Creation of Pre-Crash Time-Series Data using Chinese In-Depth Crash Database from SHUFO

Junaid Shaikh, Bo Sui, Nils Lubbe

Abstract In 2021, 21% of globally estimated road traffic deaths occurred in China. The current study has three aims: first, to create a pre-crash time-series data (PCTSD) that can be used for assessing safety systems for China-specific crash scenarios; second, to understand the national representativeness of PCTSD; and third, to list recommendations to improve future PCTSD creation. The information (such as participant details, vehicle dynamics, environment details) from the Shanghai United Road Traffic Safety Scientific Research Centre (SHUFO) data during 2011–2021 was processed and converted to a PCTSD format. Of the 1,231 cases in SHUFO, 529 were converted. Car-to-powered two-wheeler and car-to-car are the top collision types, followed by single car crashes. Compared with national data, crashes in SHUFO and PCTSD at intersections are over-represented, and those occurring in the evening are under-represented. Recommendations include loading participants during reconstruction as per the participant's number in the dataset, accurate positioning of multibody while switching it from rigid body, extending the vehicle path backward during reconstruction, and deriving weight factors for each case. PCTSD data can be used to evaluate active safety systems, predict future crashes, and provide inputs to the active human body model simulations for China.

Keywords Crash database, PCM, pre-crash data, crash avoidance, driver assistance.

I. INTRODUCTION

Globally, 1.2 million people die on the road every year, and tens of millions more are injured or disabled in road traffic crashes [1]. This global concern must also be a priority for China: there were 62,218 reported road traffic deaths nationally in China in 2021, whereas the World Health Organization (WHO) estimates 248,099 deaths annually, representing 5% and 21% of global road traffic deaths, respectively [1]. Vulnerable road users (VRUs: pedestrians, motorcyclists and bicyclists) sustain the largest number of road traffic deaths in China [2-3].

The China New Car Assessment Program (C-NCAP) and China Insurance Automotive Safety Index (C-IASI) were introduced in 2006 and 2017, respectively. To reduce road traffic deaths, these assessment programs introduced physical testing methods to rate the safety systems: Automated Emergency Braking (AEB), Forward Collision Warning (FCW), and Lane Keep Assist (LKA). These systems are regarded as effective at preventing crashes and reducing road traffic injuries [4-6].

Several kinds of data can also be used to assess the effectiveness of driver assistance and crash-avoidance systems in virtual settings. Pre-crash data have been applied worldwide by many researchers in many studies [7-9]. The pre-crash matrix (PCM), based on the German In-Depth Accident Study (GIDAS), started in 2011 and was the first of its kind (pre-crash data in a database format) [10]. The basic PCM data consist of five tables (global, participant, dynamics, traffic, objects); in more recent versions, space-saving object libraries are used instead of individual tables [11]. As the importance of pre-crash data is increasingly understood, various in-depth databases have recently created PCM-like data: the Initiative for the Global Harmonization of Accident Data (IGLAD, known as IGLAD PCM) [12] and the Road Accident Sampling System India (RASSI, known as RASSI–PCTSD) [13].

Supporting the sustainable development goal (SDG) *Target 3.6 Halve the number of global deaths and injuries from road traffic accidents* in China calls for a better understanding of its traffic situation. Thus it is important to have data providing pre-crash information about crashes. There are several in-depth crash datasets available in China, such as the China In-Depth Accident Study (CIDAS) and data from the Shanghai United Road Traffic Safety Scientific Research Centre (SHUFO). The current work focusses on SHUFO data, obtained from crashes in the Jiading district of Shanghai city [14]. The database includes information from the three main crash phases: pre-

J. Shaikh (email: junaid.shaikh@autoliv.com; tel: +91-9821617552) is a Project Lead Traffic Safety Research Engineer with Autoliv India Pvt. Ltd, Bangalore, India. B. Sui was a Senior Traffic Safety Research Engineer with Autoliv (Shanghai) Vehicle Safety System Technical Center Co., Ltd., China. N. Lubbe is Director of Research with Autoliv Development AB, Sweden.

crash; in-crash; and post-crash. However, the pre-crash information is not available in a time-series format. This study has three objectives: first, to develop a method to convert SHUFO cases into time-series data according to the previous method defined by Shaikh and Sander [13]; second, to assess the quality and national representativeness of the resulting pre-crash time-series data (PCTSD); and third, to provide recommendations to improve the quality and quantity of the PCTSD cases in future.

