Practical validation of synthetic pre-crash scenarios

Jian Wu[a,b,*], Ulrich Sander[a], Carol Flannagan[b,c], Jonas Bärgman[b]

[a] *Safety Center, Volvo Cars, 41878 Göteborg, Sweden,*
[b] *Department of Mechanics and Maritime Sciences, Chalmers University of Technology, 41756 Göteborg, Sweden,*
[c] *University of Michigan Transportation Research Institute, Ann Arbor, Michigan 48109, USA,*

Abstract

The representativeness of synthetic pre-crash scenarios is crucial for assessing the safety impact of Driving Automation Systems through virtual simulations. However, a gap remains in the robust evaluation of synthetic pre-crash scenarios' practical equivalence to their real-world counterparts; that is, whether they are similar enough for the intended assessment purpose. Conventional significance testing is inadequate, as it focuses on detecting differences rather than establishing practical equivalence. This study addresses the research gap by extending our previous work on a Bayesian Region of Practical Equivalence (ROPE)-based equivalence testing framework by introducing a binning-based approach to define appropriate statistics and equivalence criteria. Two binning-based statistics are proposed to measure practically meaningful distributional differences between datasets in the context of safety impact assessment. The framework's applicability is demonstrated through a case study, which tests the practical equivalence of two synthetic rear-end pre-crash datasets with a previously developed reference dataset in the context of the safety impact assessment of an Automatic Emergency Braking system. The results show that the framework provides informative quantitative assessments of practical equivalence as well as diagnostic insights into the divergence of datasets. Although the demonstration focuses on rear-end pre-crash scenarios, the framework is generic and extensible to broader validation contexts, providing an interpretable and principled basis for practical equivalence assessment across diverse synthetic data applications.

Keywords: Practical Equivalence Testing, Synthetic Pre-Crash Scenarios, Driving Automation Systems, Virtual Simulations, Safety Impact Assessment.

1. Introduction

Driving Automation Systems (DAS) (On-Road Automated Driving Committee, 2021), including Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS), are expected to reduce crash risk and improve traffic safety (Pradhan et al., 2022). To assess their safety impact, *virtual safety assessment* has become a primary procedure due to its low cost and high efficiency compared to conventional field tests (Donà and Ciuffo, 2022; Cai et al., 2022; Szalay, 2023; Wu et al., 2025a). In this virtual paradigm, pre-crash scenarios—short time sequences describing driver, vehicle, and environmental dynamics leading up to a potential collision—are simulated in two conditions: baseline (without the DAS under assessment) and treatment (with the DAS). These pre-crash scenarios must align with the assessment objective and should encompass all relevant elements that may impact the performance of the technology under assessment (Wimmer et al., 2023). The safety effects of a system can be estimated by comparing the outcomes of the simulated baseline and treatment scenarios (Bärgman et al., 2017; Baron et al., 2020; Szalay, 2023; Wimmer et al., 2023).

The assessment procedure requires that the pre-crash scenarios used are adequate and accurately represent real-world conditions, to ensure a statistically sound and unbiased comparison (Donà and Ciuffo, 2022; Cai et al., 2022; Wimmer et al., 2023; Wu, 2024). However, real-world pre-crash data are typically limited in quantity and suffer from sampling bias and coverage issues. For example, naturalistic driving studies, such as the Second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS), provide extensive exposure to everyday driving behavior, recording millions of kilometers of real-world traffic interactions (Hankey et al., 2016). However, despite their large scale, these studies capture relatively few crashes, particularly high-severity ones, over-representing low-severity crashes (Bärgman et al., 2017; Wu et al., 2025a). In contrast, in-depth crash databases, such as the Crash Investigation Sampling System (CISS) in the United States and the German In-Depth Accident Study (GIDAS) Pre-Crash Matrix (PCM), provide detailed recorded or reconstructed kinematic data and associated behavioral and environmental factors, but represent only a small subset of crashes; high-severity crashes are over-represented (Schubert et al., 2017; Zhang et al., 2019).

To overcome these limitations, a common strategy is to create synthetic pre-crash scenarios using statistical and/or behavioral models derived from real-world data (Scanlon et al., 2021; Gambi et al., 2019; Wang et al., 2022; Wu et al., 2025a). This approach ensures the generation of a sufficient number of scenarios. However, a critical gap remains in validating the representativeness of the resulting synthetic scenarios. In partic-

*Corresponding author
Email address: jian.wu@chalmers.se (Jian Wu)

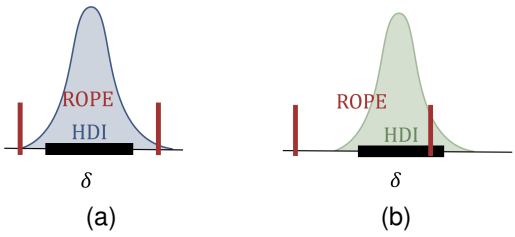


Figure 1: Illustration of the ROPE concept in Bayesian equivalence testing: (a) equivalence and (b) non-equivalence. The shaded curves represent the posterior distributions of a parameter or statistic, δ. Red vertical lines denote the ROPE boundaries, and black bars indicate the HDIs. Equivalence is supported when the HDI lies entirely within the ROPE.

ular, it is especially important to determine whether the synthetic scenarios are practically equivalent to their real-world counterparts for the intended assessment, meaning they are similar enough that any remaining differences are negligible in practice. Without proper validation, biases such as the over-representation of severe crashes may remain undetected, leading to misleading or biased assessments (Olleja et al., 2022; Bärgman et al., 2017; Hamdane et al., 2015).

Conventional statistical significance tests are commonly used as formal validation tools in traffic safety research (Lord and Mannering, 2010; Daamen et al., 2014). However, these methods are inherently difference-oriented: they are designed to detect differences rather than to establish equivalence (Anderson et al., 2000). Specifically, the absence of a statistically significant difference does not imply equivalence, nor does a statistically significant difference necessarily indicate a practically meaningful discrepancy in a given application context (Gibbs, 2013; Wu et al., 2025b).

To address this gap, a promising solution lies in practical equivalence testing. It is a set of statistical approaches aimed at determining whether two treatments, processes, or groups are sufficiently similar that any observed differences are small enough to be ignored in practice (Limentani et al., 2005; Greene et al., 2000). While such methods are well established in fields such as medicine and psychology (Lakens et al., 2018), their adoption in traffic safety and automated driving research remains limited to date.

Among these methods, Bayesian approaches based on the Region of Practical Equivalence (ROPE) (Schwaferts and Augustin, 2020) offer a straightforward interpretation of similarity and naturally integrate domain-specific thresholds (Kruschke, 2018; Wu et al., 2025b), making them particularly well-suited in virtual safety impact assessments of DAS. Rather than asking whether two samples originate from exactly the same distribution, ROPE-based methods assess whether the highest density intervals (HDIs) of posterior statistic estimates—such as the difference between means or variances—fall entirely within a predefined region of values considered practically equivalent. Equivalence is quantified as the probability that a parameter falls within a region of negligible difference, thereby providing an interpretable probabilistic statement of similarity that also integrates prior information. The underlying logic of this procedure is illustrated in Fig. 1, where the relative positions of the HDI and ROPE determine whether practical equivalence can be concluded.

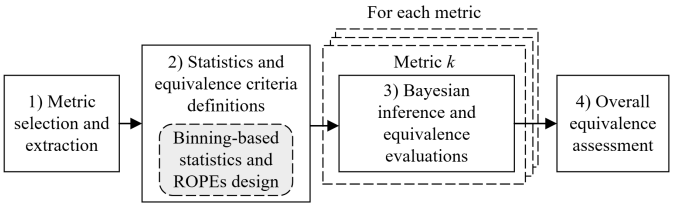


Figure 2: Flowchart of the proposed equivalence testing method. The gray component in Step 2 represents the part newly developed in this work.

Our earlier work proposed a Bayesian ROPE-based equivalence test framework (shown in Fig. 2) for validating synthetic pre-crash scenarios (Wu et al., 2025b). Step 2 of the framework is intentionally flexible, allowing users to define the statistics and equivalence criteria based on the specific needs of the assessment.

