Thesis for The Degree of Doctor of Philosophy

Generation and Validation of Representative Pre-Crash Scenarios

Toward Accurate and Credible Safety Impact Assessments of Driving Automation Systems

Jian Wu

Department of Mechanical Engineering
Gothenburg, Sweden, 2026

Generation and Validation of Representative Pre-Crash Scenarios
Toward Accurate and Credible Safety Impact Assessments of Driving Automation Systems

Jian Wu

ISBN 978-91-8103-387-8
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5844.
ISSN 0346-718X
https://doi.org/10.63959/chalmers.dt/5844

Department of Mechanical Engineering
Division of Vehicle Safety
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Cover image:
A schematic depiction of how heterogeneous pre-crash data are used to generate many representative synthetic scenarios.

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

To my family

To Yuan

To the past, present, and future

Abstract

Driving automation systems (DAS), including Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS), are expected to substantially improve traffic safety. Virtual safety assessment is the primary approach for quantitatively evaluating the prospective safety impacts of these systems, but its validity depends critically on the availability of comprehensive and representative pre-crash scenarios. Existing real-world data are limited in quantity and coverage and often suffer from sampling bias, making the generation of synthetic pre-crash scenarios necessary. However, current generation approaches face challenges such as biased or incomplete data and difficulties in validation. In particular, the absence of systematic methods for validating the representativeness of the synthetic scenarios remains a critical knowledge gap.

To address these challenges, this thesis develops an integrated methodological framework for generating and validating representative synthetic pre-crash scenarios for (prospective) safety impact assessment (SIA) of DAS. The framework consists of two complementary components: 1) a novel approach for generating representative synthetic pre-crash scenarios, and 2) an assessment-oriented framework for validating their representativeness.

Papers I and II present the proposed scenario generation approach that combines heterogeneous empirical data through model-based parameterization and weighting to construct reference pre-crash datasets. Synthetic scenarios are generated using parametric multivariate models and reweighted to match the reference distributions. The underlying generation logic can, in principle, be applied to conflict-based scenarios with or without collision, but the empirical implementation focuses on rear-end pre-crash scenarios with purely longitudinal dynamics, reflecting current limitations in available datasets.

To address the validation gap, Papers III and IV introduce a Bayesian Region of Practical Equivalence (ROPE)-based framework to assess whether synthetic pre-crash scenarios are practically equivalent to their real-world counterparts for SIA purposes. The framework emphasizes assessment-relevant metric selection, interpretable statistics, and explicitly defined equivalence criteria, and provides diagnostic insight into the sources and implications of non-equivalence.

Overall, the thesis contributes a transparent, reproducible methodology for generating representative synthetic rear-end pre-crash scenarios and a general, assessment-oriented framework for validating scenario representativeness, supporting more accurate and credible SIAs of DAS.

Keywords

Driving Automation Systems, Virtual Safety Assessment, Safety Impact Assessment, Synthetic Pre-Crash Scenario Generation, Data Combination, Equivalence Testing.

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **Wu, J.**, Flannagan, C., Sander, U., & Bärgman, J. (2024). Modeling lead-vehicle kinematics for rear-end crash scenario generation. *IEEE Transactions on Intelligent Transportation Systems*, 25(9), 10866–10884. https://doi.org/10.1109/TITS.2024.3369097

[**Paper II**] **Wu, J.**, Flannagan, C., Sander, U., & Bärgman, J. (2025). 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(10), 15932–15950. https://doi.org/10.1109/TITS.2025.3573386

[**Paper III**] **Wu, J.**, Sander, U., Flannagan, C., & Bärgman, J. (2025, September). Practical equivalence testing and its application in synthetic pre-crash scenario validation. In *2025 IEEE International Automated Vehicle Validation Conference (IAVVC)* (pp. 1–8). IEEE. https://doi.org/10.1109/IAVVC61942.2025.11219586

[**Paper IV**] **Wu, J.**, Sander, U., Flannagan, C., & Bärgman, J. (2026). Practical validation of synthetic pre-crash scenarios. (under review at Accident Analysis & Prevention).

Author's contribution (for all manuscripts): designed the methods, performed the simulations, did the analysis, led the conclusion work, and wrote the original draft of the manuscript.

Acknowledgment

I would like to express my deepest gratitude to my supervisors, Jonas Bärgman and Carol Flannagan, and my industrial supervisor, Ulrich Sander, for their inspiring guidance and continuous support throughout this journey. I vividly remember the moments when we arrived at the same idea and voiced it almost simultaneously, as well as the times when we disagreed—but never hesitated to challenge one another's perspectives. Those discussions shaped not only my research but also the way I think as a researcher and scientist. You have been exceptional mentors, collaborators, and role models, and I am sincerely thankful for all that I have learned from you.

I am grateful to András Bálint and Mikael Ljung Aust for their support and guidance at the beginning of my doctoral studies, and to Marco Dozza, my examiner, for his constructive feedback and insightful comments that strengthened this thesis. I would also like to thank my former manager at Volvo Cars, Mattias Robertson, for his early support and encouragement, and my current manager, Ashok Chaitanya Koppisetty, for his continued understanding and flexibility, which helped me maintain a healthy balance between research and life.

My heartfelt thanks go to all my colleagues and friends at Chalmers and Volvo Cars for creating an inspiring and collaborative research environment filled with ideas, curiosity, and humor.

This research was made possible through the support of the Fordonsstrategisk forskning och innovation (FFI) program, funded by Vinnova—the Swedish governmental agency for innovation—under the QUADRIS project, Grant 2020-05156. Parts of this work also benefited from outcomes developed within the V4SAFETY project, funded by the European Commission under grant number 101075068. I sincerely acknowledge the support, which provided the resources and collaboration framework that enabled this work.

Finally, I am deeply grateful to my family for their unwavering support and encouragement. Most of all, I thank my caring, wise, and farsighted wife, Yuan Liao, who has stood beside me with patience, insight, and love in every step of this journey.

Gothenburg, May 2026
Jian Wu

Declaration on the Use of Generative AI

During the preparation of this thesis, the author used generative large language models (ChatGPT and Microsoft Copilot) to improve grammar, wording, and clarity in selected sentences and paragraphs of the manuscript. The use of these tools was strictly limited to language refinement and did not involve the generation of scientific content, ideas, methods, analyses, results, or conclusions. All text assisted by these tools was carefully reviewed and, where necessary, edited by the author, who takes full responsibility for the content of the thesis.

Contents

Part I

Summary

Chapter 1

Introduction

1.1 Enhancing traffic safety through Driving Automation Systems

Road traffic crashes remain one of the most significant challenges for modern society, causing substantial fatalities, injuries, and economic losses worldwide [1]. Over recent decades, vehicle safety research has evolved from passive protection systems (e.g., seat belts and airbags) to Driving Automation Systems (DAS) [2], which include Advanced Driver Assistance Systems (ADAS) [3, 4, 5] and Automated Driving Systems (ADS) [6, 7, 8]. These systems are designed to support or replace human drivers in safety-critical moments by monitoring the environment, identifying potential conflicts, and intervening when necessary to prevent or mitigate crashes [9, 10]. In particular, many DAS primarily operate during the crash-imminent phase—the short time window immediately preceding a potential crash—when they can correct driver errors, reduce crash speed, or avoid crashes altogether [11, 12].

Systems such as Adaptive Cruise Control (ACC) [3], Lane Keeping Assist (LKA) [5], and Automatic Emergency Braking (AEB) [4] represent key ADAS technologies that support drivers by monitoring the driving environment, assessing potential conflicts, and providing timely warnings or control interventions to enhance safety and situational awareness [13]. ADS, in contrast, differ primarily in the level of automation and control authority: they are designed to perform the driving task (partly or fully) within a defined operational design domain by maintaining continuous control and handling the perception–decision–action loop with reduced reliance on immediate driver supervision [6, 7, 8]. Ranging from driver assistance to full automation, these DAS aim to complement or replace human control in safety-critical situations, thereby avoiding or mitigating crashes that may otherwise occur due to factors such as distraction [14], fatigue [15], or impairment [16].

Empirical research has demonstrated the safety impact of such systems: AEB and LKA, for example, have been shown to substantially reduce crash and injury rates [17, 18, 19, 20]. Nevertheless, accurately quantifying the overall

safety performance of DAS in real traffic remains a major challenge. Crashes are rare events; comprehensive field testing would require extremely large driving exposure and prohibitively high costs to obtain statistically significant safety evidence [21]. Moreover, safety performance depends on a multitude of contextual factors (such as traffic composition, weather, infrastructure, and driver behavior) that are difficult to control or replicate experimentally. There is a need for efficient, reproducible, and systematically controlled approaches to prospective safety evaluation that address these challenges—that is, methods that estimate safety impacts before large-scale real-world deployment, rather than relying solely on observed crash outcomes [22].

1.2 Virtual simulations for accurate and credible safety impact assessment

To overcome the limitations of field testing, researchers and industry practitioners increasingly employ simulation-based assessment approaches, collectively referred to as *virtual safety assessment*, to evaluate the safety performance of DAS [22]. These methods enable controlled experimentation across a wide range of traffic, driver, and environmental conditions, including rare or hazardous crash events that would be infeasible, unsafe, or prohibitively expensive to reproduce in real traffic [23, 22]. They can be broadly categorized into two types based on their purpose.

The first type of assessment, *safety assurance*, focuses on system-level validation. It aims to ensure that a DAS performs safely and reliably across its operational design domain, even under challenging and rare conditions [24]. This type of assessment explores the system's functional behavior, decision-making robustness, and capacity to detect, manage, and mitigate hazardous situations before they result in a crash [25, 26, 27]. It is primarily used in the development and verification stages to build confidence in the system's operational safety and to identify edge-case scenarios that may expose system limitations [28, 29]. It also plays a regulatory role, as reflected in United Nations Economic Commission for Europe (UNECE) Regulation No. 157 on Automated Lane Keeping Systems [30].

The second type, *safety impact assessment* (SIA), is the focus of this thesis. It aims to quantify the expected safety impact/benefits of a system at the population level—typically in terms of reductions in crash frequency, impact speed, or injury risk [31]. In SIA, *pre-crash scenarios* (short time sequences describing the dynamics of involved traffic participants and the environment leading up to a crash) are simulated under baseline (without the DAS under assessment) and treatment (with the DAS) conditions [32, 33]. The prospective safety impact of a system can be estimated by comparing the outcomes of these simulated scenarios [31, 34, 32, 33]. It should be noted that SIA is often used as a component of safety assurance and safety argumentation; it is also widely applied for other purposes, such as system development, comparative evaluation of design alternatives, and policy-oriented analysis [21, 35]. Additionally, in this thesis, SIA is understood as a prospective, scenario-based evaluation

conditioned on defined pre-crash conflicts, rather than as a comprehensive assessment of continuous driving or conflict genesis.

To ensure a statistically sound and unbiased comparison, the SIA procedure requires that the pre-crash scenarios are adequate and represent real-world conditions accurately [31]. However, the available empirical data for creating these scenarios suffer from limited sample sizes, incomplete coverage, and sampling bias [36, 31, 37, 38]. To address these limitations, a common strategy is to use statistical, behavioral, or combined models derived from empirical data to create diverse, reproducible, and scalable sets of synthetic pre-crash scenarios suitable for the assessment [39, 28, 40]. While running more simulations using this strategy can easily generate a statistically sufficient number of scenarios, the resulting synthetic scenarios may not be representative. The creation of synthetic scenarios relies heavily on empirical data; any limitations or biases in the datasets—such as incomplete coverage or unbalanced sampling—will be reflected in the synthetic scenarios. Therefore, ensuring that the generated scenarios accurately represent their real-world counterparts for the intended SIA is a critical challenge.

1.3 Challenges in generating and validating representative pre-crash scenarios

The creation of representative pre-crash scenarios involves two major challenges: 1) generating representative synthetic pre-crash scenarios and 2) validating their *representativeness*. These challenges have been widely recognized in prior research on crash reconstruction, naturalistic driving analysis, and simulation-based safety evaluation [41, 31, 42, 23, 43, 44].

1.3.1 Generation challenges

Two main approaches have been widely employed to generate synthetic pre-crash scenarios: traffic-simulation-based and in-depth-crash-data-based (IDC-based) [45]. Both have inherent limitations that restrict their ability to produce representative scenarios across the full severity range, from physical contact to severe injuries or fatalities.

The traffic-simulation-based approach aims to replicate daily driving activities with traffic agents to generate crashes in a virtual, modeled driving environment [46, 47, 48, 45]. Typically, traffic agents are built using naturalistic driving data (NDD; studies that unobtrusively collect data from participants' daily driving over weeks, months, or even longer periods [49]), which often contain a limited number of crashes, typically minor in severity [48]. While this method can produce a wide variety of low-risk driving scenarios, it has two main limitations. First, safety-critical events that lead to crashes are rare during normal driving, so simulations must run for very long durations to gather enough crash scenarios for statistical analysis [41, 22]. This makes the approach inefficient and computationally expensive. Second, crashes generated through traffic simulations may differ systematically from real-world crashes in terms

of their kinematic characteristics and overall severity distributions [38, 50]. This is because the traffic agents are primarily based on NDD, which underrepresent severe crashes and may therefore fail to reproduce the complex and abrupt pre-crash maneuvers, especially in severe crashes [51]. Therefore, while traffic-simulation-based approaches offer broad exposure coverage, they rarely produce realistic, representative pre-crash behaviors across the full severity range.

In contrast, IDC-based approaches typically generate synthetic scenarios either by resampling from empirical distributions of relevant pre-crash and crash-related characteristics, such as pre-crash kinematics [40], or by systematically modifying and varying real-world pre-crash scenarios [31, 52]. These approaches rely on detailed reconstructed or recorded pre-crash data from in-depth crash datasets. Notable examples of such datasets include the U.S. Crash Investigation Sampling System (CISS) and the German In-Depth Accident Study (GIDAS), together with its Pre-Crash Matrix (PCM) extension. These datasets provide valuable information on vehicle dynamics and potential contributing factors to accidents [42, 23]. However, SIA generally requires more detailed real-world crash instances than are currently available [6, 27]. An additional challenge is that the in-depth crash data are often biased toward severe crashes. This bias arises not only from their explicit inclusion criteria (e.g., requirements related to injuries or towed vehicles), but also from the severity-dependent investigation and reporting probabilities, as low-severity crashes are substantially less likely to be investigated in depth or systematically reported. For instance, CISS focuses on crashes with at least one light vehicle towed from the scene [23], and GIDAS gathers on-scene accident cases with personal injury across several cities in Germany [42]. Moreover, generating crashes using reconstructed crash data can present challenges. The reconstruction of pre-crash kinematics is heavily influenced by the analysis software used and the assumptions made about the behavior of the road users involved, particularly when detailed pre-crash recordings are lacking. In fact, the resulting pre-crash scenarios may depend more on these assumptions and the software than on real-world scenarios [39, 50]. As a result, IDC-based methods offer detailed severity information but lack statistical balance and generalizability.

In summary, existing synthetic pre-crash scenario generation approaches capture only a portion of the pre-crash severity spectrum. Traffic-simulation-based methods provide broad exposure coverage but struggle to reproduce realistic and representative pre-crash behaviors across the full severity range. In contrast, IDC-based methods provide detailed severity information but are limited by small sample sizes and a bias toward severe crashes. These limitations pose fundamental challenges for SIA. Addressing them requires methods that achieve both individual-level realism/plausibility and population-level representativeness, so that synthetic pre-crash scenarios are physically and behaviorally realistic while collectively reflecting real-world conditions across the full severity range in a manner that is accurate, credible, and assessment-relevant.

1.3.2 Validation challenges

Even when synthetic pre-crash scenarios are carefully generated, their representativeness must be validated by assessing whether they are practically equivalent (or "similar enough") to their real-world counterparts for the intended assessment purpose. This issue is central to the accuracy and credibility of SIA: without explicit validation, biases such as the over- or under-representation of severe crashes can lead to misleading conclusions about system effectiveness [31, 38, 53].

In much of the existing literature, emphasis is placed on verifying the individual-level *plausibility* of synthetic scenarios. Plausibility checks typically assess whether a scenario is physically feasible and internally consistent, as well as realistic in terms of driver and vehicle behavior; its kinematic, dynamic, and temporal components must be compatible, in terms of (for example) realistic accelerations, speeds, and trajectories. These checks, a necessary first step in validation, primarily address internal realism rather than population-level representativeness [35]. Consequently, whether a set of generated scenarios collectively reflects the statistical properties of real-world pre-crash conditions relevant to a given assessment objective is often left implicit. This situation is reinforced by the absence of standardized validation criteria or regulatory guidance on representativeness, as well as by methodological practices that prioritize event-level realism or difference detection over assessment-oriented similarity.

The validation of representativeness requires methods that assess the *practical equivalence* between synthetic and real-world (or reference) pre-crash scenarios—that is, whether observed differences are sufficiently small to be negligible for the given SIA, with predefined, assessment-relevant tolerances [54, 55, 56]. However, such equivalence-oriented approaches are largely absent from current validation practice [35]. In their absence, virtual SIA is still validated predominantly on descriptive summaries, visual comparisons, or difference-oriented statistical tests [57, 40, 58].

Descriptive statistics and visualization techniques (such as histograms, scatter plots, and dimension-reduction methods) are frequently used to provide exploratory insights into distributional similarity and structural patterns, and correlation analyses are sometimes employed to examine whether key relationships between variables are preserved [59, 40, 58]. While useful for exploratory analysis and plausibility checking, these approaches do not provide explicit decision rules or acceptance criteria for determining whether observed differences are small enough to be negligible for the intended SIA, and therefore cannot, on their own, support a formal validation of practical equivalence. Similarly, difference-oriented statistical tests (such as the Kolmogorov–Smirnov, Anderson–Darling, and chi-square tests) are not designed to establish whether the differences are negligible for practical purposes [60, 57]: failure to reject a difference does not imply practical equivalence, and detected differences may still be practically negligible [61, 62, 53]. Overall, existing validation practices are primarily intended to support plausibility checking and difference detection, rather than explicitly evaluating whether synthetic scenarios are practically

equivalent to real-world pre-crash data in ways that matter for crash outcomes and the intended SIA.

To address this gap, a potential solution lies in *equivalence testing*, a family of statistical approaches designed to determine whether observed differences fall within predefined, practically acceptable bounds [54, 55]. Equivalence testing reverses the logic of classical hypothesis testing by treating equivalence as the null hypothesis and rejecting it only when evidence suggests a meaningful difference. In practice, one or more statistics quantify differences between datasets—such as differences in means or medians—and are compared against the predefined equivalence bounds. If the observed differences fall entirely within those bounds, the datasets are considered practically equivalent for the intended purpose. Although this approach is well-established in fields such as pharmacology and psychology [63], its application in transportation safety and the validation of synthetic scenarios for SIA remains limited to date [62, 53]. Moreover, applying equivalence testing in this context poses three main challenges: 1) selecting what to measure, 2) measuring differences in an interpretable and practically relevant manner, and 3) defining and justifying criteria for declaring equivalence.

The first challenge is selecting what to measure. Pre-crash data typically consist of multivariate time series describing the dynamics and trajectories of the road users involved. A direct comparison of these data is impractical due to differences in duration, temporal alignment, dimensionality, and noise, as well as the lack of interpretable similarity measures that link signal (time-series)-level differences to assessment-relevant safety outcomes [64, 65, 27]. As a result, the data are typically characterized by a set of quantitative variables—referred to as *metrics*—that capture key aspects of the dynamics of the road users involved and the crash outcomes [53]. It is important to note that not all metrics are equally relevant to a specific SIA. For example, the metrics of lateral dynamics are crucial for LKA, while they are less relevant for AEB, which is primarily focused on the vehicle's longitudinal maneuvers. Further, while including many metrics in the equivalence test enhances comprehensiveness, it also increases the risk of false non-equivalence. In such cases, non-equivalence in less relevant metrics (such as those related to lateral control in the study of AEB) may dominate the validation outcome, leading to an erroneous rejection of an otherwise representative dataset. To date, there is no established guidance on how to prioritize or weight validation metrics according to their relevance to the assessment objective.

The second challenge is measuring differences in an interpretable and practically relevant manner. Assessing practical equivalence requires *validation statistics* that quantify both statistical differences and their practical implications. A statistic is a numerical quantity computed from sample data used for summarization or inference; formally, it is any function of the observed data that does not depend on unknown population parameters [66]. Unlike conventional statistics—such as absolute differences in means or medians—that focus only on statistical variations, interpretable validation statistics explicitly link these variations to their practical relevance for the specific assessment, enabling analysts to evaluate whether and how discrepancies between synthetic

and empirical datasets influence the assessment outcome. However, general and practically applicable approaches for defining these statistics are largely undeveloped to date, limiting the consistency and interpretability of validation results across studies.

