression parameters for all three SUR pairs. The table also includes estimates that describe the middle 95% of the distributions, which are indicated by the range of the 2.5 and 97.5 percentiles. Log exposure is fit on the right side of the model equations for both the crash and the surrogate regression. There is some similarity in the directions and the magnitudes for certain variables. For example, in the crash–LDEV regression, the posterior mean for the curve variable is –0.642 for crash and –0.558 for LDEV surrogate. From the coding of the curve variable, this suggests that crashes and LDEV events were more likely on curves. In addition, the area variables are both negative and of similar magnitude. This suggests the protective effects of urban areas relative to rural areas. The shoulder variables are both positive, although of somewhat different magnitudes between the two regressions. The freeway variable is not significant at the 0.05 level in the LDEV regression.

The posterior estimates from the regression parameters for the curve variable in the crash–LDW model are also negative, suggesting that crashes and LDW events are more likely on curved road segments. The shoulder coefficients are also both positive. This model contains an interaction term between the freeway and area variables that is significant in the LDW regression equation and marginally significant in the crash regression. The negative coefficients suggest that the additive effects of freeway and area are somewhat reduced in urban areas when not on a freeway.

For regression estimates between the crash and surrogate measures, the crash–TTEC model shows the best agreement among the three models. The log exposure, curve, area, and shoulder variables are not only in the same direction between the crash and the TTEC regressions, but the magnitudes also tend to be reasonably close. The intercept is of no interest because it captures only the difference on an absolute scale of numbers of crashes and TTEC events. All parameters in this crash regression are significant, but the freeway and area variables in the TTEC regression do not meet the significance criteria at 0.05.

In all three SUR models, the signs of the coefficients for the shoulder width in the crash regressions look counterintuitive: apparently, crash risk is higher when the shoulders are wider. Care is needed in interpretation. This does not imply that increasing shoulder width on a particular road segment will increase crash risk; rather, it indicates that within the resolution of the statistical model used here, there is a systematic effect that more road-departure crashes occur under conditions in which shoulders are wider than in

TABLE 1 Parameter Estimates for SUR Models of Crash and Candidate Surrogate Events

	Crash and LDEV		Crash and LDW		Crash and TTEC	
	Mean (SD)	95% CI	Mean (SD)	95% CI	Mean (SD)	95% CI
Crash Parameter						
Intercept	2.095 (0.419)	1.263 – 2.933	1.918 (0.420)	1.044 – 2.720	2.017 (0.438)	1.136 – 2.836
Log exposure	0.469 (0.042)	0.386 – 0.554	0.463 (0.039)	0.391 – 0.546	0.478 (0.045)	0.394 – 0.567
Curve	–0.642 (0.072)	–0.782 – –0.500	–0.629 (0.069)	–0.766 – –0.494	–0.638 (0.077)	–0.787 – –0.488
Freeway	0.262 (0.126)	0.012 – 0.510	0.580 (0.213)	0.169 – 1.018	0.285 (0.130)	0.038 – 0.545
Area	–0.534 (0.216)	–0.967 – –0.113	–0.240 (0.234)	–0.729 – 0.194	–0.579 (0.230)	–1.033 – –0.129
Shoulder2	0.523 (0.129)	0.267 – 0.778	0.486 (0.119)	0.258 – 0.728	0.541 (0.135)	0.283 – 0.805
Shoulder3	0.327 (0.145)	0.040 – 0.615	0.315 (0.134)	0.057 – 0.591	0.351 (0.152)	0.063 – 0.653
Freeway × area	—	—	–0.367 (0.195)	–0.764 – 0.008	—	—
Surrogate Parameter						
Intercept	3.981 (0.203)	3.569 – 4.374	1.536 (0.654)	0.285 – 2.833	4.557 (0.341)	3.843 – 5.213
Log exposure	0.553 (0.030)	0.494 – 0.613	0.422 (0.087)	0.238 – 0.589	0.464 (0.054)	0.362 – 0.576
Curve	–0.558 (0.062)	–0.680 – –0.433	–0.522 (0.174)	–0.864 – –0.184	–0.594 (0.100)	–0.788 – –0.399
Freeway	–0.153 (0.079)	–0.306 – 0.005	0.866 (0.498)	–0.071 – 1.885	0.072 (0.150)	–0.218 – 0.368
Area	–0.568 (0.141)	–0.849 – –0.290	0.428 (0.580)	–0.709 – 1.565	–0.469 (0.259)	–0.996 – 0.026
Shoulder2	0.658 (0.090)	0.477 – 0.836	0.388 (0.250)	–0.109 – 0.870	0.462 (0.150)	0.172 – 0.765
Shoulder3	0.794 (0.103)	0.594 – 1.006	0.643 (0.293)	0.062 – 1.212	0.466 (0.180)	0.119 – 0.829
Freeway × area	—	—	–0.964 (0.485)	–1.919 – –0.045	—	—

