Response to Reviewers' Comments for TRC_2019_1394

We would like to thank the three anonymous reviewers for their comments, which significantly improved our manuscript. In this document, we detail how we have addressed all the comments made by the reviewers. For your convenience, we first reproduce your comments in a grey box , then highlight our response in **bold**. Changes in the manuscript are highlighted in blue, with the exception of minor edits on typos and language issues.

1 Response to Reviewer 1's Comments

1.1 General Comments for Part 1 and Part 2:

> The paper reads well and this study explores the association between truck crashes and safety-critical events using crash reports and naturalistic driving data. The topic is interesting. However, the reviewer has some concerns.
>
> 1. Justification of the research gap is weak, especially the introduction. Besides, what are the safety-critical events (SCEs)? Could you provide a definition?

Thank you for your comments and suggestions. We have integrated the literature review and some part of the introduction to be a new subsection (Research gaps) in the Introduction. We hope this version of research gaps is better. The definition of safety-critical events (SCEs) has been given in the second paragraph on Page 3:

Such surrogates are usually referred to as safety-critical events (SCEs) and are selected to intuitively represent (more numerous) "near misses", i.e., special types of accident precursors that have all features of accidents, except that potentially catastrophic outcomes were avoided by last-second evasive maneuvers (Dingus et al., 2011; Saleh et al., 2013).

> 2. The introduction is not good and some sentences make them confused. The authors would be suggested to improve it largely.

Thank you for your suggestion. We have completely rewritten the Introduction, and we hope that this version is more clear and convincing. If there are specific topics that are not clear, please let us know and we will strive to improve the text.

3. Authors also kept silent regarding data quality & integrity.

What is the percentage of drivers excluded? How about the accuracy of the GPS data? As we know, sometimes, the coordination of the GPS may be far away from the actual location, even the other side of the road.

How to choose the thresholds? Any references?

The quality of crash data in this manuscript should be reported. MC: We do not have any measure for the quality of the crash data.

Thank you for your comment. We have added quality check statistics of GPS data, as well as data quality & integrity. Here are the relevant sentences:

The included 31,828 commercial truck drivers had a total of 1, 494, 678, 173 pings. Based on the GPS quality indicator, 98.7% of these pings were "good quality", with at least five decimal places and an resulting expected accuracy of 0.1-1.0 meters.

The crashes and SCEs data were not collected by the research team, but by the company. We provided more details about sensor system and the thresholds in the revised manuscript:

- Headway, which signals an instance of tailgating for $\geqslant$ 118 seconds at an unsafe distance of $\leqslant$ 2.8 seconds (Grove et al., 2015).
- Hard brakes, which are defined as instances of deceleration rate $\geqslant$ 9.5 miles per hour per second.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix®, 2007).
- Activation of the forward collision mitigation system.

4. In 4.1, why not report the driving experience (e.g., driving years)? Why not report the statistical summary of the variables?

Thank you for your suggestion. We have added tables of the summary statistics for the predictor variables (Table 2 and Table 4). Company A did not collect the driving years until the recent three years, so the missing rate of driving years is more than 30% and quality of the driving experience is not good. Given that driving years is highly correlated with driver age and driver age was reported in this study, we did not report driving experience in the study.

> 5. Were all variables included in the models? So the authors did not consider the multicollinearity? And the tables made the reviewer confused.

Thank you for your comments. We used different combination of predictors for different models, which are clarified in the newly added Table 5. We have added a correlation matrix plot (Figure 3) and variance inflation test of multicollinearity (Table 9) in the revised manuscript. The results seem to suggest that there is no serious issue of multicollinearity.

> 6. "In the two models using the number of fatalities as the outcome variable (column 4 and 5), all 95% CIs of IRRs included one and the CIs were very wide", it means that the variables are not statistically significant?

Yes. To clarify the credible intervals in Bayesian inference, we added the following statement into the methods section.

