Automated Vehicle Safety Consortium™ Best Practice

AVSC00012202308

Issued 2023-08

AVSC Best Practice for Developing ADS Safety Performance Thresholds Based on Human Driving Behavior

Citation: Automated Vehicle Safety Consortium. 2023. AVSC Best Practice for Developing ADS Safety Performance Thresholds Based on Human Driving Behavior. SAE Industry Technologies Consortia.

Rationale

Automated driving system (ADS) developers need a way to describe safe and competent driving for automated driving system-dedicated vehicles (ADS-DVs) in a way that is relatable to how stakeholders interpret safe driving today. Metrics informed by competent and safe human behavior could improve understanding and confidence in ADS-DVs. One way to make ADS safety performance relatable to stakeholders is to adopt an intuitive comparison to behaviors displayed on the road by human drivers.

It is important to acknowledge that not all aspects of human driving are socially acceptable. Therefore, the behaviors observed in naturalistic driving studies (NDS) are likely to include both desirable and undesirable qualities. Consequently, caution should be exercised when utilizing NDS data as one of several inputs to inform safety performance standards for ADS. To ensure safety, ADS developers are expected to incorporate various data sources, define parameter sets, and establish safety thresholds throughout the design, construction, verification, and validation (V&V) processes. This best practice should be viewed as an illustrative approach that can be further generalized and replicated for other data sources and datasets beyond those discussed in this best practice.

Preface

The Automated Vehicle Safety Consortium™ (AVSC) is an industry program of SAE Industry Technologies Consortia® (SAE ITC). The AVSC shares information to inform and accelerate industry-wide standards and advance the safe development, deployment, and fleet operations of automated driving systems (ADSs). The members of this consortium have decades of accumulated experience, including millions of cumulative miles of physical and simulated ADS testing focused on safer, reliable, high-quality transportation. They are committed to applying their experience and combined knowledge to earn public confidence in the safe operation of SAE level 4 and level 5 automated vehicles.

The wide range of technologies, use cases, and operating domains create unique challenges with public perception of ADSs. The consortium recognizes the beneficial role best practices and information reports can have for the industry and for the safe operation of SAE level 4 and level 5 automated driving system-dedicated vehicles (ADS-DVs). These technology-neutral documents provide key considerations for safely deploying ADS-DVs on public roads. AVSC documents are based on current state-of-the-art technology and the experiences of the AVSC members. AVSC members currently support, or intend to support, the best practices or equivalent measures to set a bar for other industry participants to meet.

Technology advances rapidly, and new information is becoming available at an increasing rate. The AVSC's best practices and information reports are living documents. As knowledge and experience grow, our publications will be revisited and updated, as needed, to continue to support the safer on-road use of ADS-DVs. Comments and open discussion on the topics are welcome in appropriate industry forums.

Introduction

Historically, safety performance measures have played a vital role in evaluating the state and progress of road traffic safety. In the United States, the National Highway Traffic Safety Administration (NHTSA) has been responsible for reporting road traffic safety statistics and conducting analyses using data from on-road vehicles; to date, mostly consisting of human drivers. These measures serve as important indicators that inform Congress, the general public, and other relevant parties (e.g., regulatory agencies) about the status of road safety [1]. These traffic safety performance measures provide a good and relatable reference point for understanding traffic safety changes year over year.

NHTSA's databases have traditionally focused on lagging indicators of safety (i.e., collision statistical analysis). This best practice aims at providing a complementary approach to inform safety by focusing specifically on leading behavioral evaluation [2] [3]. The approach proposed here involves comparing the on-road driving behavior of ADSs to that of human drivers, aiming to gain a deeper understanding of ADS behavior and its implications. The assessment centers on the predictability and similarity of ADS behavior relative to observed human driving behavior. By adopting this approach, a more comprehensive evaluation of the performance of ADS fleets can be achieved. This empowers developers to identify areas for improvement and potentially unknown unsafe behaviors.

Reference points for assessing the relative safety performance of ADS fleets are crucial. The safety performance of some ADS behaviors can then be measured and compared to NDS data from human drivers to help characterize the socially acceptable balance between safety, lawful driving, efficiency, and comfort. For instance, developers can utilize human-driver data to determine an appropriate minimum passing distance when vulnerable road users (VRUs) are present. This process enables them to enhance the safety performance of ADS fleets by aligning with human-relative benchmarks and considerations.

The recommended set of safety performance metrics from AVSC00006202103 and the behavioral competencies from AVSC00008202111 provide a starting point for assessing ADS behavior in a specific operational design domain (ODD). This best practice builds upon those previous efforts by providing ADS developers with a third component for safety assurance by identifying:

- The specific behavioral competency of interest for evaluation.
- The applicable safety performance metrics.
- The data that can be leveraged to define safety performance reference values.

This best practice outlines a process for leveraging human driving data to establish safety performance targets for ADS-DV behaviors. The targets within the specific use-case, exemplified in this best practice, are based on naturalistic driving data from manually driven vehicles, in the hope of aiding understanding from a broad audience of stakeholders.

Several studies have suggested that using human drivers as a reference point for an ADS-DV safety approach provides a relatable framing for behavioral evaluation. For instance, the RAND Corporation conducted research that involved interviewing a diverse group of autonomous vehicle (AV) stakeholders and surveying the general public to determine the impact of using human driving data on AV safety [4]. It concluded that using human drivers as a reference point for AV safety is generally consistent with public expectations. People are already accustomed to the risks associated with human-driven vehicles, and they expect ADS-DVs to outperform average human drivers.

Throughout the document, we leverage an example use-case associated with a specific behavioral competency and analyze human drivers' behaviors through the usage of NDS. As mentioned previously, the outlined process can be generalized and abstracted to remain applicable to other data sources and comparison points to establish safety performance thresholds. The data and statistical examples used are for illustrative purposes, only.

Table of Contents

1. Scope

Human driving data[1] provides one possible source for establishing reference values (i.e., performance threshold) for ADS-DVs. This best practice describes a framework to establish human driving data reference values for ADS safety performance metrics. This framework provides a way to assess different behaviors and is broadly generalizable. Key characteristics of datasets that can be used to establish these reference values for ADS are described in an objective, repeatable, and explainable way.

This best practice offers guidance on ensuring data quality and conducting appropriate analyses, including sample size determination, error analysis and interpretation, variance, standard deviation, segmentation, and normalization. The guidelines established in this best practice specifically concentrate on interactions among road users, such as vehicles and pedestrians. Other factors (e.g., presence of objects like debris or construction equipment) and other aspects of safety behavior that do not involve the interaction between two road users, are beyond the scope of this best practice.

1.1 Purpose

This best practice is intended for use by the technical community (developers, manufacturers, testers, etc.) to aid in the development, validation, and safe deployment of SAE level 4 and level 5 ADSs. It may also be useful to public agencies and stakeholders, including standards bodies and governmental decision-makers, who have interest in better understanding the safety posture of ADS deployments.

This best practice supports public and private organizations in preparing for and deploying ADS-DV systems. For example, it may be used by ADS manufacturers and developers to establish quantitative performance baselines based on AVSC safety metrics for various driving behaviors. It is intended to garner discussion, foster public understanding, and promote acceptance of ADS-DVs.

2. References

2.1 Applicable Documents

The following publications were referenced during the development of this document. Where appropriate, documents are cited.

2.1.1 SAE Publications

Unless otherwise indicated, the latest issue of SAE publications applies. Available from SAE International, 400 Commonwealth Drive, Warrendale, PA 15096-0001, Tel: 877-606-7323 (inside USA and Canada) or +1 724-776-4970 (outside USA), www.sae.org.

AVSC00002202004 AVSC Best Practice for Describing an Operational Design Domain: Conceptual Framework and Lexicon

AVSC00004202009 AVSC Best Practice for Data Collection for Automated Driving System-Dedicated Vehicles (ADS-DVs) to Support Event Analysis

AVSC00006202103 AVSC Best Practice for Metrics and Methods for Assessing Safety Performance of Automated Driving Systems (ADS)

[1] Other relevant data sources may include crash datasets to understand the causes of crashes or driver simulator data to simulate a wide range of driving scenarios in a controlled environment for ADS testing or performance validation purposes and other public datasets like nuScenes, nuPlan, or Waymo Open dataset.

AVSC00008202111	AVSC Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (ADS-DVs)
AVSC00009202208	AVSC Best Practice for Interactions Between ADS-DVs and Vulnerable Road Users (VRUs)
SAE J3016_202104	Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles
SAE J3216_202107	Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles
SAE J3164_202301	Ontology and Lexicon for Automated Driving System (ADS) - Operated Vehicle Behaviors and Maneuvers in Routine/Normal Operating Scenarios

2.1.2 Other Documents

[1] NHTSA. (2018). *NHTSA's safety performance measures selection criteria* (Report No. DOT HS 812 628). National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812628.

