IDENTIFYING LINKAGES BETWEEN DEMOGRAPHICS, BEHAVIORS, AND ROAD ACCIDENT FREQUENCY: A MACHINE LEARNING APPROACH IN ENGLAND (1108)

Dakota Aaron McCarty[1]
[1] Department of Urban Policy & Administration, Incheon National University, Yeonsu-gu, Incheon, Korea

Abstract. This study addresses the challenge of collecting comprehensive data on individual drivers' behavior, which has been found to contribute to over 70% of road accidents. It focuses on demographic factors and their indirect impact on accident rates by leveraging existing literature on the connection between demographics and risky driving behaviors. A review of literature identifies demographic characteristics correlated with risky driving behaviors. Using regression-based machine learning models, the investigation covers all of England, UK, aiming to establish connections between driving behaviors, demographics, and accident frequency. The study's results align with previous findings and provide a valuable methodology to investigate behavior-accident links on a broader scale despite data limitations. The findings support utilizing readily available demographic information to estimate road accident rates.

Keywords: Driving behavior; Traffic accidents; Cohort analysis; SHAP values; Extra Trees Regressor.

1. Introduction

The study of human behavior presents inherent complexities within the field of behavioral sciences, primarily due to the challenges associated with measuring behavior across diverse populations and timeframes. Comprehensive research in this area necessitates a thorough understanding of various influences and relationships, aiming to control for a wide range of external factors (Kelly and Barker, 2016). While extensive research exists on other behavioral issues, such as substance abuse and eating habits, there has been relatively less attention given to risky driving behavior. This is a notable issue as driving behavior is dissimilar from many other potentially risky behaviors. For instance, modifying behavior related to ‘rare events’ like road accidents poses challenges since individuals engaging in such behavior often do not experience immediate negative consequences in their daily lives. Additionally, habits—both positive and negative—such as exercising or smoking, gradually develop and become ingrained in daily routines, whereas the need to change driving behavior is typically prompted by accidents or reflexive responses (Kelly and Barker, 2016).

When studying the general population, further challenges arise when attempting to control for factors like exposure and experience (Brown, 1982). These complications are particularly relevant when attempting to investigate the impacts of risky driving behaviors on road safety. To overcome some of these limitations, researchers often narrow the scope of their investigations by focusing on specific demographics, geographic areas, or other controllable factors that facilitate data collection (e.g., through surveys or in-car monitoring devices). However, this narrowing of focus inadvertently reduces the generalisability of the research to new areas and populations. Consequently, there is a need to explore new variables that can serve as proxies to estimate driver behavior and the risk of road accidents, as highlighted in previous studies. Thus, the primary objective of this research is to identify potential proxies for estimating risky behaviors that can be utilised in future road accident prediction models.

Numerous studies have established the significant role of human behavior in causing road traffic accidents (Andrew Yockey and Barroso, 2023; Bener et al., 2009). Moreover, an expanding body of research supports the effectiveness of behavioral science approaches in identifying and mitigating accident-prone behaviors (Gielen and Sleet, 2003). This study aims to explore readily obtainable features, mitigating a limitation in the field around data collection and availability, which have been found to have connection with risky driving behaviors. The study builds upon existing literature that has established connections between demographics and propensity to engage in risky driving behaviors and further explores how the related features may influence road accident frequency. Through the use of regression-based machine learning modeling methods, this study seeks to develop potential proxies that can approximate risky behavior, utilising a comprehensive literature review as the foundation for our analysis.

2. Human Behavior, Demographics, and Road Accidents

The literature presents consistent findings indicating that human behavior is significantly responsible for road accidents, attributing over 70 percent of such incidents to human factors (Abdullah and Sipos, 2022; Chand et al., 2021). In order to achieve a substantial reduction in road accident rates, it is critical to understand and investigate human behavior in specific areas. This approach could lead to more accurate predictions and risk assessments, hence forming the primary goal of this research.

The web of human behaviors that notably impact road accident risks is intricate. Each factor, though discussed individually, is part of a complex interrelated system with shared causes, highlighting the complexity of human behavior (Brown, 1982; Kelly and Barker, 2016).