A 95% credible interval is the posterior probability that the parameter of interest falls into that range given the data is 95%. If the 95% CI of the IRR includes one, then one is plausible value for the true IRR, so we will consider the parameter of this variable as statistically insignificant. On the other hand, if the 95% CI of the IRR excludes one, we will consider the parameter as statistically significant.

> 7. The analysis on the model results, such as the association between four different types of SCEs and crashes, the relationship between the SCEs (e.g., Headways) of the variables, etc., are simple and weak. The authors would be suggested to add the deeper analysis. Otherwise, the contributions of this manuscript would be limited.

Thank you for your comment. We have reorganized the results section and regression tables according to our three research questions. The regression results include different models for three different outcome variables (crashes, injuries, and fatalities), different predictor combinations, and subgroup analyses stratified by different business units and driver types. We hope that this version of the results is clearer. If there is specific additional analyses that could further improve our results, please feel free to let us know and we will add them into our results.

8. The discussion and conclusions would be also suggested to be improved substantially.

Thank you for your suggestion. We have combined and integrated our discussion and conclusions into one section. We tried to clarify and highlight our contributions, extend our results to the trucking and general commuting population, and point out the limitations and future research direction. We hope this revised discussion and conclusion Please let us know if there are certain aspects we could do better.

9. Many sentences are weak / improper and make readers confused.

For instance, page 11: One unit increase in the number of any type of SCEs per 10,000 miles was associated with 8.4% (95% CI: 8-8.8%) increase in the number of crashes per mile. How do you know 8.4%?

Thank you for your comment. The description of our findings in the original manuscript was written in a style that is more familiar in the field of epidemiology studies. A bit of math could reach this 8.4%. Since the number of SCEs per mile $\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$**, the rate change of** μ **when the predictor** x_1 **is increased by one unit can be calculated as:**

$$\frac{\mu'}{\mu} = \frac{\exp[\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_K x_K]}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K} \quad (1)$$

$$= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K) \times \exp(\beta_1)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)} \quad (2)$$

$$= \exp(\beta_1) \rightarrow \text{which is the incidence rate ratio (IRR).} \quad (3)$$

Therefore, if the IRR is 1.084, the number of SCEs per mile μ **is increased by 8.4% when the predictor** x_1 **is increased by one unit. To clarify the results and avoid the confusion in different tables, we reorganized the regression tables according to our three research questions and rewrote the results section.**

10. The website of the data sets provided in the manuscript does not work.

Thank you for your information. Here are the three links we provided:

- website: https://caimiao0714.github.io/Github-SCE-crash/
- data sets: https://github.com/caimiao0714/Github-SCE-crash/blob/master/data/sample_ping.csv

We have tested it on different computers and browsers and they work out well. Please check if there is any internet connection issues and try different web browsers and see if they work.

> 11. Lastly, the review would suggest that the authors re-check their grammar and text as there are many spelling mistakes in the manuscript. For instance, line 2 in page 5: "The found that ......". This manuscript should be revised properly.

Thank you for your suggestion. We have corrected this spelling mistake, and we also conducted spelling and grammar check in this revised manuscript. Please let us know if there is any obvious spelling mistakes in this revised version.

2 Response to Reviewer 2's Comments

> Overall, this is a high-quality paper and is a worthy addition to literatures with some improvement. The study using emerging telematics data and the combination The paper used a large dataset to evaluate the relationship between SCE and crash risk. The topic is certainly worth investigation given the importance of the validity of crash surrogates. Some specific comments are below:

> 1. Please refrain from using acronyms and abbreviations. There are too many of them and make the paper difficult to follow.

Thank you for your suggestion. We have reduce the frequency of acronyms in our manuscript, and only commonly used acronyms are kept in this revised manuscript.

> 2. Please provide more details on the information about the "ping". For example,
>
> - Is a ping a single data point or several points?
> - The paper mentioned "ping data every couple of seconds to around 5 minutes." For a particular truck/driver, is the interval fixed or varying?
> - If a vehicle was not on, will a ping still be sent? Figure 1 seems implying so.
> - One important issue is that if a SCE occurred not during at ping period, will it still show in the dataset? Some trigger based system will catch all such events, for example, when the acceleration is above 0.5G, a record will be automatically

generated. It is not clear from the description whether this is the case for the data used.

