Evaluation of adaptive sampling methods in scenario generation for virtual safety impact assessment of pre-crash safety systems

Xiaomi Yang*
Division of Vehicle Safety, Chalmers University of Technology
and
Henrik Imberg
Department of Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg
and
Carol Flannagan
University of Michigan Transportation Research Institute
Division of Vehicle Safety, Chalmers University of Technology
and
Jonas Bärgman
Division of Vehicle Safety, Chalmers University of Technology

*Corresponding author: Xiaomi Yang, email: xiaomi.yang@chalmers.se

Abstract

Virtual safety assessment plays a vital role in evaluating the safety impact of pre-crash safety systems such as advanced driver assistance systems (ADAS) and automated driving systems (ADS). However, as the number of parameters in simulation-based scenario generation increases, the number of crash scenarios to simulate grows exponentially, making complete enumeration computationally infeasible. Efficient sampling methods, such as importance sampling and active sampling, have been proposed to address this challenge. However, a comprehensive evaluation of how domain knowledge, stratification, and batch sampling affect their efficiency remains limited.

This study evaluates the performance of importance sampling and active sampling in scenario generation, incorporating two domain-knowledge-driven features: adaptive sample space reduction (ASSR) and stratification. Additionally, we assess the effects of a third feature, batch sampling, on computational efficiency in terms of both CPU and wall-clock time. Based on our findings, we provide practical recommendations for applying ASSR, stratification, and batch sampling to optimize sampling performance.

Our results demonstrate that ASSR substantially improves sampling efficiency for both importance sampling and active sampling. When integrated into active sampling, ASSR reduces the root mean squared estimation error (RMSE) of the estimates by up to 90%. Stratification further improves sampling performance for both methods, regardless of ASSR implementation. When ASSR and/or stratification are applied, importance sampling performs on par with active sampling, whereas when neither feature is used, active sampling is more efficient. Larger batch sizes reduce wall-clock time but increase the number of simulations required to achieve the same estimation accuracy.

In conclusion, applying ASSR and stratification in importance sampling and active sampling, where applicable, significantly improves efficiency, enabling the reallocation of computational resources to other safety initiatives.

Keywords: virtual safety impact assessment, active sampling, importance sampling, machine learning, domain knowledge, crash-causation model, glance behavior.

1 Introduction

Approximately 1.2 million people die in traffic annually (World Health Organization, 2023). To address this, both pre-crash safety systems and in-crash protection systems have been developed to prevent or mitigate the consequences of crashes (Cicchino, 2016; Eichberger et al., 2011; Jermakian, 2011; Evans, 1986; Crandall et al., 2001). Verifying the performance of safety systems is essential before their market deployment and throughout the development process. Increasingly, these assessments are performed virtually through simulations (Wimmer et al., 2023).

One virtual assessment method is counterfactual simulations, in which the specific system under assessment is virtually applied to a set of baseline crash kinematics, and the outcomes (e.g., number of crashes, impact speed, or injury risk) are compared between the baseline crashes and the simulations where the pre-crash safety systems are implemented. Consequently, baseline crashes are a prerequisite for counterfactual simulations. More precisely, a set of pre-crash kinematics is required to describe how the involved road users moved prior to the crash—without the safety system under assessment. One source of these kinematics is original reconstructed crashes from in-depth crash databases (Rosén, 2013; Stark et al., 2019; Deng et al., 2013; Yang et al., 2024). However, these databases are often small due to the cost of in-depth crash investigations. To create a larger set of crashes that better cover more of the possible parameter space and more comprehensively reflect real traffic scenarios(Riedmaier et al., 2020), new crashes can be generated from those reconstructed crashes. This can be done by applying computational crash-causation-behavior models and critical-event-response-behavior models to them. For example, Bärgman et al. (2024) validated a combination of crash-causation and critical-event-response models to generate new rear-end crashes. Their crash-causation model is in part based on drivers' off-road glances and their tendency to brake less aggressively than vehicle's full braking capacity, even in crash situations. This type of counterfactual simulation is called a behavior-model-based counterfactual simulation. Behavior-model-based counterfactual simulations typically start with a relatively small number of reconstructed crashes where the evasive maneuver of one of the road users has been removed (Rosén, 2013). Each of these events is referred to as a prototype event. The prototype event is then simulated in a virtual environment, with the driver model replacing the removed maneuver. A range of events can be constructed by varying the model parameters, for example: vehicle speed (Afsaneh et al., 2021), driver behavior (e.g., glance behavior and brake deceleration; Lee et al., 2018),

and environmental conditions (Ruiz et al., 2018).

A common problem with scenario generation for counterfactual simulations is that the number of potential crash scenarios increases exponentially as the number of parameters being varied increases. Given the already high computational cost of running the simulations, complete enumeration becomes practically infeasible—and subsampling becomes necessary. Many researchers turn to importance sampling methods to reduce the number of samples while obtaining accurate estimates (Wang et al., 2021; de Gelder and Paardekooper, 2017; Zhao et al., 2016). Importance sampling is a variance reduction technique employed in subsampling problems to increase estimation precision (Tokdar and Kass, 2010). In the context of scenario generation, the input required for this technique is a proposal distribution on the parameters used, which is a probability distribution that determines the likelihood of sample selection, guiding the sampling process towards more important and informative regions of the input space. A random sample of scenarios is then generated according to the proposal distribution, and an unbiased estimate of the characteristic of interest can subsequently be obtained by inverse probability weighting (Mansournia and Altman, 2016). Ideally, the importance sampling probabilities should be proportional to the outcome of interest for optimal performance (Bugallo et al., 2017). However, specifying the proposal distribution is challenging because the actual outcome is unknown before running the simulation, often resulting in suboptimal performance. Consequently, efficient sampling strategies to improve upon traditional importance sampling methods are needed.

Imberg et al. (2024) introduced active sampling, a machine-learning-assisted subsampling method that combines adaptive importance sampling with predictive modeling to optimize sample selection. While this method addresses many of the limitations inherent in traditional importance sampling, certain aspects and extensions of active sampling in the context of crash-causation scenario generation remain unexplored. For instance, the potential benefits of incorporating domain knowledge to reduce the sample space and minimize the number of simulations required for accurate safety impact assessment have not been investigated. Additionally, previous work by Imberg et al. (2024) focused on the average safety impact over the entire input space without considering stratification, leading to an imbalance among the original prototype events, as different prototype events generate different number of crashes under varying scenarios. An important consideration in safety impact assessment is that each prototype event should contribute equally to the evaluation, as these prototypes represent an empirical distribution of crash scenarios and thus should be weighted equally. Case weighting (through post-stratification) and stratification techniques

could help ensure a balanced representation of the prototype events in the overall safety impact assessment. Stratification has been shown to reduce variance in stochastic simulations (Park et al., 2024), yet the relative efficiency of stratification versus post-stratification in crash-causation scenario generation remains unclear. Finally, Imberg et al. (2024) did not explore the potential reduction in overall computational time that could be achieved through parallel computing—an approach that could be particularly beneficial for large-scale virtual safety assessment. To address these gaps, this study evaluates the impact of three previously unexplored features: domain knowledge integration, post-stratification versus stratification, and parallel computing, for both importance sampling and active sampling. Despite its limitations, importance sampling remains widely used and is often easier to implement than active sampling, making it an important benchmark for comparison.

1.1 Aim

The overall aim of this work is to evaluate three implementation features of two adaptive sampling methods. To determine the effectiveness of the features, we applied them to scenario generation-based virtual impact assessments of a pre-crash safety system. Specifically, we aim to

- Assess the sampling efficiency of proposed adaptive sample space reduction (ASSR) logic, which applies domain-knowledge-based logical constraints to simulation outcomes, in the context of importance sampling and active sampling.
- Evaluate the effectiveness of stratification within importance sampling and active sampling.
- Analyze the impact of batch size on efficiency in active sampling, particularly in parallel computing environments.
- Provide practical guidelines for utilizing ASSR, stratification, and batch sampling in importance sampling and active sampling for scenario generation in virtual safety impact assessments.

2 Methods

Data, models, and simulation setup are described in Section 2.1. The implementation of the three features—ASSR logic, stratification, and batch size—is detailed in Section 2.2. The

criteria for sampling stopping conditions are presented in Section 2.3. Finally, simulation procedures and performance metrics are described in Section 2.4.

2.1 Data, models, and simulation setup

The data used for scenario generation in this study consist of the reconstructed pre-crash kinematics of 44 rear-end crashes from a crash database provided by Volvo Car Corporation. The Volvo Cars Traffic Accident Database (VCTAD) contains information about crashes involving Volvo vehicles that occurred in Sweden with repair costs exceeding €4,000 (Isaksson-Hellman and Norin, 2005). The 44 crashes were selected from a total of 344 rear-end frontal crashes recorded between 2006 and 2017 on roads with a speed limit of ≥ 70 km/h (highways or expressways).

For this work, the braking action by the driver of the following vehicle (FV; the striking vehicle) in each prototype event was removed and replaced by a driver response model. This model triggers braking using a threshold on looming (optically defined time-to-collision, or TTC) as the onset of driver braking (Lee, 1976; Bärgman et al., 2024). The threshold was selected based on work by Markkula et al. (2016). In addition to the driver response model, the counterfactual model for scenario generation in this study included a crash-causation model consisting of two sub-models: an off-road glance-based crash causation sub-model, and a deceleration-based crash causation sub-model. As noted, the latter is based on the fact that, even in crash situations, drivers do not always brake to the full capacity of their vehicles (Bärgman et al., 2024). See Section 2.1.2 for details.