Vehicle speed, influenced by a mixture of human and environmental factors, plays a

critical role in accidents. With human factors being a combination of driver, passenger, and also pedestrian condition, age, experience and other defining factors; and environmental factors encompassing the actual environment that the humans are acting in, made up of features such as road quality, urban form, traffic flow, weather, time, and so forth (Khayesi et al., 2005; Yusuf et al., 2016). Furthermore, socioeconomic status has been identified as an additional, and notable, influencing human factor having an influence on risky behaviors such as drunk driving (Li et al., 2013).

Fatigue is another prominent risk factor in road accidents. Certain demographics, including younger drivers, shift workers, individuals with sleep disorders, and those operating vehicles during the early morning hours, are particularly (Bharadwaj et al., 2021). Similarly, distractions, which can arise from various factors such as cell phone use, loud music, or roadside advertisements, have been recognised to increase road accident risk (Papantoniou et al., 2017; Zokaei et al., 2020).

Reckless driving, often associated with alcohol consumption, exerts a substantial influence on accident risk (Borgialli et al., 2000). Emotional states, particularly depression and anxiety, have also been found to increase the risk of road accidents (Alavi et al., 2017).

Age and gender also play significant roles in accident risk. Younger drivers, particularly men, are more prone to engage in risky driving behaviors, while older drivers generally exhibit safer habits (Bener et al., 2009; Rhodes and Pivik, 2011; Russo et al., 2014). However, older individuals, especially pedestrians, face a higher risk of injury. Women, regardless of road conditions, are less likely to be involved in accidents or engage in dangerous driving behaviors (Morgan and Mannering, 2011; Russo et al., 2014).

One previous study investigated the relationship between the level of deprivation in residential areas and the rate of road accidents, finding that injury rates in the poorest areas were double compared to the wealthiest areas, particularly with pedestrians (Abdalla et al., 1997). Another comprehensive study examined the number of pedestrian injuries in English wards and revealed that road accidents involving child pedestrians were four times more frequent in the poorest wards than in wealthier wards, with adult pedestrian deaths nearly twice as high in impoverished areas (Graham et al., 2005).

Despite the second study utilising improved methods, both concluded the same findings of there being a higher risk of casualties related to accidents in areas with lower socioeconomics statuses. In summary, understanding human behavior and demographic factors, as well as their interconnected nature, is crucial in developing effective interventions and strategies to reduce road accidents. Future research is needed to continue refining our models and approaches, ensuring that we can predict and mitigate risks efficiently.

3. Behavioral Factors Influencing Road Accidents in the United Kingdom

This study examines the 2019 data from the UK Department of Transportation, representing the most recent year with complete and normalised data prior to the COVID-19 pandemic. The investigation is focused on understanding the behavioral elements contributing to road accidents in the United Kingdom.

Demographic data suggests a significant gender disparity in road accidents across the United Kingdom, with male drivers implicated in nearly twice as many accidents as their female counterparts (Figure 1). Specifically, male drivers accounted for 283,291 accidents, compared to 141,699 for female drivers. Age was also shown to play s a significant role in road accident frequency. Through an analysis of accident frequency by age in the study area, it was found that those aged 20 to mid-30s were most likely to be involved in an accident, while also finding that the elderly and teenagers were less likely to be involved in an accident (Figure 1).

