COMPARING MOTOR-VEHICLE CRASH RISK OF EU AND US VEHICLES

CAROL A. FLANNAGAN[1], ANDRÁS BÁLINT[2],
KATHLEEN D. KLINICH[1], ULRICH SANDER[2],
MIRIAM A. MANARY[1], SOPHIE CUNY[3],
MICHAEL MCCARTHY[4], VUTHY PHAN[3],
CAROLINE WALLBANK[4], PAUL E. GREEN[1], BO SUI[2],
ÅSA FORSMAN[2], HELEN FAGERLIND[2]

[1]UNIVERSITY OF MICHIGAN TRANSPORTATION RESEARCH INSTITUTE, ANN ARBOR, MI
[2]SAFER VEHICLE AND TRAFFIC SAFETY CENTRE AT CHALMERS, GOTHENBURG, SWEDEN
[3]CENTRE EUROPEEN D'ETUDES DE SECURITE ET D'ANALYSE DES RISQUES, NANTERRE, FRANCE
[4]TRL (TRANSPORT RESEARCH LABORATORY), CROWTHORNE, UNITED KINGDOM

1. Report No. UMTRI-2015-1	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Comparing Motor-Vehicle Crash Risk of EU and US Vehicles	5. Report Date May 2015 6. Performing Organization Code
7. Author(s) Carol A. C. Flannagan, András Bálint, Kathleen D. Klinich, Ulrich Sander, Miriam A. Manary, Sophie Cuny, Michael McCarthy, Vuthy Phan, Caroline Wallbank, Paul E. Green, Bo Sui, Åsa Forsman, Helen Fagerlind	8. Performing Organization Report No.
9. Performing Organization Name and Address University of Michigan Transportation Research Institute 2901 Baxter Rd. Ann Arbor MI 48109 in association with: SAFER Vehicle and Traffic Safety Centre at Chalmers, Gothenburg, Sweden Centre Européen d'Etudes de Sécurité et d'Analyse des Risques, Nanterre, France TRL (Transport Research Laboratory), Crowthorne, United Kingdom	10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Alliance of Automobile Manufacturers	13. Type of Report and Period Covered Final, May 2014-January 2015 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This study examined the hypotheses that vehicles meeting EU safety standards perform similarly to US-regulated vehicles in the US driving environment, and vice versa. The analyses used three statistical approaches to "triangulate" evidence regarding differences in crash and injury risk. Separate analyses assessed crash avoidance technologies, including headlamps and mirrors. The results suggest that when controlling for differences in environment and exposure, vehicles meeting EU standards offer reduced risk of serious injury in frontal/side crashes and have driver-side mirrors that reduce risk in lane-change crashes better, while vehicles meeting US standards provide a lower risk of injury in rollovers and have headlamps that make pedestrians more conspicuous.

17. Key Word Crashworthiness, comparison EU US, injury risk, front-side crashes, rollover, logistic regression, SUR, likelihood surface, Bayes factor, crash avoidance, headlamps, side mirrors	18. Distribution Statement

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price

Contents

List of Figures

List of Tables

Acknowledgements

This work was carried out by the University of Michigan Transportation Research Institute (UMTRI) in association with SAFER – Vehicle and Traffic Safety Centre at Chalmers, Sweden; Centre Européen d'Etudes de Sécurité et d'Analyse des Risques (CEESAR), France, and Transport Research Laboratory (TRL), United Kingdom. UMTRI, CEESAR, TRL and three SAFER partner organizations were involved in this project: Chalmers University of Technology (András Bálint and Helen Fagerlind), Autoliv Research (Ulrich Sander and Bo Sui) and VTI, the Swedish National Road and Transport Research Institute (Åsa Forsman).

We would like to thank the following agencies for use of their data:

This report used accident data from Great Britain collected as part of Co-operative Crash Injury Study (CCIS) during the period 2003-2009. CCIS was managed by TRL Limited, on behalf of the DfT (Transport Technology and Standards Division) who funded the project along with Autoliv, Ford Motor Company, Nissan Motor Company and Toyota Motor Europe. Previous sponsors of CCIS have included Daimler Chrysler, LAB, Rover Group Ltd, Visteon, Volvo Car Corporation, Daewoo Motor Company Ltd and Honda R&D Europe (UK) Ltd. Data were collected by teams from the Birmingham Automotive Safety Centre of the University of Birmingham; the Vehicle Safety Research Centre at Loughborough University; TRL Limited, and the Vehicle & Operator Services Agency of the DfT. CCIS has subsequently been superseded by the UK's Road Accident In-Depth Studies (RAIDS) programme and all the data is stored in the RAIDS database. More information on the study can be found at https://www.gov.uk/government/publications/road-accident-investigation-road-accident-in-depth-studies.

For the present study, accident data from GIDAS (German In-Depth Accident Study) was used. GIDAS, the largest in-depth accident study in Germany, is funded by the Federal Highway Research Institute (BASt) and the German Research Association for Automotive Technology (FAT), a department of the VDA

(German Association of the Automotive Industry). Use of the data is restricted to the participants of the project. Further information can be found at http://www.gidas.org.

CEESAR acknowledges the French National Research Agency and Foundation MAIF for funding the VOIESUR project. We also acknowledge the members of the VOIESUR consortium, LAB (Laboratoire d'Accidentologie, de Biomécanique et d'Etudes du comportement humain), IFSTTAR (Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux) and CEREMA (Centre d'étude et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement).

The National Highway Safety Transportation Administration (NHTSA) funds the National Automotive Sampling System and Fatality Analysis Reporting System data collection activities. The three component databases are collected continuously and datasets are made free and open to the public on NHTSA's ftp website: ftp://ftp.nhtsa.dot.gov/

We thank the European Commission for funding the PENDANT project under the 'Competitive and Sustainable Growth' Program (1998-2002) and the consortium members of the project for their contributions to the resulting crash database.

Crash data from the Swedish INTACT and the Indian RASSI databases were used in this project to address a question regarding data harmonization. The relevant data were extracted from the databases by Jan Martinsson (Chalmers University of Technology) and Dr. Avinash Penumaka (Mercedes Benz Research and Development India, MBRDI), respectively.

We thank the developers of informedforlife.org for sharing their data on US NCAP ratings.

This research was sponsored by the Alliance of Automotive Manufacturers. The opinions expressed within this report represent those of the authors and not necessarily those of the funding agency.

Executive Summary

This report describes Phase II (implementation of analysis) of a project to develop and implement a statistical methodology to investigate the hypothesis that passenger vehicles meeting EU safety standards would perform equivalently to US-regulated passenger vehicles in the US driving environment, and that vehicles meeting US safety standards would perform equivalently to EU-regulated vehicles in the EU driving environment. To answer this question, it was necessary to separate risk from exposure because EU and US drivers drive in different environments. Risk is the probability of injury given a particular set of circumstances; exposure is the particular collection of those circumstances. In addition, regulation affects both risk of a crash (crash avoidance) and risk of injury *given* a crash (crashworthiness). These analyses were carried out separately because the relevant datasets and outcomes are different.

For the crashworthiness analysis, we represented risk (of injury in a particular crash) using a statistical model that could be applied to different environments. For crash avoidance, we selected a crash subpopulation and control crashes to adjust for any exposure differences between the EU and US. For both crashworthiness and crash avoidance, the comparison of injury risk given a particular set of crash characteristics, respectively the comparison of crash involvement, was then argued to be driven by differences between the vehicles themselves.

Analysis of Crashworthiness
The basic process first involved identifying appropriate databases that include in-depth crash information, such as estimation of crash severity using Delta-V and injury outcome based on medical records. The next step was to harmonize variable definitions and sampling criteria so that the data could be combined and compared using the same parameters. Logistic regression models of injury risk in EU-regulated and US-regulated vehicles were constructed and evaluated using three different approaches. Method 1 tested the hypothesis that all coefficients in the best-fit EU and US risk models are the same (i.e., the models are the same as a whole). Method 2 evaluated the injury risk predictions of the best EU model and the best US model (even if different), each applied to both the US and EU standard populations. Method 3 evaluated the strength of the evidence for a variety of levels of overall risk difference between the two vehicle groups (EU-regulated and US-regulated) compared to the evidence for no risk difference.

Datasets used were the National Automotive Sampling System-Crashworthiness Data System (NASS-CDS or CDS) for the US, the Co-operative Crash Injury Study (CCIS) from Great Britain, the Véhicule Occupant Infrastructure Etudes de la Sécurité des Usagers de la Route - Vehicle Occupant Infrastructure and Road Users Safety Studies (VOIESUR) from France, and the German In-Depth Accident Study (GIDAS) from Germany. In addition, a sample from the European Pan-European Co-ordinated Accident and Injury Database (PENDANT) project was included. PENDANT covered eight EU countries; cases were removed that could be duplicated in other datasets. For weighting of EU datasets, we also used the Community Road Accident Database (CARE). CARE contains aggregated national crash data (police-reported crashes) from all 28 EU countries plus Iceland, Liechtenstein, Norway and Switzerland.

Sampling restrictions used in any of the datasets were applied to all datasets to avoid sampling bias. Key restrictions were: 1) at least one occupant in the crash had an Abbreviated Injury Scale injury of 1 or greater (AIS1+); 2) at least one vehicle had a damage extent of 2 or greater according to its Collision

Damage Classification (CDC) for the crash; and 3) if available, at least one vehicle was towed away from the accident site. The analysis was conducted at the occupant level, and additional restrictions were applied to focus on risk that could be associated with vehicle design related to regulatory requirements. These restrictions included: 1) Vehicle model years 2003+; 2) front outboard occupants age 13+ with known belt use status; 3) vehicles with reconstructed Delta-V (does not apply to rollover); 4) cases with non-missing values of predictors; and 5) vehicles with front or side damage (based on the CDC for the most harmful event) or vehicles that experienced a rollover. The injury outcome used in analysis was based on the Maximum Abbreviated Injury Scale score. Occupants whose worst injury had a score of 3 or higher or those who were fatally injured were classified as "MAIS3+F injured"; those who were uninjured or whose worst injury had a score of 2 or less were classified as "not MAIS3+F injured." This injury level was selected because it is typically used for regulatory analysis to define targets and assess vehicle performance. Thus, stated precisely, the goal of the statistical modeling was to predict MAIS3+F injury risk to front outboard occupants ages 13+ in front, side, or rollover towaway crashes in which at least one occupant in the crash sustained an injury and at least one vehicle was towed or damaged at extent level 2 or greater.

To estimate overall injury risk in the crash population for each model, we required a standard population for each region. The EU standard population consisted of the combined EU datasets used for model development (the in-depth data from each country) weighted to the EU crash population based on the CARE dataset using the most recent years per country (2009 to 2013). The US standard population was the CDS crash years 2007-2012 with previously identified restrictions applied. Assessment of overall injury risk was carried out in parallel: once on the US standard population and once on the EU standard population.

After harmonization of sampling, we identified a master list of potential predictors that were available in all of the in-depth datasets. For each predictor, the definitions and measurement methods used in the datasets were compared, and a harmonized definition was developed. In many cases, this required categorization of cases (e.g., intrusion was categorized in to none, minor, and major). In others (e.g., age), harmonization was straightforward.

For Delta-V, reconstruction was done using two different methods: crush-based and trajectory-based. To assess the comparability of these methods, we found cases with data that allowed both reconstruction methods to be applied. The two reconstructions were compared separately for frontal and side impacts, and found to be generally similar. From these comparisons, we developed a simple linear transformation to apply to crush-based reconstruction cases to harmonize them with the trajectory-based reconstructions. Thus, the Delta-V values used throughout this study can be considered to be equivalent to trajectory-based reconstructed Delta-V.

The first step in the model development process was to generate injury risk models individually using each dataset. Frontal, near-side, and far-side crashes were analyzed together (termed "front/side crashes"). Analyzing these crashes together served the original goal of maximizing comprehensiveness of the analysis and to maximize sample size. A separate model was developed for rollover because Delta-V is generally not reconstructed for rollover. The starting list of harmonized predictors for front/side crashes included: Delta-V (log and square transformations considered), crash type (front, near side, far side), age, age^2 (to allow a quadratic relationship), belt use, road type, vehicle type, model year group, principal direction of force (PDOF) (relative to side of impact), intrusion (relative to side of impact), airbag deployment, crash partner, presence of multiple impacts, and interactions of Delta-V and crash direction. For rollover, the starting list included: age (including a quadratic term), gender, roof intrusion, ejection, belt use, road type, model year, light condition, and seat position. For each dataset, non-significant model parameters were dropped. In marginal cases, changes to Akaike Information Criteria (AIC) were considered in deciding whether to include a parameter or not.

Based on the results from the individual models, all predictors significant in any set were included in the final models. Although model year group was not significant in any of the models, we included a two-level model-year predictor (2003-2006 vs. 2007+) to account for regulatory changes that occurred in the US between 2006 and 2007. The final predictor list for front/side crashes was: Delta-V, age, age^2, crash type (front, far-side, near-side), belt use (belted, unbelted), Delta-V*crash type interaction, intrusion (none, minor, major), principal direction of force (PDOF; 0, 30, >30 relative to side of damage), crash partner (car, narrow, wide, other), model year (2003-2006, 2007+), and road type at accident location (rural, urban). For rollover, the final predictor list was: age, belt use, roof intrusion, model year, road type at accident location, and gender (male, female).

The best-fit US risk models for front/side and rollover were developed using logistic regression. Case weights were used in analysis, and survey methods (Taylor series) were used to account for the sample survey design and estimate the variance-covariance matrix for the coefficients. The best-fit EU model was also developed using logistic regression. Cases in the four EU development datasets were weighted based on CARE, and weights were normalized to the raw sample size to appropriately estimate the variance-covariance matrix for the coefficients. All models used the same set of 18 predictors (including an intercept) for front/side and 9 predictors (including an intercept) for rollover.

In addition to the best-fit models, we also calculated the log-likelihood for a large number of alternative models. The log-likelihood for these models was used in the development of Bayes Factors in Method 3. Because the EU raw data could not be shared because of use agreement restrictions, we could not use traditional iterative search methods for the EU model. The assessment of log-likelihood for the large set of alternative models also facilitated the search for the best-fit model for the EU dataset. That is, log likelihood was calculated for 193,563 possible models for front/side and 164,865 possible models for rollover. The highest log-likelihood among these models was selected as the best-fit model and the remaining models were used to compute the Bayes Factors (Method 3) comparing evidence for different levels of overall risk differences. The latter was also done for the US data, though the best-fit models were selected using standard iterative search methods in the statistical software SAS.

Three approaches were used to evaluate equivalence of the risk models. Method 1 tested the null hypothesis that all coefficients in the EU and US injury risk models are the same. A Type I error occurs if the null hypothesis of no difference between coefficients is rejected when it is actually correct. Type II error occurs if the null hypothesis is accepted when it is, in fact, incorrect. The original proposed methodology planned to balance between these types of errors using power analysis. However, the results proved to be conclusive without reference to power. Using seemingly unrelated regression (SUR), tests were conducted to determine if individual coefficients are significantly different for EU and US models and if all coefficients as a whole are significantly different for the EU and US models. For frontal/side crashes, nine of the 18 coefficients were found to be statistically different, as was the overall set of model coefficients (p=0.0001). For rollovers, the belt use coefficient was the only one that reached the 0.05 level of significance, but it was so different that the null hypothesis of overall model equivalence was also rejected (p=0.00016). For both injury models examined using Method 1, we reject the null hypotheses that the EU and US injury models are the same.

Method 2 evaluated the predictions of the best US model compared to the best EU model, using the same predictor set for both models. Each best model was applied to both the EU and US standard populations. For each standard population, we find the risk difference (arbitrarily defined as subtracting US from EU injury risk), and find the variance of the risk difference. Positive values indicate lower risk for US vehicles; negative values indicate lower risk for EU vehicles.

In general, variance in these estimates was higher for the EU risk models than the US risk models, which is consistent with their relative raw sample sizes. To convey a sense of the magnitude of uncertainty, we present confidence intervals (CIs) along with point estimates here. *However, the reader is cautioned that the fact that the confidence intervals contain 0 cannot be interpreted as a proof of no risk difference.* This is discussed further in the section “Interpretation of Crashworthiness Results” at the end of the Executive Summary.

When applied to the EU front/side population, the US model predicted a 0.065 risk and the EU model predicted a 0.052 risk; the absolute difference was -0.013 (95% CI: (-0.084, 0.059)). For the rollover model applied to the US standard population, the US model predicted a risk of 0.071 and the EU model predicted 0.13 risk; the most likely absolute difference was 0.057 (95% CI: (-0.064, 0.179)). When applied to the EU rollover standard population, the US model predicted a 0.067 risk and the EU model predicted 0.10 risk, with a difference of 0.036 (95% CI: (-0.055, 0.128)). So using Method 2, EU models predicted lower risk in front/side impacts, but higher risk in rollovers.

To better understand the source of risk differences, we used the best-fit models to estimate EU and US injury risks for certain subsets of each population (for both standard populations). The results for the two standard populations were consistent. For front/side crashes, the largest risk differences were seen in near-side crashes, occupant ages from 31-70, and unbelted occupants. In addition, the risk difference increased with increasing Delta-V such that predicted risk was the same for Delta-V<20 km/h and the difference was largest for Delta-V≥60 km/h. For rollovers, both belted and unbelted occupants were at lower estimated risk in US vehicles compared to EU vehicles, but the difference was largest for unbelted occupants. Similarly, both ejected and unejected occupants in US vehicles were at lower risk compared to those in EU vehicles, but the difference was largest for ejected occupants (who make up a very small proportion of the sample).

Method 3 evaluated evidence for a variety of hypotheses compared to the hypothesis of no risk difference. For each standard population, we defined a series of specific risk differences, and for each risk difference, we computed the evidence as compared to the evidence for zero difference. Evidence in this context is defined as the likelihood and the ratio of likelihoods for two hypotheses is called the Bayes Factor. In this application, we estimated log Bayes Factors using the Schwarz Criterion. Log Bayes Factors greater than 1 indicate positive evidence for a particular risk difference (as compared to zero difference) and log Bayes Factors greater than 3 indicate strong evidence for the risk difference. As before, risk difference was arbitrarily defined as EU risk – US risk. For the frontal/side US population, the strongest evidence (log Bayes Factors > 3) was for the hypotheses associated with risk differences from -0.018 to -0.004, all of which are more supported than the zero-difference hypothesis. For the frontal/side EU population, the Bayes Factors indicated strongest evidence for risk differences from -0.018 to -0.009. All of the most supported hypotheses indicated that injury risk in EU vehicles is lower than US vehicles in front/side crashes. For both the US and EU rollover population, the evidence strongly supports the hypothesis that injury risk is lower in US vehicles than EU vehicles.

Crash Avoidance

Crash avoidance analysis focused on headlamps for visibility of pedestrians and mirrors for prevention of lane change/merge behaviors because sufficient data were not available to analyze other crash avoidance equipment. For the headlamp comparison, Daylight Savings Time (DST) analyses were performed to compare the dark/light ratios for pedestrian fatalities for the EU and US. In principle, using a time window on either side of DST holds pedestrian exposure constant while the light level changes substantially. The dark/light ratio of pedestrian fatalities for these time periods should reflect the relative risk to pedestrians in dark compared to light. These ratios for the EU and US can then be compared. Note that this analysis does not consider the effect of glare.

Data from the US and eight EU countries were available for the analysis. The overall estimate for the US/EU ratio of dark/light risk was 0.67 (95% CI: 0.41 to 1.11), which represents a 30% lower risk in the in the US. One explanation for this is that US headlamps illuminate pedestrians better than EU headlamps. The variance is fairly large and the 95% confidence interval does contain the neutral value of 1.

For mirrors, the US specifies a planar mirror on driver side, while the EU allows non-planar mirrors on both sides. Thus, if we compare driver-side lane changes to passenger-side lane changes, the US ratio would be expected to reflect differences in the effectiveness of the different mirror types as well as differences in the exposure to lane changes on the two sides, whereas the EU ratio would reflect only exposure differences. If the relative exposure to driver-side vs. passenger-side crashes can be argued to be similar in the two regions, the ratio of the US ratio to the EU ratio would reflect a performance difference in the planar vs. non-planar mirror. Only two EU countries provided usable data for this analysis; the US/EU ratio of driver and passenger lane change crashes was 1.24 (95% CI: 1.18 to 1.30), suggesting that mirrors in EU vehicles on the driver's side prevent lane-change/merge crashes on the driver's side better than those in US vehicles. However, the small number of EU countries included in the analysis limits the possibilities of drawing conclusions regarding the entire EU based on these results. The reader is also cautioned that we do not know how differences in overtaking behavior in the UK and US might influence the results.

Summary of Results

The project results support the following conclusions:

- The EU and US injury risk models are different for both front/side crashes and rollovers.
- Overall risk across the US front-side crash population (given the selection criteria for this study) is likely lower for EU vehicles. Though the range of estimates is wide, the best estimate of the risk difference is -0.012.
- Overall risk across the EU front-side crash population (given the selection criteria for this study) is likely lower for EU vehicles. Though the range of estimates is wide, the best estimate of the risk difference is -0.013.

- Overall risk across both EU and US rollover crash populations is lower for US vehicles. The best estimate of the risk difference for the US population is 0.057. The best estimate of the risk difference for the EU population is 0.036.
- Risk differences in front/side crashes are largest for near-side crashes, middle occupant ages (31-70), unbelted occupants, and higher Delta-Vs. In rollovers, risk differences were highest for unbelted occupants and ejected occupants.
- US ratio of pedestrian fatalities in dark vs. light is estimated to be lower than in the EU, though the 95% CI contains 1; one possible explanation for this is that headlamps in US vehicles may illuminate pedestrians better than those in EU vehicles.
- EU ratio of driver-side lane changes compared to passenger-side lane changes, based on data from only two EU countries, is lower than in the US. One possible explanation for this is that driver-side mirrors in EU vehicles reduce risk in lane-change crashes better than those in US vehicles.

Interpretation of the Crashworthiness Results

The goal of this study was to address the equivalence of the real-world safety performance of passenger vehicles developed in two separate regulatory environments. In principle, the approach is designed to evaluate evidence related to the elements of relative field performance of EU and US vehicles that can be attributed to regulatory differences (rather than environmental differences). In practice, the causal tie between regulatory differences and observed field performance differences cannot be made without randomized controlled trials. Thus, the modeling approach used here can identify observed differences and can eliminate as many alternative explanations as possible, but analysis of observational field data cannot establish cause with certainty.

Two steps in the data analysis served to remove as many alternative explanations as possible. First, we constrained the inclusion criteria for all of the samples to be the same. This way, we sampled from the same population of crashes, even though they may arise very differently in the two regions. Second, we used the same set of predictors to build risk models that estimate injury risk under a specified set of circumstances of the crash, vehicle, or occupant. The circumstances (e.g., occupant age, crash severity, crash direction) were designed to isolate risk from exposure as much as possible. That is, injury risk should not be affected by whether a crash was caused by speeding, texting, or falling asleep at the wheel if the nature of the crash (its direction and severity, indicating the forces acting on the vehicle occupants) is the same. We seek to take these into account in the model.

Although the risk model approach is a good way to separate risk from exposure, it does not perfectly eliminate all possible alternative explanations. (As noted earlier, only randomized controlled trials can demonstrate cause.) In this case, we argue that regulatory differences are the primary mechanism to explain differences between the risks from the two populations. However, because regulation provides a *minimum* standard, one alternative explanation for differences is that one population of vehicle owners tends to purchase safer vehicles (i.e. vehicles higher above the minimum standards) than the other. This cannot be controlled or measured with our datasets and could produce overall differences in risk. A related alternative explanation is that consumer ratings systems, which are also different in the two regions, drive vehicle design, and differences are related to the elements emphasized by the ratings rather than the base regulations. Finally, the possibility exists that data artifacts not accounted for by the models are influencing the results. Significant effort was put into removing foreseeable artifacts, but unforeseen issues are always possible in analysis of observational data.

Finally, we caution the reader in interpreting significance tests and confidence intervals. Standard hypothesis testing, which relies on the p<0.05 rule, considers the question: "What is the probability of getting my results, *if the null hypothesis of no difference were true."* When results are significant, as with Method 1, the no-difference hypothesis is highly unlikely (less than a 5% chance of being true). However, failure to reach significance, including risk-difference confidence intervals that contain 0, is not evidence *for* the null hypothesis. The test establishes an arbitrary (but mathematically convenient and logical) hypothesis as a "straw man." If the statistic found is highly unlikely under the null hypothesis, then the null hypothesis can be rejected. However, many other hypotheses remain untested using this approach (e.g., a risk difference of -0.001 or a risk difference of +0.02).

In this context, where evidence *for* equivalence is sought, other methods must be considered. In particular, Method 3 approaches the question without setting any hypothesis as the default. Instead, it computes and compares evidence for a wide variety of hypotheses (with comparisons made two at a time). Similarly, the distributions of probable risk differences in Method 2 give a more complete picture of the uncertainty in the analysis and the relative support for different risk differences. Thus, for Method 2, we present CIs to present a more complete picture of both the best estimate (the mean) and the level of uncertainty (the CI).

Recommended Next Steps

To our knowledge, this is the first side-by-side comparison of predicted risk for EU-regulated and US-regulated vehicles. As such, further work should be done to replicate the results, identify artifacts that may have influenced the patterns seen, and/or seek evidence for mechanisms linking the results to vehicle design differences that result from regulatory differences. We recommend two primary paths for next steps in research.

First, we recommend additional analyses of the field data. In particular, some patterns seen in the breakdowns of subgroups were unexpected. For example, the EU model shows very similar overall predicted risk in near- and far-side crashes while the US model shows higher risk in near-side crashes compared to far-side crashes. Because of the proximity of the occupant to the source of the impact, near-side crashes would be expected to result in greater injury risk. Similarly, the potential effect of the substantially greater share of SUVs and pickup trucks in the US population than in the EU should be examined. Datasets with a rollover severity measure should be used to look at whether different ESC penetration in the two populations could have influenced the rollover results. Finally, detailed investigation of injury patterns should make mechanisms of injury (as they related to regulation) clearer. Both unexpected and expected results should be looked at closely to identify those that are most robust and those that may be influenced by dataset or population artifacts.

Second, we recommend using computational models of typical US-regulated and EU-regulated vehicle designs to investigate potential physical mechanisms of the differences seen. Crash testing is only done in extreme conditions, but most crashes in the field data are lower severity. Computational models allow investigation of injury mechanisms over a wide range of field conditions. When combined with crash data analysis, this approach can help find mechanisms for the results seen in the field.

Finally, in this project, the use of crash data in various contexts has been demonstrated and at the same time, certain gaps in data availability have been identified. Future reproductions and extensions of this study would greatly benefit from the availability of harmonized accident data, hence further data collection and data harmonization efforts are encouraged.

Introduction

At the time of this writing, the United States and the European Union have entered into negotiation of the Transatlantic Trade and Investment Partnership (TTIP). This agreement is designed to reduce barriers to trade between the two economic units. One barrier to trade is the differing safety standards testing and requirements for vehicles sold in the EU and the US. Testing the same make/model under both regimens and adapting design to each can be expensive, and negotiation of common standards may be difficult and time-consuming.

An alternative to item-by-item harmonization is mutual recognition, an approach that has been implemented to some degree in the airline domain. Under this solution, vehicles that meet EU regulations would be recognized for sale in the US, and vehicles that meet US regulations would be recognized for sale in the EU. To justify mutual recognition, it would be helpful (or possibly even necessary) to demonstrate that safety in EU- and US-regulated vehicles is essentially equivalent.

The TTIP trade negotiations prompted the current research project to analyze crash data to compare the crash injury risk of US and EU vehicles. In Phase 1 of the project, a methodology was proposed to investigate the hypothesis that vehicles meeting EU safety standards would perform equivalently to US-regulated vehicles in the US driving environment, and that vehicles meeting US safety standards would perform equivalently to EU-regulated vehicles in the EU driving environment. In Phase 2, the analysis was carried out. This document contains a description of the Phase 1 methodology *as implemented* (this was done with only minor changes) and presentation of the results of the Phase 2 analysis.

A key challenge in evaluating safety performance for EU- and US-regulated passenger vehicles is that the two types of vehicles are driven in different driving environments, and crash datasets contain events involving only one group of vehicles. Thus, crash datasets represent the combination of risk and exposure for a given environment and vehicle population. Risk is the probability of injury or crash involvement given a particular set of circumstances, while exposure is the particular collection of those circumstances. If a vehicle is moved to a different driving environment, its risk characteristics are carried with it, but the exposure to different crash characteristics changes with the change in environment. To answer the question posed, we must separate risk from exposure. Because EU vehicles and US vehicles are separated geographically, their risk is represented with a statistical model, which is then applied to the other region's exposure population. The risk model based on EU vehicle performance can be applied to the US crash environment and compared to the performance of US vehicles in the US crash environment, and vice versa. As the risk models generated from each region are applied to both regions' environments, the question is then asked: What is the evidence that vehicle safety performance is (or is not) essentially equivalent?

In this project, analysis of crashworthiness and crash avoidance are performed separately, as the relevant datasets and outcomes are different. In-depth crash databases with harmonized injury outcomes are needed to assess crashworthiness, defined as a risk of injury given that a crash occurred. Databases of police-reported crashes and exposure data are needed for crash avoidance, defined as the risk of a crash occurring.

The methods section of this report contains details on datasets, treatment of the data (inclusion criteria and variable definitions), and analytical methods. Some statistical details are included in appendices.

The approach for analyzing crashworthiness uses three methods to better understand the comparison between the two vehicle groups. The first method tests the basic hypothesis that the two best-fit risk models (one for EU-regulated vehicles and one for US-regulated vehicles) are the same. The second method applies the two Method 1 risk models side-by-side to the EU crash data, which represent the EU driving environment, and again to the US crash data, which represent the US driving environment. This creates two separate direct comparisons of risk, which allows for a more detailed look at the groups of crashes (such as frontal or side impacts or rollovers) for which predicted injury risk is similar or different within each environment. Finally, the third method compares the overall weight of evidence for models that predict some risk difference vs. models that predict no risk difference. This approach uses Bayes Factors to compare evidence for two hypotheses and does not depend on the single best-fit model. As with the second method, the comparisons are done separately for the EU crash population and the US crash population.

The methods also include description of how crash avoidance was considered. Data in the relevant EU and US datasets were only sufficient to address two crash avoidance countermeasures: headlamps (in relation with pedestrian crashes at nighttime versus daytime) and mirrors (where the analysis is based on the proportion of lane-change/merge crashes to the driver’s side versus the passenger’s side).

Methods

General Approach

The general analysis approach uses statistical models to separate risk from exposure in each region. The statistical model of risk, which is a logistic regression model made up of a series of estimated coefficients of predictors, is used to represent predicted risk in EU-regulated or US-regulated vehicles. The EU and US environments are represented by "standard populations," which are defined in this section. These standard populations are a representative collection of crashes in each environment that future vehicles are likely to encounter in that environment. Thus, when a given risk model (e.g., EU-regulated vehicles) is applied to a given standard population (e.g., US standard population), the combination is an estimate of the overall effect of having that vehicle group (e.g., EU-regulated vehicles) being driven in the driving environment (e.g., US).

Because of the different nature of regulations for crashworthiness and crash avoidance, these analyses were separated. Crashworthiness is associated with injury risk, *given that a crash has occurred, and given its characteristics,* whereas crash avoidance is associated with the risk of a crash event occurring in the first place. Although many new crash avoidance systems are becoming available and are being considered for regulation now, equipped vehicles and the effects of regulation are not in the datasets available for this project. Thus, the crash avoidance analysis was limited to two regulation-relevant vehicle components (headlamps and side mirrors) that are designed to keep drivers out of crashes. The primary focus of our work was on crashworthiness and injury risk given a crash, and most of the methodological presentation below is focused on that area.

A key requirement for comparing risk models is that the datasets on which they are built are sampled from underlying crash populations that are defined in the same way. In addition, the variables used to develop the models must be defined in the same way so that when they are applied to different datasets, the coefficients are being appropriately applied. A simple example is that velocity must be measured in the same units (e.g., km/h not mi/h). However, some of the harmonization issues described below are more complex.

As discussed above, the populations of crashes being sampled are inherently different (representing the crashing environment). The modeling approach we use, logistic regression, produces unbiased coefficients even when the underlying sample is biased (Prentice & Pyke, 1979). However, the intercept of these models will be influenced by the overall injury rate in the sample. Thus, to ensure that the models are producing comparable estimates that are not biased by the fact that they were built from different samples, the inclusion criteria for the samples must be harmonized as well.

Once the datasets were harmonized, we made use of maximum likelihood for all three methods. The details are given in the sections below.

Datasets

The US is a single country, and national crash datasets are made available to the general public for free. There are three major national datasets of crashes: 1) the Fatality Analysis Reporting System (FARS); 2) the National Automotive Sampling System—Crashworthiness Data System (NASS-CDS or CDS); and 3) the National Automotive Sampling System—General Estimates System (NASS-GES or GES). FARS is a census of fatal crashes on public roads in the US. CDS is an annual probability sample of approximately 3500-4500 tow-away crashes involving light vehicles. The CDS data collection includes in-depth crash investigation and estimation of Delta-V using the software WinSmash (an enhanced and updated version of the accident reconstruction software CRASH3), as well as details on injury outcome. Finally, GES is an annual probability sample of approximately 50,000 police reported crashes. The basis for the data in GES is information contained in state police crash reports, but the data elements are coded to a national standard. To perform estimates of injury risk for occupants of US vehicles, the CDS dataset was used because it contains measures of crash severity and injury level. GES and FARS were used for analysis of crash avoidance.