The off-road glance-based crash causation sub-model is based on research showing that when the driver's eyes are on the road ahead, the driver is typically able to react to looming of the lead vehicle (LV), braking early enough and hard enough to avoid a crash. As mentioned, in this crash-causation model-based scenario generation, the main causation factors are drivers' eyes-off-road (EOFF) glance durations in everyday (non-critical) driving and the braking intensity just prior to crashing. Both factors have been identified as important contributors to the occurrence of rear-end crashes (Wang et al., 2022). Naturally, glance data are required as input to such a model. Similarly, data describing drivers' braking behavior, particularly the maximum deceleration reached before impact, are needed to appropriately model driver braking as part of the crash-causation model. In the combined glance and deceleration-based crash-causation model, we treat glances and deceleration as independent variables, as there is no evidence indicating a correlation between the two.

2.1.1 An EOFF-based crash-causation sub-model

In this study, the EOFF distribution is based on baseline epochs (30s segments) from the Victor et al. (2015) study, which investigated the relationship between the LV driver's visual inattention and distraction (operationalized as off-road glances) and crash risk. EOFF glances were extracted from these 30s baseline segments and matched to the rear-end crashes and near-crashes in the SHRP2SOA8 study (Victor et al., 2015), in order to ensure that they were relevant for car-following scenarios. See Victor et al. (2015); Bärgman et al. (2024) for more details about baseline extraction and glance annotations.

For an EOFF-based crash causation model, not only is the duration of the glance important, but so is the placement of EOFF glances in relation to the critical event. Following Bärgman et al. (2024), we use an EOFF glance anchor of τ^{-1} of 0.2s^{-1}, where τ^{-1} is the inverse of the optically defined TTC ($\tau^{-1} = \theta/\dot{\theta}$, where θ is the optical angle of the width of the LV on the driver's retina, and $\dot{\theta}$ its time-derivative). As in Bärgman et al. (2024), we assume that a) the LV has an equal probability of braking at any time, b) the FV driver follows the EOFF glance distribution described above until τ^{-1} reaches $0.2s^{-1}$, and c) if the FV driver is looking at the road ahead when $\tau^{-1} = 0.2s^{-1}$, they will not look away anymore. These assumptions are based on a study of naturalistic braking in critical events reported in Markkula et al. (2016). This results in a glance anchoring scheme where only glances overlapping $\tau^{-1} = 0.2s^{-1}$ (just before the crash) are considered relevant for crash causation. Consequently, the original EOFF glance distribution can be transformed into an overshot distribution, which describes the probabilities of off-road glances exceeding the anchor point at $\tau^{-1} = 0.2s^{-1}$—hereafter referred to as OEOFF. See Bärgman et al. (2024) for a description of the overshot distribution. Once the OEOFF is obtained, the crash-causation model assumes that OEOFF glances are simply "placed" with their starting point at $\tau^{-1} = 0.2s^{-1}$. Note that OEOFF durations range from 0s to 6.6s. For the simulations, we used a bin size of 0.1s, resulting in a total of 67 bins of EOFF and OEOFF glance durations.

2.1.2 A driver maximum deceleration crash-causation sub-model

The second element of the crash-causation model is the maximum deceleration of the FV driver when there is a critical event. In safety-critical situations, deceleration generally increases quickly until it reaches a maximum. Surprisingly, the literature has shown no clear correlation between the urgency of a situation and maximum deceleration (although

urgency is correlated with reaction time and jerk; Markkula et al.,2016). That is, drivers do not seem to use the vehicle's or roadway's full potential, even if they are about to crash. This lack of optimal deceleration by drivers is the second part of our crash-causation model: drivers who could have braked more aggressively, but did not, contribute to crash causation and severity. Consequently, different maximum braking levels are used in the simulations, with a jerk value of approximately 20 m/s^3. Specifically, a maximum deceleration distribution was extracted by fitting a piecewise linear model to the pre-crash kinematics of rear-end crashes in the SHPR2 naturalistic driving dataset. For the simulations, discrete decelerations ranging from -10.25 m/s^2 to -3.75 m/s^2 with bin widths of 0.5 m/s^2 were chosen from the fitted deceleration distribution.

2.1.3 Combining two crash-causation sub-models

As mentioned, we assume that OEOFF glances and maximum decelerations are independent. That is, the probability of any specific combination of OEOFF duration and maximum driver deceleration is simply the product of the two marginal probability density functions. This joint probability distribution is the core of the crash-causation model, and the joint and marginal distributions are shown in Figure 1. Since most glances occur at 0s with probability 0.854, this category is excluded from the glance distribution illustration in Figure 1.

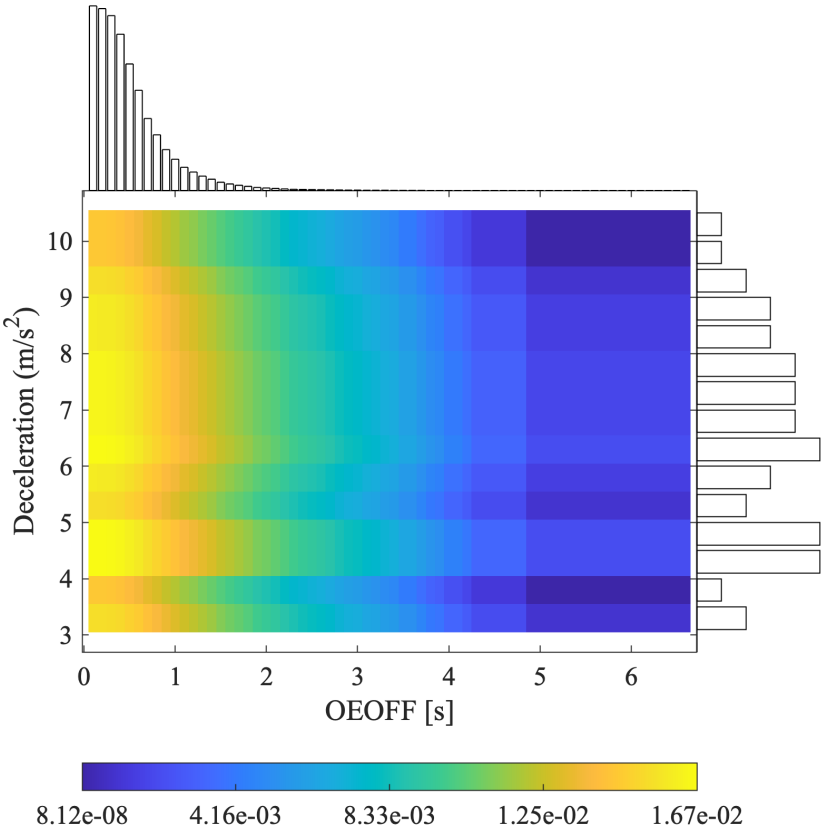

Figure 1: Heatmap of the joint probability of OEOFF and deceleration, with their marginal distributions above and to the right, respectively. Glances of 0s are excluded in this plot.

2.1.4 An advanced driver assistance system

This paper applied sampling methods to assess a specific safety system and estimate its performance, using counterfactual simulations. The system evaluated is an automated emergency braking (AEB) system provided by Zenseact (Zenseact, 2025). The AEB system assessed is a company-internal system used for testing purposes, not a production system. The counterfactual simulations compare the crashes outcomes in baseline simulations without the system (generated using the crash-causation and driver response models) to those in simulations where the AEB is applied, evaluating its safety impact.

2.1.5 Virtual simulation framework

The virtual simulations were carried out in the Volvo Car Corporation's virtual simulation tool chain, which integrates a combination of vehicle models (including brakes, chassis,

powertrain, and tires, etc.) as well as driver models. The simulations in this study were executed from a Python script. The parameter settings for each prototype event were created by the script, which combined OEOFF durations with maximum deceleration values while keeping the information about the probability of each combination. The script then instantiated simulations in Esmini (esmini, 2025), which loaded and executed the individual concrete scenarios. Specifically, for each simulation, the parameters of the crash-causation models (glance and deceleration inputs) were set. The simulation output was the relative impact speed at impact between the LV and the FV, where a relative impact speed of zero indicated a successfully avoided collision. Injury risk was calculated based on an injury risk function as a function of Δv from Stigson et al. (2012), for both baseline and countermeasure simulations (i.e., simulations where the safety system was virtually applied to baseline scenarios). An injury risk function for Maximum Abbreviated Injury Scale equal or greater than two (MAIS2+) was used. The calculation of Δv was simplified by assuming that the two vehicles had the same mass, and that the coefficient of restitution was zero (i.e., a perfectly inelastic collision). Therefore, Δv was taken as half of the relative speed at impact (hereafter referred to as impact speed). We then calculated reductions in impact speed and injury risk for each original prototype event and OEOFF-maximum deceleration pair. Finally, the simulation results was exported to `R` (R Core Team, 2024) for empirical evaluation of the adaptive sampling methods.

2.1.6 The ground truth dataset

In this work, a ground-truth dataset was created, and its safety metrics were used to assess the sampling methods. The dataset comprised all generated scenarios based on all possible combinations of parameter values across all prototype events. It was constructed by running virtual simulations for all 1,005 possible pairs of OEOFF durations (67 levels, ranging from 0.0 to 6.6s) and deceleration values (15 levels, from 3.75 to 10.25 m/s^2) for each of the 44 prototype events. Each simulation was conducted both under normal driving conditions (baseline scenario) and with the virtual AEB system applied, resulting in a total of 88,440 simulations (Table 1), same as in Imberg et al. (2024).

In contrast to Imberg et al. (2024), this study evaluates the safety impact as the average across all prototype events while ensuring an equal contribution of each prototype event to the overall safety impact assessment. Equal contribution is achieved either through stratification, where the target characteristic is first estimated separately for each prototype event and then averaged across cases, or through post-stratification, where each instance is

weighted after each sampling iteration using case-specific weights computed as the inverse of the total OEOFF–deceleration probability within the crash region of that particular case. When applied to the full dataset (ground truth), both approaches are mathematically equivalent and yield identical results.