[2] Inouye, J. (2015). *Practical guide to leading indicators: Metrics, case studies & strategies*. Campbell Institute, National Safety Council. https://www.thecampbellinstitute.org/wp-content/uploads/2017/05/Campbell-Institute-Practical-Guide-Leading-Indicators-WP.pdf.

[3] CCPS. (2011). *Process safety leading and lagging metrics… You don't improve what you don't measure*. Center for Chemical Process Safety. https://www.aiche.org/sites/default/files/docs/pages/CCPS_ProcessSafety_Lagging_2011_2-24.pdf.

[4] Blumenthal, M.S., Fraade-Blanar, L., Best, R., and Irwin, J.L. (2023). *Safe enough: Approaches to assessing acceptable safety for automated vehicles*. RAND Corporation. https://doi.org/10.7249/RRA569-1.

[5] ISO. (2018). *Road vehicles — Functional safety — Part 1: Vocabulary* (Standard no. ISO 26262-1:2018). International Organization for Standardization. https://www.iso.org/standard/68383.html.

[6] Fraade-Blanar, L., Blumenthal, M.S., Anderson, J.M., and Kalra, N. (2018). *Measuring automated vehicle safety*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2662.html.

[7] NHTSA. (2021, October 19-21). *ADS research* [Virtual conference presentation]. NHTSA Safety Research Portfolio Public Meeting. https://downloads.regulations.gov/NHTSA-2021-0060-0010/attachment_4.pdf.

[8] Pollard, J,H. (1979). *A handbook of numerical and statistical techniques: With examples mainly from the life sciences*. Cambridge University Press.

[9] VTTI. (2022). *Project highlights: Automated mobility partnership*. Virginia Tech Transportation Institute. https://www.vtti.vt.edu/projects/amp.html.

[10] Helwig, N.E. (2020). *One sample t test*. University of Minnesota. http://users.stat.umn.edu/~helwig/notes/OneSampleTest.pdf.

[11] NCSL. (2022). *Safety passing bicyclists chart*. National Conference of State Legislatures. https://www.ncsl.org/research/transportation/safely-passing-bicyclists.aspx.

[12] NIST. (2012). Critical values of the student's *t* distribution. *NSIT/SEMATECH e-handbook of statistical methods.* National Institute of Standards and Technology. https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm.

[13] Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*(2), 73-80. https://doi.org/10.1016/j.apergo.2004.01.002.

[14] Seltman, H.J. (2018). *Experimental design and analysis*. Carnegie Mellon University. https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf.

[15] Karayel, D., Topakci, M., Ünal, İ., Šarauskis, E., and Canakci, M. (2012). Using real-time kinematic (RTK) global positioning system (GPS) for the determination of seedling distributions over the field. *Zemdirbyste*, *99*(4), 425-430.

[16] Mohammed, T.H.H. and Trilaksono, B.R. (2018). Integrated INS/GPS navigation system. *International Journal on Electrical Engineering and Informatics*, *10*(3) 491-512. https://doi.org/10.15676/ijeei.2018.10.3.6.

[17] Hasegawa-Johnson, M. (2017, September 5). *ECE 417 Lecture 3: 1-D Gaussians*. University of Illinois, Urbana Champaign. https://courses.engr.illinois.edu/ece417/fa2017/ece417fa2017lecture3.pdf.

3. Definitions

3.1 Operational Design Domain (ODD) (SAE J3016_202104)

Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics.

3.2 [Safety] Metric

A measurement used to evaluate and track safety performance.

3.3 Behavioral Competency

Expected and measurable capability of an ADS feature operating a vehicle within its ODD.

NOTE: Competency refers to the term “expected” in the definition. Using skills, knowledge, and abilities, an ADS executes behaviors competently according to performance criteria set by the ADS developer.

3.4 Safety Outcome Metrics

A direct measurement of actualized outcomes or adherence to societal norms.

NOTE 1: Safety outcomes temporally lag deployment. It can take considerable time to collect a sufficient sample size to establish statistically significant measurements.

NOTE 2: Societal norms may differ by industry, geographic regions, and application.

NOTE 3: Although an ADS is expected to have benefits to other societal outcomes, such as mobility and accessibility, the focus of this work is on safety outcomes only.

NOTE 4: ADS safety outcome metrics concern a variety of ADS market penetration rates but are generally assumed to be commercial-scale deployments, and not pilots with vehicles operated by highly trained drivers employed by ADS technology developers or small-scale demonstrations, such as very low volume deployments, limited geographic regions, or tightly limited ODD.

3.5 Scene

A snapshot of the environment including the scenery, dynamic elements, and all actor and observer self-representations, and the relationships between those entities.

NOTE 1: This best practice specifically references operational scenes encountered by an ADS-DV.

NOTE 2: Only a scene representation in a simulated world can be all-encompassing (i.e., an objective scene or ground truth). In the real world, the scene is incomplete, incorrect, uncertain, and from one or several observers' points of view (i.e., a subjective scene).

NOTE 3: A scene is a descriptive representation of the state of the world at a point in time. A scenario consists of a sequence of scenes.

Example: At an instant in time, an ADS-DV is traveling at 35 km/h, in the right-hand lane on an arterial roadway in clear conditions, while another human-operated vehicle travels in the adjacent left lane at 33 km/h with the ADS-operated vehicle located in the blind spot of the human-operated one.

3.6 [Operating] Scenario

A description of the temporal development through several consecutive scenes in a sequence of scenes.

NOTE 1: Every scenario starts with an initial scene. In contrast to a scene, a scenario spans a certain amount of time. Actions and events can be specified as transitions between scenes to characterize the temporal development within a scenario. Scenes in a scenario can also be augmented with goals, values, and beliefs of the traffic participants, resulting in a sequence of situations.

NOTE 2: Scenarios may be defined over varying durations. A scenario may overlap with or be completely contained within another scenario. For example, an overtaking scenario may be decomposed into three scenarios: lane change scenario, followed by lane maintenance scenario, followed by lane change scenario.

NOTE 3: Scenarios may be defined at varying levels of abstraction, ranging from individual quantitative scenarios to quantitative classes of one or more scenarios, to qualitative scenario classes with narrative descriptions.

NOTE 4: This best practice specifically references emergency and nonemergency scenarios encountered by ADS-operated vehicles.

NOTE 5: The term "operating" in this scenario definition refers to dynamic driving task (DDT) performance (as opposed to, for example, a post-crash scenario in which a first responder is interacting with an ADS-equipped vehicle that is no longer performing the DDT). It also comprehends all types of operating scenarios, such as test scenarios (whether on track or in simulation), as well as scenarios encountered on road.

3.7 Safety Envelope

A kinematically defined state space around a vehicle that represents a buffer between the subject vehicle and other objects in the scene. A safety envelope may vary with specific context.

3.8 Hazardous Event

The combination of a vehicle-level hazard and an operational situation [5] of the vehicle with potential to lead to a harmful event (e.g., a collision) if not controlled.

4. Establishing Reference Values for Safety Performance Metrics

ADS developers and manufacturers should establish context-relevant[2] reference values and their uncertainties for safety performance metrics. In AVSC00006202103, several recommendations were brought forward that help ensure a robust evaluation of performance is undertaken.

Metrics employed in safety evaluations need to be complemented with reference values (and associated confidence intervals) that qualify the ADS performance as a comparison to such targets. For instance, if a metric focuses on the minimum passing distance between an ADS vehicle and VRUs, a target buffer would need to be set to enable the correct identification of notable events and the evaluation of the ADS's behaviors. This best practice outlines a process that can be followed to establish such targets, which, in turn, enable a clear evaluation of potential violations.

Reference values should be commonly understandable to all stakeholders involved in the development and deployment of ADS systems, including regulatory bodies, developers, manufacturers, and end users. They should be valid, reliable, and feasible, such that they accurately reflect the desired safety performance and can be consistently measured and achieved in real-world scenarios. In addition, they should be non-manipulatable to prevent alteration to a desired outcome. These values should be determined objectively using repeatable methods of evaluation [6]. NDS data of human drivers can be one such input to setting these values.

This best practice focuses on normal/routine driving conditions in which the ADS and other road users are expected to adhere to the rules of the road and broader societal expectations, as well as reasonably foreseeable situations that may arise during driving, such as traffic congestion or pedestrian crossings.

A specific safety performance metric and behavioral competency, as noted in 5.1.1 (maintaining a lane while passing a pedestrian), is utilized throughout this best practice as an illustrative example of a method to establish reference values. This method is applicable for all safety metrics and competencies.

NOTE: This best practice provides an example concentrating on interactions between two road users only (such as a vehicle and a pedestrian) and excludes interactions with other road users that may be involved in an event.

Incorporating data from NDS should be done with caution, as it represents a mix of desirable and undesirable human driving behaviors, making it just one input among others to inform ADS safety performance standards. For example, imitating speeding behaviors would not be advisable. Comparisons to behaviors observed from NDS can provide an important starting point for behavioral evaluation but may not be the sole basis for setting reference values. Section 5 describes a methodology for setting reference values using NDS for behavioral evaluation.