The *German In-Depth Accident Study (GIDAS)* is the largest database of its kind in Europe. Data collection commenced in 1999 and was initiated by the German Federal Highway Research Institute (BASt) and the German Association for Research on Automotive Technology (FAT), which unites all German passenger and commercial vehicle manufacturers as well as numerous suppliers. Today GIDAS has 16 sponsors who have exclusive access to the database. Crash data is collected by two teams, one at the Hannover Medical School (MHH) and one at the Traffic Accident Research Institute (VUFO) of Technische Universität Dresden (TU Dresden). A statistically developed sampling plan defines the work shifts for the teams, which cover 12 hours per day. If an accident occurs with at least one injured person suspected, the GIDAS team is notified directly by the local police or rescue service via radio communication. Sample criteria for the GIDAS database are that at least one accident participant has been injured and the accident occurs within the shifts and the specified regions. After 15 years of continuous data collection, the database includes over 28,000 injury crashes (i.e. crashes in which at least one person was injured) investigated in-depth. Delta-V values are reconstructed using a method based on the conservation of momentum in the crash, predominantly with the software PC-Crash.

The *Cooperative Crash Injury Study (CCIS)* is a major crash database in Great Britain in which data collection, funded by the UK Department for Transport and industrial partners, started in 1983 and ended in 2009. The sponsors have exclusive access to the database, which contains more than 15,000 crashes. Crash events are collected according to a stratified sampling procedure, which favors cars containing fatal or seriously injured occupants. More specifically, the inclusion criteria in CCIS require that at least one passenger car which is younger than 7 years has been involved in the crash (or younger than 5 years if the injured occupant was only slightly injured) and towed from the scene and that at least one crash-involved occupant was injured, according to the police report. Data were collected retrospectively (several days after the crash) by teams of investigators from Birmingham Automotive Safety Centre (BASC) based at the University of Birmingham, Vehicle Safety Research Centre (VSRC) based at Loughborough University, Transport Research Laboratory (TRL) and Vehicle Operations and Standards Agency (VOSA) from various locations in England. Delta-V reconstruction in CCIS is damage-based, using the software AI-Damage which is, similar to WinSmash, based on the CRASH3 algorithm.

VOIESUR (Véhicule Occupant Infrastructure Etudes de la Sécurité des Usagers de la Route - Vehicle Occupant Infrastructure and Road Users Safety Studies) is a project funded by the French National Research Agency and Foundation MAIF. In this project, a database of more than 9000 crashes is built from the in-depth analysis of police reports in France in 2011. More specifically, the database contains the following crashes from 2011: all fatal crashes in France, 5% of the injury crashes in France, and every crash in the Rhône region. Data from the Rhône region are used to develop case weights for the remainder of the dataset but were not used for analysis in this study. The data come from expert investigations of police reports, sketches and photos. However, police-coding of variables is not automatically accepted – instead, police information is used to understand what happened. Delta-V in the crash is reconstructed using a method based on the vehicle trajectories when there is sufficient data available to do so. The VOIESUR database has been developed by a consortium of four French research organizations: CEESAR[1], CEREMA[2], IFSTTAR[3] and LAB[4], and the agreement of all members is required for data access.

PENDANT, the *Pan-European Co-ordinated Accident and Injury Database*, was developed between 2003 and 2005 in a project co-funded by the European Commission. The main objective of PENDANT was to support EU vehicle and road safety policy making. The resulting database contains approximately 1100 crashes collected in eight EU countries (Austria, Germany, Spain, Finland, France, The Netherlands, Sweden and the United Kingdom). An inclusion criterion is that at least one vehicle occupant was injured in the crash. A further requirement was that at least 20% of cases from each country to be of MAIS 3+ injury severity and a maximum 10% of the required case-load for each partner could comprise pedestrian crashes. Although an inclusion criterion was that the crash includes a vehicle with model year 1998 or later, only a subset of the crashes in the dataset will meet the current project's model year restriction of 2003 and later. In addition, PENDANT crashes from Germany and the UK will not be used because it is possible those cases would be duplicated in other datasets being analyzed. (Note that PENDANT crashes in France are not duplicated in VOIESUR because of the different data collection period.) Figure 1 shows the countries that contribute to PENDANT, including those that were not used in the current analysis.

[1] Centre Européen d'Etudes de Sécurité et d'Analyse des Risques

[2] Centre d'étude et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement

[3] Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux

[4] Laboratoire d'Accidentologie, de Biomécanique et d'Etudes du comportement humain

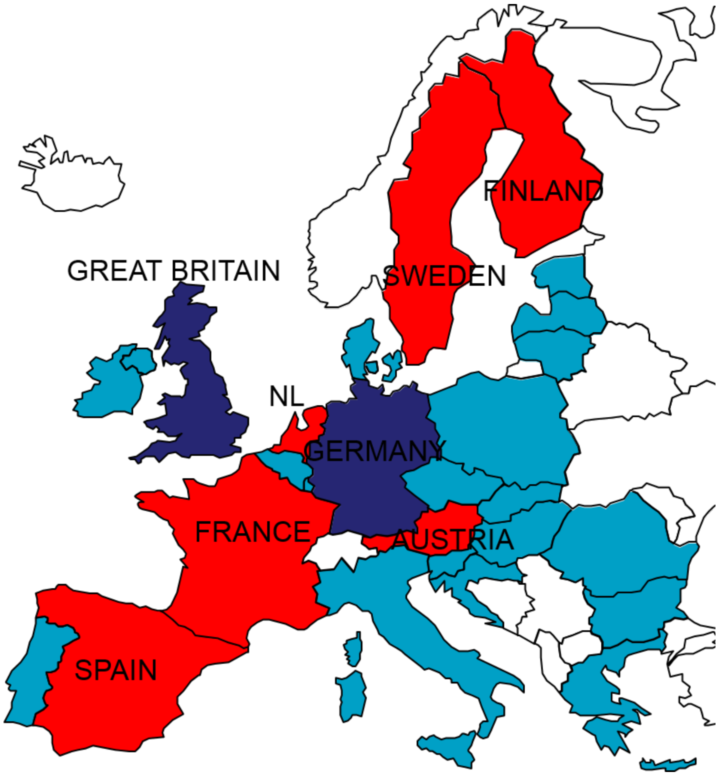

Figure 1. Summary of PENDANT dataset. Red: data collected and used in analysis. Dark blue: data collected but not used to avoid possible duplication. Light blue: EU country not included in PENDANT dataset.

The *INTACT* database was developed in consecutive research projects funded by the vehicle industry, the Swedish Governmental Agency for Innovation Systems (VINNOVA), the Intelligent Vehicle Safety Systems (IVSS) program, the European Commission (EC) and the Swedish Research Council (VR). These projects included both methodology development and data collection addressing different applications. The INTACT methodology, developed in the IVSS-funded project *Investigation Network and Traffic Accident Collection Techniques* during 2007-2010, was adapted by the EU project *Road Safety Data, Collection, Transfer and Analysis* (DaCoTA) in 2010 as the method to be used for in-depth crash investigation on a European level (Hill et al. 2012). Data collection using this methodology is ongoing in a VR-funded project; the resulting database currently contains approximately 300 crashes. Data collection is conducted in Gothenburg, Sweden and the six surrounding municipalities; the inclusion criterion is that at least one passenger car, bus or truck was involved in the crash and an ambulance was called to the crash scene. The software PC-Crash is used for Delta-V reconstruction.

Due to the relatively small number of cases in INTACT, the dataset is not being analyzed in the same manner as the other European datasets because there are not enough injured occupants in the dataset to estimate injury risk once the inclusion criteria are applied. However, the cases in the INTACT dataset contain enough information to compute Delta-V using both crush-based and trajectory-based measures. Thus this dataset was used to identify possible differences in estimating Delta-V using different methods.

To provide more cases for Delta-V harmonization, the *Road Accident Sampling System – India* (RASSI) dataset was also used. The data collection was established by JP Research in 2011 and is supported by a consortium of automotive OEMs and suppliers. In-depth, on-scene data collection is conducted in two cities in India, Coimbatore and Prune. The geographical area covers urban, semi-urban and rural regions including national and state highways. Inclusion criteria for data sampling are that at least one motorized vehicle involved and the accident must have happened on public road within sampling region. To date the database includes more than 400 cases with about 700 coded variables. Accident reconstruction is conducted with PC-Crash. The crush profiles of accident vehicles are also measured according to the NASS field investigation protocol in a retrospective vehicle investigation.

Figure 2 summarizes the in-depth crash data sources used in this study. Eastern Europe is not well represented in the in-depth datasets that were available for this study. As a result, a weighting system was developed to adjust observations to better represent the whole of the EU. This process is described later in the methods section.

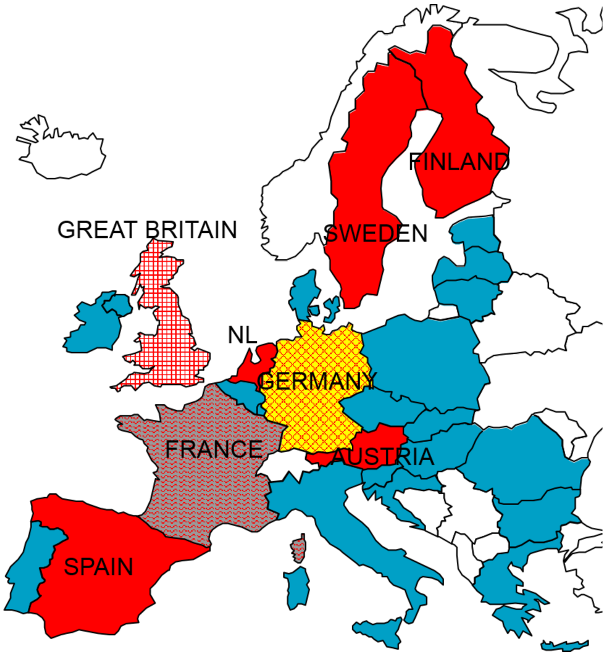

Figure 2. Summary of EU data sources. GIDAS (yellow), CCIS (hatched red), VOIESUR (wave red), PENDANT (red). Remaining EU countries are shown in blue.

The most comprehensive source for aggregated national crash data (police-reported crashes) in the EU is the Community Road Accident Database (CARE) which contains national data from all 28 EU countries plus Iceland, Liechtenstein, Norway and Switzerland. CARE has no data collection activity of its own but the data come from the member states; such data are re-coded according to uniformization protocols (CAREPLUS and CADaS) to obtain a standardized data set. CARE does not contain Delta-V or MAIS values since those are generally not included in national crash data, and the inclusion criteria are typically less restrictive than those of in-depth databases. In this study, we used CARE to develop weighting factors for the EU datasets being analyzed in detail so they estimate the injury risk for all EU countries and to conduct crash avoidance analyses.

Harmonization

Harmonization of both inclusion criteria and variable definitions is a critical element of the methods for this study. To develop comparable risk models for both crashworthiness and crash avoidance, we must ensure that the datasets are sampled from comparable populations and that variable definitions are comparable as well.

Key Variable Definitions

Crash type. When categorizing crashes by impact direction, rollovers were first extracted. Any rollover, whether the first or subsequent event and whether tripped or untripped, was defined as a rollover crash. Remaining crashes were classified according to the deformation location (first letter) of the initial impact using the Crash Damage Classification (CDC) code according to SAE J224, which was used in all datasets. A crash was designated a frontal if the CDC area of deformation was F. For side impacts, crashes were also classified with respect to occupant location, since the analysis is conducted at the occupant level. A right-side impact with right-side occupant or left-side impact with left-side occupant was classified as a near-side crash. Alternatively, if the occupant was on the opposite side of the vehicle from the worst impact, that occupant's crash was classified as a far-side crash. Occupants in the same vehicle can have different crash classifications under this system.

Delta-V. Delta-V measures the change in speed (in km/h) in a crash experienced by each vehicle's occupants, taking into account the relative masses of the vehicle and its crash partner. Delta-V in all datasets was estimated (rather than measured by a crash data recorder) and two different types of estimation methods were used. In CDS, CCIS, and some PENDANT cases, a crush-based reconstruction was conducted, where Delta-V is derived from the deformation energies of the collision partners, their masses, and their stiffness. For GIDAS, VOIESUR, and some PENDANT cases, a trajectory-based reconstruction was conducted which calculates the Delta-V from the difference of immediate post- and pre-crash momentum.

To compare the methods, we used Delta-V values derived from the same cases using different methods. Cases were extracted from two datasets. In the Swedish INTACT database, trajectory-based reconstruction was conducted with PC Crash, but the crush profile and position were also measured and coded according to the WinSmash and AI Damage protocols. In the Indian RASSI database, data were also available to allow comparison of Delta-V calculations using both methods. When data were input into WinSmash from these datasets, the generic stiffness coefficients based on wheelbase and vehicle type (Sharma et al. 2007) were used to characterize the stiffness of vehicles, since these are European or Indian vehicles, which do not have specific stiffness coefficients in WinSmash.

Comparison of Delta-V calculated using different methods was performed separately for frontal and side impacts. Data were available to calculate Delta-V using both methods for 35 frontal impacts and 14 side impacts. This is shown in Figure 3.

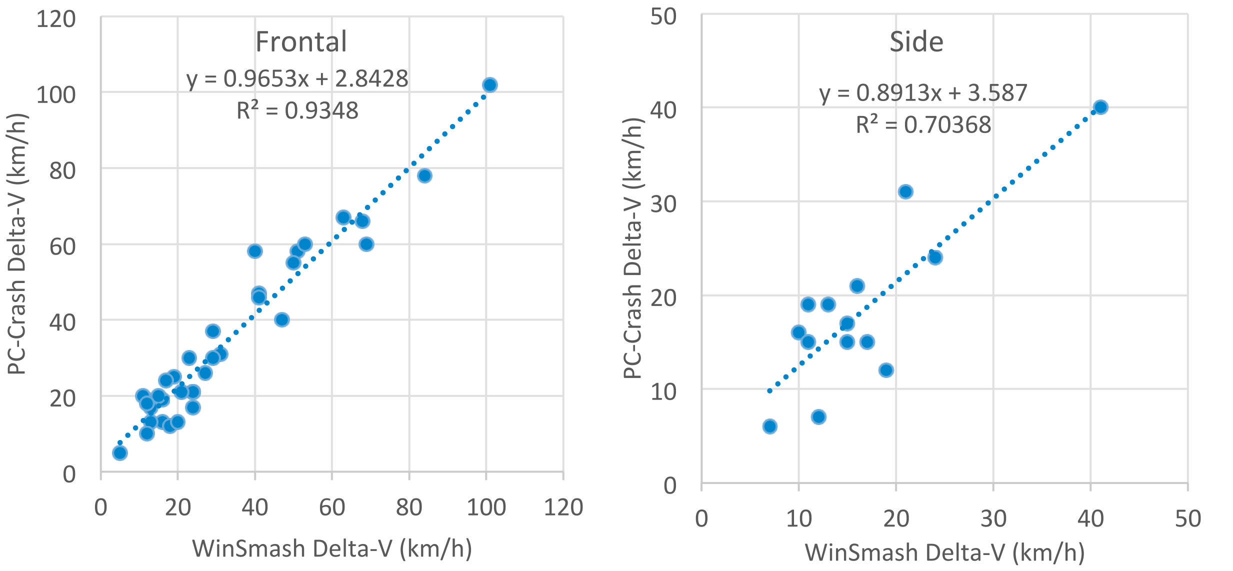

Figure 3. Relationship between Delta-V calculated with PC-Crash and with WinSmash for frontal (left) and side (right) impacts.

The methods produced similar results in both crash directions, and the relationships between them fell relatively close to the identity line (especially in front impact). However, to maximize comparability of the methods, we applied a transfer function to the crush-based reconstruction cases to assign a Delta-V value to those cases that was better harmonized with trajectory-based Delta-V values. The functions applied are shown on the graphs in Figure 3.

PDOF. No PDOF restrictions were applied. However, PDOF was grouped into categories as 0^o, 30^o, or $>30^o$ relative to the direction perpendicular to the side of impact. The 0^o category covers force directions from -15^o to $+ 15^o$, the 30^o category from ($+15^o$ to $+45^o$) and (-15^o to -45^o) and the $>30^o$ category every angle greater than +/- 45^o, For example, frontal PDOF are relative to a head-on crash of 0^o, while side impacts are relative to a T-bone crash of $\underline{+}$ 90^o.

Occupant age. Initial analyses considered categorical groupings of age. However, the final model used age as a continuous predictor, measured in years.

Occupant gender. Gender was divided into males and females. Pregnant females were grouped with other females.

Belt Use. Belt use was divided into 3-point belt or no belt use. Unknown, lap only, and shoulder only were not included in the analysis. Because some datasets did not indicate whether the 3-point belt was used properly, all use of 3-point belt was included.

Intrusion. Intrusion definitions varied among datasets. To harmonize the categories used in analysis, levels of none, minor, and major intrusion were categorized using the definitions from each dataset listed in Table 1. Maximum intrusion to the front, left, right, or roof were noted separately. Intrusion to the front was coded over the whole length of the front, independent of the occupant's position.

Intrusion in side impact was defined only for the same side as the occupant's position. Thus, in far-side impacts, intrusion would usually be considered "none" unless intrusion was so large it passed the centerline of the vehicle. This was done to ensure that as a predictor, intrusion would better reflect a mechanism of injury, rather than serving as a proxy for crash severity (which would be true if struck-side intrusion were used as a predictor for far-side impacts). For rollovers only the roof intrusion was considered.

Table 1. Definitions of intrusion level from each dataset (cm).

	None	**Minor**	**Major**
CDS	0	3-15	16+
PENDANT	0-5	6-15	16+
GIDAS	0-5	6-15	16+
CCIS	0-5	6-15	16+
VOIESUR	0%	1-25%	25%+

Vehicle Type. Because of the limited overlap in vehicle types between the EU and US passenger-vehicle fleet, vehicle types were split into those with 7 or more seating positions and those with 6 or fewer.

Vehicle Model Year. Vehicle model year was initially grouped in two-year increments from 2003-2013 for exploratory analysis. In the final risk models, model year was grouped into two groups: 2003-2006 and 2007 and later. The cut point between 2006 and 2007 was chosen because of the upgrade to US FMVSS No. 208 that took effect in 2007. In some of the EU datasets, the year of first registration was used instead of vehicle model year. In CCIS, both registration and model year were available, and they were the same year in over 98% of cases. The primary purpose of the model year variable was to allow the model to adjust for changes in regulation over time. Although some regulations were phased in, the model year of a vehicle represents the regulations in effect at the time, and each vehicle met (or exceeded) those regulations at that time.

Crash Partner. Coded crash partner varied across datasets. To limit the total number of categories and ensure similarity of the variable across datasets, crash partner was categorized into: passenger vehicles, wide objects (including heavy good vehicles & busses), narrow objects, and other.

Light Condition. For light conditions, dusk, twilight, and day were grouped in one category; the other category was night. Because the CCIS dataset does not explicitly code light condition, the time of crash relative to sunset or sunrise for a given date were used to classify crashes as light or night.

Ejection. Partial ejection, complete ejection, and ejection to an unknown degree were considered ejected. All other known cases were considered not ejected.

Crash Location/road type. The definition and available categories for road type varied the most widely across datasets. To address variation, final categories of crash location were labeled "rural" and "urban" as defined in in Table 2.

Table 2. Definitions of crash location/road type from each dataset.

	Rural	Urban
CDS	Undivided road with speed limit > 40 mi/h	All other roads
PENDANT	("Local area" rural) or ("Local area" mixed, "carriageway type" motorway and speed limit >90 km/h) or ("Local area" mixed, "carriageway type" not motorway and speed limit >50 km/h)	("Local area" urban) or ("Local area" mixed, "carriageway type" motorway and speed limit <=90 km/h) or ("Local area" mixed, "carriageway type" not motorway and speed limit <=50 km/h)
GIDAS	Out of city	In city
CCIS	Speed limit > 40 mi/h	Speed limit ≤ 40 mi/h
VOIESUR	Outside urban area	Inside urban area

Selection Criteria

After key variables were harmonized, all cases were filtered according to uniform selection criteria as follows.

1) All crashes have at least one occupant sustaining an AIS-1+ level injury or fatality. While CDS does not require injury for inclusion in the dataset, most of the EU datasets do. Occupants with unknown injury severity are excluded.
2) All crashes have at least one vehicle that was towed, based on the CDS inclusion requirement.
3) Delta-V is known (does not apply to rollovers, for which delta-V is not estimated).
4) All crashes have at least one vehicle with damage severity greater than level 1 according to the collision damage classification (CDC) coding. This requirement was added because the VOIESUR dataset does not have a variable indicating towaway. Analysis of other datasets indicated that applying this damage-level criterion was a reasonable substitute for filtering towaway crashes.
5) Vehicle model year is 2003 or later (because US datasets no longer include vehicles older than 10 years at the time of the crash).
6) Front impacts, side impacts, and rollovers are included.
7) Occupant ages 13 years and older (to maximize dataset size while eliminating those using child restraints); age must be known.
8) Occupants in driver or outboard front passenger seating position; position must be known.
9) Belt use status must be known.
10) Only ECE-class M1 passenger cars with equal or less than 9 seats inclusive the driver (EU) or passenger vehicles (body type code <50) in US.

The decision was made not to include rear impacts in the dataset because there were very low numbers of occupants with MAIS3+F injuries in rear impacts. In addition, there were insufficient rear impacts to determine a relationship between Delta-V calculated by trajectory-based and crush-based methods.

The injury outcome under consideration in analyses is a Maximum Abbreviated Injury Scale score of 3 or greater or a fatality (MAIS3+F). This injury level was selected because it is typically used for regulatory analysis to define targets and assess vehicle performance. In addition, after selection criteria were applied, the total unweighted sample size was sufficiently large to use this outcome as shown in Table 3.

Table 3. Unweighted number of eligible cases with known Delta-V in each dataset.

Dataset	Unweighted N Front-Side	Unweighted N Rollover	Unweighted N MAIS3+F Front-side	Unweighted N MAIS3+F Rollover
CDS (US)	9,245	1,877	1019	447
PENDANT (AT, ES, FI, FR, NL, SE)	89	35		
VOIESUR (France)	503	55		
GIDAS (Germany)	2,131	112		
CCIS (Great Britain)	723	367		
Total EU	3446	569	448	123

Weighting European Datasets

The main use of CARE in the crashworthiness analysis is to specify the standard crash-involved occupant population in Europe and develop weighting factors for the EU datasets being analyzed in detail so they estimate the injury risk for all EU countries. The goal of the weighting process is to address the concern that the European data used for detailed injury analysis come from more westerly/northerly, wealthier countries. The data from Great Britain, France, and Germany need to be weighted to better represent the EU as a whole, while making sure weighting factors are reasonable.

The CARE dataset was reviewed to identify a set of variables with high quality that are available for most countries and are present in each in-depth database considered in the analysis. The relevant set of variables include urban/rural area, motorway (y/n), junction (y/n), vehicle registration year, road surface conditions, and lighting conditions.

While hypercube clustering is frequently used to identify weighing factors, small-N or even empty cells in the dataset are often problematic when using this approach. With a small number of cases in a cell there is the risk that the distribution of values is biased which leads to large and overly influential weights. Instead, decision tree algorithms allow the classification of the data in an effective way by specifying separation variables and one target variable. Stop criteria for data splitting, e.g. minimum bucket size, can be set to avoid clusters that are too small. The GIDAS data and the R software package RPART were used to derive the decision tree. As the decision tree had to be applied later to CARE, CCIS, PENDANT, and VOIESUR data, the only filter criteria was set to passenger cars, as this could be identified in all data sets. The following separation variables were selected as they were available in the CARE database and all of the in-depth databases:

- Accident location & road type (motorway, rural road, urban road)
- Light condition (day, night, twilight, unknown)
- Vehicle registration year (1960-1992, 1993-1997, 1998-2002, 2003-2006, 2007-2014)

The variables were recoded as dummy variables with possible values of zero or one, so that twelve binary variables were available for data separation. The injury severity (uninjured, slight injury, severe injury, fatal injury) was used as the target variable for data splitting. In the RPART control settings the minimum cluster size was set to 200 with a complexity parameter of zero. An additional parameter for the classification function was the use of the generalized Gini index of impurity for the splitting index.

The application of the decision tree to the GIDAS data resulted in 14 categories. Except for PENDANT, which is a relatively small database, each of the clusters of the other databases remained a meaningful sample size.

Not all countries in the CARE database included in the weighting had complete data on all weighting variables. Missing data points were estimated using the relationships between known variables from a "similar country". More precisely, if data for country X was unavailable at a given node, the distribution of the splitting variable in a "similar country" Y was used as a substitute to distribute the number of occupants for country X that were present at the given node between the children nodes. Similarity was measured using the so-called χ^2-distance (Niebuhr et al. 2011), which was computed for each injury severity level, and the country with the smallest χ^2-distance was taken as a substitute.

Each of the countries' crash populations was reviewed to identify which countries had similar characteristics based on the following variables:

- Area type (Urban/rural),
- Junction (Crash in junction Yes/No)
- Road condition (Dry road Yes/No)
- Light conditions

In particular, for each crash severity level and all pairs of EU countries, χ^2-distances were computed separately using the following sets of variables:

1) Area type, Light conditions, Junction;
2) Area type, Junction, Road condition;
3) Area type, Light conditions, Road condition.

Then, for each pair of countries, the average of these distances were computed (i.e. the average of set 1, 2 and/or 3 values depending on which combinations of variables were available for both countries in CARE). The resulting values provide a measure of similarity between EU countries with respect to the distributions of the above variables: small χ^2-distances indicate similar distribution while large χ^2-distances suggest substantial differences. Having defined a similarity metric this way, the country with the smallest[5] χ^2-distance was taken as a substitute for a country with missing data, and thus the cluster distribution in CARE was computed.

In-depth accident data was weighted in two steps to achieve better representation of US and EU level crash data as indicated in Figure 4. In a first step, each national dataset used in this project was weighted to represent that country, using weights and weighting approaches developed specifically for those datasets. The CCIS dataset was weighted to national crash statistics for Great Britain using

[5] The smallest among the set of those countries that had all variables available along the unique path in the tree from the root to the given node

national Stats19 data. To develop weights, the CCIS sampling criteria were matched using accident type (frontal, nearside, farside, rollover) and casualty severity. For GIDAS, weighting to the German DeStatis data (national statistics) was computed by hyper-cube clustering on the accident level for each data collection year separately using the accident severity (slight, severe, fatal), accident type (seven categories defined by the institute for road traffic in Cologne, ISK), and accident time (day, night, twilight). PENDANT was weighted to a regional level for all 8 countries where data were collected by computing at an occupant level the injury severity and type of area (rural-urban). VOIESUR was weighted to France's national levels to consider underreporting of crashes using injury severity, type of road user (car occupant, P2W, bike, other or pedestrian), the number of vehicles in the accident (one or more than one), type of police force investigating the accident, type of area (rural-urban), and road type.

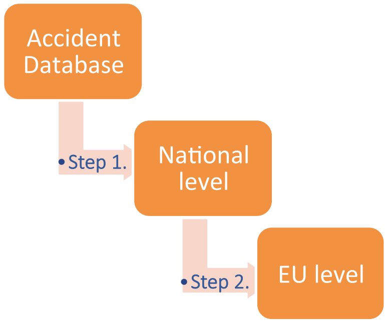

Figure 4. Steps used in weighting EU datasets.

In a second step, EU weighing factors were derived by application of the decision tree categories. The relative category size and their relative injury severity distribution in each national representative in-depth data set was compared to CARE to define the weights that scale the national weighted dataset to a EU standard population. If a dataset did not contain any cases in a category (such as PENDANT for categories with model years 2007+), it was not necessary to define weights. For nonempty categories the weights were computed using the following formula:

$$EUweight_{dataset[\,inj.sev,category]} = \left(\frac{freq_{CARE[inj.sev,category]}}{freq_{CARE[total]}}\right) \Big/ \left(\frac{wtdfreq_{dataset[inj.sev,category]}}{wtdfreq_{dataset[total]}}\right)$$

Total weights for description of the EU standard population through the in-depth data samples have been computed by the multiplication of the national weights and the EU weights. The variables that are used to generate the decision tree are defined as follows (the ones in boldface were included in the final decision tree shown in Figure 5):

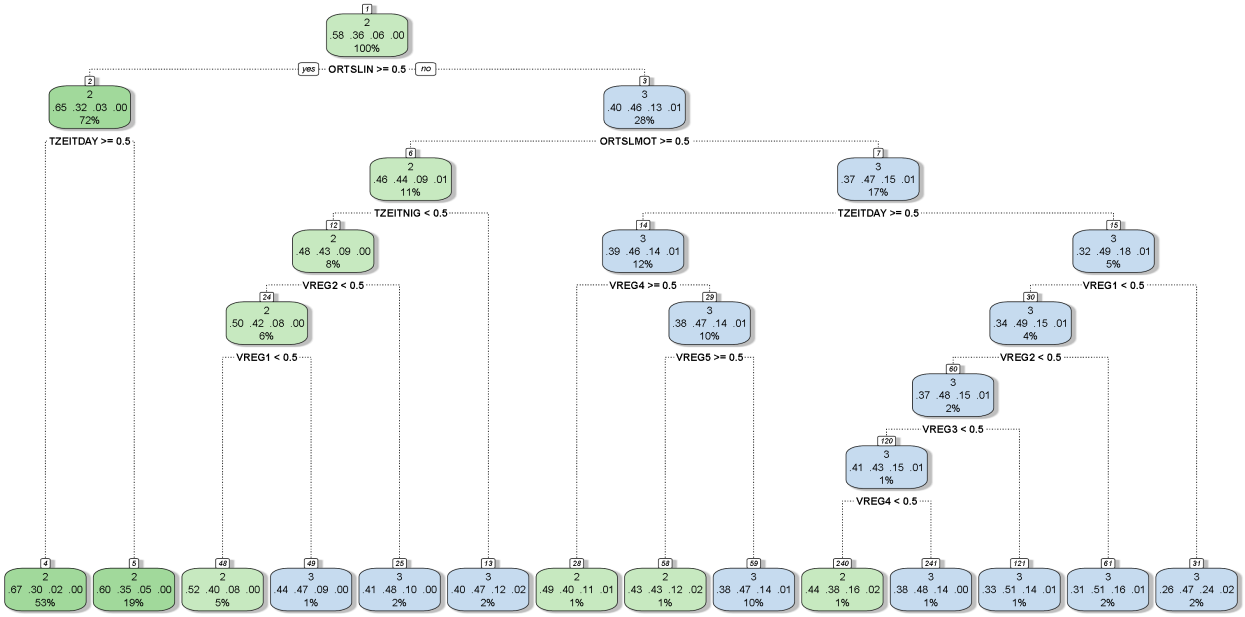

Figure 5. Decision tree used to determine EU weights.

Accident location

ORTSLIN = People involved in accidents inside the city, exclusive motorway;

ORTSLOUT = People involved in accidents outside the city, exclusive motorway;

ORTSLMOT = People involved in accidents on motorways;

Accident light conditions

TZEITDAY = People involved in accident during daytime;

TZEITNIG = People involved in accidents during night time;

TZEITTWI = People involved in accidents during twilight;

TZEITUNK = People involved in accidents during unknown light conditions;

Vehicle registration

VREG1 = People in vehicles registered between 1960 – 1992;

VREG2 = People in vehicles registered between 1993 – 1997;

VREG3 = People in vehicles registered between 1998 – 2002;

VREG4 = People in vehicles registered between 2003 – 2006;

VREG5 = People in vehicles registered between 2007 – 2014;

VREG6 = People in vehicles with unknown registration.

Figure 6 shows in green the countries for which CARE data was used for the weighting. The countries colored red had insufficient data to be used, either because there were many variables with missing data or unknown rate higher than 20%.

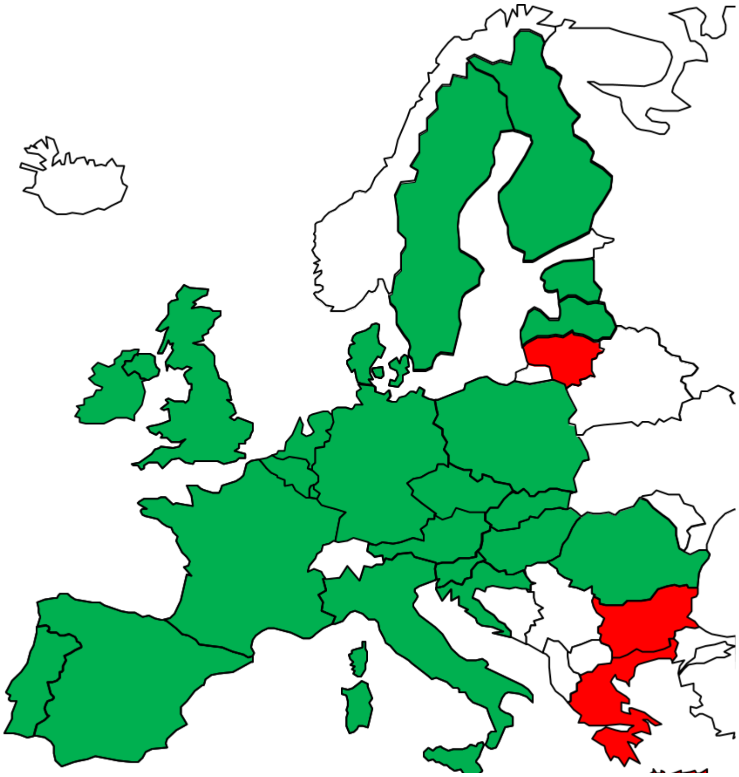

Figure 6. EU countries used in weighting (green) and not used (red).

In the final EU models, the weights applied in Step 2 were normalized to the sample size of each dataset before the results were combined (see Appendix B for details). This meant that although the cases in each dataset were weighted to represent the distribution of these crashes in the EU, the contribution of each dataset to the model was commensurate with the raw number of cases included rather than the size of the country they represent.

Table 4 shows the end result of weighting for both datasets. For the EU combined dataset, the weighted sample injury rate was cut in half or more for both front/side and rollover, relative to the raw injury rate in the samples. The high injury rate in the raw sample results from oversampling of injury and fatal crashes in some of the EU datasets. Similarly, the US injury rate for the unweighted samples are more than double the weighted injury rates. Note that after weighting, the sample injury rate is higher in the EU compared to the US. This indicates that the population of crashes defined by the inclusion criteria is more severe in the EU, on average, than the same population of crashes in the US.