NOTE: SD = standard deviation, CI = confidence interval.

which shoulders are narrower. Because of NDD limitations, a single model was used for both urban and rural areas, and only a limited set of highway variables was included. In urban areas, with high traffic density and occasional congestion, single-vehicle road-departure crashes are relatively rare; curbs typically define the road edges (shoulder width is zero), and risk is low. On rural highways with higher traffic speeds (and shoulders present), the risk is expected to be higher. The urban–rural variable accounts for some of this variation, but if the effect noted is particularly strong and the population-based area variable in only partially correlated with road conditions, it is not surprising that the presence of shoulders is associated with higher crash risk. It would be fruitful to increase the number of explanatory variables so that the shoulder variable is not confounded with other factors. It will also be beneficial to implement separate models for urban and rural areas. It is expected that in larger-scale naturalistic driving studies, confounding effects can be removed, and the shoulder width coefficients will provide a more direct indicator of relative risk.

The real focus of this analysis is on the RR measures. Crash rates are expected to be considerably smaller than rates derived from surrogate measures. This makes the RR an attractive exposure-based measure, because rates are compared not on an absolute scale but on a relative scale. If the log RR of a crash and that of the candidate surrogate are the same, then it is argued that the candidate is a good surrogate for the crash. Accordingly, for each model Markov chain Monte Carlo is used to generate a sample from the posterior distribution of the log RR difference. The hypothesis of interest is whether 0 is contained in the middle 95% of this distribution.

The log RRs of a crash and of each candidate surrogate for the three models for a road segment with a curve compared to a road segment without a curve on a nonfreeway rural road with shoulders greater than 3 ft but less than 8 ft and the distribution of log RR difference were obtained. Each distribution was drawn from samples of size 30,000 and is shown, respectively, in Figures 1 through 3.

Figure 1 shows the distribution of the log RRs for crashes and the LDEV surrogate and the posterior distribution of the log RR difference between them. The estimate for the log crash RR is 1.15 with a 95% confidence interval of (0.98, 1.33). The estimate for the log LDEV RR is 0.77 with a 95% confidence interval of (0.63, 0.92). Neither confidence interval contains 0, suggesting that the risks of crashes and LDEV events are greater on curves holding the other variables fixed. However, the mean of the distribution of the log RR difference is 0.38 with a 95% confidence interval of (0.15, 0.61). Since 0 is not contained in the confidence interval, the conclusion is that lane deviation is a poor surrogate for lane-departure crashes.

Figure 2 shows the log RRs for crashes and the LDW surrogate. The estimate for the log crash RR is 1.00 with a 95% confidence interval of (0.84, 1.16). The estimate for the log LDW RR is 1.09 with a 95% confidence interval of (0.65, 1.53). Neither confidence interval contains 0, suggesting that the risks of crashes and LDW events are greater on curves holding the other variables fixed. The mean of the posterior distribution of log RR difference is –0.08 with a 95% confidence interval of (–0.51, 0.33). The 95% confidence interval for the log RR difference includes 0, indicating that LDW could be useful as a surrogate for crashes on rural nonfreeway roads.