II. METHODS

The methods are presented in three sections. First, a brief section describes the SHUFO in-depth crash data. The second section describes in detail how the pre-crash time-series data were created. The third section describes the created data's quality assessment.

In-depth crash data

SHUFO has been collecting in-depth crash data since 2005 [14]. They are collected retrospectively by an investigation team operating mostly in Jiading, a district of Shanghai city which covers an area of 463 sq. km with a population of 1.6 million inhabitants. Crashes involving passenger vehicles which meet one of the following criteria are investigated: (a) at least one of the involved persons has a serious or fatal injury; or (b) at least one airbag was deployed; or (c) the total damage cost more than 3,500 USD [14]. Annually, 150–200 cases are collected and coded in the dataset. The information includes a general crash description, vehicle specifications and deformations, personal details and injuries, and details about the surrounding infrastructure [15]. In addition, each case contains a scaled crash-scene sketch (in DWG format) that includes traffic lines, zebra crossings, traffic lights, skid marks and debris. For cases with sufficient detail, a reconstruction file is created which contains delta-v, impact speeds and other relevant information related to the participants. Reconstruction of a crash is performed to estimate the crash characteristics and post-crash kinematics of VRUs using a multibody system, which is a complex group of virtual rigid bodies (connected through different forms of joints) that can mimic VRU kinematics, which a single rigid body cannot do. However, VRUs are often loaded as single rigid bodies during the pre- and in-crash phases and then switched to a multibody during the post-crash phase in order to understand the post-crash kinematics. This switch can disrupt the flow of events, which is important when creating a time series of the pre-crash phase (as explained below).

SHUFO data from 2011 to 2021 were queried for all cases with a reconstruction file. This resulted in 1,231 cases (an additional 11 cases were missed at the time, due to inappropriate folder structure). Primary parameters were extracted from each case: case numbers, participant numbers, participant type and weight, pre-crash movement, road surface type, and road condition. The coefficient of friction was estimated based on the combination of the road surface type and road condition (Appendix A). The pre-crash movement parameter was transformed to a Boolean data type: True – moving backwards, and False – moving forward, stationary, making a U-turn, turning, lane changing, and others or unknown. The primary parameters were used for the pre-crash time-series data creation, which is explained in the next sub-section.

The national data were compared with both the SHUFO and the newly created PCTSD data to understand the biases in the sample. Future analyses adjusted to account for these biases will provide results that are more representative of China than the original sample. The Road Traffic Accidents Annual Report 2017, released by the Traffic Management Research Institute of the Ministry of Public Security, was used to understand the national distribution (population) [16]. The latest report available for public usage was for the year 2017. A set of secondary parameters that were in common between national and SHUFO data – the hour of the crash, the weather during the crash, and the road type on which the crash had happened – were extracted from the national data.

Pre-crash time-series data

The cases with reconstruction files passed through four phases: data extraction; data processing; data extrapolation; and pre-crash time-series data creation. These four phases are explained in detail and an illustrative flowchart is shown in Fig. 1.

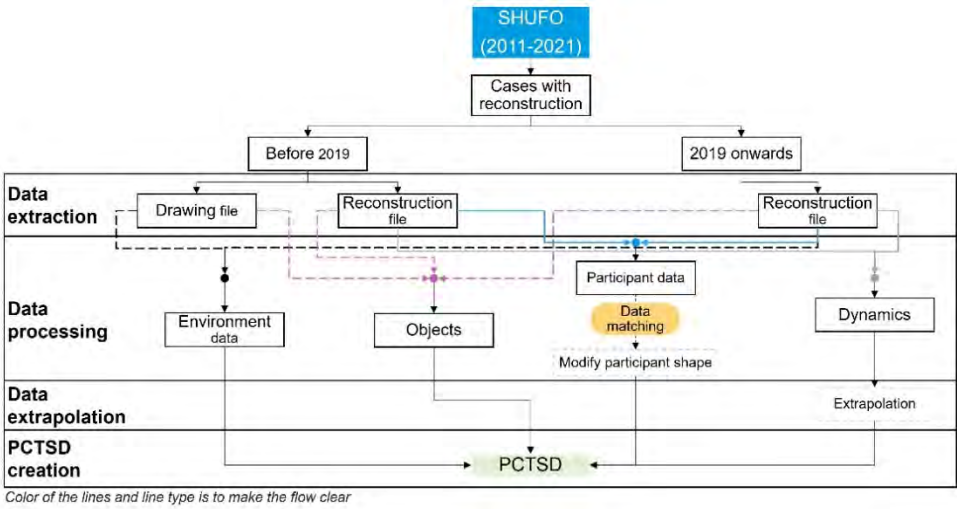


Fig. 1. Flowchart of the phases in the creation of pre-crash time-series data.