The third challenge is defining and justifying criteria for declaring equivalence. These criteria, which specify the acceptable level of difference between compared datasets, may be applied at different levels (individual metrics, subsets of interdependent metrics, or jointly across all metrics). When equivalence is assessed at the metric or subset level, additional *overall equivalence criteria* are required to aggregate multiple validation outcomes into a single, interpretable conclusion. The choice of equivalence criteria involves an inherent trade-off: those that are too stringent may cause normal variability to be misinterpreted as meaningful non-equivalence, while overly lenient criteria may mask important discrepancies and lead to false equivalence conclusions. Establishing appropriate criteria requires combining statistical reasoning with domain knowledge to make the underlying design choices explicit and to justify the accepted level of similarity for the specific assessment. However, practical guidance on defining, aggregating, and justifying equivalence criteria remains lacking, limiting both transparency and comparability across validation studies.

1.4 Aims and objectives

Together, these challenges underscore the need for systematic, domain-specific methods to generate and validate representative pre-crash scenarios. Accordingly, the overarching aim of this thesis is to develop such methods to improve the accuracy and credibility of SIAs for DAS.

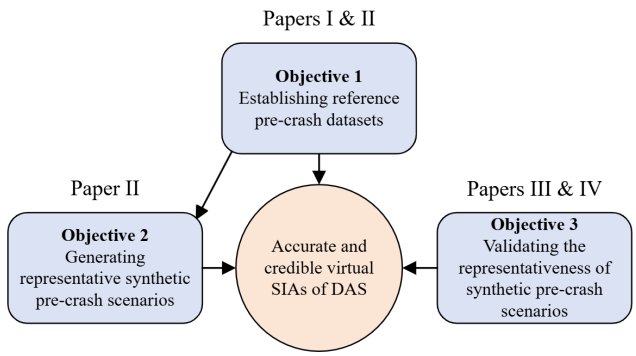

Figure 1.1: An overview of the four papers included in this thesis and their connections to the three objectives and the overarching aim (orange circle).

To achieve this aim, three main methodological research objectives (covered by four papers) have been established, as illustrated in Figure 1.1.

1. **Establishing reference pre-crash datasets.** This objective, addressed in Papers I and II, focuses on developing methods to establish these datasets. A *reference dataset* is defined as an empirically grounded dataset

intended to approximate the real-world population of a particular crash type and serving as the baseline against which synthetic scenarios are generated and validated. Because existing empirical sources (such as naturalistic driving data and in-depth crash databases) each capture only part of the spectrum of pre-crash scenarios, multiple complementary reference datasets are typically required to describe the full range of conditions relevant to SIA. The first step toward enabling systematic analysis and integration across heterogeneous sources is to transform multivariate pre-crash time-series data via model-based parameterization into a common, finite set of interpretable parameters that capture the essential dynamics of each scenario. Based on this parameterized representation, reference datasets are then constructed by combining empirical data from multiple sources and applying *sample weighting* techniques. This process does not generate new scenarios; rather, it reweights and integrates existing observations to provide adequate coverage of driving exposure and crash severity while mitigating sampling and selection biases.

2. **Generating representative synthetic pre-crash scenarios based on the reference datasets.** This objective, addressed in Paper II, focuses on developing generation methods that overcome the limitations of existing traffic-simulation-based and IDC-based approaches. Building on the reference pre-crash datasets established in the first objective, the work constructs probabilistic distribution models that define *reference distributions* for selected scenario parameters or parameter subsets. Synthetic pre-crash scenarios are subsequently generated through model-based simulation using parameter configurations sampled from these reference distributions. Because scenarios obtained through model-based simulation rather than direct resampling may not inherit the statistical properties of the reference population, sample weighting is applied to improve alignment with the reference distributions. The generated scenarios are designed to reproduce the diversity of real-world pre-crash dynamics and crash severities while maintaining physical plausibility as well as statistical consistency with the reference distributions relevant to the intended SIA.

3. **Validating the representativeness of synthetic pre-crash scenarios.** This objective, addressed in Papers III and IV, focuses on developing a generic, interpretable validation framework to assess whether synthetic pre-crash scenarios are practically equivalent to their real-world counterparts for the intended assessment purpose, as well as two practical guidelines for the framework's application. The first guideline addresses how to measure differences in an interpretable and practically relevant manner, and the second addresses how to define and justify the criteria for declaring practical equivalence. The framework and guidelines together support the consistent, transparent, and meaningful validation of synthetic scenarios for accurate and credible SIA.

Collectively, these objectives define a unified methodological framework

for generating and validating representative pre-crash scenarios. The scenario generation approach proposed in this thesis can, in principle, be applied to conflict-based scenarios with or without collision, but the empirical implementation focuses on rear-end pre-crash scenarios with purely longitudinal dynamics. This focus is a direct consequence of current data availability: existing empirical datasets provide sufficiently detailed and consistent longitudinal kinematic information, but do not offer the reliability required to model lateral dynamics, lane changes, multi-agent interactions, infrastructure influences, or environmental conditions. In contrast, the validation framework developed in this thesis is not restricted to a specific pre-crash scenario type and is formulated to be generally applicable to representativeness assessment.

Chapter 2

Methodology

This chapter presents the methodological framework developed in this thesis for generating and validating representative pre-crash scenarios for accurate and credible SIAs of DAS. The framework comprises three logically connected stages: 1) establishing reference pre-crash datasets, 2) generating representative synthetic pre-crash scenarios, and 3) validating their representativeness using a practical equivalence testing framework. These stages correspond directly to the three methodological objectives outlined in Section 1.4 and together form a unified workflow linking empirical observations, synthetic scenario generation, and scenario validation.

The chapter begins with an introduction to the empirical pre-crash data used in this thesis, including their sampling characteristics, severity coverage, and available pre-crash information. Each methodological stage is then described with a workflow diagram illustrating the main steps involved, followed by key considerations and high-level rationales; available approaches and their typical strengths and limitations are also listed. Finally, the concrete methods and models used in this thesis are presented. This structure ensures transparency and traceability in how empirical evidence is transformed into representative synthetic pre-crash scenarios and how their validity is assessed within the context of SIA.

2.1 Empirical pre-crash data

This thesis uses pre-crash data from three empirical sources: the Second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) [49], CISS, and GIDAS-PCM. These datasets differ in their sampling design, measurement precision, and severity coverage, providing complementary inputs for establishing reference pre-crash datasets.

2.1.1 SHRP2

The SHRP2 dataset is one of the largest naturalistic driving datasets in the world, involving more than 3,300 instrumented vehicles operated under everyday

driving conditions across six U.S. sites [49]. Each vehicle was equipped with a data acquisition system that recorded multi-channel information, including vehicle kinematics, driver behavior (via interior video), and roadway context (via forward and rear-facing cameras).

Collected between 2010 and 2013, the SHRP2 dataset provides millions of kilometers of driving data and a broad spectrum of traffic events, ranging from routine car-following to near-crashes and low-severity collisions. Event identification algorithms and manual annotations were used to classify incidents into severity levels [49]. While the SHRP2 dataset offers excellent coverage of normal and moderately critical driving conditions, high-severity crashes remain rare due to the nature of naturalistic sampling [67]. Thus, it provides rich exposure information but limited representation of severe crash dynamics.

2.1.2 CISS

CISS is a nationally representative crash investigation program in the United States designed to provide detailed information on police-reported crashes involving at least one towed passenger vehicle [23]. CISS includes on-scene documentation, vehicle damage assessments, and contextual factors; it has been widely used to study crash and injury mechanisms [20, 68, 69].

A key feature of CISS is the integration of Event Data Recorder (EDR) information, which contains records of the vehicle's speed signal for the five seconds prior to the crash [70]. However, EDR data are available only for a subset of crashes, and usable paired-vehicle EDR cases (i.e., when EDR data are available from both vehicles involved in the crash) are even more scarce. For example, among 1,125 rear-end crashes in CISS, only 10% (113 cases) contain EDR data for both vehicles. Additionally, the sampling frequency of most EDRs is low (1–10 Hz), with only 0.4% of cases at 5 Hz or higher. This limited temporal resolution constrains the ability to capture rapid changes in vehicle dynamics and driving behavior leading up to a crash.

Thus, CISS EDR data contribute detailed crash-involved cases with recorded kinematics, but with restricted temporal fidelity and limited sample sizes for detailed pre-crash reconstruction.

2.1.3 GIDAS-PCM

GIDAS is a long-standing on-scene crash investigation program covering police-reported injury crashes in several regions of Germany. The GIDAS-PCM, introduced in 2011, is a subset of GIDAS that provides reconstructed trajectories for up to two involved traffic participants over the five seconds preceding the crash [42]. This dataset, which includes pre-crash dynamics, environmental context, and roadway geometry, is widely used for accident causation analysis [71, 72], active safety technology evaluation [73, 34, 74], and predictive crash modeling [75].

Unlike SHRP2 or CISS EDR data, which contain directly recorded kinematics, GIDAS-PCM relies on reconstruction methodologies combining physical evidence, expert assessment, and simulation tools. As a result, there are uncer-

tainties, particularly in cases involving rapid dynamics such as hard braking or evasive maneuvers [39]. For this reason, this thesis uses only robustly reconstructed parameters—such as the minimum accelerations of both vehicles—to ensure consistency with the recorded datasets [45].

2.1.4 Summary

Each dataset has limitations. However, together, they can provide complementary information:

- **SHRP2**: naturalistically collected, recorded pre-crash data with a broad exposure to everyday and moderately critical driving, mainly low-severity crashes.
- **CISS (EDR)**: recorded pre-crash data for police-reported crashes with at least one towed passenger vehicle, mainly high-severity crashes.
- **GIDAS-PCM**: reconstructed pre-crash data for injury-involved crashes with comprehensive contextual detail, mainly high-severity crashes, less reliable because of reconstruction.

Both CISS (EDR) and GIDAS-PCM provide data for the five seconds leading up to a crash, whereas SHRP2 data covers a longer duration. To analyze rear-end pre-crash scenarios, data were extracted from all three sources for the five seconds preceding the crash. These data included time-series information on the longitudinal distance between the lead and following vehicles, as well as their speeds, where available.

2.2 Establishment of reference pre-crash datasets

Reference pre-crash datasets serve as the foundation for generating and validating pre-crash scenarios. Since none of the currently available empirical datasets alone can serve as a reference dataset due to their limitations, such as sampling bias and limited coverage [50], multiple complementary reference datasets or parameter distributions are required to accurately describe pre-crash scenarios across the full severity range.

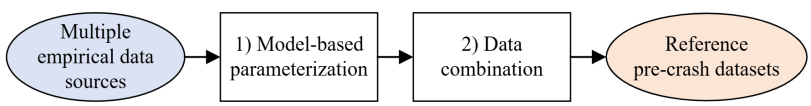

Figure 2.1: Workflow for establishing reference pre-crash datasets.

Figure 2.1 illustrates the workflow used to establish the reference pre-crash datasets in this thesis. First, a model-based parameterization approach was applied to simplify multivariate pre-crash time-series data and represent scenarios from multiple empirical sources with a common set of interpretable parameters. Second, a *data combination* procedure was employed to integrate

the parameterized data across sources and establish reference pre-crash datasets that represent the real-world crash population.

2.2.1 Model-based parameterization

Parameterization is a fundamental component of quantitative modeling that represents complex system behavior through a finite set of interpretable parameters [76], thereby supporting calibration, sensitivity analysis, and simulation while enabling models to remain tractable and adaptable across scenarios [77]. Effective parameterization is therefore essential for capturing real-world variability while maintaining analytical clarity in data-driven and simulation-based studies [77, 78].

In the context of pre-crash analysis, empirical data are typically available as multivariate time series describing the kinematics and trajectories of road users involved in a developing conflict. The data may include vehicle speeds, accelerations, relative distances, and driver control inputs sampled at high temporal resolution. While these time-series data are rich in detail, direct comparison, integration, or statistical modeling across heterogeneous sources is impractical, particularly when datasets differ in sampling frequency, signal availability, noise characteristics, and temporal alignment [79, 80]. These discrepancies are further amplified when the objective is large-scale scenario generation and representativeness validation rather than the analysis of individual events.

Model-based parameterization addresses these challenges by simplifying complex pre-crash trajectories into a compact, consistent representation. By modeling the behavior or kinematics of the involved road users, each pre-crash scenario can be distilled into a finite set of parameters that capture its essential characteristics while abstracting away unnecessary temporal detail. This parameterized representation enables systematic statistical analysis, facilitates integrating data from heterogeneous sources, and allows individual scenarios to be recreated or re-simulated by instantiating the underlying models with a given parameter vector [81, 82]. As such, parameterization provides a practical link between raw empirical observations and the generation of synthetic scenarios.

A further methodological consideration concerns the suitability of existing road-user behavior models for pre-crash scenario generation. Many commonly used behavior models—such as car-following and lane-changing models—were originally developed for traffic flow analysis or nominal driving conditions [80]. While valuable for reproducing typical behavior and aggregate traffic patterns, these models are often not designed to represent the short, safety-critical pre-crash phase, during which driver responses may be abrupt or intermittent, and may be influenced by distraction, delayed reaction, or surprise [83, 82]. Moreover, such models are typically calibrated on normal driving data, further limiting their ability to accurately represent the pre-crash dynamics observed in real-world crashes [35].

For the purpose of establishing representative pre-crash scenarios, parameterization models should therefore be explicitly grounded in empirical pre-crash data, focusing on accurately representing the observed behavior. In this context, the goal of modeling is not to prescribe how road users should behave, but

to encode how they actually behaved prior to crashes in a form suitable for statistical modeling and simulation-based scenario generation.

In this thesis, a specific model-based parameterization was developed for rear-end pre-crash scenarios, which are dominated by longitudinal interactions between a lead vehicle and a following vehicle. Accordingly, separate models were built for the longitudinal motion of the two involved vehicles: a lead-vehicle kinematics model and a following-vehicle behavior model. Together, these models transform empirical pre-crash time-series data into a finite-dimensional parameter representation that forms the basis for the subsequent data combination, *distribution modeling*, and synthetic scenario generation.

The lead-vehicle kinematics model

In rear-end pre-crash scenarios, the behavior of the lead vehicle is, to a large extent, independent of that of the following vehicle, particularly during the early stages of conflict development. This property allows the lead vehicle's longitudinal motion to be modeled separately from the following vehicle's behavior [84].

To parameterize lead-vehicle motion in this thesis, a piecewise linear model was used to represent the pre-crash speed profile as a sequence of consecutive linear segments, resulting in a six-parameter vector describing the lead vehicle's longitudinal kinematics. Figure 2.2 illustrates an example with three segments.

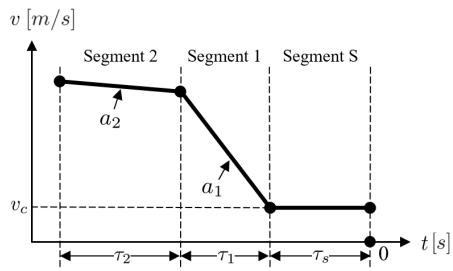

Figure 2.2: An illustration of a piecewise linear model with three segments (time zero is the impact moment) [84]. Each segment assumes constant acceleration, with Segment S representing a constant-speed phase.

Compared with conventional constant acceleration or deceleration models [85, 86, 87], the piecewise linear model provides a more faithful representation of lead-vehicle speed profiles observed in real-world rear-end pre-crashes [84]. In particular, it captures a wider range of empirically observed behaviors, including sustained constant-speed phases, multi-stage braking, and non-monotonic speed changes. Additional details are provided in Section III-A of Paper I.

The following-vehicle behavior model

In contrast to the lead vehicle, the behavior of the following vehicle in a rear-end pre-crash scenario is strongly influenced by the motion of the lead vehicle

and by the driver's response to the developing conflict. Parameterizing this behavior, therefore, requires a model capable of representing both nominal car-following dynamics and safety-critical pre-crash braking responses.

In this thesis, a following-vehicle behavior model with four parameters was constructed by combining two established driver models. A modified Intelligent Driver Model (IDM) [88] provides a continuous car-following framework, while a pre-crash driver brake-response model [82] captures the timing and modulation of braking (including its jerk) based on looming-related visual cues [83]. The integration of these models enables a unified representation of longitudinal following and pre-crash braking behavior.

In addition, the model was extended to account for a specific rear-end pre-crash pattern observed in the GIDAS-PCM dataset and incorporated into the weighted reference dataset, where it represents approximately 9.2% of crashes [45]. In these cases, both vehicles were initially stationary; after a while, the following vehicle accelerated until it hit the lead vehicle. The driver of the following vehicle seemed to ignore the lead vehicle completely, possibly due to distraction. The model includes the possibility of generating this 'abnormal', but empirically observed, driver acceleration behavior. Further details are provided in Section III-A of Paper II.

Parameterization of the rear-end pre-crash scenario

The lead-vehicle kinematics model contributes six parameters, and the following-vehicle behavior model contributes four parameters. The only remaining elements in a rear-end pre-crash scenario are two initial states: the following vehicle's speed and the initial following distance. Consequently, a rear-end pre-crash is parameterized as a twelve-dimensional (12-D) vector using the modeling framework adopted in this thesis. Additional details can be found in Section III-B of Paper II.

2.2.2 Data combination

Following the model-based parameterization, rear-end pre-crash scenarios from multiple empirical sources—SHRP2, CISS, and GIDAS-PCM—are represented as sets of parameters. However, as mentioned, none of these datasets alone can serve as a reference dataset for rear-end pre-crashes. Each source is affected by limitations such as sampling bias, incomplete parameter availability, and varying data quality. In particular, only GIDAS-PCM and a subset of SHRP2 contain information on both vehicles, rather than only the lead or the following vehicle [45]. As a result, establishing reference datasets that adequately represent the real-world rear-end pre-crash population across the full severity range requires combining information from multiple complementary sources.

In fact, the term data combination [89] refers to the integration of multiple datasets to leverage complementary information while compensating for individual limitations. It is widely used in statistics, econometrics, and data science to improve coverage, reduce bias, and increase inferential power [90, 91]. In the context of pre-crash analysis, data combination must resolve issues of

compatibility, consistency, and bias to ensure that the resulting dataset remains meaningful for downstream modeling, validation, and subsequent SIA.

The core objective of data combination in this thesis is to integrate complementary empirical information while mitigating systematic biases present in individual datasets [50]. Naively merging datasets is insufficient, as empirical crash data are often biased with respect to key parameters—most notably crash severity [42, 23, 84, 45]. When a reliable reference marginal or joint distribution is available for a subset of parameters, the biases can be reduced by reweighting the raw samples so that the weighted data align with the reference distribution [92]. The resulting weighted dataset can then be treated as a reference dataset for the parameters it contains.

A commonly used approach to mitigate bias is *post-stratification weighting* [92], which is widely applied in survey research. It involves dividing the target population into strata based on certain variables, collecting data within each stratum, and then assigning weights (based on the target population distribution within each stratum) to the observations.

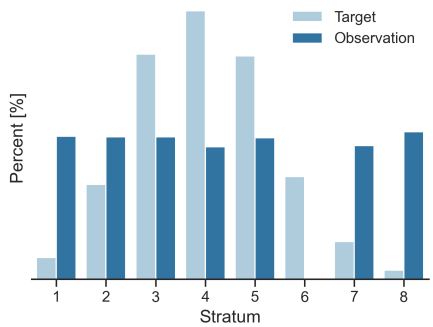

Figure 2.3: A binning example for unidimensional data [50]. Stratum 6 is the omitted stratum.

Typically, binning is used to create strata when the variables are continuous [93]. The weight assigned to observations in each stratum is calculated by dividing the target population total by the number of observations in the stratum. The raw samples (i.e., observations) are grouped into discrete bins based on the known reference distribution (i.e., target population). This method assumes that no strata are omitted, meaning that observations must cover all bins spanned by the target population. Otherwise, the process will attempt to divide by zero. However, in our case, omitted strata exist in the combined data. (See Figure 2.3 for an example: Stratum 6 contains no observations.) Consequently, conventional post-stratification methods cannot be directly applied. Therefore, we developed a *k-nearest neighbors* (KNN)-based sample weighting method. It can be interpreted as a post-stratification approach with a dynamic, data-driven binning strategy. Each sample drawn from a known reference distribution is assigned a unit weight. Each reference sample's k nearest neighbors in the raw empirical dataset are identified and

grouped to share this weight based on their distance from that sample. Raw samples that are never selected as neighbors of any reference sample receive a weight of zero, reflecting their limited relevance for the target population.