Figure 3 shows the log RRs for crashes and the TTEC surrogate. The estimate for the log crash RR is 1.00 with a 95% confidence interval of (0.82, 1.18). The estimate for the log TTEC RR is 1.12 with a 95% confidence interval of (0.83, 1.36). Neither confidence interval contains 0, suggesting that the risks of crashes and TTEC

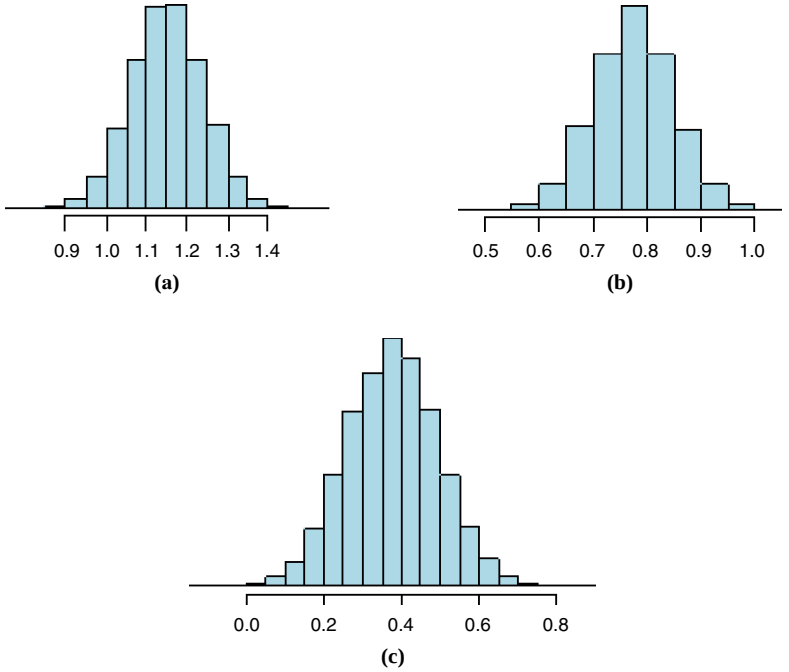

FIGURE 1 Posterior distributions, curve versus no curve: (*a*) log RR crash, (*b*) log RR LDEV, and (*c*) log RR difference.

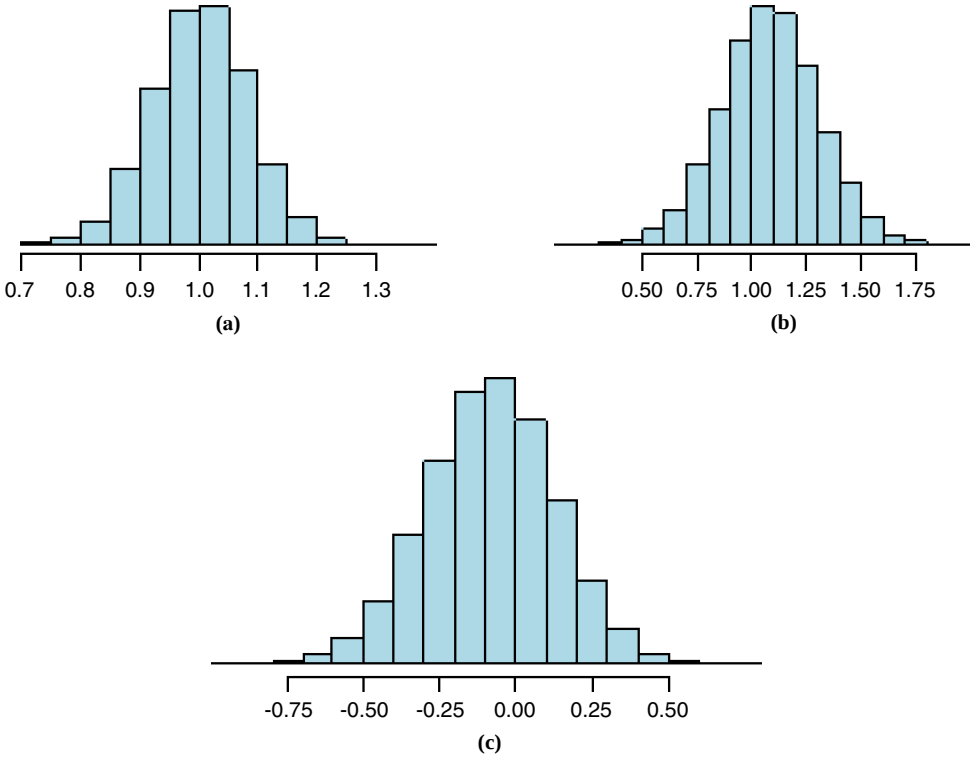

FIGURE 2 Posterior distributions, curve versus no curve: (*a*) log RR crash, (*b*) log RR LDW, and (*c*) log RR difference.