1) Data extraction

Each PCTSD case consists of a reconstruction file and a drawing file. The drawing file contains the crash scene (environment) information: road dimensions, positions of skid marks, point of impact, final rest position, road furniture, road signage, and debris of participants. The reconstruction file contains four crucial elements: (1) participant details: type, height, weight, length, centre of gravity, the moment of inertia, and other inter-related information; (2) dynamics history: time, position, velocity, and accelerations; (3) sequence information: collision time and collision partner details; and (4) additional environment information: the size and position of objects that were not included in the drawing file. For cases from 2019 onwards, the drawing exchange format (DXF) file was loaded directly into the reconstruction file. However, for cases before 2019, the environment information could only be retrieved from the drawing file. Because the file had been loaded as a bitmap file during the reconstruction, positional details (coordinates) were lost and could not be used for retrieving the environment information. The three steps of the data extraction process are described below.

a) Extract details from the reconstruction file

The time step for generating dynamics data from the reconstruction file was set to one millisecond, ensuring that the dynamics data retrieved were in the one millisecond interval and uniform for all the cases. For cases from 2019 onwards, the scene information in the reconstruction file was converted to the DXF. For cases before 2019, the scene information from the DWG file was used. The participant details were exported to the extensible markup language (XML) format. Further, participant dynamics were converted to a text file, which was later converted to a comma-separated-value (CSV) format for easy usage in the data processing phase. For each multibody present (if any), the position, velocity and heading angle were extracted for each time step up to the collision point. The multibody dynamics and the initially extracted participant dynamics were merged into a single file for ease of use. The sequence and report files for each case, converted to CSV format, were used to retrieve collision time and appropriate collision partner information. The complete extraction process was automated using AutoIT software [17].

b) Extract details from the drawing file

At the end of the previous step, each case included either a DWG file (cases before 2019) or a DXF file (cases from 2019 onwards). These files were converted to a CSV format using CAD2Shape software [18], in line with the earlier work [13]. The data were processed as environment or object data (explained in the later section).

c) Offset calculation

Offset calculation was an essential step for cases before 2019. Scene information from cases before 2019 was retrieved from the drawing file separately (not from the reconstruction file). In these cases, data from the reconstruction (dynamics and additional environment information) and the drawing file (scene information) are offset. In 2019, SHUFO started using drawing files directly in the reconstruction file, so there was no longer any offset between the positions of the participants and the environment information, eliminating the need for an offset calculation. To map the offsets for cases before 2019, the offset values were calculated in two steps: first,

marking a point on the scene diagram in the drawing file and identifying the same point in the reconstruction file; and secondly, using the difference between the coordinates of the two marked points to provide the offset value for that case.

2) Data processing

The data extracted in the previous phase were not directly usable for PCTSD creation and required processing. The collision time was recorded in the report file. If the file did not have a collision time, then the collision time was assumed to be 0 seconds. This happens when a multibody is involved in the first collision, a vehicle is involved in a rollover event, or the details of the dynamics were unavailable due to the file not being locked appropriately before delivery. These cases were reported to the database team. Further, the cases with negative collision time, indicating insufficient pre-crash simulation time, were discarded while the cases with positive collision time were passed to the next stage. The later stages were in line with the earlier work [13]; detailed steps are shown in Fig. 2. However, the data-matching process differed: in the earlier work [13], participants were matched by vehicle make and model, and pedestrians were counted directly. Since SHUFO datasets are created in the Chinese language, converting the make and model from Chinese to English in the current work and then matching them with the make and model defined in the reconstruction file was not feasible.