An alternative strategy based on a KNN imputation of missing parameters [94] was also evaluated. However, empirical simulation results demonstrated that the KNN-based sample weighting approach yielded superior performance for the purpose of scenario generation and validation [50].

In this thesis, the SHRP2 dataset provided coverage from physical contact events to higher-severity crashes, whereas CISS and GIDAS-PCM are strongly biased toward severe crashes [42, 23]. Accordingly, the distribution of lead-vehicle Delta-v (i.e., speed change during the impact [95]) observed in the SHRP2 dataset was used as the reference distribution for crash severity [84]. Data from CISS and GIDAS-PCM were then incorporated to enrich the representation of higher-severity scenarios that are sparsely observed in SHRP2.

In summary, data combination, together with sample weighting, enables the construction of reference datasets that exploit the complementary strengths of multiple data sources while mitigating individual biases. Initially, multiple intermediate reference datasets were constructed for different subsets of parameters. These intermediate datasets were then used to derive one or more reference datasets that maximize parameter coverage, since a single empirical dataset covering the entire parameter set may not always be available. Further methodological details are provided in Section III-C of Paper II.

Reference rear-end pre-crash datasets

In this thesis, two reference rear-end pre-crash datasets were constructed, as illustrated in Figure 2.4. The first dataset (blue) contains all six lead-vehicle-related parameters that define the lead-vehicle speed profile. The second dataset (green) contains five parameters describing both vehicles. These two datasets share two parameters and jointly cover nine parameters. As discussed earlier, a rear-end pre-crash scenario is parameterized as a twelve-dimensional vector. The remaining three parameters, associated with the following-vehicle behavior model, were characterized using separately acquired reference marginal distributions.

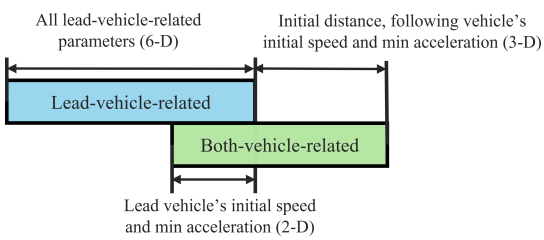

Figure 2.4: Two reference rear-end pre-crash datasets constructed through data combination and sample weighting.

2.3 Generation of representative synthetic pre-crash scenarios

The next step is to generate representative synthetic pre-crash scenarios using the reference datasets. The objective is to produce a sufficiently large and diverse set of scenarios suitable for SIA that preserves the multivariate characteristics of real-world rear-end pre-crashes across the full severity range.

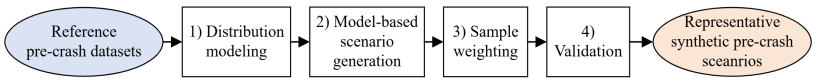

Figure 2.5: Workflow for generating representative synthetic pre-crash scenarios.

Figure 2.5 illustrates the overall methodological workflow adopted in this thesis. Distribution models were first built from reference pre-crash datasets to characterize the underlying parameter space. Synthetic pre-crash scenarios were then generated through model-based simulation using parameter configurations sampled from the distributions. The resulting scenarios did not automatically reproduce the reference population—due to simulation constraints and parameter matching procedures—so sample weighting was subsequently applied to align the synthetic dataset with the reference distributions. The final validation step would ideally focus on assessment-oriented methods to evaluate representativeness. However, in this case, validation was conducted through statistical comparisons with available reference datasets, as no comprehensive reference dataset covering the full parameter space was available.

2.3.1 Distribution modeling

To enable the generation of synthetic pre-crash scenarios, the reference datasets are first represented by probabilistic models from which an arbitrary number of samples can be drawn.

Parametric [96, 97], *nonparametric* [98, 99], and *copula-based* [100, 101, 102, 103] methods are commonly used to model a multivariate distribution. The parametric methods assume a specific parametric form (such as the multivariate normal distribution [96]) for the joint distribution, which can be advantageous when the underlying distribution is known or can be reasonably assumed. They often have fewer parameters than other methods, making them computationally efficient and less prone to overfitting when dealing with sparse data. However, they rely on strong assumptions about the underlying distribution, which might not hold in real-world scenarios. If the assumptions are violated, the parametric models can provide biased estimates. In addition, if the available data are sparse, they might not provide enough information to accurately estimate the chosen distribution's parameters.

Nonparametric methods, such as kernel density estimation [104] or nearest neighbor methods [105], make minimal assumptions about the underlying distribution and instead estimate it directly from the data. The methods might require more data to achieve reliable estimates than parametric methods, and

they may not always provide accurate density estimates, especially for sparse or irregularly sampled data [106]. In addition, nonparametric methods can produce biased estimates, particularly near boundaries and in distributions with long tails [107]. Finally, interpreting the results can be more challenging (compared to parametric methods) since the distribution does not have an explicit functional form.

Copula-based methods allow the flexible modeling of dependence structures between random variables; they are often used when the marginal distributions are known or estimated separately [103]. However, these methods usually require a relatively large amount of data to accurately estimate the parameters of the copula function, especially for high-dimensional datasets [101]. They are also generally less interpretable than simpler parametric methods.

In our situation, the amount of data available is limited, and some parameters are sparsely represented. Additionally, it is crucial to be able to interpret the distribution models and grasp the correlations among parameters, as these correlations have physical meaning and directly influence the generated scenario populations. Interpretable dependencies improve transparency and confidence in the assessment results because the relationships among parameters can be related to meaningful pre-crash dynamics.

For these reasons, in this thesis, a parametric multivariate distribution modeling method was adopted, with explicit emphasis on simplicity and interpretability to reduce the risk of overfitting and to facilitate transparent reasoning about parameter dependencies.

The adopted method models marginal distributions together with linear correlations among parameters. Specifically, linear correlations are first estimated for each parameter pair. Each parameter is classified as correlated if it exhibits a significant, non-weak correlation with at least one other parameter, and as uncorrelated otherwise. Uncorrelated parameters are modeled independently by fitting a set of candidate distributions (e.g., normal and gamma), with the optimal model selected using the Akaike Information Criterion [108]; the parameters exhibiting point-mass behavior—i.e., a concentration of observations at a specific value—are modeled using mixture distributions [84]. Correlated parameters are modeled jointly: their marginal distributions are first transformed to standard normal form via quantile transformation [109], after which a multivariate normal distribution is fitted. Additional information regarding the distribution modeling method can be found in Section III-C of Paper I.

Given the heterogeneity of empirical pre-crash data and the complexity of the multivariate parameter space, a single global multivariate distribution model would likely need to be highly complex to represent all observed behavioral and kinematic patterns. This challenge has been widely documented in traffic safety and crash modeling, where heterogeneous populations are often represented using stratified or finite mixture models rather than a single joint distribution [110, 111, 112]. Such approaches have been used to account for regime-dependent behavior and unobserved heterogeneity in crash processes, for example, by separating distinct crash types, severity regimes, or behavioral patterns prior to modeling.

To address this issue, each reference dataset was categorized into multiple

sub-datasets based on observed patterns, allowing simpler, more interpretable distribution models to be fitted. This strategy is consistent with the traffic simulation and validation literature, which recommends regime-specific modeling when global distributions obscure meaningful structure [35]. Details of the categorization procedure are provided in Section IV-C of Paper I and Section IV-B of Paper II.

The overall distribution model was accordingly constructed as a finite mixture of these sub-models, combined according to the proportions of their corresponding sub-datasets in the reference dataset [110]. In contrast to prior applications, which primarily use mixture models to improve statistical fit or predictive performance, the mixture representation in this thesis serves a methodological role: it provides an empirically grounded, interpretable basis for synthetic scenario generation and subsequent validation. Two multivariate distribution models were constructed, one for each of the two reference datasets obtained in the data combination step, forming the basis for the synthetic pre-crash scenario generation.

2.3.2 Model-based scenario generation

Once the reference datasets have been represented through probabilistic distribution models, the next methodological step is to generate synthetic pre-crash scenarios that are dynamically consistent, physically plausible, and statistically aligned with the reference data. *Model-based scenario generation* has been widely adopted in traffic safety and automated driving research as a means to reconstruct complete pre-crash trajectories from abstract scenario descriptions or parameter sets [31, 38].

Unlike purely data-driven resampling or trajectory-stitching approaches [22, 27], model-based scenario generation relies on simulation to translate parameter configurations into continuous-time vehicle motion. This approach allows for the enforcement of physical constraints, interaction dynamics, and behavioral assumptions during scenario generation. Simulation-based reconstruction is particularly important when scenarios are represented at the parameter level rather than as complete time-series trajectories, as is common in assessment-oriented scenario modeling and testing frameworks [113]. By combining probabilistic sampling in the parameter space with deterministic or stochastic simulation models, model-based generation provides a systematic framework for generating large numbers of pre-crash scenarios with internal consistency and physical feasibility [21, 33].

As described in Section 2.2.1, each rear-end pre-crash scenario in this thesis is represented as a twelve-dimensional parameter vector. These parameters describe the initial conditions and behavioral characteristics of the involved vehicles but do not directly encode full trajectories. The parameters must therefore be instantiated in a simulation environment to reconstruct physically meaningful pre-crash time series with coherent temporal evolution. In this way, the generated synthetic scenarios remain consistent with vehicle kinematics and the assumed behavioral constraints that cannot be guaranteed by direct statistical sampling of time-series data.

A key methodological consideration at this stage is how to construct coherent parameter combinations when no single joint reference distribution is available across all twelve dimensions. Direct sampling from an assumed full multivariate distribution would require strong assumptions and extensive data support, neither of which is available in this context. Instead, scenario generation should combine information from multiple partial distribution models to construct twelve-dimensional parameter configurations, preserving observed dependencies where joint reference information is available and avoiding unsupported higher-dimensional dependence assumptions for parameters lacking a joint reference distribution.

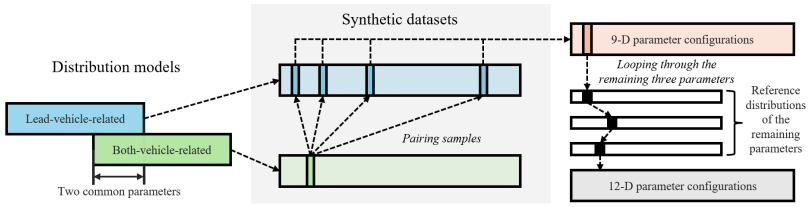

Figure 2.6: Procedure for constructing the twelve-dimensional parameter configurations for simulation-based scenario generation.

Based on this consideration, the procedure illustrated in Figure 2.6 was used. Two synthetic datasets were first generated by sampling from two established distribution models: a six-dimensional model describing all lead-vehicle-related parameters, which fully determine the lead-vehicle speed profile, and a five-dimensional model capturing (a subset of) parameters for both vehicles. The two datasets share two common parameters, which were used to pair samples across the datasets, resulting in nine-dimensional parameter configurations that preserve observed dependencies.

The remaining three parameters, which pertain to the following-vehicle behavior model only and lack a joint reference distribution, were populated by iterating over their respective reference marginal distributions. This strategy allows the scenario space to be explored without imposing unsupported assumptions about higher-dimensional dependence structures.

Each resulting twelve-dimensional parameter configuration was then instantiated in simulation to generate a synthetic pre-crash trajectory. A simulation was considered valid if it resulted in a rear-end crash occurring approximately five seconds after the start of the simulation, matching the temporal coverage of the empirical pre-crash data. Through this process, a synthetic dataset consisting of 5,000 valid pre-crash scenarios was generated and used in subsequent weighting and validation steps.

2.3.3 Sample weighting

When synthetic scenarios are obtained through model-based simulation rather than direct resampling, the resulting dataset may not inherit the statistical properties of the reference population. Selection effects can arise at multiple

stages of the generation framework—for example, when only a subset of parameter configurations leads to valid simulations, or when scenarios are assembled by combining information from multiple probabilistic models. Without correction, these effects can distort the parameter distributions and undermine the representativeness of the synthetic dataset.

Sample weighting provides a principled mechanism for addressing these distortions. By adjusting the relative contribution of individual synthetic scenarios, it aligns the synthetic dataset with the target reference distributions, even when the generation process is biased or incomplete. This step is particularly important when generating scenarios for SIA, because conclusions are drawn at the population level and therefore depend on the representativeness of the scenario population.

In many practical settings, the full joint distribution of all scenario parameters is unknown or cannot be reliably estimated due to data sparsity, heterogeneity, or partial observability. In such cases, weighting approaches that calibrate the weighted sample to known marginal distributions offer a pragmatic and widely accepted solution [114, 115]. *Iterative Proportional Fitting* (IPF) is a classical statistical technique developed for this purpose, originally in the context of contingency tables and later adopted in survey calibration, demography, and population synthesis [114, 116]. IPF iteratively adjusts sample weights so that the weighted dataset matches a set of target marginal distributions, without requiring explicit specification of the full joint distribution.

In this thesis, sample weighting was required because synthetic pre-crash scenarios were generated by simulating sampled parameter configurations rather than by directly resampling reference data. Not all sampled configurations resulted in valid simulations—for example, some did not lead to a crash within the modeled time window. As a result, retaining only valid simulations introduced selection effects that altered the resulting parameter distributions. Moreover, complete parameter vectors were constructed by matching samples drawn from different distribution models, which does not guarantee that the resulting set of simulated scenarios accurately reflects the reference population. Sample weighting was therefore applied to correct for these distortions by aligning the synthetic dataset with the reference marginal distributions obtained during the distribution modeling step (Section 2.3.1).

An additional consideration is that each reference dataset was modeled as a mixture of multiple sub-datasets, reflecting heterogeneous patterns in the empirical data. Consequently, weighting needs to preserve not only the marginal distributions of individual parameters within each sub-dataset, but also the relative proportions of the sub-datasets themselves. Therefore, an IPF-based weighting procedure was developed that enforces alignment with the reference marginal distributions at the sub-dataset level while maintaining the mixture structure of the synthetic dataset. Further details of this procedure are provided in Section III-E of Paper II.

After applying the sample weighting process, a weighted synthetic dataset consisting of 5,000 rear-end pre-crash scenarios was obtained and made publicly available [117].

2.3.4 Statistical validation against available reference data

Validation, the final methodological step in the proposed framework, aims to evaluate whether the weighted synthetic pre-crash scenarios are sufficiently consistent with empirical evidence for the purposes of SIA. Ideally, this involves assessing whether the synthetic scenarios are representative of real-world pre-crash conditions for the intended assessment.

However, in practice, validation of synthetic pre-crash scenarios is constrained by the availability and scope of the reference data. As discussed in Section 1.3.2, ideally, validation would be performed against a comprehensive reference dataset that covers the full set of scenario parameters. When such data are unavailable, validation necessarily relies on partial reference datasets and parameter subsets and must align with how the synthetic scenarios were constructed. Validation conducted under these constraints, as in this thesis, is therefore difference-oriented. Rather than evaluating practical equivalence, the objective is to test whether the weighted synthetic dataset is statistically distinguishable from the available reference distributions for the corresponding parameter subsets. Accordingly, the validation should be performed at the level of those parameter subsets for which reliable reference data exist, using statistical hypothesis tests to identify significant differences between synthetic and reference distributions.

In this thesis, distribution modeling and synthetic scenario generation were performed separately for multiple sub-datasets, each defined over a subset of parameters. Validation was therefore conducted at the sub-dataset level, ensuring coherence between the modeling assumptions, the generation process, and the available empirical evidence. This approach provides a conservative diagnostic check for distributional mismatches, while acknowledging that statistical difference detection alone cannot establish assessment-oriented representativeness.

The complete validation process combined descriptive analysis, visualization, and difference-oriented statistical testing. Descriptive statistics and visualization techniques, such as comparisons of means and variances or t-distributed stochastic neighbor embedding (t-SNE) [59], provided exploratory insight into distributional similarity. In addition, the statistical tests were applied from three complementary perspectives:

1. Marginal distributions: For each parameter within each sub-dataset, weighted two-sample Kolmogorov–Smirnov (KS) tests were conducted to assess whether the weighted synthetic marginal distributions were (statistically) significantly different from the corresponding reference distributions at the 0.05 significance level.

2. Multivariate distributions: For each sub-dataset containing multiple parameters, t-SNE was used to project the multidimensional data into a unidimensional representation. Weighted two-sample KS tests were then applied to the transformed data to test whether the synthetic and reference datasets were significantly different at the 0.05 significance level.

3. Crash severity distribution: The distribution of crash severity, represented

by the lead-vehicle Delta-v, was compared between the synthetic and reference datasets using a weighted two-sample KS test at the 0.05 significance level.

Again, while difference-based analyses are useful for identifying mismatches between synthetic and reference data or distributions, they do not by themselves indicate whether such mismatches are practically meaningful for the intended SIA. In this thesis, assessment-oriented validation is constrained by the absence of a comprehensive reference dataset covering the full parameter space of the target pre-crash population. As a result, validation at this stage combines statistical comparison against available reference distributions with cautious interpretation [45]. More generally, determining whether observed differences are acceptable for a given assessment purpose requires explicit criteria that relate statistical deviations to their practical relevance. When complete reference data are unavailable, such criteria cannot be derived empirically from the target population alone and must instead rely on additional assumptions or external justification. The implications of this limitation, and how assessment-oriented validation can be approached under different data availability conditions, are discussed in Section 4.2.4.

2.4 Validation of representativeness

Assessing the representativeness of synthetic pre-crash scenarios requires methods that go beyond detecting differences and instead evaluate whether observed discrepancies are practically acceptable for the intended SIA. Equivalence testing methods are well suited to this application, since they are specifically designed to determine whether differences between datasets fall within predefined, practically acceptable bounds [54, 55]. Unlike conventional null-hypothesis significance testing, which is designed to detect deviations from equality, equivalence testing reframes the validation question as whether two datasets are sufficiently similar for a given purpose.

Equivalence testing methods have been developed in both frequentist and Bayesian settings. Frequentist approaches typically rely on interval-based procedures, such as two one-sided tests (TOST), to assess whether observed differences lie within prespecified equivalence margins [118]. Bayesian approaches, in contrast, quantify equivalence probabilistically by evaluating posterior uncertainty relative to practically meaningful thresholds [119, 56]. A review of existing equivalence testing approaches and their applicability to scenario validation was conducted in Paper III.

Bayesian approaches based on the *Region of Practical Equivalence* (ROPE) [56] are particularly well suited for validating the representativeness of pre-crash scenarios used in SIA. These methods offer an intuitive interpretation of similarity and allow domain-specific relevance thresholds to be incorporated directly into the analysis [119, 62]. Rather than testing whether two samples (in this case, synthetic and reference datasets) originate from exactly the same distribution, these methods explicitly assess whether differences between them

are sufficiently small to be considered practically negligible for the intended assessment.

In a ROPE-based framework, equivalence is evaluated by comparing the *highest density interval* (HDI) of a chosen statistic with a predefined ROPE representing practically acceptable values. The HDI is the shortest interval of the posterior distribution containing a specified proportion (e.g., 95%) of the posterior probability mass. The chosen statistic may, for example, be a difference in means, variances, or exceedance probabilities. Equivalence is supported when the entire HDI lies within the ROPE, yielding a probabilistic, uncertainty-aware statement of practical similarity directly interpretable in the context of SIA.

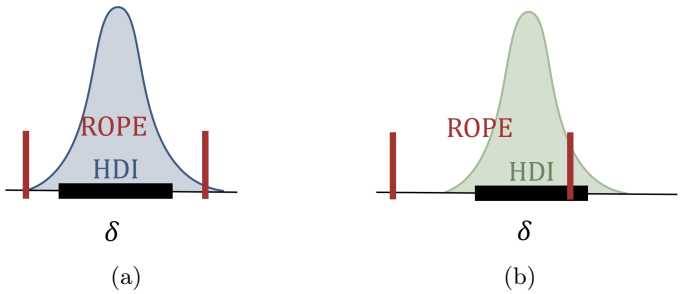

Figure 2.7: Illustration of the ROPE concept in Bayesian equivalence testing: (a) equivalence and (b) non-equivalence [53]. The shaded curves represent posterior distributions of a statistic, δ. The red vertical lines denote the ROPE boundaries, and black bars indicate the HDIs. Equivalence is supported when the HDI lies entirely within the ROPE.