This study aims to provide clear and practical guidelines for implementing this step by introducing a novel approach that uses binning-based statistics and ROPEs designed in a transparent and interpretable manner (illustrated by the gray component in Fig. 2). The approach is illustrated through a case study, which tests the practical equivalence of two synthetic rear-end pre-crash datasets with a previously developed reference dataset (Wu et al., 2025a) in the context of the safety impact assessment for an Automatic Emergency Braking (AEB) system. The case study demonstrates how the framework can support both quantitative assessments of practical similarity and diagnostic analysis, such as identifying which regions of the data contribute most to non-equivalence. It also illustrates how sample weighting can mitigate sampling bias within a defined assessment scope, although its effectiveness depends on the underlying data structure and must be evaluated on a case-by-case basis.

Although rear-end pre-crashes are used in the case study, the validation framework is inherently generic and can be readily extended to contexts beyond safety impact assessments, such as model validation in traffic flow simulation, comparative evaluation of behavioral models, or assessment of synthetic data used for system testing. By providing an interpretable and probabilistically sound foundation for assessing practical equivalence, the framework is a flexible tool for the systematic validation of synthetic data against a reference.

2. Methodology

This section describes the extended Bayesian ROPE-based practical equivalence testing framework. Its primary objective is to assess whether a synthetic pre-crash dataset is practically equivalent to a representative reference dataset for an intended safety impact assessment.

2.1. Proposed Equivalence Testing Method

The proposed equivalence testing method contains the four steps shown in Fig. 2. Pre-crash data typically consist of time series outlining the dynamics and trajectories of (at a minimum) the road users involved in the crash. Because direct comparison of time-series data often is impractical, they are characterized by a set of derived variables, referred to here as "metrics". In this study, metrics are the quantitative variables used to characterize a scenario for equivalence testing. These may include scenario-descriptive parameters such as initial speeds and relative positions, obtained either directly from the time series or through a parameterization of the scenario, as well as outcome-based quantities such as Delta-v or estimated injury risk (when relevant to the safety impact assessment). The framework does not depend on a specific parameterization; it accommodates any set of metrics deemed meaningful for evaluating representativeness in the intended assessment context.

2.1.1. Step 1: Metric selection and extraction

This step involves selecting the metrics most relevant to the intended assessment and extracting or deriving them from both the reference and synthetic datasets. The reason for selecting only the most relevant metrics is that, while using more metrics can make the test more comprehensive, it also heightens the risk of making a false non-equivalence error. In such cases, non-equivalence in less relevant metrics may dominate the validation outcome and lead to an erroneous rejection of an otherwise representative dataset. Therefore, the assessment only considers the selected metrics.

2.1.2. Step 2: Statistics and equivalence criteria definitions

For each metric, two statistics, θ and Θ, are defined to assess the differences between its distributions in the two compared datasets. (More details about these statistics will be provided in Section 2.2.)

Subsequently, practical equivalence criteria are established for the statistics by specifying both the posterior probability threshold (α) and ROPEs ($[0, \theta_{\text{thd}}]$ and $[0, \Theta_{\text{thd}}]$, where θ_{thd} and Θ_{thd} are ROPE thresholds). Typically, a 95% posterior probability threshold is selected, and the ROPE is determined based on expert judgment and domain-specific knowledge relevant to the application. For a metric to be deemed practically equivalent between datasets, the 95% HDIs of both statistics' posterior distributions must lie entirely within the defined ROPEs.

2.1.3. Step 3: Bayesian inference and equivalence evaluations

For each metric, a set of Bayesian distribution models (e.g., exponential, normal, log-normal, gamma, or mixture models) is chosen based on expert judgment. The models are fitted separately to empirical distributions of the metric in reference and synthetic datasets. If the datasets are weighted, sample weights are incorporated directly into the Bayesian model through a weighted likelihood formulation. Specifically, each observation contributes to the total log-likelihood in proportion to its assigned weight, ensuring that the posterior distribution reflects the representativeness of the weighted data. Leave-one-out cross-validation (Vehtari et al., 2017) is employed to select the optimal distribution model for each dataset.

Posterior samples, each representing a draw of the predictive distribution parameters, are obtained from the optimal Bayesian models fitted to the two datasets. The two binning-based statistics θ and Θ are computed for each paired set of posterior samples from both distribution models. Finally, the HDI of each statistic's posterior distribution is evaluated against the practical equivalence criteria defined in Step 2. If each HDI lies entirely within its corresponding ROPE, the data support the practical equivalence of the metric between the compared datasets. The Python implementation of the Bayesian distribution fitting and binning-based statistics computation functions used in this study is publicly available in the `bayes-binned-equivalence` repository (Wu, 2026).

2.1.4. Step 4: Overall equivalence assessment

In the previous steps, multiple equivalence tests have been conducted for the selected metrics. This step aims to establish clear overall equivalence criteria that synthesize these individual results into a single conclusion. The criteria should effectively integrate the outcomes of individual tests with expert-based weighting of the relative importance of different metrics, ensuring that the synthetic dataset adequately represents the real-world data in the aspects most relevant to the intended assessment.

In theory, two datasets are considered equivalent only if all metrics individually demonstrate equivalence. However, in practice, less stringent approaches may be justified. For instance, overall equivalence may still be concluded if equivalence is achieved for a subset of the selected metrics—specifically those identified as most critical within the already relevance-filtered set—provided that the remaining less-critical metrics do not exhibit substantial deviations.

Two additional considerations are important here. First, the overall equivalence criteria should be clearly defined in advance. If applicable, the definitions should also include a quantitative explanation of what constitutes a "substantial deviation" for less-critical metrics. Importantly, the rationale for adopting less stringent criteria should be based on the trade-off between strictness and inclusiveness. As the number of evaluation metrics increases, so does the likelihood that at least one of them will fail to meet the equivalence criteria. Allowing some flexibility for less critical metrics enables a more holistic and practically meaningful judgment of equivalence, supported by a transparent understanding of where and how the synthetic data diverges from the reference data.

Second, it is essential to emphasize that practical equivalence testing relies heavily on expert judgment and reasoning, rather than purely numerical outcomes. It is generally not possible to define practical equivalence without making decisions based on some level of expertise. Therefore, the rationale for declaring equivalence or non-equivalence must be thoroughly documented, including the motivations behind the selection of metrics, statistics, ROPEs, and overall equivalence criteria.

2.2. Two Binning-Based Statistics

The two proposed statistics, θ and Θ, are designed to quantify practical differences between the (one-dimensional) reference and synthetic distributions within a specific safety impact assessment context. These statistics are derived through a binning process that strikes a balance between interpretability and robustness. Interpretability is achieved by decomposing distributional differences into deviations across explicit, localized regions of the metric's range, which can be directly related to assessment-relevant regimes. Robustness is achieved by aggregating data within bins, thereby reducing sensitivity to small-scale noise (random fluctuations that do not carry meaningful information for the assessment) while still capturing systematic differences (consistent and structured deviations across the distribution that reflect genuine discrepancies between the datasets).

The reference distribution is first partitioned into N bins, each containing (approximately) the same proportion of the total data. The same bin boundaries are then applied to the synthetic distribution, resulting in paired bin proportions that enable a direct, one-to-one comparison between the two distributions.

It is important to note that not all bin groups are equally relevant for the intended safety impact assessment. The practical relevance of each bin depends on the specific objective of the assessment. For instance, under many safety-impact objectives, bins corresponding to safety-critical ranges of a metric or to conditions in which the assessed system performs poorly should have a greater influence on the overall assessment outcome. Bin weights can reflect this heterogeneity in practical relevance. They link statistical deviations across bins with their potential impact on assessment results, thereby bridging the gap between purely statistical differences and practically meaningful consequences. Bin weighting thus constitutes a key component of the proposed method. Details regarding the bin weighting design and implementation are provided in Section 2.3.

The two proposed statistics, θ and Θ, integrate the weighted deviations across bins to quantify the overall degree of practical difference between the reference and synthetic distributions. Let ω_i denote the weight of the i-th bin and $P_{\text{ref},i}$ and $P_{\text{syn},i}$ denote the proportions of data in the i-th bin for the reference and synthetic distributions, respectively. $\Delta P_i (= P_{\text{syn},i} - P_{\text{ref},i})$ denotes the proportion difference for the i-th bin between two distributions. Then, the two statistics θ and Θ are defined as:

$$\theta = \max_{1 \le i \le N} \left(\left| \frac{\Delta P_i}{P_{\text{ref},i}} \right| \cdot \omega_i \right), \tag{1}$$

$$\Theta = \sum_{i=1}^{N} |\Delta P_i| \cdot \omega_i. \tag{2}$$

The maximum weighted absolute relative deviation across bins (i.e., a worst-case perspective) is captured by θ, whereas Θ represents the total weighted absolute deviation (i.e., an aggregate perspective). Together, they provide complementary views of the practical distributional difference between the synthetic and reference data.