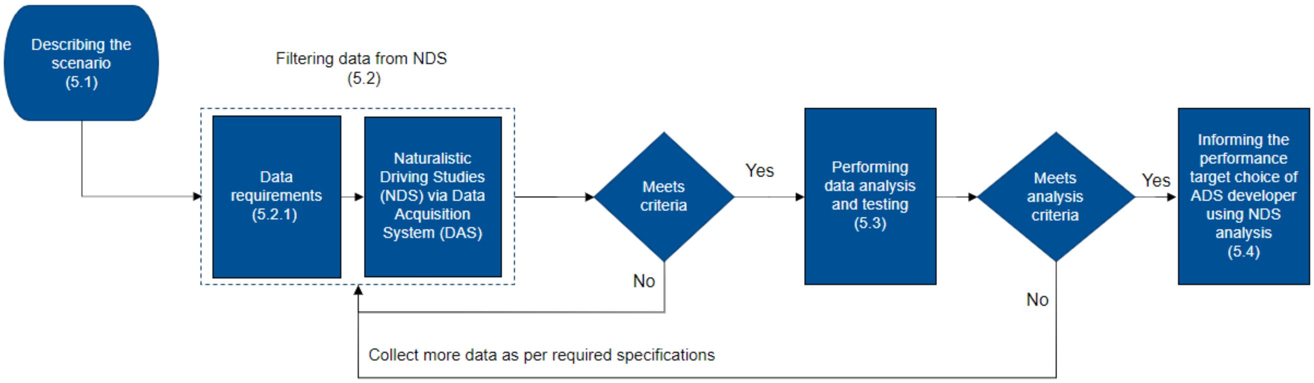
[2] "Context-relevant," in this document, is defined in terms of ODD and specific scenarios.

5. Methodology for Setting Safety Metric Reference Values Using Human Driving Data

This section provides a methodology for setting reference values using human driving data. A four-step, high-level methodology for setting metric reference values is shown in FIGURE 1 and includes:

- **Step 5.1**: Describing the scenario of interest based on a sought behavioral competency.
- **Step 5.2**: Filtering data from NDS.
- **Step 5.3**: Performing data analysis and testing.
- **Step 5.4**: Informing the performance target choice of ADS developer using NDS analysis.

FIGURE 1 Analysis framework



5.1 Describing the Scenario of Interest Based on a Sought Behavioral Competency

Scenarios provide context for ADS behavioral competencies which are constructed using terms that describe ODD conditions, object and event detection and response (OEDR), and maneuvers [AVSC00009202208]. Contextual ODD information, such as the environment, geography, time of-day restrictions, and specific traffic or roadway characteristics, should be considered when determining reference values. ADS developers should comprehensively define the scenario within the context of their ODD to ensure the validity of results and conclusions derived from the remaining steps in this process.[3]

ADS performance in various scenarios can be evaluated using a combination of behavioral competencies and metrics. AVSC00008202111 and AVSC00006202103 describe these, respectively.

[3] It is important to note that the intention of using behavioral competencies is to imply that certain levels of performance can be transferable to other ODDs. Therefore, an ADS developer can focus on defining a representative set of scenarios rather than an exhaustive list, striking a balance between comprehensiveness and practicality.

5.1.1 Example Scenario: Maintaining a Lane When Passing a Pedestrian

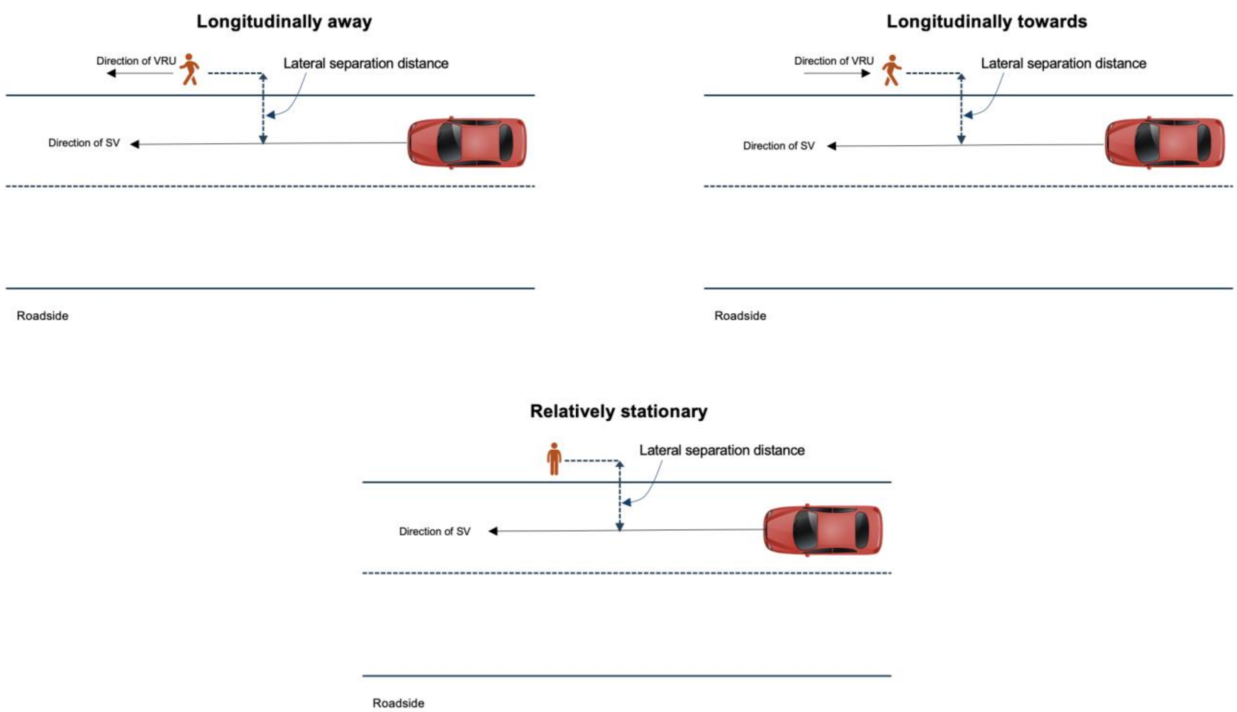
A specific safety performance metric and behavioral competency, as described in the scenario[4] below, is used to illustrate the principles of the framework outlined in Section 5. NDS data used for the example scenario was collected by a manually driven vehicle with various sensors and cameras to capture data on the driver's behavior, including acceleration, braking, turning, etc. These vehicles are referred to as subject vehicles (SVs) in the rest of the document.

Scenario (NDS Data)

SV maintaining a lane while passing a pedestrian where SV attempts to maintain proper lane position with respect to designated lane markings and speed limits while maintaining a safety envelope with the pedestrian in the form of lateral distance at the time of passing.

The goal of using this scenario for data analysis is to demonstrate a process to establish a baseline for minimum passing distance (i.e., lateral separation distance) for an ADS while passing a VRU such as a pedestrian on the roadside either moving longitudinally or staying relatively stationary, as shown in FIGURE 2.

FIGURE 2 Example scenario: maintaining a lane when passing a pedestrian



An ADS manufacturer can assess the safety impact of the ADS-DV relative to a human performance reference value. This scenario is defined in TABLE 1. This method can also be followed for other scenarios.

4 An ADS developer may choose to leverage existing standards to describe scenarios. For example, EuroNCAP (European New Car Assessment Programme) and other Advanced Driver Assistance System (ADAS) standards describe a set of standardized test scenarios and protocols to evaluate the safety performance of vehicles, including their interaction with VRUs, such as pedestrians and cyclists.

TABLE 1 Example of describing the scenario of interest based on a sought behavioral competency

SAFETY PERFORMANCE METRICS (from AVSC00006202103)		
Category	**Safety Performance Metrics**	**Description**
Maintain a safety envelope	Lateral distance	A violation of a kinematically defined state space around a vehicle that represents a buffer/lateral distance threshold between the subject vehicle and pedestrian in the environment
BEHAVIORAL COMPETENCIES (from AVSC00008202111)		
Category	**Behavior**	**Specification**
Roadway infrastructure	Maintaining a lane	Driving along roads predictably and consistently maintaining proper lane position with respect to designated lane markings and speed limits
Dynamic conditions	Responding to VRUs	Maintaining a safety envelope with respect to VRUs
OPERATIONAL DESIGN DOMAIN (ODD) CHARACTERISTICS (from AVSC00009202208)		
Considerations		**Specification**
Geographical		Urban
Roadway type		Local roads, arterials, and collectors[5]
Time of day		All
Weather conditions		All

5.2 Filtering Data from NDS

Human driver performance metrics, derived from NDS, can be used to establish reference values for ADS performance. However, all datasets have limitations and care should be taken to understand and account for them when using the data to derive reference values. These limitations and any assumptions should be documented for any NDS used for this purpose. Below is a sample list of data factors an ADS developer can consider while characterizing an NDS dataset:

- **Data type**: Video and dynamic performance (e.g., kinematic) data via a network of sensors distributed around the vehicle. AVSC0004202009 makes recommendations for the collection, storage, and retrievability of onboard motor vehicle ADS event data. It provides the data type and characteristics that may be useful in developing reference values for safety performance. The data collected should be representative of real-world scenarios and behaviors that the ADS is expected to encounter.[6]

- **Sensor suite on vehicle**: Real-time video compression; a multiplexed video channel permitting multiple video inputs; lane tracker; sound level meter; three-axis gyroscopes; three-axis accelerometers; and radar.