Thank you for your suggestion. Below are our answers to your questions:

- **One ping is a one data point including latitude, longitude, date, time, real-time speed, and others.**
- **For a particular truck/driver, the interval is still varying.**
- **Yes, the pings are still recorded when a vehicle was not on.**
- **The SCEs were collected in a system independently from the ping collecting system. These SCEs were collected whenever the kinematic thresholds were triggered by the driver. Therefore, SCE will still be recorded even when there was no ping at that time. The sensor system and kinematic thresholds is introduced in the revised Introduction.**

3. I have major concerns regarding the hard break, which is defined based on the speed decrease: "the speed decrease within a unit time is larger than a preset threshold value." How long is a unit time? I can only specular it is based on the time interval, couple of seconds to around 5 minutes. In reality a hard break typically only takes less than a second. Speed change over more than a few seconds most likely only represents whether the truck got to a stop instead of a hard brake. Can authors clarify how the Hard Brake and headway SCEs were defined. For example, acceleration threshold?

Thank you for your comment. We provided the sensor system used by the company and some of the thresholds in the revised introduction:

The sensor-based monitoring system used by the company captures four different kinematic events:

- Headway, which signals an instance of tailgating for $\geqslant$ 118 seconds at an unsafe distance of $\leqslant$ 2.8 seconds (Grove et al., 2015).
- Hard brakes, which are defined as instances of deceleration rate $\geqslant$ 9.5 miles per hour per second.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix®, 2007).
- Activation of the forward collision mitigation system.

4. For the regression model, I suspect the four SCEs could be correlated. Can you check whether multicolineality is an issue? The results on page 15" "Four SCEs" model was not significantly better than the "Pooled" model." Could be a result of this?

Thank you for your suggestion. We have added a correlation matrix plot for the four different types of SCEs. We also conducted a variance inflation factor test for potential multi-collinearity. The Pearson correlation coefficients were all less than 0.2 for the four SCEs (Figure 3), and the VIF scores were all less than 1.3 (Table 9). The two tests suggest there were minor positive correlation among the rates of four types of SCEs, but did not cause significant multicollinearity issues.

5. Please provide appropriate summary statistics for the four SCEs, e.g., mean, standard deviation, and correlation among them. Hard Brake should happen at much higher frequency than other types of SCEs, which could dominate the "All SCEs" variable.

Thank you for your suggestion. The added correlation matrix indicates that there is only minor correlation between different types of SCEs (Pearson correlation coefficient less than 0.2). In our study, the predictor variables are not the number of SCEs, but the rate of SCEs (the number of SCEs per 10,000 miles). This is because the number of events highly depends on miles driven, while the rate of events is less dependent on miles driven. The newly added Table 4 shows that the number of hard brakes, headways, collision mitigation, and rolling stability per 10,000 miles were 6.86, 5.35, 1.74, 0.21. The SCEs were dominated by hard brakes and headways, not by hard brakes alone.

6. Page 2: "Since NDS data are typically collected every 30 seconds to 5 minutes, the amount of NDS data are generally very large, which provides both an opportunity and a challenge for data analytics. " This statement is not accurate. Typical NDS studies used continuous data collection method, e.g., SHRP2 NDS and 100-Car NDS collect data from ignition-on to ignition-off at 10HZ for video and acceleration data.

Thank you for your correction. We have completely rewritten the Introduction and the inaccurate statement has been revised.

7. Figure 3 and 4 are difficult to understand. Why only check the percentage of zeros? Careful exam of Figure 4 show that "the observed proportion of zero crashes located almost exactly at the center of the simulated distributions" is not true: the observed is still far from center for "DED LOC" and "DEC OTR". Overall, I don't think Figure 3 and 4 provide indispensable information on the model performance. The benefit seems to be overwhelmed by the difficulty in understanding them. Unless the authors can substantially improve its clarity.