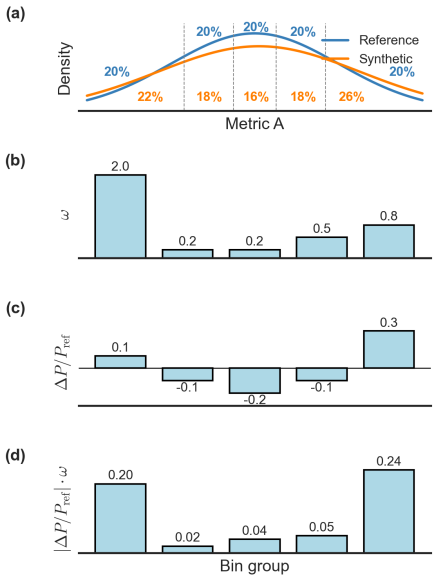


Figure 3: Illustration of the procedure to compute θ.

An illustration of the procedure to compute θ for the reference and synthetic distributions is shown in Fig. 3.

1. Fig. 3(a) illustrates the binning process used in the computation of θ and Θ. The reference distribution (blue) is divided into five quantile-based bins, each containing 20% of the reference data. The same bin boundaries are then applied to the synthetic distribution (orange), resulting in bin proportions of 22%, 18%, 16%, 18%, and 26%.
2. The weights ω (illustrated in Fig. 3(b) as 2.0, 0.2, 0.2, 0.5, and 0.8) represent the relevance of each bin to the intended assessment. The method for setting these weights is described in Section 2.3.
3. The relative deviations $\Delta P/P_{\text{ref}}$ (illustrated in Fig. 3(c) as 0.1, -0.1, -0.2, -0.1, and 0.3) measure the divergence between the compared bin pairs in the reference and synthetic distributions.
4. The weighted absolute relative deviations $|\Delta P/P_{\text{ref}}| \cdot \omega$ (illustrated in Fig. 3(d) as 0.20, 0.02, 0.04, 0.05, and 0.24) then scale the statistical divergence by the practical relevance of each bin, represented by the corresponding weight illustrated in Fig. 3(b), emphasizing differences in high-weight regions. In this example, the last bin shows the maximum weighted absolute relative deviation, which determines the value of θ. Accordingly, $\theta = 0.24$.

It is essential to note that the bin boundaries and bin weights are intrinsic components of the definitions of θ and Θ. As described in Section 2.1.3, these statistics are computed for each paired set of posterior samples from the Bayesian models fitted to the two datasets being compared. Consequently, for every posterior draw, the bin boundaries and the corresponding bin weights are recalculated based on the posterior samples from the reference distribution model.

2.3. Bin Weights

We propose an empirical procedure for determining bin weights:

1. Re-simulate the reference pre-crash scenarios with a virtual representation of the DAS under assessment to generate system-specific outcome data for each scenario. The result may be a sparse distribution due to the limited size of the reference dataset.
2. Identify a set of re-simulation outcome measures most relevant to the assessment purpose, such as residual impact speed or estimated injury risk, denoted as $[x_1, x_2, \ldots, x_l]$.
3. Define a weight function $f(X)$ that assigns a bin weight based on these outcome measures:

$$\omega_i = f(X_i), \tag{3}$$

where ω_i is the weight of the i-th bin, and $X_i = [x_1^{(i)}, x_2^{(i)}, \ldots, x_l^{(i)}]$ represents the selected outcome measures for the i-th bin.

The choice of outcome measures and the design of $f(X)$ should be guided by domain knowledge and expert judgment in addition to the assessment objective, to ensure that the resulting bin weights accurately capture the practical relevance of the corresponding bins. By combining statistical deviations between datasets with bin-level weights that reflect their importance within the assessment scope, the two statistics effectively quantify the practical differences between the compared datasets for the intended assessment. An example of the bin-weight function is presented in Section 3.3.

2.4. Parameter Settings

The parameters N, α, θ_{thd}, and Θ_{thd}, described below, influence the behavior and sensitivity of the proposed testing method.

The choice of the number of bins N is particularly critical, since it directly determines the granularity of the comparison between the reference and synthetic distributions, and consequently affects the evaluation of the two statistics θ and Θ. A larger N yields a more stringent test by enabling finer-grained comparisons, thereby improving the ability to detect localized differences. However, increasing N also heightens the sensitivity to variance and amplifies the cumulative effect of multiple comparisons. In practical terms, using more bins increases the likelihood that at least one bin will fail to meet the equivalence criteria, and thus that the overall test will conclude the datasets are not equivalent. Conversely, a smaller N reduces noise but may fail to catch localized distributional differences.

To balance robustness and granularity, we adopt a simple, practical guideline for selecting N: each bin should contain a sufficient number of reference data points to ensure stable proportion estimates and reliable equivalence evaluation. Further, the number of bins should not exceed some predefined threshold to avoid the cumulative effect of too many comparisons. Accordingly, the number of bins is determined as:

$$N = \min\left(\lfloor \frac{n}{m} \rfloor, N_{\max}\right), \tag{4}$$

where n is the number of reference samples, m is the desired minimum number of samples per bin, and $N_{\max}$ is the maximum allowable number of bins. Following conventional statistical recommendations based on the Central Limit Theorem, we suggest $m \in [30, 50]$. To control test sensitivity and mitigate multiple-comparisons effects, we recommend capping the number of bins at $N_{\max} = 20$.

The remaining three parameters α, θ_{thd}, and Θ_{thd} jointly define the practical equivalence criteria. Conventionally, α is set to 0.95. The parameters θ_{thd} and Θ_{thd} are thresholds shaping the ROPEs for the statistics θ and Θ, respectively. These thresholds should be set based on expert judgment to reflect the relevance and priorities of the intended assessment. A practical example is provided in Section 3.4.

It is important to note that a common pair of ROPE thresholds can be applied across different metrics. This is possible because, although the bins are partitioned separately based on the reference distribution of each metric, the bin weights are consistently determined by the same weight function, ensuring comparability of the two statistics across metrics.

3. Demonstration

This study demonstrates the proposed practical equivalence testing method by comparing two rear-end pre-crash scenario datasets with a reference dataset in the context of a safety impact assessment for an AEB system (Hay et al., 2025) developed within the V4SAFETY Project (V4SAFETY, 2022).

3.1. Datasets

3.1.1. Reference dataset

The first dataset comprises 200 rear-end pre-crash scenarios randomly sampled with replacement from the QUADRIS pre-crash dataset (Wu, 2025). The full QUADRIS dataset contains 5,000 weighted rear-end pre-crash scenarios generated by modeling real-world rear-end pre-crash data and is designed to represent the U.S. rear-end crash population across the full severity range, from minor physical contact to severe injuries and fatalities (Wu et al., 2025a). We consider the QUADRIS dataset to be the most comprehensive representation to date of U.S. rear-end crashes across all severity levels. Because real-world pre-crash datasets are typically much smaller than the full QUADRIS dataset, a random sample of 200 scenarios was drawn to approximate realistic dataset sizes; this subset is treated as the "reference" dataset in the present study.

3.1.2. PCM dataset

The second dataset consists of 866 reconstructed concrete rear-end pre-crash scenarios from the GIDAS-PCM dataset (Schubert et al., 2017; Wu et al., 2025b), initiated in 2011. Hereafter referred to as the "PCM" dataset, it is a dedicated subset of the GIDAS database, which collects on-scene accident investigations involving personal injury in Hannover and Dresden, Germany (Schubert et al., 2017).

Unlike the reference dataset, which spans the full severity range, the PCM dataset includes only injury-involved crashes,

with lower inclusion probabilities for less severe injuries (Wu et al., 2025b). To mitigate this known bias, a weighted version of the PCM dataset was created, as detailed in Section 3.5.1. Both the raw and weighted PCM datasets were compared with the reference dataset.

3.1.3. SCM-based dataset

We also wanted to include a dataset of rear-end pre-crash scenarios generated via traffic simulation, an approach that contrasts with the crash reconstruction approach used to create the PCM dataset (Shah et al., 2018; Baron et al., 2020; Fries et al., 2022). Ideally, such a dataset would be generated either by simulating everyday multi-agent traffic and extracting crash events as they naturally occur or by scenario-based simulation, where selected critical scenarios are simulated repeatedly, and the resulting outcomes are weighted by their real-world occurrence frequencies. However, to our knowledge, no such dataset is currently available. Therefore, we created one for demonstration purposes, based on the QUADRIS dataset and the Stochastic Cognitive Model (SCM) (Fries et al., 2022; BMW Group).