- **Data collection frequency**: Sufficient to capture underlying signal to calculate safety metrics.

- **Data sample size**: Sample size large enough to calculate confident safety metric distributions.

- **Fields of view**: Maximum area sensors/instruments can capture/observe.

- **Measurement error in data collection system**: Validate by comparing NDS with empirical data.

- **Locality of data collection**: Operating in various urban, rural, residential, nonresidential, etc.

---

5 These roadway classification labels are appropriate for this use because scenarios are generically described and may not necessarily be specific to one geographic location. Refer to AVSC00009202208 for a more-detailed discussion on describing an ODD.

6 An ADS developer might consider sub-selecting behaviors that are desired to be emulated or performed by the ADS, particularly if population being modeled is not representative of the target use case. Techniques like nominal model sub-selection could be used. However, this best practice does not cover principles of sub-selecting behaviors.

- **Road type**: e.g., arterials and collectors and local roads.
- **Safety performance metrics** (from AVSC00006202103): e.g., speed, jerk, distance, time to collision (TTC) (or another analogue for safety envelope), OEDR, reaction time.
- **Minimum segment length**: The video segment length that corresponds to the time interval of data capture.
- **Weather conditions**: Documented or observable weather conditions including rain, snow, etc.
- **Location of VRU**: Precise location of VRU on roadside or sidewalk.

5.2.1 Example Data Requirements

For the development of the example used in this best practice, the AVSC utilized real-world NDS from the Virginia Tech Transportation Institute (VTTI) Automated Mobility Partnership (AMP) Program. AMP has built a library of crashes, near-crashes[7], and driving cases[8] [9]. In TABLE 2, an illustrative example is provided to demonstrate the data requirements that an ADS developer may consider, including the high-level features of AMP data that are relevant to the specific scenario being analyzed.

TABLE 2 Example of data requirements for assessing performance in the described scenario

Data Factors	Example Requirements	AMP Data for the Chosen Scenario
Data type	Relevant saliency and sensing data as recommended by AVSC00004202009	Video and dynamic performance (i.e., kinematic) data via a network of sensors distributed around the vehicle
Data collection frequency	Data collection frequency (e.g., 10 Hz or higher) as recommended by AVSC00004202009	10 Hz
Data sample size	Sample size can be determined based on parameters set by developers (risk tolerance, standard deviation, confidence interval, performance of subsystems, and the hypothesized effect size)	77
Fields of view (FOV)	Preferably, unrestricted	Restricted—FOV restrictions leads to unavailability of exact position of VRU relative to the car while passing; however, VRU position can be estimated using basic kinematics
Measurement error in data collection system	Ground truth analysis by comparing NDS with empirical data	Not available
Locality	Operating in various urban, rural, residential, nonresidential, etc.	Urban
Road type	Arterials and collectors, local roads including its characteristics	Arterials and collectors, local roads
Safety performance metrics	Speed, jerk, distance, TTC (or another analogue for safety envelope), OEDR, reaction time	Speed, jerk, distance, TTC (or another analogue for safety envelope), OEDR, reaction time
Minimum segment length	20 seconds	20 seconds
Weather conditions	Any	Any

7 “Near crashes” refer to situations where a driver takes an evasive action, such as jerking or braking, to avoid a collision, but the actual collision is narrowly avoided.
8 “Driving cases,” in the AMP database, are segments of nominal driving under various conditions.

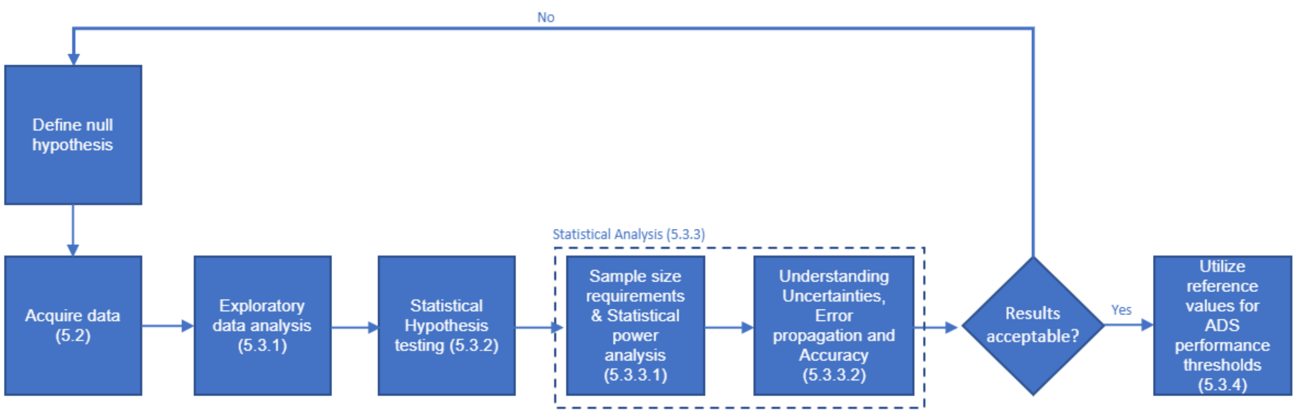
5.3 Performing Data Analysis and Testing

The following subsections provide a guideline for performing data analysis on an NDS. The initial step for an ADS developer is to establish a null hypothesis for targeted data collection and analysis. For instance, the null hypothesis for the selected scenario in which a pedestrian is walking within the lane of travel of the ADS-DV could be, "The lateral distance does not increase between the SV and pedestrian from the time of initial detection to the time of passing." Based on this initial hypothesis, an ADS manufacturer can opt to procure NDS scenarios from a third party, collect the data themselves, or filter data relevant to the scenario from a larger dataset. Regardless of the source of data, the analysis process should include the following steps, as illustrated in FIGURE 3.

- **Step 5.3.1**: Performing an exploratory data analysis to ensure that the data needed to test a hypothesis is available and applicable, as well as relevant to the ODD of interest. The scope of this data should be consistent with the scenario definition and ODD context from 5.1; otherwise, the data collection will need to be augmented to fill any gaps.
- **Step 5.3.2**: Execute the appropriate statistical hypothesis testing.
- **Step 5.3.3**: Statistical analysis.
  - **Step 5.3.3.1**: Sample size requirements and statistical power analysis.
  - **Step 5.3.3.2**: Understanding uncertainties, error propagation, and accuracy.
- **Step 5.3.4**: Utilize reference values for ADS performance thresholds.

If the result of testing is consistent with the desired accuracy and acceptable error[9], the results may be used as reference values for safety metrics in similar scenarios. If not, the ADS developer should acquire more data, test alternative hypotheses, and repeat 5.3.1 through 5.3.3.

FIGURE 3 Dataset analysis process



[9] Acceptable error or margin of error refers to the degree of uncertainty or inaccuracy that is considered acceptable in a measurement or analysis.

5.3.1 Exploratory Data Analysis

The goal of this step is to verify that the data required for the scenario is available within the dataset. Vehicle data collected in 5.2 should be filtered based on the scenario description. This step helps identify patterns and relationships in the data, which can guide the formulation of relevant hypotheses for testing.

AMP provided 77 samples of naturalistic driving data for the scenario defined in this best practice in 5.1.1. The samples are broken down based on the type of roadway and the direction of pedestrian movement as follows:

- Arterials and Collectors (46)
  - Pedestrian moving longitudinally[10]—21
  - Pedestrian relatively stationary[11]—25
- Local Roads (31)
  - Pedestrian moving longitudinally—25
  - Pedestrian relatively stationary—6

5.3.2 Hypothesis Testing

AVSC recommends hypothesis testing[12] and trend analyses on any driving dataset (including NDS) to establish data quality, accuracy, and an understanding of how errors can propagate during analysis. The following high-level steps can be followed by an ADS developer to define the hypothesis to be tested.[13]

- Develop a theoretical framework by drawing on existing knowledge of patterns and trends of human driving and decision-making. Throughout this best practice, the AVSC uses "passing a pedestrian" as an example scenario to create its theoretical framework.

- Formulate testable hypotheses that can be evaluated using empirical data. Hypotheses should be specific and measurable and should clearly articulate the relationship between variables. For example, a hypothesis relevant to this best practice: "The lateral distance does not increase between the SV and pedestrian from the time of initial detection to the time of passing."

- Define the variables and measures to test the hypothesis. This may involve identifying specific aspects of the phenomenon being studied that are of particular interest, such as driver behavior or ADS performance. For example, lateral distance, time of detection, and time of passing are the variables used in the above example hypothesis.