The original participant numbers could not be used for matching either, since the numbers in the reconstruction file were assigned in the order they were loaded into the simulation environment; they could be different from the participant numbers in the database. As a result, the participants were matched using the weights recorded in the dataset and the reconstruction file. Ideally, these weights are the same. In practice, however, it was not true, therefore a tolerance of ±30 kg was provided. If the weight of a participant in the reconstruction file was within this tolerance for the same case in the dataset, then the case was accepted and recorded in the passed case list. The participant number in the reconstruction file was then changed to the participant number in the dataset. If the weights were not within 30 kg, the case was moved to the dropped case list. With this tolerance, 54% of the non-matching files were accepted, yielding enough cases for the current work. Increasing the tolerance would have provided more samples but would have complicated the participant matching process, since the probability of multiple participants within the same case having weights within the increased tolerance would increase. Data matching was not performed for participants that were loaded as multibodies. Instead, these cases were recorded directly in the passed case list with pre-defined values, as the multibody has a complex geometry which made it difficult to retrieve participant's weights. Further, in the cases with multibody and multiple vehicles, only the first vehicle was selected to avoid selecting the wrong vehicle. In the cases when the collision time was unavailable, it was impossible to match the data. The complete process is illustrated in Fig. 2.

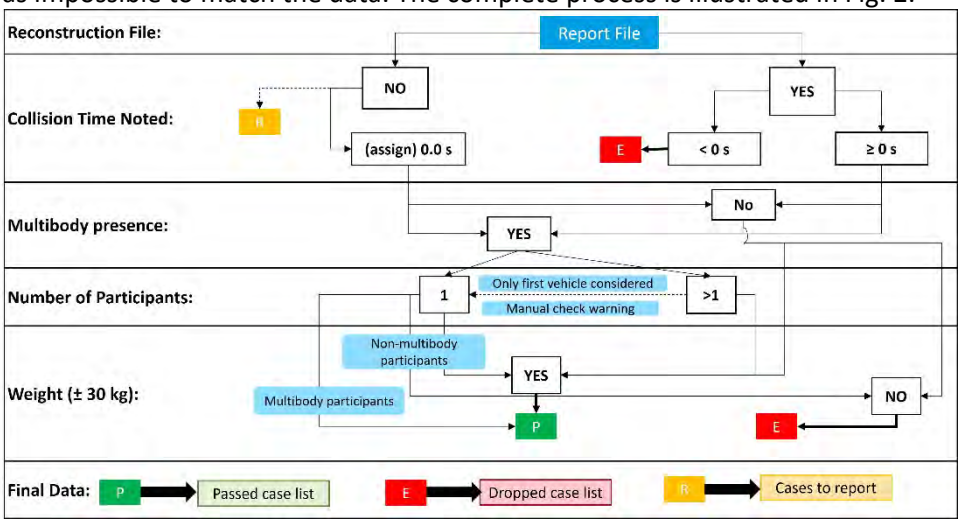


Fig. 2. Data-matching process using the report file from the reconstruction file. The "P" category represents passed cases considered for further processing, the "E" category represents the dropped (excluded) case list, and the "R" category represents cases reported to the database team, which were considered for conversion after appropriate changes.

The cases that passed the data-matching process were considered for record creation. The records in the current PCTSD version are in line with the earlier work (RASSI–PCTSD) [13]. PCTSD has six records: "global data",

"participant data", "dynamics data", "objects data", "environment data" (as available in a PCM), and "dropped-cases data" (not available in a PCM).

a) Global data

Global data is the simplest of all records, with only two variables: case number (FALL); and the total number of participants (PARTICIP; see sample in Table BI of Appendix B). In the previous work, the PARTICIP variable included only those participants involved in the first collision [13]. However, because more information could be useful for traffic simulation applications in the future, the current work records all participants in the crash, not just those in the first event.

b) Participant data

The participant data were created using participant details obtained from each participant involved in the first collision event. If the participant was involved only in the later events, then it was passed to the object data, where it was processed further. If the participants in the first collision event were loaded as a multibody, then the participant details were not available and pre-defined values were assigned (Appendix C).

The reconstruction file was limited to these specific participant types: car, motorcycle, truck, tram, pedestrian, occupant, tree, wall, trailer. In the SHUFO dataset, however, the participant types were: car, powered two-wheeler (motorcycle or e-bike), bicycle, pedestrian, truck, bus, powered three-wheeler, trailer. Due to the limited participant types in the reconstruction tool itself, often the researchers grouped powered two-wheelers and bicycles as motorcycles, buses as trucks, and powered three-wheelers as either cars, trucks or motorcycles in the reconstruction environment. The participant type in the reconstruction file was segregated and matched to the participant type in the dataset.