This dataset consists of 7,888 synthetic rear-end pre-crash scenarios that were obtained by re-simulating each of the 5,000 pre-crash scenarios (hereafter referred to as the "seed cases") in the full QUADRIS dataset 2,000 times. The lead-vehicle behavior and initial following distance were fixed as in the seed cases, and the following vehicle was governed by the SCM.

The SCM has been under development by BMW and partners since 2014 and is used in the openPASS simulation framework (Fries et al., 2022). By modeling cognitive processes, the SCM aims to ensure a realistic representation of the traffic in the simulation across a wide range of scenarios, including possible collision scenarios (BMW Group). Given the same input, SCM generates various behavioral outcomes by sampling driver perceptions and actions from probabilistic distribution models; thus, the 2,000 re-simulation outcomes of a single seed case are not identical. Only around 0.08% of all the simulations resulted in a crash. Of the 5,000 seed cases, 651 generated at least one re-simulated crash, with the number of crashes per seed case ranging from 0 to 959 out of the 2,000 simulations.

For each seed case, the number of re-simulated crashes depends on the number of simulation runs and the SCM crash probability in that scenario. Ideally, the number of simulation runs should be chosen—or the resulting crashes weighted—in proportion to the real-world occurrence frequency of the corresponding seed case. Formally,

$$n_{\mathrm{sim},i} \cdot \omega_{\mathrm{sim},i} \propto f_i, \tag{5}$$

where $n_{\mathrm{sim},i}$ is the number of simulations conducted for the i-th seed case, $\omega_{\mathrm{sim},i}$ is the weight assigned to all re-simulated crashes originating from that seed case, and f_i denotes the real-world occurrence frequency of the i-th seed case scenario. This condition ensures that the weighted dataset of all re-simulated pre-crash scenarios reflects the true prevalence of seed-case types in real traffic, thereby preventing the over-representation of seed cases that are frequently simulated but rare in reality.

However, the real-world frequencies f_i of the seed case scenarios are not available. Therefore, we utilize a property of the QUADRIS dataset: the weight of each pre-crash scenario is proportional to the real-world frequency of that crash type. Formally,

$$\omega_i \propto f_i \cdot p_{\mathrm{c},i}, \tag{6}$$

where ω_i is the weight of the i-th seed case in the QUADRIS dataset and $p_{\mathrm{c},i}$ denotes the crash probability of human drivers in that scenario. Combining (5) and (6) gives:

$$n_{\mathrm{sim},i} \cdot \omega_{\mathrm{sim},i} \cdot p_{\mathrm{c},i} \propto f_i \cdot p_{\mathrm{c},i} \propto \omega_i. \tag{7}$$

If we assume that, in seed cases with at least one re-simulated crash, SCM has the same crash probability as human drivers, that is:

$$\hat{p}_{\mathrm{c,SCM},i} = p_{\mathrm{c},i} \quad \text{for seed cases with } \hat{p}_{\mathrm{c,SCM},i} > 0, \tag{8}$$

then we obtain:

$$n_{\mathrm{sim},i} \cdot \omega_{\mathrm{sim},i} \cdot \hat{p}_{\mathrm{c,SCM},i} \propto \omega_i \quad \text{for } \hat{p}_{\mathrm{c,SCM},i} > 0, \tag{9}$$

where $\hat{p}_{\mathrm{c,SCM},i}$ is the observed SCM crash probability for the i-th seed case, estimated from the 2,000 simulation runs.

Rearranging (9) yields the weight for re-simulated crashes:

$$\omega_{\mathrm{sim},i} \propto \frac{\omega_i}{n_{\mathrm{sim},i} \cdot \hat{p}_{\mathrm{c,SCM},i}} = \frac{\omega_i}{n_{\mathrm{sim,c},i}} \quad \text{for } \hat{p}_{\mathrm{c,SCM},i} > 0, \tag{10}$$

where $n_{\mathrm{sim,c},i}$ is the number of re-simulated crashes originating from the i-th seed case.

Finally, the weights for all re-simulated crashes were computed using (10) and scaled so that their sum equals the total number of collected crashes (7,888). The resulting weighted dataset of 7,888 scenarios, hereafter referred to as the "SCM-based" dataset, is compared with the reference dataset.

It is important to note that the SCM-based dataset represents only a finite sampling of the SCM's stochastic behavior, which introduces several limitations. First, seed cases that yield zero re-simulated crashes are effectively treated as having zero crash probability and are therefore excluded, even though the absence of observed crashes may simply be a finite-sample effect. Second, for seed cases with one or more observed crashes, the weighting procedure relies on the strong assumption that the empirically observed SCM crash probability is a reasonable approximation of the corresponding real-world crash probability. As a consequence, both the set of included seed cases and their associated weights are sensitive to the number of simulation runs and may change with additional re-simulations. While increasing the number of runs would yield more stable and robust estimates of SCM-induced pre-crash behavior, the total number of re-simulations was constrained by the substantial computational effort. Given these limitations and because the primary aim is to illustrate the proposed equivalence-testing framework, the SCM-based dataset is used here primarily for demonstration purposes, rather than to assess the validity of the SCM for crash generation.

3.2. Metrics

In this demonstration, the general scope is to assess the impact of an AEB system on Maximum Abbreviated Injury Scale

(MAIS) 2+ injuries (Gennarelli and Wodzin, 2006) in rear-end pre-crash scenarios. Following our previous study (Wu et al., 2025b), we selected three of the previously used metrics and added one new metric, replacing the lead vehicle's Delta-v with its (estimated) injury risk (P_{inj}). They provide a balanced basis for comparing reference/real-world and synthetic pre-crash scenario datasets in the safety impact assessment context.

- P_{inj}: Estimated MAIS 2+ injury risk for the lead vehicle's driver. This metric represents the severity of potential crash outcomes and serves as a severity-weighted measure for comparing scenario datasets.
- t_{nr} [s]: No-return time. It is defined as the point of no return beyond which a collision is unavoidable even if the following vehicle applies the maximum deceleration of -9 m/s^2. Time zero corresponds to the impact moment; thus, t_{nr} is negative. This metric measures scenario criticality and the temporal margin available for evasive action.
- $a_{\text{l,min}}$ [m/s^2]: Minimum acceleration (maximum deceleration) of the lead vehicle. This value characterizes the harshness of the lead vehicle's braking maneuver, a common contributing factor in rear-end crashes. Capturing its distribution helps ensure that synthetic scenarios, to some extent, reflect realistic lead-vehicle deceleration behavior as it is relevant to AEB performance.
- $a_{\text{f,min}}$ [m/s^2]: Minimum acceleration of the following vehicle. Capturing its distribution helps ensure that synthetic scenarios reflect, to some extent, the realistic deceleration behavior of the following vehicle.

The four metrics collectively capture outcome severity (P_{inj}), scenario criticality (t_{nr}), and braking responses for the leading and following vehicles ($a_{\text{l,min}}$ and $a_{\text{f,min}}$, respectively). With this selection, practical equivalence testing considers factors that are physically meaningful and relevant to the assessment scope for which the synthetic scenarios are intended. Note again that this is an example and that specific assessment scopes may need to capture other features of the pre-crash kinematics.

3.3. Bin-Weight Function

An empirical approach for designing bin weights is described in Section 2.3. In this approach, the pre-crash scenarios in the reference dataset are first re-simulated with a specific DAS system (e.g., AEB in this work). A set of re-simulation outcome measures most relevant to the assessment is then selected and used to define a bin-weight function based on prioritization judgment and domain knowledge of the assessment scope.

In the context of safety impact assessments, relevant outcome measures may include the re-simulated crash rate, the average re-simulated injury risk of the lead-vehicle driver, the reduced average re-simulated injury risk of the lead-vehicle driver, or other comparable measures.