Table 4. Weighted and unweighted sample MAIS3+F injury rates

Dataset	Unweighted Sample Injury Rate Front/Side	Unweighted Sample Injury Rate (Rollover)	Weighted Sample Injury Rate Front/Side	Weighted Sample Injury Rate (Rollover)
EU Combined	0.130	0.216	0.053	0.105
US	0.110	0.238	0.033	0.074

Standard Populations

Standard populations are needed for Methods 2 and 3 to provide a testbed of crash-involved occupants, complete with their crash, vehicle, and occupant characteristics, one for each driving environment (EU and US). The standard population represents the crash environment that would be encountered by any vehicle driving in a region, and it allows for a side-by-side comparison of predicted risk for the two models.

In the Phase 1 report (Flannagan et al., 2014), we planned to use a standard population for the EU based on the GIDAS dataset and a simplified weighting scheme that is used for the EuroNCAP Advanced Technology Assessment. However, in looking more closely at that approach, we saw benefits in the usage of the development datasets weighted to the EU populations as described in the previous section. This ensures that the EU standard population is as large as possible and generally representative of EU crashes. It also significantly simplified the assessment of the models procedurally.

Since the EU weighting was based on the latest available data year possible, more recent years were also used to define US standard population as well. Thus, CDS for crash years 2007-2012 were used for this purpose.

Maximum Likelihood Models of Injury Risk

All of the statistical methods in this report make use of the likelihood surface associated with models of injury risk. The models considered are constrained to be logistic regression models, which are a type of general linear model. Details of logistic regression are given in Appendix A. However, some key information is presented here.

The logistic regression equation is given in Equation 1 below.

$$\hat{p} = 1 \Big/ \left(1 + e^{-\sum_{i=0}^{r} \hat{\beta}_i x_i}\right) \qquad (1)$$

where $\hat{p}$ is predicted risk of MAIS3+F injury, x_i is the *i*th predictor value, and $\hat{\beta}_i$ is the *i*th coefficient in the model (i=0..r).

A given set of coefficients (β_i) define a single model. Using those coefficients, predicted risk for that model can be computed for every observation in a dataset, and these can, in turn, be compared to the actual outcome.

To choose from among the infinitely many possible models, we use likelihood as a way of scoring the comparison between the predicted risk and the actual outcome. Likelihood is the probability of getting the data that were observed, given the model under consideration. Although there may be a small level of dependence between occupants of the same vehicle, we treat each observation in these datasets as independent, and thus the probability of the data given the model is the product of the probabilities of each outcome under the model. For our applications, it is easier to use the log of the likelihood, which is shown in Equation 2.

$$\mathcal{L} = \sum_{j=1}^{n}(y_j \log(\hat{p}_j) + (1 - y_j)\log(1 - \hat{p}_j)) \qquad (2)$$

where $\mathcal{L}$ is log likelihood, y_j is the outcome (1 is MAIS3+F injured; 0 is MAIS<3) and $\hat{p}_j$ is the predicted risk of injury for the *j*th observation.

Log-likelihood can be thought of as a score with higher values indicating that the observed data are more likely to have been observed. The parameters in each model describe the relationship between predictors and outcome.

A common use of likelihood is to select the model associated with the highest likelihood as the best model of the observed data. This approach is called *maximum likelihood* and is the model selection method for a wide variety of statistical models. In Methods 1 and 2 in this analysis, we select the maximum likelihood models for the US data and the EU data and evaluate them. In Method 3, we compare the likelihoods for a variety of models that fall into groups of hypotheses about the risk difference between the two groups of vehicles (EU-regulated and US-regulated).

One of this project’s challenges was the inability to share and combine raw data from the EU datasets. In a typical analysis using logistic regression, raw data in a single file would be analyzed using statistical software that takes advantage of efficient iterative search techniques to find the maximum likelihood. In this project, we could only calculate log-likelihood using summary statistics from the component EU datasets. As a result, we had to compute log-likelihood for a large number of alternative models within the space of all possible models. The assessment of log-likelihood for the large set of alternative models facilitated the search for both the best-fit model for the EU dataset (used for Methods 1 and 2) and the computation of Bayes Factors in Method 3. That is, log-likelihood was calculated for 193,563 possible models for front/side and 164,865 possible models for rollover. The highest log-likelihood among these models was selected as the best-fit model and the remaining models were used to compute the Bayes Factors (Method 3) comparing evidence for different levels of overall risk differences.

Unlike the EU datasets, the US dataset is a complex sample survey requiring specialized methods for estimating variance. To find the best-fit US model, we used SAS PROC SURVEYLOGISTIC with Taylor series estimation of the variance-covariance matrix to account for clustering and stratification in the sample design. Weights provided with the dataset were applied. Fortunately, Method 3 does not depend on variance estimates, but simply measures the probability of the observed outcomes for each model. Thus, for Method 3, the same process was applied to the weighted US dataset as for the EU datasets.

The details of the computation of log-likelihood and the variance-covariance matrix for the US and EU models are provided in Appendix B. The next sections describe each of the three likelihood-based methods used in this project.

Method 1: Seemingly Unrelated Regression

Seemingly Unrelated Regression (SUR) was proposed by Zellner (1962) and used in Gordon et al. (2011) with Poisson regression to assess whether two models built on different datasets are different. The extension of SUR to logistic regression used in this project is described in detail in Appendix C. Conceptually, the SUR framework creates a single model for all of the data, in which separate parameters are estimated for the US occupants and the EU occupants. Hypothesis tests focus on the null hypothesis of the form that one or more coefficients of common predictors for the two models are the same. Although not all predictors must be the same for SUR, in this application, we used a single set of predictors for both populations. This way, hypothesis tests were conducted to compare each individual parameter estimate for the two populations. In addition, we tested the multi-degree-of-freedom (multi-df) null hypothesis that *all* parameters are the same. This null hypothesis is equivalent to stating that the risk models for EU and US vehicles are the same (as a whole).

The first step in this process was to identify the final set of predictors to be used in the models. To do this, individual logistic regression models were built on each of the five datasets. Models were built separately for front/side crashes and rollovers since Delta-V is not estimated for rollovers in many of the datasets. Front and side crashes were analyzed together to increase the sample size. The starting set of predictors for the front/side model were Delta-V (log and square transformations considered), crash type (front, near side, far side), age, age squared (to allow a quadratic relationship), belt use, road type, vehicle type, model year group, PDOF (relative to side of impact), intrusion (relative to side of impact), airbag deployment, crash partner, and the presence of multiple impacts. In addition, interactions of Delta-V and crash direction were also considered. For each dataset, non-significant model parameters were dropped. In marginal cases, changes to Akaike Information Criteria (AIC) were considered in deciding whether to include a parameter. The final set of predictors included any predictor that was significant in any of the individual models. In addition, model year (grouped into 2003-2006 and 2007+) was retained, even though it was not significant in any individual model, to account for regulatory changes that occurred during that time. Airbag and vehicle type were not significant in any model, and transformations of Delta-V were not found to be appreciably better than Delta-V as a linear predictor. Vehicle mass was evaluated as a potential predictor but does not improve the model once other variables that contain the effect of vehicle mass (notably, delta-V) were included. Variables included in the model were also checked for collinearity.

The initial set of predictors for the rollover model were age (including a quadratic term), gender, roof intrusion, ejection, belt use, road type, model year, light condition, and seat position. Seat position and light condition were not significant and were eliminated from the final parameter set. The final predictors for both models are listed in Table 5.

Table 5. Final predictors in logistic regression models for all methods

Front-Side Model	Rollover Model
Delta-V	Age (no quadratic term)
Age, Age squared	Belt Use (Belted, Unbelted)
Crash Type (Front, Far, Near)	Intrusion (None, Minor, Major)
Belt Use (Belted, Unbelted)	Model Year (2003-2006, 2007+)
Delta-V*Far	Ejection (Ejected, Not Ejected)
Delta-V*Near	Accident location (Rural, Urban)
Intrusion (None, Minor, Major)	Gender (Male, Female)
PDOF 30 (0, 30, >30)	
Crash Partner (Car, Narrow, Wide, Other)	
Model Year (2003-2006, 2007+)	
Accident location (Rural, Urban)	

Log-likelihood was computed for a large number of combinations of parameter values for each of the models (front-side and rollover) for each vehicle population (US and EU). The parameters for the models with the highest log-likelihood were selected as the best models for SUR. The variance-covariance matrices for each of the SUR analyses were constructed based on the best models as described in Appendix C.

Using the model parameters and variance-covariance matrices, hypothesis tests were conducted as follows. First, simple tests of H_0: $\beta_{i_{EU}} = \beta_{i_{US}}$ for each parameter, *i*, were conducted, where r=17 (in the terminology of Equation 1) for front-side, r=8 for rollover, and *i=0* indicated the test of the intercept. This resulted in 18 tests for the front-side models and 9 for the rollover models. Finally, the multi-df test of the null hypothesis that all of the parameters are the same was conducted (H_0: $\beta_{0_{EU}} = \beta_{0_{US}}, \beta_{1_{EU}} = \beta_{1_{US},} \ldots, \beta_{r_{EU}} = \beta_{r_{US}}$).

Method 2: Best Models Applied

The null hypothesis tested in Method 1 represents a strong definition of equivalence of EU and US injury risk. Under that definition, the predicted injury risk for each occupant will be effectively the same, regardless of the population of crashes. Even without this level of equivalence, it is possible for *overall* predicted injury risk to be the same for a given population of crashes. Method 2 investigates this less stringent definition of equivalence by applying the best models to standard populations for the EU and the US.

There were two elements to Method 2. First, we assess the best models side-by-side on each standard population to compute the mean and variance of the estimated risk difference between the two models across the *whole* of each population. Second, we looked at the estimated injury risk for specific subsets of each population (such as frontal or side impacts and rollovers) to understand the nature of any differences in predicted risk. Risk differences were applied rather than risk ratios, because risk ratios can cause mathematical problems when baseline risks are close to zero. Also, a given risk difference always means the same thing in terms of associated number of injuries, regardless of baseline risk.

Assessment of the overall mean and variance of predicted risk differences relied on asymptotic normality of predicted injury risk in these models. This result is shown in Appendix D.

Method 3: Bayes Factors

Method 3 uses Bayes Factors, which are ratios of evidence for two different hypotheses. Evidence is measured as the likelihood of the data, given a hypothesis. In this application, a hypothesis is defined in terms of a particular risk difference between EU and US vehicles for one standard population. Since we are interested in evaluating evidence with respect to equivalence, we compare each risk-difference hypothesis to the hypothesis of no difference.

One particular benefit of this approach is that it does not assume a null hypothesis (as Method 1 does). Instead, it compares the evidence for each of a number of hypotheses about the true state of the world. It is also not influenced by the particular coefficients of the best model, but instead reflects the extent to which the likelihood surface as a whole is very peaked (i.e., a great deal of evidence for a few models at the peak and much less evidence for other models) or relatively shallow (i.e., many models are similarly likely to have produced the observed data). Thus, as a companion to the other methods, Method 3 provides a different view of the information available to us.

The basic equation for Bayes Factors is shown in Equation 3.

$$B_{i0} = \frac{p(\boldsymbol{D}|H_i)}{p(\boldsymbol{D}|H_0)} \tag{3}$$

where B_{i0} is the Bayes Factor comparing a hypothesized risk difference of *i* to a risk difference of zero, ***D*** is the observed data, H_i is the group of models that result in a risk difference of *i*, and H_0 is the group of models that result in a risk difference of zero. (Zero actually denotes an interval around zero whose width is agreed upon based on a reasonable definition of practically no difference.) Note that the hypothesis of zero risk difference is not treated as a null hypothesis in the same way as in Method 1. However, it is treated as the comparison hypothesis for all other hypotheses. In principle, any risk-difference hypothesis can be compared to any other risk-difference hypothesis using this method.

In applications such as this one, each hypothesis can be represented by a large number of specific models. For example, many models in this space result in zero risk difference, and many other models result in a risk difference of 0.001. In this situation, the probability of the data given the hypothesis is shown in Equation 4.

$$p(\boldsymbol{D}|H_k) = \int p\,(\boldsymbol{D}|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k \tag{4}$$

where θ_k is a set of coefficients (i.e., a model) that result in a risk difference of *k*, and $\pi(\theta_k|H_k)$ is the prior probability of θ_k given the hypothesis H_k.

The direct computation of Equation 4 can be difficult, especially on a large dataset. As a result, Bayes Factors are generally estimated rather than computed directly. Different estimation approaches employ different approaches to defining the prior probabilities. However, in this analysis, we have no clear means of assigning prior probabilities, and thus prefer an estimation method for which priors will have little or no effect on the estimated Bayes Factors. The specific estimation approach we selected is the Schwarz Criterion, which is ideal for this application because 1) it uses log-likelihood, which we already need to compute for a large set of models for Methods 1 and 2; and 2) it does not make strong assumptions about the prior probability of each model within a hypothesis. Instead of introducing prior probabilities for each potential model, the Schwarz Criterion uses the log-likelihood of the *best* model within each hypothesis. Further details on the estimation of the log Bayes Factors in this study are given in Appendix E.

Crash Avoidance

During the course of the project, four areas of technologies and regulations related to crash avoidance have been considered:

- Headlamps
- Mirrors
- Electronic Stability Control (ESC)
- Brakes and stopping distance

Data in the relevant EU and US datasets are only sufficient for the analysis of headlamps (in relation to pedestrian crashes at nighttime versus daytime) and mirrors (where the analysis is based on the proportion of lane-change/merge crashes to the driver's side versus the passenger's side). Therefore, the corresponding issues and analysis approach will be described in greater detail. As for ESC and Brakes, a brief description of the planned analysis method and the obstructions preventing the execution of the analysis will be provided.

Headlamps

Analyses based on dark/light risk ratios of crashes around Daylight Savings Time (DST) changes have been used to investigate headlamp performance with respect to pedestrian safety; the basic idea is described in Sullivan and Flannagan (2007) as follows: *"The influence of natural light on crash risk is determined by the dark/light risk ratio—the number of crashes in a certain period of darkness divided by the number of crashes during a comparable period of daylight. A dark/light ratio greater than 1 indicates that darkness is more risky than daylight […] If we suppose that some improvement in artificial lighting at night could create conditions more like daylight, we would expect the dark/light ratio to approach 1."* Note that this analysis does not consider the effect of glare.

A key element of the analysis conducted in Sullivan and Flannagan (2007) is that the changes between DST and winter time create a 1-hour time period in which the light conditions just before the time change and right after the change are different while other factors are essentially unchanged. In particular, it is assumed that exposure before and after the change is similar (motivated by the argument that the number and distribution of road users is to a large extent governed by the hour of the day; for example, by the work hours). This way, the effect of light condition can be isolated from the other factors.

Sullivan and Flannagan (2007) showed that the influence of light conditions is most obvious for the number of pedestrian fatalities. Therefore, dark/light ratios for pedestrian fatalities in the US were compared to similar ratio in the EU. Under the working hypothesis that a dark/light ratio closer to 1 corresponds to a better imitation of daylight conditions, a comparison of the US and EU dark/light ratios can be related to headlamp performance.

The regional datasets relevant for the analysis are FARS in the US and CARE containing EU data. In these databases, the number of pedestrian fatalities in crashes with the involvement of passenger cars by hour and by light condition was investigated for the crash years 2007-2012. This time period was chosen because CARE has complete data for the greatest number of EU countries for this period; this way, there were eight EU countries included in the analysis: Austria, France, Greece, Poland, Portugal, Romania, Spain, and the UK[6]. Sweden and Finland were excluded from the analysis due to the different geographical positions and the resulting different light patterns compared to other EU countries.

Data in CARE contains month and hour of the crash only (not day or minute); therefore, instead of the actual 1-hour period with the changed light conditions (which can be specified by the precise times of sunrise or sunset), whole hours and full months were considered. DST ends on the last Sunday of October in EU and the first Sunday in November in the US (since 2007), hence the months October and November were analyzed. This approach was replicated using FARS for the US.

As a first step of the analysis, we examined whether the light conditions in the EU and the US are comparable in the different 1-hour periods in October and November. This was done by considering the light conditions in all crashes in the given periods in GES and CARE (here and later, only the above specified eight EU countries are considered in the analysis). Not only fatalities are considered here because the quantity of interest is purely the categorization of light condition in different time periods and not the usual attributes of crashes (such as injury severity). The results in Figure 7 and Table 6. show that while the general patterns are similar, there are differences in the light conditions. This can also be measured by a comparison of the odds ratios of light in October versus November in the EU and the US as shown in Table 6. This table also shows that the light conditions are most affected by the time change in the time frame 4:00pm-6:59pm.

Table 6. Ratio of the relative odds of light in October by the relative odds of light in November in the US and eight countries of the EU by hour, based on the classification of light conditions in road crashes in GES and CARE, respectively

Hour	Relative odds of light Oct / Nov in the US	Relative odds of light Oct / Nov in the EU
1:00pm-1:59pm	2.00	1.35
2:00pm-2:59pm	3.01	1.60
3:00pm-3:59pm	2.90	5.08
4:00pm-4:59pm	13.23	16.98
5:00pm-5:59pm	21.89	14.81
6:00pm-6:59pm	9.54	11.96
7:00pm-7:59pm	2.23	4.47
8:00pm-8:59pm	0.93	1.70
9:00pm-9:59pm	1.24	1.19

The differences in the ratio “relative odds of light in October by relative odds of light in November” between EU and US may affect the dark/light ratio in a way that is difficult to quantify. To eliminate this

[6] Data years 2009-12 are used for the UK because of coding errors for the light condition variable for the years 2007-08.

effect, only ”Light” crashes are considered in October and only ”Dark” crashes are considered in November. These crashes happen approximately in the same time period due to the time change, which means that the exposure is still similar. However, slight changes (within time period and geographical distribution of crashes) are possible. To minimize the change in the temporal and geographical distributions, the 6:00pm-6:59pm time interval was used because that interval gave the smallest difference between the ratios ”relative odds of light in October / relative odds of light in November” in the EU and US in the relevant time frame.

The number of pedestrian fatalities in crashes involving passenger cars for the given light condition (light for October and dark for November) was queried from FARS (US) and CARE (EU). These databases contain a census of fatalities, hence no weighting is required[7].

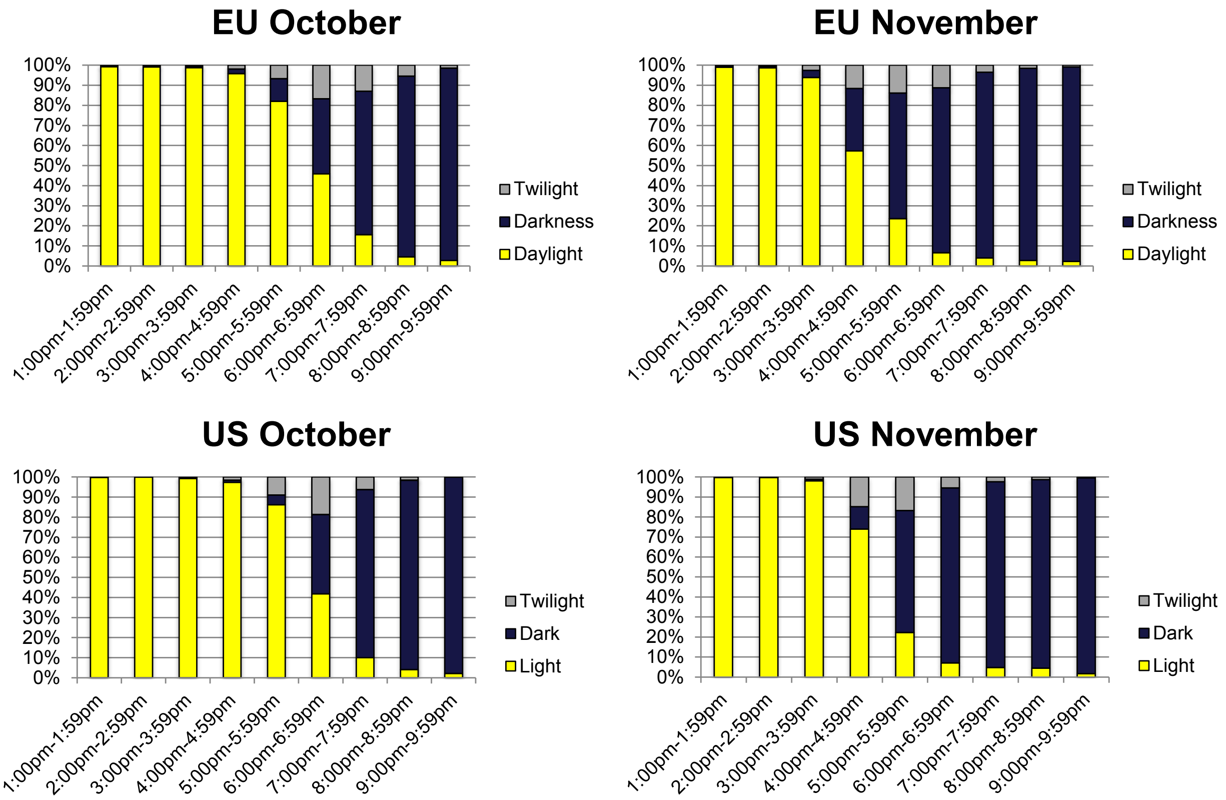

Figure 7. Proportion of light, dark, and twilight by hour for the EU and US in October and November.

The ratios and the relevant confidence intervals were computed using the same terminology as in http://www.biostat.umn.edu/~susant/Fall10ph6414/Lesson14_complete.pdf,

where the computation of odds ratios in a case-control study is described by the following table:

[7] Correction factors are used in CARE to make up for the differences in the definitions used by the EU countries. In particular, the data for Spain and Portugal have been multiplied by a correction factor to comply with the 30-day rule for the registration of road fatalities (death within 30 days of the crash as a consequence of the crash) used in the other EU countries.

	Cases	Controls
Exposed	a	b
Not exposed	c	d

The odds ratio can be computed using the following formula:

$$OR = \frac{a}{b} \div \frac{c}{d}$$

The confidence interval is computed on the logarithmic level and the exponents are taken afterwards. As a result, the confidence interval given in the formula below is not symmetric about OR.

$$\mathrm{Exp}\left(\mathrm{Ln(OR)} \pm 1.96 * \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

Using the corresponding formulas gives an estimate and a 95% confidence interval for the odds ratio. The result is significant if and only if the interval does not contain one. *Note, however, that a non-significant result is not synonymous with a conclusion of "no difference"*; Hauer (2004) discusses the fallacies of the .05 significance testing in the context of traffic safety.

Mirrors

The second aspect of crash avoidance that was investigated is concerned with side mirrors, because of a difference in regulatory requirements. While in the EU the mirrors on both the driver's and the passenger's side are non-planar, the driver-side mirror in the US is planar while the passenger side may be non-planar (and generally is). The reason for this difference is that both mirror types have advantages compared to the other type: according to Luoma et al. (2000), the blind zone with a non-planar mirror is smaller than with a planar mirror; at the same time, the distances and relative speeds of the vehicles approaching from behind are easier to assess using a planar mirror. The difference in the driver's side mirrors in the EU and the US becomes relevant in lane-change maneuvers. Therefore, the analysis is based on a comparison of lane-change crashes to the driver's side in US and EU using passenger-side lane-change crashes as control.

The first step in the analysis is the identification of the relevant data elements in the EU database CARE, including the identification of EU countries with non-missing data. A careful reading of (CADaS, 2013) shows that the potentially relevant variables in CARE for identification of lane change crashes to the left/right are the following:

Accident type A-11: At least two vehicles – no turning;

U-11: Traffic unit maneuver.

Unfortunately, the potentially relevant types A-11.01 and A-11.03 do not differentiate between changing lane to the left and changing lane to the right; therefore, only the pre-crash maneuver U-11 could be used, with the potentially relevant values marking the direction of movement being:

U-11.13 Changing lane to left;

U-11.14 Changing lane to right;

U-11.16 Overtaking vehicle on its left;

U-11.17 Overtaking vehicle on its right.

It is unclear *a priori* whether U-11.16 and U-11.17 are relevant for the analysis. This could be checked using crash data from Great Britain where the first point of impact is available in CARE. The corresponding data shows that "overtaking" crashes have substantially higher percentage of frontal impacts than "changing lane" crashes (48% vs 27%) which indicates that these pre-crash maneuvers result in crash types with different characteristics. Hence, exclusively U-11.13 and U-11.14 were used in the analysis.

The US database used for this analysis was GES. Right and left lane changes leading to a crash were coded in the accident type variable (acc_type codes 46 and 47).

There are only two EU countries that register "Changing lane to left" and "Changing lane to right", namely Portugal and the United Kingdom, and data years 2010-12 are available. Moreover, the UK data made up 90% of the sample from the two countries and thus dominates the results. Consequently, the results do not represent the EU broadly. The quasi-induced exposure analysis was performed to give an indication of performance differences related to differences in the driver-side mirror *per se*, but the reader is cautioned that we do not know how differences in overtaking behavior in the UK and US might influence the results.

ESC

Only a preliminary analysis was conducted for ESC. The analysis plan was to compare rollover and single-vehicle run-off-road rates for model year groups before and after ESC was installed. The model year groups would be identified using fleet penetration estimates, because datasets do not reliably indicate the presence of ESC on vehicles. However, when examining the CARE dataset, only the 2012-2013 datasets had crash records indicating the vehicle registration year and the type of crash including run-off-road or rollover. Results were only available for a limited number of countries (Finland, Latvia, Luxembourg, UK had both rollover and run-off-road data; Germany and Hungary had only run-off-road). Information about fleet penetration of ESC indicated that only Germany has higher than 80% penetration for model year 2007, but Germany has had greater than 50% fleet penetration for all model years since 2003. An effort was made to group vehicle registration year ranges into 1999-2002 and 2010-2013 using the available data from crash years 2012-2013. However, the driver population of vehicles in these categories may have very different characteristics, which might influence the distribution of single crashes vs. general crashes. For example, a young driver might be less likely to afford a newer vehicle, and is also more likely to be in a run-off road crash. Therefore, the data available do not support the planned analysis.

Brakes and Stopping Distance

The initial plan was to analyze vehicle brake performance between the two regions. Unfortunately, brake failure information and stopping distance in crash events were not sufficiently available in the datasets to allow any comparisons. The regulations related to brake performance have achieved a high level of harmonization in the US and EU. In particular, Antilock Braking Systems (ABS) have been required in the EU since 2007 while in the US, ABS are required in conjunction with Electronic Stability Control per FMVSS 126 as of mid 2011.

A Note on the Role of Significance Tests

The Phase I final report (Flannagan et al., 2014) included a detailed discussion of the role of hypothesis testing in this project (also see Hauer, 2004). To summarize, the difficulty with standard hypothesis testing in this context is that it is designed for questions such as "Are two groups different?" but not for questions about whether two groups are the same. To address that question, which is posed here, we have to find other ways to convey the comparison of different hypotheses.

A p-value, or test of significance, measures the probability that one would have gotten the statistics computed from the observed data *if the null hypothesis of no difference were true.* It does not measure the probability that the no-difference hypothesis is true. Thus, when p=0.20, this is not equivalent to there being an 80% chance of no difference being true. In fact, for a test where p=0.20, there will be many highly likely hypotheses (various differences), and the most likely one will be the difference that was observed (regardless of the significance test).

For questions of equivalence, there is no clear comparable approach. Thus, we take several different approaches to identify what the evidence favors. In the case of Method 1, we do use significance tests to directly compare the two models. The Phase 1 Final report (Flannagan et al. 2014), describes how we proposed to take into account the probability of accepting the null hypothesis when the alternative (the two models are different) is, in fact, true. (This is called Type II error and the probability of avoiding a Type II error is called the power of the test.)

Method 2 relies on measurement of variance to describe what we know about risk differences between the two models (for a given population). Here, we do not present p-values for parameters because we do not include and remove parameters on the basis of p-values (parameter selection is described in the Method 1 section on p. 27). If we were to do this, fewer parameters would reach significance in the smaller (EU) dataset, regardless of the true value of the parameters. Instead, all of the variance not accounted for by the model is captured in the variance estimate. Thus, the greater the goodness-of-fit, the smaller the variance in the estimate of overall risk for a population. In addition, the larger the sample size, the smaller the variance. Since the US dataset has a larger sample size, there is likely to be less uncertainty in the overall predicted risk.

To help interpret the Method 2 results, we can construct confidence intervals (CIs) on the risk difference distribution. However, the standard interpretation of confidence intervals—that a CI containing 0 indicates a non-significant result—is based on the same logic of hypothesis testing. Thus, in this study, CIs are intended only to convey the magnitude of uncertainty. A CI containing 0 is not evidence *for* the null hypothesis. Other approaches to interpretation are presented in the results section, including one modeled after the bioequivalence testing (e.g., Committee for Medicinal Products for Human Use, 2010).

Finally, Method 3 treats all hypotheses equally. Although we compare each difference hypothesis to the no-difference hypothesis, the measurement of evidence is done exactly the same way for all hypotheses. Thus, we measure evidence, form a ratio, and use that to identify hypotheses that are more or less likely than the zero-difference hypothesis. Such ratios could be formed between any two hypotheses.

pIn general, we try to minimize use of significance testing because of the nature of the question being asked. The three places it arises are: 1) in Method 1, which tests whether there is difference between coefficients for the EU and US models, 2) in the initial selection of parameters for inclusion in the models, and 3) in the analyses of crash avoidance (which are also subject to the same problems of interpretation as for crashworthiness). The reader is cautioned against interpreting failure to reach significance as evidence that there is no difference.

Results

Maximum Likelihood Injury Risk Models

Table 7 lists the coefficients of the best EU and US models for the front-side and rollover populations. The front-side models use 18 coefficients, while the rear models use 9 coefficients.

Table 7. Coefficients of best models

Variable	**EU: frontal/side**	**US: frontal/side**	**EU: Rollover**	**US: Rollover**
Intercept	-6.099	-9.353	-3.386	-4.454
Delta-V	0.072	0.075		
Age	-0.075	0.073	0.014	0.027
Age*Age	0.081	-0.031		
Far	0.715	-1.522		
Near	0.759	-0.353		
Unbelted	0.361	1.498	2.145	0.866
Delta-V*Far	0.037	0.069		
Delta-V*Near	-0.024	0.050		
Intrusion: minor	0.662	1.249	-0.835	0.268
Intrusion: major	1.790	1.607	0.447	0.693
PDOF 30	-0.344	0.141		
PDOF >30	-1.692	-0.509		
Partner: narrow	1.171	1.227		
Partner: wide	2.363	0.789		
Partner: other	1.115	1.036		
Model year 2007+	-0.413	-0.175	0.069	-0.557
Rural	1.383	0.598	0.385	0.637
Ejection			1.587	1.740

Method 1: Compare Injury Models

The results of Seemingly Unrelated Regression hypothesis testing for the front-side crashes are shown in Table 8. Nine of the eighteen coefficients were significantly different between the US and EU injury models with $p< 0.05$. If the coefficients are not significantly different, it means that after accounting for all the other variables, we cannot reject the null hypothesis that the variable has a similar effect on injury for both US and EU vehicles. The test evaluating the overall model (simultaneous comparison of all coefficients) was also significantly different with $p< 0.0001$.

Table 8. Comparison of EU and US coefficients in injury model for frontal/side crashes

Variable	Chisquare	DF	P-value	Conclusion
Intercept	14.30	1	0.00015	Sig
Delta-V	0.14	1	0.71	NS
Age	24.20	1	<0.0001	Sig
Age*Age	22.20	1	<0.0001	Sig
Far	5.00	1	0.025	Sig
Near	1.20	1	0.26	NS
Unbelted	6.80	1	0.0089	Sig
Delta-V*Far	4.90	1	0.027	Sig
Delta-V*Near	3.90	1	0.047	Marginal
Intrusion: minor	0.15	1	0.70	NS
Intrusion: major	0.47	1	0.49	NS
PDOF 30	2.40	1	0.12	NS
PDOF >30	2.50	1	0.12	NS
Partner: narrow	0.84	1	0.36	NS
Partner: wide	10.30	1	0.0014	Sig
Partner: other	0.01	1	0.92	NS
Model year 2007+	0.76	1	0.38	NS
Rural	3.10	1	0.078	Marginal
All (including intercept)	141.3	18	<0.0001	Sig

The results summarizing differences between EU and US model coefficients for rollovers are shown in Table 9. The only variable that is significantly different is the unbelted parameter, but its significance was sufficiently high (p=0.0029) that the overall models were also significantly different (p=0.00016). The odds ratio for unbelted vs. belted occupants in rollovers is higher for EU vehicles compared to US vehicles.

Table 9. Comparison of EU and US coefficients in injury model for rollover crashes

Variable	Chisquare	DF	P-value	Conclusion
Intercept	2.70	1	0.099	NS
Age	1.50	1	0.21	NS
Unbelted	8.90	1	0.003	Sig
Intrusion: minor	4.00	1	0.046	Marginal
Intrusion: major	0.34	1	0.56	NS
Model year 2007+	1.90	1	0.17	NS
Rural	0.28	1	0.60	NS
Ejection	0.08	1	0.78	NS
Female	0.03	1	0.86	NS
All (Including intercept)	32.50	9	0.0002	Sig
All but Unbelted	12.00	8	0.15	NS

Method 2: Apply Best Models

The goal of Method 2 is to estimate the risk difference for EU and US vehicles (as represented by their injury risk models) in the EU (i.e., using the EU standard population) and the US (i.e., using the US standard population). We arbitrarily define risk difference in all cases as EU risk minus US risk. Thus, negative risk differences indicate that estimated risk for EU vehicles is *lower* than that of US vehicles, and positive values indicate that estimated risk for EU vehicles is *higher* than that of US vehicles. Because the distribution of estimated risk is asymptotically normal, the difference between the two distributions is also normal.

