BUILDING SCIENTIFIC CONSENSUS ON THE CRASH SAFETY PERFORMANCE OF AUTOMATED DRIVING SYSTEMS

John M. Scanlon
Kristofer D. Kusano
Timothy L. McMurry
Tilia Gode
Trent Victor
Waymo, LLC
United States

Paper Number ESV26-252

ABSTRACT

The deployment of Automated Driving Systems (ADS) offers a potential solution to reduce injury burden on US roadways. Rigorous retrospective safety assessments quantifying their efficacy is one key component in building scientific consensus around ADS safety. Unlike prior technologies, ADS evaluation is uniquely challenging due to its control over the entire driving task across diverse scenarios, use-case specific driving exposure profiles, and the unprecedented volume of real-time data generated.This paper focuses on improvements in data sources, implementation, and reporting to enable more credible performance assessments. Current public ADS data, like NHTSA SGO reports, lack the necessary exposure metrics (like Vehicle Miles Traveled, or VMT) for rate computation. We recommend ADS developers release detailed VMT and supplemental crash data to enable rate computation and rigorous analysis. Greater data granularity includes exposure confounders, comprehensive crash outcome units, and the inclusion of relevant performance lenses. Robust research implementation is critical. Researchers must align ADS and baseline data to account for temporal and geographic differences and are encouraged to follow best practices like the RAVE Checklist. To mitigate the potential for bias, we advocate for transparency through proactively published benchmarks and the use of justified reporting windows. Finally, independent oversight—including peer review, data access for independent researchers, and cooperative study partnerships—is essential for ensuring unbiased evaluation. The substantial influx of ADS data necessitates that the research community builds the empirical evidence for achieving consensus. Rigorous, continuous safety impact evaluations are a key catalyst; if ADS effectiveness is widely recognized, faster adoption could lead to expedited harm reduction on public roadways.

Keywords: Automated driving systems, safety impact, safety performance assessment, retrospective analysis, historical crashes

INTRODUCTION

Baseline Safety Burden Accumulation

There is a steady accumulation of injury burden in our global vehicle fleet. In the United States, the current fatality burden (using the latest data from 2024) on US roadways sits at a total of 39,345 fatalities over 3.3 trillion vehicle miles traveled (VMT) [1, 2], which results in 1.20 fatalities are occurring for every 100 million VMT. This puts the US national fatality accumulation rate at approximately 108 fatalities per day in 2024. Nearly 50 years ago, in 1976, the United States had 44,525 fatalities with a societal fatality rate of 3.35 fatalities for every 100 million VMT driven. The 1976 per-mile fatality rate is nearly three times what it was in 2024.

Historically Successful Safety Interventions

The cost of not implementing new safety technologies is the continual accumulation of that safety burden. Vehicle safety interventions have played a key role in lowering US fatality rates. *Successful* safety technologies follow a

common pattern (shown ***Figure 1.***).

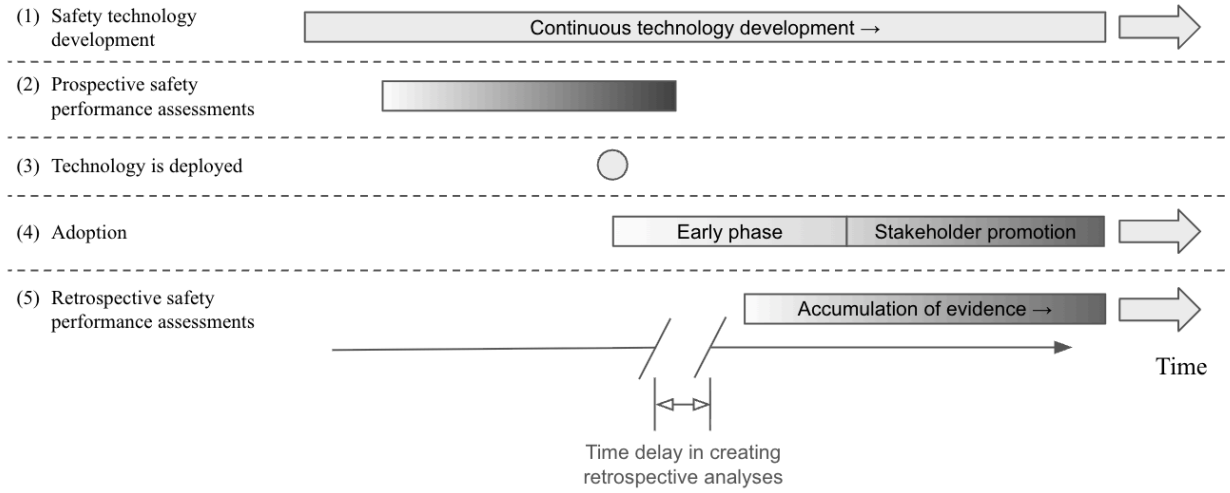

Figure 1. Depicting the role of retrospective safety performance assessments within the scope of technology development and adoption.

The development of safety technologies is iterative, continuous, and can take many years. Initially, the research, design, and validation phases can take many years before the technology is deemed ready for deployment in the public domain. For example, airbag concept designs were proposed as early as the 1950s [3]. The first public deployment of airbags (by General Motors) did not occur until the early 1970s [4]. Once deployed, the pace of fundamental, step-change improvements to the core technology tends to slow down, reflecting the maturation of the underlying engineering principles [5, 6, 7]. Looking again at airbags, early designs, for instance, sometimes posed risks to certain occupants. Design improvements were subsequently implemented to mitigate risks associated with out-of-position occupants, improper seating posture, or system malfunctions like unexpected ruptures [5]. These iterative, data-driven modifications have collectively resulted in significantly increasing safety benefits over time [8]. Viano [5] estimated that through January 1, 2009, there were 320 deployment deaths attributed to airbags [9]. NHTSA estimated that frontal airbags have saved a total of 50,457 US lives between the years 1975 and 2017 [10].

During early stages of development, projections are made about a technology's potential safety impact. This helps motivate where the technology may be effective, where the technology may be limited, and how it should be designed and tested. This has historically been accomplished through a combination of target population analyses and testing. ISO/TS 21934-1 [11] and ISO/TS 21934-2 [12] each provide methodology overview and guidelines for "prospective" safety impact assessment of pre-crash technology, that is the "predictive assessment of the future performance of given technologies before their deployment into a vehicle population."

The technology is then deployed in some limited capacity, where many factors influence how quickly adoption occurs. Among those factors driving adoption are scientific consensus around the safety performance of the technology. At initial deployment, there is no accumulation of historical performance, so performance estimates are made based on testing performance (i.e., reliance solely on prospective analysis). The gold standard for assessing safety performance is ultimately what can be determined from its actual, on-road safety performance. There is a natural pause, however, from the initial deployment of the technology and this retrospective safety performance assessment as mileage accumulates and statistical analysis can be performed.

Scientific consensus is ultimately dictated by:

1. *Making Data available*. Public crash databases have been commonly used, and those have a set cadence for release. For some technologies, the developers themselves are the only entities that can provide the necessary information for tracing performance, so evaluation is dependent on their making this information available to researchers.

2. *Conducting credible research*. Each technology has nuances in how it can be studied, which often require custom methodological practices. There are also limitations in the data itself, which often necessitate creative solutions. As researchers and stakeholders interrogate the data, new research questions and methodology needs arise.
3. *Accumulation of peer-reviewed and/or independent scientific reporting*. That research must accumulate until the point the community reaches scientific consensus, which, ultimately, drives promotion of the technology and *accelerates adoption*.