Although the car in the reconstruction file is usually represented as a rectangle (top view), cars have curved bumpers, which are represented by bevelled corners [19]. To account for bevelled corners in the front bumper, a fixed-width ratio of 0.6 is defined in the participant data. Further, two-wheelers in the reconstruction file are also rectangles; but in real life, when viewed from above, they appear more like rhombi [19]. To account for this shape difference ratios of the distance from the handlebar to the leading edge were introduced: 0.4 for powered two-wheelers and 0.27 for bicycles. The processed participant data consist of 18 variables, described in Table BII, Appendix B.

c) Dynamics data

The dynamics data were created from the details of the dynamics in the reconstruction file. The steps were the same as those in the earlier work [13], except for offset accounting. The participant's position in the dynamics data was translated to the new position as per the offset values (only for pre-2019 cases). The parameters of the dynamics data are shown in Table BIII, Appendix B.

d) Objects data

Objects, such as walls, trees, trailers and stationary participants from the reconstruction file were considered for the object data [13]. The participants that were in motion and not involved in the first event were ignored. Similar to dynamics data, the positions of objects were translated to the new positions after accounting for the offset values. SHUFO has a detailed sketch illustrating objects such as median barriers, trees, poles, etc., appropriately classified in the objects layer in the drawing file. The details of all the objects from the object layers were retrieved from the converted CSV file as well. The objects data from the reconstruction file and the CSV file were combined for each case. The parameters of the objects data are shown in Table BIV, Appendix B.

e) Environment data

The drawing data in the CSV format were used to create the environment data, as in the earlier work [13]. In SHUFO, the environment data has appropriately defined layers; for example, lines related to roadside edges were defined in the "roadside edge" layer, which simplifies the data processing for environment data. The layers in the data were assigned appropriate object types, such as roadside, continuous line, marks interrupted, etc. However, for some cases, the researchers had numbered the layers instead of using the name. For example, the lines related to roadside edge were assigned to the layer "101" and not to the "roadside edge" layer. For such cases, actual layer numbers were copied to the object types. Table BIV provides the list of object types along with the possible codes that are observed in the PCTSD.

f) Dropped-cases data

The cases that were dropped during data extraction, data processing, or later (during data extrapolation and PCTSD creation) were recorded here, along with the reasons for dropping them.

3) Data extrapolation

Two extrapolations were performed on the dynamics data: backward and forward. Backward extrapolation was performed to have sufficient pre-crash simulation time. Any case with a pre-crash simulation time less than 4.9 seconds underwent backward extrapolation. This threshold was defined by the author of the current work. For backward extrapolation, there were two assumptions:

- the velocity will remain that of the first available time step (furthest away from the collision) in the pre-crash phase without backward extrapolation;
- and the participant will continue backwards on a straight path based on the yaw angle at the first available time step in the pre-crash phase.

The new positions are calculated using Equation 1, where v_0 is the resultant velocity of the participant in the local coordinate system at the first available time step, φ_0 is the yaw angle of the participant at the first available time step, x_0 and y_0 are the x- and y-coordinates of the participant at the first available time step, t is the time step, and x_t, y_t are the x- and y- coordinates of the participant at the t^{th} time step. If the participant was reversing, then the right side of Equation 1 was multiplied by a negative sign.

$$\{x_t, y_t\} = v_0 * -t * \{\cos(\varphi_0), \sin(\varphi_0)\} + \{x_0, y_0\} \quad (1)$$

Forward extrapolation was performed when participants did not collide at the final time step (pre-crash simulation time of 0 seconds). This occurrence was due to the switching from rigid body to multibody in SHUFO. Multibodies were often introduced at the in- or post-crash phase: participants were loaded as a rigid body for the pre-crash phase, and at the collision point switched to a multibody. In practice, however, the switch between a rigid body and a multibody is not perfect. The multibody is approximately positioned and does not necessarily overlap with the rigid body coordinates at the collision point. Due to this approximate placement, some data (such as dynamics and trajectories) are lost. In such instances, as a result of the current data processing, only the rigid body is considered and the multibody is discarded (the participant sequence number of the rigid body precedes that of the multibody). As there are no dynamics data for the rigid body during the collision phase, the participants do not collide. Also, a change in the participant shape was introduced while the participant data were created, which caused some participants to miss the collision point. There were three assumptions in the forward extrapolation:

- the velocity will remain that of the final available time step (without forward extrapolation);
- the participant will go straight if the yaw rate was zero, left if the yaw rate was greater than zero, and right if the yaw rate was less than zero;
- the change in the yaw angle for the future path was maintained as that of the last two steps in the pre-crash phase.