Figure 2.7 illustrates the underlying logic of ROPE-based equivalence testing. When the HDI (black bar) is fully contained within the ROPE (red vertical lines), the observed difference is considered practically negligible; when it extends beyond the ROPE, equivalence is not supported.

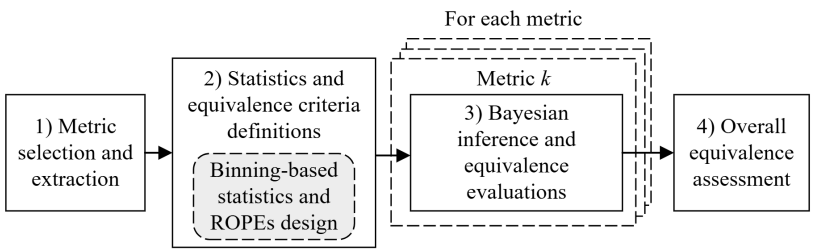

Figure 2.8: Workflow of the practical equivalence testing framework [53].

Building on this general principle, a structured practical equivalence testing framework is proposed to validate the representativeness of synthetic pre-crash scenarios for SIA, as illustrated in Figure 2.8. The framework is explicitly assessment-oriented: the selection of metrics, statistics, and equivalence criteria

should be guided by the intended use of the scenarios, such as estimating changes in crash risk or crash severity after the introduction of a DAS. Accordingly, practical equivalence testing requires an appropriate reference dataset or reference distribution for the metrics under consideration.

Note that, in this thesis, the ROPE-based equivalence testing framework is applied with different intents across Papers III and IV. In Paper III, the primary objective is methodological: to contrast conventional difference-oriented significance testing with equivalence-based reasoning. The ROPE thresholds in that paper were therefore intentionally specified to highlight potential discrepancies between statistical significance and practical equivalence, rather than to represent realistic acceptance criteria for a specific safety application. In Paper IV, the focus shifts from methodological comparison (aimed at highlighting a specific issue) to application-driven validation (aimed at demonstrating how validation can be carried out in practice). Accordingly, ROPEs are defined to reflect realistic tolerances for a concrete safety impact assessment task—specifically, the evaluation of an AEB system—by linking equivalence bounds to assessment-relevant outcome measures.

2.4.1 Metric selection and extraction

The first step in the proposed framework is the selection and extraction of scenario metrics that are relevant for the intended SIA. The metrics serve as the quantitative interface between pre-crash scenarios and assessment outcomes: they characterize aspects of scenario kinematics, conflict evolution, or severity that influence system behavior or predicted safety effects. For this reason, metrics are derived either from the pre-crash time series or from the parameterized scenario representation, yielding quantities that are directly comparable across reference and synthetic datasets.

Metric selection in the context of practical equivalence testing is inherently purpose-driven rather than exhaustive. Although including many metrics may appear to strengthen validation coverage, doing so increases the probability that at least one metric will fail the equivalence criterion due to random variation or minor, assessment-irrelevant discrepancies. When overall representativeness is assessed by combining individual tests, this multiplicity effect can lead to systematically conservative conclusions [55].

In classical hypothesis testing, multiple comparisons are often handled by formal multiplicity corrections (e.g., Bonferroni) that adjust the significance level [120]. In Bayesian ROPE-based equivalence testing, an analogous adjustment targets the posterior probability threshold (i.e., the posterior mass required within the ROPE) rather than the ROPE itself. The ROPE reflects domain-specific tolerances for practical equivalence and should be prespecified; widening it solely to offset multiple comparisons would alter the substantive definition of practical equivalence. Accordingly, multiplicity can be addressed by releasing the posterior probability threshold, while careful, relevance-based metric selection serves as a principled mechanism for mitigating unnecessary multiplicity and preserves the assessment-oriented meaning of the equivalence criteria [54, 56].

As a result, the validation framework deliberately focuses on a limited number of metrics that are most informative for the assessment objective. Metric selection is guided by three core considerations: 1) safety relevance (whether the metric influences predicted safety outcomes or system interventions); 2) interpretability (ensuring that deviations can be meaningfully related to driving behavior or system performance); and 3) robustness (the metrics should be stable under measurement noise, minor alignment differences, and modeling uncertainty) [38, 53]. The framework prioritizes relevance over quantity, reducing the risk that inconsequential metrics dominate the validation outcome.

In practice, selected metrics may span multiple conceptual levels: 1) kinematic descriptors (e.g., speeds, accelerations, headways, relative speed), 2) conflict and criticality indicators (e.g., time-to-collision-related measures), 3) crash severity proxies (e.g., Delta-v, injury risk estimates), and 4) categorical or event-based quantities (e.g., occurrence of harsh braking or the share of high-risk scenarios) [31, 38]. Comparability is ensured by extracting the metrics in both reference and synthetic datasets using identical definitions and preprocessing steps, including consistent filtering, inclusion criteria, and temporal alignment [62].

By constraining validation to a carefully chosen set of assessment-relevant metrics, the framework balances coverage, interpretability, and statistical reliability. This approach supports transparent equivalence judgments while limiting the influence of multiple-comparison effects that could otherwise obscure the overall representativeness assessment.

2.4.2 Statistics and equivalence criteria definitions

For each selected metric, one or more statistics are defined to quantify differences between the synthetic and reference datasets. Typical choices include differences in central tendency (e.g., means or medians), differences in dispersion or quantiles, ratios or differences in exceedance probabilities (such as the proportion of high-severity scenarios), and other distributional summaries that are directly interpretable in terms of system behavior or safety outcomes [55, 38]. The selection of statistics should therefore be guided by assessment relevance rather than statistical convenience.

Equivalence criteria are specified by combining a *posterior probability threshold* with a ROPE. Within a Bayesian framework, posterior samples of each statistic are obtained by propagating uncertainty from the fitted reference and synthetic models. From these samples, an HDI is computed to summarize posterior uncertainty [119]. Practical equivalence is supported if the HDI corresponding to the chosen posterior probability threshold (95% in this thesis) lies entirely within the predefined ROPE. The ROPE itself is specified based on domain knowledge, sensitivity analyses, or assessment requirements and represents the range of differences considered practically negligible for the intended SIA.

In this thesis, equivalence testing was conducted independently for each metric. This design choice prioritizes transparency and interpretability by avoiding the need to define equivalence regions for high-dimensional joint

statistics, which would require substantially more data and stronger modeling assumptions [55]. Metric-level testing also facilitates diagnostic interpretation, as non-equivalence can be directly attributed to specific aspects of scenario behavior. Equivalence conclusions are therefore reported on a per-metric basis, while the implications of not explicitly modeling dependencies between metrics are addressed in the discussion.

2.4.3 Bayesian inference and equivalence evaluations

Once metrics and equivalence criteria have been defined, the next step is to quantify uncertainty in the selected statistics and evaluate practical equivalence between the synthetic and reference data for each metric. From a methodological perspective, the goal is not merely to obtain point estimates of differences between datasets, but to characterize the range of plausible differences and assess whether this range lies within assessment-relevant tolerances [55].

Bayesian inference provides a natural, coherent framework for this purpose. By treating model parameters and derived statistics as random variables, Bayesian methods allow uncertainty in the data and modeling assumptions to be propagated consistently into the equivalence evaluation [119]. This uncertainty propagation is crucial in situations where empirical data is limited, and validation decisions must be made under uncertain, rather than ideal, large-sample conditions. Moreover, Bayesian inference aligns naturally with equivalence testing, as it enables probability statements about whether differences fall within predefined regions of practical equivalence [56, 119].

In this step, Bayesian inference is applied independently to each selected metric. For a given metric, a set of candidate probabilistic models (e.g., exponential, normal, log-normal, gamma, or mixture distributions) is specified based on prior knowledge of the data and expert judgment. These models are fitted separately to the reference and synthetic data. When weighted data are used, the weights are incorporated directly into the inference via a weighted likelihood formulation. Each observation contributes to the log-likelihood in proportion to its assigned weight, ensuring that the resulting posterior distributions reflect the intended representativeness captured by the weighting scheme. Model comparison techniques, including the Widely Applicable Information Criterion (WAIC) [121] and leave-one-out cross-validation (LOO) [122], are used to select an appropriate model for each dataset and metric.

Posterior predictive samples are then drawn jointly from the selected reference and synthetic models, producing paired draws for the selected metric. Each paired draw reflects how the metric varies under parameter uncertainty in the two models. The comparison statistics are then computed directly from these paired draws, ensuring that uncertainty in both posterior predictive distributions is consistently propagated into the evaluation of differences between the datasets.

Equivalence is evaluated by comparing the HDI of each statistic's posterior distribution with the corresponding ROPE. If the entire HDI lies within the predefined ROPE for all statistics associated with a metric, the metric is regarded as practically equivalent across the synthetic and reference datasets

(for the intended SIA); otherwise, it is classified as non-equivalent.

2.4.4 Overall equivalence assessment

The final step of the framework synthesizes the metric-level equivalence results into an overall assessment of representativeness that is explicitly aligned with the intended use of the synthetic dataset. Rather than relying on an implicit or universal acceptance criterion, the overall decision rule is formulated to reflect the specific priorities, tolerances, and objectives of the intended SIA.

In its most conservative form, overall practical equivalence is supported only if all selected metrics satisfy their respective equivalence criteria. Depending on the assessment context, however, less restrictive decision rules may be applied. For example, 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 deviations observed in the remaining metrics are demonstrably limited and not expected to materially influence the assessment outcomes. Crucially, both the prioritization of metrics and the conditions under which deviations are acceptable must be explicitly specified in the validation protocol, rather than inferred post hoc.

The overall equivalence assessment should be accompanied by structured documentation to ensure that the assessment objective is transparent, reproducible, and traceable. At a minimum, this documentation should clearly state: 1) the selected metrics and the rationale for their inclusion, 2) the specified ROPEs and posterior probability thresholds together with their justification, and 3) the resulting posterior evidence supporting or rejecting equivalence at both the metric and aggregate levels. When appropriate, additional analyses—such as sensitivity checks examining the robustness of conclusions to reasonable alternative choices (e.g., variations in ROPE widths, prior assumptions, or metric prioritization)—may also be included to further strengthen transparency and interpretability. Explicit documentation of these elements enables the validation process to be critically examined and ensures that acceptance decisions are clearly grounded in the stated SIA purpose.

2.4.5 Binning-based statistics and ROPEs design

A central methodological challenge in representativeness validation is that many comparison approaches implicitly treat all parts of a metric's distribution as equally important [54, 55, 35]. However, this assumption is rarely justified for synthetic pre-crash scenarios used in SIA. Different regions of a distribution often correspond to qualitatively different safety regimes, and discrepancies in these regions can have markedly different implications for system evaluation [53]. Uniform aggregation of deviations therefore risks either obscuring assessment-critical differences or overstating negligible ones.

This challenge is closely tied to how equivalence criteria are specified. Equivalence bounds placed directly on conventional global-summary statistics—such as differences in means, variances, or overall distributional distances—are often difficult to interpret in safety-relevant terms. For example, a small

difference in the mean time-to-collision may appear negligible overall yet mask substantial discrepancies in the tail of the distribution, where rare but safety-critical events occur.

At the other extreme, purely data-driven distance statistics—such as KS statistics or divergence measures—quantify overall dissimilarity without requiring explicit equivalence bounds. Although useful for detecting differences, these measures typically offer limited insight into where discrepancies arise or whether they matter for the intended assessment. A large distance may reflect widespread but inconsequential differences, whereas a small distance may conceal localized mismatches in critical regions.

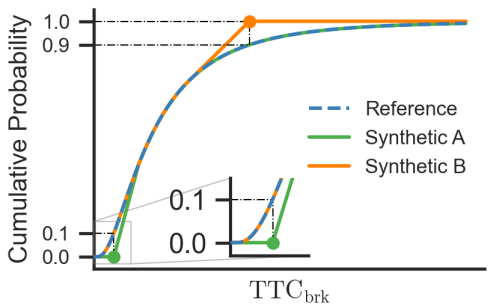

Figure 2.9: Cumulative distributions of TTC_{brk} for a reference dataset and two synthetic datasets. Both synthetic datasets have the same KS statistic (0.1), but their deviations occur in different regions of the distribution, leading to different practical implications for SIA.

From a methodological perspective, these statistics—whether conventional global summaries or data-driven distances—provide only indirect insight into how statistical deviations relate to their practical relevance for the intended SIA, thereby limiting the defensibility of the resulting validation conclusions. To make this issue concrete, Figure 2.9 illustrates this issue using time-to-collision at braking onset (TTC_{brk}) under normal driving conditions. TTC_{brk} is defined as the instantaneous time-to-collision evaluated at the moment when the driver initiates braking, and it serves as an interpretable indicator of temporal criticality at the onset of conflict mitigation [123]. The figure compares the cumulative TTC_{brk} distributions of a reference dataset with two synthetic datasets. By construction, both synthetic datasets yield the same KS statistic (0.1) with respect to the reference, indicating an identical maximum vertical distance between the cumulative distributions; yet, the nature of the deviations is qualitatively different. For Synthetic A, the maximum deviation occurs in the lower range of TTC_{brk}, corresponding to higher urgency conditions. For Synthetic B, the same KS value arises from deviations in the upper range of TTC_{brk}, associated with lower urgency conditions.

From an SIA perspective, however, these two deviations are not equivalent. Differences in the lower tail of TTC_{brk} directly affect the frequency of

high-urgency situations that are most influential for system triggering, residual crash risk, and estimated safety benefits, particularly for functions such as AEB. In contrast, deviations of similar magnitude in the upper tail primarily affect benign driving conditions and are typically of limited consequence for assessment outcomes. Thus, the same numerical statistical difference can correspond to markedly different practical impacts. This example highlights that global, unweighted statistics can obscure whether observed differences occur in assessment-critical or assessment-irrelevant regions, thereby providing limited guidance for determining practical equivalence.

This observation motivates the binning-based statistics introduced below. By decomposing a metric's distribution into localized, interpretable regions and scaling deviations according to their relevance to the intended assessment, the proposed approach makes explicit the value judgments that conventional global summaries and data-driven distance statistics leave implicit—namely, which parts of the distribution are considered more consequential for the assessment objective. This structured decomposition links deviations to specific regions and their assessment relevance, enabling transparent alignment between empirical discrepancies and assessment intent while retaining robustness through controlled bin-level aggregation.

Importantly, the framework applies two generic, metric-independent *binning-based statistics* together with two corresponding ROPEs uniformly across all metrics. Assessment specificity is introduced through the design of the binning scheme and its associated bin weights, which determine how localized deviations contribute to the overall equivalence assessment. This separation between what is compared (the generic statistics) and how assessment priorities are expressed (the binning and weighting design) enhances both interpretability and consistency across metrics.

The design presented in Paper IV, therefore, strikes a balance between flexibility and standardization. It avoids reliance on opaque, single-number discrepancy measures while eliminating the need to redefine statistics and equivalence criteria for each metric. The following sections describe the binning procedure, the construction of the two statistics, and the corresponding ROPE design in detail.

Binning procedure

Binning is performed based on the fitted Bayesian reference distribution model for a given metric, rather than directly on the raw reference data. This choice propagates uncertainty in the reference distribution and avoids sensitivity to sampling noise. Specifically, for each paired posterior draw from the Bayesian models fitted to the reference and synthetic datasets, the posterior predictive sample from the reference model is partitioned into N bins defined by quantiles, so that each bin contains approximately the same proportion of the posterior predictive probability. The resulting bin boundaries are then applied unchanged to the paired posterior sample from the synthetic model. This procedure ensures that binning reflects the posterior-implied reference distribution and that uncertainty in both the reference and synthetic distributions is consistently

propagated into the binning-based statistics.

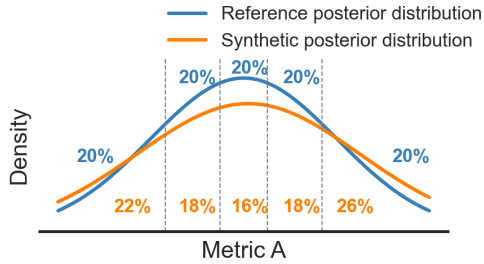

Figure 2.10: Illustration of the binning procedure for a single metric with $N = 5$. Bin boundaries (dashed lines) are defined by equal-probability partitioning of the reference posterior distribution. Blue and orange percentages denote the posterior bin probabilities $P_{\mathrm{ref},i}$ and $P_{\mathrm{syn},i}$, respectively.

Figure 2.10 illustrates this procedure for a single metric. Applying these boundaries to the synthetic distribution yields paired bin proportions that enable a direct, one-to-one comparison between the two distributions. For each posterior draw, the reference distribution is re-partitioned in this manner, and the resulting bin boundaries are used to compute the bin probabilities. Note that although the bins contain equal probability mass for the reference distribution (e.g., 20% per bin in Fig. 2.10), the corresponding bin proportions for the synthetic distribution may differ, revealing localized over- or under-representation across the metric's range.

Statistics

Let $P_{\mathrm{ref},i}$ and $P_{\mathrm{syn},i}$ denote the proportions of posterior samples falling into the i-th bin for the reference and synthetic data, respectively, and define the bin-level proportion difference as

$$\Delta P_i = P_{\mathrm{syn},i} - P_{\mathrm{ref},i}.$$

Because both the bin boundaries and the resulting bin probabilities are recomputed for each posterior draw, ΔP_i is inherently stochastic, and its uncertainty is fully reflected in the posterior distributions of the derived equivalence statistics.

Not all bins are equally relevant for a given SIA. The practical relevance of each bin depends on the specific assessment objective. To capture this heterogeneity, bin weights (further described in Section 2.3 of Paper IV) are introduced. The bin weights are determined empirically by linking deviations in each bin to their potential impact on downstream assessment outcomes. The steps involved in this empirical approach are as follows:

1. Re-simulate the reference scenarios under a virtual representation of the DAS.

2. Identify a set of re-simulation outcome measures most relevant to the assessment purpose (e.g., residual impact speed or injury risk).

3. Define a weight function $f(\cdot)$ that maps these outcome measures to a bin weight given the binning done for each posterior draw of the fitted reference model,
$$\omega_i = f(X_i),$$
where X_i denotes the identified set of assessment-relevant outcome measures associated with the i-th bin.

As with the bin boundaries, the bin weights are recalculated for each posterior draw, ensuring that uncertainty in the system response and in the assessment's relevance is propagated into the equivalence statistics.

Building on the bin-level deviations and weights, the two statistics are defined as
$$\theta = \max_{1 \le i \le N} \left(\left| \frac{\Delta P_i}{P_{\mathrm{ref},i}} \right| \cdot \omega_i \right), \tag{2.1}$$
and
$$\Theta = \sum_{i=1}^{N} |\Delta P_i| \cdot \omega_i. \tag{2.2}$$
Here, θ captures the maximum weighted absolute relative deviation across bins, providing a worst-case localized perspective on practical difference, while Θ captures the total weighted absolute deviation, providing an aggregate measure of distributional discrepancy.

Crucially, the bin boundaries and bin weights are intrinsic components of the definitions of θ and Θ. Because both are recomputed for each posterior draw based on the posterior-implied reference distribution, the resulting posterior distributions of θ and Θ fully reflect uncertainty in the underlying distributions, the binning process, and the assessment-relevant weighting.

ROPEs

ROPEs are defined on the statistics θ and Θ. The common thresholds θ_{thd} and Θ_{thd} can be applied across metrics because, although the bins are partitioned separately according to the posterior-implied reference distribution of each metric, the bin weights are consistently determined by the same weight function that links statistical differences to practical influence. This consistency ensures the comparability of the two statistics across metrics.

Additionally, a practical rule for setting ROPEs is proposed (see Section 3.4 of Paper IV), which specifies tolerances for a baseline bin and derives ROPE thresholds from them. A baseline bin—defined as the bin assigned a standard weight of $\omega_{\mathrm{b}} = 1$—serves as a reference point for interpreting deviations; its weight reflects neither amplification nor attenuation of the practical relevance.

For the baseline bin, the user should first interpret what this condition represents in practical terms under the chosen bin-weight function, ensuring that any specified tolerance is meaningful within the assessment context. The next step is to decide the acceptable range of deviations—namely, the absolute

relative deviation $|\Delta P/P_{\text{ref}}|$ and the absolute deviation $|\Delta P|$ (based on the definitions of θ and Θ)—that should be regarded as practically equivalent for the baseline bin. The tolerances are denoted as $|\Delta P/P_{\text{ref}}|_{\text{thd}}$ and $|\Delta P|_{\text{thd}}$.