- Evaluate the results to determine if a hypothesis is supported. Supported hypotheses can validate the theoretical framework and lead to new research questions, while unsupported hypotheses may require revisions to the framework or consideration of new variables or measures.

---

10 Due to the restricted field of view of the sensor systems on the vehicles used in the AMP dataset, the precise position of the VRU relative to the vehicle at the moment of passing is unavailable. However, given the slow pace of movement of the pedestrian, the change in its position relative to the ADS-DV during the brief interval of passing is negligible.

11 "Relatively stationary" refers to the case where pedestrian is either not moving or moving at a negligently slow speed with respect to vehicle.

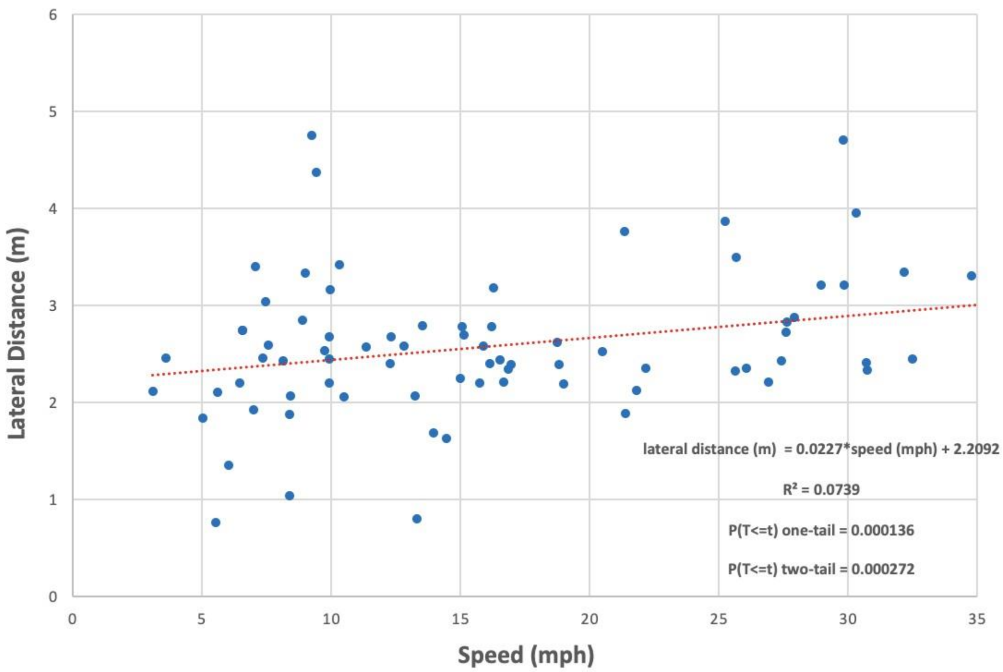
12 Hypothesis testing is one statistical approach for data analysis of NDS. ADS developers can utilize other approaches to analyze human data to inform ADS behavior. For example, NHTSA research work on testable cases highlights model-based predictive analytics approaches. These approaches involve the creation of mathematical models that simulate driving scenarios, enabling predictions of ADS behavior. By leveraging these models, ADS developers can assess and evaluate the performance of their systems in different testable cases. This broader range of analytical methods goes beyond hypothesis testing and provides additional tools for understanding and improving ADS behavior based on human data [7].

13 More information on defining hypotheses is available in [8].

Hypothesis testing as outlined in this section can be done for any scenario defined by ADS developers with an available dataset. For the example scenario chosen in this best practice, the following hypotheses were tested:

Hypothesis 1: Lateral distance between the subject vehicle (SV)[14] and pedestrian increases with increase in SV speed. As shown in FIGURE 4, the lateral distance between SV and pedestrian tend to increase[15] as the initial speed (velocity) recorded at time of detection of the SV increases to the point of passing the VRU.

FIGURE 4 Lateral distance increases with speed



The approach used in this analysis has limitations indicated by the low values of R-square and p-values. The low R-square value suggests that only a small proportion, approximately 7%, of the variation in lateral distance can be explained by changes in speed. Additionally, the low p-values indicate that the observed relationship between speed and lateral distance may not be statistically significant. To address these limitations, a larger and more diverse dataset may be required that includes a wider range of speeds and corresponding lateral distances.

Hypothesis 2: Lateral distance between the SV and pedestrian increases while the speed decreases from the time of initial detection to the time of passing. To test this hypothesis, flow diagrams were developed to visualize the SV movement from the initial point that the human driver of the SV detected the pedestrian on the roadside to the point of last detection before the pedestrian goes out of SV's FOV. FIGURE 5 shows the flow diagrams broken down by four possible combinations of lateral distance and speed recorded at the time of detection and time of passing.

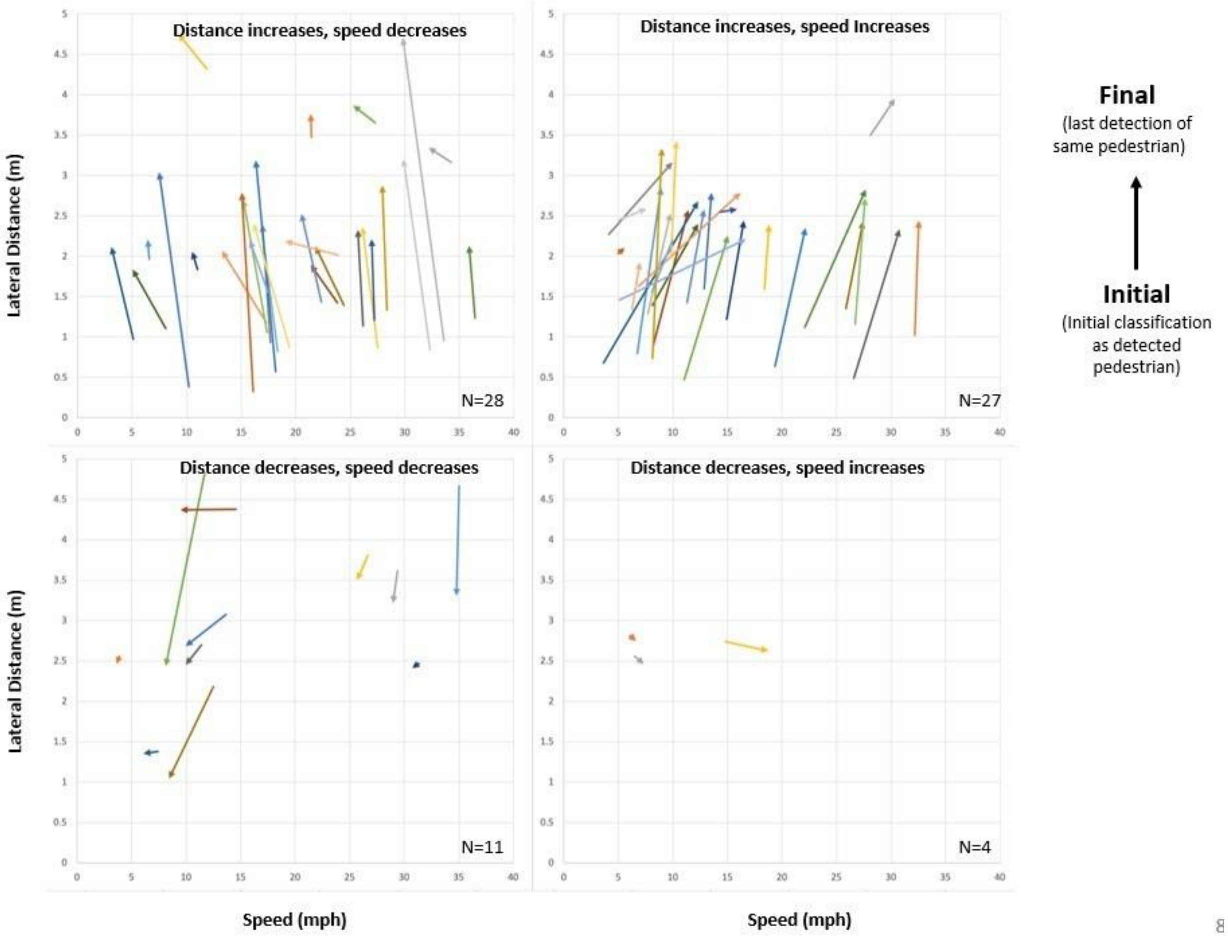
- Lateral distance increases and speed of SV decreases.
- Lateral distance and speed increase.
- Lateral distance and speed decrease.
- Lateral distance decreases and speed increases.

---

14 The term "subject vehicle" pertains to the vehicle operated by human drivers in the NDS dataset provided by AMP, within a presumed fixed infrastructure of arterial and local roads, such as a fixed road width.

15 While a linear regression is presented in this section, the selection of the best-fit model should be informed by the exploratory data analysis shown in the previous sections.