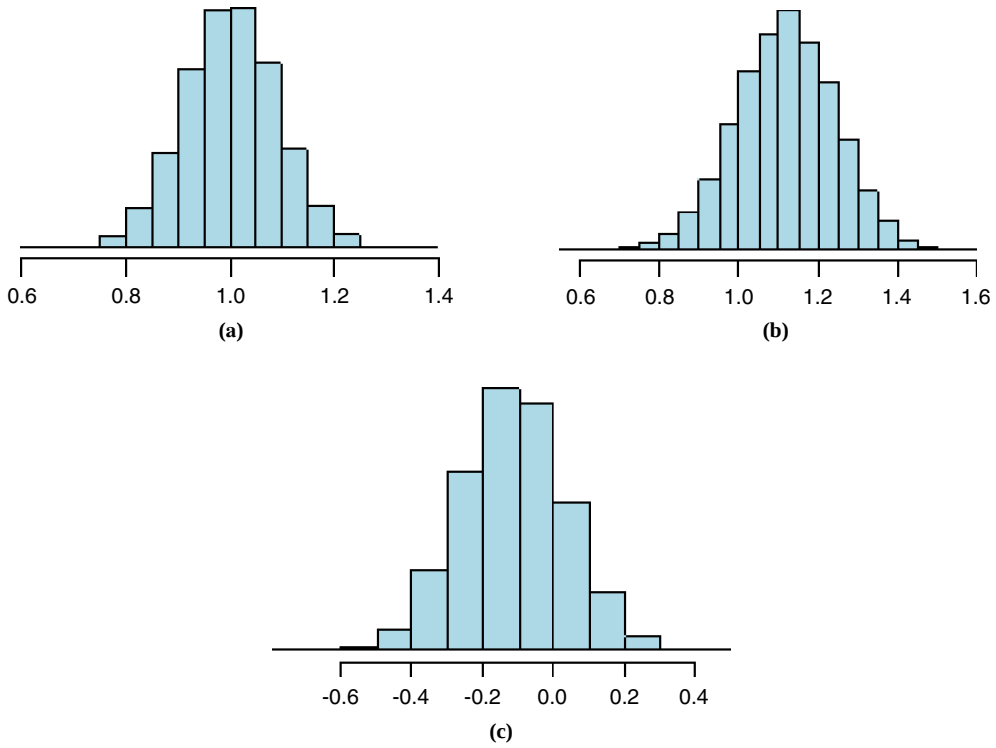

FIGURE 3 Posterior distributions, curve versus no curve: (*a*) log RR crash, (*b*) log RR TTEC, and (*c*) log RR difference.

events are greater on curves holding the other variables fixed. The mean of the posterior distribution of log RR difference is –0.11 with a 95% confidence interval of (–0.40, 0.18). The 95% confidence interval for the log RR difference includes 0, indicating that TTEC could be useful as a surrogate for crashes on rural nonfreeway roads.

Overall, different candidate surrogates have different qualities for fidelity to the crash model. In a comparison of log-relative risk differences, as well as signs and magnitudes of the regression parameters, it is clear that LDEV is the worst candidate of the three, TTEC is the best candidate, and LDW is intermediate. The results not only help confirm surrogacy but also provide a possible tool for guiding future studies in reducing risk: given a valid surrogate—one that mimics relative risk in crashes well—an intervention such as road widening, improving lane markings, changing signage, or adding rumble strips can be evaluated with the surrogate in a relatively short time. The effect on relative risk then represents a predicted safety benefit, and although data are still needed for evaluation of the effect on the surrogate, this tool is potentially much more useful, sensitive, and repeatable than counting crashes at a single treated location.