Specifically, in the demonstration presented in Section 3 of Paper IV, the baseline bin corresponds to an average baseline Maximum Abbreviated Injury Scale (MAIS) 2+ injury risk $P_0 = 0.02$. We considered the bin's tolerances as $|\Delta P/P_{\text{ref}}|_{\text{thd}} = 10\%$ and $|\Delta P|_{\text{thd}} = 5\%$. This means that, for the baseline condition with an expected injury risk of $P_0 = 0.02$, differences within $\pm 10\%$ in relative and $\pm 5\%$ in absolute terms are regarded as practically negligible–i.e., deviations too small to materially influence the estimated safety outcomes.

These baseline tolerances are then used to derive the ROPE thresholds for the aggregated statistics θ and Θ (i.e., θ_{thd} and Θ_{thd}):

$$\theta_{\text{thd}} = |\Delta P/P_{\text{ref}}|_{\text{thd}} \cdot \omega_{\text{b}} = |\Delta P/P_{\text{ref}}|_{\text{thd}}\,, \tag{2.3}$$

$$\Theta_{\text{thd}} = |\Delta P|_{\text{thd}} \cdot \omega_{\text{b}} = |\Delta P|_{\text{thd}}\,. \tag{2.4}$$

In the mentioned demonstration, 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 [53]. This balance follows the general principle that ROPE boundaries should be defined based on domain-relevant effect sizes—that is, differences deemed practically negligible for the assessment purpose, rather than on arbitrary statistical thresholds [119, 56, 53].

Because bin weights scale deviations according to their practical relevance, the resulting ROPEs consistently reflect both the magnitude of distributional differences and their assessment-specific importance. This approach provides a transparent and interpretable mechanism for defining ROPEs while maintaining comparability across metrics and assessment objectives.

Diagnostic capability

Beyond producing an overall equivalence judgment, the binning-based statistics are designed to support method development and refinement by providing structured diagnostic insight. By decomposing distributional differences into bin-level deviations and their weighted contributions, the framework reveals where discrepancies arise within a metric's distribution and how they contribute to non-equivalence. This enables a clear distinction between localized mismatches in specific regimes (e.g., extreme or rare conditions) and more diffuse shifts affecting broader portions of the distribution.

Conventional statistics, such as the KS statistic, offer only limited localization by identifying the point at which the maximum difference between cumulative distributions occurs. However, such measures reduce distributional differences to a single extreme deviation and provide no insight into why non-equivalence arises—whether it is driven by a narrow region, by several moderate deviations dispersed across the distribution, or by systematic shifts in assessment-relevant regimes. In contrast, the binning-based approach provides

a richer diagnostic decomposition by quantifying deviations across predefined regions of the metric space and explicitly linking them to assessment-oriented weighting schemes.

Such diagnostic capability is particularly valuable for the iterative development of scenario generation, weighting, and modeling strategies, as it allows specific sources of representativeness loss to be identified and addressed rather than being obscured by aggregate summary measures. In this sense, validation is not treated as a terminal pass/fail check, but as an integral component of a feedback loop that informs systematic improvement of synthetic scenario generation methods.

Chapter 3

Summary of Included Papers

3.1 Paper I

Modeling lead-vehicle kinematics for rear-end crash scenario generation

Introduction

The use of virtual safety assessment as the primary method for evaluating DAS has emphasized the importance of crash scenario generation. One of the most common types of crashes is the rear-end crash, which involves a lead vehicle and a following vehicle. Most studies have focused on the following vehicle, assuming that the lead vehicle maintains a constant acceleration/deceleration before the crash. However, there is no evidence for this premise in the literature. This study aims to address this knowledge gap by thoroughly analyzing and modeling the lead vehicle's behavior as a first step in generating rear-end crash scenarios.

Methods

A piecewise linear model was employed to represent the lead-vehicle speed profile during the pre-crash phase, offering a more accurate digital representation of the lead-vehicle kinematics compared to the conventional constant acceleration/deceleration model. Two datasets were combined to produce a comprehensive dataset of rear-end critical incidents (crashes/near-crashes) that captures the full severity range, from physical contact to severe injuries or fatalities. Multivariate distribution models were constructed to generate synthetic lead-vehicle speed profiles, which were compared with the raw speed profiles.

Results

The results show that the piecewise linear model has good fitting performance. The raw and synthetic incidents display a notable alignment. Moreover, a range of different lead-vehicle speed patterns was revealed, indicating that

the proposed piecewise linear model is more accurate than the conventional constant acceleration/deceleration model. For example, the lead vehicle could exhibit harsh braking followed by gentle braking or even acceleration; however, it does not necessarily brake harshly. In fact, in many cases, the lead vehicle maintains a constant speed or remains stationary for a considerable time (up to five seconds) prior to the crash.

Conclusions

The proposed lead-vehicle kinematics model accurately matches lead-vehicle kinematics from in-depth pre-crash/near-crash data across the full severity range, outperforming previously existing lead-vehicle models in terms of both severity range and precision. Furthermore, in addition to generating simulated rear-end crash scenarios, this model has the potential to aid substantially in the reconstruction of individual real-world crashes. That is, the model provides a means of generating a distribution of realistic speed profiles during the reconstruction process, rather than generating only a single speed profile.

3.2 Paper II

Model-based generation of representative rear-end crash scenarios across the full severity range using pre-crash data

Introduction

Generating representative rear-end pre-crash scenarios is crucial for the SIAs of DAS. However, existing methods face challenges such as sparse, severity-biased pre-crash data and difficulties in formal validation. This study sought to overcome these challenges by combining naturalistic driving data and pre-crash kinematics data from rear-end pre-crashes to create a representative distribution of rear-end pre-crash scenarios in the United States. The resulting distribution can be used not only for the SIAs of DAS, but also as a benchmark when evaluating the representativeness of scenarios generated through other methods.

Methods

The process of generating synthetic rear-end pre-crash scenarios consists of three steps: 1) parameterizing the rear-end pre-crashes through modeling the two involved vehicles, 2) building distribution models for the parameterized crash data, and 3) generating representative synthetic crash scenarios.

In the first step, a following-vehicle behavior model was developed by combining two existing driver models. A twelve-dimensional vector representing the kinematics of the pre-crash phase of a rear-end pre-crash was then created by combining the newly created model with the lead-vehicle kinematics model from a previous study and the initial states of rear-end crash scenarios.

In the second step, parameterized pre-crash data from multiple pre-crash datasets were combined and weighted to create a reference dataset of the initial states and minimum fitted accelerations of both vehicles (REF_b). A synthetic dataset containing these data (REF_sb) was then created by sampling from the distribution model built for REF_b.

Lastly, simulations were conducted using the following-vehicle behavior model and two synthetic datasets, REF_sb and REF_sl. (The latter is a representative synthetic rear-end pre-crash lead-vehicle speed profile dataset created in a previous study.) Valid simulations were gathered and weighted using an IPF-based weighting procedure to create a representative synthetic rear-end pre-crash dataset.

Results

Sixty-one weighted two-sample KS tests were conducted to compare the synthetic dataset with reference datasets. The only two showed a significant difference, because the following vehicle's acceleration model could not imitate aggressive acceleration in certain situations. Comparing the weighted datasets for Δv_l showed no significant differences.

Conclusions

None of the available pre-crash datasets in this study contains all the necessary signals or is free of significant bias; these issues are commonly encountered in data-driven studies. To resolve them, we proposed a set of methods to combine and weight data from multiple pre-crash datasets. These methods create a reference dataset of the initial states and minimum accelerations of the following and lead vehicles across the full severity range. Moreover, the data were weighted to match a reference dataset using the KNN-based sample weighting method, thereby reducing bias. This weighting method is particularly noteworthy because, unlike conventional post-stratification methods, it can be used to weight biased data to match a reference dataset even when some strata are omitted. The data combination methods we propose can also be applied to other situations with biased datasets and/or incomplete signals that require a multivariate joint distribution.

In addition, the representative rear-end pre-crash dataset created can be used for the safety assessments of DAS, as well as serving as a benchmark when evaluating the representativeness of scenarios generated through other methods.

3.3 Paper III

Practical equivalence testing and its application in synthetic pre-crash scenario validation

Introduction

Reliable SIAs of DAS require baseline pre-crash scenarios that accurately represent real-world crash conditions. However, a gap remains in the robust evaluation of the similarity between synthetic and real-world pre-crash scenarios and their crash characteristics, such as Delta-v and injury risk. One reason for this validation gap is the lack of focus on methods to confirm that the synthetic test scenarios are practically equivalent (or "similar enough") to their real-world counterparts for the intended assessment purpose. Existing validation practices rely heavily on statistical significance testing, which can detect differences but cannot establish equivalence. This study addresses the validation gap by proposing a practical equivalence testing framework based on ROPE, which is demonstrated through the testing of practical equivalence between two rear-end pre-crash datasets.

Methods

The proposed ROPE-based framework aims to assess whether a synthetic pre-crash dataset is practically equivalent to a reference dataset in terms of scenario metrics relevant for SIA. Specifically, it consists of four steps: 1) metric selection and extraction, 2) statistics and equivalence criteria definitions, 3) Bayesian inference and equivalence evaluations, and 4) overall equivalence assessment.

In the first step, metrics most relevant to the intended assessment are selected and extracted from both datasets. Because pre-crash data consist of multivariate time-series signals, the analysis focuses on derived metrics—such as crash severity indicators, temporal criticality measures, and braking behavior—that capture how the scenarios influence the performance of the system under assessment.

In the second step, for each metric, at least one statistic (e.g., mean differences, KS statistics, or ratios of high-risk scenarios) is defined, together with practical equivalence criteria. These criteria are specified through ROPEs and a posterior probability threshold (typically 95%), ensuring that equivalence is judged based on differences that are practically negligible within the assessment scope.

In the third step, 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 separately fitted to the reference and synthetic data. Posterior samples of each statistic are then computed based on the selected optimal model, and practical equivalence is established if the 95% highest-density interval of the statistic lies entirely within its ROPE.

Finally, in the fourth step, the results across all metrics are integrated into an overall equivalence assessment following predefined decision rules. In general,

two datasets are considered practically equivalent if all metrics individually demonstrate equivalence. However, in practice, a less stringent approach may be justified. Regardless of the approach taken, it is essential that the rationale for declaring equivalence or non-equivalence is thoroughly documented, including the motivations behind the selection of metrics, statistics, ROPEs, and the overall equivalence criteria.

Results

The proposed framework was demonstrated by comparing two rear-end pre-crash datasets: a reference dataset consisting of 5,000 scenarios and the GIDAS-PCM dataset, which furnished 866 reconstructed pre-crash cases. A conventional two-sample KS significance test was also conducted on the datasets.

Four general metrics were selected for evaluation: the lead-vehicle Delta-v, the no-return time (which indicates temporal criticality), and the minimum accelerations of both the lead and following vehicles. The equivalence test results indicated that only one metric, the lead-vehicle minimum acceleration, satisfied all practical equivalence criteria. In contrast, all significance tests, except for the one concerning the lead-vehicle Delta-v, revealed a statistically significant difference at the 0.05 significance level.

The comparison between the equivalence and significance testing results reveals no clear connection: a statistically non-significant difference does not imply practical equivalence, while a statistically significant difference may still be practically negligible. This finding highlights the limitations of significance testing in the context of scenario validation.

Conclusions

This study introduces a structured, transparent Bayesian ROPE-based framework for validating synthetic pre-crash scenarios against real-world data. Unlike significance testing, the method explicitly evaluates the practical similarity of assessment-relevant metrics and accommodates complex distributional shapes through flexible Bayesian modeling.

Although demonstrated for rear-end crashes, the approach is general and can be applied to other scenario types and validation tasks. It is an important initial step toward a systematic, transparent practical equivalence test between synthetic and real-world pre-crash datasets. The method provides practitioners with a systematic, rigorous tool for validating synthetic scenario generation processes by clearly quantifying similarities and discrepancies between synthetic and real-world datasets. It enables the targeted identification of aspects that need refinement, thereby guiding iterative improvements in scenario realism and representativeness.

3.4 Paper IV

Practical validation of synthetic pre-crash scenarios

Introduction

Building on the ROPE-based practical equivalence testing framework introduced in Paper III, this study addresses a key remaining challenge: how to define assessment-relevant statistics and equivalence criteria in a transparent, interpretable, and practically applicable manner. While the framework in Paper III establishes whether equivalence can be tested, its application in practice requires concrete guidance on what should be compared and how to set equivalence bounds.

This paper proposes and demonstrates practical guidelines for designing statistics and ROPEs for synthetic pre-crash scenario validation. Specifically, it introduces binning-based statistics that balance robustness and interpretability, a weighting mechanism that links distributional deviations to assessment relevance, and a pragmatic procedure for setting ROPEs based on interpretable baseline conditions. Together, these contributions extend equivalence testing beyond a binary decision and provide diagnostic insight into where and why synthetic scenarios deviate from real-world data in ways that matter for SIA.

Methods

The proposed validation framework builds on the ROPE-based equivalence testing methodology introduced in Paper III and introduces a novel binning-based approach that facilitates the design of appropriate statistics and the setting of ROPE thresholds in a transparent and interpretable manner.

For each metric, Bayesian distribution models are fitted separately to the reference and synthetic datasets. Posterior samples drawn from the reference distribution are partitioned into bins with approximately equal posterior mass, and the same bin boundaries are applied to the synthetic distribution.

However, not all bin groups are equally relevant for the intended SIA. The practical relevance of each bin depends on the specific objective of the assessment; consequently, bin weights are introduced to reflect the heterogeneous practical relevance of different regions of a metric's distribution. An empirical procedure is proposed which determines these weights by re-simulating reference scenarios with a virtual representation of the system under assessment and linking bin-level deviations to assessment-relevant outcome measures, such as residual impact speed or injury risk. Based on the resulting bin-level proportions and the bin weights, two complementary statistics are defined: a worst-case statistic that captures the maximum weighted relative deviation across bins, and an aggregate statistic that captures the total weighted absolute deviation. Practical equivalence criteria are then defined by applying ROPEs to the aggregated statistics, with thresholds derived from the tolerances specified for a baseline bin.

Results

The framework was demonstrated by comparing a reference and two synthetic rear-end pre-crash datasets for the SIA of a given AEB system. The results show that binning-based statistics enable clear differentiation between localized worst-case deviations and more diffuse distributional shifts, providing insights unavailable from conventional significance tests or global distance measures.

The ROPE-based evaluation revealed cases in which statistically significant differences were practically negligible, as well as cases where non-equivalence was driven by deviations in assessment-relevant regions of the distribution. The diagnostic analysis further identified which bins contributed most to non-equivalence and how these discrepancies related to system performance outcomes, illustrating the value of the proposed weighting and binning strategy.

Conclusions

This study presents a practical and assessment-oriented framework for validating the representativeness of synthetic pre-crash scenarios. By combining Bayesian modeling, binning-based statistics, empirically grounded bin weights, and ROPE-based equivalence testing, the framework bridges the gap between purely statistical comparisons and practically meaningful validation.

Beyond supporting binary equivalence decisions, the proposed approach provides diagnostic insight into where and why synthetic scenarios deviate from reference data, enabling targeted refinement of scenario generation methods. Although demonstrated for rear-end crashes, the framework is general and can be applied to other scenario types and validation tasks involving the use of synthetic pre-crash scenarios in SIA.

Chapter 4

Discussion

This chapter focuses on how the proposed methods collectively advance the virtual SIAs of DAS and how they inform assessment-oriented thinking about the concept and validation of representativeness. A central theme is that the use and interpretation of pre-crash scenarios in SIA involve not only modeling and simulation choices, but also fundamental assessment design decisions. Choices regarding the construction of reference data, the generation of synthetic scenarios, and the evaluation of representativeness all shape the validity, transparency, and interpretability of assessment outcomes. Accordingly, these choices are treated in this thesis as integral components of the methodological framework rather than as secondary implementation details.

The chapter is structured in four main parts. Section 4.1 discusses the three main methodological contributions of the thesis and their implications for SIA. Section 4.2 then examines several key assessment design considerations for SIA. Finally, Section 4.3 reflects on limitations and Section 4.4 outlines directions for future work.

4.1 Main contributions and their implications

The main contributions of this thesis consist of three tightly connected aspects of virtual SIA: the establishment of reference pre-crash datasets, the generation of synthetic pre-crash scenarios that are statistically consistent with those datasets, and the validation of representativeness through assessment-oriented equivalence testing. Together, these contributions form a coherent methodological framework that explicitly links empirical data, synthetic scenario populations, and validation criteria.

Rather than proposing isolated modeling, generation, or validation techniques, the thesis advances an integrated, assessment-oriented perspective in which representativeness, plausibility, and decision relevance are treated as interdependent design considerations. This proposal addresses the three objectives defined in Section 1.4 and reflects a shift away from descriptive or difference-oriented analyses toward methodologies that explicitly support population-level reasoning and acceptance decisions when using synthetic scenarios for SIA.

4.1.1 Establishing reference pre-crash datasets as an explicit methodological construct

The first major contribution is the explicit treatment of reference pre-crash datasets as a distinct methodological construct for generating and validating synthetic pre-crash scenarios used in SIA. In much of the existing literature [65, 28, 124], empirical pre-crash data are used implicitly—either as direct inputs to simulation or as limited validation examples—without a clear definition of what constitutes an appropriate reference for assessment purposes. Traffic-simulation-based approaches [47, 48] often rely on behavioral models calibrated to sparse or indirect data, while IDC-based approaches [42, 23] typically depend on single data sources that lack exposure information and are biased toward particular severity ranges. As discussed in Papers I and II, these practices complicate representativeness claims and weaken the empirical basis of prospective safety assessment. To address this issue, the present work formalizes the establishment of reference pre-crash datasets as a prerequisite for both scenario generation and validation, with two key components.

The first is model-based parameterization of pre-crash data. As mentioned in Section 2.2.1, parameterization is a simplification process necessitated by the complexity of multivariate pre-crash time-series data and the limited size of available empirical datasets. These constraints preclude the direct use of data-hungry machine learning approaches for trajectory generation and instead motivate a structured representation of pre-crash behavior. While many forms of parameterization are possible, the choice among them can affect the results and should be guided by the purpose of the analysis. Importantly, for the purposes of SIA, parameterization should not focus solely on static states at the moment of impact, such as crash configurations [52, 40], but should also capture the evolution of dynamics during the pre-crash phase. The pre-crash phase is where driver assistance systems operate and exert their influence, while impact states and outcomes remain essential for defining assessment endpoints.

This thesis adopted a behavior- and kinematics-based parameterization of the rear-end pre-crash scenario, in which such a scenario is distilled into a finite set of parameters that describe the relative motion and behavioral relationships between the involved road users. This representation mostly captures the essential dynamics of the scenario while abstracting away largely unnecessary temporal detail. By capturing pre-crash dynamics while preserving the information required to derive impact conditions, the adopted parameterization aligns the construction of reference datasets with both the behavior of DAS during the pre-crash phase and outcome-based scenario evaluation.

Within this parameterization framework, two vehicle behavior models were utilized to characterize pre-crash dynamics. Unlike typical models used in traffic simulations, which are predominantly calibrated on normal driving data [80, 48], these two models were specifically developed from empirical pre-crash data. For the lead vehicle, a kinematics model derived from pre-crash data captures empirically observed diverse speed-change patterns that are not adequately represented by commonly used constant-acceleration or constant-deceleration assumptions [85, 86, 87]. For the following vehicle, a

behavior model tailored to the pre-crash phase combines a conventional car-following model with a pre-crash driver brake-response model, while additionally accounting for abnormal acceleration behavior observed in a non-negligible fraction of rear-end pre-crashes but not represented by the brake-response model (Section 2.3.1). Grounding both models in pre-crash observations rather than nominal driving data means that the parameterization better reflects the behavioral regimes critical to safety assessment.

The other component of this contribution is the data combination and sample weighting methods used to construct reference datasets. Given the inherent limitations and biases of individual empirical pre-crash data sources, this thesis adopted a principled approach to integrating data from multiple sources to better represent the intended evaluation domain. Sample weighting strategies were used to account for differences in sampling mechanisms, coverage, and scenario prevalence, allowing the influence of each data source to be adjusted without assuming that any single dataset constitutes a complete or unbiased reference.

Importantly, these data combination and weighting methods are formulated in a general manner and are not tied to the specific datasets used in this thesis. These methods can be applied broadly in situations where a joint multivariate reference dataset is required but only biased or partially observed datasets are available. This generality facilitates the reuse of the methodology in other SIA studies and supports transparent documentation of how reference datasets are constructed and justified.