Thank you for your suggestion. Since we are concerned how likely our model predicts crashes when the driver has SCEs, comparing the the proportion of zeros in simulated and observed data is informative in prediction accuracy. Figure 3 and 4 show that the simulated distributions (light blue histograms) are generally on the right of the observed zero proportion (solid black vertical line), suggesting that our Bayesian negative binomial models tend to over-predict non-crashes but under-predict the crashes, and the magnitude of the prediction bias is small. We have revised the paragraph describing the graphical presentation of the posterior predictive checks in subsection 4.3 and refrained from claiming "the observed proportion of zero crashes located almost exactly at the center of the simulated distributions" in Figure 4. Please see if the revised edition is clearer:

To investigate the models' predictive accuracy, we adopt the approach of Gelman et al. (2013, Section 6.3) who suggested simulating some function of the data and parameter and comparing it with the observed value of a particular quantity. For our trucking safety application, we examined the proportion of zero crashes since it corresponds to a crash-free trip, which is of interest to truck drivers and operators alike. The probability of having zero crashes is, of course, an unknown quantity, but its posterior distribution can be estimated by simulating samples using Hamiltonian Monte Carlo. In this section, we limit our analysis to the models whose outcomes were crashes since the accident and fatality models indicated that our observed events were insufficient for statistical inference (based on the size of the credible interval in Table 6).

Figure 4 shows the posterior distributions, which are indicated by the histograms in light blue, for the posterior probability of zero crashes under each of the six models considered in Table 7. The observed proportion of zero crashes is indicated by the vertical line in each part of Figure 4. For all six models, the observed proportion of zero crashes was considerably less than what would be predicted by the model. Note that the magnitude of this prediction bias is small, usually around 0.015. The deviations from the simulated posterior distributions suggests that some aspect of the model may involve unreasonable assumptions.

Based partly on the result from Figure 4, we ran the model with all four SCEs (model 2) separately for each of the seven business units and driver types. The corresponding posterior predictive check for zero crashes is shown in Figure 5. Here, the vertical lines are much closer to the simulated posterior distribution. This suggests that different business units and driver types should be accounted for in the model.

Minor issues:

1. Equation (1), should subscript "i" be in the right of equation for "y"as well?
2. Page 3: "The large sample size can yield statistically significant results and conclusions." Large sample is a major strength of this paper. However, large sample size does not necessarily yield statistically significant results (if there is no relationship, we probably don't want statistically significant results). More accurate statement is needed, for example: The large sample size provides high statistical power to detect potential relationship between SCEs and crashes.
3. Abstract: second line: should it be "used to measure safety" since "outcome" is a little vague.
4. Figure 3: please label the X-axis. What is PPC in the caption of the figure?

Thank you for your suggestions. We have addressed the issues as suggested.

1. **A subscript "i" has been added to the ys in the right side of Equation (1).**
2. **The sentence has been changed into: "The large sample size provides a large statistical power to detect potential relationship between SCEs and crashes."**
3. **The corresponding sentence has been changed into "In NDSs, safety-critical events (SCEs) are commonly used to measure safety since crashes are very rare.".**
4. **PPC is graphical posterior predictive checks, which has been spelled out in the caption.**

3 Response to Reviewer 3's Comments

This paper investigated the association of the surrogate safety metrics and crashes using the NDS data collected from instrumented trucks. Below comments can further improve the quality of the manuscript.

Please contact the journal in case of clarifying the data source and data reliability step.

Thank you for your suggestion. We have clarified the data sources and quality of the collected data in the revised Data section. A fake sample of our ping data has also been provided in our supplementary materials section.

While authors reviewed some NDS studies, the biggest NDS study in the US and Europe i.e. "SHRP2" and UDRIVE are missing from the review. Below papers can provide more info about the latest and the largest scale NDS studies with more than 3200 drivers. Please update the Table 1, accordingly.