The full-grid impact speed and injury risk distributions for the baseline scenarios across all 44 cases are shown in Figure 2. The median (range) of the maximum impact speed across the 44 prototype events was 53.6 (14.9–109.9) km/h. Approximately 38.5% of baseline simulations resulted in crashes. The ground truth impact speeds and injury risks for the baseline scenarios (excluding non-crashes) are shown in Figure 2, while Table 2 presents a comparison between baseline and the countermeasure simulations.

Table 1: Parameter levels and values used in the crash causation-based scenario generation in this study.

Parameter	Specification
OEOFF glance duration, number of levels (range)	67 (0.0–6.6 s)
Maximum deceleration, number of levels (range)	15 (3.75–10.25 m/s^2)
Number of prototype events	44
Total parameter combinations per prototype event	1,005
Total number of simulation scenarios (baseline + countermeasure)	88,440
Maximum impact speed per prototype event, median (range)	53.6 (14.9–109.9 km/h)
Proportion of crashes in baseline scenario per prototype event, median (range)	0.34 (0.0004–1.00)
Overall proportion of crashes in baseline scenario	0.385

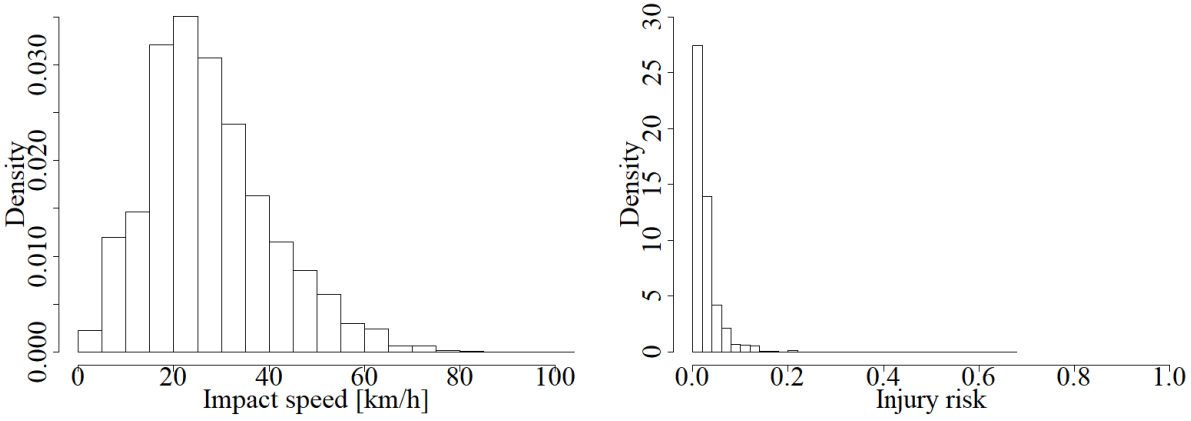

Figure 2: Empirical distributions of baseline impact speed (left) and MAIS2+ injury risk (right) in the full ground-truth dataset of simulated crashes.

Table 2: Descriptive statistics of simulation outcomes in the crash-causation-based scenario generation, comparing manual driving (baseline scenarios) and countermeasure scenarios with a automated emergency braking (AEB). In countermeasure scenarios, instances where crashes were completely avoided are included in both the impact speed and the injury risk calculations.

	Baseline	AEB	Difference
Mean impact speed (km/h)	26.3	3.0	-23.3
Mean MAIS2+ injury risk	0.027	0.002	-0.025
Proportion of crashes	100%	15.5%	-84.5%

2.2 Sampling methods

We implemented two versions of importance sampling along with the active sampling proposed by Imberg et al. (2024). To improve efficiency, we applied ASSR logic with the goal of bypassing scenario sampling and simulation runs when the outcomes could be inferred from previous simulations. Additionally, we evaluated the impact of stratification on sampling efficiency. General implementation details are provided in Section 2.2.1. The importance sampling methods are described in Section 2.2.2, followed by descriptions of ASSR logic, stratification, post-stratification, and active sampling in Sections 2.2.3 to 2.2.5.

2.2.1 General implementation

All sampling methods were implemented iteratively, selecting simulation scenarios (i.e., OEOFF and maximum deceleration parameter values) in batches according to a multinomial distribution. Here, a batch refers to the number of samples chosen in one iteration. For importance sampling, the sampling probabilities within the joint distribution of OEOFF and maximum deceleration remained fixed throughout the experiment, whereas for active sampling, they were updated iteratively.

2.2.2 Importance sampling methods

We implemented two importance sampling schemes: density importance sampling and severity importance sampling. Density importance sampling assigns sampling probabilities proportional to the probability of the OEOFF–deceleration pair. This approach aligns with the scenario generation framework, which prioritizes instances with high OEOFF–deceleration probabilities, as these contribute more to the estimation of key outcome variables, including the number of crashes, impact speed, and injury risk.

Severity importance sampling aims to further reduce variance in safety impact evaluation by oversampling high-severity instances. Given that the safety impact of an AEB

system increases with impact severity, this scheme assigns sampling probabilities proportional to $w_i \times o_i \times d_i \times m_i$, normalized to sum to 1. Here w_i represents the probability of the OEOFF–deceleration pair for instance i, o_i is the corresponding off-road glance duration, d_i is the maximum deceleration, and m_i is an *a priori* known maximum possible impact speed of instance i. To account for differences in variable scaling, all continuous variables (off-road glance duration, deceleration, and maximum impact speed) were normalized to the range $[0.1, 1]$ before computing the severity sampling probabilities.

To determine the maximum impact speed for each prototype event, severity sampling was initialized with a deterministic sample of 44 baseline-scenario simulations (one per prototype event), observed at a maximum glance duration of 6.6s and a minimal deceleration of 3.3 m/s^2 (see Section 2.2.3 for further details).

2.2.3 Using ASSR logic to reduce sampling space

In this application, crash severity is monotonically related to both glance duration and deceleration level. Some scenarios may not even result in a crash—for instance, when glance duration is short and/or braking is sufficiently strong. For a given deceleration level, as OEOFF duration increases from 0s to 6.6s, there exists a threshold at which the first crash occurs. Similarly, as glance duration further increases, there may come a point where it is too late for the driver to react at all, meaning that under the given response model, the crash happens before any braking is applied. Beyond this point, for all longer glance durations, the impact speed of crashes remains constant, as no braking occurs. Consequently, large regions of the simulation parameter space may consist entirely of non-crashes or crashes at maximum impact speed. Based on this structure, we can deduce the following:

i) **Eliminating unnecessary baseline simulations**: If a non-crash is observed in the baseline scenario for a specific prototype event, off-road glance duration, and deceleration level, then all less severe variations of that case (i.e., shorter OEOFF durations and larger decelerations) will also not result in a crash. Since our interest lies in assessing the system's effectiveness in avoiding or mitigating baseline crashes, simulations in such regions do not need to be conducted.

ii) **Skipping countermeasure simulations for non-crash scenarios**: A simulation with a countermeasure (e.g., an AEB system) will never result in a crash if the corresponding baseline scenario (without the countermeasure) did not result in a

crash. Therefore, baseline simulations should be executed first, and countermeasure simulations only need to be run for instances where a crash occurs in the baseline scenario.

iii) **Reducing countermeasure simulations for mitigated crashes**: If a crash is observed in the baseline scenario but does not occur with the countermeasure for a given prototype event, OEOFF duration, and deceleration level, then we know that none of the less severe variations (i.e., shorter OEOFF duration and greater deceleration) with the countermeasure will produce a crash. Thus, it suffices to run simulations only for the baseline scenario in that sampling space.

iv) **Inferring impact speed in extreme scenarios**: Each prototype event has a maximum impact speed, which occurs in the baseline scenario when the OEOFF duration is at its maximum and the deceleration is at its minimum. If a collision at maximum impact speed is observed at a shorter OEOFF duration or greater deceleration, then all more extreme variations (i.e., longer OEOFF durations or smaller decelerations) will also result in collisions at the same impact speed. Consequently, the impact speed in these regions can be inferred without running additional simulations.

By sampling simulation scenarios in small iterative batches, we can progressively exclude regions from the sampling space where non-crashes are observed in the baseline scenario. This approach minimizes unnecessary computations by leveraging previously run simulations to deduce outcomes, thereby improving efficiency in the sampling process.

2.2.4 Stratification and post-stratification

In this study, both stratification and post-stratification were applied to ensure balanced contributions from prototype events in the overall safety impact assessment and to correct biases arising from variations in crash frequencies across scenarios.

Stratification ensures balance at the sampling stage by dividing the parameter space into predefined strata and allocating samples proportionally. Without stratification, prototype events with a higher crash likelihood in high OEOFF–deceleration probability regions would disproportionately influence the results. By ensuring equal representation, stratification ensures that each prototype event contributes equally to the assessment, regardless of its inherent crash frequency. Estimation was performed by computing the target characteristic separately for each prototype event and then averaging across cases.

Post-stratification, in contrast, applies post hoc reweighting (Franco et al., 2017) when sampling is not explicitly stratified. Since crash frequencies and associated probabilities vary across the prototype events, each instance is assigned a weight equal to the inverse of the total OEOFF–deceleration probability within its crash region, ensuring proportional representation in the final estimates.