Taken together, these two components highlight that establishing reference pre-crash datasets is a non-trivial task that involves balancing behavioral fidelity, data availability, and analytical tractability. By making both the parameterization and data integration steps explicit, this contribution provides a structured basis for constructing reference datasets that are aligned with the objectives of scenario-based safety assessment, rather than treating reference data as fixed or self-evident inputs.

An important implication of this contribution is that it provides a generalizable way of conceptualizing reference pre-crash datasets, rather than a dataset- or model-specific solution. Researchers and practitioners can adopt the proposed approach by first making explicit the purpose of their assessment and then selecting or constructing a parameterization that captures the aspects of pre-crash dynamics most relevant to that purpose. In this sense, the models developed in this thesis should be viewed as instantiations of a broader principle, demonstrating how pre-crash behavior can be represented in a compact, interpretable manner under practical data constraints.

In practice, however, the construction of reference datasets is highly dependent on the available data sources and their limitations. Combining empirical data from different sources often involves handling incomplete variable coverage, heterogeneous sampling mechanisms, and multiple forms of bias, making data integration non-trivial. Even when this process is undertaken, it is both possible and likely that the resulting reference dataset will cover only a subset of the parameters desired for a given assessment. The proposed methodology accommodates this reality by allowing reference datasets to be explicitly defined

and documented at varying levels of completeness, supporting transparent comparisons across studies and clarifying how reference definitions depend on data availability, system design, and assessment objectives.

4.1.2 Generating representative synthetic pre-crash scenarios from reference distributions

The second major contribution concerns the generation of synthetic pre-crash scenarios that are both plausible at the individual level and representative at the population level. While individual scenario realism is necessary, it is insufficient for SIA, since conclusions are drawn from aggregated outcomes across large sets of scenarios. In this context, it is necessary not only to generate realistic-looking trajectories but also to ensure that the generated scenarios collectively reflect the behavioral and kinematic characteristics of the reference population relevant to the intended assessment. The three key components of this contribution are listed below.

The first is the parametric multivariate distribution modeling method developed to represent reference pre-crash datasets under practical data constraints. As mentioned in Section 2.3.1, empirical pre-crash datasets are typically limited in size, with some parameters sparsely represented. Nonetheless, it is crucial to interpret the distribution models and understand the correlations among parameters, as these correlations have physical meaning and directly influence the generated scenario populations. Interpretable dependencies support transparency and confidence in assessment results by linking statistical structure to underlying pre-crash dynamics. Under these conditions, direct empirical modeling approaches, such as kernel density estimation [104], or data-intensive generative methods, such as copula-based models [103], become less suitable, as they tend to require dense data coverage to produce stable estimates and often provide limited interpretive insights [106, 107, 101].

In contrast, the proposed method provides a controlled and interpretable way to represent the reference population while explicitly accounting for data sparsity. By retaining marginal distributions and linear dependencies between parameters, the resulting models capture key population-level structure observed in the data without relying on high-dimensional density estimation. An important aspect of this approach is the use of sub-datasets to reflect heterogeneity in pre-crash conditions. Rather than forcing a single global model to represent all scenarios, the data are partitioned into subsets corresponding to distinct behavior patterns, allowing simpler distribution models to be applied.

This distribution modeling approach can be adapted to other pre-crash scenarios, as well as to analyses that require a reference distribution model when only a relatively limited dataset is available but the underlying distribution is reasonably understood. For example, in the case of intersection pre-crash scenarios, separate sub-datasets could be constructed for different conflict types, such as crossing, turning, or merging. The corresponding distribution models would then focus on variables relevant to the interaction type (such as approach speeds, relative arrival times, gap acceptance measures, or acceleration responses) while preserving interpretable dependencies between them.

Such an approach allows intersection-specific dynamics to be represented in a structured manner, while acknowledging that the resulting reference models remain conditional on the available data and the chosen stratification (this issue is revisited in the discussion of limitations).

The second key component of this contribution is the model-based scenario generation with statistical alignment to the reference distributions. In this thesis, scenario generation is not treated as a purely kinematic sampling problem, but as a behaviorally constrained process grounded in the models used for parameterization. The use of explicit vehicle behavior models ensures the individual-level plausibility of the generated scenarios, as the temporal evolution of vehicle behavior is governed by empirically derived behavioral dynamics.

At the same time, statistical alignment with the constructed reference distributions aims to achieve population-level statistical consistency with the reference data. This distinction highlights an important methodological point: individual-level plausibility does not automatically imply that a set of scenarios is suitable for population-level analysis (e.g., SIA). By separating the roles of behavioral modeling and distributional alignment, this component enables explicit reasoning about how individual scenarios contribute to the properties of the generated dataset as a whole.

From an SIA perspective, this framing supports analyses that depend on aggregated outcomes across large scenario sets, where biases in scenario composition can lead to systematic distortions even if individual cases appear realistic. The proposed combination of model-based generation and statistical alignment, therefore, provides a structured way to balance behavioral plausibility at the scenario level with statistical consistency at the dataset level, aligning the generation process with the population-oriented nature of SIA.

The third component is the constructed representative synthetic rear-end pre-crash dataset and its role within the overall assessment workflow. Rather than serving solely as a static validation benchmark, the synthetic dataset should be treated as an operational reference that supports multiple stages of SIA. For instance, from an application perspective, the dataset can be used directly for the SIAs of specific DAS.

Beyond direct evaluation, the dataset also provides a meaningful benchmark for examining the outputs of alternative scenario generation approaches, including those based on traffic simulation or machine learning. Comparing the statistical properties of externally generated scenarios with those of this dataset makes it possible to identify systematic deviations and better understand the limitations and biases associated with different scenario generation methods. Note that, as demonstrated in Papers III and IV, the synthetic dataset was used not only as a reference in the equivalence testing but also as a reference for adapting existing data sources to the intended assessment context. In particular, it served as the basis for practical equivalence testing and for reweighting an existing dataset (i.e., the GIDAS-PCM dataset) so that its population-level characteristics are statistically consistent with the reference dataset. Thus, a synthetic reference dataset can support both evaluation and data harmonization, reinforcing its role as an integral component of assessment-oriented

analysis rather than a standalone artifact.

Taken together, these three components illustrate the potential benefits of transforming how synthetic pre-crash scenarios are constructed and used for SIA. Rather than treating scenario generation, validation, and dataset preparation as separate or sequential steps, this contribution emphasizes their interdependence through a common reference defined at the population level. By grounding generation and downstream use in explicitly modeled reference distributions, the framework maintains a clear connection between individual scenario dynamics and the statistical properties of the scenario set as a whole.

An important implication of this contribution is that it provides a structured way to handle synthetic scenarios as population-level artifacts, rather than as isolated test cases. Researchers and practitioners can adopt this perspective by first defining reference distributions aligned with their assessment objectives, and then using model-based generation and statistical alignment to construct scenario sets consistent with those references. Doing so supports analyses in which conclusions depend on aggregate outcomes, such as changes in risk measures or performance distributions, rather than on the behavior observed in a small number of selected scenarios.

4.1.3 Reframing representativeness validation as practical equivalence testing

The third major contribution of this thesis is the reframing of representativeness validation as a problem of practical equivalence rather than difference detection. As noted in Section 1.3.2, validation practices in the existing literature predominantly rely on descriptive summaries, visual inspection, and/or difference-oriented statistical tests [57, 40, 58]. While effective at identifying discrepancies, such approaches provide limited guidance on whether the observed differences are practically negligible or consequential for a given SIA.

The first key component of this contribution is the adoption of ROPE-based equivalence testing as a principled validation framework for the pre-crash scenarios used in SIA. Instead of testing whether synthetic and reference data are significantly different, the ROPE-based approach explicitly evaluates whether observed differences fall within a region of practical equivalence defined by the assessment objective. This shift reframes validation away from individual-level plausibility checking toward population-level assessment, aligning it with the decision-oriented nature of SIA, where the central question is whether remaining discrepancies are small enough to permit meaningful population-level inference.

The second key component concerns the use of binning-based statistics with diagnostic capability. By partitioning the reference distribution into quantile-based bins, the proposed statistics enable the localized assessment of equivalence across different regions of the distribution. As mentioned in Section 2.4.5, the binning-based approach provides a structured decomposition of discrepancies, in contrast to conventional global summary measures (such as mean differences, aggregate distance metrics, or the KS statistic), which condense distributional differences into a single scalar value. Although additional information (such as the location of the maximum deviation in the KS statistic)

can be reported, these measures do not provide a systematic regional breakdown that supports diagnostic insight. This approach instead identifies which regions contribute to non-equivalence and distinguish between isolated extreme deviations and broader, systematic differences affecting multiple regions. This level of diagnostic resolution is particularly important in assessment contexts, such as SIA and model validation; understanding the sources and structure of discrepancies complements the interpretation of overall outcomes and supports informed decisions about model adequacy and applicability.

The third key component is the design of assessment-oriented bin weighting schemes and ROPE definitions. Rather than treating all regions of the distribution equally, the proposed framework weights bins according to their practical relevance to the intended assessment, so that the resulting weighted statistics quantify practical differences between datasets in an assessment-relevant manner. Since the same bin-weight function is applied to all metrics, the connection between statistical deviations and their practical implications is defined consistently. This consistency allows common ROPE thresholds to be applied across metrics, providing a unified criterion for determining what constitutes an acceptable magnitude of practical difference in the assessment context. These design choices make explicit the link between validation criteria and assessment objectives, rather than embedding such priorities implicitly in generic statistical tests.

Taken together, these components position representativeness validation as a structured inferential task that is explicitly aligned with the objectives of SIA. An important implication is that the representativeness of synthetic pre-crash scenarios used in SIA is inherently assessment-dependent. Synthetic pre-crash scenarios should therefore be evaluated as fit-for-purpose rather than representative in an absolute or context-free sense. Their adequacy depends on the specific DAS under evaluation, the metrics of interest, and the decision context in which the assessment is conducted.

Recognizing representativeness as assessment-dependent has important consequences for validation practice. If representativeness depends on the intended assessment, then validation criteria must also be defined with respect to that assessment and its practical tolerances. Difference-oriented approaches, such as statistical significance testing or generic goodness-of-fit measures, provide limited guidance for acceptance decisions because their thresholds are defined in terms of statistical discernibility rather than the practical relevance of differences for the assessment. As demonstrated in Papers III and IV, this mismatch can lead to misleading conclusions: non-significant differences do not imply practical equivalence, while statistically significant differences may still be negligible for the intended safety assessment. By reframing validation as a question of practical equivalence, this thesis aligns validation more closely with the logic of safety argumentation [125, 126], where the emphasis is on acceptable tolerances defined in advance rather than on the detectability of differences.

More broadly, this contribution highlights that validation of synthetic pre-crash scenarios used in SIA is not a purely technical exercise but an integral part of methodological design. Metrics selection, binning strategy, bin weighting, and

equivalence criteria implicitly encode assumptions about which discrepancies matter for the assessment at hand. By making these choices explicit through ROPE-based equivalence testing with diagnostic binning, the validation process becomes transparent and aligned with the population-level reasoning used to construct reference datasets and generate synthetic scenarios.

From a practical and regulatory perspective, this integrated view of validation supports more consistent, defensible acceptance decisions across datasets and scenario generation methods. Extending beyond binary pass/fail judgments, the proposed framework provides diagnostic insight into how and why the synthetic dataset deviates from the reference dataset, revealing whether non-equivalence is driven by, for example, rare but safety-critical conditions, systematic shifts in common scenarios, or specific behavioral regimes. This diagnostic capability supports the targeted refinement of scenario generation methods and facilitates constructive dialogue among developers, researchers, and regulators, reinforcing the role of validation as a decision-support mechanism within SIA rather than an end in itself.

Additionally, the proposed validation framework is inherently generic and can be readily extended to contexts beyond the specific case studies in this thesis, such as model validation in traffic flow simulation, comparative evaluation of behavioral models, and assessment of synthetic data used for system testing [53]. Let us use the validation of car-following models in traffic flow simulation as an example. Traditional validation typically focuses on trajectory-level accuracy or aggregate traffic properties, such as the root mean square error of speed, reproduction of fundamental diagrams, and stability characteristics [80, 81, 79]. Such approaches primarily assess goodness-of-fit or detect differences between simulated and observed behavior. However, they do not explicitly evaluate whether the simulated behavior is representative for a specific downstream assessment. The following high-level example illustrates how the proposed equivalence testing framework can instead be applied in an assessment-oriented manner. In this context, the objective is typically not to reproduce individual vehicle trajectories exactly, but to assess whether the population of simulated behavior is practically equivalent to the corresponding empirical population for a given application, such as capacity analysis, stability assessment, or safety-related evaluation. The four validation steps are described below.

Step 1: Construct a reference dataset aligned with the intended assessment purpose. Depending on the application, this dataset may be derived from naturalistic driving studies, video-based trajectory datasets, or other empirical traffic observations, and it may focus on specific traffic states or operating conditions. Importantly, the reference dataset should be explicitly defined and documented in relation to the evaluation objective, as the validity of subsequent conclusions depends on this choice.

Step 2: Select metrics for comparing the simulated and reference data. In the car-following context, these may include kinematic measures (e.g., speeds, accelerations, and headways), conflict or criticality indicators (e.g., time-to-collision-related measures), and other relevant variables.

These metrics do not need to correspond to model parameters; instead, they should capture the aspects of behavior or interaction that matter for the assessment.

Step 3: Conduct equivalence tests for the selected metrics. This step follows the general procedure described in Section 2.4. In the present context, particular attention should be given to defining binning strategies, weighting schemes, and ROPE thresholds in a way that reflects the intended assessment objective. For example, bin weighting schemes may emphasize traffic regimes or interaction conditions that are especially relevant to the assessment. Similarly, ROPE thresholds should represent differences in the selected metrics that are considered practically negligible for the application at hand.

Step 4: Interpret the test results in relation to the assessment objective. Rather than producing a single pass/fail decision, the framework reveals where the car-following model is consistent with empirical behavior and where there are discrepancies. This information can support transparent model evaluation, guide targeted model refinement, and clarify the operational domains in which the model is fit for purpose. In this way, the proposed equivalence testing framework can serve as a practical, assessment-oriented validation tool for, for example, traffic flow simulation models.

In summary, this example illustrates the high-level steps for applying the proposed framework to a domain different from those considered in the application examples of this thesis, highlighting both the generality of the framework and the importance of assessment-oriented validation design.

4.2 Safety impact assessment design considerations

4.2.1 Evolving regulatory and assessment context

The importance of representative reference data is increasingly aligned with ongoing developments in both regulatory frameworks and consumer-oriented vehicle safety assessment programs (commonly referred to as New Car Assessment Programs, NCAPs) for DAS [127, 128]. Historically, road safety evaluation has relied primarily on physical crash testing and injury-focused in-depth crash databases. While such datasets are indispensable for understanding severe outcomes and injury mechanisms, they are not, by themselves, sufficient for prospective virtual SIA. Injury-based sampling typically over-represents higher-severity crashes and under-represents lower-severity conflicts and near-crash situations. Yet these pre-crash interactions are precisely the situations in which many DAS operate. As demonstrated in Papers III and IV, reliance on injury-oriented datasets without adjustment can therefore lead to biased population-level safety conclusions.

Recent regulatory developments reflect a gradual but clear shift toward scenario-based and function-oriented safety evaluation. Within the UNECE, the development of the New Assessment/Test Method (NATM) for DAS explicitly incorporates scenario-based validation alongside traditional physical testing pillars, emphasizing traceability of assumptions and quantitative performance evaluation across defined operational conditions [129]. Related work under UNECE working groups on virtual testing and scenario categorization further highlights the importance of structured scenario databases and transparent validation methodologies in supporting credible safety arguments [130, 131]. Together, these developments signal an increasing expectation that safety claims be supported by systematically defined and validated scenario populations, rather than by isolated test cases alone.

Consumer-assessment programs exhibit similar trends. For example, Euro NCAP's Vision 2030 roadmap outlines a strategic expansion from crash protection toward crash avoidance and safe driving performance, broadening the scope of evaluation beyond injury outcomes alone [132]. Current protocols already include structured near-crash and imminent-collision test scenarios (e.g., for AEB and other DAS). However, when safety impact is assessed using virtual simulation and scenario-based methods, the question extends beyond whether specific test cases are covered. It also concerns whether the underlying scenario populations reflect the broader range of situations in which the DAS operates, including near-crash and low-severity situations that may not result in recorded injuries but are central to system effectiveness.

In this broader context, an "optimal" reference dataset should not be understood as a universal benchmark. Rather, it must be defined relative to a specific assessment objective, operational design domain, and outcome metric. Practically, this implies that reference construction should aim to 1) capture relevant exposure conditions and interaction regimes, 2) include near-crash and low-severity events in addition to injury crashes, 3) document contextual scope and known data limitations transparently, and 4) provide principled mechanisms for combining heterogeneous data sources when no single dataset offers complete coverage. From this perspective, the primary contribution of this thesis is not the proposal of a fixed benchmark dataset. Instead, it lies in providing a structured methodology for constructing, validating, and adapting reference populations in a manner explicitly aligned with assessment objectives under realistic data constraints. By framing representativeness as an assessment-dependent and decision-relevant property, the proposed approach supports transparent, defensible SIA within evolving regulatory and consumer-evaluation environments.

4.2.2 Risks and biases associated with non-representative scenarios

Failure to explicitly define and validate representativeness can have substantial consequences for research conclusions, system development decisions, and regulatory decision-making. When synthetic scenario sets are generated without reference to an empirically grounded target population, SIA results may be

dominated by artifacts of the scenario generation process rather than by the real-world scenario population [35, 21]. However, existing work on SIA typically emphasizes scenario generation methods and plausibility checks, while population-level representativeness is often treated implicitly or left unverified [65, 126, 58].

One important consequence is distorted estimation of system effectiveness. Scenario sets derived primarily from naturalistic driving data tend to over-represent low-severity scenarios, leading to overly optimistic estimates of crash avoidance and injury risk reduction performance [38, 31]. Conversely, scenario sets constructed mainly from in-depth crash databases can over-emphasize severe outcomes, potentially overstating residual risk or obscuring improvements in more frequent low-severity scenarios [23, 50]. In both cases, the absence of an explicit reference distribution makes it difficult to determine whether observed results reflect genuine system performance or merely imbalances in scenario composition. As a result, SIA outcomes may be misinterpreted and cannot be reliably generalized to the broader operational design domain or the target scenario population.

Ignoring representativeness can also hinder iterative system development. Without diagnostic insight into which regions of the scenario space are under- or over-represented, developers may inadvertently optimize systems for idiosyncratic or unbalanced scenario distributions, improving simulated performance without corresponding real-world safety benefit [45, 38]. Explicitly defining and validating representativeness enables more targeted, risk-informed refinement by linking observed performance gaps to specific, assessment-relevant regions of the scenario space. The framework proposed in this thesis supports more accurate, credible SIA and more effective system improvement by making representativeness an explicit and assessable property.

From a broader methodological perspective, the requirement to consider representativeness extends beyond SIA to any context in which a dataset, model, or simulation output is used to support inference or decision-making. Whenever conclusions about a target population, operational domain, or real-world system are drawn from a finite dataset or model-generated sample, implicit assumptions about representativeness are being made. For example, a vehicle behavior model trained on data collected predominantly under low-risk traffic conditions may achieve high apparent performance while failing systematically under high-severity or rare but safety-critical situations. In such cases, deficiencies arise not from the modeling technique itself but from a mismatch between the training data distribution and the intended application domain. If these assumptions remain unexamined, analytical rigor at the modeling level cannot compensate for systematic biases in the underlying data population. Making representativeness explicit is therefore not only a technical refinement but a fundamental condition for credible, responsible application of data-driven methods.

4.2.3 Assessment-oriented metric selection

Beyond the validation framework itself, this thesis highlights that metric selection fundamentally shapes the interpretation of representativeness and the conclusions drawn from a given SIA. Metrics are not neutral descriptors of scenarios; they encode assumptions about which aspects of pre-crash scenarios matter for safety, system performance, and decision-making.

Different classes of metrics emphasize different aspects of the scenario. Kinematic metrics (e.g., speeds, accelerations, and headways) primarily reflect physical feasibility and motion characteristics, while conflict and criticality indicators (e.g., time-to-collision-related measures) emphasize temporal proximity to a crash and the urgency of the driver or system response. Crash-severity-related metrics (e.g., Delta-v and estimated injury risk) quantify outcome consequences rather than interaction dynamics, whereas categorical or event-based metrics (e.g., the occurrence of harsh braking or delayed response) capture specific behavioral patterns. Selecting one class of metrics over another implicitly prioritizes certain safety mechanisms, system functions, and severity outcomes.