Figure 8 shows the distributions of estimated overall population injury risk for EU and US front-side injury risk models applied to the US front-side standard population, while Figure 9 shows the EU and US front-side injury models applied to the EU standard population. The resulting distributions of risk differences are shown in Figure 10 for the US population and Figure 11 for the EU population. Note that in Figure 8, the EU population includes a non-trivial proportion of cases at 0. This occurs because asymptotic normality is violated for this dataset and the variance is large enough that the distribution should extend into negative values and must be cut off at 0.

When applied to the US front-side standard population, the mean estimated risk for the US-vehicle model is 0.035 with a standard deviation of 0.012, and the mean estimated risk for the EU-vehicle model is 0.023 with a standard deviation of 0.016. The most likely risk difference is -0.012, indicating that risk would be lower on the US front-side population when the EU model is applied. The standard deviation of the risk difference is 0.020 and the 95% CI is (-0.051, 0.027). This mean risk difference represents a 33% reduction in risk for EU vehicles over the US mean injury rate.

To illustrate a possible way of interpreting the figures taken from the bioequivalence literature (e.g., Committee for Medicinal Products for Human Use, 2010), the blue-shaded box represents an *arbitrarily* defined region of “essential equivalence.” The boundaries shown here, from -0.02 to +0.02 risk difference, are for illustration only—values used in application must be determined by agreement. (Some guidelines for selecting boundaries are discussed in the section on fleet penetration models on page 59.) In this example, since 59% of the area under the curve lies within the blue box, there is a 59% probability that the risk difference lies between -0.02 and +0.02.

When applied to the EU front-side standard population, the mean estimated risk for the US-vehicle model is 0.052 with a standard deviation of 0.025, and the mean estimated risk for the EU-vehicle model is 0.065 with a standard deviation of 0.027. As shown in Figure 11, the most likely risk difference is -0.013. The standard deviation of the predicted risk difference is 0.037 and the 95% CI is (-0.084, 0.059). There is a 39% probability that the risk difference falls between -0.02 and +0.02.

Comparable results for the rollover models are shown in Figure 12 through Figure 15. The rollover models applied to the US population are in Figure 12 and the rollover models applied to the EU population are in Figure 13. For the US standard population, the predicted mean risk is 0.071 (sd=0.024) for the US-vehicle model and 0.128 (sd=0.057) for the EU-vehicle model. The best estimate of the risk difference applied to the US population is 0.057, with a standard deviation of 0.062. The 95% CI is (-0.064, 0.179). As shown in Figure 14, only 17% of the area below the curve falls within the range of -0.02 to +0.02.

For the EU rollover standard population shown in Figure 15, the mean predicted risk for the US-vehicle model is 0.067 (sd=0.024) and for the EU-vehicle model the mean is 0.103 (sd=0.040). The most likely risk difference is 0.037, with a standard deviation of 0.047. The 95% CI is (-0.055, 0.128), and 25% of the area below the curve falls within the range of -0.02 to +0.02.

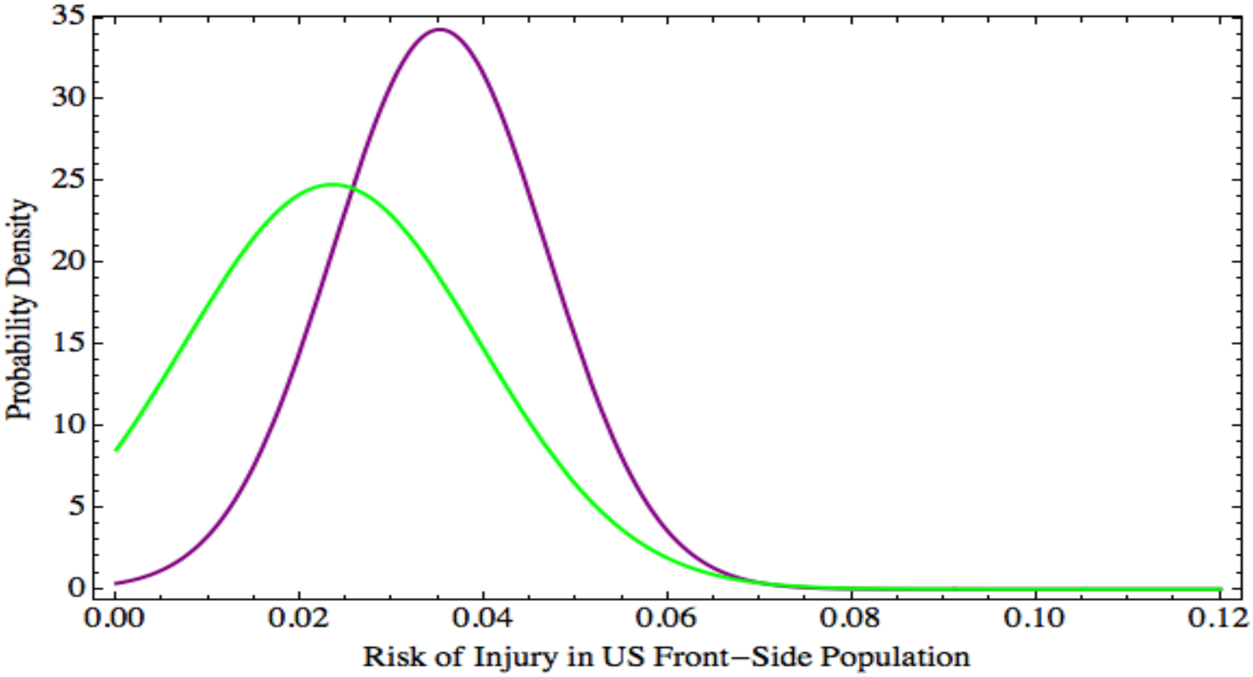

Figure 8. EU (green) and US (purple) front-side injury models applied to the US front-side population.

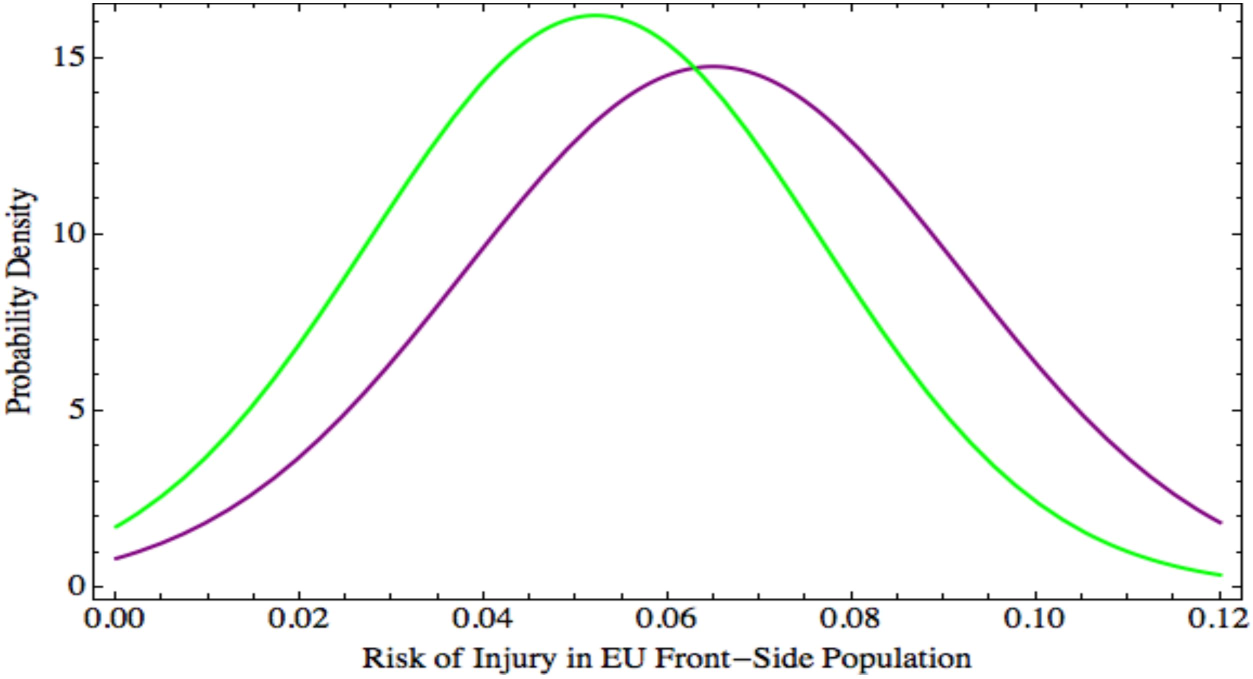

Figure 9. EU (green) and US (purple) front-side injury models applied to the EU front-side population.

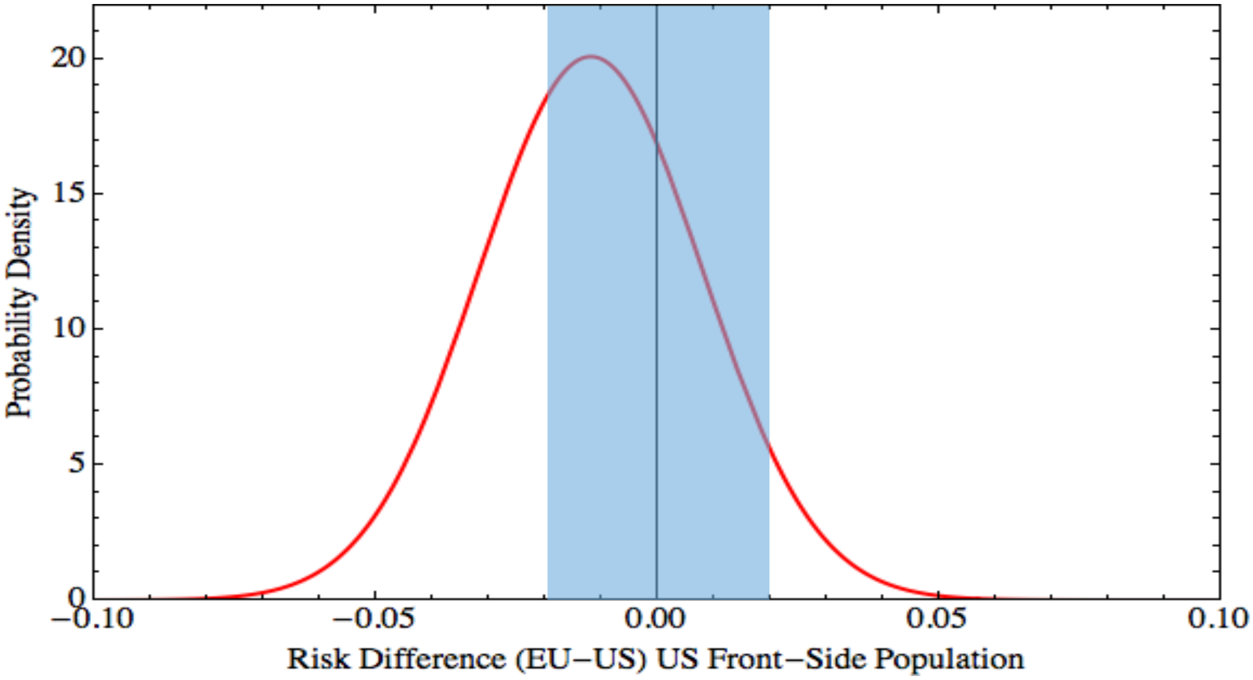

Figure 10. Difference in risk between EU and US models applied to the US front-side population.

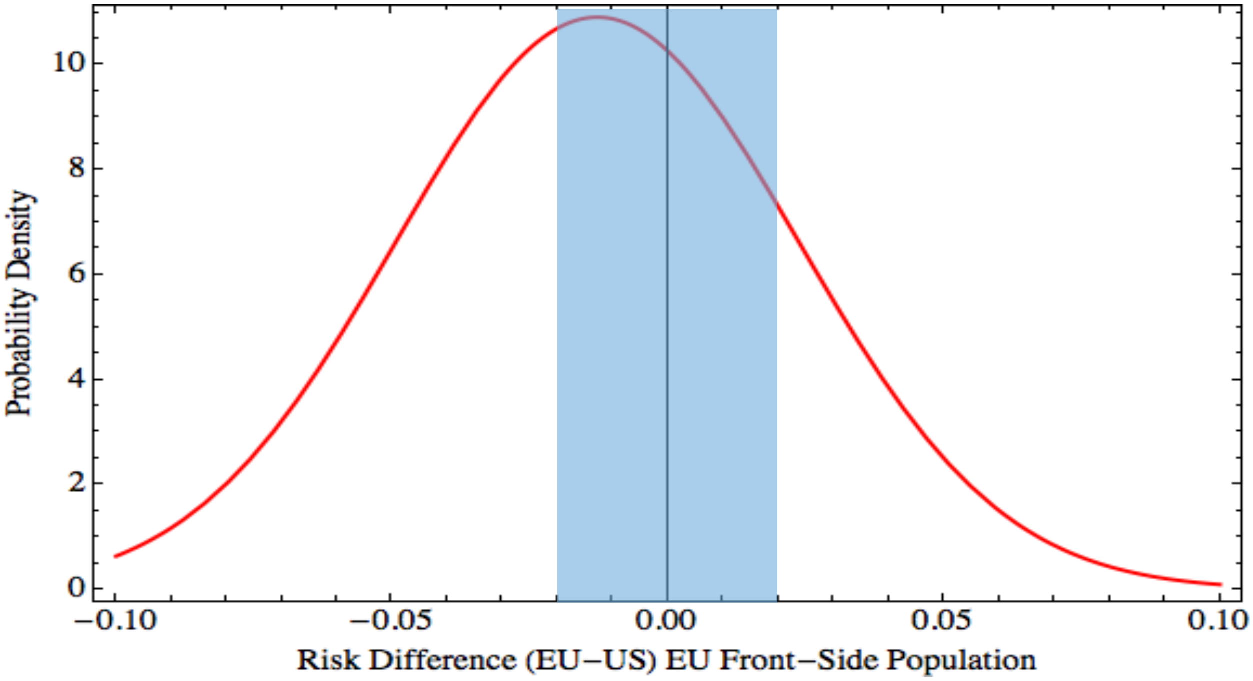

Figure 11. Difference in risk between EU and US models applied to the EU front-side population.

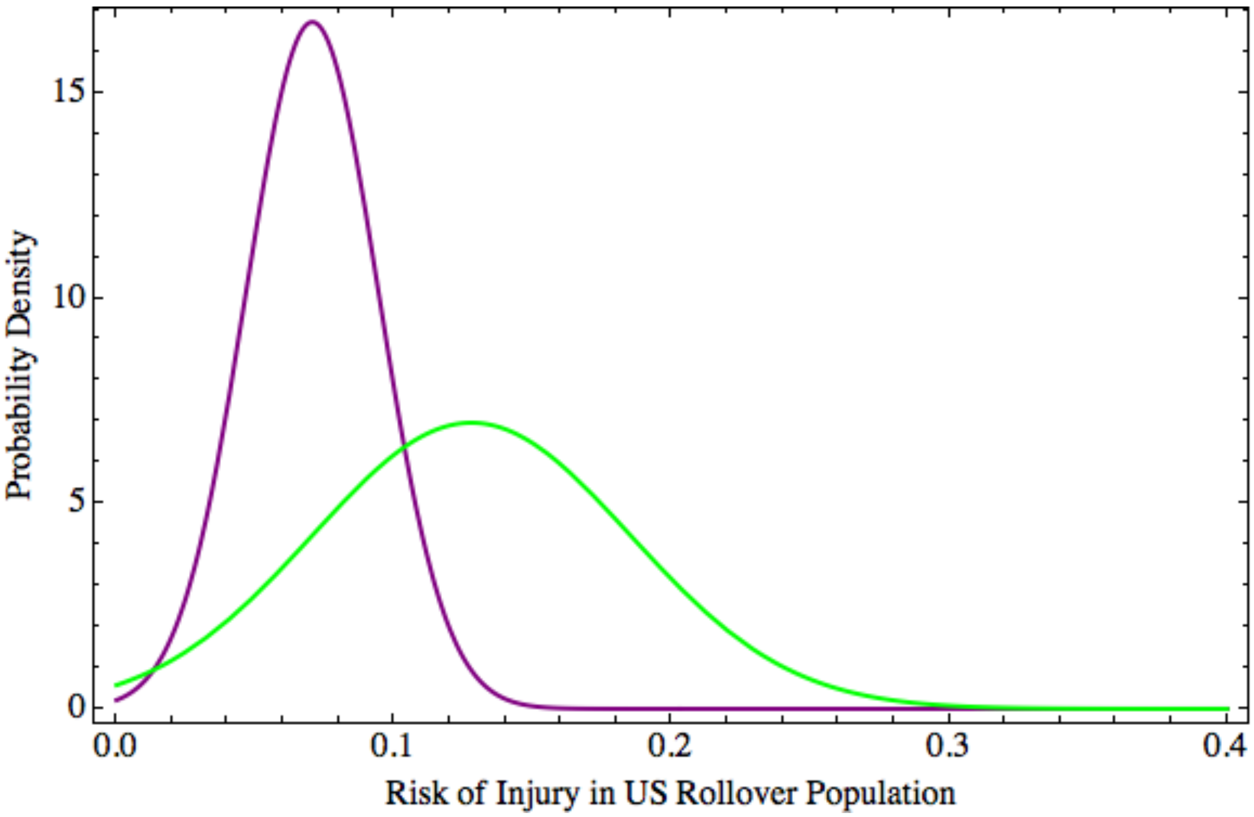

Figure 12. EU (green) and US (purple) rollover models applied to the US rollover population.

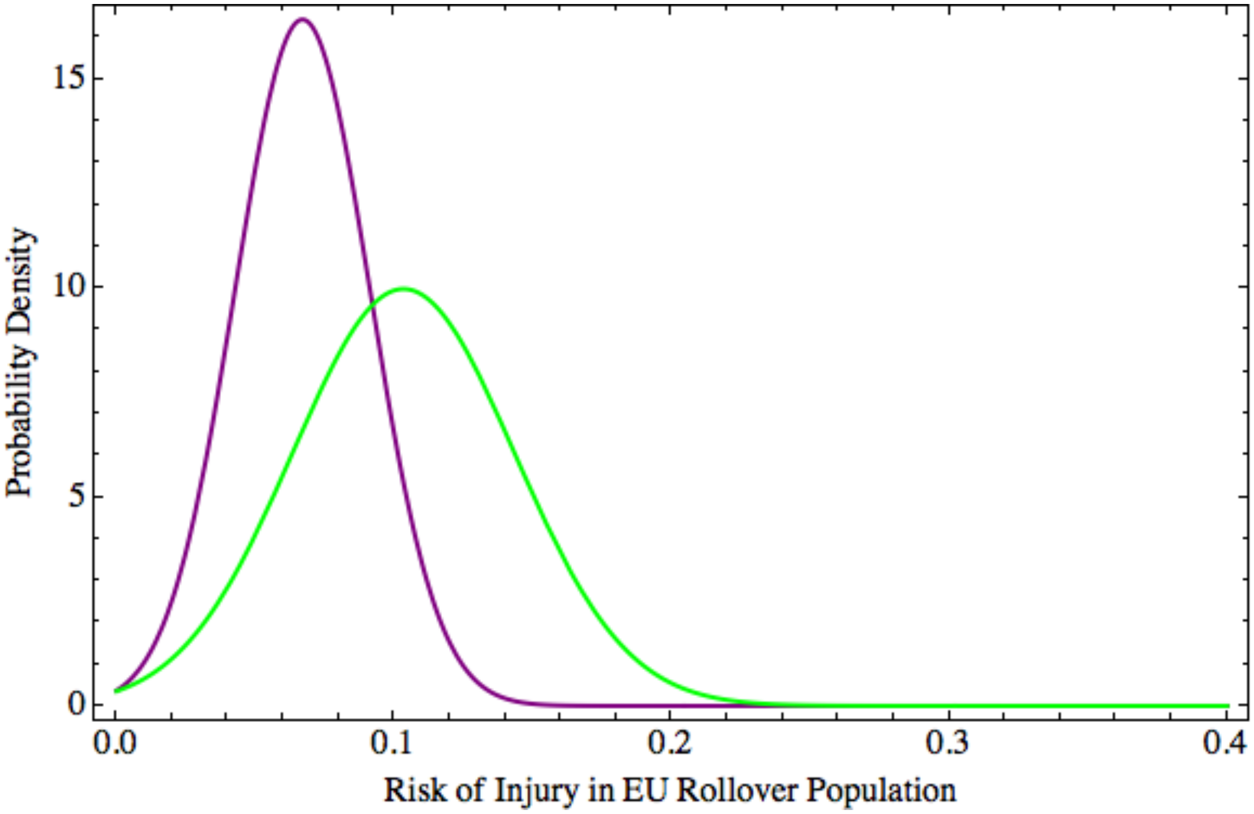

Figure 13. EU (green) and US (purple) rollover models applied to the EU rollover population.

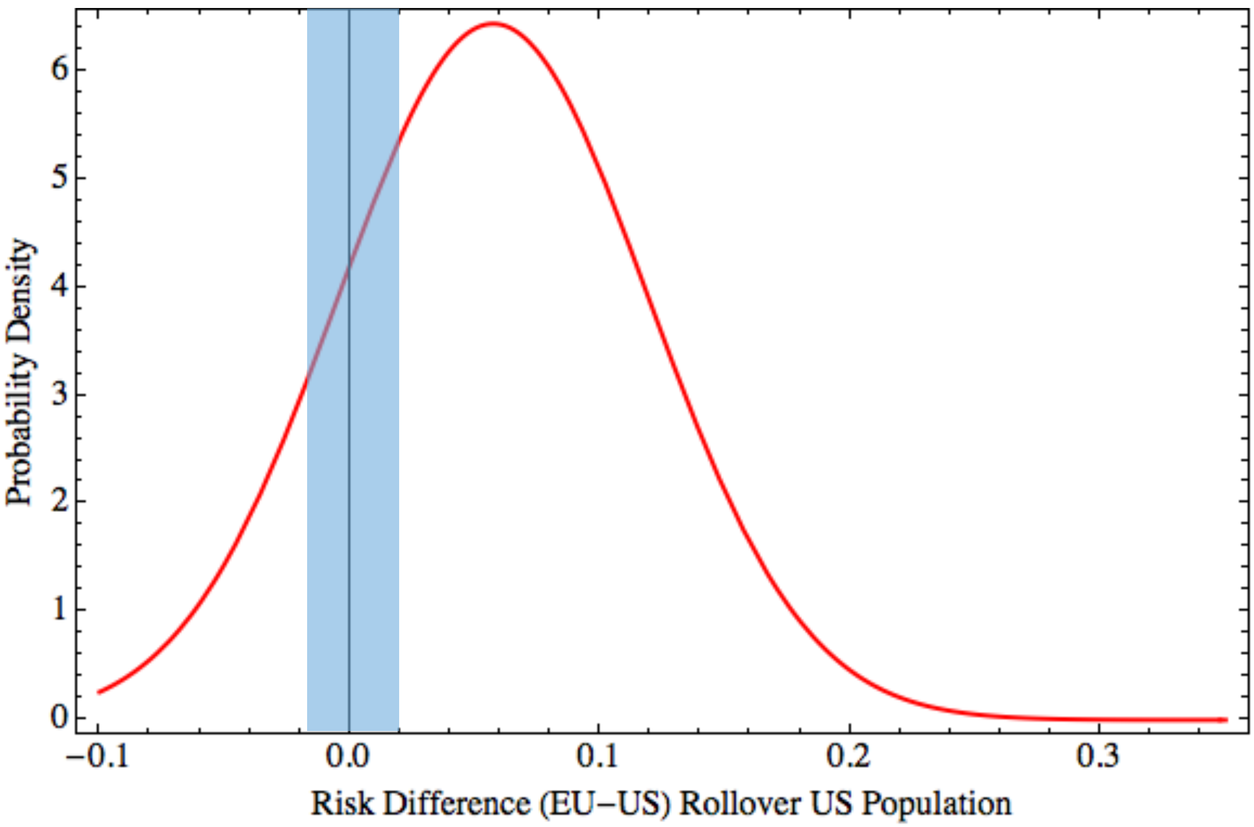

Figure 14. Difference in risk between EU and US models applied to the US rollover population.

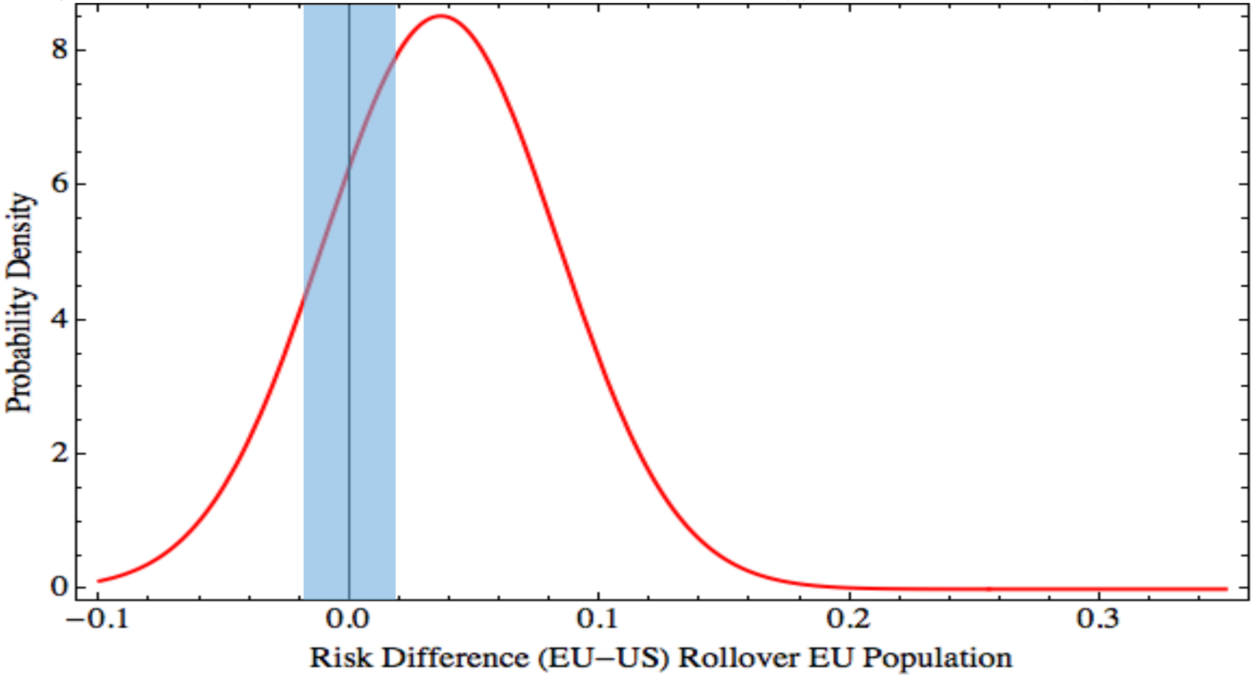

Figure 15. Difference in risk between EU and US models applied to the EU rollover population.

Method 3: Bayes Factors

Figure 16 shows a series of log Bayes Factors assessed for the US front-side population. Each point represents the log Bayes Factor comparing the hypothesis of a given risk difference (x-axis) to the hypothesis of no risk difference. A series of ranges are marked on the plot based on guidelines in Kass & Raftery (1995). These are described in Table 10.

Table 10. Interpretation guide for Log Bayes Factors

Log Bayes Factor Range	Interpretation (Kass & Raftery, 1995)
>5	Very strong evidence favoring risk-difference hypothesis
3 to 5	Strong evidence favoring risk-difference hypothesis
1 to 3	Positive evidence favoring risk-difference hypothesis
-1 to +1	No evidence favoring either hypothesis
-3 to -1	Positive evidence favoring no-risk-difference hypothesis
-5 to -3	Strong evidence favoring no-risk-difference hypothesis
<-5	Very strong evidence favoring no-risk-difference hypothesis

For the US standard population, hypotheses that are more likely than the zero difference model range from -0.018 to -0.004. When evaluated for the EU front-side population in Figure 17 hypotheses more likely than the zero difference model range from -0.018 to -0.009; those not distinguishable from the zero difference model range from -0.024 to 0.003. The evidence supports the hypothesis that EU risk models produce lower risk in both the US and EU front-side populations.

Similar results for the rollover populations are shown in Figure 18 for the US and Figure 19 for the EU. For the US population, hypotheses more likely than the zero difference models indicate the EU model would produce higher risk, with a difference ranging from 0.015 to 0.093. For the EU population, evidence also supports hypotheses for risk differences ranging from 0.018 to 0.062 compared to the zero-difference model.

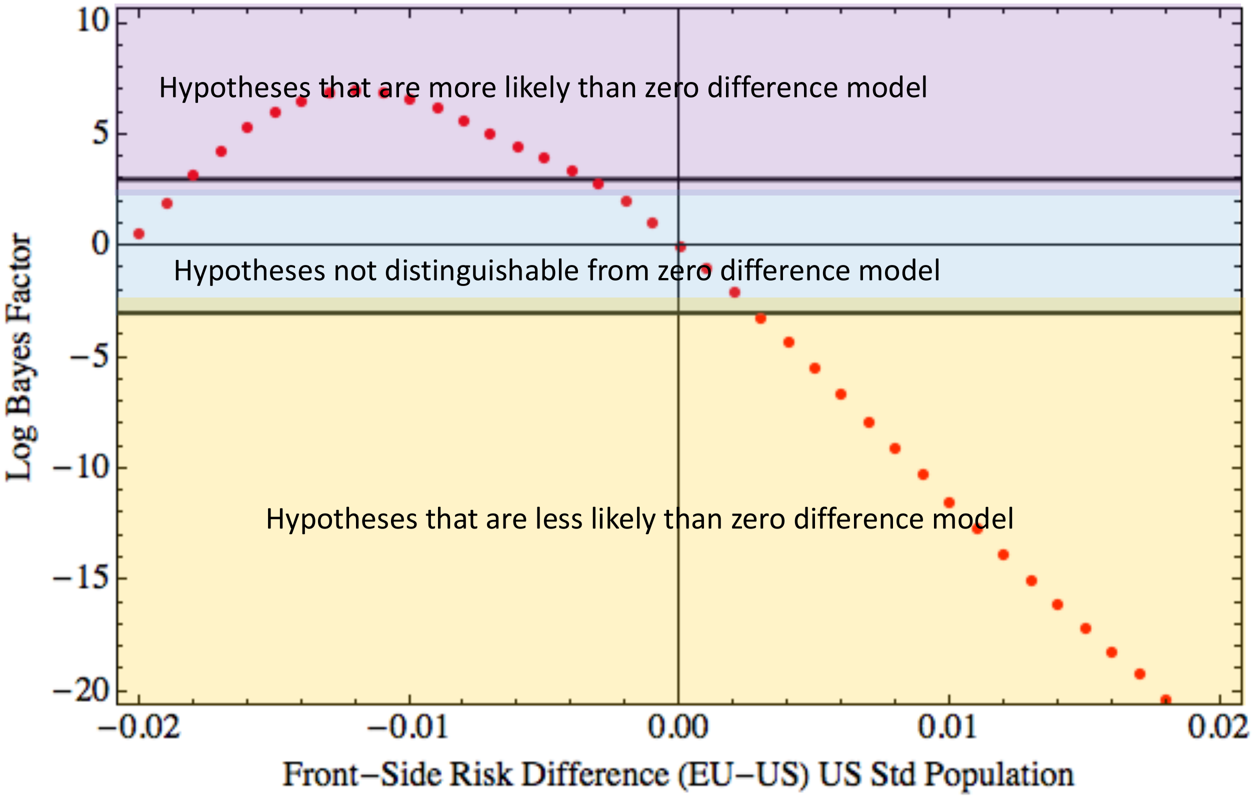

Figure 16. Distribution of Bayes Factors vs. the EU-US risk difference applied to the US front-side population.

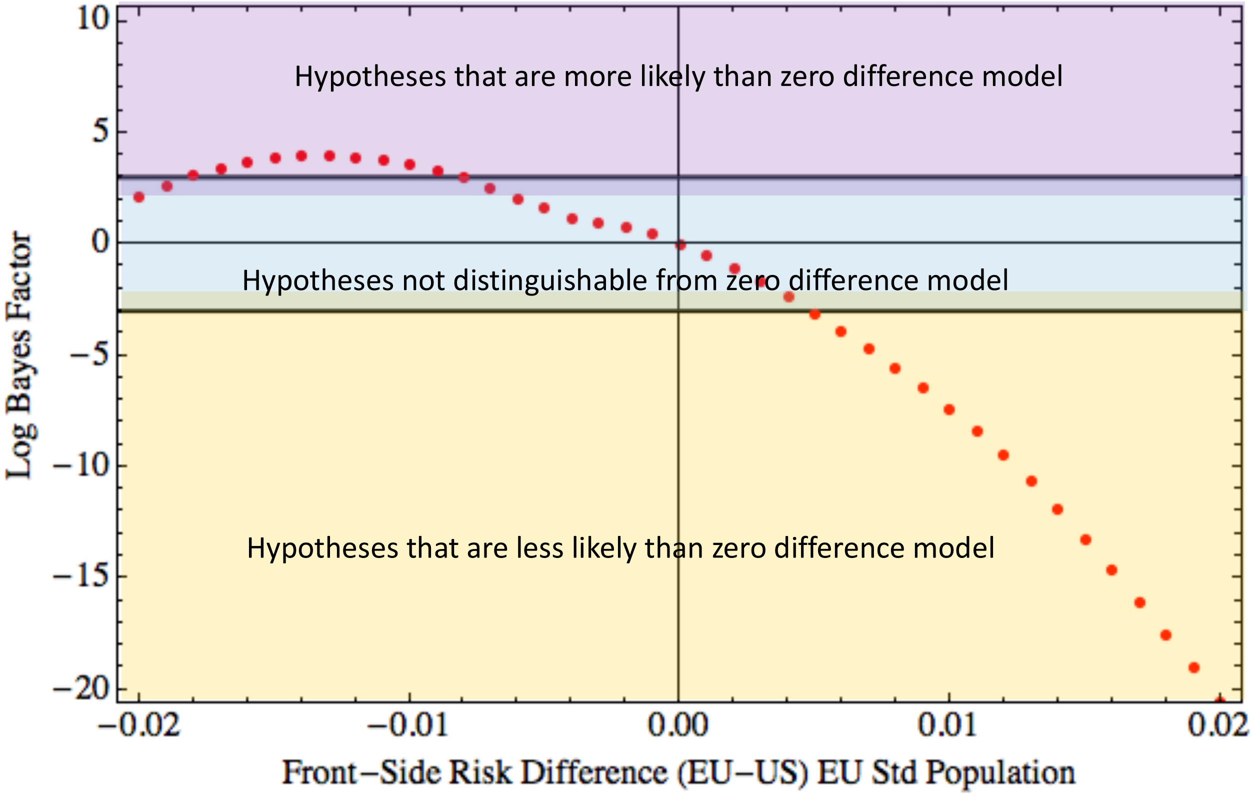

Figure 17. Distribution of Bayes Factors vs. the EU-US risk difference applied to the EU front-side population.