2.2.5 Active sampling

The active sampling method, introduced by Imberg et al. (2024), is illustrated in Figure 3. Unlike traditional 'passive' importance sampling methods, active sampling employs machine learning to iteratively optimize the sampling scheme, offering an adaptive and data-driven approach to sample selection. Below, we provide a brief overview of the active sampling method and its implementation in this study.

i) **Input**: The input to the algorithm consisted of a dataset of potential simulation scenarios, defined by prototype events and combinations of OEOFF duration and deceleration. The algorithm was optimized for a target characteristic, such as the mean impact speed reduction, crash avoidance rate, or mean injury risk reduction. Additional input parameters included batch size n_t, maximum number of iterations T, and a target precision δ, which determined the stopping condition.

ii) **Initialization**: The active sampling algorithm was initialized with a deterministic sample of 44 instances, corresponding to prototype events with the maximum OEOFF duration (6.6s) and minimum deceleration (3.3m/s^2). This allowed the maximum impact speed for each prototype event to be deduced (see Section 2.2.3), since it was expected to be an important predictor of the case-specific safety impact profile. Consequently, maximum impact speed was incorporated into subsequent learning and optimization steps to enhance predictive accuracy during the learning phase and improve the overall efficiency of the sampling algorithm.

iii) **Learning**: Observed data were used to train models predicting the probability of collision in the baseline scenarios, as well as the expected impact speed reduction, injury risk reduction, and probability of collision in the countermeasure scenarios, depending on the target characteristic. Modeling was performed using the random forest method (Breiman, 2001), implemented via the `ranger` package (version 0.14.1) (Wright and Ziegler, 2017), with hyperparameter tuning conducted through cross-validation using the `caret` package (version 6.0-92) (Kuhn, 2022) in R (R Core Team,

2024). The explanatory variables included OEOFF duration, maximum deceleration, and case-specific maximum impact speed. The learning step was implemented from the second iteration onward.

iv) **Optimization**: Based on the trained models, optimal sampling probabilities were computed according to Equation (1) in Section 2.2.6. If the learning step failed due to insufficient data (e.g., no crashes were generated) or if the machine learning model's accuracy fell below a pre-defined threshold (e.g., R-squared <0 or classification accuracy <0 on hold-out data—a portion of the dataset reserved for validation and not used during training; James et al.), density importance sampling was used as a fallback instead of active sampling. This fallback approach is theoretically optimal in the active sampling algorithm when no auxiliary information is available.

v) **Sampling**: A batch of n_t new simulation scenarios was selected at random according to a multinomial sampling design. Both baseline scenarios and corresponding countermeasure scenarios were executed. Simulation outputs included impact speed in both baseline and countermeasure scenarios, as well as the calculated outcomes: impact speed reduction, injury risk reduction, and crash avoidance with the countermeasure.

vi) **Estimation**: The target characteristics were estimated using inverse probability weighting. Additionally, the standard error of the current estimate was calculated to assess its precision. Methods for parameter and variance estimation were detailed in Imberg et al. (2024), with adjustments for stratification and post-stratification outlined in Section 2.2.4 above.

vii) **Termination**: The algorithm terminated if the standard error of the current estimate of the target characteristic (i.e., mean impact speed reduction, mean injury risk reduction, or crash avoidance rate) fell below the pre-determined target precision δ or the maximum number of iterations T was reached. If neither condition was met, the process returned to the Learning step for another iteration.

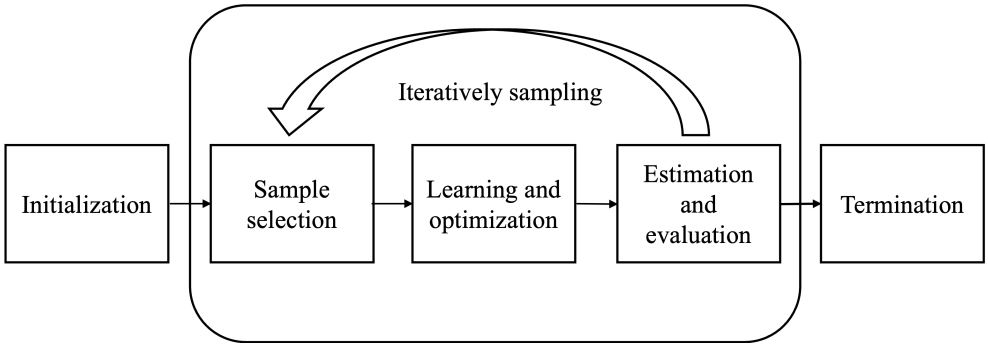

Figure 3: Flowchart illustrating the active sampling process.

2.2.6 Optimal sampling probabilities

The optimal sampling probabilities are given by

$$\pi_i \propto c_i, \tag{1}$$
$$c_i = \sqrt{\hat{p}_i w_i^2 \left[(\hat{y}_i - \hat{\mu})^2 + \sigma^2\right]}.$$

Here, i represents the index of a simulation scenario (baseline-countermeasure pair), which is defined by a combination of prototype event, OEOFF duration, and deceleration. The quantity $\hat{p}_i$ denotes the predicted probability of a collision in the baseline scenario. The term w_i represents the probability of the simulation scenario i, as determined by the crash-causation model described in Section 2.1.2, while $\hat{y}_i$ represents the predicted outcome. The predicted outcome may correspond to impact speed reduction, injury risk reduction (with the countermeasure compared to manual baseline driving), or the probability of collision in the countermeasure scenario. The term $\hat{\mu}$ is the estimated mean outcome, which may represent the mean impact speed reduction, mean injury risk reduction, or crash avoidance rate, depending on the target characteristic of interest. Finally, σ represents the residual standard deviation or root mean squared error (RMSE) of the corresponding prediction model. For a continuous outcome, σ was taken as the RMSE on hold-out data, whereas for a binary outcome y_i, it was defined as $\sigma = \sqrt{\hat{y}_i(1 - \hat{y}_i)}$. A formal proof and derivation of this formulation can be found in Imberg et al. (2024).

In the stratified setting, the target characteristics and sampling probabilities given by Equation (1) were evaluated separately for each prototype event, with the sampling probabilities normalized within each case to sum to 1. The prediction models, however, were trained across all cases to maximize the use of available data. To stabilize performance

in case-specific estimation of the mean impact speed reduction, mean injury risk reduction, or crash avoidance rate in small samples, a Bayesian-inspired shrinkage procedure was employed. Each case-specific estimate $\hat{\mu}_k$ was adjusted using a weighted combination of the observed case-level characteristic and the overall characteristic, with increasing weight assigned to the observed data and decreasing weight assigned to the grand mean as the number of generated crashes per case increased:

$$\hat{\mu}_k^{\text{shrunk}} = \rho_k \hat{\mu}_k + (1 - \rho_k)\bar{\mu}, \quad \rho_k = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_k^2/n_k}$$

where $\hat{\mu}_k$ is the case-specific safety impact characteristic, and $\bar{\mu}$ represents the grand mean or the combined target characteristic across all cases. The quantity σ_u^2 represents the variance of the estimated target characteristic across cases, σ_k^2 the variance within case k, and n_k is the sample size (i.e., the number of crashes in the baseline scenario) for case k. If no crashes were generated for a specific case, ρ_k was set to 0, ensuring that the estimate defaults to the overall mean. This approach stabilizes estimates for cases with limited data while allowing those with sufficient observations to remain closer to their empirical values, mimicking the maximum a posteriori (MAP) estimate of the mean in a Bayesian normal model.

In the non-stratified setting, sampling probabilities were calculated similarly to Equation (1) but with the values of c_i for scenarios i corresponding to a common prototype event k pre-multiplied by a case-specific normalization factor $u_k = 1/\sum_j \hat{p}_j w_j$ before calculating the sampling probabilities, where the summation over j includes all indices corresponding to case k. As above, $\hat{p}_j$ denotes the predicted probability of a collision in the baseline scenario. This adjustment accounts for the re-weighting procedure used in estimation, ensuring that each prototype event contributes equally to the overall estimation. Sampling was performed over the entire input space, with probabilities normalized to sum to 1 across all prototype events.

2.3 Stopping conditions

To conserve resources, a stopping condition is required to terminate sampling once sufficient data have been collected. By monitoring the confidence interval width—or equivalently, the standard error of the estimate—we can determine when the desired precision has been reached. When the confidence interval becomes sufficiently narrow, or the standard error falls below a predefined threshold, sampling can be terminated.

Stopping conditions are typically informed by domain knowledge and may be based on metrics such as the region of practical equivalence (ROPE) or the coefficient of variation (Schwaferts and Augustin, 2020; Kukuková et al., 2008; Carrasco and Jover, 2003). The appropriate thresholds for these metrics should be chosen based on expert judgment. A ROPE value defines the range within which differences are considered practically equivalent and can be determined through, for example, expert consensus in a workshop setting. For instance, if experts agree that a crash avoidance rate within $\pm$ 0.05 (five percentage points) is practically equivalent, then a threshold of 0.025 for the standard error or 0.05 for the confidence interval half-width can be chosen as the stopping threshold. In contrast, the coefficient of variation is a percentage representing the allowable uncertainty of the estimate relative to the quantity of interest. A typical value, such as 2.5%, ensures that the standard error does not exceed 2.5% of the estimated parameter value, providing a relative measure of precision. Thus, the stopping condition is met when the standard error drops below this percentage of the target characteristic estimate.

Another practical stopping criterion is the limitation imposed by available resources, such as computational capacity or constraints. If only a fixed proportion, for example 10% of the total possible simulations, can be run due to resource limitations, sampling must stop once this limit is reached. Similarly, if there is a fixed amount of CPU time, memory or wall-clock time allocated, sampling must conclude when the allocated resources are exhausted.

2.4 Metrics and simulation performance evaluation

Sampling performance was evaluated using the RMSE of the estimator for the mean impact speed reduction, crash avoidance rate, or mean injury risk reduction. RMSE was calculated as the square root of the mean squared deviation of the estimate from the ground truth across 200 independent subsampling experiments. This was analyzed and presented graphically as a function of the number of virtual crash-causation model simulations. Unless otherwise stated, the batch size was set to ten simulations per iteration.