First, each case was checked to determine if the participants collided at the final simulation step by matching the body profiles of the participants. If a collision was not detected, then the case was marked and processed for forward extrapolation. This check was performed only on cases with two or more participants. Secondly, based on the yaw rate, participants were classified as going straight or turning. In the case of a straight-moving participant, the new positions were calculated using Equation 2, where v_f is the resultant velocity of the participant in the local coordinate system at the final simulation step, φ_f is the yaw angle of the participant at the final simulation step, x_f and y_f are the x- and y-coordinates of the participant at the final simulation step, t is the simulation step, and x_t and y_t are the x- and y- coordinates of the participant at the t^{th} simulation step. If the participant was reversing, then the right side of Equation 2 was given a negative sign.

$$\{x_t, y_t\} = v_f * t * \{\cos(\varphi_f), \sin(\varphi_f)\} + \{x_f, y_f\} \quad (2)$$

If the participant was turning, then the radius of curvature (ρ) was calculated using Equation 3, where v_f is the resultant velocity at the final simulation step and $\dot{\omega}$ is the yaw rate.

$$\rho = \frac{v_f}{\dot{\omega}} \quad (3)$$

Finally, the new x- and y- coordinates of the participant (x_t, y_t) during turning were calculated using Equation 4, where pol2cart is a function to convert the polar coordinate system to a cartesian coordinate system using the radius of curvature (ρ) and yaw angle (φ), and x_f and y_f are the x- and y- coordinates of the participant at the final simulation step, respectively. The values were added when turning right and subtracted when turning left.

$$\{x_t, y_t\} = \{x_f, y_f\} \pm pol2cart\{\rho, \varphi\} \quad (4)$$

The forward extrapolation was performed for the next 100 time steps. Each extrapolated simulation step was checked for a collision. When a collision was detected, that simulation step was marked and labelled as the "new simulation step". The time step between the final simulation step and the new simulation step was added in the dynamics data. The added data was marked to differentiate between the actual reconstructed and extrapolated information in the dynamics data. If, even after iterating over 100 extrapolated time steps, the collision was not detected, then the case was dropped. It was the authors' choice to extrapolate for 100 time steps based on iterations performed during the work. One hundred steps appear sufficient as increasing the number of time steps to "1,000" did not affect the results.

4) Pre-crash time-series data creation

The final step of the PCTSD creation consisted of checking the data types of each column within the database, removing any extraneous text characters in the case numbers, and converting the text-based case numbers to numeric. The details of data types for each column are provided in Appendix (Tables BI–BIV). The case numbers were six digits: the first two indicate the year of the crash, making each case distinguishable from crashes in different years. The data were then combined into one database file. The complete process was performed in R [20].

Quality assessment

A quality check was performed on the created PCTSD in three steps. First, the participant's type (according to the participant number) was checked to ensure that the participant type and other relevant details assigned to the participant number were the same for both the PCTSD and SHUFO data. Secondly, the cumulative distribution of pre-crash simulation times for the dynamics data was plotted (with and without backward extrapolation). Ideally, no case has less than 4.9 seconds of pre-crash simulation time. Lastly, the yaw angles of all participants in all cases were checked for any irregularities across the complete pre-crash simulation. There are two possible reasons for an irregularity: (1) the participant did not follow a smooth path profile during the reconstruction; or (2) the participant changed velocity direction and heading angle in the pre-crash phase. These yaw angle irregularities would affect the simulations in future work by causing inappropriate path predictions. The yaw angle for each participant was checked by calculating the difference in yaw angles across the complete series of available time steps. The minimum and maximum of the 2.5th and 97.5th percentiles, respectively, were calculated across all participants and considered threshold values. Since the irregularities can be handled by smoothening the yaw angle profile, the cases with yaw angles beyond the threshold were not excluded.

III. RESULTS

Pre-crash time-series data

Of 1,231 applicable cases, 529 were converted to PCTSD format and the remaining 702 cases were dropped. The collision partners in the converted cases were most frequently car-to-powered two-wheeler (27%), followed by car-to-car (25%), and car (single vehicle, 17%); see Fig. 3.