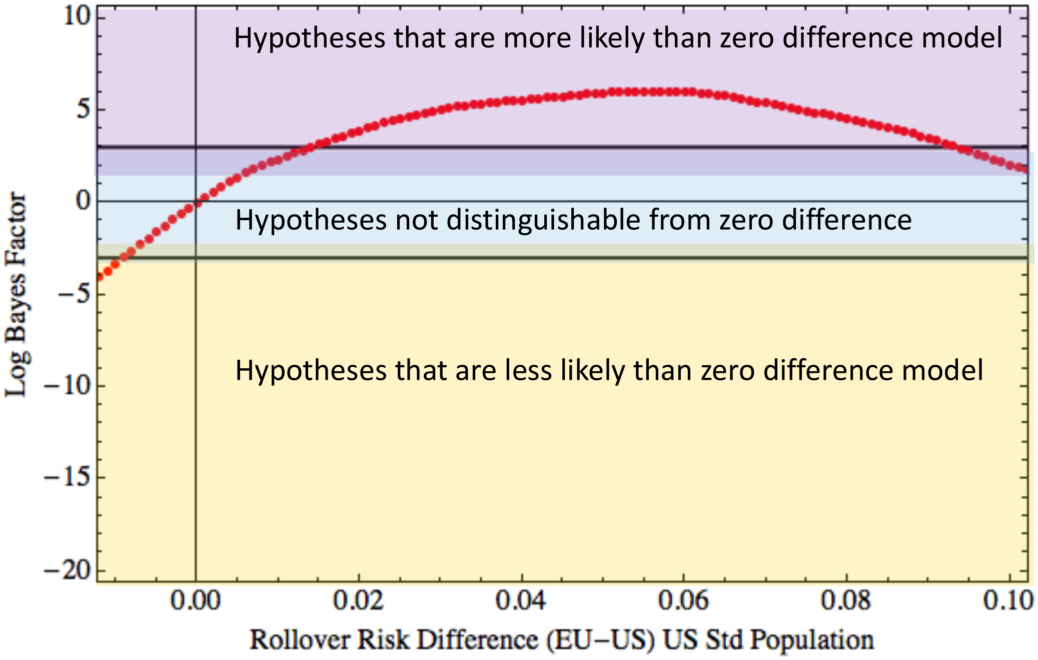

Figure 18. Distribution of Bayes Factors vs. the EU-US risk difference applied to the US rollover population.

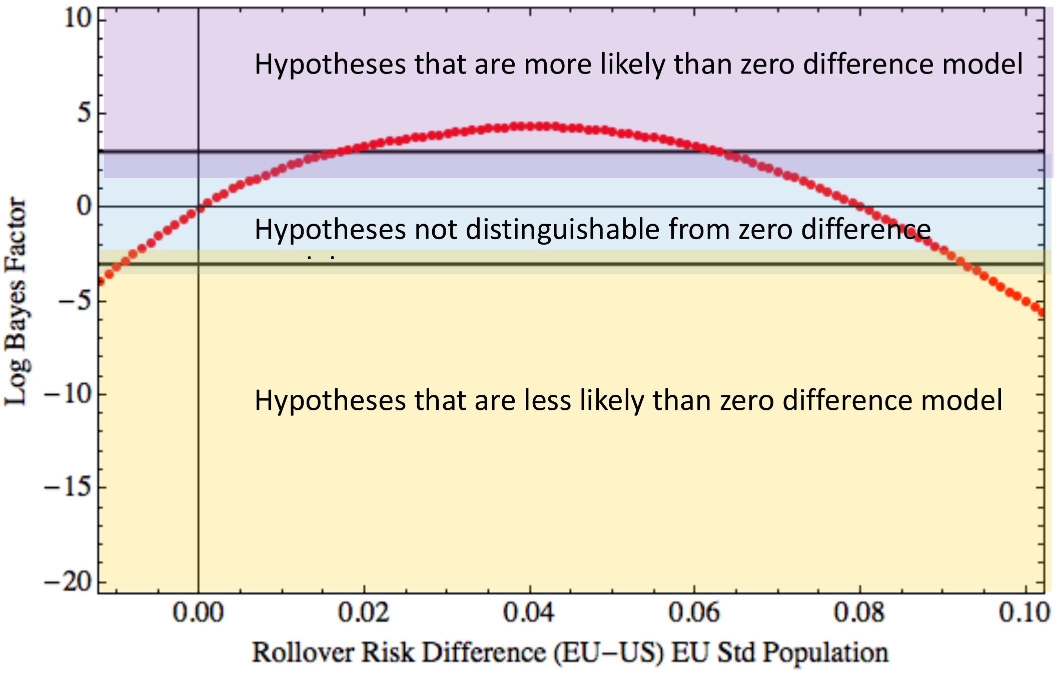

Figure 19. Distribution of Bayes Factors vs. the EU-US risk difference applied to the EU rollover population.

Crash Subgroups

The three analytical methods show that the EU and US models are different in certain ways, resulting in evidence supporting differences in overall risk within front-side impacts and within rollovers. This begs the question of which crashes and conditions are driving the overall differences. One of the advantages of Method 2 is that model predictions for crash subgroups can be compared in the same way predicted risk for the whole population can be compared. Although a comprehensive analysis of this is not feasible, we present a set of breakdowns that aid in understanding how the models differ most. Only estimated mean risk for each subgroup is presented here, though variances can be computed in principle. Note that because mean risk is compiled across all events in each subgroup, the overall estimated risk shown reflects both the effect of the model parameter(s) pertaining to the subgroup *and* the exposure of that subgroup to different severities of crashes. For example, suppose that young drivers tend to experience more severe crashes but are at lower risk of injury *given* a particular crash severity. The mean risk shown for that subgroup will reflect the combination of the greater crash severity experienced by young drivers and their reduced risk of injury relative to older drivers.

Comparisons of Subgroups within Front/Side Populations

Figure 20 through Figure 29 show the mean predicted risks of the EU and US models in each population for front and side crashes broken down by category. In each of these plots, the contribution of each subgroup to the crash population is shown on the horizontal axis, and the overall mean predicted risks for the whole population are shown for comparison. For example, Figure 20 shows the mean predicted risks for the EU and US models for front, near-side, and far-side crashes in the US population. Frontals make up 69% of the US crash population, while near-sides and far-sides make up 17% and 14% of the population respectively. In general, frontal crashes mirror the overall results, but the largest risk differences are seen in near-side impacts. Far-side impacts show little or no risk difference when averaged across all far side cases in this population. The results in Figure 21 for the EU standard population show a similar pattern in that the greatest difference is seen in near-side crashes. However, overall risk for frontals is much more similar to that of near-sides for this population than for the US population. This reflects differences in the underlying populations of frontal crashes (for example) that are seen in the EU vs. the US.

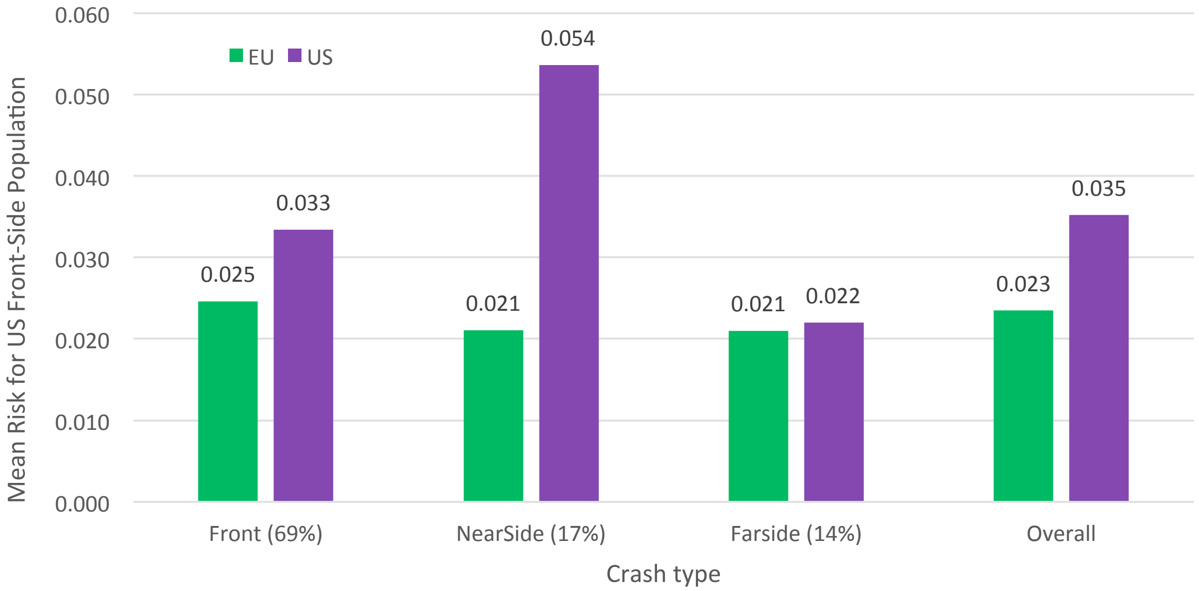

Figure 20. Mean predicted risk for cases in US front-side standard population broken down by crash type.

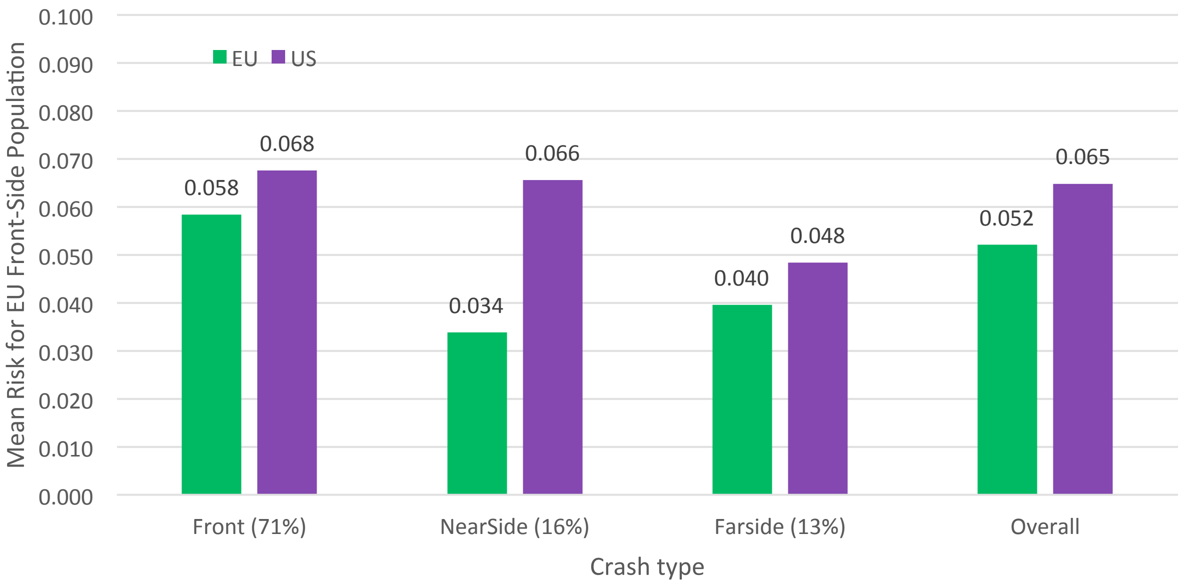

Figure 21. Mean predicted risk for cases in EU front-side standard population broken down by crash type.

Figure 22 and Figure 23 shows the mean predicted risks by age group for the US and EU populations, respectively. In both populations, the youngest and oldest occupants have similar mean risk. However, in the middle age range (31-70), predicted risk for the US model is much higher, and the difference is most pronounced for the 51-70 age group in the EU population. In both models, age was a quadratic function, but the form of the quadratic was different for each. In the EU model, the age effect starts slowly and then accelerates; in the US model, the age effect accelerates early and then slows. It is interesting to note that the crash population in this analysis is generally younger in the US than in the EU.

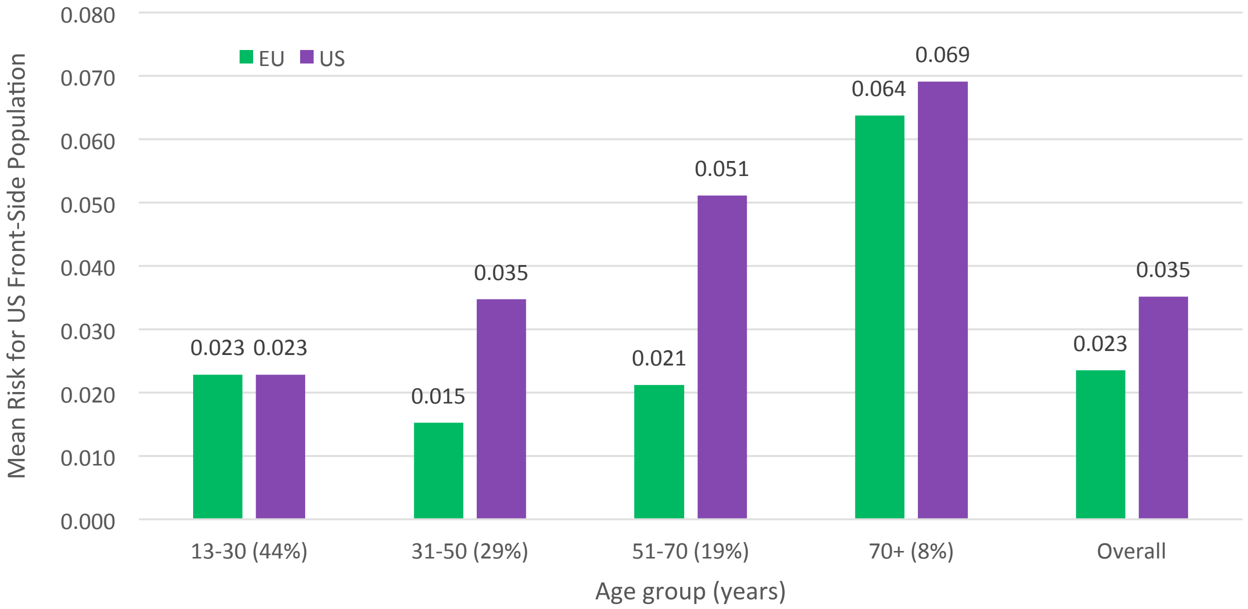

Figure 22. Mean predicted risk for cases in US front-side standard population broken down by age group.

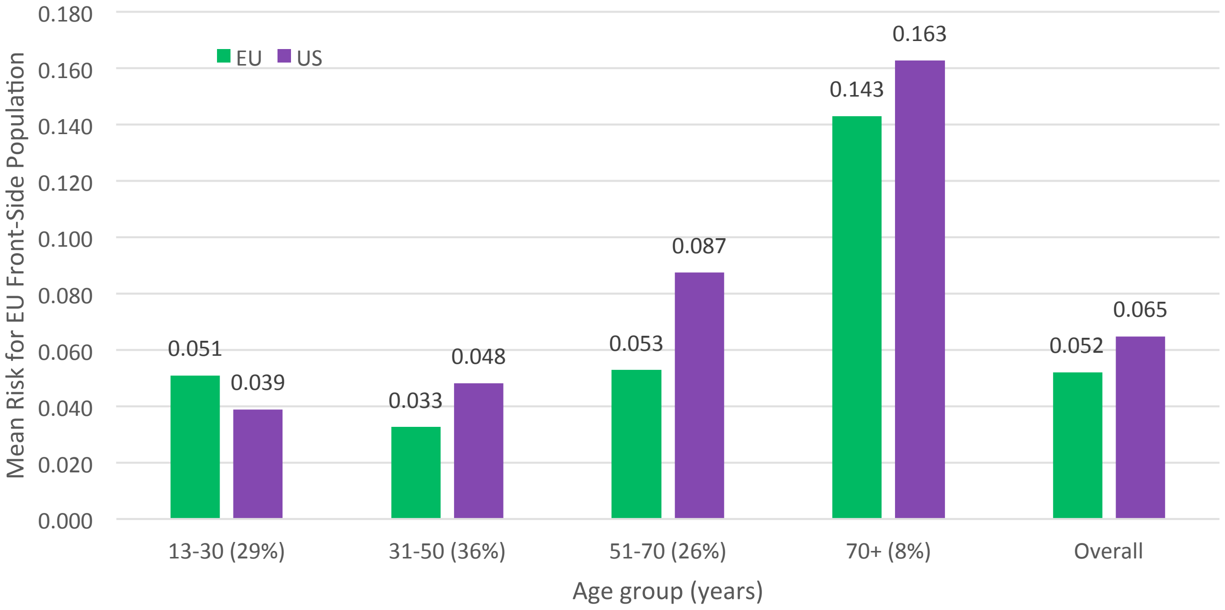

Figure 23. Mean predicted risk for cases in EU front-side standard population broken down by age group.

The breakdowns by belt status for the US and EU populations are shown in Figure 24 and Figure 25. In both populations, the mean risk for belted occupants is somewhat higher for the US model. However, the largest differences are among unbelted occupants. Even though unbelted occupants make up a very small proportion of the EU population (4%), their much higher risk still influences the overall risk difference to a non-trivial extent. The risk difference for belted occupants in the EU is -0.08 while the overall risk difference is -0.013. In the US population, the influence of the unbelted on the overall risk difference is greater because they make up a greater percentage of the population.

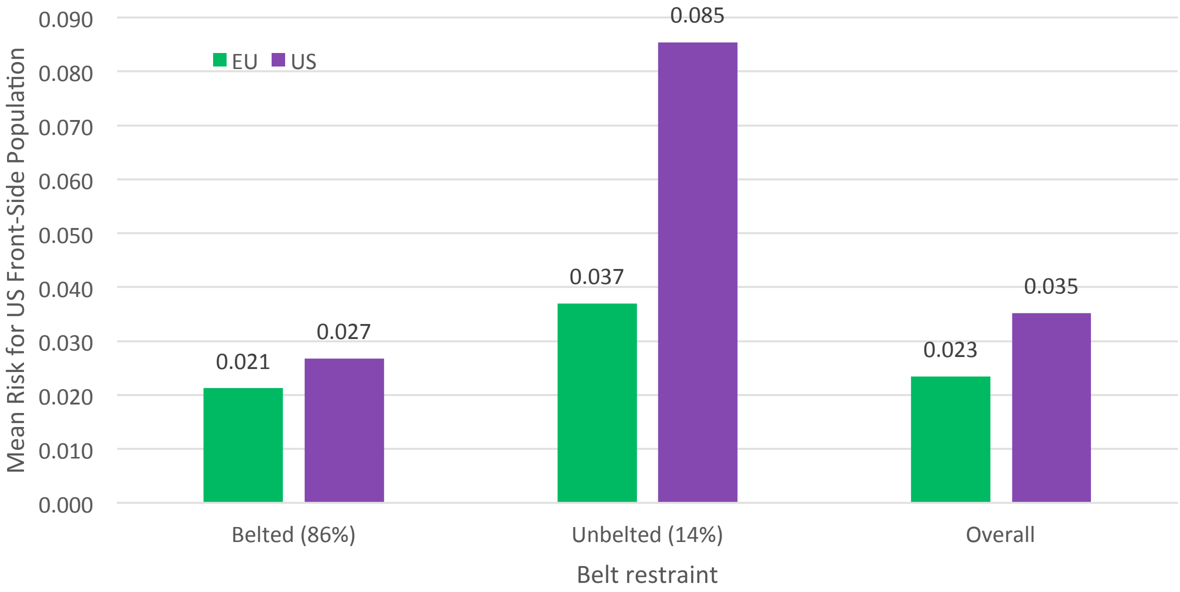

Figure 24. Mean predicted risk for cases in US front-side standard population broken down by belt use.

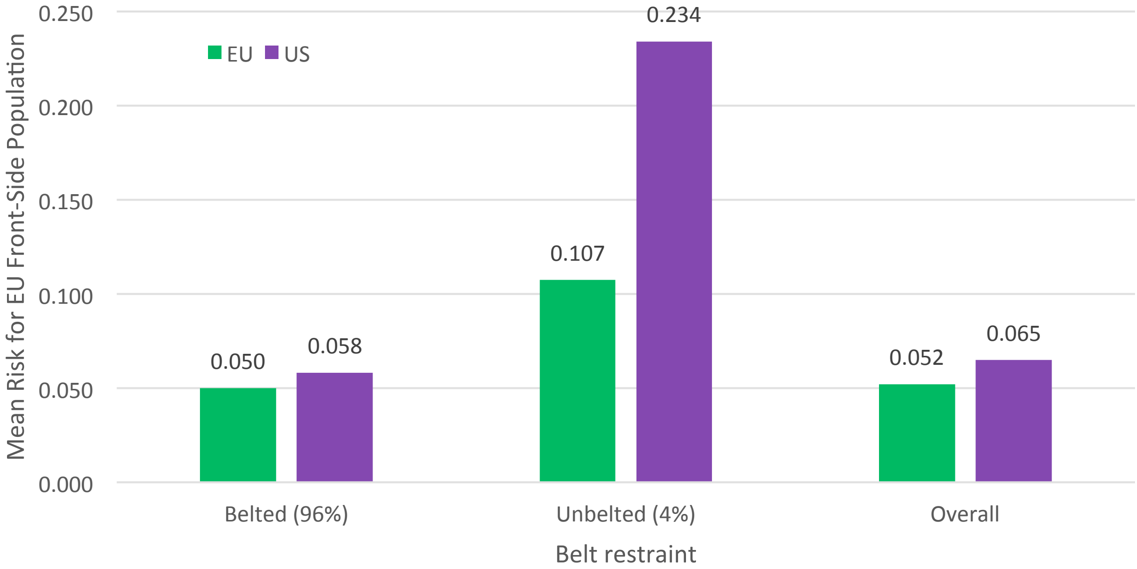

Figure 25. Mean predicted risk for cases in EU front-side standard population broken down by belt restraint.

Figure 26 and Figure 27 show predicted risks by roadway location/type for the US and EU populations. In the US population, the risk differences for rural and urban roads are not distinguishable and both are similar to the overall risk difference. In the EU population, the risk difference is somewhat larger for rural roads, but not dramatically so. Interestingly, crashes on rural roads are more prevalent in the EU crash population and their risk is very high. The mean risk for urban locations/roads is very similar for the two populations, whereas the rural risks and the overall risks for the EU population are much higher than for the US population.

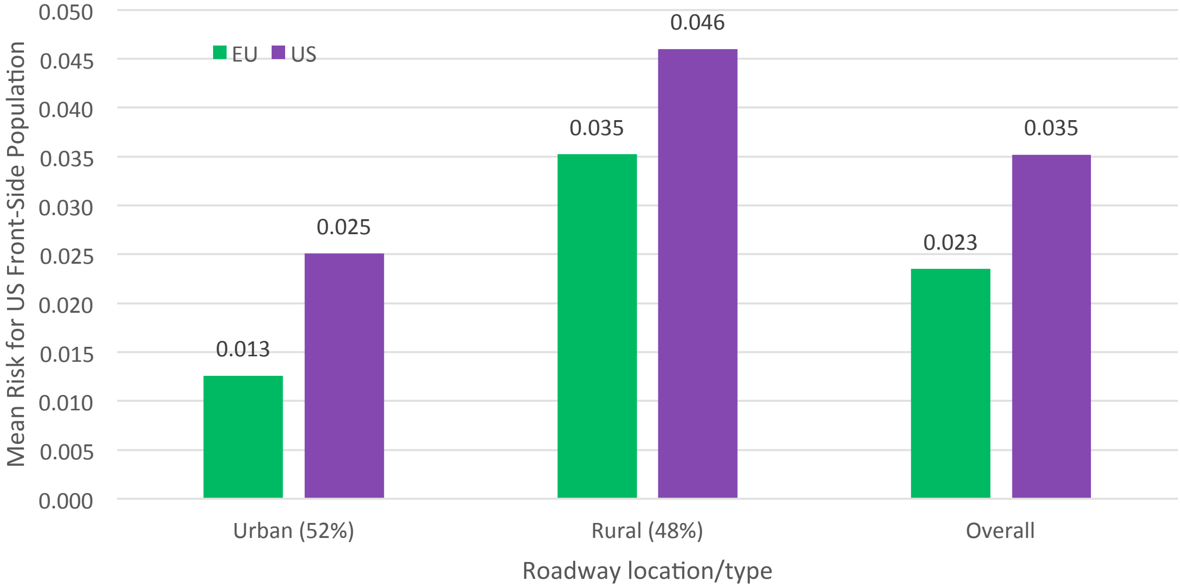

Figure 26. Mean predicted risk for cases in US front-side standard population broken down by rural-urban road type.

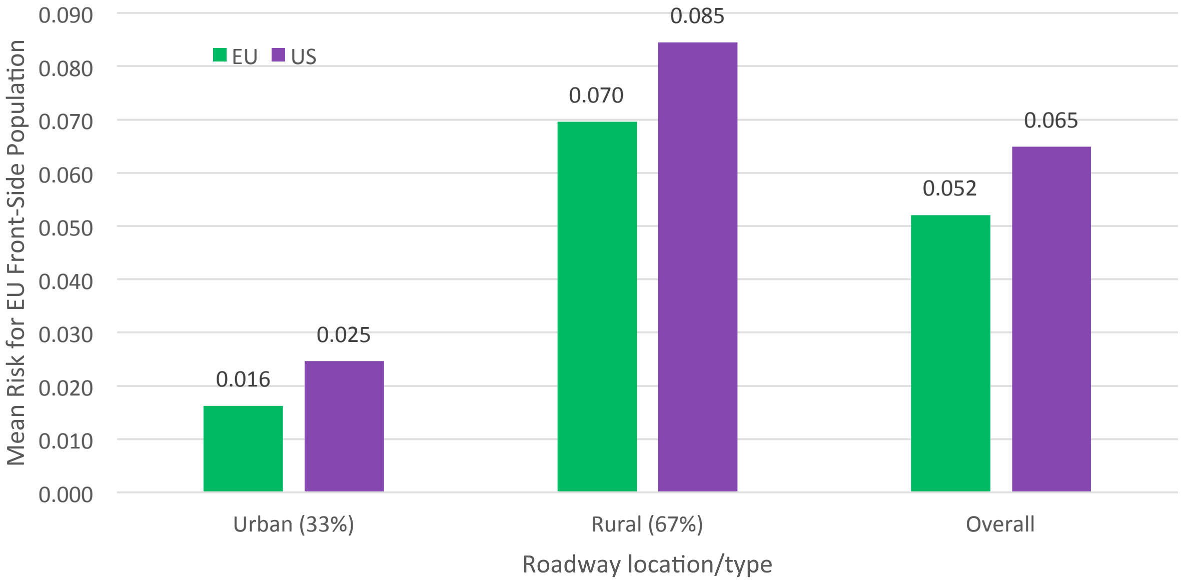

Figure 27. Mean predicted risk for cases in EU front-side standard population broken down by roadway location/type.

Predicted risks broken down by Delta-V group for the US and EU populations are shown in Figure 28 and Figure 29. Note that these groups are not separated by crash type. The risk differences increase as Delta-V increases for both populations.

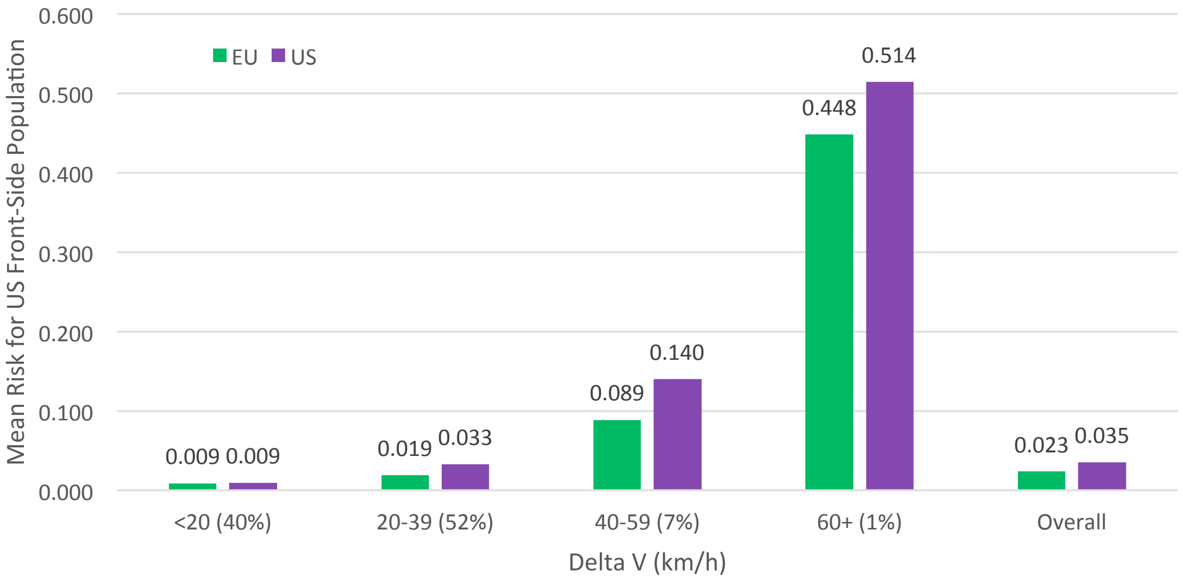

Figure 28. Mean predicted risk for cases in US front-side standard population broken down by Delta-V category.

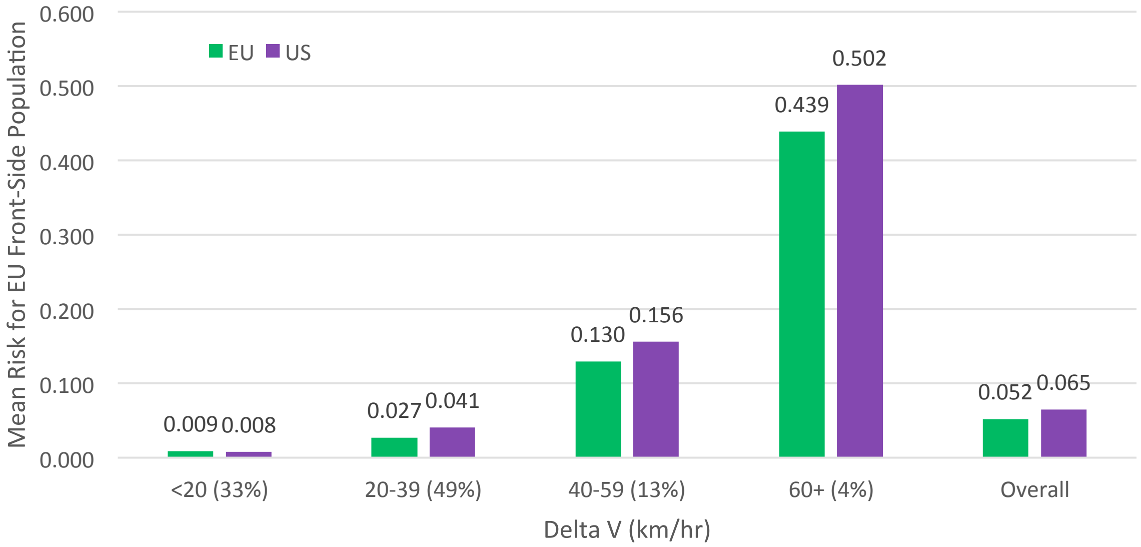

Figure 29. Mean predicted risk for cases in EU front-side standard population broken down by Delta-V.

Comparisons of Subgroups within Rollover Population

For the US rollover population, Figure 30 and Figure 31 show very similar results for the US and EU rollover populations broken down by belt restraint use. Although both belted and unbelted occupants are at lower risk in US vehicles in rollovers, the difference is much larger for unbelted occupants. Since unbelted occupants make up a larger proportion of the US population than the EU population, the effect of the unbelted risk difference on the overall risk difference is greater for the US population.

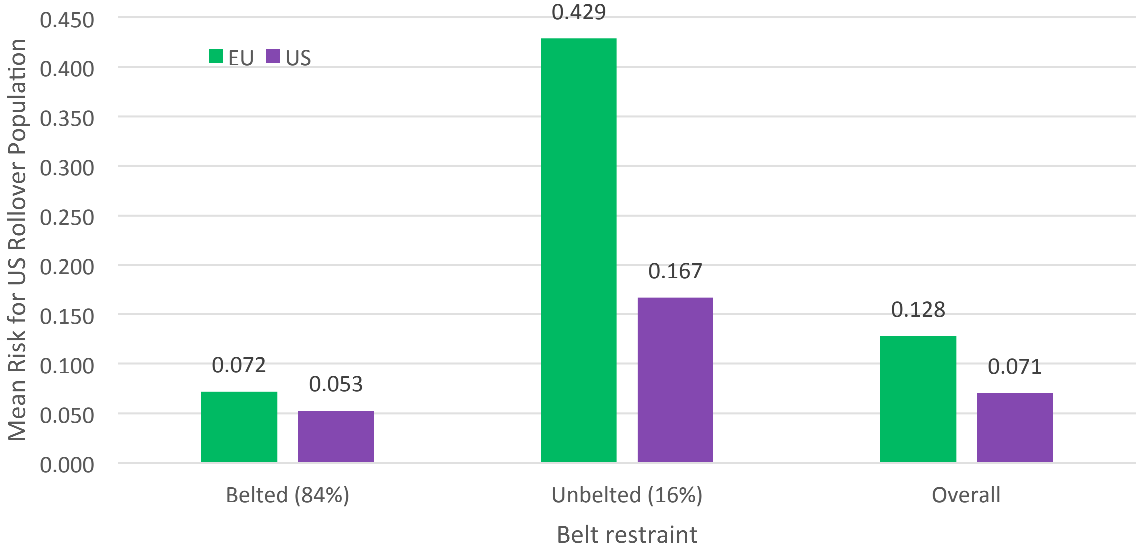

Figure 30. Mean predicted risk for cases in US rollover standard population broken down by belt use category.