Accordingly, any conclusion about representativeness should be understood as conditional on the specific metric set used in the validation. A synthetic scenario set may appear representative with respect to kinematic distributions while deviating substantially in terms of conflict timing or outcome severity, for example. Without explicit reflection on what each metric captures—and what it omits—validation results risk being interpreted as general statements about dataset representativeness, rather than specific statements about representativeness with respect to the specific metric set used in the validation. This distinction has direct implications for how validation results should inform decisions about the use and interpretation of scenario datasets in SIA.

This dependency also has important implications for SIA. Metrics aligned with system triggering logic and intervention mechanisms are more suited to evaluating functional performance, whereas outcome-oriented metrics are more relevant for estimating the safety impact (e.g., injury reduction) of a system/function at the societal level, or for estimating residual risk. Using metrics that are misaligned with the intended assessment question can therefore lead to misleading conclusions.

The proposed framework makes these dependencies visible by structuring validation around explicitly selected, assessment-relevant metrics and evaluating their practical equivalence both individually and by means of overall equivalence criteria. This transparency permits a more nuanced interpretation of validation results, enabling stakeholders to understand which aspects of representativeness are supported by the evidence and which remain uncertain. Such clarity is particularly important in regulatory and comparative contexts, where conclusions about dataset representativeness may depend on the validation metrics used, and where unclear or inconsistent metric selection can lead to conflicting or non-reproducible validation outcomes.

4.2.4 Validation under incomplete reference data

Assessment-oriented validation depends critically on the availability and scope of the reference data. When a reference dataset covering the full parameter space is available, validation design parameters—such as region-specific relevance weights and tolerable deviation bounds—can be derived from re-simulations of the reference scenarios with the DAS under assessment. As a result, the validation procedure explicitly accounts for how different regions of the scenario space influence the system performance and contribute to assessment outcomes.

In contrast, when reference data are incomplete, validation must rely on limited empirical evidence. Under such conditions, it is still possible to move beyond purely difference-oriented testing. However, many validation design parameters—particularly the equivalence criteria—must be specified using external information, such as expert judgment, prior studies, sensitivity analyses, or assessment requirements derived from system design or regulatory context.

This distinction between complete and incomplete reference data has important implications for both methodology and interpretation. Validation results obtained from incomplete reference data should be interpreted as conditional on the limited empirical coverage of the available reference dataset, rather than as definitive evidence of the representativeness of the synthetic dataset for the intended assessment. Explicitly acknowledging this limitation is essential for the transparent, reproducible, and meaningful use of validation results in safety assessment and regulatory dialog.

Viewed in this light, the value of this thesis is not limited to a specific validation technique, but lies in clarifying the relationship between reference data availability, validation design, and the strength of the representativeness statements that can be made about synthetic datasets. The work supports a more nuanced and defensible use of SIA across different stages of system development, evidence maturity, and, ultimately, regulation and system deployment.

4.3 Limitations

Despite the advances presented in this thesis, several limitations remain. The first limitation concerns the availability, coverage, and consistency of the empirical pre-crash data used to establish the reference datasets, and this limitation also defines the empirical scope of the thesis. Although the methodological framework developed in this work is general, its empirical implementation is restricted to rear-end pre-crash scenarios with purely longitudinal dynamics. This restriction is not conceptual but practical. None of the available empirical datasets used in this thesis—SHRP2, CISS (EDR), and GIDAS-PCM—provides sufficiently reliable, temporally resolved, and consistently measured information on lateral maneuvers, multi-agent interactions, lane changes, steering inputs, road geometry, weather conditions, or traffic controls during the pre-crash phase. In contrast, these datasets do contain detailed and comparable longitudinal kinematic information for the two-vehicle rear-end configuration. As a result, all parameterization, reference dataset construction, scenario generation, and representativeness validation conducted in this thesis should be interpreted

strictly within this scenario class and not as a comprehensive representation of all pre-crash behaviors. Even with careful data combination and the use of sample weighting to mitigate known limitations such as severity imbalance and sampling bias, the resulting reference datasets remain approximations of the real-world population. They are shaped by the constraints, assumptions, and heterogeneity of the underlying empirical sources and cannot fully eliminate the uncertainties introduced by dataset-specific sampling mechanisms and reconstruction procedures.

The second limitation concerns the scope of scenario parameterization and behavioral modeling. The parameterization developed in this thesis captures the dominant longitudinal dynamics of rear-end pre-crash events through a piecewise-linear lead-vehicle kinematics model and a following-vehicle behavior model combining a modified IDM with a pre-crash braking response formulation. This abstraction supports tractable modeling and systematic scenario generation, but it necessarily omits additional behavioral dimensions such as lateral maneuvers, evasive steering, and interactions with more than two vehicles. Consequently, the parameterization represents a stylized abstraction of pre-crash dynamics rather than a complete reconstruction of real-world behavior. The suitability of the parameterization, therefore, depends on its alignment with the specific objectives of the safety impact assessment.

The behavioral models used for scenario generation introduce further simplifications. The lead-vehicle model imposes piecewise sections with constant acceleration, capturing many empirical patterns but producing abrupt changes in acceleration that may not fully reflect real-world smoothness. Similarly, although the following-vehicle model reproduces key aspects of observed braking behavior, it cannot replicate the most extreme acceleration behaviors observed in a small subset of empirical cases. These limitations stem from the dual need to avoid model overfitting under sparse data and to preserve interpretability for safety assessment, but they restrict the fidelity achievable in certain edge cases.

The third limitation concerns the distribution modeling approach. The thesis uses parametric models with largely linear dependency structures and subdataset stratification to manage heterogeneity. This provides interpretability and robustness under limited data availability but cannot capture more complex nonlinear or multimodal dependencies that may exist in real-world pre-crash behavior. More expressive semi-parametric or hierarchical approaches could address these limitations, but would require richer datasets and would reduce transparency, which remains essential for safety assessment contexts.

The fourth limitation relates to the metric-level formulation of the representativeness validation framework. Practical equivalence testing is performed independently for each selected metric, and equivalence decisions are aggregated using predefined rules. This preserves transparency and avoids high-dimensional assumptions unsupported by available data. However, it does not explicitly model dependencies among metrics, which may lead to conservative or potentially incomplete validation outcomes.

Finally, several elements of the validation design—including the selection of metrics, the binning scheme, the construction of bin weights, and the

specification of ROPE thresholds—necessarily involve expert judgment. While this thesis provides explicit methodological procedures and a certain level of practical guidance for binning, weighting, and ROPE specification, both metric selection and the design of the bin weight function are addressed primarily at a conceptual level by emphasizing their dependence on assessment objectives and safety relevance. Consequently, some degree of expert judgment remains mandatory, particularly when deciding which metrics to prioritize and how to weight different regions of the distribution for a given safety function and decision context.

4.4 Future work

The limitations identified above point to several promising directions for future research. First, expanding the empirical scope of the framework beyond purely longitudinal rear-end scenarios represents an important opportunity. As richer, more diverse pre-crash datasets become available—particularly those capturing lateral motion, multi-agent interactions, infrastructure features, and environmental context—future work may extend the parameterization and scenario generation framework to a broader set of crash and conflict types. Such extensions will likely require new behavioral models, revised parameterizations, and more flexible distribution modeling techniques capable of describing higher-dimensional and more heterogeneous domains.

Second, future research may enhance the scenario generation process by introducing more expressive modeling frameworks that retain interpretability while capturing nonlinear dependencies and multimodal patterns. Examples include semi-nonparametric and semi-parametric mixture models [133, 134], as well as hierarchical or Bayesian nonparametric formulations that integrate sub-dataset structures more explicitly [135]. As larger and more diverse pre-crash datasets become available, more expressive multivariate models and data-driven model selection strategies may become feasible, potentially reducing the need for manual partitioning while preserving robustness and assessment relevance.

Third, the validation framework can be further strengthened. Potential extensions include multivariate or grouped-metric equivalence testing, hierarchical ROPE definitions, or dependence-aware statistics that reflect joint deviations across multiple assessment-relevant metrics. Such developments would require larger reference datasets and careful attention to methodological tractability and interpretability.

Fourth, future work may aim to increase the standardization and reproducibility of validation design choices. This includes formalizing metric selection criteria, developing systematic procedures for bin-weight construction, establishing recommended ROPE ranges for different types of safety assessments, and integrating sensitivity analyses into validation protocols.

Finally, the broader conceptual shift proposed in this thesis—from difference-oriented to assessment-oriented practical equivalence—suggests important opportunities for integration with emerging regulatory and consumer assessment frameworks. Embedding the methods developed here into standardized safety

assessment workflows and applying them to additional DAS functions and operational design domains may support more transparent, credible, and reproducible virtual prospective safety evaluations.

Chapter 5

Conclusions

This thesis has investigated the generation and validation of representative synthetic pre-crash scenarios for the virtual SIAs of DAS. Rather than treating representativeness as an implicit consequence of simulation or data-driven generation, this thesis formulated it as an explicit methodological challenge that spans both scenario generation and evaluation. As a consequence, two components were generated: the first to determine how representative synthetic scenarios should be generated from empirical evidence, and the second to assess the scenarios in terms of their representativeness of real-world pre-crash conditions for a given SIA purpose. Based on this perspective, the thesis draws five conclusions.

The first conclusion is that representative scenario generation and validation benefit strongly from the use of explicitly constructed reference pre-crash data, rather than reliance on a single empirical source. Existing scenario generation approaches typically draw on either naturalistic driving data or in-depth crash databases, each of which captures only a partial, biased view of the real-world pre-crash population. Naturalistic datasets provide rich coverage of normal and moderately critical interactions but contain very few high-severity crashes, while in-depth crash databases lack exposure information and are dominated by severe outcomes due to case-selection mechanisms. When used in isolation, such sources implicitly define a distorted version of the population of interest, which limits the representativeness of generated scenarios and thus, inevitably, the interpretability of validation results. This thesis demonstrates that explicitly constructing reference pre-crash datasets through model-based parameterization, data combination, and sample weighting is essential to make the target population for assessment explicit. These reference datasets do not merely support scenario generation; they also provide a necessary empirical basis for validation. Without a clearly defined reference population, it is impossible to determine whether observed similarities or differences between synthetic and real-world data are meaningful for SIA.

The second conclusion is that representativeness must be assessed at the population level, not inferred from individual-level similarity or plausibility alone. Much existing work focuses on verifying the realism or plausibility of

individual scenarios by, for example, ensuring that trajectories are physically feasible or consistent with known driver and vehicle behavior. While such checks are necessary, they are insufficient for SIA, because the accuracy and credibility of assessment outcomes depend on whether the overall scenario set reflects the distribution of real-world pre-crash conditions, not merely on the realism of individual scenarios. Synthetic scenarios that appear realistic individually may still be systematically biased as a set (e.g., over-representing benign interactions or safety-critical conditions). This thesis argues that population-level similarity between synthetic and real-world pre-crash data is therefore a necessary condition for representativeness. Individual-level fidelity must be complemented by an explicit assessment of how well the collective statistical properties of synthetic scenarios align with those of the reference population. Treating population-level representativeness as a requirement for scenario evaluation shifts the focus from the realism of isolated example events to distributions, frequencies, and severity patterns—properties that ultimately determine the accuracy and credibility of virtual SIA outcomes.

The third conclusion is that representativeness validation should be framed as a problem of practical equivalence rather than difference detection. Existing validation practices in virtual SIA rely predominantly on descriptive summaries, visual comparisons, or difference-oriented statistical tests. While these approaches are effective for identifying mismatches, they do not address whether observed differences are, for example, negligible for the intended assessment purpose. This thesis argues that validation should assess practical equivalence by determining whether differences between synthetic and reference data fall within predefined, assessment-relevant tolerances. Representativeness validation with a Bayesian equivalence testing perspective shifts from difference detection to acceptance-based reasoning. This approach explicitly accounts for uncertainty in data and models, assessing whether observed discrepancies are small enough to be acceptable (given the intended use).

The fourth conclusion is that representativeness validation benefits from diagnostic insight beyond binary acceptance or rejection. The binning-based statistics and weighting strategies developed in this thesis demonstrate how to make equivalence testing more interpretable and informative. Rather than collapsing validation into a single pass/fail outcome, the framework reveals which regions of a metric's distribution drive any non-equivalence and how these deviations relate to assessment-relevant outcomes. This diagnostic capability supports the targeted refinement of scenario generation methods and promotes a constructive role of validation as part of an iterative development process rather than a one-time gatekeeping step.

The fifth and final conclusion is that the methodological contributions of this thesis are general and extensible. Although the empirical implementation focused on longitudinal rear-end pre-crash scenarios, the underlying principles—explicit reference construction, behavior-based scenario generation coupled with statistical alignment, and assessment-oriented practical equivalence validation—apply broadly to scenario-based virtual SIAs of DAS. As richer datasets become available and assessment needs expand, the proposed framework can be extended to lateral, multi-agent, and context-rich scenarios, and potentially integrated

into emerging regulatory processes for virtual testing and safety argumentation.

Taken together, the contributions of this thesis advance virtual SIA by shifting the focus of scenario generation from implicitly assuming representativeness to explicitly generating and validating representative scenarios across the entire severity range. This thesis treats reference dataset conduction, scenario generation, and validation as interdependent components of a single methodological framework, thus providing a more transparent, defensible basis for using synthetic scenarios in safety assessment. While the focus of this thesis has been on rear-end pre-crash scenarios, the methodological contributions are general and can be applied to other crash types, DAS, and regulatory contexts. Some of the methods should also be applicable well beyond the automotive domain. As empirical data become more readily available and assessment requirements become more explicit, the concepts and methods developed in this thesis can lead to more accurate and credible virtual SIAs. By enabling assessments explicitly grounded in representative real-world conditions and supported by defensible validation arguments, the framework supports more accurate, reliable safety evaluations of DAS. In the longer term, such assessments can contribute to better-informed system design, more robust regulatory decision-making, and ultimately to the deployment of safer and more trustworthy automated vehicles.

References

[1] W. H. Organization, *Global status report on road safety 2018*. World Health Organization, 2019.

[2] SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International Std. J3016_202104, Apr. 2021. doi: 10.4271/J3016_202104., sAE Standard J3016, Issued April 2021.

[3] L. Xiao and F. Gao, "A comprehensive review of the development of adaptive cruise control systems," *Vehicle system dynamics*, vol. 48, no. 10, pp. 1167–1192, 2010. doi: 10.1080/00423110903365910.

[4] E. Coelingh, A. Eidehall, and M. Bengtsson, "Collision warning with full auto brake and pedestrian detection-a practical example of automatic emergency braking," in *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2010, pp. 155–160. doi: 10.1109/itsc.2010.5625077.

[5] W. Chen, W. Wang, K. Wang, Z. Li, H. Li, and S. Liu, "Lane departure warning systems and lane line detection methods based on image processing and semantic segmentation: A review," *Journal of traffic and transportation engineering (English edition)*, vol. 7, no. 6, pp. 748–774, 2020. doi: 10.1016/j.jtte.2020.10.002.

[6] G. Bathla, K. Bhadane, R. K. Singh, R. Kumar, R. Aluvalu, R. Krishnamurthi, A. Kumar, R. Thakur, and S. Basheer, "Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities," *Mobile Information Systems*, vol. 2022, 2022. doi: 10.1155/2022/7632892.

[7] S. Abbasi and A. M. Rahmani, "Artificial intelligence and software modeling approaches in autonomous vehicles for safety management: A systematic review," *Information*, vol. 14, no. 10, p. 555, 2023. doi: 10.3390/info14100555.

[8] A. Giannaros, A. Karras, L. Theodorakopoulos, C. Karras, P. Kranias, N. Schizas, G. Kalogeratos, and D. Tsolis, "Autonomous vehicles: Sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions," *Journal of Cybersecurity and Privacy*, vol. 3, no. 3, pp. 493–543, 2023. doi: 10.3390/jcp3030025.

[9] S. E. Shladover, "Connected and automated vehicle systems: Introduction and overview," *Journal of Intelligent Transportation Systems*, vol. 22, no. 3, pp. 190–200, 2018.

[10] E. Thorn, V. Knisley, and J. Auchter, "Advancing verification and validation for ADAS and ADS," SAE Technical Paper, Tech. Rep., 2025.

[11] I. I. Hellman and M. Lindman, "Estimating the crash reducing effect of advanced driver assistance systems (ADAS) for vulnerable road users," *Traffic Safety Research*, vol. 4, pp. 000 036–000 036, 2023. doi: 10.55329/blzz2682 .

[12] Y. Chen, Y. Xie, C. Wang, L. Yang, N. Zheng, and L. Wu, "Time-dependent effect of advanced driver assistance systems on driver behavior based on connected vehicle data," *Analytic Methods in Accident Research*, vol. 45, p. 100370, 2025. doi: 10.1016/j.amar.2025.100370 .

[13] D. P. Bui, S. Balland, C. Giblin, A. M. Jung, S. Kramer, A. Peng, M. C. P. Aquino, S. Griffin, D. D. French, K. Pollack Porter, S. Crothers, and J. L. Burgess, "Interventions and controls to prevent emergency service vehicle incidents: A mixed methods review," *Accident Analysis & Prevention*, vol. 115, pp. 189–201, 2018. doi: 10.1016/j.aap.2018.01.006 .

[14] A. Kashevnik, R. Shchedrin, C. Kaiser, and A. Stocker, "Driver distraction detection methods: A literature review and framework," *IEEE Access*, vol. 9, pp. 60 063–60 076, 2021. doi: 10.1109/access.2021.3073599 .

[15] J. F. May and C. L. Baldwin, "Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies," *Transportation research part F: traffic psychology and behaviour*, vol. 12, no. 3, pp. 218–224, 2009. doi: 10.1016/j.trf.2008.11.005 .

[16] R. B. Voas and J. C. Fell, "Preventing impaired driving opportunities and problems," *Alcohol Research & Health*, vol. 34, no. 2, p. 225, 2011.

[17] Y. Zhao, D. Ito, and K. Mizuno, "AEB effectiveness evaluation based on car-to-cyclist accident reconstructions using video of drive recorder," *Traffic injury prevention*, vol. 20, no. 1, pp. 100–106, 2019. doi: 10.1080/15389588.2018.1533247 .

[18] H. Tan, F. Zhao, H. Hao, and Z. Liu, "Estimate of safety impact of lane keeping assistant system on fatalities and injuries reduction for China: scenarios through 2030," *Traffic injury prevention*, vol. 21, no. 2, pp. 156–162, 2020. doi: 10.1080/15389588.2020.1711518 .

[19] H. Tan, F. Zhao, H. Hao, Z. Liu, A. A. Amer, and H. Babiker, "Automatic emergency braking (AEB) system impact on fatality and injury reduction in China," *International journal of environmental research and public health*, vol. 17, no. 3, p. 917, 2020. doi: 10.3390/ijerph17030917 .

[20] M. E. Dean and L. E. Riexinger, "Estimating the real-world benefits of lane departure warning and lane keeping assist," SAE Technical Paper, Tech. Rep., 2022. doi: 10.4271/2022-01-0816 .

[21] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation research part A: policy and practice*, vol. 94, pp. 182–193, 2016.

[22] J. Cai, W. Deng, H. Guang, Y. Wang, J. Li, and J. Ding, "A survey on data-driven scenario generation for automated vehicle testing," *Machines*, vol. 10, no. 11, p. 1101, 2022.

[23] F. Zhang, R. Subramanian, C.-L. Chen, and E. Y. Noh, "Crash investigation sampling system: Sample design and weighting," National Highway Traffic Safety Administration, Washington, DC, USA, Tech. Rep. DOT HS 812 706, 2019.

[24] H. Weber, J. Bock, J. Klimke, C. Roesener, J. Hiller, R. Krajewski, A. Zlocki, and L. Eckstein, "A framework for definition of logical scenarios for safety assurance of automated driving," *Traffic injury prevention*, vol. 20, no. sup1, pp. S65–S70, 2019. doi: 10.1080/15389588.2019.1630827 .

[25] X. Zhang, J. Tao, K. Tan, M. Törngren, J. M. G. Sánchez, M. R. Ramli, X. Tao, M. Gyllenhammar, F. Wotawa, N. Mohan, M. Nica, and H. Felbinger, "Finding critical scenarios for automated driving systems: A systematic mapping study," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 991–1026, 2023. doi: 10.1109/TSE.2022.3170122 .