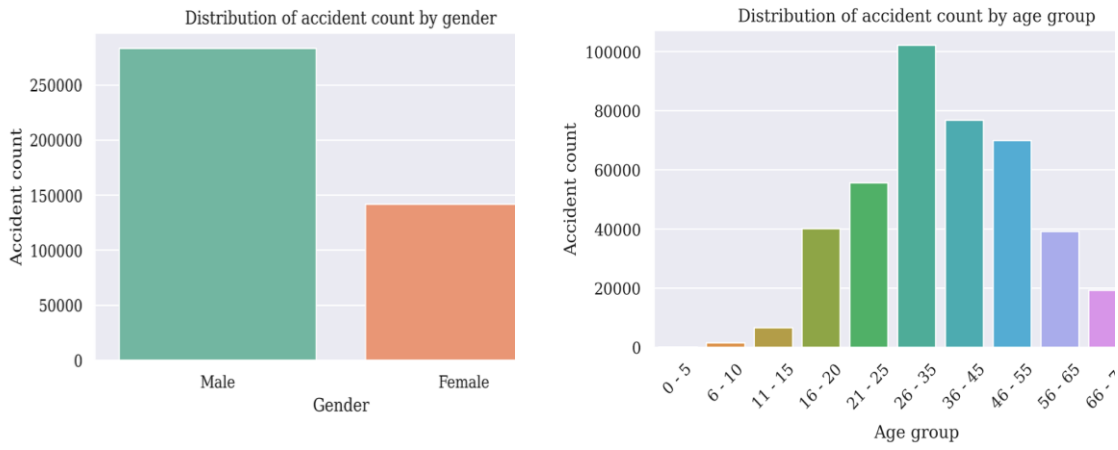

Figure 1. Gender and age distribution of accident counts

When visualizing the data by gender and age cohort together, the trend is still present showing that in each of the age groups that male drivers have a higher count of accidents. With those aged 26 to 35 being the group with the highest number for both genders (Figure 2).

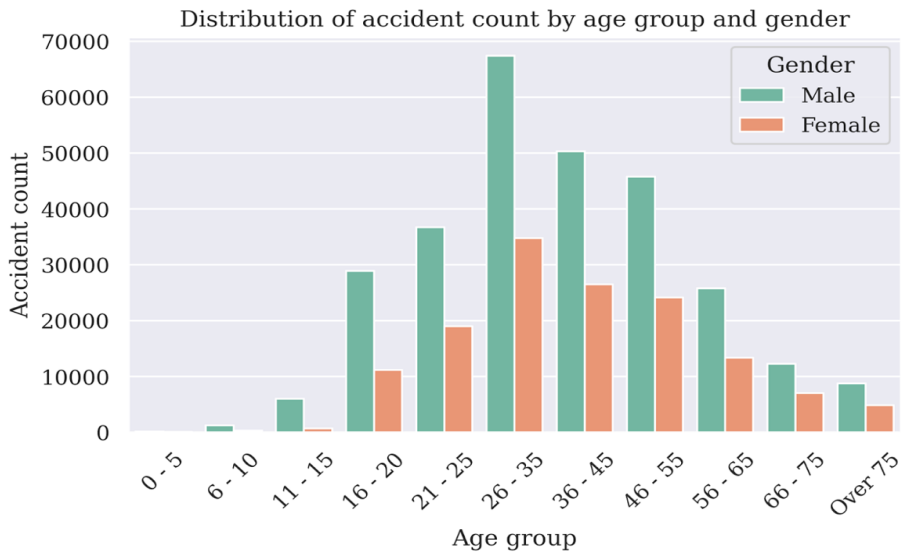

Figure 2. Distribution of accident counts by age and gender splits

While basic, these trends do align with previous literature on the correlations between gender, age, and road accident counts. This initial analysis makes the basis for the remainder of this paper, which aims to explore this relationship more in-depth.

The Annual Report on Road Casualties provides valuable insights into the primary causes of road accidents in England, with behavioral elements such as driver error, impairment, injudicious actions, reaction, distraction, and inexperience identified as the most prevalent contributors. Among these, driver behavior stands out as a leading cause, underscoring the importance of effectively estimating this feature in road accident prediction models.

Speeding, for instance, is a significant contributing factor to road accidents in the United Kingdom. It accounted for 7 to 18 percent of road accident casualties in 2019 across various subcategories such as exceeding speed limits, traveling too fast for conditions, and driver carelessness or recklessness. Similarly, fatigue driving directly caused only 2 percent of all accidents, while those involving drivers under the influence of alcohol or drugs made up 6 percent and 2 percent of all accidents, respectively.

Despite accounting for a relatively small percentage of accidents in the dataset, distracted driving remains a concern, including distractions from cell phone use, within the vehicle, and outside the vehicle. The lack of policy enforcement and detection could underlie these statistics.

Inexperience stands out as a significant causal factor, implicated in 5 percent of all road accidents. Meanwhile, reckless driving, defined by driver carelessness, recklessness, or being in a hurry, contributed to 18 percent of all casualties. Related behaviors, such as following too closely, sudden braking, and swerving, also pose substantial risks.