Retrospective Safety Assessments of Automated Driving Systems

Automated Driving Systems (ADS), specifically SAE Level 4 technology that maintains the entire dynamic driving task (DDT) and DDT fallback without any expectation that a user will need to intervene" [13], are being proposed as a crucial safety technology solution to mitigate the persistent injury burden on public roadways. The current status of emerging ADS deployments is diverse. The core objective of this paper is not to provide a real-time assessment of active deployments. Rather, it is to strategically plan for a future, mature state of deployment where comprehensive safety measurement is feasible, provided that necessary preparatory steps are taken in data availability, methodological rigor, and transparent reporting. To offer a general overview of the industry's status at the time of this publication (December 2025), the following points warrant consideration:

A. Many companies have planned ADS deployments in the US,
B. Very few companies have deployed with enough accumulated mileage for statistical testing, and
C. Even fewer (one) companies have provided sufficient public data for rigorous statistical analysis of safety performance.

ADS technology is unlike any other prior safety intervention when it comes to how we measure their safety performance. Like ADAS, these technologies navigate (steer, brake, accelerate) the vehicle. Unlike ADAS, the entire DDT is handled by the technology without any "human fallback-ready user" within the scope of how it is deployed [13]. This creates an evaluation space that is not limited to specific scenario evaluation. For example, frontal airbags are designed to function in frontal collisions. When evaluating airbags, historical performance in frontal crashes were used, because that is where the technology was designed to reduce risk of injury [5, 14]. Similarly, frontal crash prevention technologies (frontal crash warning and AEB) are designed for preventing crashes with other road users in front of the vehicle (e.g., vehicles and pedestrians). When evaluating these technologies, reductions in frontal crashes are examined [15, 16]. Because ADS handles, by definition, the entire driving task, a more holistic evaluation of crash rates needs to be employed. This provides an opportunity to (1) directly quantify the ADS overall risk across all driving versus some baseline and (2) examine the performance of ADS within subsets of that driving (e.g., specific roads, collision partners, weather conditions, etc.).

A further, distinctive consideration surrounding ADS technology resides in the expectation of behavioral driving responsibility, a feature markedly absent in preceding ADAS technologies. With ADAS, the driver retains responsibility for the DDT, and the assessment of ADAS functionality is narrowly focused on its capacity to *aid* the human operator in preventing and mitigating collisions. ADAS operates by autonomously detecting an imminent collision threat and initiating a corrective response, irrespective of the driver's antecedent behavior. Consequently, the historical evaluation of ADAS has been confined to a more holistic assessment of its safety impact, referred to later as performance relative to the "status quo" of driving. These prior research questions sought to determine: to what extent are human drivers made safer by the integration of ADAS into their vehicles? Conversely, ADS (SAE Level 4) maintains complete authority over the entirety of the DDT. Consistent with the expectation placed upon human drivers, there is a commensurate expectation that the ADS will not only diminish instances of crash initiation but also proactively exhibit conflict avoidance behaviors to preclude potential responder crashes [17, 18]. As explored in the "Implementation" section, researchers must grapple with the question of how to expand their analytical focus beyond merely improving the "status quo" to rigorously consider the *societal expectations* placed upon this technology (discussed later in "Research Question Forcus").

Data availability with ADS technology is also unprecedented for safety impact analysis, and, alone, requires a paradigm shift in how we approach the opportunity for retrospective safety performance assessments. Safety impact measurement relies on both measures of exposure and crash outcomes. Historically, crash investigations have been post-hoc, depending on disparate physical evidence, such as scene reconstruction or vehicle damage analysis, to

approximate the precise sequence of events preceding and following a contact event. Correspondingly, estimates of driving exposure have frequently been characterized by imprecision. In contrast, ADS-equipped vehicles utilize an extensive array of sensors that capture the exact movements of the vehicle as it operates on public roadways. In the event of a collision, these sensors can provide more insights into the precise nature of the impact by providing detailed tracking of actors potentially involved in the collision sequence than traditional vehicles. As will be discussed in subsequent sections, new methodologies and reporting practices are needed in order to utilize this wealth of information in ADS safety assessments. A critical component of this data enablement involves translating the proprietary system information into a format that is directly comparable to existing benchmark data, thereby facilitating methodologically sound, "apples-to-apples" risk comparisons. Because of the wide array of insights possible with ADS data due to its information richness, the limiting factor in safety performance assessments are more often going to be in what can be gleaned from the available benchmark data. As will be discussed, concurrent work on processing benchmark data sources is needed to unlock new analysis opportunities.

Finally, the deployment modality of ADS technology is fundamentally distinct from that of prior vehicle safety technologies. Historically, safety interventions-such as airbags, seatbelts, backup cameras, and ADAS-were developed for and integrated into the general consumer vehicle fleet. In sharp contrast, Level 4 ADS deployments are primarily confined to purpose-built, commercial vehicle fleets. Examples of this include long-haul trucking (Class-8 vehicles) and ride-hailing and goods delivery services (Class 1 and 2 vehicles). In both scenarios, the geographic scope and the operational intent (the movement of freight and people) are meticulously tailored to the specific use case. This specificity raises a critical methodological challenge: how can one generate a methodologically rigorous, "apples-to-apples" comparison against a baseline human driver fleet when the Automated Driving System operates within such a highly constrained and specialized Operational Design Domain (ODD)?

STUDY SCOPE

This paper critically evaluates the current status of available data, methodology, and reporting for conducting retrospective safety performance assessments of ADS operating on public roadways. The study identifies needed improvements and proposes necessary corrective actions to facilitate credible, rigorous, and timely analyses. The methodology systematically dissects each phase of the retrospective safety research process-from initial data sourcing and methodological implementation to the eventual transparent communication of findings. By deconstructing this process, the paper intends to furnish a comprehensive roadmap for enabling robust safety evaluations of ADS technology.

STUDY EXECUTION

This paper breaks the retrospective safety impact research process down into three main components:

(1) *Data Sources*: The starting point for any retrospective analysis. Studies can only be done if benchmark and ADS data is made available. That data must be provided with enough information that it can be useful.

(2) *Implementation*: The systematic approach and set of procedures used to conduct the study or achieve a specific objective. This includes detailing the research design, analytical techniques, and validation processes.

(3) *Reporting*: The communication of plans, findings, and conclusions derived from the analysis.

Data Sources

Retrospective safety performance assessments are fundamentally impossible without the requisite data on both the human-driven benchmark and ADS sides. All rigorous retrospective safety performance analyses hinge upon the judicious combination of crash outcomes and driving exposure. When calculating crash rates, the established benchmark necessitates the use of crash outcomes (events meeting or exceeding a prespecified severity threshold) correlated with driving exposure (typically quantified as VMT). Furthermore, researchers require the necessary data granularity to meticulously investigate diverse research questions and to appropriately account for potentially confounding variables [19].

Currently within the US, NHTSA has issued a standing general order (SGO) regarding ADS crash reporting [20]. This order mandates that ADS operators submit crash reports for incidents that satisfy certain predefined severity thresholds. These reports include a standardized set of data fields broadly capturing the details of the collision event. NHTSA subsequently publishes the data for analysis, such as for use by the research community. Crucially, the SGO does *not* require the reporting of VMT or any other exposure-related metrics (e.g., hours of operation).

The selection of benchmark data is an equally pivotal consideration in conducting retrospective safety performance assessments. Like ADS data, the benchmark requires both robust exposure and crash data. The most common benchmark has historically been a human driver baseline, although comparisons against prior ADS versions or between different ADS technologies are also methodologically possible. While a comprehensive examination of all possible data sources for generating a human baseline, including their specific data features and inherent limitations, is beyond the scope of this paper, successful examples have been demonstrated using: (a) telematics data [21], (b) police reported crash data [22, 23, 24], (c) naturalistic driving studies [25], (d) dash cameras [26], (e) insurance data [27, 28], and (f) professional driving fleets [29].

Making Data Available

ADS crash rates cannot currently be directly computed from the data required to be submitted under the SGO. Crashes per VMT (or kilometers traveled) is the most widely reported crash rate form. Prior research conducted by the authors has cataloged nearly 20 ADS benchmarking studies over the past decade, all of which utilize VMT as the exposure metric, with no viable alternative proposed in the literature to date [23].