[26] B. Zhu, Y. Sun, J. Zhao, J. Han, P. Zhang, and T. Fan, "A critical scenario search method for intelligent vehicle testing based on the social cognitive optimization algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 7974–7986, 2023.

[27] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6971–6988, 2023.

[28] J. M. Scanlon, K. D. Kusano, T. Daniel, C. Alderson, A. Ogle, and T. Victor, "Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain," *Accident Analysis & Prevention*, vol. 163, p. 106454, 2021.

[29] A. Ivanov, S. Shadrin, and D. Makarova, "The analysis of international standards in the field of safety regulation of highly automated and autonomous vehicles," in *2022 Systems of Signals Generating and Processing in the Field of on Board Communications*. IEEE, 2022, pp. 1–6.

[30] United Nations Economic Commission for Europe (UNECE), "UN regulation No. 157: Automated lane keeping systems (ALKS)," Tech. Rep., 2021. [Online]. Available: https://unece.org/transport/documents/2021/03/un-regulation-no-157-automated-lane-keeping-systems-alks

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

[32] P. Wimmer, O. Op den Camp, H. Weber, H. Chajmowicz, M. Wagner, J. L. Mallada, F. Fahrenkrog, and F. Denk, "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*, 2023, pp. 3–6.

[33] F. Fahrenkrog, A. Das, D. Sander, J. Bärgman, M. Urban, M. Pohl, C. Glasmacher *et al.*, "Prospective Safety Assessment Framework – Instruction," Horizon Europe Project V4SAFETY, Project Deliverable D2.1, 2024, deliverable D2.1. [Online]. Available: https://v4safetyproject.eu/outputs/deliverable_d2_1_framework_instruction_v11.pdf

[34] U. Sander, *Predicting Safety Benefits of Automated Emergency Braking at Intersections-Virtual simulations based on real-world accident data.* Chalmers University of Technology, 2018.

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

[36] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016. doi: 10.1073/pnas.1513271113 .

[37] A. Abdulhafedh, "Road traffic crash data: an overview on sources, problems, and collection methods," *Journal of transportation technologies*, vol. 7, no. 2, pp. 206–219, 2017.

[38] P. Olleja, J. Bärgman, and N. Lubbe, "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*, vol. 83, pp. 139–151, 2022. doi: 10.1016/j.jsr.2022.08.011 .

[39] A. Gambi, T. Huynh, and G. Fraser, "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*, 2019, pp. 257–267.

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

[41] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan, and M. Dozza, "How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 35, pp. 152–169, 2015. doi: 10.1016/j.trf.2015.10.011 .

[42] A. Schubert, H. Liers, and M. Petzold, "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*, vol. 117, 2017. [Online]. Available: https://trid.trb.org/View/1482288

[43] J. Bärgman, S. Jokhio, M. Svärd, J. Östh, F. Denk, C. Klein, J. Iraeus, A. Paliotto, J. Beckman, L. Fonseca Alexandre De Oliveira, K. Adjenughwure, W. Leitgeb, A. Fries, F. Fahrenkrog, R. Davidse, R. de Zwart, X. Yang, A. Das, and M. Hammouda, "Guidelines for the development, quality, and use of models of road-user behaviour and models for in-crash simulations in virtual safety assessment," Horizon Europe Project V4SAFETY, Deliverable D3.1, 2025. [Online]. Available: https://doi.org/10.17196/V4SAFETY/2025/D3.1

[44] C. Glasmacher, J. Beckmann, D. Sander, M. Wisch, J. Bärgman, X. Yang, T. Menzel, T. Amler, M. Urban, and G. Schermers, "Generating baseline scenarios for predictive safety assessment," Horizon Europe Project V4SAFETY, Deliverable D4.2, 2025.

[45] J. Wu, C. Flannagan, U. Sander, and J. Bärgman, "Model-based generation of representative rear-end crash scenarios across the full severity range using pre-crash data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 10, pp. 15 932–15 950, 2025. doi: 10.1109/TITS.2025.3573386 .

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

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

[48] S. Feng, X. Yan, H. Sun, Y. Feng, and H. X. Liu, "Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment," *Nature communications*, vol. 12, no. 1, p. 748, 2021.

[49] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets," Virginia Tech Transportation Institute, Tech. Rep. S2-S31-RW-3, 2016.

[50] J. Wu, *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), 2024.

[51] A. Arun, M. M. Haque, A. Bhaskar, S. Washington, and T. Sayed, "A systematic mapping review of surrogate safety assessment using traffic conflict techniques," *Accident Analysis & Prevention*, vol. 153, p. 106016, 2021.

[52] A. Leledakis, M. Lindman, J. Östh, L. Wågström, J. Davidsson, and L. Jakobsson, "A method for predicting crash configurations using counterfactual simulations and real-world data," *Accident Analysis & Prevention*, vol. 150, p. 105932, 2021.

[53] J. Wu, U. Sander, C. Flannagan, M. Zhao, and J. Bärgman, "Practical validation of synthetic pre-crash scenarios," *Accident Analysis & Prevention*, 2026, under review.

[54] W. L. Greene, J. Concato, and A. R. Feinstein, "Claims of equivalence in medical research: are they supported by the evidence?" *Annals of Internal Medicine*, vol. 132, no. 9, pp. 715–722, 2000.

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

[56] P. Schwaferts and T. Augustin, "Bayesian decisions using regions of practical equivalence (ROPE): Foundations," Department of Statistics, University of Munich, Technical Report 235, 2020. doi: 10.5282/ubm/epub.74222 .

[57] N. M. Razali, Y. B. Wah *et al.*, "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[58] A. Demetriou, H. Alfsvåg, S. Rahrovani, and M. Haghir Chehreghani, "A deep learning framework for generation and analysis of driving scenario trajectories," *SN Computer Science*, vol. 4, no. 3, p. 251, 2023. doi: 10.1007/s42979-023-01714-3 .

[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[60] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967. doi: 10.2307/2283970 .

[61] N. Gibbs, "Errors in the interpretation of 'no statistically significant difference'," pp. 151–153, 2013. doi: 10.1177/0310057x1304100203 .

[62] J. Wu, U. Sander, C. Flannagan, M. Zhao, and J. Bärgman, "Practical equivalence testing and its application in synthetic pre-crash scenario validation," in *2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*. IEEE, 2025, pp. 1–8. doi: 10.1109/IAVVC61942.2025.11219586 .

[63] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence testing for psychological research: A tutorial," *Advances in methods and practices in psychological science*, vol. 1, no. 2, pp. 259–269, 2018.

[64] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern recognition letters*, vol. 42, pp. 11–24, 2014.

[65] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE access*, vol. 8, pp. 87 456–87 477, 2020. doi: 10.1109/access.2020.2993730 .

[66] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[67] D. G. Kidd and A. T. McCartt, "The relevance of crash type and severity when estimating crash risk using the SHRP2 naturalistic driving data," in *International Conference on Driver Distraction and Inattention, 4th, 2015, Sydney, New South Wales, Australia*, 2015.

[68] J. Schoner, R. Sanders, and T. Goddard, "Effects of advanced driver assistance systems on impact velocity and injury severity: An exploration of data from the crash investigation sampling system," *Transportation Research Record*, p. 03611981231189740, 2023. doi: 10.1177/03611981231189740 .

[69] D. I. Swedler, B. Ali, R. Hoffman, J. Leonardo, E. Romano, and T. R. Miller, "Injury and fatality risks for child pedestrians and cyclists on public roads," *Injury epidemiology*, vol. 11, no. 1, pp. 1–11, 2024. doi: 10.1186/s40621-024-00497-2 .

[70] K. Böhm, T. Kubjatko, D. Paula, and H.-G. Schweiger, "New developments on EDR (event data recorder) for automated vehicles," *Open Engineering*, vol. 10, no. 1, pp. 140–146, 2020. doi: 10.1515/eng-2020-0007 .

[71] H. Johannsen, D. Otte, and M. Urban, "Pre-crash analysis of accidents involving turning trucks and bicyclists," in *IRCOBI Council (Hg.): 2015 IRCOBI Conference Proceedings. IRCOBI*, 2015, pp. 09–11.

[72] F. Char and T. Serre, "Analysis of pre-crash characteristics of passenger car to cyclist accidents for the development of advanced drivers assistance

systems," *Accident Analysis & Prevention*, vol. 136, p. 105408, 2020. doi: 10.1016/j.aap.2019.105408 .

[73] E. Rosen, "Autonomous emergency braking for vulnerable road users," in *Proceedings of IRCOBI conference*, 2013, pp. 618–627.

[74] F. Char, T. Serre, S. Compigne, and P. Puente Guillen, "Car-to-cyclist forward collision warning effectiveness evaluation: a parametric analysis on reconstructed real accident cases," *International journal of crashworthiness*, vol. 27, no. 1, pp. 34–43, 2022. doi: 10.1080/13588265.2020.1773740 .

[75] R. Putter, A. Neubohn, A. Leschke, and R. Lachmayer, "Predictive vehicle safety—validation strategy of a perception-based crash severity prediction function," *Applied Sciences*, vol. 13, no. 11, p. 6750, 2023. doi: 10.3390/app13116750 .

[76] O. Kempthorne, *Design and analysis of experiments: Advanced experimental design.* Wiley-Interscience, 2005.

[77] S. M. Ross, *Introduction to probability models.* Academic press, 2014. doi: 10.1016/b978-0-12-407948-9.00012-8 .

[78] S. Wolfram, *Mathematica: a system for doing mathematics by computer.* Addison Wesley Longman Publishing Co., Inc., 1991.

[79] T. Toledo, "Driving behaviour: models and challenges," *Transport Reviews*, vol. 27, no. 1, pp. 65–84, 2007.

[80] M. Treiber and A. Kesting, *Traffic flow dynamics.* Springer, 2013, vol. 1.

[81] S. P. Hoogendoorn and P. H. Bovy, "State-of-the-art of vehicular traffic flow modelling," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 215, no. 4, pp. 283–303, 2001.

[82] M. Svärd, G. Markkula, J. Bärgman, and T. Victor, "Computational modeling of driver pre-crash brake response, with and without off-road glances: Parameterization using real-world crashes and near-crashes," *Accident Analysis & Prevention*, vol. 163, p. 106433, 2021.

[83] D. N. Lee, "A theory of visual control of braking based on information about time-to-collision," *Perception*, vol. 5, no. 4, pp. 437–459, 1976.

[84] J. Wu, C. Flannagan, U. Sander, and J. Bärgman, "Modeling lead-vehicle kinematics for rear-end crash scenario generation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 10 866–10 884, 2024.

[85] J. D. Lee, D. V. McGehee, T. L. Brown, and M. L. Reyes, "Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator," *Human factors*, vol. 44, no. 2, pp. 314–334, 2002.

[86] G. Li, W. Wang, S. E. Li, B. Cheng, and P. Green, "Effectiveness of flashing brake and hazard systems in avoiding rear-end crashes," *Advances in Mechanical Engineering*, vol. 6, p. 792670, 2014. doi: 10.1155/2014/792670 .

[87] X. Wang, M. Zhu, M. Chen, and P. Tremont, "Drivers' rear end collision avoidance behaviors under different levels of situational urgency," *Transportation research part C: emerging technologies*, vol. 71, pp. 419–433, 2016. doi: 10.1016/j.trc.2016.08.014 .

[88] O. Derbel, T. Peter, H. Zebiri, B. Mourllion, and M. Basset, "Modified intelligent driver model for driver safety and traffic stability improvement," *IFAC Proceedings Volumes*, vol. 46, no. 21, pp. 744–749, 2013.

[89] G. Ridder and R. Moffitt, "The econometrics of data combination," *Handbook of econometrics*, vol. 6, pp. 5469–5547, 2007. doi: 10.1016/s1573-4412(07)06075-8 .

[90] M. Sandelowski, "Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies," *Research in nursing & health*, vol. 23, no. 3, pp. 246–255, 2000.

[91] C. A. Gotway and L. J. Young, "Combining incompatible spatial data," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 632–648, 2002.

[92] D. Holt and T. F. Smith, "Post stratification," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 142, no. 1, pp. 33–46, 1979.

[93] D. Sloane and S. P. Morgan, "An introduction to categorical data analysis," *Annual Review of Sociology*, vol. 22, no. 1, pp. 351–375, 1996.

[94] R. Malarvizhi and A. S. Thanamani, "K-nearest neighbor in missing data imputation," *Int. J. Eng. Res. Dev*, vol. 5, no. 1, pp. 5–7, 2012.

[95] S. G. Shelby *et al.*, "Delta-v as a measure of traffic conflict severity," in *3rd International Conference on Road Safety and Simulation September*, 2011, pp. 14–16.

[96] T. W. Anderson and I. Olkin, "Maximum-likelihood estimation of the parameters of a multivariate normal distribution," *Linear algebra and its applications*, vol. 70, pp. 147–171, 1985. doi: 10.1016/0024-3795(85)90049-7 .

[97] A. G. Stephenson, "High-dimensional parametric modelling of multivariate extreme events," *Australian & New Zealand Journal of Statistics*, vol. 51, no. 1, pp. 77–88, 2009. doi: 10.1111/j.1467-842x.2008.00528.x .

[98] A. Kottas, P. Müller, and F. Quintana, "Nonparametric Bayesian modeling for multivariate ordinal data," *Journal of Computational*

and Graphical Statistics, vol. 14, no. 3, pp. 610–625, 2005. doi: 10.1198/106186005x63185 .

[99] N. Yang, Y. Huang, D. Hou, S. Liu, D. Ye, B. Dong, and Y. Fan, "Adaptive nonparametric kernel density estimation approach for joint probability density function modeling of multiple wind farms," *Energies*, vol. 12, no. 7, p. 1356, 2019. doi: 10.3390/en12071356 .

[100] A. Sancetta and S. Satchell, "The bernstein copula and its applications to modeling and approximations of multivariate distributions," *Econometric theory*, vol. 20, no. 3, pp. 535–562, 2004. doi: 10.1017/s026646660420305x .

[101] R. B. Nelsen, *An introduction to copulas*. Springer, 2006. doi: 10.1007/0-387-28678-0 .

[102] I. Kojadinovic and J. Yan, "Modeling multivariate distributions with continuous margins using the copula r package," *Journal of Statistical Software*, vol. 34, pp. 1–20, 2010. doi: 10.18637/jss.v034.i09 .

[103] H. Kazianka and J. Pilz, "Copula-based geostatistical modeling of continuous and discrete data including covariates," *Stochastic environmental research and risk assessment*, vol. 24, pp. 661–673, 2010. doi: 10.1007/s00477-009-0353-8 .

[104] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," *Biostatistics & Epidemiology*, vol. 1, no. 1, pp. 161–187, 2017. doi: 10.1080/24709360.2017.1396742 .

[105] J. Orava, "K-nearest neighbour kernel density estimation, the choice of optimal k," *Tatra Mountains Mathematical Publications*, vol. 50, no. 1, pp. 39–50, 2011. doi: 10.2478/v10127-011-0035-z .

[106] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[107] A. Z. Zambom and D. Ronaldo, "A review of kernel density estimation with applications to econometrics," *International Econometric Review*, vol. 5, no. 1, pp. 20–42, 2013.

[108] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 3, p. e1460, 2019.

[109] H. A. Panofsky, G. W. Brier, and W. H. Best, *Some application of statistics to meteorology*. Mineral Industries Extension Services, College of Mineral Industries, Pennsylvania State University, 1958.

[110] G. J. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2000.

[111] B.-J. Park and D. Lord, "Application of finite mixture models for vehicle crash data analysis," *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 683–691, 2009.

[112] D. M. Cerwick, K. Gkritza, M. S. Shaheed, and Z. Hans, "A comparison of the mixed logit and latent class methods for crash severity analysis," *Analytic Methods in Accident Research*, vol. 3, pp. 11–27, 2014.

[113] A. A. B. da Costa, P. Irvine, X. Zhang, S. Khastgir, and P. Jennings, "Ontology-based scenario generation for automated driving systems verification and validation using rules of the road," *IEEE Transactions on Intelligent Vehicles*, pp. 1–11, 2024. doi: 10.1109/TIV.2024.3377534 .

[114] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, 2007. doi: 10.1007/978-0-387-72806-3 .

[115] S. Kolenikov, "Calibrating survey data using iterative proportional fitting (raking)," *The Stata Journal*, vol. 14, no. 1, pp. 22–59, 2014. doi: 10.1177/1536867x1401400104 .

[116] A.-A. Choupani and A. R. Mamdoohi, "Population synthesis using iterative proportional fitting (IPF): A review and future research," *Transportation Research Procedia*, vol. 17, pp. 223–233, 2016.

[117] J. Wu, "Quadris project pre-crash and near-crash dataset," 2025, accessed: May 2025. [Online]. Available: https://github.com/JianWu09/QUADRIS-project-Pre-crash-near-crash-database

[118] D. Schuirmann, "On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval," in *Biometrics*, vol. 37, no. 3. INTERNATIONAL BIOMETRIC SOC 808 17TH ST NW SUITE 200, WASHINGTON, DC 20006-3910, 1981, pp. 617–617.

[119] J. K. Kruschke, "Rejecting or accepting parameter values in Bayesian estimation," *Advances in methods and practices in psychological science*, vol. 1, no. 2, pp. 270–280, 2018.

[120] H. Abdi *et al.*, "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia of measurement and statistics*, vol. 3, no. 01, p. 2007, 2007.

[121] S. Watanabe and M. Opper, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of machine learning research*, vol. 11, no. 12, 2010.

[122] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017. doi: 10.1007/s11222-016-9696-4 .

[123] R. Chen, R. Sherony, and H. C. Gabler, "Comparison of time to collision and enhanced time to collision at brake application during normal driving," in *SAE 2016 World Congress and Exhibition*. SAE Technical Paper, 2016.

[124] B. Schütt, J. Ransiek, T. Braun, and E. Sax, "1001 ways of scenario generation for testing of self-driving cars: A survey," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[125] S. Riedmaier, D. Schneider, D. Watzenig, F. Diermeyer, and B. Schick, "Model validation and scenario selection for virtual-based homologation of automated vehicles," *Applied Sciences*, vol. 11, no. 1, p. 35, 2020.

[126] R. Donà and B. Ciuffo, "Virtual testing of automated driving systems. a survey on validation methods," *IEEE Access*, vol. 10, pp. 24 349–24 367, 2022.

[127] C. A. Hobbs and P. J. McDonough, "Development of the European new car assessment programme (Euro NCAP)," *Regulation*, vol. 44, no. 3, pp. 2439–2453, 1998.

[128] L. L. Hershman, "The US new car assessment program (NCAP): Past, present and future," in *International Technical Conference on Enhanced Safety of Vehicles*. National Highway Traffic Safety Administration, 2001.

[129] Informal Working Group on VMAD (GRVA), "New assessment/test method for automated driving (natm) guidelines for validating automated driving system (ads)," United Nations Economic Commission for Europe (UNECE), Geneva, Switzerland, Working Document ECE/TRANS/WP.29/2023/44/Rev.1, 2023, accessed: 2026-02-23. [Online]. Available: https://unece.org/sites/default/files/2023-06/ECE-TRANS-WP.29-2023-44-r.1e%20.pdf

[130] United Nations Economic Commission for Europe (UNECE), "Virtual testing in un r 152 - introduction and credibility assessment," Working Party on Automated/Autonomous and Connected Vehicles (GRVA), Tech. Rep. GRVA-18-55, jan 2024, 18th GRVA session, 22-26 January 2024. [Online]. Available: https://unece.org/sites/default/files/2024-01/GRVA-18-55e.pdf

[131] E. de Gelder, O. O. den Camp, and N. de Boer, "Scenario categories for the assessment of automated vehicles," *CETRAN, Singapore, Version*, vol. 1, p. 1, 2020.

[132] Euro NCAP, "Euro NCAP Vision 2030: A Safer Future for Mobility," European New Car Assessment Programme, Brussels, Belgium, Tech. Rep., 11 2022, accessed: 2026-02-23. [Online]. Available: https://cdn.euroncap.com/media/74468/euro-ncap-roadmap-vision-2030.pdf

[133] S. Xiang, W. Yao, and G. Yang, "An overview of semiparametric extensions of finite mixture models," *Statistical Science*, vol. 34, no. 3, pp. 391–404, 2019.

[134] G. Sfeir, M. Abou-Zeid, F. Rodrigues, F. C. Pereira, and I. Kaysi, "Latent class choice model with a flexible class membership component: A mixture model approach," *Journal of choice modelling*, vol. 41, p. 100320, 2021.

[135] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, pp. 1566–1581, 2006.