Emotions, though challenging to quantify due to their subjective nature, also play a role in road accidents. Aggressive driving, often linked to emotional stimulation, accounted for 4 to 8 percent of all road accidents in the study area.

The Annual Report on Road Casualties paints a complex picture of the contributing factors to road accidents, with behavioral aspects such as driver error, impairment, and inexperience playing significant roles. Speeding, reckless driving, and distractions emerge as particularly potent threats to road safety. It's essential to continue efforts to raise awareness of these issues, enhance law enforcement and develop policies that address these risk factors more effectively. The data also highlight the need for predictive models that can better consider these variables to effectively mitigate the occurrence of accidents. While certain aspects, such as emotional factors, remain challenging to measure and integrate into these models, their acknowledged influence on road safety necessitates further research and methodological innovation. With targeted intervention strategies that are informed by comprehensive and nuanced understandings of these contributors, there's potential to significantly reduce road casualties in the future.

4. Materials and Methods

At the time of this study, the Department for Transport does not provide non-aggregate data on road accidents and their respective causes, which poses a challenge for quantitative analyses. Therefore, there is a heavy reliance by this study on existing literature, taking a more inferential approach to examining the relationships between risky driving behavior and demographic factors and their subsequent relationship with road accident frequency.

The study utilises a dataset created by combining accident frequencies and gender-age cohorts. Tobler's pycnophylactic interpolation method was employed to aggregated the data to an hexagonal grid, with each cell covering an area of approximately 730 square meters. This cell resolution is approximately the of the urban Lower Layer Super Output Areas (LSOAs) in the United Kingdom. The dataset includes 34,554 hexagons in total being areas with a minimum of 5 accidents per hexagon cell. To analyse the spatial correlations between the risk of road accidents and demographic characteristics of an area, the accident counts per hexagon were divided by the total population of the area to calculate the accident rate per capita, which also makes the dependent variable in the machine learning regression analysis.

The dataset includes various features such as accident per capita, population share across different age cohorts, and the ratio of men to women within each age cohort. The population features were normalised as proportions and the gender-age cohort data

was calculated as ratios of men to women. This study has elected to use rates (accident per capita) instead of frequencies helps account for variations in population size.

To better analyse the relationship between accidents per capita and demographic characteristics, the dataset was used to build multiple regression-based models. PyCaret, an open-source AutoML tool, was used to generate several models for further analysis (Table 1). The Extra Trees Regressor model performed the best, with an R-squared (R2) value of 0.2695, a Mean Absolute Error (MAE) of 0.0024, and a Root Mean Square Error (RMSE) of 0.0039. For both the MAE and RMSE, all models performed relatively the same. However, for R2, the top performing model was over double that of the baseline (Linear Regression) model. Considering the limited scope of data, the model was fed for these predictions, the results can be considered meaningful. With only demographic features of an area, the model was able to account for nearly 27% of the variance in road accident frequency.

Table 1. Modeling results

MODEL	MAE	RMSE	R2
EXTRA TREES REGRESSOR	0.0024	0.0039	0.2695
LIGHT GRADIENT BOOSTING MACHINE	0.0024	0.0040	0.2391
RANDOM FOREST REGRESSOR	0.0025	0.0040	0.2134
GRADIENT BOOSTING REGRESSOR	0.0024	0.0041	0.1986
K NEIGHBORS REGRESSOR	0.0025	0.0042	0.1699
BAYESIAN RIDGE	0.0026	0.0042	0.1365
RIDGE REGRESSION	0.0026	0.0042	0.1329
LINEAR REGRESSION	0.0026	0.0042	0.1326

5. Results and Discussion

This research has followed a conceptual model of an in-depth literature review to identify risky driving behaviors along with their links to demographic characteristics; an analysis of accident rates per capita across gender and age cohorts through the development of machine learning models; and the generation of feature importance and SHapley Additive ExPlanation (SHAP) plots for visual examination of the results and comparison to the previous literature.

In the context of an Extra Trees Regressor model (a type of decision tree ensemble algorithm), a feature importance plot (Figure 3) illustrates the contribution of each

feature to the predictive power of the model. Specifically, in an Extra Trees Regressor, feature importance is typically computed from the average reduction in impurity that results from splits on that particular feature, averaged over all trees in the ensemble.