In this demonstration, the weight of the i-th bin is defined as:

$$\omega_i = f(\bar{P}_{\text{inj,rs},i}) = \frac{\bar{P}_{\text{inj,rs},i} + \varepsilon}{P_0 + \varepsilon}, \tag{11}$$

where $\bar{P}_{\text{inj,rs},i}$ denotes the average probability of the lead-vehicle driver sustaining a MAIS 2+ injury for the re-simulated scenarios in the i-th bin group, P_0 is a pre-defined baseline MAIS 2+ injury risk, and ε is a small positive constant introduced to ensure that the minimum possible weight is above zero. A hypothetical baseline bin, defined as having a standard weight $\omega_{\text{b}} = 1$, corresponds to the baseline injury risk P_0 (i.e., $\omega_i = 1$ when $\bar{P}_{\text{inj,rs},i} = P_0$). Note that, in this study, the MAIS 2+ injury risk of the lead-vehicle driver in non-crash cases is set to zero, and in crash cases it is computed using the model proposed by Wang (Wang, 2022):

$$P_{\text{inj}} = \frac{1}{1 + e^{6.1818 - 0.3315\Delta v_{\text{l}}}}, \tag{12}$$

where Δv_{l} (m/s) is the speed change of the lead vehicle during the impact, estimated as described in Section III-F of our previous study (Wu et al., 2025a).

With this bin-weight function, higher weights are assigned to scenarios with higher average re-simulated injury risks, thereby placing a greater emphasis on conditions in which the AEB system is less effective. Conversely, scenarios associated with lower average injury risks receive smaller weights, reflecting their comparatively limited influence on the assessment. This ranking aligns with the objective of safety impact assessment: to prioritize the accurate representation of safety-critical scenarios that most influence the estimated overall safety benefit of the system.

We set $\varepsilon = 1 \times 10^{-4}$ and $P_0 = 0.02$. The choice of P_0 is motivated by conventional practice in injury risk modeling; a MAIS 2+ probability below 2–5% is regarded as the lower bound of clinically meaningful injury risk (Forman et al., 2012). Anchoring at 2% avoids over-weighting near-zero cases, while ensuring that bins with higher risks are emphasized relative to a conservative but nontrivial reference point. This value represents what is generally considered a meaningful baseline (Forman et al., 2012). Note that the baseline is primarily used to scale the bin weights; one can choose any baseline. Yet, in general, a baseline with an interpretable physical meaning is recommended, as it is more straightforward to set realistic ROPE thresholds accordingly (see Section 3.4).

3.4. Parameters

The four parameters N, α, θ_{thd}, and Θ_{thd}, described in Section 2.4, need to be specified beforehand. Considering the small sample size $n_{\text{ref}} = 200$ of the reference dataset, we set $m = 40$, thereby obtaining the number of bins $N = 5$.

The remaining three parameters jointly define the equivalence criteria. In this demonstration, each metric has to satisfy the equivalence criteria individually in order to establish practical equivalence. For each metric, the posterior probability threshold is set to $\alpha = 0.95$. A common pair of ROPE thresholds (θ_{thd}, Θ_{thd}) is applied across all metrics, ensuring comparability of the equivalence decisions (see Section 2.4).

We propose a practical rule for setting ROPEs: specify tolerances for a hypothetical baseline bin and derive ROPEs from them. As mentioned in Section 3.3, the baseline bin is assigned

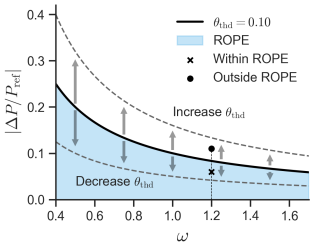


Figure 4: Illustration of the ROPE for θ. The x-axis represents the weight of the bin with the maximum weighted absolute relative deviation, while the y-axis represents the absolute relative deviation. The cross marker at (1.2, 0.06) lies within the ROPE, whereas the circle marker at (1.2, 0.11) lies outside. The solid black curve corresponds to $\theta_{\text{thd}} = 0.10$, with the shaded region representing the ROPE

a weight of $\omega_{\text{b}} = 1$. The user should first interpret what this baseline condition represents in practical terms according to the chosen bin-weight function, ensuring that the tolerance specified is meaningful for the specific assessment. The next step is to decide the acceptable range of deviations. The absolute relative deviation $|\Delta P/P_{\text{ref}}|$ and the absolute deviation $|\Delta P|$ (based on the definitions of θ and Θ) should be regarded as practically equivalent for the hypothetical baseline bin. The tolerances are denoted as $|\Delta P/P_{\text{ref}}|_{\text{thd}}$ and $|\Delta P|_{\text{thd}}$.

In our demonstration, the hypothetical baseline bin corresponds to an average baseline MAIS 2+ injury risk $P_0 = 0.02$. We assigned tolerances of $|\Delta P/P_{\text{ref}}|_{\text{thd}} = 10\%$ and $|\Delta P|_{\text{thd}} = 5\%$ to this bin. These values mean that for the baseline bin, its proportion in the synthetic distribution P_{syn} must satisfy the two conditions:

$$0.9P_{\text{ref}} \leq P_{\text{syn}} \leq 1.1P_{\text{ref}}, \tag{13}$$

$$P_{\text{ref}} - 0.05 \leq P_{\text{syn}} \leq P_{\text{ref}} + 0.05. \tag{14}$$

Differences within ±10% relative deviation or ±5% absolute deviation are considered practically negligible and too small to materially influence the intended safety impact assessment.

Once these tolerances are defined, the corresponding ROPE thresholds for θ and Θ can be derived directly from (1) and (2) as:

$$\theta_{\text{thd}} = |\Delta P/P_{\text{ref}}|_{\text{thd}} \cdot \omega_{\text{b}} = |\Delta P/P_{\text{ref}}|_{\text{thd}}\,, \tag{15}$$

$$\Theta_{\text{thd}} = |\Delta P|_{\text{thd}} \cdot \omega_{\text{b}} = |\Delta P|_{\text{thd}}\,. \tag{16}$$

Based on the chosen tolerances, the ROPE thresholds are thus $\theta_{\text{thd}} = 0.10$ and $\Theta_{\text{thd}} = 0.05$. A 10% localized deviation tolerance ensures sensitivity to meaningful discrepancies in critical bins, while a 5% aggregate tolerance maintains a conservative standard for overall equivalence. This balance follows the general principle that ROPE boundaries should be defined based on domain-relevant effect sizes (that is, differences deemed practically meaningful for the assessment purpose) rather than on arbitrary statistical thresholds (Kruschke, 2018; Schwaferts and Augustin, 2020). In summary, this practical procedure offers an intuitive approach to translating user-specified practical tolerances into consistent formal ROPE thresholds that can be applied across bins.

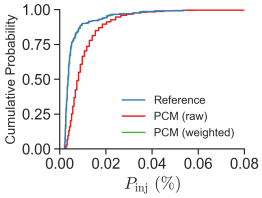


Figure 5: Cumulative distribution functions (CDFs) of the injury risk P_{inj} for the reference dataset (blue), the raw PCM dataset (red), and the weighted PCM dataset (green). The green curve closely overlaps with the blue one.

Fig. 4 illustrates the defined ROPE for θ. The arrows and dashed gray curves demonstrate how variations in θ_{thd} affect the ROPE: an increase in θ_{thd} results in a larger ROPE, while a decrease results in a smaller ROPE. At $\omega = 1.2$, the threshold for $|\Delta P/P_{\text{ref}}|$ is $\theta_{\text{thd}}/\omega = 0.10/1.2 \approx 0.083$, indicating that the cross marker at (1.2, 0.06) lies within the ROPE, whereas the circle marker at (1.2, 0.11) lies outside. This visualization illustrates that the equivalence decision is sensitive to the definition of tolerances, underscoring the importance of transparent justification when setting thresholds.

3.5. Results

3.5.1. Weighted PCM dataset

As noted in Section 3.1, the PCM dataset includes only pre-crash data that resulted in the injury of at least one person. To mitigate this limitation, we generated a weighted version of the dataset using the k-nearest neighbors (KNN)-based sample weighting method developed in our previous study (Wu et al., 2025a), aligning the injury risk distribution with that of the reference dataset. Figure 5 shows the cumulative distributions of P_{inj} for the reference, raw PCM, and weighted PCM datasets, illustrating how the weighting procedure effectively corrects for the sampling bias in this application. Equivalence tests were then performed to compare both the raw and weighted PCM datasets against the reference dataset.

3.5.2. Re-simulation

As described in Section 3.3, to calculate bin weights, the pre-crash scenarios in the reference dataset are re-simulated with the AEB system; of the 200 crashes, 41 remain. Fig. 6 shows the re-simulation results: higher re-simulated injury risks are associated with greater baseline injury risk, shorter no-return times, and harsher braking maneuvers. This information provides the basis for assigning larger bin weights to more critical scenarios in the equivalence tests. The weights are used to determine the two statistics θ and Θ for each metric.