The most direct mechanism for bridging the data availability deficit is for developers to proactively release their VMT data in a format that permits researchers to seamlessly integrate this information with the publicly accessible SGO crash data. Furthermore, the existing SGO crash data fields are often insufficient for comprehensive safety impact analysis, necessitating greater data granularity (which will be explored in the subsequent section). Strategic data-sharing pathways are critically needed for ADS developers to provide crash and exposure data to enable research. Naturally, such sharing must be executed with rigorous consideration for privacy concerns and the unintentional disclosure of sensitive information. Therefore, careful methodological planning is imperative to enable this data release. One crucial consideration is that while SGO data is provided at the event level, more sensitive data, such as highly granular crash or VMT information, may be better suited for aggregation. For instance, the release of high-frequency breadcrumb GPS data from individual trips presents numerous potential security and privacy challenges. One pathway is for ADS developers to voluntarily share their crash and VMT data for researchers. To the knowledge of the authors, only Waymo has, to date, done this with their deployed ADS fleet, where they provided both VMT aggregated by level 13 S2 cell and crash data for download beginning in late 2024 [30].

Some retrospective assessments for earlier safety technologies have employed the concept of "induced exposure," a methodological approach warranting brief discussion here. Induced exposure utilizes specific, carefully chosen crash subsets as exposure surrogates, operating under the assumption that these subsets are unaffected by the safety intervention being studied and increase proportionally across both the treatment and control groups [31, 32, 33, 34]. For instance, Kullgren et al. [33], in their investigation of Pedestrian and Cyclist Automatic Emergency Braking (AEB) effectiveness, utilized rear-end struck counts as an exposure proxy, hypothesizing that the AEB system would not influence this specific crash type. Although, more recent, larger scale analysis by the Partnership for Analytics Research in Traffic Safety (PARTS) has shown (using VMT as an exposure metric) that rear-end struck rates are potentially slightly lower for AEB-equipped vehicles, which calls into question the applicability of induced exposure using rear-end struck for these systems [35]. The application of induced exposure for assessing ADS safety performance has yet to be demonstrated, and it is likely ill-suited for this purpose. This limitation arises because an ADS assumes control of the entire dynamic driving task, implying that it would be challenging to identify a crash scenario that is both (a) identifiable in the crash data (that has limited data fields) and (b) has direct, identical proportionality with crash risk (for both ADS and baseline). In actuarial science, analogous criteria exist for

establishing an exposure base [36]. These guidelines stipulate that the metric: (1) must be "directly proportional to expected loss," (2) should be "practical," meaning it is "objective and relatively easy and inexpensive to obtain and verify," and (3) a less pertinent best practice suggests the measure should maintain "historical precedence" to avoid introducing significant analytical complications (e.g., concerning the management of dependent rating schemes predicated on prior methodology) [36]. To illustrate, consider the aforementioned AEB example that relied on rear-end struck rates as an exposure proxy. Kusano et al. [37], analyzing approximately 57 million VMT of ADS operation, identified notable differences in the mechanisms of rear-end struck crashes when compared to a human-driven fleet.

Providing Data Granularity

Simply providing crash and mileage data is insufficient for enabling rigorous, credible scientific inquiry. Data granularity, the specific features embedded within the dataset, is the factor that ultimately enables robust safety research capable of accurately identifying the technology's strengths and limitations. Data granularity can be systematically categorized into the following informational components:

1. Exposure Units: A foundational metric quantifying the total driving activity performed (e.g., VMT).
2. Exposure Confounders: Characteristics that potentially or demonstrably influence the observed crash rates.
3. Crash Outcome Units: A measure specifying whether a given crash event satisfies a defined severity or reporting criterion.
4. Performance Lenses: Features that furnish insight by segmenting and describing specific subsets of the overall driving crash risk.

All driving is not equal in terms of crash risk. For instance, a study by Chen et al. [38] investigated the initial deployment of an ADS in San Francisco, contrasting it with the subsequent expansion of its ODD to encompass the entire county as a commercial rideshare service. As anticipated, police-reported crash rates exhibited a clear dependency on areas within the city. During the early testing phase, the selected operational areas were characterized by a baseline crash rate lower than the county average. Following the commercial launch, however, the estimated baseline driving exposure was over 30% higher than the county rate due to shifting fleet VMT to more densely populated, and higher crash risk areas. Consequently, a broad array of confounding variables must be rigorously considered when analyzing ADS crash rates. These features include, but are not limited to, road type, vehicle type, and spatial and temporal dimensions (i.e., where and when) of the driving [e.g., 23, 39, 40, 41]. Given that the NHTSA SGO does not mandate the reporting of exposure data, establishing a clear precedent for which exposure confounders must be provided to support robust analysis is imperative. It is notable that many of the established features influencing crash rates can be incorporated simply by knowing the precise spatiotemporal coordinates of the accumulated exposure.

Crash outcome units span a broad spectrum of severities, providing the critical signal that directly reflects the potential for harm in an incident. Without comprehensive information regarding the crash outcome, researchers are unable to appropriately align baseline data for comparison. Furthermore, restricting the available crash outcomes to a narrow set of values significantly impedes the ability of researchers to rigorously examine safety performance across the full range of events capable of inflicting harm. Generally speaking, there is a need to establish outcomes that are explainable, unbiased, and safety-relevant.

Performance lenses are integral to enabling research that meticulously examines the various components of the driving task. These features are primarily expected to be tailored to the specific mechanisms of crash occurrence. In the analysis of baseline risk, understanding the composition of crashes that result in the most significant harm provides a clearly defined target population for ADS to demonstrate their effectiveness in mitigation (or lack thereof) [23, 37]. Conversely, crash mechanisms that are overrepresented within ADS data may serve as crucial

indicators of potential system limitations, thereby highlighting key opportunities for subsequent technological improvement. An example of a performance lens is examining safety performances by individual crash types.

All of these data features require compatibility with available baseline datasets. The information must be provided in both, and the features themselves should be aligned. If the information is missing or unclear, there is a risk that the safety performance assessment will misrepresent the actual system effectiveness, or the analysis may just be foregone altogether. Given the early stages of ADS, we are in a key position to examine the available ADS data and make decisions about what information needs to be provided. Researchers, evaluators, and developers need to collaborate to identify what baseline data is available and could be used, and then, from there, identify what data features should be included in ADS release.

Prior research demonstrating the potential to use publicly available data is what motivated the design of the previously discussed voluntary ADS download [30]. Specifically, multiple peer reviewed studies using public crash and VMT data were used to design the downloadable content. That downloadable information has also been updated over time using the latest available public baseline data and associated research demonstrating what is possible. The crash data download provides event-level supplemental information to the SGO that can be directly merged by researchers. The crash information has multiple fields, including KABCO compatible scoring (as assigned by the police reporting), crash type analysis (identifying VRU partner and general crash configuration consistent with police reporting), and air bag deployment status of any vehicle (a commonly reported field in police reporting). The mileage data download provides S2-cell level (a geospatial indexing system) VMT aggregations, which enables detailed exposure comparison to baseline driving fleets. For example, the Federal Highway Administration (FHWA) Highway Performance Monitoring System (HPMS) provides road-level AADT estimates that can be used for VMT estimation across all fifty states [42]. Chen et al. [38] demonstrated that this information can be easily used to distribute surface street and freeway mileage VMT totals according to where that mileage is being accumulated. In another example, Flannagan et al. [21] made use of GM telematics data, which would similarly have GPS breadcrumb data that could be used to do both spatial and temporal modeling of exposure risk.

Implementation

The attainment of future scientific consensus is predicated on the execution of robust and credible research. Given the availability of comprehensive data sources, it is imperative that the implementation of research methodologies accurately assesses these technologies. Consequently, researchers must possess a clear understanding of how foreseeable methodological flaws can lead to erroneous conclusions and, further, how the community can collectively refine its scientific approaches to progress toward a unified scientific consensus.