If a feature has a high value in the feature importance plot, that suggests the feature has a significant role in the model's prediction. Conversely, a lower value suggests that the feature has a lesser impact on the model's decision making.

In Figure 3, we can see that the top four most important features were related to age, being, in order of importance, the total population aged 20 to 34, aged 35 to 49, under 10 years of age, and over 65 years of age. This shows that those aged 20 to 34 and aged 35 to 49 have a higher impact on predicting road accident frequency with those aged 20 to 34 having a slightly higher importance in the model.

The next three most importance features are related to the male to female ration at different age groups. With areas having more males than females aged 20 to 34 having the highest importance among these.

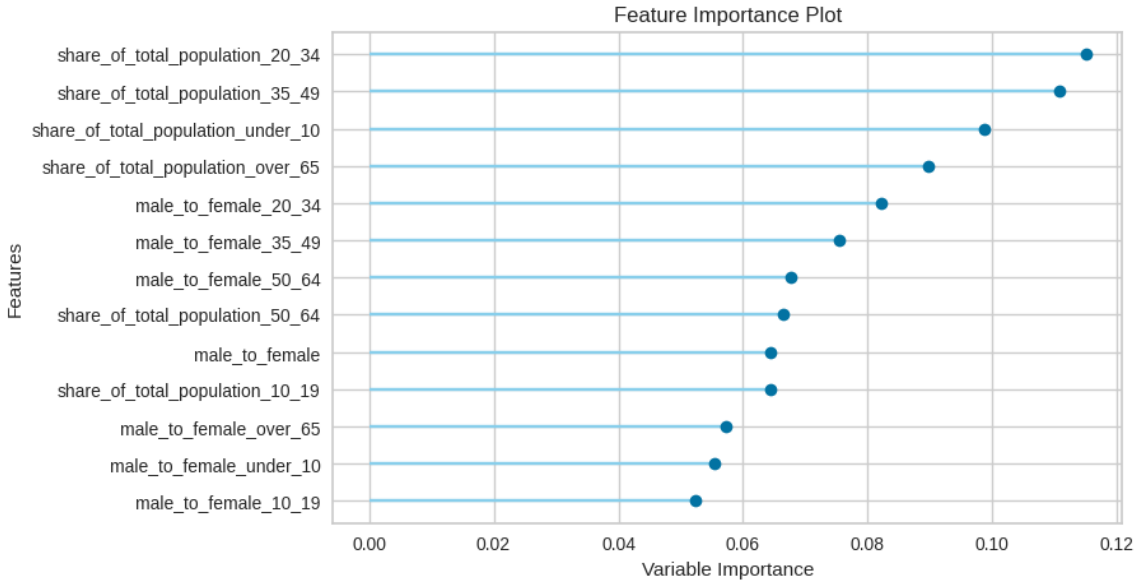

Figure 3. Feature Importance Plot

It's worth noting that while feature importance can give a good sense of which features are most predictive, it doesn't provide insight on how changes in those features impact the target variable - for that, other methods such as Partial Dependence Plots or SHAP (SHapley Additive exPlanations) values can be useful.

For this study, the SHAP values are utilised (Figure 4) to provide a more precise interpretation of the results. The SHAP values explain the predictions of the machine learning model by assessing the impact of each feature. The plot shows the features arranged in descending order of importance, with each point representing a sample

point of the data. The colors in the plot indicate the value of the point, with red indicating a higher value and blue indicating a lower value.

The SHAP value plot generated by the Extra Trees Regressor model supports the trends observed in previous analyses regarding the relationship between road accident risk and gender-age cohorts. The plot indicates that road accident rates increase when there is a greater proportion of population aged 20 to 34 and aged 35 to 49. When looking at those aged 35 to 49, it is interesting to note that there is a nearly equal relationship between the increase and decrease of road accident frequency corresponding to the increase or decrease in this age cohort. Again, as seen in the feature importance plot, the gender of the population also plays a role, with road accidents being more common in areas with a higher proportion of men between 35 and 49 years of age but also a slightly lower frequency of accidents in areas where there are more men aged 50 to 64 than women.