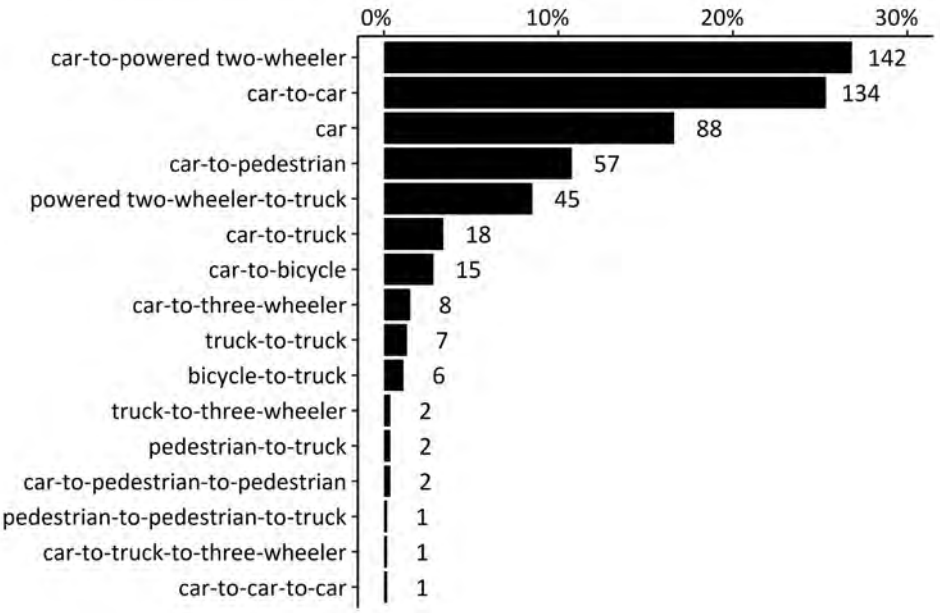


Fig. 3. Distribution of collision partners in the PCTSD cases (N=529).

The conversion rate is higher for more recent years (Fig. 4), from around 3% for 2011 to as high as 53% for 2021. The conversion rate can be further increased by analysing and treating the dropped cases. The reasons for dropping the 702 cases are presented in the Pareto chart. The bar indicates the presence of an individual reason while the line represents the share of the combination of reasons. The two major reasons for dropping cases (which together cover 76% of dropped cases) are: a difference in participant's weights between the reconstruction file and the dataset; and a lack of dynamics data.

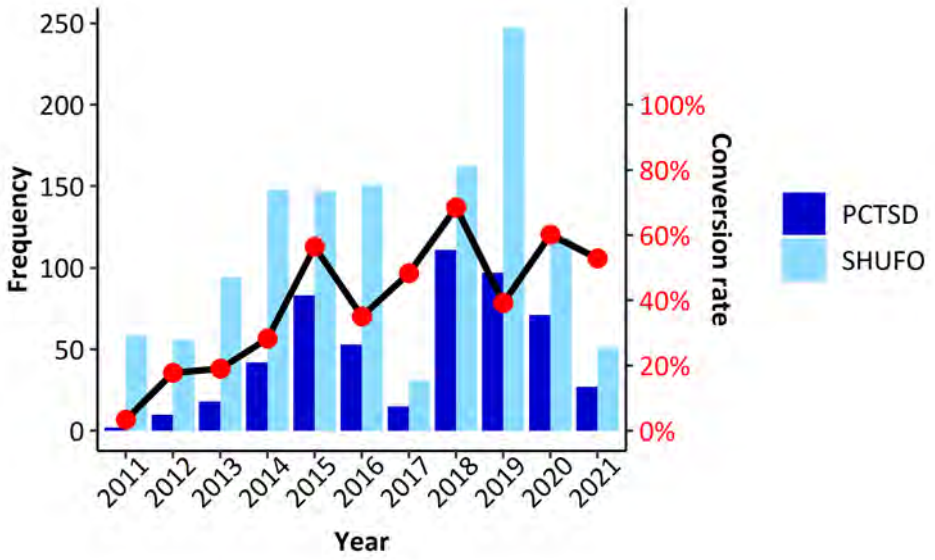


Fig. 4. Distribution of SHUFO reconstructed cases and PCTSD converted cases (left y-axis), and the PCTSD conversion rate (right y-axis) over ten years.