Following and Advancing Best Practices

The retrospective availability of ADS crash data has, in the past decade, prompted a small surge of research studies. A disconcerting trend within this work, however, involves researchers relying on national, pre-processed aggregate statistics to evaluate the safety performance of highly localized ADS deployments. A summary of these studies can be found in [23]. Tools offered by entities like NHTSA also allow for high-level aggregation of crash data [43]. These statistics are very useful for understanding societal trends in crash risk. There is an understandable inclination to utilize these readily available, processed, secondary data sources, likely due to (a) the inherent credibility and independence of the source (e.g., NHTSA) and (b) the sheer accessibility of the figures. Critically, these pre-processed statistics were not derived with the specific intent of analyzing ADS within the constrained and specialized ODD in which they function.

The better alternative is to rely on the primary data source that is the full data downloads that are available for use, such as what is downloadable directly from NHTSA [44] or from many states throughout the US [22]. Several factors may account for the limited adoption of these primary data sources. First, researchers may be unaware of

their existence. Second, a knowledge deficit often exists regarding the proper techniques for utilizing this data; while crash data manuals are generally accessible, the requisite tooling and processing steps may not be widely known within the research community. Third, there may be unfamiliarity with the range of crash risk confounders that necessitate a highly curated and specific analysis. These aggregate statistics are typically compiled at the national or a broad regional level and segmented by characteristics useful for identifying macro-level trends.

One well documented challenge has been the units used in the crash rates presented in these aggregate stats (RAVE 1C) [19]. In one example, the crash outcome used is represented as the total number of crash events, where the crash events could involve many vehicles [45, 46, 47]. In a second instance, the total number of fatalities are counted, where many vehicles could have been involved in the crash that led to the fatality [48]. . In both scenarios, the represented rates are fundamentally misaligned in their units for comparison against vehicle-level crash rates, which are calculated as the count of crashed vehicle instances divided by a measure of VMT. Specifically, in both instances, the outcome (numerator in the rates) is representative of a crash-level statistic, and is not representative of each involved vehicle's perspective. The use of these off-the-shelf benchmark rates, therefore, simply creates a misleading, mathematically-invalid comparison. This misalignment can, generally, be corrected by using the raw data and computing the statistic at the vehicle-level. For example, rather than count the "crashes" or "fatalities", one would count the instances of "crashed vehicles" or "crashed vehicles in fatal-occurring crashes" [22].

Foundational efforts have already been undertaken to establish best practices for safety performance evaluation. The RAVE Checklist [19], collaboratively developed by a cross-sector group of researchers, including academics, technology developers, independent research organizations, and insurance entities, defined key challenges inherent in current and future research and identified a comprehensive set of recommendations for evaluating the rigor of any given study. It covered all facets of study implementation, broadly classified into categories encompassing (a) quality and validity, (b) transparency, and (c) interpretation. The RAVE Checklist was written, in part, as a way to communicate the potential pitfalls of an overly simplistic approach and motivate individuals to add rigor to their analysis. The authors strongly advocate for a meticulous review of the RAVE checklist, alongside the established body of literature in this domain, as a prerequisite for both conducting and evaluating retrospective safety performance assessments.

The domain of ADS retrospective safety performance assessment is currently the subject of ongoing ISO standardization work (ISO/NP TS 25536). The RAVE checklist served as an important impetus for this developing ISO standard. Although this work is in its preliminary phase, it presents a crucial opportunity for collaboration to develop robust standards that underpin the responsible evaluation of this technology. Global cooperation on these best practices is essential, particularly given the potential variations in available data and specific research questions across different jurisdictions.

Rather than thoroughly review the various issues that might limit study quality and validity, it is worth noting two of the most important implementation requirements when dealing with ADS data and potential baseline data: the alignment of *driving outcomes* (RAVE 1A) and *exposure* (RAVE 1B). As discussed in the prior section, ADS data and baseline data are not necessarily designed to be compatible. Steps must be taken to make the data comparable before meaningful statistics can be derived. Some early work has demonstrated some methodological techniques that can support these goals. Teoh and Kidd [24] and Scanlon et al. [22] demonstrated how state-reported VMT and police crash data can be leveraged to get a vehicle, roadway, and geographic area specific benchmark. Chen et al. [38] presented a benchmarking technique using only public data that can reweigh police reported crash rates according to where and when the vehicle operates. For example, a vehicle only operating at certain lower risk times of day and locations within the city should have a different baseline for comparison. Likewise, comparison between ADS technologies with different temporal and spatial driving exposure requires some alignment before a comparison can be made. Multiple studies [22, 23, 24, 37, 49] have shown that crash outcomes and crash types can be matched between ADS reported data and police reported. This alignment is only possible if the requisite data is shared, as discussed in the previous Data Sources section.

Research Question Focus

The inherent elegance of scientific inquiry into a nascent domain resides in the continuous propagation of novel and consequential research questions. In alignment with any rigorous scientific endeavor, the systematic pursuit and resolution of these inquiries serve to illuminate the extant gaps in our collective knowledge. We remain in the nascent phases of comprehending the full impact that ADS will ultimately exert upon public roadways. Anecdotally, the authors have collectively published a total of eight studies detailing the implementation of retrospective ADS safety impact analysis over the last two years. These publications have consistently generated a substantial influx of expert inquiries concerning the methodologies employed (or omitted) and the trajectory for subsequent research. Thus, these research studies are intentionally posited as foundational stepping stones for future work, and the resultant inquiries serve as a critical roadmap for the entire safety research community.

These inquiries generally coalesce into two distinct categories.

The first set of questions focuses on the "status quo" of driving. The questions directed at investigators examine how the performance of an ADS deployment contrasts with the existing fleet of human drivers currently operating on public roadways. In the Introduction, we referred to the existing fleet's performance as the "Baseline Safety Burden." Measuring the performance relative to this established burden provides critical insight into the technology's overall impact on traffic safety. Historically, safety performance assessments for all previous interventions have prioritized answering these specific types of questions. For example, researchers have traditionally concentrated on inquiries such as, "To what extent does AEB reduce front-to-rear crashes?", where the efficacy of the intervention is being measured against the entirety of the current driving fleet [50, 51].

In the context of these "status quo" research inquiries, our collective experience suggests that the questions generally fall into two distinct categories: (a) was sufficient methodological consideration given to confounding variables such as X, Y, and Z? and (b) has the analysis examined performance across specific, delineated sets of conditions (i.e., safety performance lenses)? Regarding confounding variables, the research community must collaboratively enhance its understanding of the factors that can influence baseline and ADS crash risk, subsequently identifying the requisite data and developing the methodologies necessary to implement the required adjustments. For safety performance lenses, inquiries are almost invariably concentrated on either scenarios known to represent a prevalent safety burden for the human baseline or conditions hypothesized to pose significant challenges for ADS technology. For "status quo" research, the continued implementation of rigorous safety performance assessments on these new questions enable the scientific community to systematically narrow the scope of unknowns regarding the technology's true efficacy.

The second set of questions focuses on expectations of ADS technology. Surpassing the "status quo" constitutes an inherent expectation for ADS technology, particularly when considering that the current vehicle fleet is plagued by high-risk behaviors such as driving without seatbelts, speeding, and impairment [52]. Fraade-Blanar et al. [53] organized the behavioral expectations placed upon ADS vehicles into three distinct categories:

(a) *Empirical expectations:* Beliefs on which behavior will be expected,
(b) *Normative expectations:* Beliefs on which behavior ought to be exhibited, and
(c) *Furtherance expectations:* Beliefs on which behavior could be exhibited based on experiences.