- Complementary Methodologies to Identify Weather Conditions in Naturalistic Driving Study Trips: Lessons Learned from the SHRP2 Naturalistic Driving Study & Roadway Information Database
- Eenink, R., Barnard, Y., Baumann, M., Augros, X., & Utesch, F. (2014). UDRIVE: the European naturalistic driving study. In Proceedings of Transport Research Arena. IFSTTAR.
- The impacts of heavy rain on speed and headway behaviors: an investigation using the SHRP2 naturalistic driving study data (TRC).
- The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters

Thank you for your suggestion. We have read through the provided three references. These papers did use large NDS data sets, but they did not investigated the association between crashes and SCEs as the papers listed in Table 1. Therefore, we did not add them in Table 1. Here are a brief summary of the recommended papers:

1. Ghasemzadeh et al. (2019) used the SHRP2 NDS database with more than 3,500 drivers and described three methods to merge NDS data sets with weather information: wiper status, National Climate Datta Center, and weather-related crashes. They also provided data reduction and processing procedures for the SHRP2 NDS data. The paper did not investigate the relationship between crashes and SCEs.
2. Eenink et al. (2014) and Barnard et al. (2016) introduces the overarching goal, methodology, research questions, data to be collected, and expected outcomes of the European naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environment (UDRIVE) project. This paper is more like a research proposal than an evaluation of the data, and the association between crashes and SCEs is a proposed research question but unanswered.

3. Ahmed and Ghasemzadeh (2018) is a data analytics paper that quantifies the association between weather conditions and driver speed and headway selection behaviors using the SHRP2 NDS database. They did not investigate the association between driver speed, headway selection, and real crashes.

Please introduce the data ping. What frequency of data a data ping is representing?

Thank you for your suggestion. We have updated our description of the data ping, including the frequency of the data pings in the revised manuscript:

Our study is based on data captured from April 1, 2015 to March 31, 2016 by the company's sensor-based monitoring system on their entire fleet. Our dataset includes intermittently collected real-time driving *ping* records, which ranges between every couple of seconds to approximately 15 minutes. Over 50% of the time intervals between two pings were less than 5 minutes and over 95% of them were less than 15 minutes. The time intervals varied among drivers, places, and trips, and there were no clear patterns explaining the variations in interval lengths. Each ping is a data point that includes the exact date and time of the record (year, month, day, hour, minute, and second), GPS (latitudes and longitudes with five decimal place recordings), GPS quality, speed, and drivers' anonymized unique ID.

In the meanwhile, a fake sample of ping data (in Comma Separated Values format) is provided in our supplementary materials. The readers are encouraged to check out the ping data if they feel unclear/confused by our description.

- Please explain the collision mitigation surrogate in Table 1.
- The authors should explain the method that they calculated headway. Were the vehicles instrumented with radar?
- More explanation of the hard brakes and the threshold that was used should be added. What was the threshold for the 231101 hard brakes? Clarify whether this number represents the events or data pings.
- Rolling stability should be defined.
- Description of the headway calculation and the threshold for critical headway SCE should be added.
- What are the present thresholds on page 8?

Thank you for your suggestion. We did not collect/define these SCEs. They were collected by our partner company as a part of their own routine monitoring program using the Bendix® Wingman® Advanced™ monitoring system. The definitions of the four SCEs are defined in the revised Introduction:

- Headway, which signals an instance of tailgating for $\geqslant$ 118 seconds at an unsafe distance of $\leqslant$ 2.8 seconds (Grove et al., 2015).
- Hard brakes, which are defined as instances of deceleration rate $\geqslant$ 9.5 miles per hour per second.
- Activation of the rolling stability system, which intervenes by applying brake pressure (in addition to potentially applying trailer pressure) assisting the driver in aligning the vehicle when the system's critical thresholds are approached (Bendix®, 2007).
- Activation of the forward collision mitigation system.

> Regarding rolling stability, any NDS study that consider adverse weather and driver performance?