FIGURE 5 Lateral distance versus speed distribution



Visual inspection of FIGURE 5 shows, in most cases, the SV moved away from the pedestrian while passing, which supports the hypothesis. To analyze the SV driver's behavior when moving away from pedestrians, differences between their initial and final position and speed was calculated. In more than 50% of cases, the driver marginally reduced their speed in addition to moving away from the pedestrian. The proportion of cases where this occurred was determined through individual calculations of the differences and their comparison. These trends are consistent with the expected behavior of the SV and provided a validation of the NDS selected for this scenario.

Hypothesis 3: Lateral distance between the SV and pedestrian should be higher for cases when a pedestrian is walking while not facing the vehicle. There were three types of cases analyzed.

- **Case 1—Pedestrian moving longitudinally away**: This case involves a pedestrian who is moving away from the SV in the same direction the SV is traveling. This scenario can arise in various situations, such as when the pedestrian is walking along a roadside that runs parallel to the road where the SV is traveling. In this case, the pedestrian's back is facing the SV. An ADS developer should list the edge cases to consider whether they fall within the scope of their ODD. Some examples of edge cases include:

    - The pedestrian is not walking in a straight line but moving erratically, such as zigzagging or walking in circles.

    - The pedestrian is walking backwards while moving away from the SV, which means their front is facing the SV. This scenario can happen when the pedestrian is walking backwards to take a picture.

- **Case 2—Pedestrian moving longitudinally towards**: This case refers to a scenario in which the pedestrian is walking towards the subject vehicle in the same direction that the vehicle is traveling. The pedestrian's front is facing the SV. Some examples of edge cases include:
  - The VRU is moving towards the SV in the direction the SV is traveling but is walking backwards while doing it (the pedestrian's back is facing the SV[16]).
  - The VRU is moving towards the SV in a wheelchair or other assistive device, which may have different movement patterns than someone walking.
- **Case 3—Pedestrian relatively stationary**: This case refers to a scenario in which the pedestrian is stationary relative to the SV, meaning the pedestrian may be facing towards or away from the SV or may not be facing it at all.
  - The pedestrian is standing still but is leaning or bending over. This can hinder the system's ability to recognize them as a pedestrian and accurately assess their behavior.

APPENDIX B shows the flow diagram for this hypothesis along with the statistics. Although the mean lateral distance for Case 1 is higher than other cases, the sample size is not large enough to support a conclusion.

Hypothesis 4: Subject vehicle maintains a minimum passing distance to pedestrians required by state law. ADS manufacturers can use a one-sample t-test to compare the mean behavior of drivers in the NDS to a standard value. This allows the developer to determine whether the sample mean is different from a specific value [10]. This also demonstrates a method a developer can use when a specific distance is cited by law.

In the chosen scenario for this best practice, the specific value is the minimum passing distance that may be recommended by individual states (e.g., 3 feet, which can be approximated to 0.9 m) [11]. The t-value, t-table value [12], and p-value for this population were found to be 20, 1.65, and <0.05, respectively. The t-value does not fall into the range of t-table values. Therefore, the null hypothesis is rejected because the SV maintains *more* distance to pedestrians than the recommended minimum passing distance. To reflect the confidence in correctly rejecting the null hypothesis, the ADS developer can also calculate statistical power as explained in 5.3.3.1.

5.3.3 Statistical Analysis

5.3.3.1 Sample Size Requirements and Statistical Power Analysis

Determining the right sample size is a function of acceptable error, standard deviation, and confidence interval (z score[17]). The uncertainty in estimating a safety metric when actual values are not available (as with lateral distance in the example scenario) depends on the intrinsic variability of the measurements as well as the number of observations available. The impact of errors in estimating any variable can typically be reduced by increasing the sample size. ADS developers can decide on an appropriate sample size based on acceptable error (or risk tolerance of an ADS developer). For example:

- For standard deviation (SD) = 0.5 m, error = 0.05 m, z = 1.96, then sample size could be 385.
- For SD = 0.5 m, error = 0.1 m, z = 1.96, then sample size could be 96.

For the chosen scenario in this best practice, based on the available dataset of 77 samples, the observed mean and median for lateral distance between the SV and pedestrian was 2.58 m and 2.45 m, respectively. Additionally, 95% of the time drivers maintained more than 1.57 m distance from pedestrians while passing.

---

16 This is an example of an edge case but it highlights wide range of scenarios that can be used to derive reference values.

17 The example analysis presented in this section uses z-values since the population standard deviation is known and it is assumed that the entire population data follows a standard normal distribution. It is important to note that t-values can also be used in similar analyses, as described earlier in this document.

Statistical power analysis[18] is another effective method to determine a minimum sample size for analysis [13] [14]. With a given statistical power (commonly accepted lower threshold of 80% or higher), the significance level (commonly accepted lower threshold of 5%) and expected effect size (calculated to be 0.4 m based on differences of mean and SD), the sample size required for a statistically significant analysis could be approximately 150[19] (see APPENDIX C).

5.3.3.2 Understanding Uncertainties, Error Propagation, and Accuracy

Ensuring the accuracy of measurements and data points is crucial, especially when conducting perception-related analysis and addressing measurement errors. One effective approach for evaluating perception accuracy is through the analysis of ground truth data. This method involves comparing recorded values, such as position, speed, or acceleration, obtained from video measurements or other sensor data to the actual values. By assessing the consistency between recorded and actual pose values, the accuracy of various measurements can be established.

It is recommended to collect and compare data from multiple sensors or sensor modalities on a single platform, or independently collect measurements from other actors in the scene concurrently, to improve precision. Real-time kinetic (RTK) corrected satellite navigation systems are commonly used in research and testing to establish ground truth for localization [15]. Ground truth data was not available for the NDS used in this analysis. AMP researchers thoroughly reviewed the video footage for each available case and determined the values used represent reasonably accurate distance measurements both longitudinally and laterally.[20]

Measurement errors can have cascading effects throughout the entire pipeline, impacting perception, path planning, and vehicle motion control. Kinematic sensors can introduce lateral position errors ranging from 0.5 to 3.0 m [14]. For example, AMP reported a maximum error of 1 m in the lateral positioning data. In the absence of ground truth analysis, an ADS developer can employ the Monte Carlo method to estimate error propagation. This approach involves repeated calculations of mean lateral distance (or other relevant quantities), while randomly varying the expected error within the stated precision limits. Alternatively, Gaussian noise can be added to the raw lateral distance to approximate the error for the real-time kinematic sensor [17]. APPENDIX D demonstrates the application of the Monte Carlo method, illustrating the distribution of mean lateral distance and highlighting the effects of measurement data imprecision.

5.3.4 Utilize Reference Values for ADS Performance

Steps 5.3.1 through 5.3.3 describe the process of testing hypotheses associated with the NDS data to derive threshold performance for ADS in specific scenarios. The testing results should be compared to desired accuracy and acceptable error levels. The results can also serve as benchmarks or criteria for evaluating safety metrics in context-relevant scenarios. The hypothesis testing aims to identify whether there is a significant change in the lateral distance between the SV and the pedestrian during a passing maneuver. If the results meet the safety performance requirements set by the ADS developer, the lateral distance values obtained can be used as reference values for safe pedestrian passing in similar scenarios.

If the results do not meet the organization's requirements, the ADS developer should collect more data or test alternative hypotheses, repeating the testing steps outlined in 5.3.1 through 5.3.3, while adjusting the hypotheses and modifying the parameters. This iterative process should continue until the desired accuracy and acceptable error levels are met for the scenario being evaluated, ultimately leading to establishing safe lateral distance reference values for passing pedestrians.

---

[18] Statistical power, or sensitivity, is the likelihood of a significance test detecting an effect. Typically, 80% power (which means that the analysis has 80% confidence in correctly rejecting the null hypothesis) is the commonly accepted confidence threshold for null hypothesis rejection [16].

[19] The sample size of 77 available for this scenario is smaller than the suggested sample size of 150, which may result in reduced statistical power and accuracy. This example was used to illustrate the process, only and not intended to show validity of the data.

[20] The accuracy process and methodology utilized by AMP researchers were reviewed and deemed acceptable by the AVSC for the purpose of this example analysis. It is expected that AMP will include a detailed description of their methodology in the final documentation to provide transparency and ensure clarity regarding the accuracy assessment process.