Instead of focusing on isolated behaviors, the assessment of safety impact evaluates aggregate performance. Research inquiries centered on ADS performance expectations have predominantly converged on a synthesis of normative and furtherance expectations. Specifically, this entails identifying the behaviors that ought to be exhibited by these advanced technologies, and determining which behaviors will catalyze substantial improvements to societal safety (i.e., augmenting the current status quo). The foundation of these expectations rests on the premise that ADS ought to epitomize the future state of our transportation network, transcending mere incremental improvements over

the present, transitory status quo. To offer a basic illustration, alcohol-impaired driving is a pervasive and unacceptable element within a safe systems approach [52], leading to decades of concerted effort aimed at diminishing its prevalence. While stakeholders advocate for *all* human drivers to operate without impairment, a logical corollary is that ADS technology should demonstrate performance that is as effective as, or superior to, that of an unimpaired driver.

This raises the fundamental question: is it possible to quantify and contrast exemplary human driving performance against ADS performance? As previously stipulated, any rigorous safety impact analysis necessitates the quantification of both exposure metrics and collision outcomes. As an example, Goodall [54] previously established "model driver" crash rate benchmarks, intended to approximate the performance associated with “sober, rested, attentive, and cautious” driving behavior. This study utilized extant findings from naturalistic driving studies to formulate odds ratios, which compare all driving activities against this idealized driving state. These resulting odds ratios were subsequently applied to adjust overall crash rates, yielding a proxy for model driving performance. In another set of studies, Di Lillo et al. [27, 28] implemented a "contribution" lens to evaluate ADS performance. This approach involved analyzing the frequency at which ADS vehicles incurred third-party property and injury liability claims relative to a baseline of human drivers. Crucially, research inquiries focused on contribution, or accountability, mirror the normative expectation that ADS should diminish instances of conflict initiation, a critical precursor to every potential collision event [17, 18]. Di Lillo et al. [28] further explored how ADS performed in comparison to a "younger generation" of human-driven vehicles. These contemporary vehicles, equipped with the latest ADAS technology, naturally represent the provisional "status quo" of an emerging fleet that will eventually constitute the majority of vehicles. Again though, these benchmarks are secondary to and not representative of the current “status quo” of driving, which must be investigated and understood during this period of active deployments.

Reporting

This section covers three key considerations around safety reporting for ADS. To ensure the robustness and utility of retrospective safety performance assessments, reporting practices should address the following areas:

(1) *Proactively developing benchmark*: This involves clearly communicating about future assessments by proactively publishing the intended benchmarks. This allows stakeholders to understand the long-term assessment strategy and prepares them for upcoming data or findings.

(2) *Reporting windows*: The specific time periods for which performance data is collected and analyzed directly relates to research questions being answered. There are many reporting window strategies options, each of which can provide unique insights into ADS performance.

(3) *Scientific Reporting*: Mechanisms should be in place to allow third parties (e.g., academia, consumer groups, regulators) to independently analyze the raw data or reported findings (e.g., through peer-reviewed publications), thereby strengthening public trust and scientific scrutiny of the safety claims.

<u>Proactively Developing Benchmarks</u>

As the volume of driving exposure expands, the analytical opportunities grow, enabling the statistical evaluation of increasingly rare, detailed, and high-severity safety prevention performance lenses. Intriguingly, a paradoxical relationship emerges between the accumulation of data and the potential for biased analyses. With greater data availability, the risk of "cherry-picking" and "p-hacking" results is amplified. Specifically, by defining an excessively broad spectrum of potential analyses, one could theoretically wait for specific results to achieve statistical significance and then present only those findings (p-hacking), or selectively report results that align with a desired narrative (cherry-picking) [55, 56]. A potential mitigating factor in this paradox is that data accumulation will, eventually, cause these trends to converge toward their true effect size as exposure increases as long as data are re-analyzed.

As noted in preceding sections, there is also the inherent possibility of introducing bias through methodological decisions. While independent post-study review offers a mechanism to detect these errors, a concern arises when an evaluator, such as the company developing the technology or some jurisdiction with regulatory authority, bypasses the peer review process in their reporting or fails to seek expert consultation from individuals knowledgeable in the relevant scientific domain.

The authors recommend that all parties engaged in safety performance research proactively develop benchmarks against which the technology can be developed. The primary goal of this proactive research is to afford experts the opportunity to provide feedback on the proposed methodologies and evaluation criteria, thereby assisting the study performers in identifying and implementing improvements to the study design. This recommendation is applicable to the entire research community, yet it is particularly focused on developers and evaluators of these technologies. Ideally, these benchmarks should be made available before any mileage is accrued or while mileage accumulation remains relatively low.

These benchmarks function as essential indicators, precisely defining the target population for which the ADS technology is designed to enhance safety. These broader analyses also frequently articulate the challenges confronting the general driving public, thereby informing a far wider array of potential use cases. By publicly releasing these benchmark metrics, the research community gains the opportunity to scrutinize their validity and furnish substantive feedback. Since these benchmarks are distinct from the impending performance analysis, the risk of introducing bias with inappropriate/incorrect benchmarks or selectively reporting favorable outcomes is substantially mitigated, allowing the focus to remain strictly on (a) the scope of what can be reported and (b) the methodological rigor of any subsequent comparisons. Key elements to discuss in that benchmarks analysis might include:

(1) The specific research questions that could be under investigation.
(2) The data sources planned for utilization.
(3) The proposed methods for accounting for confounding variables.
(4) The specific crash outcomes that will be studied.
(5) A power analysis for when statistically significant results are anticipated.

Reporting Windows

Reporting windows represent a crucial methodological consideration in the evaluation of ADS performance, with the potential to directly influence the conclusions drawn. Figure 2 illustrates five distinct examples of reporting windows that could be adopted for a safety performance assessment. Temporal effects, such as the year-over-year variation in baseline crash rates [22], are important to consider across all these reporting windows. While each approach possesses inherent merits and limitations regarding the inferences that can be credibly made, these examples are not intended to be an exhaustive catalogue of all possible reporting window designs researchers may choose to employ. Importantly, as discussed in prior sections, the reporting window choice should be made in association with the intended research question and the available data, and the reporting window should be set prior to any analysis to avoid compromising the integrity of the analysis through data fishing.