The following comparisons were performed:

i) **Comparison of sampling methods**: Active sampling compared to density importance sampling and severity importance sampling.

ii) **Effect of domain knowledge**: Comparison of active sampling with and without domain-knowledge-based ASSR logic.

iii) **Stratification versus post-stratification (without ASSR)**: Active sampling and the better-performing importance sampling method (i.e., severity sampling), both without ASSR, comparing stratification to post-stratification.

iv) **Stratification versus post-stratification (with ASSR)**: Active sampling and the better-performing importance sampling method (i.e., severity sampling), both with ASSR, comparing stratification to post-stratification.

v) **Impact of batch size**: Active sampling with batch sizes of 44, 132, and 440 per iteration (i.e., 1x, 3x, and 10x the 44 prototype events).

3 Results

This section presents the results of the sampling strategy evaluations, starting with a comparison of active sampling and importance sampling in Section 3.1, followed by the effect of ASSR logic in Section 3.2, stratification versus post-stratification in 3.3, and the influence of batch size on active sampling performance in Section 3.4.

3.1 Comparison of importance sampling and active sampling performance

The RMSE of active sampling is compared with density and severity importance sampling (Figure 4). For this comparison, post-stratification (case weighting) was used, ensuring that each prototype event contributed equally to the estimation. Active sampling outperformed importance sampling for all three targets—mean impact speed reduction, crash avoidance rate, and injury risk reduction—regardless of the optimization target. However, for small samples, the methods performed similarly. As expected, active sampling performed best for the specific target characteristic it was optimized on.

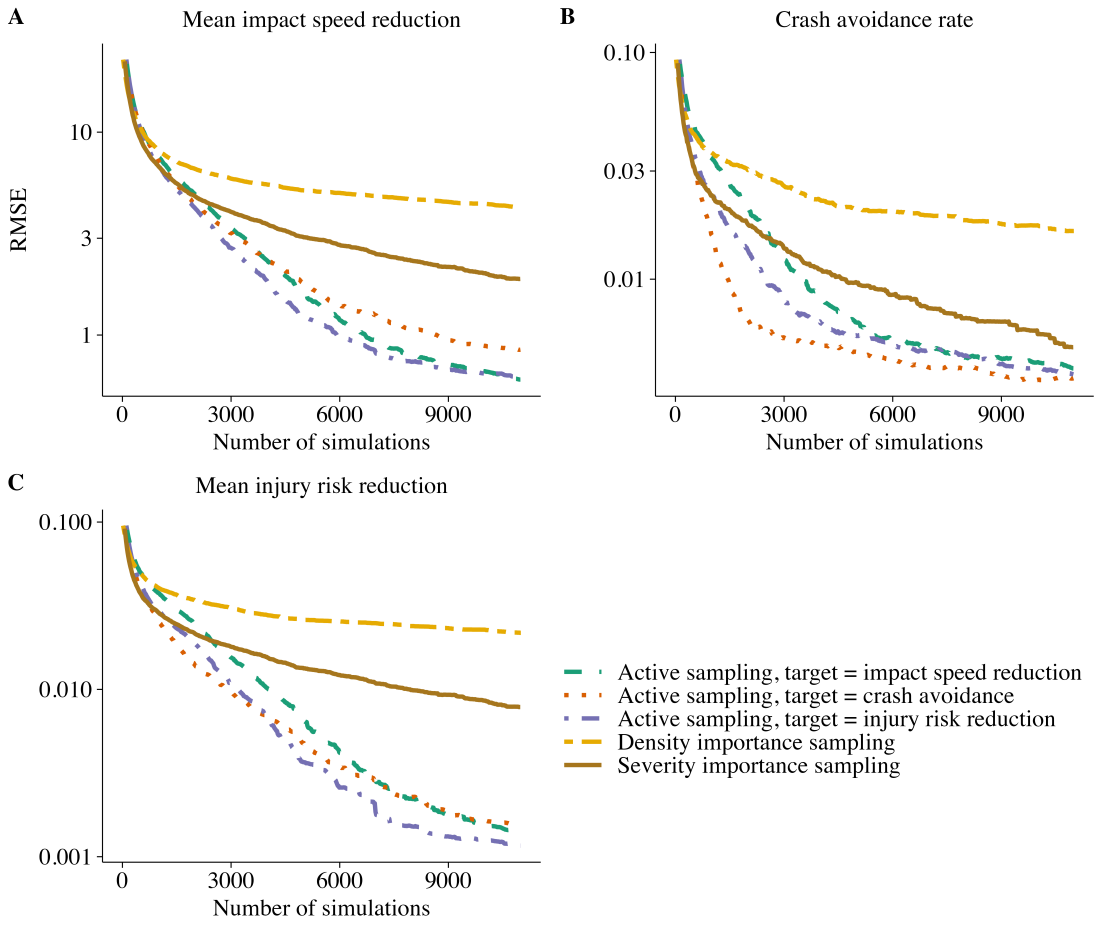

Figure 4: Root mean squared error (RMSE) for estimating (A) mean impact speed reduction, (B) crash avoidance rate, and (C) mean injury risk reduction using active sampling, compared to density and severity importance sampling. Active sampling consistently outperformed importance sampling, except for small sample sizes, where their performance was comparable. Post-stratification (case weighting) was applied to ensure equal contribution from all prototype events.

3.2 Effect of ASSR logic on active sampling performance

The RMSE of active sampling with and without domain-knowledge-based ASSR logic are compared in Figure 5. For all three target characteristics, sampling with ASSR logic outperformed sampling without it for the same number of simulations. At 6,000 simulations, for example, the RMSE was reduced by 32% to 89%, depending on the outcome.

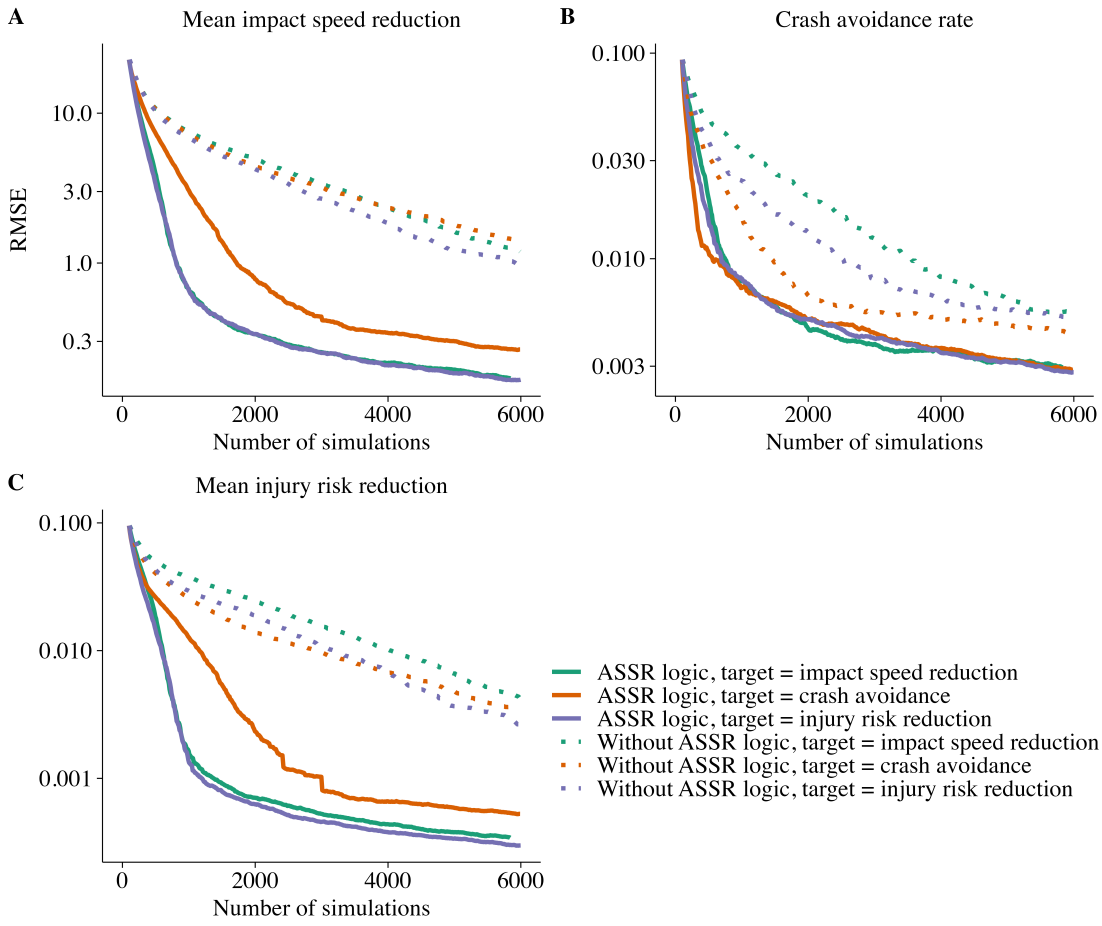

Figure 5: Root mean squared error (RMSE) for estimating (A) mean impact speed reduction, (B) crash avoidance rate, and (C) mean injury risk reduction using active sampling with and without ASSR logic. ASSR logic provided substantial performance improvements across all settings.

3.3 Effect of stratification vs post-stratification on sampling performance

Figure 6 compares the RMSE of active sampling and severity importance sampling with and without stratification, excluding ASSR logic. Both active sampling and severity importance sampling generally performed better with stratification than with post-stratification. Without ASSR logic or stratification, active sampling outperformed severity importance sampling. However, when stratification was applied, their performances were comparable. Moreover, as the sample size increased, the performance of active sampling with post-stratification approached that of stratification.

Figure 7 compares the RMSE of active sampling and severity importance sampling (the better-performing importance sampling method) using stratification versus post-stratification,

this time incorporating ASSR logic. For mean impact speed reduction and injury risk reduction, stratification and post-stratification yielded similar results. However, stratification improved performance for crash avoidance rate. Additionally, when either ASSR logic or stratification were incorporated, active sampling and severity importance sampling performed similarly (shown in Figure 6, 7).