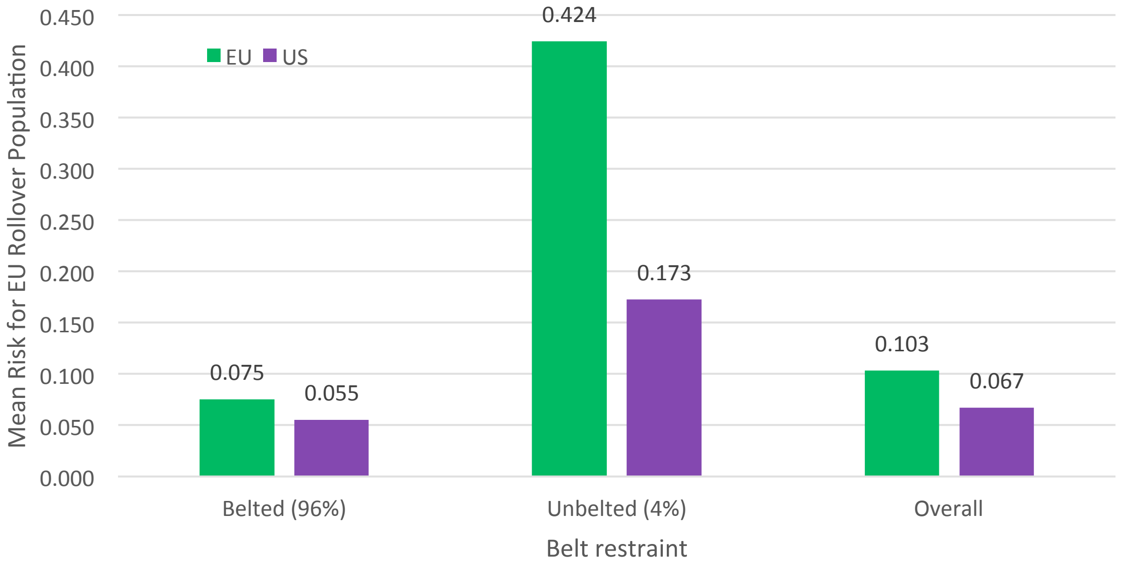

Figure 31. Mean predicted risk for cases in EU rollover standard population broken down by belt restraint.

Finally, Figure 32 and Figure 33 show the mean risks for the US and EU rollover populations broken down by ejection status. Ejected occupants show a greater risk difference in both populations, but they make up a small percentage of each and have relatively little influence on the overall risk difference.

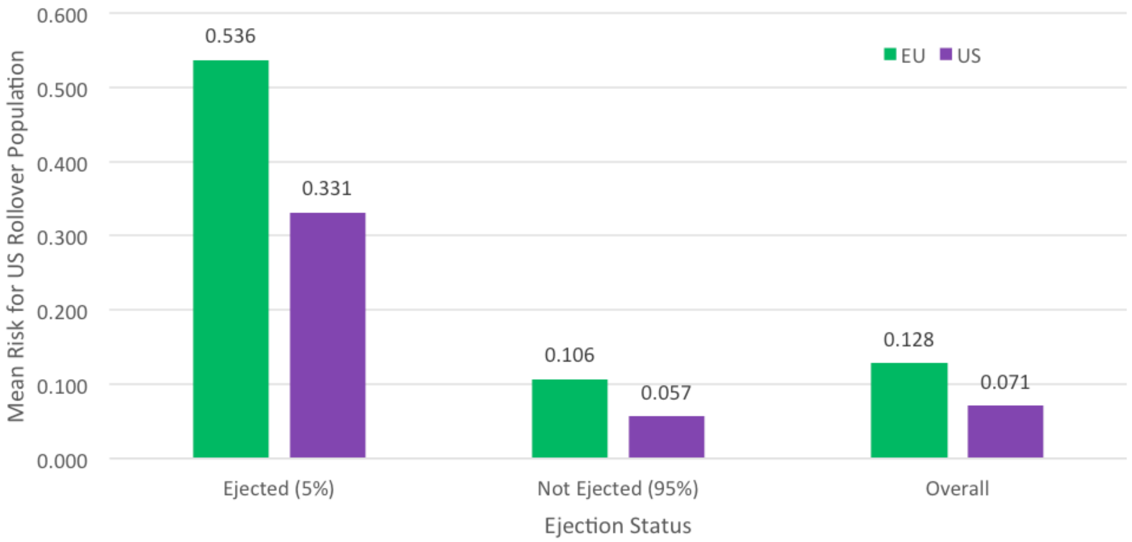

Figure 32. Mean predicted risk for cases in US rollover standard population broken down by ejection category.

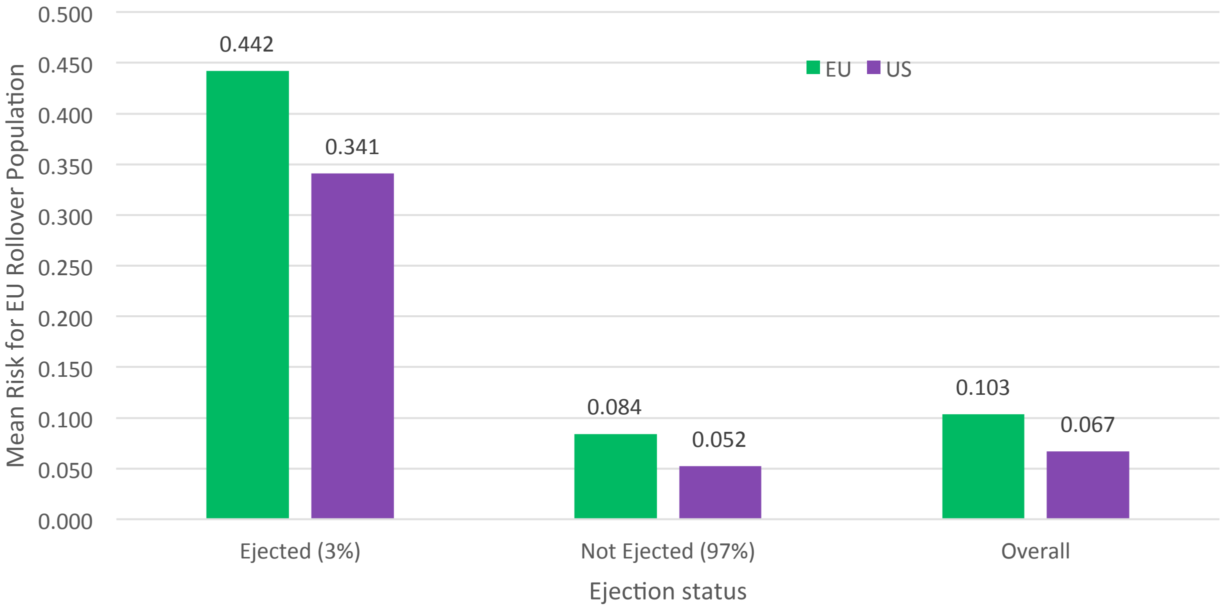

Figure 33. Mean predicted risk for cases in EU rollover standard population broken down by ejection category.

Crash Avoidance: Lighting

Table 11. shows the number of pedestrian fatalities from 6:00 to 6:59 pm that occurred in November in the dark and October in the light for the US and EU. The ratio of dark/light ratios is 0.67 (0.41, 1.11). This indicates that darkness has a smaller effect on pedestrian fatalities in the US than in the EU and suggests that US headlamps illuminate pedestrians better. However, since the confidence interval includes 1, the result is not significant (defined as $p \leq 0.05$). As discussed earlier, failure to reach significance is not evidence *for* the null hypothesis.

Table 11. Number of pedestrian fatalities between 6:00 and 6:59 for the US and EU

Pedestrian fatalities 6:00pm-6:59pm	November, Dark	October, Light
US	292	46
EU (based on included countries)	266	28

Crash Avoidance: Mirrors

The crash counts are summarized in Table 12.

Table 12. Lane changes to the driver's and the passenger's side (without any restriction of the injury level). The EU countries included in the analysis are Portugal and the UK.

Lane change crashes	To the driver's side	To the passenger's side
US	9033	6426
”EU” (PT & UK)	6040	5311

The same formula as used for the DST analysis gives the odds ratio and confidence intervals. The resulting point estimate for the US/EU ratio is 1.24, and the confidence interval of (1.18, 1.30) does not contain 1, which means that the result is significant at the 0.05 level. These results indicate a significantly higher proportion of crashes to the driver's side (compared to the passenger's side) in the US than in the combined data for these two EU countries (the only ones that distinguish between right- and left-lane change crashes). This suggests that the mirrors in the EU vehicles on the driver's side help prevent lane-change crashes better than those in US vehicles.

Fleet Penetration

The purpose of this section is to illustrate the way in which fleet penetration might affect interpretation of the results described above. In particular, if evidence for equivalence of field performance is to be evaluated, it will be necessary to define "equivalence." As in bioequivalence testing, a range of acceptable differences must be defined, along with a level of certainty about the estimated difference level. Fleet penetration will affect the extent to which any estimated risk difference affects the overall number of seriously injured occupants seen in the population over time. Thus it should be considered when identifying a range of acceptable risk differences that define "equivalence."

To understand the potential impact of mutual recognition if there are differences in risk, we looked at the population of towaway-crash-involved occupants in light vehicles. A simple simulation is provided using the US as an example. We estimated the base risk for occupants involved in towaway crashes in light vehicles per year and then evaluated the effect of various risk differences between EU and US vehicles with respect to crashworthiness and various levels of penetration in the new-vehicle fleet. We assume that new vehicles make up 5% of the fleet each year and that US-regulated vehicles are replaced at random by EU-regulated vehicles.

The results of this simulation are shown in Figure 34. The graph shows the annual change in MASI3+F injured occupants relative to current for a particular combination of risk difference across all crash types and fleet penetration among new vehicles. For illustration, the black lines show that if an estimated ±0.7% annual change in injuries were considered acceptable and fleet penetration was expected to be 25% of all new vehicles, the corresponding risk difference range would be ±0.2. This range was used for illustration in describing the results of Method 2, but choosing appropriate values is not in the purview of the research team.

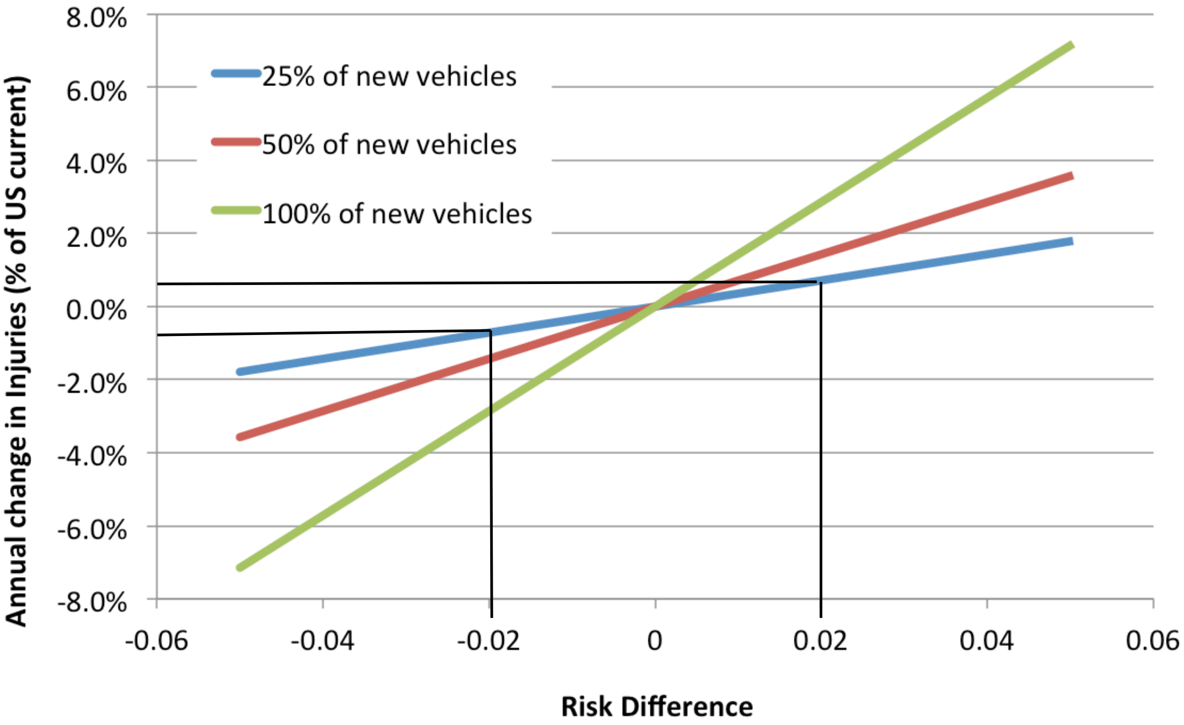

Figure 34. Simulation of overall US crash risk variation illustrating the effect of different penetration levels of EU vehicles and various levels of risk differences with respect to crashworthiness.

Finally, we caution that the assumptions made here are simplistic and unlikely to hold as we describe. In particular, vehicles are unlikely to be replaced at random, but instead, smaller EU-regulated vehicles are more likely to replace smaller US-regulated vehicles. Since a more complex simulation is speculative and beyond the scope of this project, we present only the simple model.

Consumer Ratings

The purpose of this section is to review the distribution of star ratings in the EU and US to identify how different they might be. If one population has a general tendency to purchase safer vehicles than the other, we could see risk differences that are not actually the result of regulatory differences but of purchase-habit differences. Early in the research process, we considered including star rating as a predictor in the models. However, data that would cover enough of the vehicle sample were not available. Instead, we present a brief discussion of available information on sales in the EU and US with respect to star ratings. Because the rating systems in EU and US are different, the comparisons below cannot definitively identify purchase-habit differences. Nonetheless, we include the available information to help the reader consider the extent to which consumer ratings may influence the overall field safety of a population of vehicles (outside of regulatory differences).

The European New Car Assessment Programme (Euro NCAP) was established in 1997 and has tested numerous car models since then. The aim of the testing program is to organize crash-tests and provide consumers with a realistic and independent assessment of the safety performance of some of the most popular cars sold in Europe[8]. In the beginning, Euro NCAP rated occupant protection and pedestrian protection. Later, in 2003, a child protection rating was also introduced. At this time, separate ratings were presented for each of the three areas. From 2009, the separate ratings are combined into one overall star rating for each tested vehicle. Meanwhile, the rating criteria continue to develop and as today, tests considering for example protection of whiplash injury, autonomous emergency braking (AEB), electronic stability control (ESC) and speed assistant systems are included. Kullgren et al. (2010) compared Euro NCAP test results with real-world crash data and found good correlation. Moreover, the largest difference in injury risk between 2–star and 5–star rated cars was found for risk of fatality, 68 ± 32 percent lower risk for 5–star cars.

In 2009, the European Transport Safety Council (ETSC 2009) published a comparison between countries with respect to star ratings of new cars sold in the first nine month of 2008. The results concerning occupant protection are shown in Figure 35 for France, Germany, United Kingdom and EU. In EU, 53% of the new cars sold were 5-star cars and 31% were 4-star cars with respect to occupant safety. The proportion of 5-star cars is higher in all of the three selected countries than the EU average. However, the difference is rather small; the proportion ranges from 55% in UK to 59% in France. The proportion of 4-star cars is almost the same (30–32%) as the average in EU.

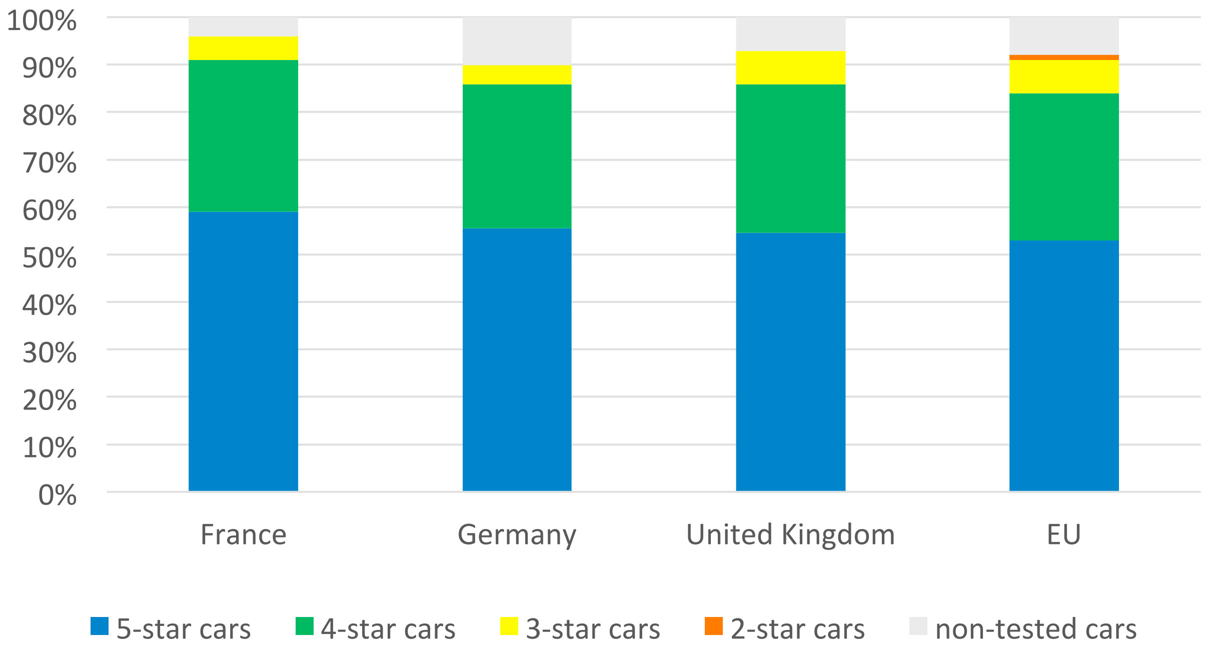

Figure 35. Distribution of 2008 vehicle sales for each country and the EU by EU star rating.

In the US, the National Highway Traffic Safety Administration established the New Car Assessment Program (NCAP) to provide consumers with additional information regarding vehicle safety. The original NCAP testing provided frontal impact results from a barrier test performed at a Delta-V of 56 km/h, higher than the 48 km/h required by regulatory testing. The latest incarnation of NCAP adopted in 2011

[8] www.euroncap.com

includes results of frontal impact testing, side impact testing, and rollover testing. The change was somewhat motivated by the high frequency of vehicles achieving 4 or 5 stars on the rating.

NHTSA provides NCAP results for individual vehicles but not the fleet. However, an independent website (informedforlife.org) has compiled results for all vehicles tested since 2011. To estimate the proportion of vehicles achieving each NCAP rating of 1 to 5 stars, sales data from the top 100 selling vehicles of 2010 and 2013 were retrieved (Automotive News 2014, 2011), which represent approximately 85% of vehicle sales in the US. NCAP scores from 2011 testing were applied to the 2010 sales figures, while scores from 2013 or 2014 were applied to the 2013 sales figures. Resulting distributions are shown in Figure 36. Because of the change in NCAP protocols starting in 2011, many of the top-selling vehicles were not tested, and there were fewer 5-star and more 3-star vehicles than in previous years. For the 2013 sales figures, the majority of vehicles had 4- or 5-star ratings.

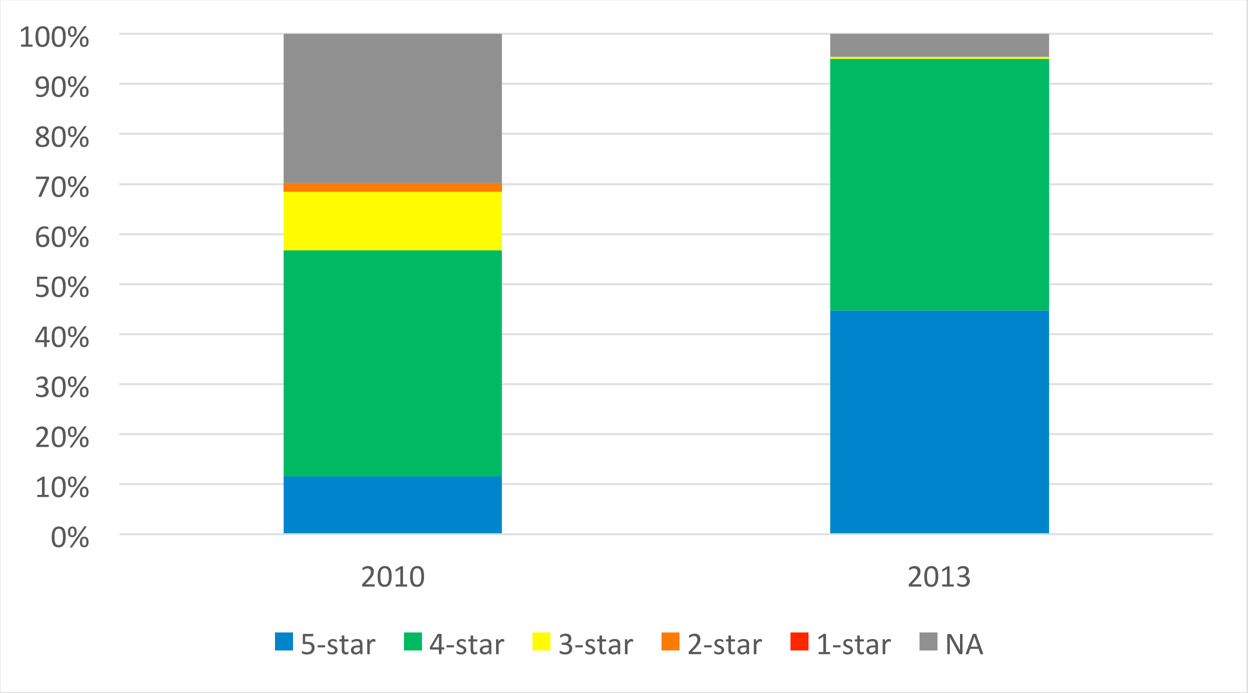

Figure 36. Distribution of top 100 vehicle sales for the US in 2010 and 2013 by NCAP star rating.

Discussion

Methods

The analysis described in this report investigated the question of whether vehicles meeting EU safety standards would perform equivalently to US-regulated vehicles in the US driving environment, and that vehicles meeting US safety standards would perform equivalently to EU-regulated vehicles in the EU driving environment. Analyses related to crashworthiness and crash avoidance standards were done separately using different datasets and methods.

The approach we chose to analyze crashworthiness was to develop statistical injury risk models for EU-regulated vehicles and US-regulated vehicles and then compare the predictions of these models on the EU crash population and the US crash population. This allows us to separate risk (which is influenced by crashworthiness-related regulations) from exposure (the collection of crashes experienced by occupants in each region). It is not useful or appropriate to compare risk of injury of US vehicles within the US population to the risk of injury in EU vehicles within the EU population, because the total injury risk in each region is a combination of the risk and exposure.

Because neither the US nor EU crash datasets allow a direct comparison of risk in the two vehicle groups (US-regulated vs. EU-regulated), we used separate datasets collected under different protocols. Moreover, crash data in the EU are collected within several countries, also under different protocols. To build risk models that could be compared on a common population, we had to ensure that the populations sampled were comparable and that variable definitions were harmonized.

The success of variable and selection-criteria harmonization is critical to the success of the approach. The use of the 1998 version of the Abbreviated Injury Scale (AIS) for injury coding and Crash Damage Classification (CDC) for damage coding in all datasets ensured that the outcome variable and critical crash descriptors were defined in the same way. Harmonization of other predictors was done on a case-by-case basis. In general, where compromises had to be made (e.g., towaway vs. damage extent criteria, registration year vs. model year), at least one dataset had enough information to confirm the correspondence between the two definitions. The one variable that is most unique to the local environment is classification of crash location/road type as rural or urban. Because this predictor was significant in analysis of most datasets it was kept as a predictor, but roads vary from country to country and their characteristics as they relate to injury risk may not translate as precisely as other variables.

A significant harmonization effort was put towards ensuring that Delta-V was comparable when reconstructed using two different methods: trajectory-based and crush-based. The presence of cases that were reconstructed using both methods allowed us to directly compare the results and develop a transformation to apply to crush-based reconstructions. The fact that the coefficient of Delta-V was so similar for EU and US risk models (see Table 8) gave some indication of the success of the harmonization.

Although logistic regression produces unbiased coefficients even when samples are biased, the intercept is not unbiased (Prentice & Pyke, 1979). Since our risk estimates depend on the intercept as well as other coefficients, it is critical to ensure that the populations being sampled in both regions are comparable. If, for example, one sample is biased towards more injured occupants for a given set of crash characteristics, then the intercept and risk estimates will be biased upwards. It should be noted, however, that while sampling from a higher-risk population of crashes (e.g., towaway crashes) results in a higher overall injury rate for the sample compared to a sample of lower-risk crashes, that is not the same as biasing the sample towards injured occupants within the crashes defined.

The US dataset is a national probability sample with weights constructed to ensure that estimates are nationally representative. Once the selection criteria were applied, the weighted sample should still be unbiased with respect to those criteria. The EU datasets are each sampled according to a different protocol. However, weights and weighting methods have been developed for each to adjust the samples to national statistics. In addition, the EU weighting based on the CARE data considered injury outcome in CARE to further adjust to the appropriate injury rate within each class of crashes.

The overall injury rate in the combined EU dataset is higher than that of the US dataset. However, when models were compared side-by-side, the risk differences for both front-side and rollover were in the same direction. This pattern suggests that the population of crashes in the EU, at least within the population studied, may be more dangerous than those in the US. However, the risk model predictions for both regions track this pattern, suggesting that the intercepts are not driving the relative risk predictions. It is not possible to know whether the sampling was perfectly comparable and unbiased, but every measure was taken to ensure comparability and the results do not suggest otherwise.

Results

Accepting that selection and variable definitions were harmonized effectively, we turn to the results. First, Seemingly Unrelated Regression (SUR) fairly definitively indicates that the two models are not the same. In the case of the front-side models, the null hypothesis of same coefficients was rejected for a number of predictors. The patterns of injury risk vary between EU- and US-regulated vehicles in a number of ways, notably in the effect of age (which shows a stronger acceleration with age in the EU than the US), unbelted (larger effect in US), rural roads (larger effect in EU), wide crash partner (larger effect in EU), and near-side and far-side risk as a function of Delta-V (both near- and far-side crashes have lower intercept and steeper slope as a function of Delta-V for US compared to EU).

The comparison of rollover models indicated that only the unbelted coefficient was significantly different between the two models. The effect of being unbelted is greater in the EU in rollovers, and the difference in the coefficients was enough that the multi-degree-of-freedom test for whether the whole model is the same was rejected.

Interpretation of the individual coefficients of these models can be challenging. The models are designed to represent the crashworthiness performance of vehicles that are subject to different regulations, which are intended to influence that performance. It is tempting to view each coefficient in terms of a mechanistic relationship between regulation and injury risk. However, these are statistical models built on observational (as opposed to experimental) data. Thus, each coefficient is influenced by the values of other coefficients and the correlation among predictors in the dataset. Moreover, the relationship between the coefficients in the models and the effect of vehicle design in response to regulation is not necessarily directly interpretable.

The results of Method 2, the side-by-side application of the two maximum likelihood models, were consistent for the two standard populations. For front and side impacts, overall estimated risk for EU vehicles was lower than for US risk, but the variability is relatively large resulting in a distribution of risk differences that extends above and below zero.

Method 3 produces a similar picture using a different approach. For the US standard population, evidence favors risk differences from -0.018 to -0.004 over the zero risk difference model, and for the EU population, the range of better-supported models was -0.018 to -0.009. In general, the likelihood surface is relatively flat indicating a wide range of fairly likely models, but the risk differences with greater evidence than the zero-difference hypothesis are all negative, indicating lower risk for EU-regulated vehicles.

For rollover results of Method 2, US vehicles have lower risk for both populations, and the distribution of risk differences, though crossing zero, strongly indicates that the risk difference is likely larger than zero. Method 3 confirms these results, showing that there is little evidence for the zero-difference hypothesis in comparison to a range of hypotheses showing lower risk in US vehicles. However, it should be noted that ESC penetration is likely higher in the EU datasets (though ESC status is unknown). If the presence of ESC results in remaining rollovers being of greater severity, then the lack of an available severity measure in the rollover data means that EU risk models would tend to predict higher risk on average. This possibility may have affected the magnitude of the difference seen and should be investigated in future work.

The breakdown of the models into subgroups provides some insight on which particular groups of crashes are affecting the overall differences. For the frontal-side population, the largest differences occurred in near-side crashes and to unbelted occupants. With respect to age, the US risk models show a gradual increase in risk for each age group, while the EU models indicate fairly steady risk across occupants less than age 70, followed by a sharp increase. Risk differences increased with increasing Delta-V. Finally, for rollovers, a larger difference was seen for unbelted occupants than belted and for ejected occupants compared to those who were not ejected.

The crash avoidance analysis, though limited to headlamps and side mirrors, replicated previous research on these areas. For pedestrian fatalities, risk in darkness is substantially higher than in light in both regions, but US headlamps reduced risk in the dark relative to the light more than did EU headlamps. Regulation regarding headlamps takes into consideration a balance between glare and illumination, but this analysis considers only the benefits of illumination to pedestrians. This choice was made because 1) previous research identified pedestrian illumination as a particular safety problem associated with darkness (e.g., Sullivan & Flannagan, 2007), and 2) identification of glare-related crashes was not feasible with these data. In addition, the presence/absence of road lighting is not available in all of the datasets and thus, could not be considered in the analysis.

In contrast, driver-side lane-change crashes were more prevalent in the US relative to passenger-side lane-change crashes, as compared to the EU (based on data from two EU countries). In the EU, both mirrors can be non-planar, and thus differences between driver- and passenger-side lane changes should not be related to the mirrors themselves. In contrast, US mirrors on driver and passenger side are different and thus the relative difference in passenger- and driver-side lane-change crashes is expected to be related to the mirror effectiveness in addition to other factors.

Results with similar implications (showing benefit of nonplanar driver-side mirrors vs planar mirrors) have been published in the traffic safety literature. For example, Luoma et al. (2000) reported a statistically significant decrease of 22.9% in lane-change crashes to the driver side for nonplanar mirrors compared to planar mirrors, and Schumann et al. (1998) found a decrease in lane-change crashes of 17.9% for spherical convex versus planar mirrors for midsize vehicles. Finally, Helmers (1992) used a simulator study to investigate this question and found benefits of multi-radius driver-side mirrors and spherical convex mirrors versus planar mirrors in terms of decreased response times for detection of cars at short distances behind in the adjacent lane.

Limitations

The primary limitations of this study arise from data limitations. First, the EU includes 28 countries, but in-depth data suitable for crashworthiness analysis were collected in only 6 of them. We adjusted using the CARE dataset to better represent EU crashes as a whole, but such weighting notably could not account for lower belt-use rates in some countries outside of the data-collection set. For example, IRTAD (2013) reports that seat belt use rates in the front seat are lower in Greece (74%-77%), Italy (63%-75%), and Hungary (82%) in comparison to France (98%), Germany (98%), and the UK (95%). Based on the subgroup breakdowns, if belt-use rates are lower in the EU than in our dataset, overall risk differences would be expected to increase in both populations (i.e., greater negative risk difference for front/side and greater positive risk difference for rollover). Further, the distribution of injury severity for several EU countries observed in CARE led to the observation that there is a tendency towards underreporting of slight or not injured occupants, which in turn may result in increased risk estimates.

Some additional artifacts might account for some of the risk differences seen. For example, the sample analyzed was the population of vehicles purchased by US and EU drivers. If drivers in one country purchase higher-end, safer vehicles on average, the overall risk for that region would be lower. Our assessment of star ratings suggests that there is not a large difference, but we cannot eliminate this possibility. Another possibility is that the inclusion criteria requiring crashes with an injured occupant, combined with higher occupancy in the EU compared to the US, might result in the population of US crashes being somewhat more severe (because multiple occupants provides more opportunities for someone to be injured). However, since the overall risk for the EU population was higher than in the US, this seems unlikely to be influencing results.

Harmonization of datasets was generally successful, but this activity introduces unquantifiable uncertainty—that is, the success of harmonization cannot be tested, so the process itself may introduce variance that cannot be measured. As a result, the likelihood surfaces are relatively flat and it is difficult to distinguish definitively among competing hypotheses. We also cannot be certain that the sampled populations are identical, though we believe that the inclusion criteria harmonization was generally successful in preventing bias.

It is also important to mention that, due to the need to harmonize the inclusion criteria, the crashworthiness analysis addresses the risk of severe or fatal (MAIS3+F) injury *in the event of an injury crash also resulting in a towaway*. This is a slightly different focus than the risk of MAIS3+F injury in case of *any* (unconstrained) crash which is addressed by the regulations. That said, the majority of injuries in the US occur in crashes that would meet these inclusion criteria.

Limitations on data access resulted in challenges that limited the number of iterations for modeling. Though it was still possible to build models that are the same as those that would be generated if data were shared, it was not feasible to explore as many different predictor combinations as we might have liked.

Finally, the headlamp analysis is based on a comparison of dark/light ratios for pedestrian fatalities for the EU and US; however, further research may be needed to link headlamp characteristics to dark/light ratio directly. As for the side mirror analysis, the main limitation was the small number of EU countries having data available, making it unfeasible to draw conclusions on EU level based exclusively on these results. Nevertheless, the results are in line with previous research on this subject; therefore, they supply further evidence for benefit of nonplanar driver-side mirrors versus planar mirrors.