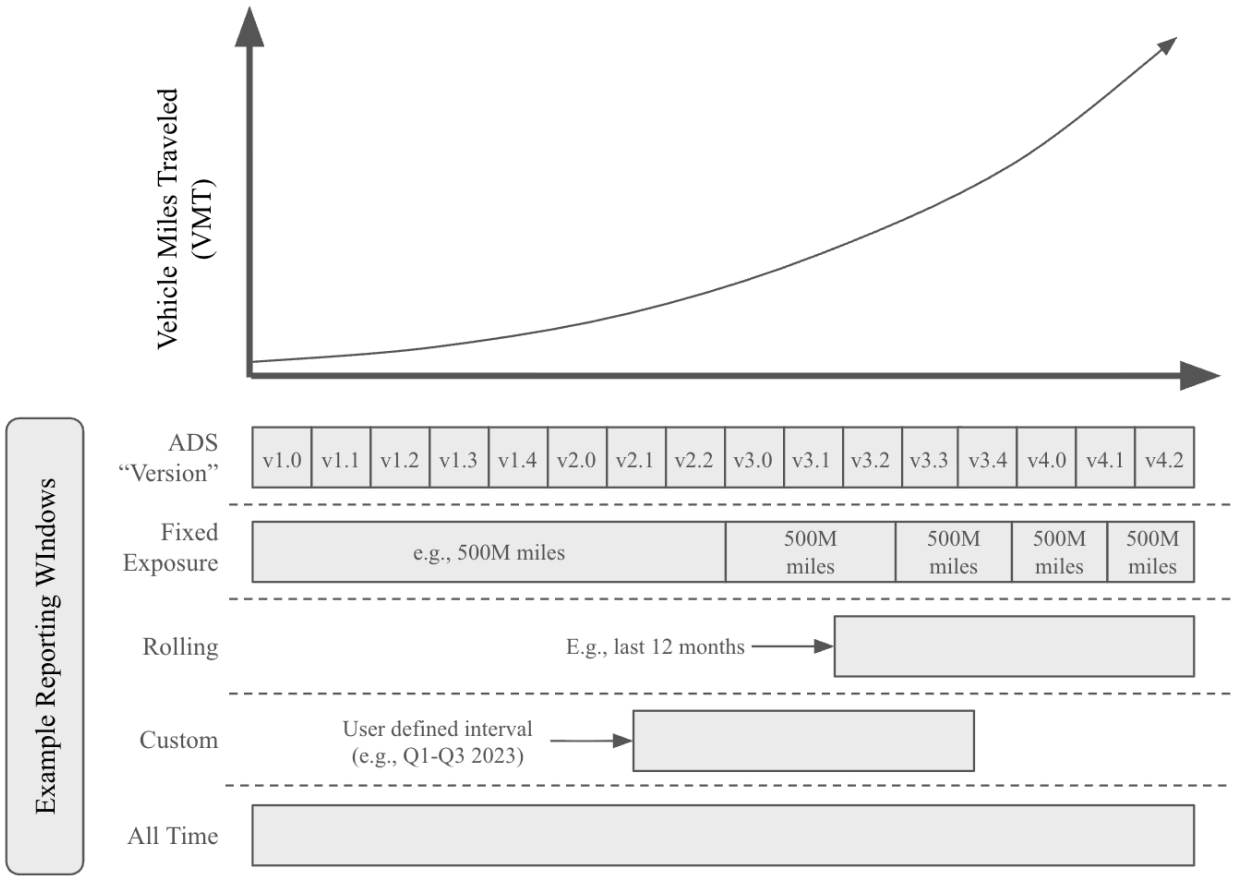

Figure 2. Various reporting windows are shown that could be used in future ADS safety performance assessments.

One consideration for researchers is determining the on-road safety performance of a specific technology version. A "version" can be delineated by various hardware and/or software specifications. Moreover, a specific version embodies the developer's preparatory readiness determination, reflecting their decision to deploy within a defined operational scope based on testing the proposed hardware and software [57, 58]. The merit of a version-based analytical approach lies in its capacity to yield precise performance evaluations for a specific version, which are essential for comparisons against prior iterations of the technology. Given the accumulation of sufficient VMT, this reporting window offers a viable means to assess a particular version currently active on public roadways. The principal constraint, however, is the requirement for adequate mileage to be accrued for each distinct version. Where mileage accumulation rates are low, a given version may never attain the necessary exposure volume for rigorous statistical analysis, particularly in early ADS versions. Furthermore, a challenge exists in establishing the threshold for system modification that officially constitutes a new "version" [59]. As discussed in the "Implementation" section, performance changes with respect to version should be made in consideration of any potential confounders. For example, if an ADS is performing better but is operating in an expanded, higher crash risk ODD, the baseline should reflect the expanded ODD to better isolate the effect of ADS performance, alone.

Another viable option is the systematic, periodic review of the ADS performance utilizing set intervals defined by accumulated exposure volume. While these exposure-volume intervals may inherently encompass multiple versions of the underlying technology, their primary advantage is the effective capture of performance trends over time. Crucially, defining the interval size based on exposure ensures the accumulation of sufficient data volume to enable statistically powered testing across a desired scope of analysis at a given expected effect size.

A rolling window analysis is also capable of providing a salient signal regarding the most contemporary system performance. This methodology incorporates a more recent subset of driving exposure to more accurately reflect the performance of the current, on-road version without necessarily limiting the window to that specific version (only a collection of its most recent forms). Crucially, a rolling window permits control over the quantity of exposure integrated into the analysis, thereby facilitating statistical comparisons across a broader array of analytical lenses. One particularly novel solution proposed involves utilizing the entirety of the performance data but differentially weighting the most recent mileage accumulation within the scoring methodology [59]. This proposition allows for the incorporation of a larger volume of mileage for enhanced statistical power while simultaneously emphasizing the signal from the most up-to-date performance.

In certain instances, a custom reporting window may be justified. For example, if the baseline data is only available for a constrained temporal window, the analysis could similarly limit the reporting window to an equivalent period to appropriately account for potential temporal shifts in baseline risk. For example, police-reported crash data is typically not disseminated in real-time, NHTSA's annual crash data release often lags many months, or even over a year, beyond the conclusion of the prior data cycle [44], etc. State crash databases typically, but not always, have a similar lagging cadence [22]. Although ADS data reporting through the SGO only has a several month data lag, prior work using claims data has needed to look at a longer timeline to provide sufficient time for insurance claims to be filed [27, 28]. Consequently, implementing custom windows that align with this data release cycle allows for a more accurate, "apples-to-apples" comparison between the datasets.

An analysis encompassing all accrued driving by the ADS, designated as "All time," furnishes a signal regarding the technology's cumulative safety performance throughout its operational history. A principal advantage of this methodological choice is the maximal utilization of accumulated mileage, thereby enabling a greater volume of robust statistical analyses. This approach is beneficial in that it does not disregard the performance of prior ADS iterations; however, a corresponding limitation is its diminished sensitivity to subsequent technological improvements as the total accumulated mileage continues to grow. Accordingly, although this reporting window may be useful to continuously monitor, its relevance in assessing the current performance of the technology diminishes over time. Generally speaking, an "all time" reporting window does not, necessarily, provide an analysis of the current version on the road, but does provide an overview of how well the technology has historically performed over many iterations of development, readiness assessment, and responsive upgrades.

Scientific Reporting

The opportunity for introducing bias into the research, like in any scientific discipline, is present. This bias can arise for a variety of reasons, including inaccurate implementation of the research (see preceding section) or simply through the personal perspective of the individual(s) conducting the research. There are ways to mitigate this potential bias. For example, as discussed previously, the methods and data used for safety impact studies should be made available for independent evaluation and extension of results. Additionally, also discussed, establishing and adhering to best practices when performing studies also minimizes the potential for bias. Beyond these data and implementation needs, there needs to be scientific reporting of results. The alternative to this is unchecked, ad hoc analyses that might contain bias. To uphold these mechanisms of transparency and accountability, it is essential that the reporting have independent oversight. We strongly recommend leveraging one or more of these three basic mechanisms:

(1) the peer review process,
(2) providing the analysis to independent parties, and
(3) cooperative study partnerships with credible third parties.

The peer review process is an important mechanism of scientific research. This is not to say that poor research cannot make it through this process. The current study even discusses some prior work that introduced bias through

methodological decisions. But, as we collectively develop best practices, the peer review process is how we uphold the continual improvement of our shared research area.

Mostly applicable to ADS developers, making the crash and exposure data and analysis details available in a format that can be used to independently reproduce the results to independent researchers is an important way to create strong, credible research. A wider dispersion of this data is even more effective as each perspective puts a new lens on the data. If accumulation or credible research is what drives, ultimately, scientific consensus around efficacy, then getting the data to independent research leaders is essential. One way to accomplish this could be leveraging third-party compliance audits, where an independent assessor has the ability to examine the analysis and verify that best practices were applied. Self assessments can also be leveraged, documented, and communicated alongside any individual study to communicate how the study performed attempted to conform with established best practices.