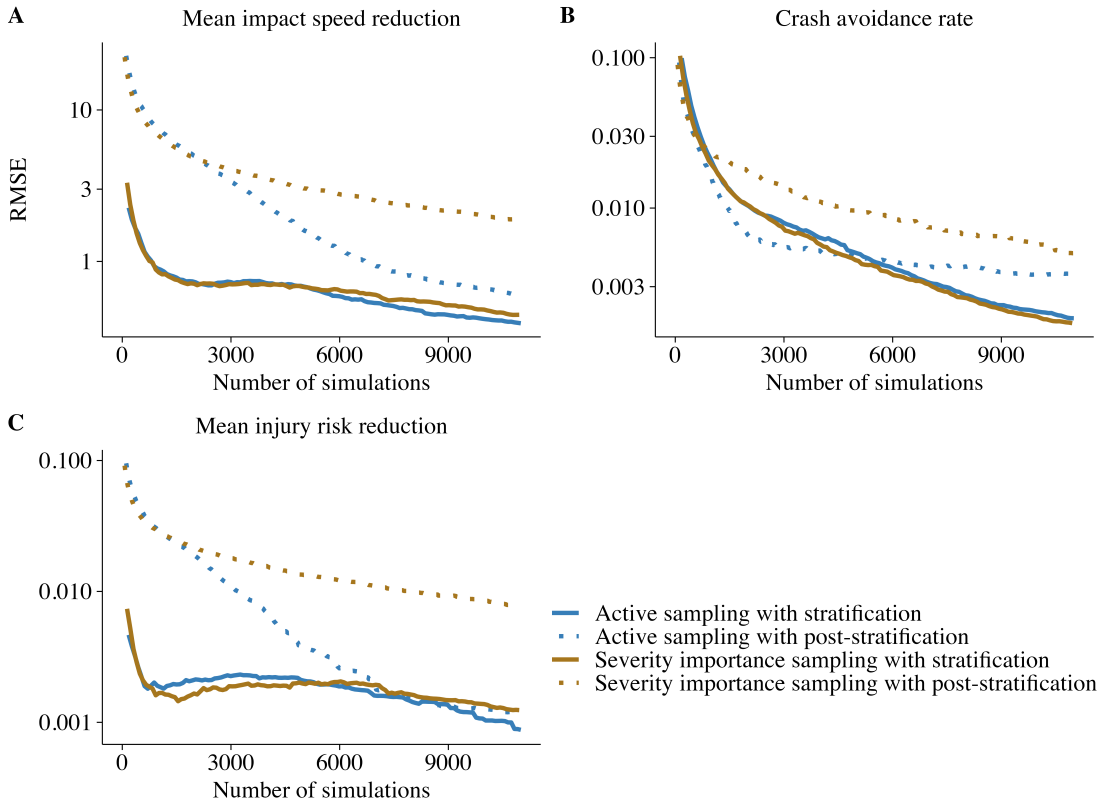

Figure 6: Root mean squared error (RMSE) for estimating (A) mean impact speed reduction, (B) crash avoidance rate, and (C) mean injury risk reduction using active sampling and severity importance sampling, each with either stratification or post-stratification. Neither method used ASSR logic. Active sampling was optimized for mean impact speed reduction in (A), crash avoidance rate in (B), and mean injury risk reduction in (C). Stratification improved performance for both methods, while active sampling with post-stratification approached the performance of stratification as sample size increased.

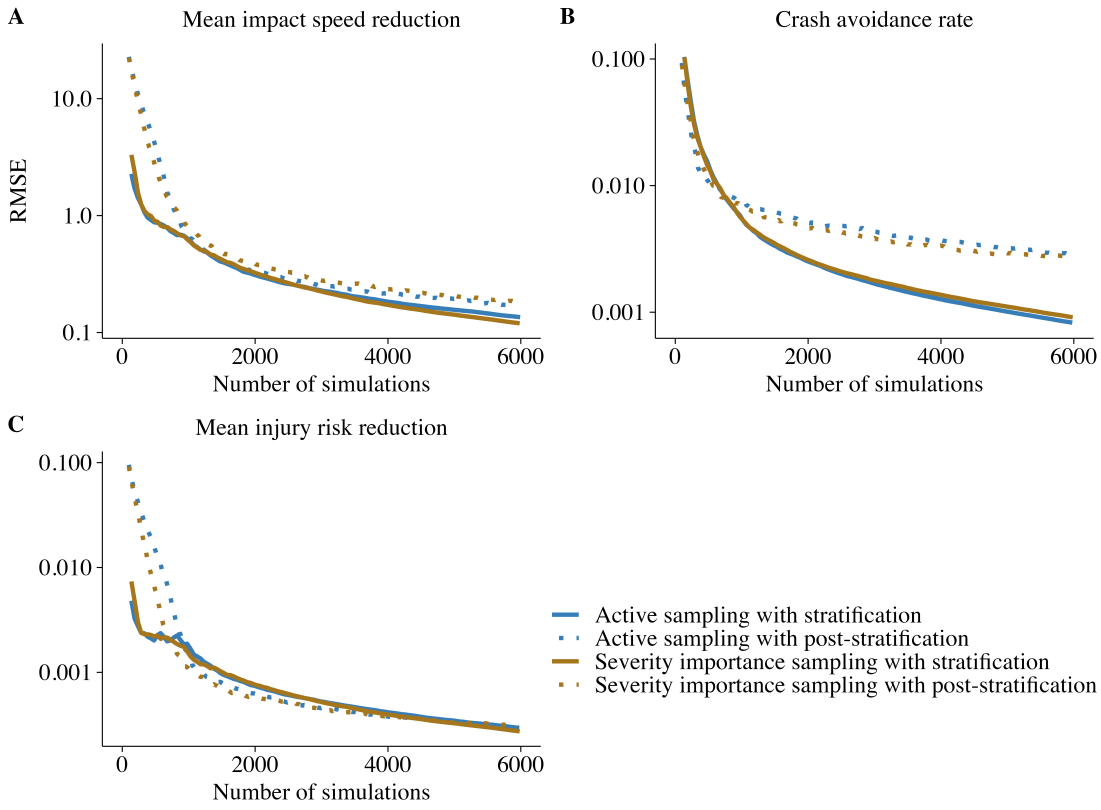

Figure 7: Root mean squared error (RMSE) for estimating (A) mean impact speed reduction, (B) crash avoidance rate, and (C) injury risk reduction using active sampling and severity importance sampling, each with either stratification or post-stratification. Both methods incorporated ASSR logic. Active sampling was optimized for mean impact speed reduction in (A), crash avoidance rate in (B), and mean injury risk reduction in (C). With ASSR logic was included, active sampling and severity importance sampling had similar performance using both stratification and post-stratification, although stratification improved performance for crash avoidance rate.

3.4 Effect of batch size on active sampling performance

Figure 8 shows the RMSE of active sampling (with ASSR logic and stratification) for varying batch sizes. Performance declined as the batch size increased, but the effect was gradual. There were minor performance losses when increasing from 44 (one per prototype event) to 132 (three per prototype event), but more noticeable reductions as batch size further increased to 440 (ten per prototype event).

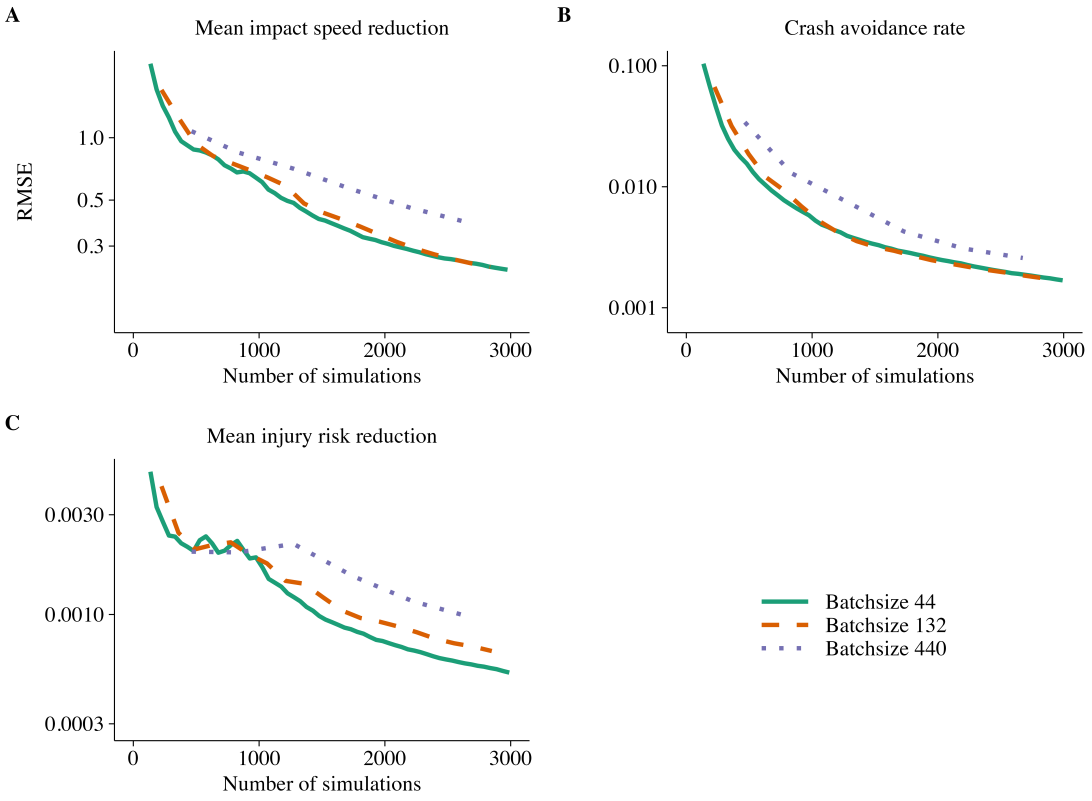

Figure 8: Root mean squared error (RMSE) for estimating (A) mean impact speed reduction, (B) crash avoidance rate, and (C) mean injury risk reduction using active sampling with batch sizes of 44 (one per prototype event), 132 (three per prototype event), or 440 (ten per prototype event) per iteration in the active sampling algorithm. Both ASSR logic and stratification were used. Active sampling was optimized for mean impact speed reduction in (A), crash avoidance rate in (B), and mean injury risk reduction in (C). Performance gradually declined as batch size increased.