Thank you for your comment. To date, we have not found any research articles that study the the association between rolling stability and crashes controlling for adverse weather and driver performance.

> It would be interesting to see challenges with the data, missing values, etc. to be explained in a paragraph as a data preparation stage.

Thank you for your suggestion. We have added more contents on missing values and data quality control in the first paragraph of Data Description subsection:

The original dataset provided by the company included 34,348 drivers. We have excluded 2,520 drivers (i.e., 7.4% of the original dataset) from our analysis based on the following criteria: (a) driver inactivity, where we required the driver to have at least 100 GPS pings in the data to be included; (b) the unique id for the driver is not found in the provided demographics table; and/or (c) the number of SCE reported were identified as obvious outliers (we only removed drivers who had an unrealistically large number of SCEs). Hereafter, all reported data will correspond to only those generated by the remaining 31,828 drivers, whose characteristics are summarized in Table 2.

> Page 9 the authors mentioned the median distance of the trip and the median number of miles per trip as 2.61 and 77.06. did the authors only considered trucks in the urban environment?

Thank you for your comment. As shown in Figure 2, this data set covers national general truck transporting environment in the United States. Although most of the ping data were in urban areas, rural areas were also covered.

> Xk should be xik

Thank you for your suggestion. We have corrected that subscript issue.

> Authors need to explain how did they come up with K values.

Thank you for your comment. The Pareto k values were not proposed by the authors, but by (Vehtari et al., 2015, 2017). We have added citations in the Methods and Results section.

> It is recommended that page 10 paragraph 1 be summarized in a table and provide stat for each category.

Thank you for your suggestion. A summary table has been added for paragraph 1 on page 10.

> Page 11 talked about table 2 and table two is presented in page 14. please keep the tables close to the description, if possible.

Thank you for your suggestion. We have reformatted the manuscript accordingly.

References

Ahmed, M. M. and Ghasemzadeh, A. (2018). The impacts of heavy rain on speed and headway behaviors: an investigation using the shrp2 naturalistic driving study data. *Transportation research part C: emerging technologies*, 91:371–384.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R., and Baumann, M. (2016). The study design of udrive: the naturalistic driving study across europe for cars, trucks and scooters. *European Transport Research Review*, 8(2):14.

Bendix® (2007). Bendix® ABS-6 Advanced with ESP® Stability System - frequently asked questions to help you make an intelligent investment in stability. Bendix Commercial Vehicle Systems LLC, a member of the Knorr-Bremse Group. https://www.bendix.com/media/documents/products_1/absstability/truckstractors/StabilityFAQ.pdf. [Published March 2007; accessed April 19, 2020].

Dingus, T. A., Hanowski, R. J., and Klauer, S. G. (2011). Estimating crash risk. *Ergonomics in Design*, 19(4):8–12.

Eenink, R., Barnard, Y., Baumann, M., Augros, X., and Utesch, F. (2014). UDRIVE: the European naturalistic driving study. In *Proceedings of Transport Research Arena*. TRA 2014, 14-17 Apr 2014, Paris, France. IFSTTAR.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Ghasemzadeh, A., Hammit, B. E., Ahmed, M. M., and Eldeeb, H. (2019). Complementary methodologies to identify weather conditions in naturalistic driving study trips: Lessons learned from the shrp2 naturalistic driving study & roadway information database. *Safety Science*, 119:21–28.

Grove, K., Atwood, J., Hill, P., Fitch, G., DiFonzo, A., Marchese, M., and Blanco, M. (2015). Commercial motor vehicle driver performance with adaptive cruise control in adverse weather. *Procedia Manufacturing*, 3:2777–2783.

Saleh, J. H., Saltmarsh, E. A., Favaro, F. M., and Brevault, L. (2013). Accident precursors, near misses, and warning signs: critical review and formal definitions within the framework of discrete event systems. *Reliability Engineering & System Safety*, 114:148–154.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2015). Pareto smoothed importance sampling.