Interpretation of the Crashworthiness Results

The goal of this study was to address the equivalence of the real-world safety performance of passenger vehicles developed in two separate regulatory environments. In principle, the approach is designed to evaluate evidence related to the elements of relative field performance of EU and US vehicles that can be attributed to regulatory differences (rather than environmental differences). In practice, the causal tie between regulatory differences and observed field performance differences cannot be made without randomized controlled trials. Thus, the modeling approach used here can identify observed differences and can eliminate as many alternative explanations as possible, but analysis of observational field data cannot establish cause with certainty.

Two steps in the data analysis served to remove as many alternative explanations as possible. First, we constrained the inclusion criteria for all of the samples to be the same. This way, we sampled from the same population of crashes, even though they may arise very differently in the two regions. Second, we used the same set of predictors to build risk models that estimate injury risk under a specified set of circumstances of the crash, vehicle, or occupant. The circumstances (e.g., occupant age, crash severity, crash direction) were designed to isolate risk from exposure as much as possible. That is, injury risk should not be affected by whether a crash was caused by speeding, texting, or falling asleep at the wheel if the nature of the crash (its direction and severity, indicating the forces acting on the vehicle occupants) is the same. We sought to take these into account in the model.

Although the risk model approach is a good way to separate risk from exposure, it does not perfectly eliminate all possible alternative explanations. (As noted earlier, only randomized controlled trials can demonstrate cause.) In this case, we argue that regulatory differences are the primary mechanism to explain differences between the risks from the two populations. However, because regulation provides a *minimum* standard, one alternative explanation for differences is that one population of vehicle owners tends to purchase safer vehicles (i.e. vehicles higher above the minimum standards) than the other. This cannot be controlled or measured with our datasets and could produce overall differences in risk. A related alternative explanation is that consumer ratings systems, which are also different in the two regions, drive vehicle design, and differences are related to the elements emphasized by the ratings rather than the base regulations. Finally, the possibility exists that data artifacts not accounted for by the models are influencing the results. Significant effort was put into removing foreseeable artifacts, but unforeseen issues are always possible in analysis of observational data.

Finally, we caution the reader in interpreting significance tests and confidence intervals. Standard hypothesis testing, which relies on the p<0.05 rule, considers the question: "What is the probability of getting my results, *if the null hypothesis of no difference were true."* When results are significant, as with

Method 1, the no-difference hypothesis is highly unlikely (less than a 5% chance of being true). However, failure to reach significance, including risk-difference confidence intervals that contain 0, is not evidence *for* the null hypothesis. In this context, where evidence for equivalence is sought, other methods must be considered. In particular, Method 3 approaches the question without setting any hypothesis as the default. Instead, it simply compares evidence for two hypotheses. Similarly, the distributions of probable risk differences in Method 2 give a more complete picture of the uncertainty in the analysis and the relative support for different risk differences.

Conclusions

Crashworthiness:

- EU and US risk models are different for front/side and rollovers. For crashes meeting the inclusion criteria, the risk of MAIS 3+ and fatal injury are significantly different in the EU and US.
- Overall risk across the US front-side crash population (given the selection criteria for this study) is likely lower for EU vehicles, though the range of estimates is wide; the best estimate of the risk difference is -0.012.
- Overall risk across the EU front-side crash population (given the selection criteria for this study) is likely lower for EU vehicles, though the range of estimates is wide; the best estimate of the risk difference is -0.013.
- Overall risk across both EU and US rollover crash populations is lower for US vehicles; the best estimate of the risk difference for the US population is 0.057, and the best estimate of the risk difference for the EU population is 0.036.
- Risk differences in front/side crashes are largest for near-side crashes, middle occupant ages (31-70), unbelted occupants, and higher Delta-Vs. In rollovers, risk differences were highest for unbelted occupants and ejected occupants.

Crash Avoidance:

- US ratio of pedestrian fatalities in dark vs. light is lower than in the EU; one possible explanation for this is that headlamps in US vehicles may imitate daylight better than those in EU vehicles.
- EU ratio of driver-side lane changes compared to passenger-side lane changes, based on data from two EU countries, is lower than in the US. Once possible explanation for this is that driver-side mirrors in EU vehicles reduce risk in lane-change crashes better than those in US vehicles.

Recommended Next Steps

To our knowledge, this is the first side-by-side comparison of predicted risk for EU-regulated and US-regulated vehicles. As such, further work should be done to replicate the results, identify artifacts that may have influenced the patterns seen, and/or seek evidence for mechanisms linking the results to vehicle design differences that result from regulatory differences. We recommend two primary paths for next steps in research.

First, we recommend additional analyses of the field data. In particular, some patterns seen in the breakdowns of subgroups were unexpected. For example, the EU model shows very similar overall predicted risk in near- and far-side crashes while the US model shows higher risk in near-side crashes compared to far-side crashes. Because of the proximity of the occupant to the source of the impact, near-side crashes would be expected to result in greater injury risk. Similarly, the potential effect of the substantially greater share of SUVs and pickup trucks in the US population than in the EU should be examined. Both unexpected and expected results should be looked at closely to identify those that are most robust and those that may be influenced by dataset or population artifacts. Some specific recommended analyses include:

- The variables selected to model injury had significance in at least one of the individual datasets, and interactions between delta V and crash type were included. The effect of considering additional interaction terms or alternate variables could be explored.
- Investigate more specific injury patterns to different body regions between EU and US vehicles to understand what is driving differences.
- Conduct additional analysis to compare the differences in injury risk between near-side and far-side impacts when US and EU models are applied each standard population. Similarly, look closely at the pattern of risk by occupant age for the EU and US models to better understand why the trends differ.
- Investigate whether rollover severity is influenced by ESC, and if so, whether differing ESC penetration in the US and the EU could contribute to differences seen in rollover injury risk
- Investigate the effect on injury risk of selecting crashes based on at least one person in the crash having an MAIS 1 injury. This could be done with the US dataset.
- Identify which variables would be most critical to improve harmonization among global datasets.

Second, we recommend using computational models of typical US-regulated and EU-regulated vehicle designs to investigate potential physical mechanisms of the differences seen. Crash testing is only done in extreme conditions, but most crashes in the field data are lower severity. Computational models allow investigation of injury mechanisms over a wide range of field conditions. When combined with crash data analysis, this approach can help find mechanisms for the results seen in the field (including mechanisms that are not attributable to regulation per se).

Finally, in this project, the use of crash data in various contexts has been demonstrated and at the same time, certain gaps in data availability have been identified. Future reproductions and extensions of this study would greatly benefit from the availability of harmonized accident data, hence further data collection and data harmonization efforts are encouraged.

References

Automotive News, January 9 2012.

Automotive News, January 6, 2014.

Committee for Medicinal Products for Human Use. (2010). Guideline on the investigation of bioequivalence. *European Medicines Agency (EMA), London*, *27*.

ETSC (2009). 2010 on the Horizon. *3rd Road Safety PIN Report*. ETSC, Brussels.

Flannagan, C.A.C., Green, P.E., Klinich, K.D., Manary, M.A., Bálant, A., Sanders, U., Sui, B., Sandqvist, P., Selpi, & Howard, C. (2014). Mutual Recognition Methodology Development, UMTRI Report No. UMTRI-2014-32, http://hdl.handle.net/2027.42/111736

Gordon, T. J., Kostyniuk, L. P., Green, P. E., Barnes, M. A., Blower, D., Blankespoor, A. D., & Bogard, S. E. (2011). Analysis of crash rates and surrogate events. *Transportation Research Record: Journal of the Transportation Research Board*, *2237*(1), 1-9.

Hauer, E. (2004). The harm done by tests of significance. Accident Analysis & Prevention 36, 495—500

Hill, J., Aldah, M., Talbot, R., Giustiniani, G., Fagerlind, H., and Jänsch, M., (2012). *Final Report*, Deliverable 2.5 of the EC FP7 project DaCoTA.

Helmers, G., Flannagan, M. J., Sivak, M., Owens, D. A., Battle, D., & Sato, T. (1992). *Response times using flat, convex, and multiradius rearview mirrors* (No. UMTRI-92-20).

Informedforlife.org

International Road Traffic Accident Database (IRTAD). (2013). Road safety annual report, OECD/ITF.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, Vol. 90, No. 430, pp. 773-795.

Kullgren, A., Lie, A. and Tingvall, C. (2010). Comparison Between Euro NCAP Test Results and Real-World Crash Data. *Traffic injury Prevention*, 11:587-593.

Luoma, J., Flannagan, M. J., & Sivak, M. (2000). Effects of nonplanar driver-side mirrors on lane-change crashes. *Transportation Human Factors*, *2*(3), 279-289.

Niebuhr, T., Achmus, S. and Kreiß, J.-P. (2011). *Testing for similarity of distributions*, Technical Report

Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika, 66(3),* pp. 403-411.

Rameshkrishnan, N., Sathyakumar, A., Balakumar, A., Hassan, R., Rajaraman, R., Padmanaban, J. (2013). The New In-Depth, At-the-Scene, Accident Investigation Database in India. *Proceedings International Research Council on the Biomechanics of Impact* (IRCOBI), Gothenburg.

Saurabh, Verma (2013). *CADaS Common Accident Data Set.* Common Accident Data Set Reference Guide

Version 3.1. European Commission. Retrieved from http://ec.europa.eu/transport/road_safety/pdf/statistics/cadas_glossary.pdf

Schumann, J., Sivak, M., & Flannagan, M. J. (1998). Are driver-side convex mirrors helpful or harmful?. *International Journal of Vehicle Design*, *19*(1), 29-40.

Sharma, D., Stern, S., Brophy J., Choi, E-H. (2007) An Overview of NHTSA's Crash Reconstruction Software WinSMASH. ESV Conference Paper 07-0211.

Sullivan, J.M. and Flannagan, M.J. (2007). Determining the potential safety benefit of improved lighting in three pedestrian crash scenarios. *Accident Analysis and Prevention*, 39 (2007), pp. 638–647

Weisstein, Eric W. "Convex Hull." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/ConvexHull.html, accessed Jan. 14, 2015.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association, 57*, pp. 348–368.

Appendix A Logistic Regression

Bernoulli Random Variable

A discrete random variable *Y* whose probability mass function (pmf) is given, for some 0<*p*<1, by Equation A1 is said to be a Bernoulli random variable with parameter *p*. This random variable has only two outcomes: *y*=0 or *y*=1. The outcome is typically called a "success" when *y*=1 and as a "failure" when *y*=0. However, in this application, we use "injured" for *y*=1, which corresponds to MAIS 3+ injury or fatality, and "uninjured" for *y*=0, which corresponds to MAIS<3.

$$f(y|p) = P(Y = y) = p^y(1-p)^{1-y} \quad y = 0,1 \tag{A1}$$

Note that $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$ and it can be shown that the mean and variance of *Y* are given by Equations A2 and A3.

$$E[Y] = p \tag{A2}$$

$$Var[Y] = p(1-p) \tag{A3}$$

The Bernoulli Random Variable and the Natural Exponential Family

A distribution indexed by parameter θ belongs to the natural exponential family if it can be written as in Equation A4.

$$f(y|\theta) = a(\theta)b(y)\mathrm{e}^{yQ(\theta)} \tag{A4}$$

where $Q(\theta)$ is called the natural parameter. If *Y*|*p* is Bernoulli(*p*), then Equation A5 describes its density function.

$$f(y|p) = p^y(1-p)^{1-y} = (1-p)\left(\tfrac{p}{1-p}\right)^y = (1-p)e^{y\log\left(\frac{p}{1-p}\right)} \quad y = 0,1 \tag{A5}$$

This is the exponential family with

$$a(p) = 1 - p \qquad b(y) = 1 \qquad Q(p) = \log\left(\frac{p}{1-p}\right)$$

and *Q(p)* is the natural parameter and represents the log odds of injury.

The Logistic Regression Model for a Binary Response

For a sample of *N* independent observations, the model is given in Equation A6.

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x_i}'\boldsymbol{\beta} \qquad \text{(A6)}$$

$Y_i|p_i \sim \text{Bernoulli}(p_i)$ and $Y_i|p_i$ independent for $i = 1, \ldots, N$

where $\boldsymbol{x_i}$ is an *r*X1 vector of predictor variables for subject *i*, and $\boldsymbol{\beta}$ is an *r*X1 vector of unknown regression parameters to be estimated. The left-hand side of the model equation is $Q(p_i)$. Solving for p_i and $1 - p_i$ we get Equation A7.

$$p_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \qquad 1 - p_i = \frac{1}{1+e^{x_i'\beta}} \qquad 0 < p_i < 1 \qquad \text{(A7)}$$

Estimation of $\boldsymbol{\beta}$ by the Maximum Likelihood Method

The likelihood function for *N* independent observations is given by Equation A8.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} f(y_i|p_i) \qquad \text{(A8)}$$

The maximum likelihood estimator (MLE) is the value of $\boldsymbol{\beta}$ denoted by $\widehat{\boldsymbol{\beta}}$ that maximizes the likelihood. Because the natural log function is monotonic, maximizing the log likelihood is equivalent to maximizing the likelihood. In general, it is easier to maximize the log likelihood function, which is given in Equation A9.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log f(y_i|p_i) \qquad \text{(A9)}$$

Taking derivatives of the log likelihood gives the likelihood equations in Equation A10.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \left[y_i - \left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)\right] x_{ij} = 0 \qquad j = 1, \ldots, r \qquad \text{(A10)}$$

where *r* is the number of parameters including the intercept, and $x_{i1} = 1$ for the intercept term.

The likelihood equations are a set of *r* equations and *r* unknowns, with the MLE $\widehat{\boldsymbol{\beta}}$ as a unique solution under regular conditions. Unlike the normal theory linear model, where the solution $\widehat{\boldsymbol{\beta}}$ can be written in closed form, the likelihood equations for the logistic model are nonlinear in $\boldsymbol{\beta}$ and the solution cannot be written in closed form. Statistical software packages, such as R or SAS, use algorithms to search the likelihood space iteratively until a solution is found.

The Variance-Covariance Matrix

Under certain regularity conditions, MLE's are consistent and asymptotically normal. That is, as *N* gets large, the MLE $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$, and converges in distribution to an *r*-variate normal distribution. The variance of $\widehat{\boldsymbol{\beta}}$ is estimated by the inverse of the *rXr* Fisher information matrix. The expected Fisher information is denoted by Equation A11.

$$I(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}\boldsymbol{\beta}'}\right] \text{ and } Var(\widehat{\boldsymbol{\beta}}) \approx \left[I(\widehat{\boldsymbol{\beta}})\right]^{-1} \qquad \text{(A11)}$$

To derive the Fisher Information matrix, consider Equation A10, which shows the likelihood equations of first derivatives. Since observations are independent, we can take the second derivatives for the *i*th observation, which leads to Equation A12.

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_h} = -x_{ij} x_{ih} p_i (1 - p_i) \quad j, h = 1, \dots, r \tag{A12}$$

Taking the negative value and summing across all *i* observations, we get the individual entries in the *rXr* information matrix, given in Equation A13.

$$-\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_h} = \sum_{i=1}^{N} x_{ij} x_{ih} p_i (1 - p_i) \quad j, h = 1, \dots, r \tag{A13}$$

Since the quantity in Equations A13 does not depend on the random variable Y_i, the expected information equals the observed information and there is no need to take expected values.

Appendix B Computing Log-Likelihood

One of this project's particular challenges was the inability to share and combine raw data from the EU datasets. In a typical analysis using logistic regression, raw data would be in a single datafile and would be analyzed using statistical software that takes advantage of efficient iterative search techniques to find the maximum likelihood. In this project, we could only share summary statistics from separate analyses of each dataset.

To develop an EU model without sharing raw data, we took advantage of the fact that the log-likelihood and each cell in the Fisher Information matrix are sums across observations. Thus, for any specific model, we can sum the log-likelihood and cells in the Fisher Information matrix within a dataset, share only the totals, and then add these together to replicate exactly what would have resulted from the raw data residing in a common database.

The log-likelihood is in Equation B1 and the Fisher Information matrix is in Equation B2.

$$\mathcal{L} = \sum_{i=1}^{n} (y_i \log{(\hat{p}_i)} + (1 - y_i)\log(1 - \hat{p}_i)) \qquad \text{(B1)}$$

$$-\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_h} = \sum_{i=1}^{N} x_{ij} x_{ih} p_i (1 - p_i) \quad j, h = 1, \ldots, r \qquad \text{(B2)}$$

Although we can compute log-likelihood and variance (by inverting the combined Fisher Information matrix), we cannot take advantage of powerful iterative techniques to explore the search space. Those techniques require many iterations, often hundreds, whereas in this application, one iteration could take many hours. The solution to this logistical issue was to select a large number of test points in the model space and have each group compute sums for those test points and then add the results together for all points simultaneously. This could then be repeated a handful of times with new large sets of test points, rather than hundreds of times with a single point.

For the front/side model, there were 18 coefficients, including the intercept. This meant that the model search space was an 18-dimensional space, with each dimension defined by the value of one coefficient. Each point represents a single model, defined by its location in each of the 18 dimensions (i.e., coefficient values), and Equations B1 and B2 can be computed for each point.

Computationally, Equation B1 could be somewhat time-consuming, depending on the software used by each group. (In retrospect, we would have obtained faster software to solve this problem had we known at the outset that it would be an issue.) Equation B2 was substantially more time-consuming, but was only needed for the best-fit model. Thus, to make the process manageable, we computed only the log-likelihood for groups of points that were selected judiciously as described in the next paragraphs.

Searching the Model Space

Given the constraints on the number of points that could be processed we needed to select points in an intelligent way. The simplest starting point would be to select a range of plausible coefficient values and create an 18-dimensional hypercube of points. However, this approach places a great deal of emphasis on unlikely corners of the search space (i.e., those that combined unlikely values of many parameters) and is thus inefficient.

Instead, we created a prior distribution on the whole space, initially based on the original separate models and later based on prior test points. The prior distribution was an 18-dimensional multivariate normal distribution with mean and variance selected for each coefficient. In one case, age and age-squared, we include a correlation parameter in the multivariate normal. However, we ignored correlations between other predictors because it was not critical to the activity.

The first round of multinormal parameters were selected based on the coefficients of the five separate models (one for each dataset, including the US). The mean was selected to be the mean across datasets, but the variance was enlarged to ensure that all observed coefficients plus a value of 0 were included within 2 standard deviations of the mean. Although we did not exclude any of the predictors after the initial parameter-selection stage, including zero allowed parameters to go to zero if the model called for this.

Because the probability distribution of the standard multinormal is an r-dimensional hypersphere, we selected groups of random points on spheres of varying radius. The process is very simple, and is described in Marsaglia (1972). We select *r* random observations from a standard normal distribution, where *r* is the number of parameters or dimensions (in this case, 18 for front/side and 9 for rollover). Then, points defined as in Equation B3 are uniformly distributed over a hypersphere of radius 1.

$$\frac{1}{\sqrt{x_1^2+x_2^2+\cdots+x_r^2}}\begin{bmatrix}x_1\\ \vdots\\ x_r\end{bmatrix} \tag{B3}$$

It is straightforward to transform points on a unit normal hypersphere to points in the original (coefficient value) units and points at different radii for that unit hypersphere. To further improve the point-selection process, we eliminated the 10% of randomly selected points that were closest to another point. This way, points were spread apart to maximize coverage per test point.

For the first round, we selected 3000 points at each radius in even steps of 0.1 from 0.1 to 3.1 (these are standard deviation units). After the elimination process, there were 78,330 test points for the first set in the front-side model. Since we cannot plot points in 18-dimensional space, we looked at histograms of coefficient values across the set of test points. These are shown in Figure 37. As desired, coefficient values tested clustered around the most likely values, but were still spread out. A hypercube approach would have produced flat graphs in the figure and would have been highly inefficient.

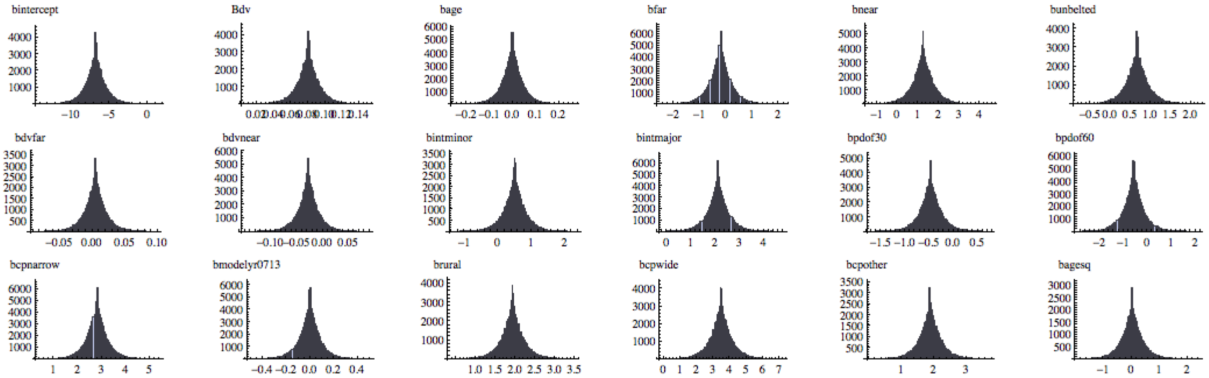

Figure 37. Initial distributions of parameters tested to generate maximum likelihood surface.

Using the initial set of points, each group calculated log-likelihood for their dataset, using the combined national and EU weights. An additional weighting factor was applied to each sum to normalize the weighted totals to the raw case count contributed by each dataset. Normalization to raw sample size is commonly used with logistic regression to ensure that variance estimates are appropriately scaled.

To visualize these results, we plotted log-likelihood against predicted risk. Example results are shown in the top four plots of Figure 38, which show results for the four component EU datasets. Models near the top of the curve (smaller negative values of log likelihood) are the best models. The combined EU likelihood is created by adding together the four independent EU likelihood surfaces as shown in the bottom of Figure 38.

After testing the initial range of parameter estimates, the focus shifted to the set of points near the peak of the log-likelihood distribution (i.e., the best models so far). This process is illustrated in Figure 39. The upper graph shows the larger-scale point cloud, and the lower-left plot zooms in on the region of the best models. From this region, we selected the top 50-75 points and used each as the center of a new multinormal hypersphere. To simplify the process, we retained the original variances. Around each point, we selected 75 new points at each of several radii, removed the closest points, and compiled a set of 48,554 points for the second set.

The pink points in Figure 39 are the values from the second point set, showing how successful the process was at filling in the space near the peak (as desired). Figure 40 shows the histograms of parameter values selected in the second iteration. Notice that the second set of histograms show multimodal distributions for many parameters. Since the second set of test points was based on many multinormal hyperspheres, rather than a single one, the points clustered around different parameter values and we get the pattern seen in the figure.

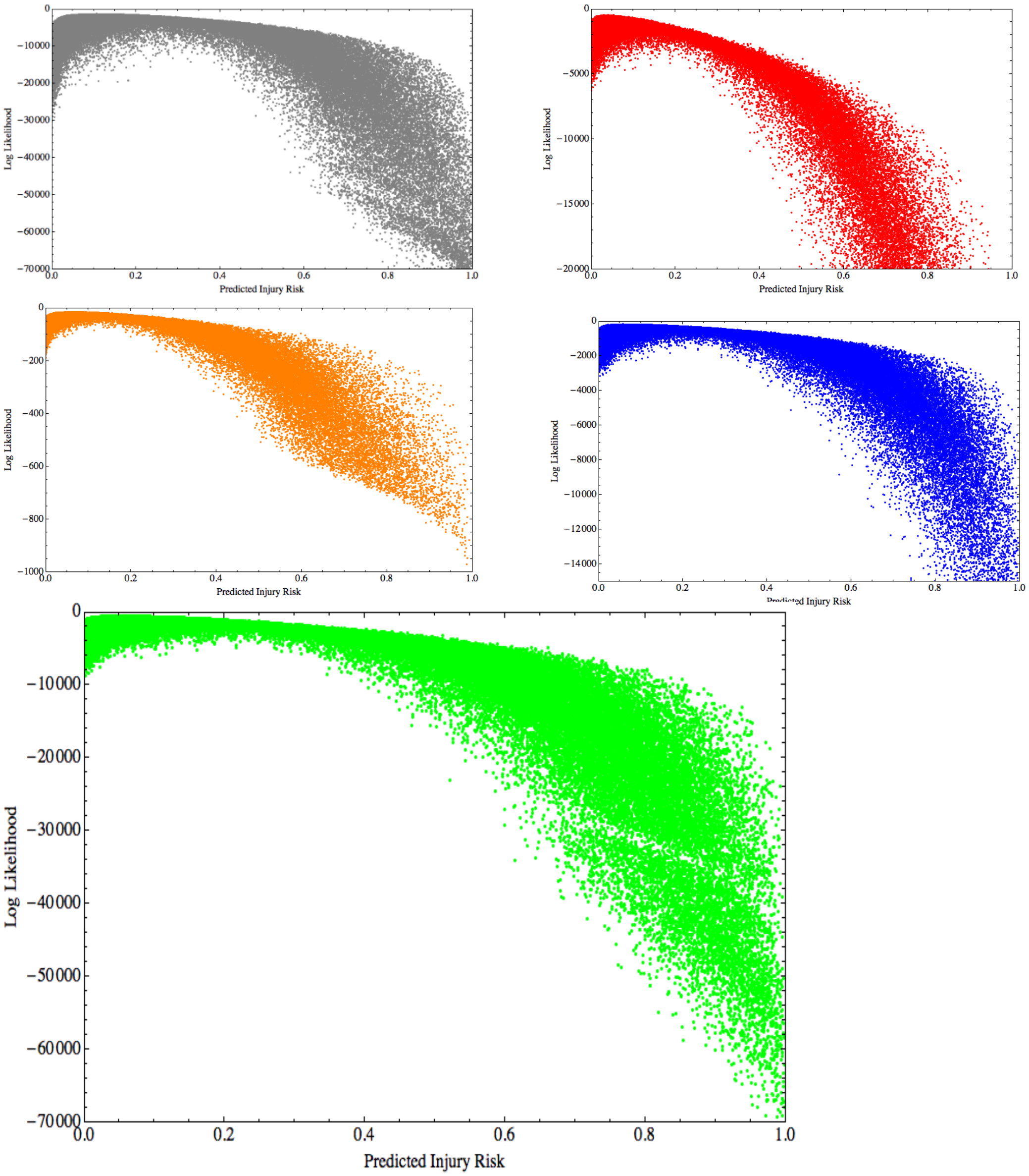

Figure 38. Four maximum likelihood models generated independently on different EU datasets added together to produce combined EU likelihood model.

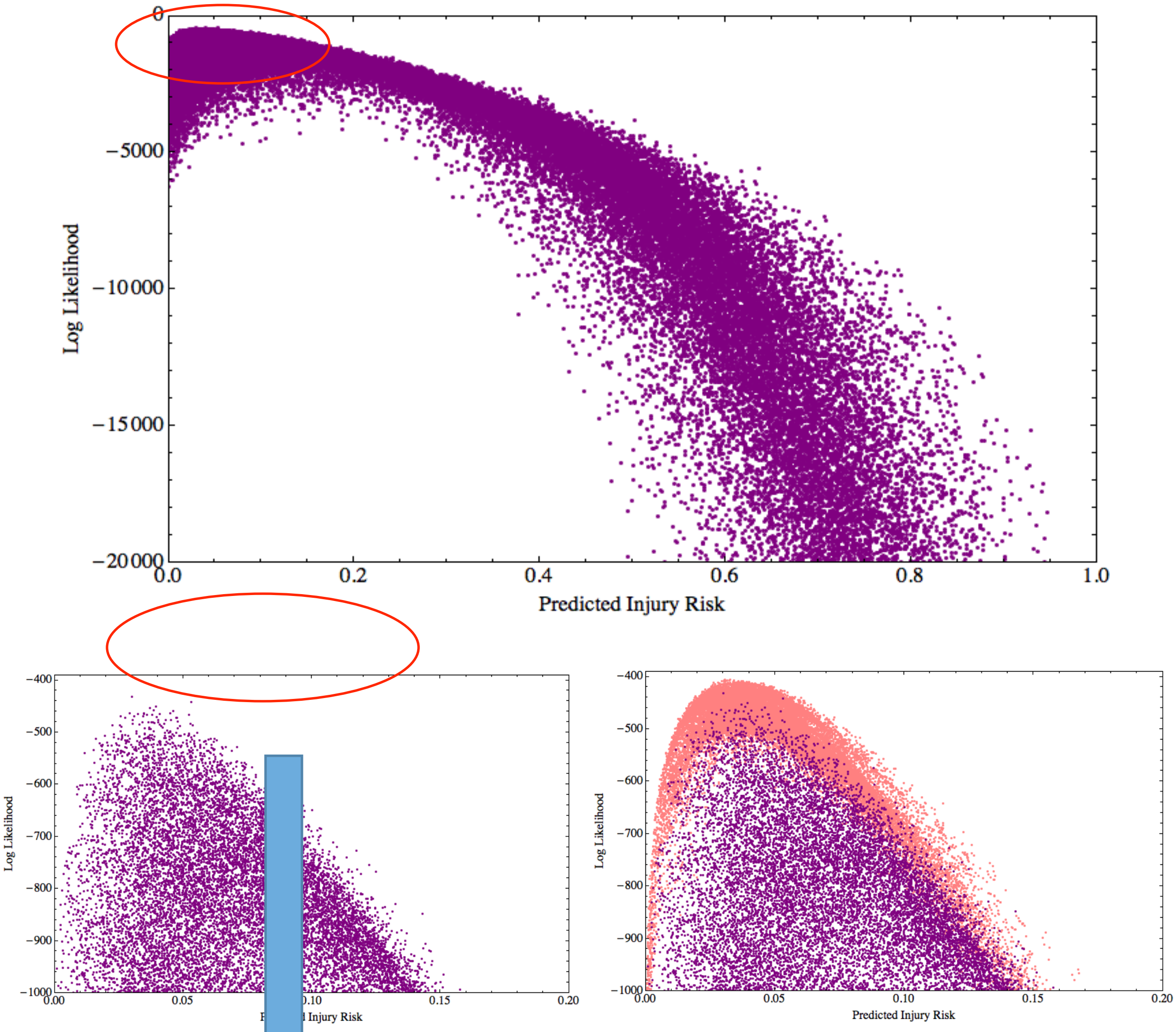

Figure 39. Iterati maximum likelihood surface. Each point represents a unique set of parameters. After generatin itial likelihood surface (top), zoom in on area represent best models (highest likelihood, botto eft). Select additional parameter sets that will increase the number of models tested in the range of highest likelihoods.

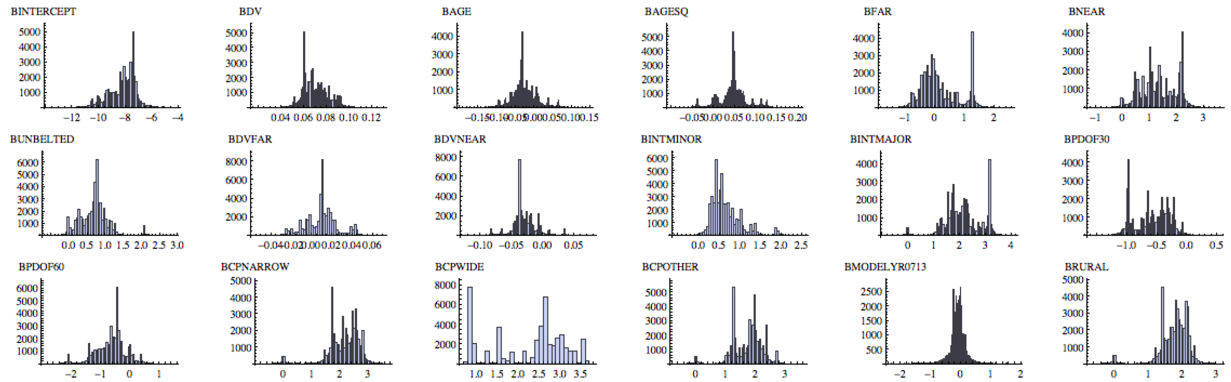

Figure 40. Range of parameter values for maximum likelihood surface after first iteration.

The process was repeated until the maximum log-likelihood changed by less than 5 units for front/side and 1 unit for rollover. For the front/side models, sufficient results were achieved after five iterations, while four iterations were performed for rollover. After the final iteration, the model parameters producing the highest maximum likelihood (Figure 41) were chosen as the best model and used in subsequent analyses.

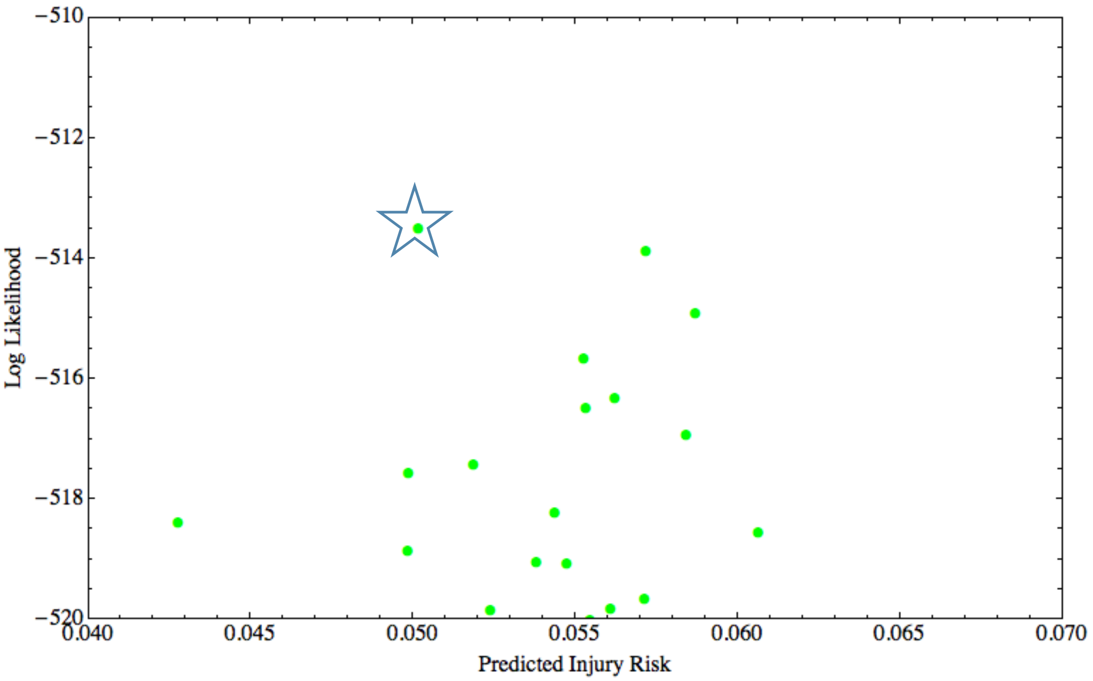

Figure 41. Close-up view of second iteration best models. Parameters used to generate the best model (star) used for the final model.