4 Discussion

This study advances the methodology for adaptive sampling in scenario-generation-based safety system assessment by refining existing approaches and integrating domain knowledge to enhance efficiency. We examined traditional importance sampling methods and adaptive importance sampling through active sampling, evaluating the impact of three key implementation features: adaptive sample space reduction (ASSR), stratification versus post-stratification, and batch size effects. Our findings indicate that ASSR improved efficiency for both importance sampling and active sampling, with importance sampling benefiting the most. Stratification consistently improved performance, while post-stratification yielded comparable results to stratification for active sampling at larger sample sizes. Active

sampling outperformed importance sampling when no domain knowledge was incorporated, but their performance was similar when ASSR and stratification were applied. The following discussion explores these results, their broader implications, and potential directions for future research.

Compared to previous work on active sampling in crash-causation-based scenario generation by Imberg et al. (2024), this study enforced a balanced contribution from the prototype events to the target characteristic of interest. In the context of safety impact assessment, crash exposure must be explicitly accounted for. Since prototype events represent crash exposure, each event should contribute equally to the overall safety impact assessment. Weighting is widely used to ensure the representativeness of the crash exposure in datasets (Babisch et al., 2023; Pfeiffer and Schmidt, 2007; Clark and Hannan, 2013). Consequently, this study employed case weighting—either by design through stratification or by estimation through post-stratification—to ensure that each prototype event made an equal contribution to the estimation of target characteristics. When ASSR and stratification were not applied, our results align with Imberg et al. (2024), confirming the superiority of active sampling over importance sampling for scenario generation when domain knowledge is limited.

One of the key challenges in implementing importance sampling for safety impact assessment is selecting an appropriate proposal distribution for generating scenarios, as performance depends heavily on how closely the proposal distribution aligns with the ground truth distribution (Bugallo et al., 2017). A poorly chosen proposal distribution may reduce performance (Swiler and West, 2010), potentially performing even worse than simple random sampling. Consequently, the two importance sampling methods considered in this study had different performance, and while severity-based sampling might be expected to perform better, its advantage cannot be determined in advance. In contrast, active sampling does not rely on a predefined proposal distribution but instead learns the optimal sampling scheme dynamically during the simulation process. However, it does require a target characteristic for optimization, which can influence performance. Extensions of active sampling to multivariate settings, where multiple characteristics are optimized simultaneously, are possible using optimal design theory (Pukelsheim, 1993; Imberg et al., 2022, 2023). When the target characteristics of interest are highly correlated, as in this study and often in scenario-generation for safety impact assessment, this has a relatively small impact on sampling efficiency. Ultimately, active sampling provides a more data-driven and robust alternative to importance sampling, particularly in settings where defining an

optimal proposal distribution is challenging.

In this work, domain knowledge was incorporated in the form of ASSR, which is closely tied to the crash-causation model that defines crash mechanisms through parameters influencing crash occurrence (Davis et al., 2011). ASSR enabled the elimination of unnecessary simulations by applying logical constraints: i) longer glances and lower decelerations increase crash risk and impact speed, while ii) shorter glances and greater decelerations increase the likelihood of crash avoidance. These constraints streamline the sampling process by avoiding redundant simulations. More broadly, this approach may be applicable to other scenario-generation problems where the relationship between input parameters and crash severity is known, as more extreme inputs typically correspond to more and more severe crashes, while less extreme inputs lead to milder outcomes. Such structured constraints can improve efficiency, particularly in knowledge-based scenario generation (Ding et al., 2023; da Costa et al., 2024; McDuff et al., 2022), which often relies on predefined expert rules or integrates external knowledge.

When ASSR logic was incorporated, the performance of all sampling methods improved, although the magnitude of the improvement varied across settings. ASSR led to greater efficiency gains for importance sampling, ultimately resulting in equal performance between active sampling and importance sampling when ASSR was applied. This may be due to the partially overlapping objectives and benefits of active sampling and ASSR. Active sampling, guided by machine learning, identifies and oversamples regions with a high likelihood of generating informative scenarios (e.g., with a high likelihood of a crash in the baseline condition). However, if a rule-based approach like ASSR can achieve similar outcomes by systematically reducing the sample space, the added value of active sampling diminishes, explaining its comparatively smaller performance gains.

Stratification is a well-established variance reduction technique, particularly effective in settings with substantial heterogeneity across strata (Singh et al., 1996; Park et al., 2024; Jing et al., 2015), which is also confirmed in this study. Across all settings, stratification significantly improved sampling performance by balancing case representation and ensuring more accurate estimation for all prototype events. However, for injury-risk reduction, RMSE remained unchanged over a substantial range of simulation counts. This may be attributed to the large non-crash or low-severity crash regions in the sampling space, combined with the exponential response curve of the MAIS2+ injury risk function, which required the generation of a sufficient number of high-severity crashes for accurate assessment. Notably, active sampling without stratification approached the performance

of stratified sampling as sample sizes increased, as the algorithm dynamically learned the underlying structure over iterations. In contrast, importance sampling consistently exhibited a performance gap between stratified and non-stratified sampling, emphasizing its reliance on explicit stratification to achieve balanced case representation and reinforcing the advantage of active sampling.

Batch size plays a crucial role in both sampling performance and computational efficiency (Citovsky et al., 2021). In our study, larger batch sizes slowed the adaptive learning process of active sampling, as the algorithm had fewer opportunities to adjust between iterations, leading to reduced sampling efficiency. However, in parallel computing environments, larger batch sizes can significantly reduce wall-clock time, enabling faster overall safety impact assessments. This trade-off highlights the need to balance simulation effort with computational efficiency when selecting batch sizes. Smaller batches are preferable in settings where minimizing the number of simulations is a priority, such as in resource-constrained settings. Conversely, larger batches are advantageous when parallelization resources are available, allowing for faster completion times despite the increased number of simulations required. An alternative to fixed batch sizes is the use of adaptive batch sizes, which dynamically adjust over iterations to balance sampling efficiency and computational performance (Ma et al., 2021). The relationship between batch size and machine learning model learning rate can also be leveraged to scale batch size adaptively across iterations (Devarakonda et al., 2017; Balles et al., 2017). Balancing batch size considerations ensures optimal performance tailored to the specific computational and resource constraints of the application.

4.1 Generalization and implications

The findings of this study have broader implications for improving computational efficiency in scenario generation. ASSR techniques informed by domain knowledge proved highly effective in reducing computational costs and can be extended to other parameterized scenario generation tasks. When such domain knowledge is available, importance sampling with ASSR is highly efficient, whereas in its absence, active sampling remains more efficient. Both methods show benefit from stratification, where applicable, further enhancing performance. Future research could explore extensions of this work to more complex scenarios, including higher-dimensional parameter spaces, continuous sampling spaces, and advanced crash-causation models, to further refine and optimize adaptive sampling techniques for

safety impact assessments.

4.2 Limitations

This study focused on estimating finite population characteristics, such as means or ratios, in the context of AEB system safety impact assessment. While the proposed methods demonstrated high efficiency for these objectives, they may be less effective for other scenario generation tasks, such as identifying corner cases or extreme scenarios. In such cases, methods designed for anomaly detection (Chandola et al., 2009) or Bayesian optimization (Frazier, 2018), may be more suitable.

Another limitation concerns the use of domain-knowledge-based logic components within the ASSR technique. While effective for parameterized scenarios, integrating logic becomes more challenging when machine learning methods generate concrete scenarios directly from data without parametrization, as in Zorin and Mercier (2024). The lower interpretability of machine learning outputs and the lack of explicit parametrization make it difficult to apply rule-based constraints directly linked to crash outcomes. This restricts the applicability of ASSR and logic-based sampling techniques in settings where explicitly parametrized input data are unavailable.

Finally, this study employed a simplified delta-v calculation, assuming equal vehicle masses and no elasticity during impact. While this simplification reduces the realism of absolute injury risk estimates, it does not affect the validity of the study's conclusions regarding the efficiency of the sampling methods.

4.3 Conclusion

This study evaluated how three key implementation features—ASSR, stratification, and batch sampling—influence the efficiency of importance sampling and active sampling for scenario generation in virtual safety impact assessment. Both ASSR and stratification substantially improved efficiency for both methods, with active sampling performing better in the absence of these features, while importance sampling performed on par with active sampling when either stratification or ASSR was applied. Batch size influenced performance, with moderate increases having negligible effects, suggesting that larger batch sizes reduce wall-clock time in parallel computing environments but increase total simulation effort, emphasizing the need for resource balancing. When ASSR and/or stratification are applicable, we recommend incorporating these features into adaptive sampling methods.

Additionally, batch size should be optimized to balance sampling efficiency and computational constraints. Implementing the three features may reduce costs and enable resource reallocation to other traffic safety initiatives, ultimately improving the overall efficiency of virtual safety impact assessment methodologies.

Acknowledgment

This work is funded by the SHAPE-IT project under the European Union's Horizon 2020 research and innovation programme (under the Marie Skłodowska-Curie grant agreement 860410). We would like to thank the European Commission for funding the work. We also extend our gratitude to Volvo Car Corporation for allowing us to use their data and simulation tool, and in particular Malin Svärd and Simon Lundell at Volvo Cars for support in the simulation setup, and Mattias Robertson for his administrative support. Furthermore, we thank Marina Axelson-Fisk and Johan Jonasson at the Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, for their valuable discussions on the methodological aspects of this work.