3.5.3. Statistics computation and diagnostic insights

As described in Section 2.1, Bayesian distribution models were fitted to the reference and synthetic datasets for each metric. The two statistics θ and Θ were then computed for every paired draw from the posterior distributions of the optimal models. Finally, the 95% HDI of each statistic's posterior distribution was compared with the corresponding ROPE.

Table 1: Statistics Across Metrics and Comparison Types

Metric	Statistic	ROPE	Raw PCM 95% HDI	Equivalence	Weighted PCM 95% HDI	Equivalence	SCM-based 95% HDI	Equivalence
P_{inj}	θ	[0, 0.10]	[0.48, 0.58]	No	[0.00, 0.06]	Yes	[0.26, 0.41]	No
	Θ	[0,0.05]	[0.10, 0.12]	No	[0.00, 0.01]	Yes	[0.05, 0.09]	No
$a_{\text{l,min}}$	θ	[0, 0.10]	[0.02, 0.05]	Yes	[0.02, 0.05]	Yes	[0.22, 0.30]	No
	Θ	[0, 0.05]	[0.01, 0.02]	Yes	[0.01, 0.02]	Yes	[0.05, 0.07]	No
$a_{\text{f,min}}$	θ	[0, 0.10]	[0.07, 0.17]	No	[0.02, 0.09]	Yes	[0.21, 0.34]	No
	Θ	[0, 0.05]	[0.02, 0.04]	Yes	[0.01, 0.02]	Yes	[0.07, 0.08]	No
t_{nr}	θ	[0, 0.10]	[0.30, 0.39]	No	[0.00, 0.05]	Yes	[0.04, 0.20]	No
	Θ	[0, 0.05]	[0.07, 0.08]	No	[0.00, 0.01]	Yes	[0.01, 0.05]	Yes

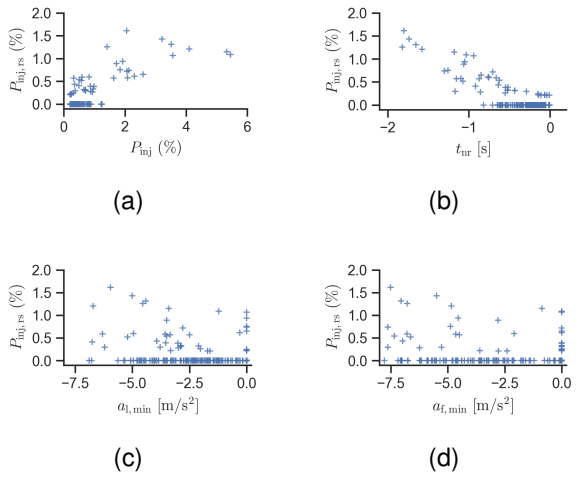


Figure 6: Re-simulated injury risk of the lead-vehicle driver against four metrics in the reference dataset: (a) P_{inj}, (b) t_{nr}, (c) $a_{l,\text{min}}$, and (d) $a_{f,\text{min}}$.

Fig. 7 provides an example of a single instance of θ for the injury risk metric P_{inj} and the lead vehicle's minimum acceleration $a_{l,\text{min}}$. The bin weights are identical across the raw and weighted PCM datasets as well as the SCM-based dataset, since they are determined by the same draw from the posterior distributions of the fitted model for the reference data. In this instance, the highest weights are assigned to the bins with the highest P_{inj} and the lowest $a_{\text{l,min}}$, which reflect the scenarios in which the AEB system is least effective at mitigating crashes.

$\Delta P/P_{\text{ref}}$ measures the relative deviation across bins between the paired distribution draws. $|\Delta P/P_{\text{ref}}| \cdot \omega$ then scales this deviation by the importance of the corresponding bin. Finally, θ is defined as the maximum of the weighted deviations across all bins; it captures the most critical (localized) distributional discrepancy, considering its magnitude as well as its relevance to the intended assessment. The computation of Θ follows the same binning process, except that instead of choosing the maximum weighted absolute relative deviation, it summarizes the weighted absolute deviations across all bins.

For diagnostic purposes, the weighted relative and absolute deviations of each bin ($|\Delta P_i/P_{\text{ref},i}| \cdot \omega_i$ and $|\Delta P_i| \cdot \omega_i$, respectively) can also be recorded and analyzed. The contribution of each bin to each statistic is useful information which can be utilized to enhance dataset generation or weighting strategies. For example, it may lead to improved sampling procedures, or more refined generative models which better represent the distributional characteristics of those bins with a highly weighted divergence.

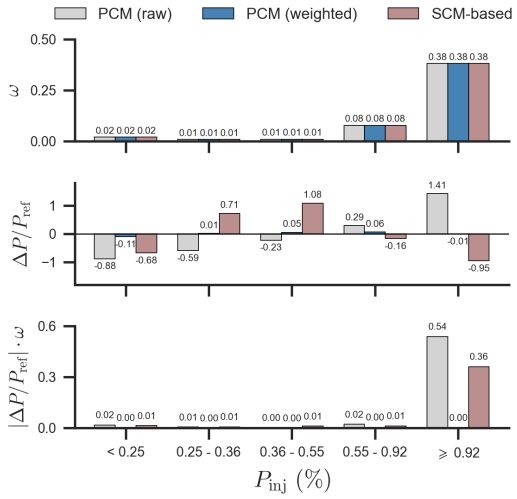
For the specific case illustrated in Figure 7a, for metric P_{inj}, among the three datasets, the raw PCM dataset contains the highest proportion of high-severity scenarios than the reference dataset, resulting in $\theta = 0.54$. Compared to the raw PCM dataset, the weighted PCM dataset aligns closely with the reference dataset, yielding a much smaller $\theta = 0.00$. Meanwhile, the SCM-based dataset contains a lower proportion of high-severity scenarios, resulting in $\theta = 0.36$.

For the case illustrated in Figure 7b, for metric $a_{\text{l,min}}$, among the three datasets, the SCM-based dataset contains the highest proportion of scenarios with low $a_{\text{l,min}}$ values than the reference dataset, resulting in $\theta = 0.24$. This difference may be explained by the relatively small sample size; seed cases involving a harsh-braking lead vehicle are more likely to produce crash events even within the limited number of simulation runs, while crashes from seed cases with a lower crash probability may simply not appear due to finite-sample limitations. Nevertheless, we cannot rule out the possibility that part of the observed discrepancy arises from limitations of the SCM itself rather than sampling effects alone. In contrast, the raw and weighted PCM datasets align closely with the reference dataset, yielding much smaller values of θ.

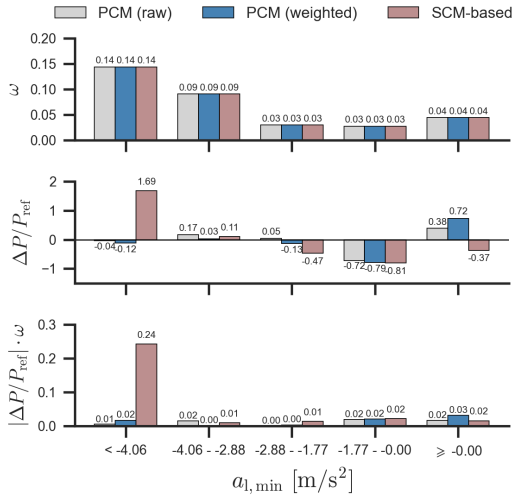
3.5.4. Equivalence testing results

Table 1 summarizes the results of the equivalence tests.

- **Raw PCM dataset (non-equivalence)**: The statistics for P_{inj} and t_{nr} both exceeded the ROPE thresholds, indicating systematic deviations from the reference dataset. In addition, $a_{\text{f,min}}$ failed the equivalence test due to the large proportion of harsh-braking events in the PCM data, which shifts the distribution toward stronger deceleration values. These results reflect the expected bias of the raw PCM dataset toward more severe cases; compared with the reference dataset, the raw PCM dataset exhibits higher injury risks (see Figure 7a), shorter no-return times, and stronger braking responses.



(a)



(b)

Figure 7: Illustrative examples of θ for (a) the injury risk metric P_{inj} and (b) the lead vehicle's minimum acceleration $a_{l,\text{min}}$. For (a) and (b), the top figures show the bin weights ω; the middle, the relative deviations $\Delta P/P_{\text{ref}}$; and the bottom, the weighted absolute relative deviations $|\Delta P/P_{\text{ref}}| \cdot \omega$ for one selected paired draw from the posterior distributions of the fitted models for the reference dataset and the dataset being compared.