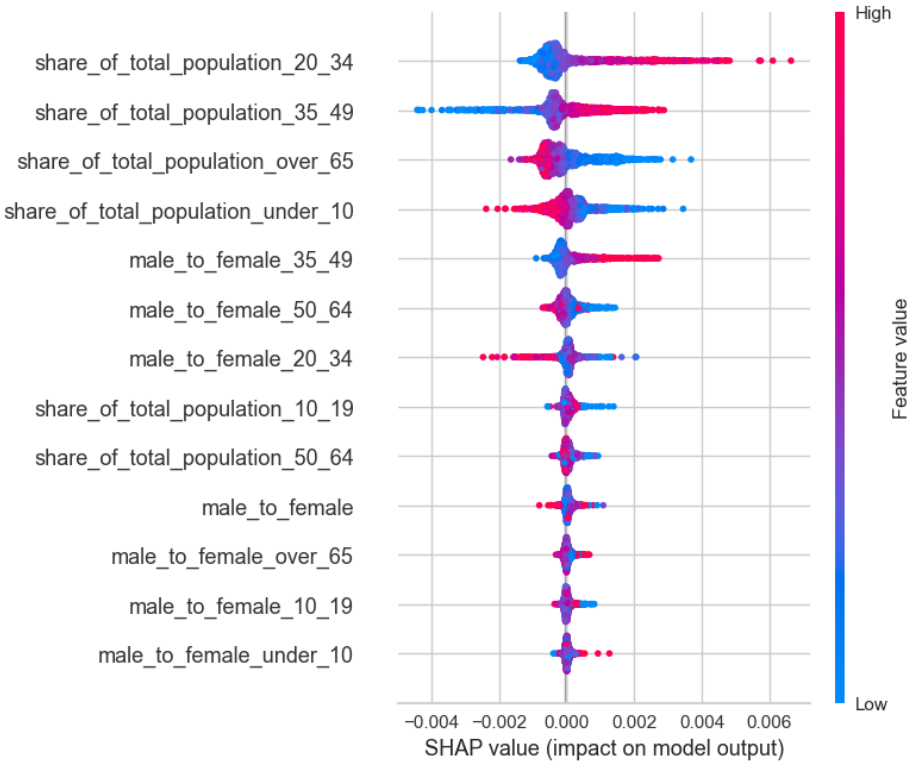

Figure 4. SHAP plot

It is important to mention that the presence of these trends does not imply a direct correlation between demographics and road accident rates. However, with that being said, with the inclusion of strong empirical connections found in the previously reviewed literatures, a conclusion can be made that there is a relationship between the features and road accident rates.

Behavior has been identified as a significant influence on road accidents, and measures targeting risky behaviors have been successfully implemented, such as drinking and driving laws, improved lighting, speed limits, and traffic calming engineering. Reducing road accidents requires a multidisciplinary approach. While obtaining detailed data on individual drivers at a large and consistent scale poses challenges, this research demonstrates the ability of gender-age cohorts to estimate road accident risk and provides a methodology for assessing the impact of new linkages between driver behavior and other characteristics. These findings contribute to the development of more accurate and powerful road accident prediction models.

In conclusion, this study establishes linkages between driving behaviors, demographic data, and road accident frequency based on existing literature. The results align with previous trends and provide a methodology for assessing the impact of new linkages. While this study does not replace more detailed research on driver behavior, it offers valuable insights and serves as a foundation for future road accident research.

6. Conclusions

Building upon existing literature, this study has applied advanced methods to thoroughly examined the relationship between risky driving behavior, demographics, and road accident rates. The findings are consistent with previous research, as demonstrated through descriptive statistical analysis, feature importance plots, and SHAP values generated by the Extra Trees Regressor model. While this study does not replace more comprehensive investigations into driver behavior, it provides a useful methodology and further strengthens the connections between driver behavior, other factors, and road accident risk.

Funding. This research was funded by xxx (International Cooperative) Research Grant in 2022, grant number 2022-0153.

References

Abdalla IM, Barker D and Raeside R (1997) Road accident characteristics and socio-economic deprivation. *Traffic Engineering & Control* 38(12): 672–676.

Abdullah P and Sipos T (2022) Drivers' behavior and traffic accident analysis using decision tree method. *Sustainability: Science Practice and Policy* 14(18). MDPI AG: 11339. DOI: 10.3390/su141811339.