References

Afsaneh, B., Yves, P., Felix, F., Hendrik, W., Elina, A., Per, H., Esko, L., Anne, S., Jonas, B., Marcel, B., Satu, I., Teemu, I., Fanny, M., Karl, P., Michael, S., Henri, S., Thomas, S., Walter, H., Thierry, H., Johannes, H., and Guilhermina, T. (2021). L3pilot deliverable d7.4 impact evaluation results. Technical report, L3Pilot. Available at: `https://l3pilot.eu/fileadmin/user_upload/Downloads/Deliverables/Update_14102021/L3Pilot-SP7-D7.4-Impact_Evaluation_Results-v1.0-for_website.pdf`.

Babisch, S., Neurohr, C., Westhofen, L., Schoenawa, S., and Liers, H. (2023). Leveraging the gidas database for the criticality analysis of automated driving systems. *Journal of Advanced Transportation*, 2023(1):1349269.

Balles, L., Romero, J., and Hennig, P. (2017). Coupling adaptive batch sizes with learning rates. In *Uncertainty in Artificial Intelligence - Proceedings of the 33rd Conference, UAI 2017*.

Bärgman, J., Svärd, M., Lundell, S., and Hartelius, E. (2024). Methodological challenges of scenario generation validation: a rear-end crash-causation model for virtual safety

assessment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 104:374–410.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34.

Carrasco, J. L. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59(4):849–858.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Cicchino, J. B. (2016). Effectiveness of forward collision warning systems with and without autonomous emergency braking in reducing police-reported crash rates. *Arlington, VA: Insurance Institute for Highway Safety*.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., and Kumar, S. (2021). Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944.

Clark, D. E. and Hannan, E. L. (2013). Inverse propensity weighting to adjust for bias in fatal crash samples. *Accident Analysis & Prevention*, 50:1244–1251.

Crandall, C. S., Olson, L. M., and Sklar, D. P. (2001). Mortality reduction with air bag and seat belt use in head-on passenger car collisions. *American journal of epidemiology*, 153(3):219–224.

da Costa, A. A. B., Irvine, P., Zhang, X., Khastgir, S., and Jennings, P. (2024). Ontology-based scenario generation for automated driving systems verification and validation using rules of the road. *IEEE Transactions on Intelligent Vehicles*.

Davis, G. A., Hourdos, J., Xiong, H., and Chatterjee, I. (2011). Outline for a causal model of traffic conflicts and crashes. *Accident Analysis & Prevention*, 43(6):1907–1919.

de Gelder, E. and Paardekooper, J.-P. (2017). Assessment of automated driving systems using real-life scenarios. In *2017 ieee intelligent vehicles symposium (iv)*, pages 589–594. IEEE.

Deng, B., Wang, H., Chen, J., Wang, X., and Chen, X. (2013). Traffic accidents in shanghai—general statistics and in-depth analysis. In *Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles*, pages 27–30.

Devarakonda, A., Naumov, M., and Garland, M. (2017). Adabatch: Adaptive batch sizes for training deep neural networks. *arXiv preprint arXiv:1712.02029*.

Ding, W., Xu, C., Arief, M., Lin, H., Li, B., and Zhao, D. (2023). A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):6971–6988.

Eichberger, A., Tomasch, E., Hirschberg, W., and Steffan, H. (2011). Potentials of active safety and driver assistance systems. *ATZ worldwide eMagazine*, 113(7):56–63.

esmini (2025). esmini github repository. Accessed: 2025-01-28.

Evans, L. (1986). The effectiveness of safety belts in preventing fatalities. *Accident Analysis & Prevention*, 18(3):229–241.

Franco, A., Malhotra, N., Simonovits, G., and Zigerell, L. (2017). Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science*, 4(2):161–172.

Frazier, P. I. (2018). A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Imberg, H., Axelson-Fisk, M., and Jonasson, J. (2023). Optimal subsampling designs. *arXiv:2304.03019*.

Imberg, H., Lisovskaja, V., Selpi, and Nerman, O. (2022). Optimization of two-phase sampling designs with application to naturalistic driving studies. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3575–3588.

Imberg, H., Yang, X., Flannagan, C., and Bärgman, J. (2024). Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. *Technometrics*, 67(1):46–57.

Isaksson-Hellman, I. and Norin, H. (2005). How thirty years of focused safety development has influenced injury outcome in Volvo cars. *Annual Proceedings. Association for the Advancement of Automotive Medicine*, 49:63–77.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jermakian, J. S. (2011). Crash avoidance potential of four passenger vehicle technologies. *Accident Analysis & Prevention*, 43(3):732–740.

Jing, L., Tian, K., and Huang, J. Z. (2015). Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition*, 48(11):3688–3702.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-92.

Kukuková, A., Noël, B., Kresta, S. M., and Aubin, J. (2008). Impact of sampling method and scale on the measurement of mixing and the coefficient of variance. *AIChE journal*, 54(12):3068–3083.

Lee, D. N. (1976). A theory of visual control of braking based on information about time-to-collision. *Perception*, 5(4):437–459.

Lee, J. Y., Lee, J. D., Bärgman, J., Lee, J., and Reimer, B. (2018). How safe is tuning a radio?: using the radio tuning task as a benchmark for distracted driving. *Accident Analysis & Prevention*, 110:29–37.

Ma, Z., Xu, Y., Xu, H., Meng, Z., Huang, L., and Xue, Y. (2021). Adaptive batch size for federated learning in resource-constrained edge computing. *IEEE Transactions on Mobile Computing*, 22(1):37–53.

Mansournia, M. A. and Altman, D. G. (2016). Inverse probability weighting. *Bmj*, 352.

Markkula, G., Engström, J., Lodin, J., Bärgman, J., and Victor, T. (2016). A farewell to brake reaction times? kinematics-dependent brake response in naturalistic rear-end emergencies. *Accident Analysis & Prevention*, 95:209–226.

McDuff, D., Song, Y., Lee, J., Vineet, V., Vemprala, S., Gyde, N. A., Salman, H., Ma, S., Sohn, K., and Kapoor, A. (2022). Causalcity: Complex simulations with agency for causal discovery and reasoning. In *Conference on Causal Learning and Reasoning*, pages 559–575. PMLR.

Park, J., Byon, E., Ko, Y. M., and Shashaani, S. (2024). Strata design for variance reduction in stochastic simulation. *Technometrics*, pages 1–12.

Pfeiffer, M. and Schmidt, J. (2007). Statistical and methodological foundations of the gidas accident survey system.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley, New York.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., and Diermeyer, F. (2020). Survey on scenario-based safety assessment of automated vehicles. *IEEE access*, 8:87456–87477.

Rosén, E. (2013). Autonomous emergency braking for vulnerable road users. In *Proceedings of IRCOBI conference*, pages 618–627.

Ruiz, N., Schulter, S., and Chandraker, M. (2018). Learning to simulate. *arXiv preprint arXiv:1810.02513*.

Schwaferts, P. and Augustin, T. (2020). Bayesian decisions using regions of practical equivalence (rope): Foundations.

Singh, R., Mangat, N. S., Singh, R., and Mangat, N. S. (1996). Stratified sampling. *Elements of survey sampling*, pages 102–144.

Stark, L., Düring, M., Schoenawa, S., Maschke, J. E., and Do, C. M. (2019). Quantifying vision zero: Crash avoidance in rural and motorway accident scenarios by combination of acc, aeb, and lks projected to german accident occurrence. *Traffic injury prevention*, 20(sup1):S126–S132.

Stigson, H., Kullgren, A., and Rosén, E. (2012). Injury risk functions in frontal impacts using data from crash pulse recorders. In *Annals of Advances in Automotive Medicine/Annual Scientific Conference*, volume 56, page 267. Association for the Advancement of Automotive Medicine.

Swiler, L. and West, N. (2010). Importance sampling: Promises and limitations. In *51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 18th AIAA/ASME/AHS Adaptive Structures Conference 12th*, page 2850.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Victor, T., Dozza, M., Bärgman, J., Boda, C.-N., Engström, J., Flannagan, C., Lee, J. D., and Markkula, G. (2015). Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk. Technical report.

Wang, X., Liu, Q., Guo, F., Xu, X., Chen, X., et al. (2022). Causation analysis of crashes and near crashes using naturalistic driving data. *Accident Analysis & Prevention*, 177:106821.

Wang, X., Peng, H., and Zhao, D. (2021). Combining reachability analysis and importance sampling for accelerated evaluation of highway automated vehicles at pedestrian crossing. *ASME Letters in Dynamic Systems and Control*, 1(1):011017.

Wimmer, P., Op_Den_Camp, O., Weber, H., Chajmowicz, H., Wagner, M., Mallada, J. L., Fahrenkrog, F., and Denk, F. (2023). Harmonized approaches for baseline creation in prospective safety performance assessment of driving automation systems'. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Yokohama, Japan*, pages 3–6.

World Health Organization (2023). *Global status report on road safety 2023*. URL `https://www.who.int/publications/i/item/9789240086517`.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Yang, X., Lubbe, N., and Bärgman, J. (2024). Evaluation of comfort zone boundary based automated emergency braking algorithms for car-to-powered-two-wheeler crashes in china. *IET Intelligent Transport Systems*, 18(9):1599–1615.

Zenseact (2025). Zenseact official website. Accessed: 2025-01-28.

Zhao, D., Lam, H., Peng, H., Bao, S., LeBlanc, D. J., Nobukawa, K., and Pan, C. S. (2016). Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques. *IEEE transactions on intelligent transportation systems*, 18(3):595–607.

Zorin, A. and Mercier, L. (2024). A new approach to ad/adas test scenario generation using open-source intelligence and large language models.