- **Weighted PCM dataset (equivalence)**: In contrast, the weighted PCM dataset satisfied the equivalence criteria across all four metrics. This finding provides evidence that carefully designed weighting strategies can restore distributional comparability between biased and representative datasets. However, it is important to be aware that the result applies specifically to the application considered here; in general, weighting may not fully correct for structural biases, unobserved confounders, or differences in underlying data-generating mechanisms. The effectiveness of sample weighting should therefore be evaluated on a case-by-case basis.

- **SCM-based dataset (non-equivalence)**: None of the metrics satisfy the equivalence criteria. As shown in Fig. 7, the SCM-based dataset under-represents high-severity scenarios and over-represents cases involving a harsh-braking lead vehicle, perhaps due to the limited number of simulation runs per seed case.

4. Discussion and Conclusions

4.1. Contributions

A critical gap exists in validating the representativeness of the synthetic scenarios used for safety impact assessment of DAS. To ensure an accurate and credible assessment, it is essential to determine whether the synthetic scenarios are practically equivalent to their real-world counterparts for the intended assessment. However, existing validation practices in traffic safety, such as conventional statistical significance tests, are inadequate for this purpose. To address this gap, this study proposes and demonstrates an extension of the practical equivalence testing framework for validating synthetic pre-crash scenarios to assess the safety impact of DAS, building on our earlier conference work (Wu et al., 2025b). The extension introduces two binning-based statistics, θ and Θ, which quantify localized and aggregate distributional discrepancies between datasets, accounting for both the magnitude of the differences and their practical relevance. Practical relevance is incorporated by weighting the bins according to system-specific re-simulation outcomes, such as injury risk or crash rate, thereby emphasizing conditions that pose the greatest challenges to the DAS under assessment.

The case study yielded several insights. First, the raw PCM dataset failed equivalence tests due to its inherent bias toward higher-severity crashes. Second, applying a KNN-based reweighting procedure to that dataset resulted in equivalence across all metrics, demonstrating that appropriate weighting can mitigate bias in injury-focused datasets and align them with the reference dataset. Third, the SCM-based dataset failed most equivalence tests. This result should be interpreted cautiously, as it may primarily reflect finite-sample effects arising from the limited number of simulation runs, in which many seed cases exhibited zero observed crashes and were thus treated as having zero probability. Therefore, the present data do not permit a firm conclusion about the adequacy of the SCM; the apparent non-equivalence may be driven by insufficient sampling rather than model limitations. Finally, the examples demonstrate that the proposed statistics offer diagnostic insight into which bin groups drive non-equivalence, enabling targeted identification of where synthetic datasets diverge most critically from the reference dataset.

Beyond the specific application demonstrated here, the framework contributes methodologically by offering a structured, transparent approach to validating synthetic datasets that goes beyond conventional significance testing. By providing interpretable, relevance-weighted measures of practical difference, the framework supports both quantitative evaluation and diagnostic analysis in safety impact assessment.

4.2. Practical implications and recommendations

The proposed framework offers a structured workflow that practitioners can follow when validating synthetic datasets for safety impact assessment. By specifying assessment-relevant metrics and applying the binning-based statistics θ and Θ, users can obtain interpretable, decision-relevant measures with practical equivalence between synthetic and reference data. The two statistics offer complementary perspectives, enabling both localized and aggregate assessment of deviations across a wide range of validation contexts: θ highlights worst-case localized discrepancies, while Θ captures aggregate distributional differences. Importantly, because they operate on distributional representations of user-defined metrics rather than on scenario-specific assumptions, the two statistics are generic and agnostic to the specific application domain.

Before conducting formal equivalence testing, however, it is essential to perform basic sanity and plausibility checks on the synthetic scenarios. Sanity checks verify that key aggregate statistics fall within reasonable ranges, such as overall crash rates, proportions of crash types, and other high-level characteristics of the scenario set. These checks help detect implementation errors or unintended biases in the scenario generation process.

Plausibility checks then ensure the synthetic scenarios are physically and behaviorally reasonable, independent of any comparison with reference data. For example, basic kinematic quantities such as speed, acceleration, and jerk should fall within physically feasible and behaviorally realistic ranges. Scenarios that violate these fundamental constraints should be identified and addressed prior to equivalence testing, as they reflect modeling or simulation artifacts rather than issues of representativeness.

Once these sanity and plausibility checks are satisfied, equivalence testing should be used to assess representativeness relative to a reference dataset. At this stage, users should consider whether key scenario types are adequately represented and whether known dataset limitations, such as finite simulation runs or uncorrected sampling biases, may lead to non-equivalent outcomes. In such cases, failing an equivalence test does not necessarily indicate deficiencies in the underlying behavioral or system models; instead, it may reflect limitations in the data generation or sampling process (see the SCM-based dataset as an example).

It is also important to note that, although appropriate weighting is demonstrated to be effective in this work, it cannot compensate for all structural differences. If the underlying data-generating mechanisms differ fundamentally, or if relevant dimensions are unobserved, reweighting may only partially mitigate bias. Therefore, the effectiveness of weighting should be evaluated on a case-by-case basis using assessment-oriented diagnostics, such as the proposed bin-level statistics.

Additionally, in practice, comprehensive reference datasets covering the entire joint parameter space are rarely available. Instead, only subsets of parameters are typically found within available reference datasets. Under such conditions, it is still possible to conduct equivalence testing. However, equivalence criteria cannot be inferred solely from population structure; instead, they must be specified using external inputs, such as expert judgment, prior evidence, sensitivity analyses, or assessment requirements linked to system design or regulatory context. Consequently, validation results should be interpreted as conditional on these assumptions, rather than as definitive statements about population-level representativeness. Explicitly documenting these assumptions is essential for transparency, reproducibility, and meaningful use of validation results in safety assessment and regulatory dialogue.

Finally, equivalence testing should be viewed as an iterative diagnostic tool rather than a one-off pass/fail criterion in the context of safety impact assessment. When non-equivalence is detected, the binning-based statistics provide insight into which parts of the distribution drive the discrepancy, guiding targeted improvements to scenario generation, weighting strategies, and/or simulation design. In this way, the framework supports the systematic refinement of synthetic datasets and their progressive alignment with assessment-relevant reference behavior.

5. Limitations and Future Work

Several limitations of the proposed framework should be acknowledged. First, the framework does not yet provide practical guidelines for selecting the most relevant metrics for a given safety impact assessment (i.e., Step 1 in Section 2.1.1). Future work should establish systematic criteria or decision rules for metric selection to ensure that equivalence testing focuses on the metrics most relevant to the scope of the intended assessment.

Second, the current implementation treats metrics independently and therefore does not account for potential correlations among them, which may result in overly conservative practical equivalence decisions and a higher incidence of false non-equivalence conclusions (Wu et al., 2025b). Expanding the framework to include multivariate distribution modeling could facilitate a more comprehensive and accurate evaluation of scenario realism.

Finally, the method is sensitive to prior assumptions of distribution models and the specification of ROPEs, which influence posterior inferences about equivalence. Sensitivity analyses of prior choices and ROPE definitions will be essential for improving transparency.

Acknowledgments

This research was supported by the Fordonsstrategisk forskning och innovation (FFI) program, sponsored by Vinnova, the Swedish governmental agency for innovation, as part of the project Improved quantitative driver behavior models and safety assessment methods for ADAS and AD (QUADRIS: nr. 2020-05156). The authors wish to thank Mikael Ljung Aust at Volvo Cars Safety Center for reviewing the manuscript.

Table A.2: Estimated power of the ROPE-based equivalence test

Metric	Statistic	Power	95% CI
P_{inj}	θ	0.870	[0.848, 0.889]
	Θ	1.000	[0.996, 1.000]
$a_{\text{l,min}}$	θ	1.000	[0.996, 1.000]
	Θ	1.000	[0.996, 1.000]
$a_{\text{f,min}}$	θ	1.000	[0.996, 1.000]
	Θ	1.000	[0.996, 1.000]
t_{nr}	θ	0.997	[0.991, 0.999]
	Θ	1.000	[0.996, 1.000]

Code availability

The core implementation of the proposed equivalence testing procedure in this study is publicly available in the `bayes-binned-equivalence` repository (Wu, 2026).

Appendix A. Bootstrap-Based Power Analysis

To assess the reliability and sensitivity of the proposed ROPE-based equivalence testing framework, a bootstrap analysis was conducted to estimate its power, defined as the probability of correctly declaring equivalence when the true parameter lies within the ROPE (Schwaferts and Augustin, 2020). The bootstrap analysis evaluates how consistently the method identifies equivalence when it is indeed true, given the adopted parameter settings (N, α, θ_{thd}, and Θ_{thd}) and sample sizes.