A third opportunity is through implementing cooperative study partnerships. This model creates a group of stakeholders, each with their own perspective, that collectively want to understand the efficacy of the technology. There are two noteworthy models that are worth discussing and have advantages. In one example, the Partnership for Analytics Research in Traffic Safety (PARTS) is a voluntary, shared research collaboration between the government and many automotive OEMs (nearly 70% of the US market) to study the effectiveness of ADAS [50, 51, 35]. The collaboration has a third party, MITRE Corporation, that acts as a data intermediary. The partnership has certain protections around IP and competitor identification, hence the inclusion of MITRE, but also has put out two independent reports on fleetwide ADAS effectiveness. This partnership provided data access for a wide array of ADAS features that have ultimately enabled an unprecedented level of insights by both technology confounders, such as driver characteristics, location, and road features. In a second partnership, Waymo collaborated with Swiss ReInsurance, one of the world's largest reinsurance companies, to study the safety performance of the Waymo ADS. As an independent evaluator, Swiss ReInsurance's participation can help provide guidance for the insurance industry on how to assess ADS risk. There is a shared interest between both parties in having a fair, accurate assessment of the technology. The partnership produced two studies; one of those studies has completed the peer review process, while the other is still under review [27, 28].

CONCLUSIONS

The research community must rapidly commence building the empirical evidence required for scientific consensus on ADS efficacy, leveraging the substantial influx of safety-relevant data. Retrospective safety performance assessments are uniquely challenging for ADS deployments due to the specific ADS exposure and unprecedented level of data available. Achieving scientific consensus requires collaboration in releasing safety data, ensuring methodological study rigor, and making sure that there is transparency and oversight in safety performance reporting. This paper critically evaluates the readiness of the community to perform ADS retrospective safety performance assessments, identifying gaps, and proposing corrective actions. The following key actions are identified:

- ADS developers, regardless of government mandate, must **release VMT data** to enable crash rate calculation.
- **Crash and exposure granularity must also supplement existing SGO data** – focusing on exposure confounders (e.g., where and when VMT is accumulated), crash outcome units (e.g., severity), and performance lenses (e.g., crash types), that can be used to compare against existing benchmark data (e.g., police reported crash data and public VMT).
- Researchers should avoid aggregate statistics, utilize primary data, and tailor analyses to **improve benchmark and crash data compatibility in a way that minimizes potential bias and relies on established state-of-the-art**.

- **Researchers should adhere to established best practices outlined in the RAVE checklist**. Stakeholders should engage to further develop methodology and update these best practices. Shared best practices directly support the promotion of credible research and mitigates risk of improper, biased results.
- **The safety community must continue to ask and address novel research questions** focused on both the "status quo" of human driving and the higher "expectations" placed upon ADS technology.
- **Benchmarks from evaluators and developers should be proactively published to solicit expert feedback**, mitigate the risk of "p-hacking" or "cherry-picking," and improve methodological rigor.
- **Multiple reporting windows should be considered** and published to enable distinct insights into ADS performance trends.
- **Independent oversight of any scientific reporting must be enacted** through (1) the peer review process, (2) providing data for independent analysis, and (3) establishing cooperative partnerships.

Collectively achieving these actions will help to generate scientific consensus. If the effectiveness of this technology can be widely recognized, the societal adoption rate can be improved and the ongoing injury burden can be more expeditiously mitigated.

REFERENCES

[1] National Center for Statistics and Analysis. (2025, April). Early estimate of motor vehicle traffic fatalities in 2024 (Traffic Safety Facts Crash Stats Brief Statistical Summary. Report No. DOT HS 813 710). National Highway Traffic Safety Administration.

[2] Federal Highway Administration (FHWA), “December 2024 Traffic Volume Trends,” US DOT, accessed November 2025, https://www.fhwa.dot.gov/policyinformation/travel_monitoring/24dectvt/page2.cfm.

[3] Bertrand, H. A. (1958). U.S. Patent No. 2,834,606. Washington, DC: U.S. Patent and Trademark Office.

[4] Campbell, D. D. (1972). Air cushion restraint systems development and vehicle application (No. 720407). SAE Technical Paper.

[5] Viano, D. C.. History of airbag safety benefits and risks. Traffic Injury Prevention, 25 no. 3 (2024): 268–287, https://doi.org/10.1080/15389588.2024.2315889

[6] Atarod, M., "Occupant Kinematics during Moderate-to-High Speed Side Impacts: An Analysis of IIHS Crash Data over the Past Decade," SAE Technical Paper 2020-01-5165 (2020), https://doi.org/10.4271/2020-01-5165.

[7] Scanlon, J. M., Isaacs, J., & Garman, C. Head and neck loading conditions over a decade of IIHS rear impact seat testing, SAE Technical Paper 2019-01-1227 (2019), https://doi.org/10.4271/2019-01-1227

[8] Teoh, E. R., How Have Changes in Front Air Bag Designs Affected Frontal Crash Death Rates? An Update. Traffic Injury Prevention, 15 no. 6 (2014): 606–61, https://doi.org/10.1080/15389588.2013.853295

[9] NHTSA, Special crash investigations counts of frontal air bag related fatalities and seriously injured persons (DOT HS 811 104). Washington, DC: National Highway Traffic Safety Administration (2009).

[10] National Center for Statistics and Analysis. (2019, April). Occupant protection in passenger vehicles: 2017 data (Traffic Safety Facts.Report No. DOT HS 812 691). Washington, DC: National HighwayTraffic Safety Administration.

[11] ISO 21934-1. 2021. Road vehicles — Prospective safety performance assessment of pre-crash technology by virtual simulation — Part 1: State-of-the-art and general method overview. ISO Standard 219341. https://www.iso.org/standard/76497.html.

[12] ISO 21934-2. 2024. Road vehicles — Prospective safety performance assessment of pre-crash technology by virtual simulation — Part 2: Guidelines and requirements for application. ISO Standard 219341. https://www.iso.org/standard/76497.html.

[13] SAE International Surface Vehicle Recommended Practice, “(R) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,” SAE Standard J3016, Revised April 2021.

[14] Pintar, F. A., Yoganandan, N., & Gennarelli, T. A. (2000). Airbag effectiveness on brain trauma in frontal crashes. In Annual Proceedings/Association for the Advancement of Automotive Medicine (Vol. 44, p. 149).
[15] Teoh, E. R. (2021). Effectiveness of front crash prevention systems in reducing large truck real-world crash rates. Traffic injury prevention, 22(4), 284-289.
[16] Cicchino, J. B. (2017). Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. Accident Analysis & Prevention, 99, 142-152.
[17] Scanlon, J. M., Kusano, K. D., Daniel, T., Alderson, C., Ogle, A., & Victor, T. (2021). Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. Accident Analysis & Prevention, 163, 106454.
[18] Scanlon, J. M., Kusano, K. D., Engström, J., & Victor, T. (2022). Collision avoidance effectiveness of an automated driving system using a human driver behavior reference model in reconstructed fatal collisions. Waymo, LLC.
[19] Scanlon, J. M., Teoh, E. R., Kidd, D. G., Kusano, K. D., et al., "RAVE checklist: Recommendations for overcoming challenges in retrospective safety studies of automated driving systems," Traffic Injury Prevention 26, no. 5 (2025): 608–621. https://doi.org/10.1080/15389588.2024.2435620
[20] National Highway Traffic Safety Administration (NHTSA). 2025. Third Amended Standing General Order 2021-01: incident Reporting forAutomated Driving Systems (ADS) and Level 2 Advanced DriverAssistance Systems (ADAS).
[21] Flannagan C, Leslie A, Kiefer R, Bogard S, Chi-Johnston G, Freeman L,Huang R, Walsh D, Joseph A. 2023. Establishing a crash rate benchmark using large-scale naturalistic human ridehail data. Ann Arbor (MI):University of Michigan Transportation Research Institute (UMTRI).
[22] Scanlon, J. M., Kusano, K. D., Fraade-Blanar, L. A., McMurry, T. L., Yin-Hsiu, C. and Victor, T., “Benchmarks for Retrospective Automated Driving System Crash Rate Analysis Using Police-Reported Crash Data.” Traffic Injury Prevention 25, no. sup1 (2024): S51–65. doi:10.1080/15389588.2024.2380522.
[23] Scanlon, J. M., McMurry, T. L., Yin-Hsiu, C., Kusano, K. D., and Victor, T., "From Stoplights to On-Ramps: A Comprehensive Set of Crash Rate Benchmarks for Freeway and Surface Street ADS Evaluation," arXiv preprint arXiv:2508.19425.
[24] Teoh, E. R., Kidd, D. G. Rage against the machine? Google’s self-driving cars versus human drivers. J Safety Res. 63 (2017):57–60. doi:10.1016/j.jsr.2017.08.008.
[25] Blanco M, Atwood J, Russell S, Trimble T, McClafferty J, Perez M. 2016.Automated vehicle crash rate comparison using naturalistic data.Final report. Blacksburg, VA: Virginia Tech Transportation Institute.
[26] Campolettano, E. T., Scanlon, J. M., Kadar, I., Lavy, L. Y., Moura, D. C., & Kusano, K. D. (2024). Baseline vulnerable road user injury risk in multiple US dense urban driving environments. Traffic Injury Prevention, 25(sup1), S94-S104.
[27] Di Lillo L, Gode T, Zhou X, Atzei M, Chen R, Victor T. 2024. Comparative safety performance of autonomous-and human drivers: a real-world case study of the Waymo one service. Heliyon.10(14):e34379. doi:10.1016/j.heliyon.2024.e34379.
[28] Di Lillo L, Gode T, Zhou X, Scanlon JM, Chen R, Victor T. 2025. Do autonomous vehicles outperform latest-generation human-driven vehicles? a comparison to waymo’s auto liability insurance claims at25.3M miles. Mountain View (CA): Waymo LLC.
[29] Chen JJ, Shladover SE. 2024. Initial Indications of Safety of Driverless Automated Driving Systems. arXiv preprint arXiv:2403.14648.
[30] Waymo, “Waymo Safety Impact: Downloads,” Waymo LLC, accessed December 2025, https://waymo.com/safety/impact#downloads.
[31] Cicchino, J. B., "Effects of automatic emergency braking systems on pedestrian crash risk," Accident Analysis & Prevention 172 (2022): 106686.
[32] Evans, L., "Double pair comparison—a new method to determine how occupant characteristics affect fatality risk in traffic crashes," Accident Analysis & Prevention 18 no. 3 (1986): 217-227.