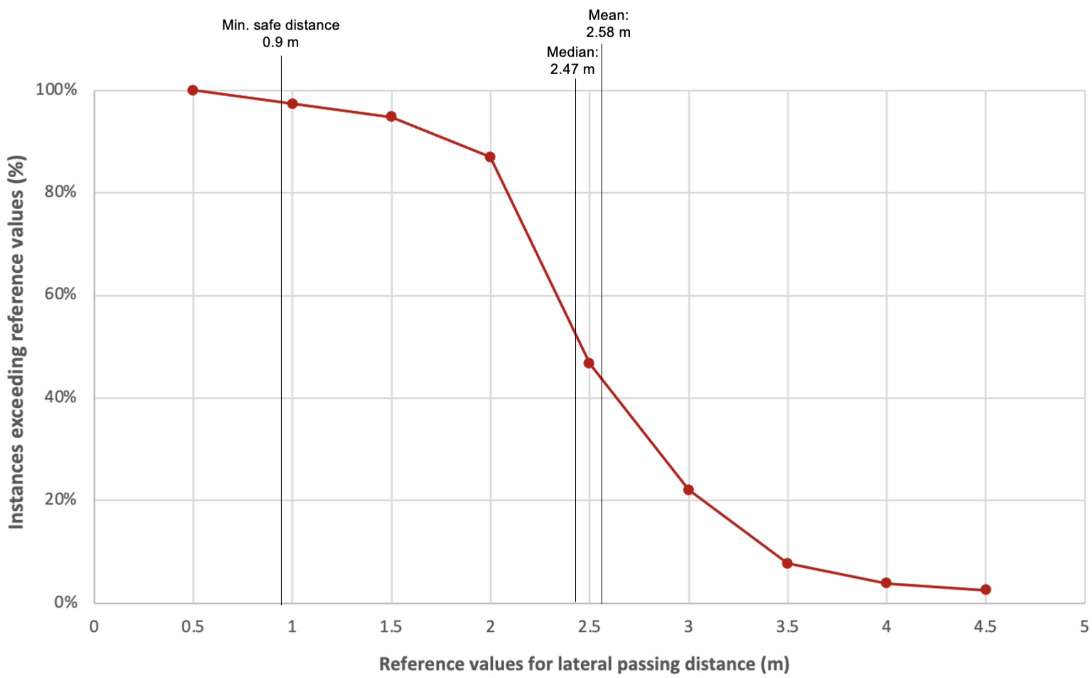
5.4 Informing the Performance Target Choice of ADS Developer Using NDS Analysis

ADS manufacturers have the option to establish more stringent thresholds than those required by traffic regulations. For instance, the mean lateral distance between SV and pedestrians during passing was determined to be 2.58 m, indicating that SVs maintain a distance greater than 2.58 m in 50% of instances. It should be noted that certain states may only mandate a passing distance of ~0.9 m (3 feet), meaning that violations of this distance in the NDS dataset represent just 3% of cases.

FIGURE 6 depicts a graph that illustrates reference value violations caused by human-driven vehicles. Assuming a minimum safe lateral distance of 0.9 m between SV and pedestrians, the NDS dataset suggests that 3% of cases present an opportunity for ADS developers to improve safety performance compared to humans. This is because ADS can be programmed to eliminate errors in judgment that human drivers may make.[21]

Once reference values are determined for specific scenarios of interest, ADS developers should engage in iterative testing and analysis to further refine these values. This iterative process can enable ADS developers to make design choices that align with the performance goals set by the reference values.

FIGURE 6 Percentage of instances exceeding example reference values



[21] While ADS present an opportunity for improving safety by eliminating human errors in judgment, they may also introduce new risks, such as software failures or sensor malfunctions, that need to be carefully managed and addressed. Such risks are out of the scope of this document.

6. Summary

This AVSC best practice provides tools for establishing quantitative performance reference values that can be integrated into a broader safety assurance framework for ADS. By using the framework and data analysis processes described in this best practice, ADS manufacturers and developers can build evidence and make a case for the safety of their system. The reference values established through this framework provide additional meaning to the safety performance metrics and behavioral competencies described in other AVSC best practices, allowing for a more comprehensive approach to evaluating the safety and performance of ADS.

The metrics and methods for assessing safety performance of ADS-DVs [AVSC00006202103] can be used in combination with the quantitative performance reference values to provide a more comprehensive picture of the system's safety. Similarly, the evaluation of behavioral competencies for ADS-DVs [AVSC00008202111] can be informed by the reference values established through the framework described in this best practice. These complementary best practices work together to ensure that the system design achieves the performance goals established by the reference values while assessing its behavioral competencies and safety performance using a range of metrics and methods.

Consistent with other AVSC best practices, this document supports industry-led, voluntary approaches in the standards development community and is expected to evolve as technology matures. Public agencies may use this best practice to better understand the safety posture of ADS deployments. In addition to the technical development community, other audiences considered in the development of this best practice include standards bodies, public agencies, and other decision-makers that may influence the deployment of ADS-DVs.

7. About Automated Vehicle Safety Consortium™

The objective of the Automated Vehicle Safety Consortium™ (AVSC) is to provide a safety framework around which automated vehicle technology can responsibly evolve in advance of the broad use of commercialized vehicles. The consortium will leverage the expertise of its current and future members and engage government and industry groups to establish best practices and provide stakeholders with ADS safety-related information. This technology-neutral content can form the foundation for key considerations for deploying SAE level 4 and level 5 automated vehicles on public roads.

AVSC Vision:
Public acceptance of SAE level 4 and level 5 automated driving systems as a safe and beneficial component of transportation through industry consensus.

AVSC Mission:
The mission of the Automated Vehicle Safety Consortium™ (AVSC) is to quickly establish safety principles, common terminology, and best safety practices, leading to standards to engender public confidence in the safe operation of SAE level 4 and level 5 light-duty passenger and cargo on-road vehicles ahead of their widespread deployment.

The AVSC will:

- Develop and prioritize a roadmap of pre-competitive topics;
- Establish working groups to address each of the topics;
- Engage the expertise of external stakeholders;
- Share output/information with the global community;
- Initially focus on fleet service applications.

8. Contact Information

To learn more about the Automated Vehicle Safety Consortium™, please visit https://avsc.sae-itc.org.

Contact: AVSCinfo@sae-itc.org.

9. Acknowledgements

The Automated Vehicle Safety Consortium™ would like to acknowledge the contributions of the member organizations during the development of this document:

Aurora Innovations, GM/Cruise, Honda, Lyft, Motional, and Volkswagen.

10. Abbreviations

ADAS	Advanced driver assistance system
ADS	Automated driving system
ADS-DV	Automated driving system-dedicated vehicle
AMP	Automated Mobility Partnership
AV	Autonomous vehicle
AVSC	Automated Vehicle Safety Consortium™
DDT	Dynamic driving task
FOV	Field of view
NDS	Naturalistic driving studies
NHTSA	National Highway Traffic Safety Administration
ODD	Operational design domain
OEDR	Object and event detection and response
RTK	Real-time kinetic
SAE ITC	SAE Industry Technologies Consortia®
SAE	Society of Automotive Engineers
SD	Standard deviation
SE	Standard error
SV	Subject vehicle
TTC	Time to collision

V&V	Verification and validation
VRU	Vulnerable road user
VTTI	Virginia Tech Transportation Institute

APPENDIX A. Quick Look

Establishing Reference Values for Safety Performance Metrics (Section 4)

- ADS developers and manufacturers should establish context-relevant reference values and their uncertainties for safety performance metrics.

- Reference values should be commonly understandable to all stakeholders involved in the development and deployment of ADS systems, including regulatory bodies, developers, manufacturers, and end users.

- Incorporating data from NDS should be done with caution as it represents a mix of desirable and undesirable human driving behaviors, making it just one input among others to inform ADS safety performance standards.

Methodology for Setting Safety Metric Reference Values Using Human Driving Data (Section 5)

- This section provides a methodology for setting reference values using human driving data which is organized into four main process steps:

  - Describing the scenario of interest based on a sought behavioral competency.

  - Filtering data from NDS.

  - Performing data analysis and testing.

  - Informing the performance target choice of ADS developer using NDS analysis.

Describing the Scenario of Interest Based on a Sought Behavioral Competency (5.1)

- Contextual ODD information, such as the environment, geography, time-of-day restrictions, and specific traffic or roadway characteristics, should be considered when determining reference values.

- ADS developers should comprehensively define the scenario within the context of their ODD to ensure the validity of results and conclusions derived from the remaining steps in this process.

Example Scenario: Maintaining a Lane When Passing a Pedestrian (5.1.1)

- A specific safety performance metric and behavioral competency is provided in an example scenario to illustrate the principles of the framework outlined in Section **5**.

Filtering Data from NDS (5.2)

- Human driver performance metrics, derived from NDS, can be used to establish reference values for ADS performance. However, all datasets have limitations, and care should be taken to understand and account for them when using the data to derive reference values. These limitations and any assumptions should be documented for any NDS used for this purpose.

Example Data Requirements (5.2.1)

- The AVSC utilized real-world NDS from the Virginia Tech Transportation Institute (VTTI) Automated Mobility Partnership (AMP) Program. In TABLE 2, an illustrative example is provided to demonstrate the data requirements that an ADS developer may consider, including the high-level features of AMP data that are relevant to the specific scenario being analyzed**.**

Performing Data Analysis and Testing (5.3)

- The initial step for an ADS developer is to establish a null hypothesis for targeted data collection and analysis.