The same reference dataset used in the main analysis was used. As described in Section 3.1, the dataset was randomly sampled from a large parent dataset, the QUADRIS pre-crash dataset. One thousand bootstrap replicates (essentially new "synthetic" datasets) were created from the parent dataset, each the same size as the PCM dataset, using the same process that created the original reference dataset. For each new synthetic dataset, the equivalence testing procedure was applied in the same manner as described in the main analysis.

Since those new datasets are truly equivalent to the reference dataset, power was estimated as the proportion of bootstrap replicates in which equivalence was declared. Binomial 95% confidence intervals for the estimated power were computed using the Wilson method (Wilson, 1927).

The resulting power estimates are summarized in Table A.2. Across all metrics and both statistics (θ and Θ), the estimated power exceeds 0.8 and, in most cases, approaches 1.0, indicating that the framework reliably detects practical equivalence under the adopted parameter setting.

References

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. The journal of wildlife management , 912–923.

Bärgman, J., Boda, C.N., Dozza, M., 2017. Counterfactual simulations applied to SHRP2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems. Accident Analysis & Prevention 102, 165–180.

Baron, W., Sippl, C., Hielscher, K.S., German, R., 2020. Repeatable simulation for highly automated driving development and testing, in: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), IEEE. pp. 1–7.

BMW Group, . About SCM: Stochastic cognitive model. `https://www.bmwgroup.com/scm-driver/en/about-scm.html`. Accessed: 2025-12-03.

Cai, J., Deng, W., Guang, H., Wang, Y., Li, J., Ding, J., 2022. A survey on data-driven scenario generation for automated vehicle testing. Machines 10, 1101.

Daamen, W., Buisson, C., Hoogendoorn, S.P., 2014. Traffic simulation and data: Validation methods and applications. CRC Press.

Donà, R., Ciuffo, B., 2022. Virtual testing of automated driving systems. a survey on validation methods. IEEE Access 10, 24349–24367.

Forman, J.L., Kent, R.W., Mroz, K., Pipkorn, B., Bostrom, O., Segui-Gomez, M., 2012. Predicting rib fracture risk with whole-body finite element models: development and preliminary evaluation of a probabilistic analytical framework, in: Annals of Advances in Automotive Medicine/Annual Scientific Conference, p. 109.

Fries, A., Fahrenkrog, F., Donauer, K., Mai, M., Raisch, F., 2022. Driver behavior model for the safety assessment of automated driving, in: 2022 IEEE Intelligent Vehicles Symposium (IV), pp. 1669–1674. doi:`10.1109/IV51971.2022.9827404`.

Gambi, A., Huynh, T., Fraser, G., 2019. Generating effective test cases for self-driving cars from police reports, in: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 257–267.

Gennarelli, T.A., Wodzin, E., 2006. Ais 2005: a contemporary injury scale. Injury 37, 1083–1091.

Gibbs, N., 2013. Errors in the interpretation of 'no statistically significant difference'. Anaesthesia and Intensive Care 41, 151–154. doi:`10.1177/0310057x1304100203`.

Greene, W.L., Concato, J., Feinstein, A.R., 2000. Claims of equivalence in medical research: are they supported by the evidence? Annals of Internal Medicine 132, 715–722.

Hamdane, H., Serre, T., Masson, C., Anderson, R., 2015. Issues and challenges for pedestrian active safety systems based on real world accidents. Accident Analysis & Prevention 82, 53–60.

Hankey, J.M., Perez, M.A., McClafferty, J.A., 2016. Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets. Technical Report S2-S31-RW-3. Virginia Tech Transportation Institute. URL: `https://vtechworks.lib.vt.edu/items/7949e150-1f1e-493b-bcdb-ba7c660fea34`.

Hay, J., Fonseca Alexandre de Oliveira, L., Schories, L., Dahringer, N., 2025. V4SAFETY automated emergency braking (AEB) model. URL: `https://openvt.eu/v4safety/aeb-technology-models`. accessed: 2025-05-12.

Kruschke, J.K., 2018. Rejecting or accepting parameter values in bayesian estimation. Advances in methods and practices in psychological science 1, 270–280.

Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological research: A tutorial. Advances in methods and practices in psychological science 1, 259–269.

Limentani, G.B., Ringo, M.C., Ye, F., Bergquist, M.L., McSorley, E.O., 2005. Beyond the t-test: statistical equivalence testing.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation research part A: policy and practice 44, 291–305.

Olleja, P., Bärgman, J., Lubbe, N., 2022. Can non-crash naturalistic driving data be an alternative to crash data for use in virtual assessment of the safety performance of automated emergency braking systems? Journal of safety research 83, 139–151.

On-Road Automated Driving Committee, 2021. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. URL: `https://www.sae.org/standards/content/j3016_202104`. SAE Standard.

Pradhan, A.K., Hungund, A., Sullivan, D.E., et al., 2022. Impact of Advanced Driver Assistance Systems (ADAS) on Road Safety and Implications for Education, Licensing, Registration, and Enforcement. Technical Report 22-027. Massachusetts. Dept. of Transportation. Office of Transportation Planning.

Scanlon, J.M., Kusano, K.D., Daniel, T., Alderson, C., Ogle, A., Victor, T.,

2021. Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. Accident Analysis & Prevention 163, 106454.

Schubert, A., Liers, H., Petzold, M., 2017. The GIDAS pre-crash-matrix 2016: Innovations for standardized pre-crash-scenarios on the basis of the VUFO simulation model VAST, in: Proceedings of the 7th International Conference on ESAR. URL: https://trid.trb.org/View/1482288.

Schwaferts, P., Augustin, T., 2020. Bayesian Decisions using Regions of Practical Equivalence (ROPE): Foundations. Technical Report 235. Department of Statistics, University of Munich. doi:10.5282/ubm/epub.74222.

Shah, S., Dey, D., Lovett, C., Kapoor, A., 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles, in: Field and Service Robotics: Results of the 11th International Conference, Springer. pp. 621–635.

Szalay, Z., 2023. Critical scenario identification concept: the role of the scenario-in-the-loop approach in future automotive testing. IEEE Access .

V4SAFETY, 2022. Virtual and validated assessment for automated and connected driving. [Online]. Available: https://v4safety.eu/. Horizon Europe Project No. 101069576.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing 27, 1413–1432. doi:10.1007/s11222-016-9696-4.

Wang, J.S., 2022. MAIS (05/08) Injury probability curves as functions of Delta V. Technical Report DOT HS 813 219. National Highway Traffic Safety Administration.

Wang, X., Peng, Y., Xu, T., Xu, Q., Wu, X., Xiang, G., Yi, S., Wang, H., 2022. Autonomous driving testing scenario generation based on in-depth vehicle-to-powered two-wheeler crash data in China. Accident Analysis & Prevention 176, 106812.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209–212. doi:10.1080/01621459.1927.10502953.

Wimmer, P., Op_Den_Camp, O., Weber, H., Chajmowicz, H., Wagner, M., Mallada, J.L., Fahrenkrog, F., 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, pp. 3–6.

Wu, J., 2024. Generation of Representative Pre-Crash Scenarios Across the Full Severity Range Using Real-World Crash Data: Towards More Accurate Virtual Assessments of Vehicle Active Safety Technologies. Chalmers Tekniska Hogskola (Sweden).

Wu, J., 2025. QUADRIS project pre-crash and near-crash dataset. URL: https://github.com/JianWu09/QUADRIS-project-Pre-crash-near-crash-database. accessed: May 2025.

Wu, J., 2026. bayes-binned-equivalence. URL: https://github.com/JianWu09/bayes-binned-equivalence.

Wu, J., Flannagan, C., Sander, U., Bärgman, J., 2025a. Model-based generation of representative rear-end crash scenarios across the full severity range using pre-crash data. IEEE Transactions on Intelligent Transportation Systems 26, 15932–15950. doi:10.1109/TITS.2025.3573386.

Wu, J., Sander, U., Flannagan, C., Zhao, M., Bärgman, J., 2025b. Practical equivalence testing and its application in synthetic pre-crash scenario validation. arXiv preprint arXiv:2505.12827 .

Zhang, F., Subramanian, R., Chen, C.L., Noh, E.Y., 2019. Crash Investigation Sampling System: Sample Design and Weighting. Technical Report DOT HS 812 706. National Highway Traffic Safety Administration. Washington, DC, USA.