SUMMARY AND CONCLUSIONS

This research performed a joint analysis of spatially referenced crash data and NDD. As a starting point, a common measure of exposure was found in the form of normalized road segment traversals; the same road segment definitions were used for both data sets, although segments with zero exposure in the NDD were excluded from the study. A unified approach was adopted for the combined analysis of crash rates and surrogate events; the seemingly unrelated regression method was adopted because it allows for the use of common explanatory variables in the two data sets and is flexible enough to include additional explanatory variables that are not available in both. This is an important property for future analysis in SHRP 2, where driver attention variables may be included in the explanatory set for NDD. Bayesian estimation was used to determine posterior distributions of the SUR model parameters and to estimate relative risk between surrogate and crashes. The posterior distributions of the log RR provided a set of validity tests of the surrogate used; the difference in log RR between crash and surrogate events should be zero for any particular comparison, meaning that zero should be contained within an associated confidence interval. It was found that the simplest surrogate, LDEV, does not satisfy this criterion in the case of a curve–no-curve comparison, and therefore it is not seen as acceptable for use as a crash surrogate. However, the corresponding log RR distributions for LDW and TTEC did satisfy this criterion. This analysis is not exhaustive and was conducted as an exemplar of the method. In the future, it will be necessary to increase the number of explanatory variables (including driver attention variables, if available) and apply multiple log RR comparisons to prioritize the wider range of metrics for lane-keeping control.

ACKNOWLEDGMENT

The authors thank SHRP 2 for providing funding for this work under the program's safety focus area.

REFERENCES

1. Schoppert, D. W. Predicting Traffic Accidents from Roadway Elements of Rural Two-Lane Highways with Gravel Shoulders. *Bulletin 158*, HRB, National Research Council, Washington, D.C., 1957, pp. 4–16.
2. Dart, O. K., and L. Mann, Jr. Relationship of Rural Highway Geometry to Accident Rates in Louisiana. In *Highway Research Record 312*, HRB, National Research Council, Washington, D.C., 1970, pp. 1–16.
3. Zegeer, C. V., and J. A. Deacon. Effects of Lane Width, Shoulder Width, and Shoulder Type on Highway Safety. *State-of-the-Art Report 6: Relationship Between Safety and Key Highway Features: A Synthesis of Prior Research.* TRB, National Research Council, Washington, D.C., 1987, pp. 1–21.
4. Shankar, V., F. Mannering, and W. Barfield. Effects of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention*, Vol. 27, No. 3, 1995, pp. 371–389.
5. Stutts, J. C., J. R. Stewart, and C. Martell. Cognitive Test Performance and Crash Risk in an Older Driver Population. *Accident Analysis and Prevention*, Vol. 30, No. 3, 1998, pp. 337–346.
6. Snyder, J. C. Environmental Determinants of Traffic Accidents: An Alternate Model. In *Transportation Research Record 486*, TRB, National Research Council, Washington, D.C., 1974, pp. 11–18.
7. Parker, M. R., Jr., and C. V. Zeeger. *Traffic Conflict Techniques for Safety and Operations.* FHWA-IP-88-027. FHWA, U.S. Department of Transportation, 1989.
8. Guo, F., S. G. Klauer, J. M. Hankey, and T. A. Dingus. Near Crashes as Crash Surrogate for Naturalistic Driving Studies. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2174*, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 66–74.
9. Gordon, T. J., L. P. Kostyniuk, P. E. Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor. Analysis of Crash Rates and Surrogate Events: Unified Approach. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
10. LeBlanc, D., J. R. Sayer, C. Winkler, R. Ervin, S. Bogard, J. Devonshire, M. Medfford, M. Hagan, Z. Bareket, R. Goodsell, and T. J. Gordon. *Road Departure Crash Warning System Field Operational Test: Methodology and Results*, Vol. 1. NHTSA, U.S. Department of Transportation, 2006.
11. Environmental Systems Research Institute, Inc. *ArcMap Geographic Software, Version 9.3.* Redlands, Calif., 2008.
12. Zellner, A. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association,* Vol. 57, 1962, pp. 348–368.
13. Agresti, A. *Categorical Data Analysis,* 2nd ed., John Wiley and Sons, New York, 2002.
14. Weisberg, S. *Applied Linear Regression,* 2nd ed., John Wiley and Sons, New York, 1985.
15. Gelman, A., J. B. Carlin, and H.S. Stern. *Bayesian Data Analysis,* 2nd ed., Chapman and Hall, New York, 2003.

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