Once the best model was selected, the cells of the Fisher Information matrix were computed for that model alone. The cell values were weighted (using the normalizing weights for the four datasets) and summed, and then the entire matrix was inverted to produce the variance-covariance matrix for the model as a whole.

Appendix C Seemingly Unrelated Regression

Seemingly Unrelated Regression (SUR) for Logistic Regression Based on Injury Outcome using NASS CDS Data

Data

Five years of CDS data (2005-2009) were assembled. Two data sets were created by randomly sampling occupants from each of the 27 Primary Sampling Units (PSUs) with probability 1/2. This procedure resulted in 26,719 occupants being randomly allocated to Data Set 1, and 26,704 occupants being randomly allocated to Data Set 2.

Analysis was restricted to the following conditions.

- Passenger cars with model year greater than or equal to 1995
- Frontal collisions (general area of damage = 'Front')
- No rollovers
- Occupant age greater than or equal to 18 years
- Drivers and front seat passengers
- Occupants not ejected

Method

Logistic regression was performed using the following variables:

Binary response variable Y:

injvar - 1= MAIS3+ and Fatal,   0= MAIS(0-2)

Predictor variables (X):

gender - 1=male,   0=female
occage - occupant age (continuous)
deltav - total delta-v (continuous)

To perform Seemingly Unrelated Regression (SUR), the response variables from Data Set 1 and Data Set 2 were stacked, as were the predictor variables from the two data sets. The notation below describes the data used in the regression model. Subscripts 1 and 2 denote data from Data Set 1 and Data Set 2, respectively.

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{X}_1 & 0 \\ 0 & \boldsymbol{X}_2 \end{bmatrix}$$

A single logistic regression model was fit to the resulting data giving rise to regression coefficients corresponding to the two data sets as shown below.

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Here $\boldsymbol{\beta}_1 = (\beta_{intcpt1}, \beta_{gender1}, \beta_{occage1}, \beta_{deltav1})'$ and $\boldsymbol{\beta}_2 = (\beta_{intcpt2}, \beta_{gender2}, \beta_{occage2}, \beta_{deltav2})'$. Having specified the model in a framework that includes data from both data sets, it is now possible to consider tests of hypotheses such as

$$H_0 : \beta_{deltav1} = \beta_{deltav2}$$

Results

The `survey` package from the R statistical software was used to fit logistic regression models, taking into account the CDS survey design and sampling weights. Before showing results from the SUR model, two logistic regression models were fit to Data Sets 1 and 2 separately and results are shown below. Results from the SUR model follow.

Regression for Data Set 1:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.907274   0.500408 -15.802 2.14e-09 ***
gender1     -0.520066   0.208872  -2.490   0.0284 *
occage1      0.035565   0.004161   8.546 1.90e-06 ***
deltav1      0.106696   0.008240  12.949 2.06e-08 ***
```

Regression for Data Set 2:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.47142    0.65441 -11.417 8.41e-08 ***
gender2     -0.28460    0.36462  -0.781     0.45
occage2      0.03256    0.00477   6.826 1.83e-05 ***
deltav2      0.09685    0.01127   8.596 1.79e-06 ***
```

SUR for combined Data:

```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Intcpt1 -7.907274   0.500408 -15.802 2.57e-07 ***
gender1 -0.520066   0.208872  -2.490 0.037529 *
occage1  0.035565   0.004161   8.546 2.71e-05 ***
deltav1  0.106696   0.008240  12.949 1.20e-06 ***
Intcpt2 -7.471423   0.654409 -11.417 3.13e-06 ***
gender2 -0.284604   0.364621  -0.781 0.457540
occage2  0.032564   0.004770   6.826 0.000134 ***
deltav2  0.096855   0.011267   8.596 2.59e-05 ***
```

Covariance and Correlation Matrices of Parameter Estimates

The covariance matrix of the parameter estimates, denoted by $(\boldsymbol{X'WX})^{-1}$, is shown below

```
          Intcpt1   gender1   occage1   deltav1   Intcpt2   gender2   occage2   deltav2
Intcpt1  0.250408 -0.020800 -0.001739 -0.003662 -0.187724  0.111955  0.000286  0.001144
gender1 -0.020800  0.043627  0.000036 -0.000160  0.074287 -0.023610 -0.000053 -0.000451
occage1 -0.001739  0.000036  0.000017  0.000023  0.000597 -0.000650  0.000001  0.000005
deltav1 -0.003662 -0.000160  0.000023  0.000068  0.002361 -0.001085 -0.000011 -0.000020
Intcpt2 -0.187724  0.074287  0.000597  0.002361  0.428251 -0.139390 -0.001210 -0.004496
gender2  0.111955 -0.023610 -0.000650 -0.001085 -0.139390  0.132948  0.000213 -0.000263
occage2  0.000286 -0.000053  0.000001 -0.000011 -0.001210  0.000213  0.000023 -0.000004
deltav2  0.001144 -0.000451  0.000005 -0.000020 -0.004496 -0.000263 -0.000004  0.000127
```

For comparison the correlation matrix is

```
         Intcpt1  gender1  occage1  deltav1  Intcpt2  gender2  occage2  deltav2
Intcpt1  1.00000 -0.19900 -0.83519 -0.88818 -0.57325  0.61359  0.11979  0.20295
gender1 -0.19900  1.00000  0.04118 -0.09284  0.54348 -0.31001 -0.05291 -0.19161
occage1 -0.83519  0.04118  1.00000  0.67284  0.21938 -0.42824  0.05928  0.10031
deltav1 -0.88818 -0.09284  0.67284  1.00000  0.43783 -0.36113 -0.27734 -0.21936
Intcpt2 -0.57325  0.54348  0.21938  0.43783  1.00000 -0.58417 -0.38774 -0.60972
gender2  0.61359 -0.31001 -0.42824 -0.36113 -0.58417  1.00000  0.12246 -0.06394
occage2  0.11979 -0.05291  0.05928 -0.27734 -0.38774  0.12246  1.00000 -0.06636
deltav2  0.20295 -0.19161  0.10031 -0.21936 -0.60972 -0.06394 -0.06636  1.00000
```

Hypothesis Testing

A Wald test is based on the large sample normal distribution of the parameter estimates.

Let $\hat{\boldsymbol{\beta}}$ be the 8×1 vector of coefficient estimates and let $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$ be the 8×8 covariance matrix. If

$$\hat{\boldsymbol{\beta}} \sim N_8(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1})$$

then for a $q \times 8$ matrix of constants $\boldsymbol{A}$

$$\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim N_q(\boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{A}')$$

and the quadratic form

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\boldsymbol{\beta})'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\boldsymbol{\beta}) \sim \chi^2_q$$

Create the matrix $\boldsymbol{A}$ to satisfy H_0 using indicators in the places corresponding to the model fit. For example, to test

$$\begin{aligned} H_0 : \beta_{deltav1} &= \beta_{deltav2} \\ \beta_{occage1} &= \beta_{occage2} \end{aligned}$$

$$\boldsymbol{A} = \left[\begin{array}{cccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \end{array} \right]$$

and

$$X^2 = (\boldsymbol{A}\hat{\boldsymbol{\beta}})'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}) \sim \chi^2_2 \qquad \text{when } H_0 \text{ is true.}$$

The p-value is the area to the right of X^2 in a chi-squared distribution on 2 df.

```
Wald test:
----------
Chi-squared test:
X2 = 0.52, df = 2, P(> X2) = 0.77
```

The test statistic $X^2 = 0.52$ on 2 df and the p-value is 0.77. According to this test, the result does not come close to significance and we fail to reject H_0.

Appendix D Asymptotic Normality of $\widehat{\boldsymbol{p}}$

According to the large sample properties of MLEs, the sampling distribution of the regression coefficients in a logistic regression model is approximately multivariate normal. In particular, for a p-vector of regression coefficients $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, [I(\boldsymbol{\beta})]^{-1})$$

where $I(\boldsymbol{\beta})$ is the $p \times p$ Fisher information matrix. The linear predictor is $\hat{\eta}_i = \boldsymbol{x}_i^{'} \hat{\boldsymbol{\beta}}$ where $\boldsymbol{x}_i$ is the p-vector of predictor variables for observation i. The variance of the linear predictor is

$$Var(\hat{\eta}_i) = Var(\boldsymbol{x}_i^{'} \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i^{'}[I(\boldsymbol{\beta})]^{-1}\boldsymbol{x}_i$$

The fitted values $\hat{p}_i$ are given by

$$\hat{p}_i = g(\hat{\eta}_i) = \frac{1}{1 + e^{-\hat{\eta}_i}}$$

By the delta method, $\hat{p}_i$ has a large sample normal distribution with variance

$$Var(\hat{p}_i) = Var(g(\hat{\eta}_i)) = [g'(\eta_i)]^2\, Var(\hat{\eta}_i)$$

Note that

$$g'(\eta_i) = \frac{e^{-\eta_i}}{(1 + e^{-\eta_i})^2} = p_i(1 - p_i)$$

and

$$Var(\hat{p}_i) = p_i^2(1 - p_i)^2\, \boldsymbol{x}_i^{'}\, [I(\boldsymbol{\beta})]^{-1}\boldsymbol{x}_i$$

giving

$$\hat{se}(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)\sqrt{\boldsymbol{x}_i^{'}\, [I(\hat{\boldsymbol{\beta}})]^{-1}\boldsymbol{x}_i}$$

Appendix E Estimating Bayes Factors Using the Schwarz Criterion

Bayes Factors are ratios of evidence for two different hypotheses, where evidence is measured as the likelihood of the data, given a hypothesis. The basic equation for Bayes Factors is shown in Equation E1.

$$B_{i0} = \frac{p(\boldsymbol{D}|H_i)}{p(\boldsymbol{D}|H_0)} \tag{E1}$$

where B_{i0} is the Bayes Factor comparing a hypothesized risk difference of *i* to a risk difference of zero, ***D*** is the observed data, H_i is the group of models that result in a risk difference of *i*, and H_0 is the group of models that result in a risk difference of zero. ("Zero" in this context actually denotes an interval around zero whose width is agreed upon based on a reasonable definition of practically no difference.) Note that the hypothesis of zero risk difference is not treated as a null hypothesis in the same way as in Method 1. However, it is treated as the comparison hypothesis for all other hypotheses. In principle, any risk-difference hypothesis can be compared to any other risk-difference hypothesis using this method.

In applications such as this one, each hypothesis can be represented by a large number of specific models. For example, many models in this space result in zero risk difference, and many other models result in a risk difference of 0.001. In this situation, the probability of the data given the hypothesis is shown in Equation E2.

$$p(\boldsymbol{D}|H_k) = \int p\,(\boldsymbol{D}|\theta_k, H_k)\pi(\theta_k|H_k)d\theta_k \tag{E2}$$

where θ_k is a set of coefficients (i.e., a model) that result in a risk difference of *k*, and $\pi(\theta_k|H_k)$ is the prior probability of θ_k given the hypothesis H_k.

The direct computation of Equation 4 can be difficult, especially on a large dataset. As a result, Bayes Factors are generally estimated rather than computed directly. Different estimation approaches employ different methods to defining the prior probabilities. However, in this analysis, we have no clear means of assigning prior probabilities, and thus prefer an estimation method for which priors will have little or no effect on the estimated Bayes Factors. The specific estimation approach we selected is the Schwarz Criterion, which is ideal for this application because 1) it uses log-likelihood, which we already need to compute for a large set of models for Methods 1 and 2; and 2) it does not make strong assumptions about the prior probability of each model within a hypothesis. Instead of introducing prior probabilities for each potential model, the Schwarz Criterion uses the log-likelihood of the *best* model within each hypothesis, as in Equation E3.

$$S = \log \mathrm{pr}\big(\boldsymbol{D}|\hat{\theta}_1, H_1\big) - \log \mathrm{pr}\big(\boldsymbol{D}|\hat{\theta}_2, H_2\big) - \frac{1}{2}(d_1 - d_2)\log(n) \tag{E3}$$

where *S* is the estimated log Bayes Factor, $\hat{\theta}_k$is the MLE under H_k, d_k is the dimension (number of df) of $\hat{\theta}_k$ and *n* is the sample size.

Since all models in this application use the same predictors, $\frac{1}{2}(d_1 - d_2)\log(n) = 0$ and *S* depends only on the likelihood of the MLE for the two hypotheses being compared.

To understand the Schwarz Criterion approach, it is useful to illustrate using the plots of log likelihood vs. injury risk. If we take a narrow vertical slice of risk, all the models within that slice are associated with that injury risk (within a small window). Figure 42 illustrates this slice for a predicted injury risk of approximately 0.20 (using the range 0.19-0.21) for the EU population. Note that this is clearly not the overall best model because an overall risk of 0.20 is less likely than smaller risk values, but given the target risk, the associated best model has a log likelihood of about -1500.

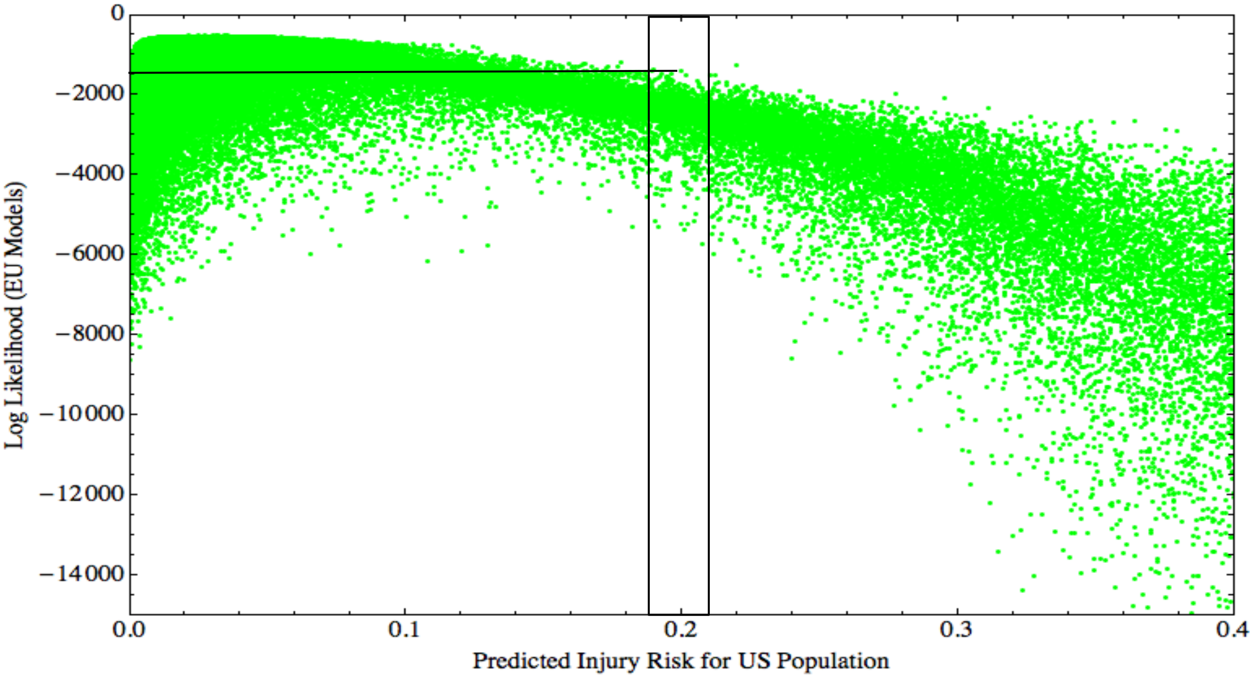

Figure 42. All points within slice represent models whose parameters predict injury risk from 0.19 to 0.21. Log likelihood value associated with highest point (~-1500) represents our best estimate for risk from 0.19 to 0.21.

To generate the best model for a given risk *difference,* we must also take slices for the US model (purple). Note that the predicted risk in each plot must be for the same standard population, whereas the likelihood is determined by the development dataset. Figure 43 illustrates the process. For the US (purple) and EU (green), with risk evaluated for the US standard population, we choose intervals I_1 and I_2 on each plot corresponding to risk windows around 0.1 and 0.2. For the US model, α_1 and α_2 are the best models that results in predicted injury risk for the window around I_1 and I_2, respectively, while β_1 and β_2 are the best models that result in predicted injury risk for the same windows.

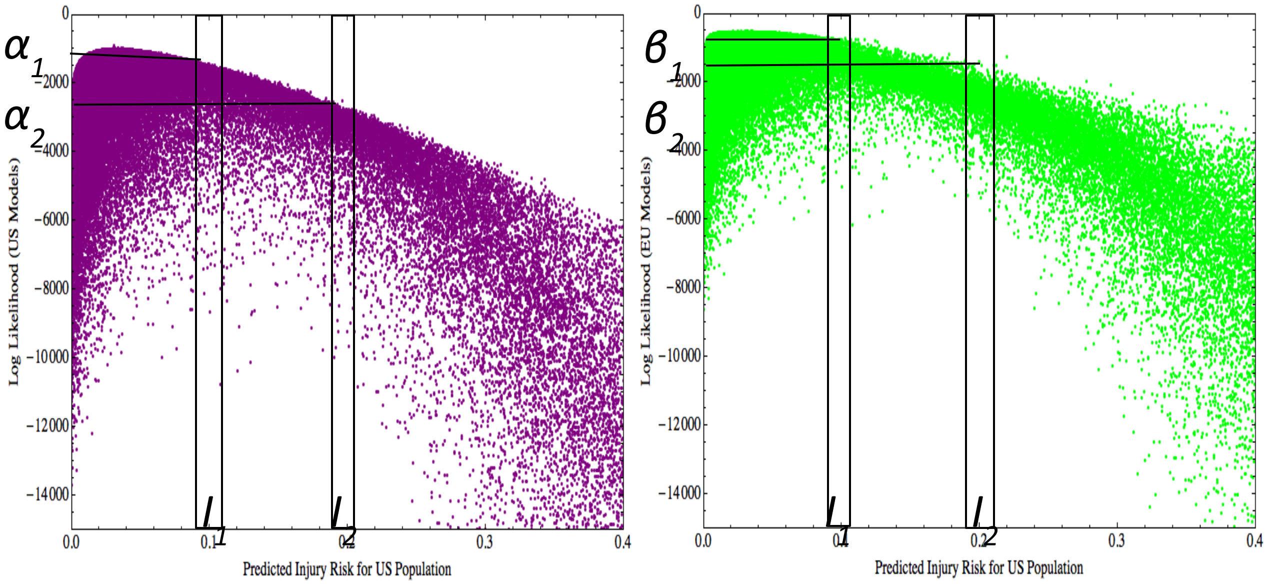

Figure 43. Calculating risk differences for zero difference

One model scenario resulting in zero difference between EU and US risk is that I_1 is the same for both models. The total log-likelihood of this model (i.e., the *evidence* for this model) is the sum of the log-likelihood of each component model, which is computed by adding α_1 and $\beta_{1.}$ A second model scenario resulting in zero difference between EU and US risk is that I_2 is the same for both models. The total log-likelihood (evidence) for this model is the sum of α_2 and $\beta_{2.}$ The process is repeated multiple times by considering different slice intervals where risk would be the same in both models. The highest total log-likelihood from all of those slice pairs associated with zero difference (θ_0) is the maximum likelihood estimator (MLE) of zero difference between the models. For the Schwarz Criterion, the total log-likelihood of the MLE for the zero-difference model is the logarithm of the estimated denominator of Equation E1.

The computation for alternative hypotheses is illustrated in Figure 44 . This time, we choose slices I_1 and I_2 corresponding to risk windows around 0.1 and 0.2 on the US plot, but choose slices I_3 and I_4 on the EU plot corresponding to risk windows around 0.15 and 0.25. Two different estimates of risk difference equal to 0.05 are I_1 & $I_3 = \alpha_1 + \beta_3$, or I_2 & $I_4 = \alpha_2 + \beta_4$. This process is repeated to estimate multiple possible total log-likelihood values for a risk difference of 0.05; the highest log-likelihood out of all of these is the evidence for a risk difference =0.05 (θ_1). This value is the logarithm of the numerator of Equation E1 when the 0.05 risk difference is considered.

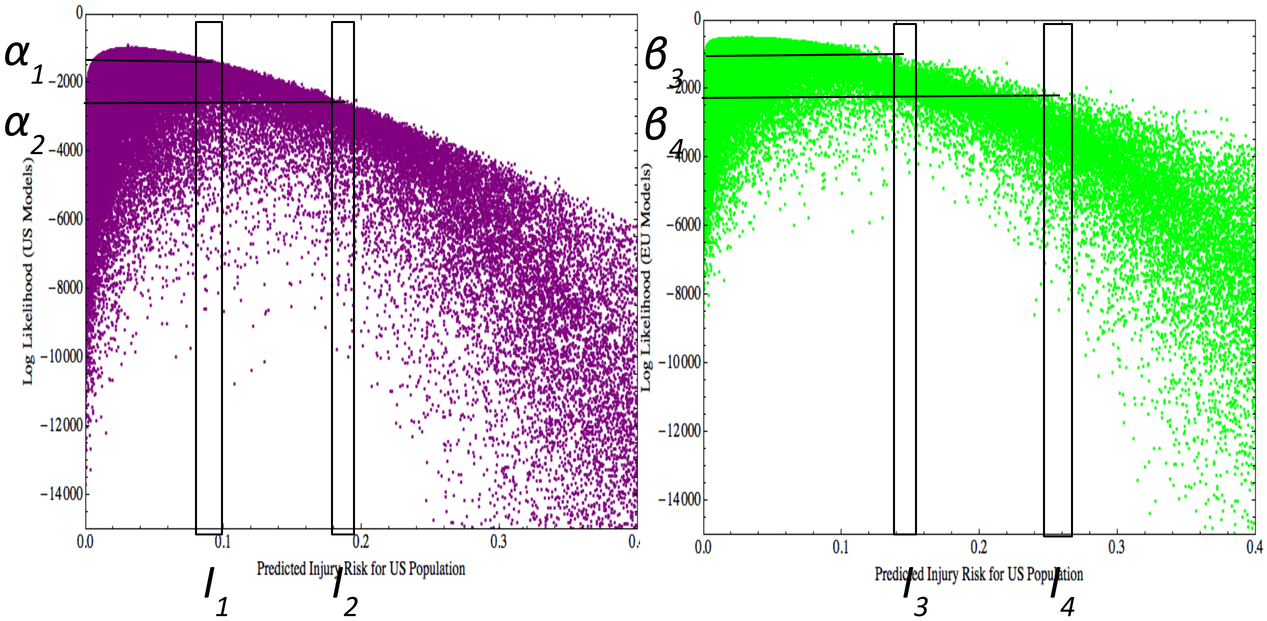

Figure 44. Calculating risk differences for 0.05 difference.

A close look at Figure 43 and Figure 44 shows that points are sparser as the models are less likely. This is because point selection for testing was focused on details at the peak. Because of this, we did not directly estimate the log likelihood for each slice using the tested points, but instead, we generated a smooth upper contour using interpolated convex hulls.

The contour estimation process is illustrated in Figure 45Figure 45 The convex hull of a set of points, *P*, is the intersection of all convex sets containing *P* (Weisstein, accessed 2015). In essence, the convex hull contains all of the outermost points of the point cloud. In Figure 45 these are the red points. Once these points are selected, then the outer contour of the likelihood cloud can be generated using linear interpolation. Thus, the interpolated line in Figure 46 was used to estimate the log likelihood for each risk value as the window was moved across the graph. This way, the components of the Bayes Factors are not influenced by the particular points that were chosen for testing in each region.

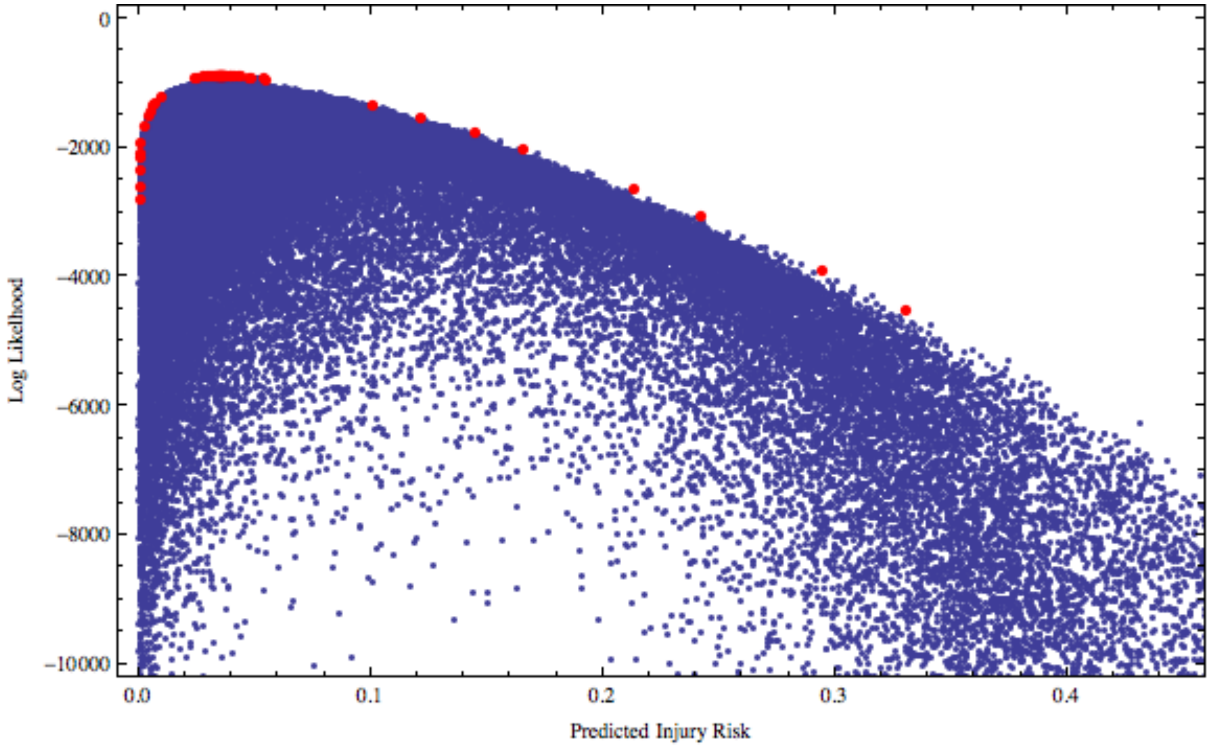

Figure 45. Convex hull point selection example. Blue points represent tested model. Red points are on the convex hull.

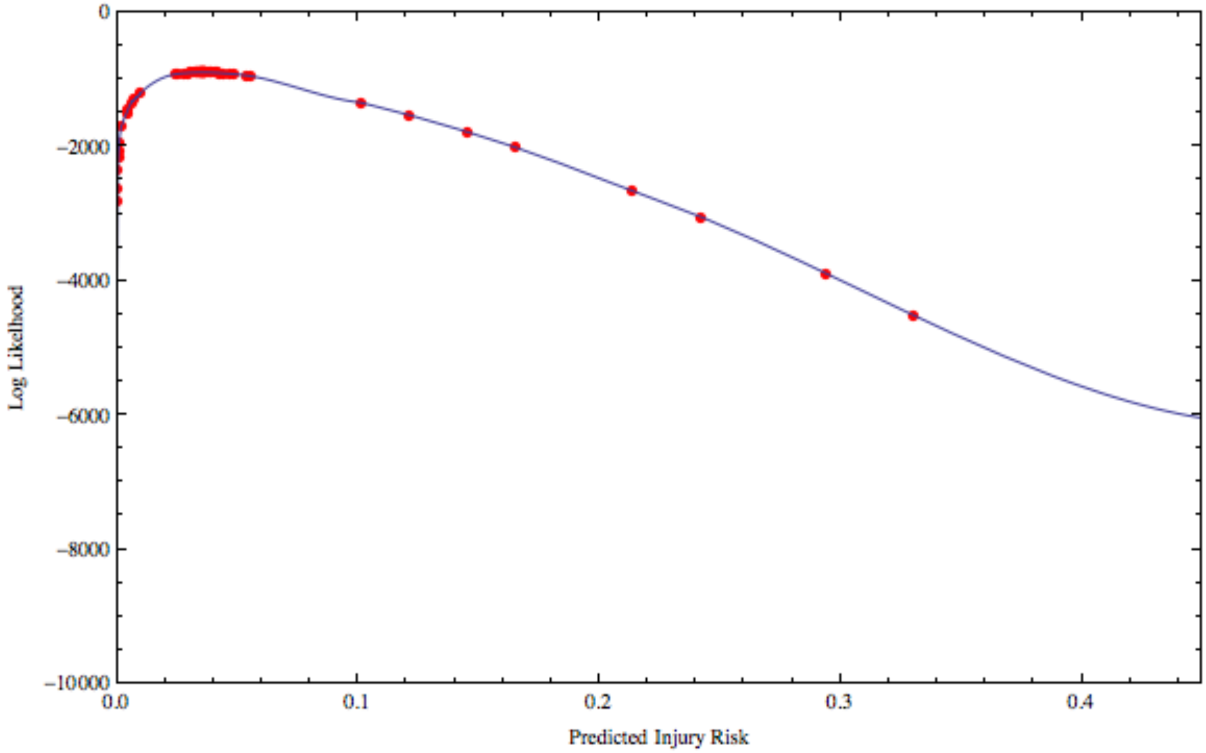

Figure 46. Interpolation between convex hull points.

Finally, using the interpolated contours, we generated the log likelihood of the MLE for risk differences ranging from -0.05 to 0.05 in increments of 0.001. Using Equation E3, we subtracted the log likelihood for the zero-difference MLE from the log-likelihood of the MLE for each hypothesized difference. The resulting estimated log Bayes Factors are shown in Figure 47.

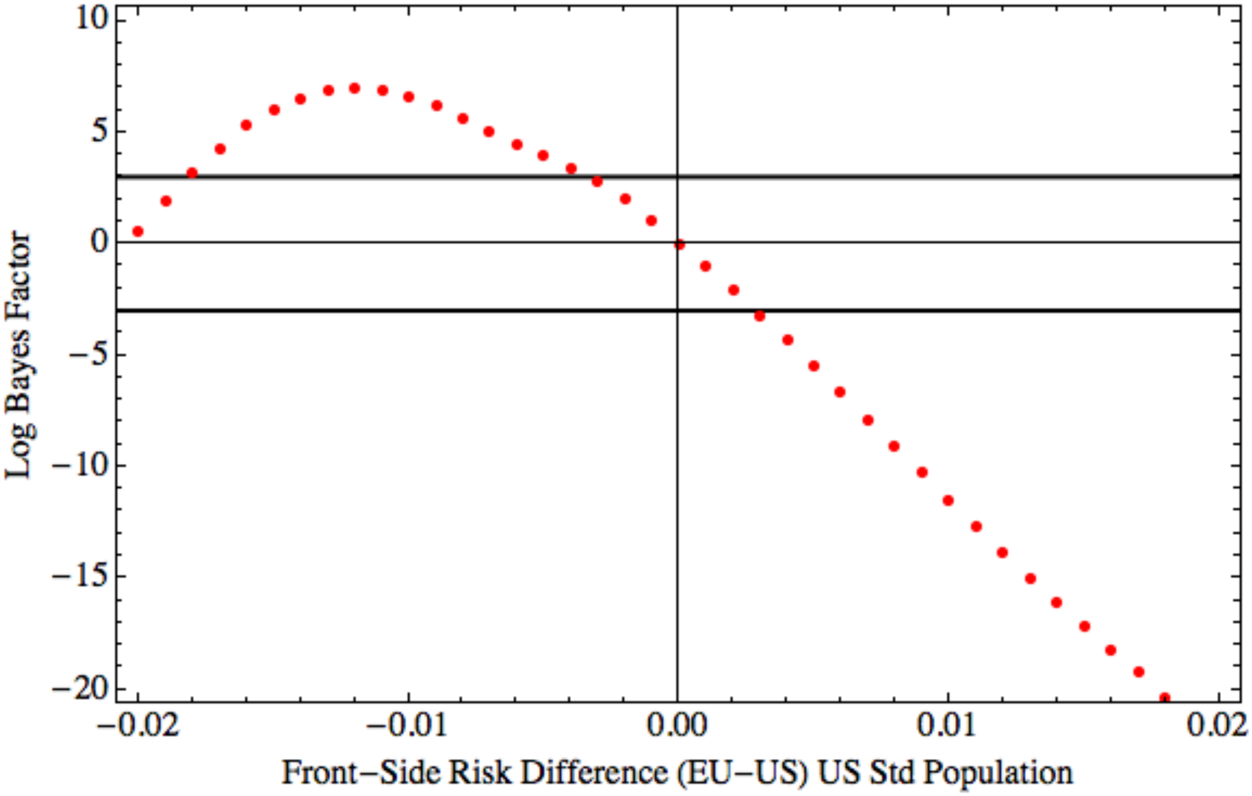

Figure 47. Calculating Bayes Factors relative to the zero difference model.