- The analysis process should include the following steps:
  - Performing an exploratory data analysis to ensure that the data needed to test a hypothesis is available and applicable, as well as relevant to the ODD of interest. The scope of this data should be consistent with the scenario definition and ODD context from 5.1; otherwise, the data collection will need to be augmented to fill any gaps.
  - Execute the appropriate statistical hypothesis testing.
  - Statistical analysis:
    - Sample size requirements and statistical power analysis.
    - Understanding uncertainties, error propagation, and accuracy.
  - Utilize reference values for ADS performance thresholds.
  - If the result of testing is consistent with the desired accuracy and acceptable error, the results may be used as reference values for safety metrics in similar scenarios. If not, the ADS developer should acquire more data, test alternative hypotheses, and repeat **5.3.1** through **5.3.3**.

Exploratory Data Analysis (5.3.1)

- Vehicle data collected in **5.2** should be filtered based on the scenario description. This step helps identify patterns and relationships in the data, which can guide the formulation of relevant hypotheses for testing.

Hypothesis Testing (5.3.2)

- AVSC recommends hypothesis testing and trend analyses on any driving dataset (including NDS) to establish data quality, accuracy, and an understanding of how errors can propagate during analysis. High-level steps can be followed by an ADS developer to define the hypothesis to be tested.

Understanding Uncertainties, Error Propagation, and Accuracy (5.3.3.2)

- Ensuring the accuracy of measurements and data points is crucial, especially when conducting perception-related analyses and addressing measurement errors.

- It is recommended to collect and compare data from multiple sensors or sensor modalities on a single platform, or independently collect measurements from other actors in the scene concurrently, to improve precision. Real-time kinetic (RTK) corrected satellite navigation systems are commonly used in research and testing to establish ground truth for localization.

Utilize Reference Values for ADS Performance (5.3.4)

- Steps **5.3.1** through **5.3.3** describe the process of testing hypotheses associated with the NDS data to derive threshold performance for ADS in specific scenarios. The testing results should be compared to desired accuracy and acceptable error levels. The results can also serve as benchmarks or criteria for evaluating safety metrics in context-relevant scenarios.

- If the results do not meet the organization’s requirements, the ADS developer should collect more data or test alternative hypotheses, repeating the testing steps outlined in **5.3.1** through **5.3.3**, while adjusting the hypotheses and modifying the parameters. This iterative process should continue until the desired accuracy and acceptable error levels are met for the scenario being evaluated, ultimately leading to establishing safe lateral distance reference values for passing pedestrians.

Informing the Design Choice of ADS Developer Using NDS Analysis (5.4)

- Once reference values are determined for specific scenarios of interest, ADS developers should engage in iterative testing and analysis to further refine these values. This iterative process can enable ADS developers to make design choices that align with the performance goals set by the reference values.

APPENDIX B. Kinematic Graphs

FIGURE B1 shows the flow diagram broken down by pedestrian movement direction for cases where lateral distance increases as SV passes the pedestrian. For hypothesis 3, it is observed that lateral distance between the SV and pedestrian is higher for cases when pedestrian is walking longitudinally away. However, due to low sample size, confidence in this observation is low.

Case 1: Pedestrian Longitudinally Away

- Mean: 2.8 m
- Median: 2.4 m
- SD: 0.83
- SE: 0.17

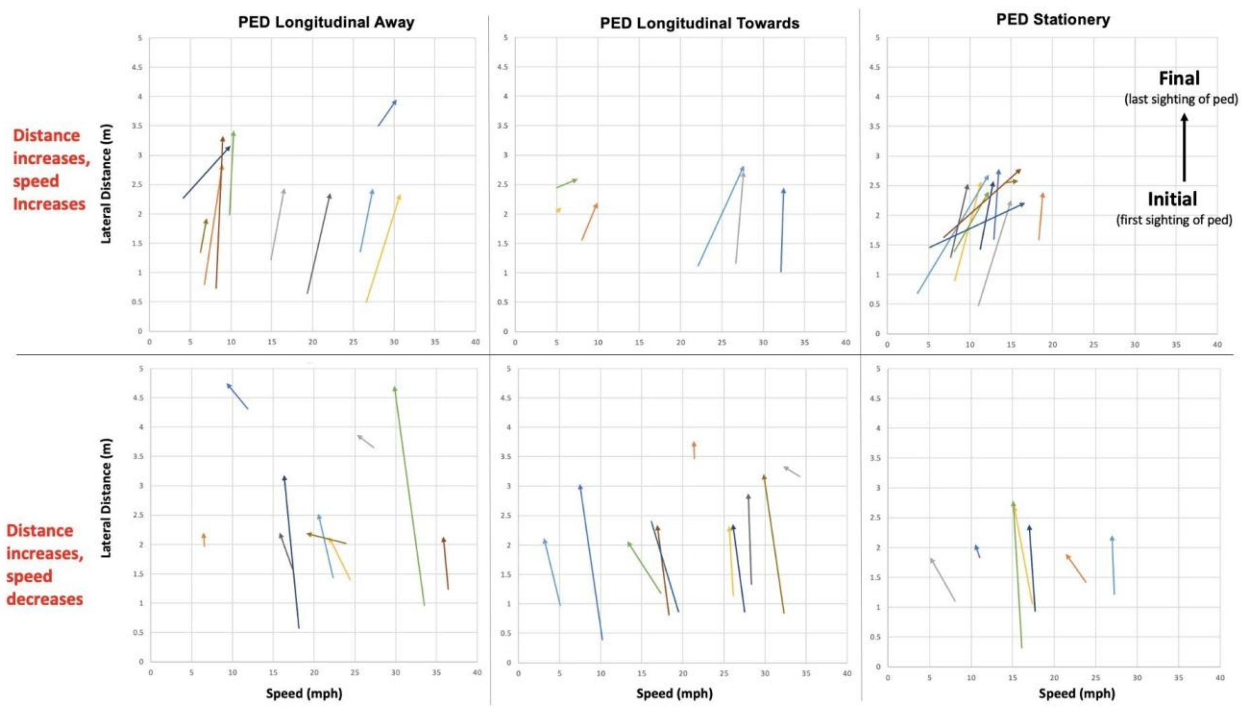
Case 2: Pedestrian Longitudinally Towards

- Mean: 2.65 m
- Median: 2.45 m
- SD: 0.77
- SE: 0.16

Case 3: Pedestrian Relatively Stationery

- Mean: 2.35 m
- Median: 2.45 m
- SD: 0.55
- SE: 0.11

FIGURE B1 Percentage of instances exceeding example reference values



APPENDIX C. Statistical Analysis

C.1 Standard Statistical Measures

The standard statistical measures relevant for:

- **Mean**: Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers. Mean = (sum of all the observations/total number of observations).
- **Median**: Middle value of the given list of data when arranged in an order. The arrangement of data or observations can be made either in ascending order or descending order.
- **Standard Deviation (SD)**: SD is a measure of variability. When we calculate the standard deviation of a sample of naturalistic driving data, we are using it as an estimate of the variability of the safety metric from which the sample was drawn.
- **Standard Error (SE)**: SE in the estimated mean is given by the sample standard deviation divided by the square root of the sample size: SE = SD/sqrt(n).
- **95% Percentile**: 95% of the time drivers maintain more than a certain distance from pedestrians while passing.
- **Statistical Power**: The likelihood that a test will detect an effect of a certain size if there is one.
  - **Type I error**: Rejecting the null hypothesis when it is true.
  - **Type II error**: Not rejecting the null hypothesis when it is false.

  Power is the probability of avoiding a Type II error. The higher the statistical power of a test, the lower the risk of making a Type II error.
- **Significance Level (Alpha)**: The maximum risk of rejecting a true null hypothesis that you are willing to take.
- **Expected Effect Size**: A standardized way of expressing the magnitude of the expected result of the study.

C.2 Determining Sample Size

- The 95% confidence interval (CI) corresponds to (approximately) 2x the error.
- For example, imagine that we have a sample of 200 observations, with a mean of 3 feet lateral distance and standard deviation of 0.5 feet. Then our 95% CI of the population mean would be given by 3 feet ± 0.069 feet.
- Standard deviation decides error.
- We can also determine the sample size needed to reach a specific level of precision, given an estimate of the standard deviation.
  - If we again assume a standard deviation of 0.5, and we want the error to be at most 0.02 (at 95% CI), then we'd need at least 2401 observations.

$$Sample\ Size\ = \frac{z^2\ x\ \ p(1-p)}{e^2}$$

where z = z-score; e = margin of error; and p = standard of deviation.

APPENDIX D. Error Propagation

Monte Carlo is an effective alternate method to understand error propagation in the overall data analysis. An ADS developer can perform repeat calculations of a safety metric, each time varying the expected error randomly within their stated limits of precision. In the chosen scenario in this best practice, errors in measurement of lateral distance were assumed to be following:

- 2 m error
- 1 m error
- 0.5 m error

The number of repeated calculations performed for each error type was limited to 1000 (typically, higher number of iterations/repetitions will produce more accurate distribution). It was observed that the impact of measurement error is largely marginal and the distribution of mean lateral distance value increases with increase in error.

FIGURE D1 Distribution of mean lateral distance