[33] Kullgren, A., Amin, K., & Tingvall, C., "Effects on crash risk of automatic emergency braking systems for pedestrians and bicyclists," Traffic injury prevention 24 no. sup1 (2023): S111-S115.
[34] Stamatiadis, N., & Deacon, J. A., "Quasi-induced exposure: methodology and insight. Accident Analysis & Prevention", 29, no. 2 (1997):37-52.
[35] Partnership for Analytics Research in Traffic Safety (PARTS), "Advanced Driver Assistance SystemCrash Rate Assessment Using Vehicle-Specific Mileage Data", The MITRE Corporation Case Number 25-1823 (2025).
[36] Werner, G., Modlin, C., & Watson W. T., "Basic ratemaking. In Casualty Actuarial Society" 5th ed. (Casualty Actuarial Society, 2016).
[37] Kusano, Kristofer D., John M. Scanlon, Yin-Hsiu Chen, Timothy L. McMurry, Tilia Gode, and Trent Victor. 2025. “Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles.” Traffic Injury Prevention, May, 1–13. doi:10.1080/15389588.2025.2499887.
[38] Chen, Y. H., Scanlon, J. M., Kusano, K. D., McMurry, T. L., Victor, T., "Dynamic benchmarks: spatial and temporal alignment for ADS performance evaluation, Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981251398744 (2025).
[39] Bergel-Hayat, Ruth, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. "Explaining the road accident risk: Weather effects." Accident Analysis & Prevention 60 (2013): 456-465.
[40] Farmer, C. M. (2019). The effects of higher speed limits on traffic fatalities in the United States, 1993–2017. Insurance Institute for Highway Safety.
[41] Hu, Wen, and Jessica B. Cicchino. "Effects of lowering speed limits on crash severity in Seattle." Journal of safety research 88 (2024): 174-178.
[42] Federal Highway Administration. Highway Performance Monitoring System Field Manual. https://www.fhwa.dot.gov/policyinformation/hpms/fieldmanual/page01.cfm, 2018. Accessed Jul. 22,2024.
[43] NHTSA, “Fatality and Injury Reporting System Tool (FIRST),” National Highway Traffic Safety Administration, accessed December 2025, https://cdan.dot.gov/query.
[44] NHTSA, “NHTSA File Downloads,” National Highway Traffic Safety Administration, accessed December 2025, https://www.nhtsa.gov/file-downloads.
[45] Schoettle B, Sivak M. 2015. A preliminary analysis of real-world crash-es involving self-driving vehicles. Report no. UMTRI-2015-34. AnnArbor, MI: University of Michigan Transportation Research Institute.
[46] Banerjee S, Jha S, Cyriac J, Kalbarczyk ZT, Iyer RK. 2018. Hands off the wheel in autonomous vehicles?: a systems perspective on over a million miles of field data. In: 2018 48th Annual IEEE/IFIPInternational Conference on Dependable Systems and Networks(DSN), IEEE. p. 586–597. doi:10.1109/DSN.2018.00066.
[47] Cummings M. 2024. Assessing readiness of self-driving vehicles. In:The 103rd Transportation Research Board (TRB) Annual Meeting.Washington, D.C.
[48] Kalra N, Paddock SM. 2016. Driving to safety: how many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp ResPart A Policy Pract. 94:182–193. doi:10.1016/j.tra.2016.09.010.
[49] Kusano, K. D., Scanlon, J. M., Chen, Y. H., McMurry, T. L., Chen, R., Gode, T., & Victor, T. (2024). Comparison of Waymo rider-only crash data to human benchmarks at 7.1 million miles. Traffic Injury Prevention, 25(sup1), S66-S77.
[50] Partnership for Analytics Research in Traffic Safety (PARTS), "A Study on Real-world Effectiveness of Model Year 2015–2020 Advanced Driver Assistance Systems", The MITRE Corporation Case Number 22-3734 (2022).
[51] Partnership for Analytics Research in Traffic Safety (PARTS), "A Study on Real-world Effectiveness of Model Year 2015–2023 Advanced Driver Assistance Systems", The MITRE Corporation Case Number 25-0114 (2025).
[52] United States Department of Transportation. (2022). National roadway safety strategy. https://www.transportation.gov/NRSS/SafeSystem

[53] Fraade-Blanar, L., Favarò, F., Engstrom, J., Cefkin, M., Best, R., Lee, J., & Victor, T. (2025). Being good (at driving): Characterizing behavioral expectations on automated and human driven vehicles. arXiv preprint arXiv:2502.08121.

[54] Goodall N. J., "Potential crash rate benchmarks for automated vehicles," Transp Res Rec. 2675 no. 10 (2021):31–40. doi:10.1177/03611981211009878.

[55] Andrade, C, "HARKing, cherry-picking, P-hacking, fishing expeditions, and data dredging and mining as questionable research practices," J Clin Psychiatry 82 vol. 1 (2021):20f13804. https://doi.org/10.4088/JCP.20f13804.

[56] Gelman, A., Loken, E., "The garden of forking paths: why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time," Depart Statist Columb Univ. 348, no. 3 (2013):1-17.

[57] Favaro, F., Fraade-Blanar, L., Schnelle, S., Victor, T., Peña, M., Engstrom, J., ... & Smith, D. (2023). Building a credible case for safety: Waymo's approach for the determination of absence of unreasonable risk. arXiv preprint arXiv:2306.01917.

[58] Favaro, F., Schnelle, S., Fraade-Blanar, L., Victor, T., Peña, M., Webb, N., ... & Smith, D. (2025). Determining Absence of Unreasonable Risk: Approval Guidelines for an Automated Driving System Deployment. arXiv preprint arXiv:2505.09880.

[59] Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). Measuring automated vehicle safety: Forging a framework.