Alavi SS, Mohammadi MR, Souri H, et al. (2017) Personality, Driving Behavior and Mental Disorders Factors as Predictors of Road Traffic Accidents Based on Logistic Regression.

Iranian journal of medical sciences 42(1): 24–31. DOI: 10.2105/AJPH.2014.301919.

Andrew Yockey R and Barroso CS (2023) Drugged driving among U.S. adolescents, 2016-2019, USA. *Journal of safety research* 84: 1–6. DOI: 10.1016/j.jsr.2022.10.002.

Bener A, Burgut HR, Sidahmed H, et al. (2009) Road traffic injuries and risk factors. *Californian journal of health promotion* 7(2). California State University Fullerton: 92–101. DOI: 10.32398/cjhp.v7i2.2017.

Bharadwaj N, Edara P and Sun C (2021) Sleep disorders and risk of traffic crashes: A naturalistic driving study analysis. *Safety science* 140: 105295. DOI: 10.1016/j.ssci.2021.105295.

Borgialli DA, Hill EM, Maio RF, et al. (2000) Effects of alcohol on the geographic variation of driver fatalities in motor vehicle crashes. *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine* 7(1): 7–13. DOI: 10.1111/j.1553-2712.2000.tb01882.x.

Brown ID (1982) Exposure and experience are a confounded nuisance in research on driver behaviour. *Accident; analysis and prevention* 14(5): 345–352. DOI: 10.1016/0001-4575(82)90012-4.

Chand A, Jayesh S and Bhasi AB (2021) Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. *Materials Today: Proceedings* 47: 5135–5141. DOI: 10.1016/j.matpr.2021.05.415.

Gielen AC and Sleet D (2003) Application of behavior-change theories and methods to injury prevention. *Epidemiologic reviews* 25: 65–76. DOI: 10.1093/epirev/mxg004.

Graham D, Glaister S and Anderson R (2005) The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. *Accident; analysis and prevention* 37(1): 125–135. DOI: 10.1016/j.aap.2004.07.002.

Kelly MP and Barker M (2016) Why is changing health-related behaviour so difficult? *Public health* 136: 109–116. DOI: 10.1016/j.puhe.2016.03.030.

Khayesi M, Mohan D, Nafukho FM, et al. (2005) *Road Traffic Injury Prevention Training Manual*. Genève, Switzerland: World Health Organization. Available at: https://books.google.at/books?id=9_dU2MFGA6UC.

Li YC, Sze NN and Wong SC (2013) Spatial–temporal analysis of drink-driving patterns in Hong Kong. *Accident; analysis and prevention* 59: 415–424. DOI: 10.1016/j.aap.2013.06.033.

Morgan A and Mannering FL (2011) The effects of road-surface conditions, age, and gender on driver-injury severities. *Accident; analysis and prevention* 43(5): 1852–1863. DOI: 10.1016/j.aap.2011.04.024.

Papantoniou P, Papadimitriou E and Yannis G (2017) Review of driving performance parameters critical for distracted driving research. *Transportation Research Procedia* 25: 1796–1805. DOI: 10.1016/j.trpro.2017.05.148.

Rhodes N and Pivik K (2011) Age and gender differences in risky driving: the roles of positive affect and risk perception. *Accident; analysis and prevention* 43(3): 923–931.

DOI: 10.1016/j.aap.2010.11.015.
Russo F, Biancardo SA and Dell'Acqua G (2014) Road safety from the perspective of driver gender and age as related to the injury crash frequency and road scenario. *Traffic injury prevention* 15(1): 25–33. DOI: 10.1080/15389588.2013.794943.
Yusuf IT, Adeleke OO, Salami AW, et al. (2016) THE FACTORS THAT AFFECT THE FREE FLOW SPEED ON AN ARTERIAL IN ILORIN, NIGERIA. *Nigerian Journal of Technology* 35(3): 473–480. DOI: 10.4314/njt.v35i3.2.
Zokaei M, Jafari MJ, Khosrowabadi R, et al. (2020) Tracing the physiological response and behavioral performance of drivers at different levels of mental workload using driving simulators. *Journal of safety research* 72: 213–223. DOI: 10.1016/j.jsr.2019.12.022.