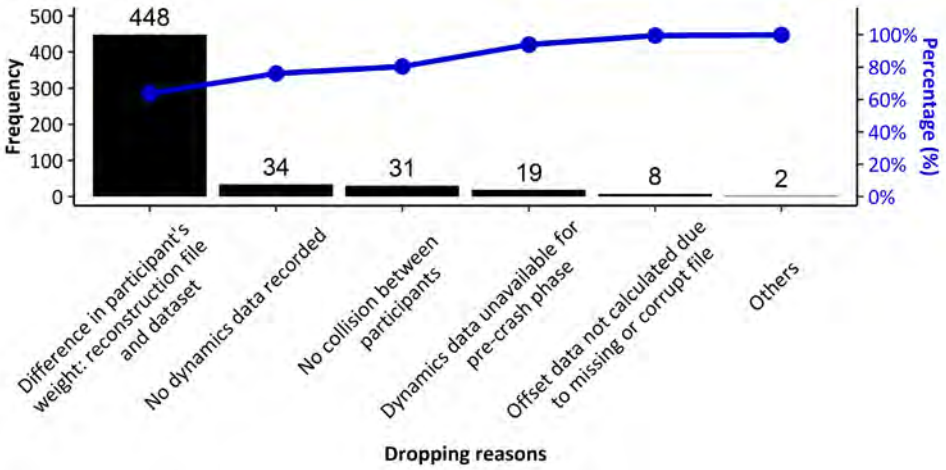


Fig. 5. Distribution of individual (bars) and cumulative (line) reasons for dropping the 702 cases in the PCTSD.

Quality assessment

The first quality check ascertained that the participant type for each case was appropriately matched between the PCTSD and SHUFO datasets against the participant number. No discrepancy was observed, indicating perfect matches. The second check monitored the quality of the backward extrapolation performed on the PCTSD data. Before extrapolation, for approximately 28% of the cases the pre-crash simulation time was less than 4.9 seconds. After backward extrapolation, no cases had a pre-crash simulation time of less than 4.9 seconds (Fig. 6). Thus, the current PCTSD version clearly provides ample pre-crash simulation time to assess any driver assistance or crash-avoidance safety systems.

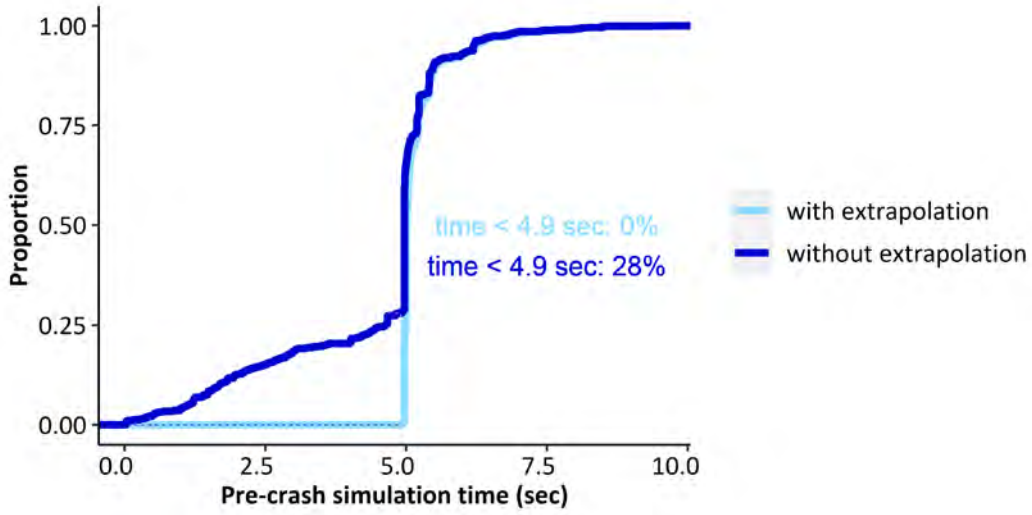


Fig. 6. Cumulative distribution of pre-crash simulation time for all 529 PCTSD cases with (blue line) and without (dark blue line) backward extrapolation.

Finally, the data quality of the dynamics data was assessed by observing the yaw angle irregularities. The lower and upper threshold values for the current PCTSD data were -0.01 radians and 0.02 radians, respectively. There were 58 cases (11%) in which the change in yaw angle of one or both participant(s) was beyond the threshold value. (Appendix D provides an example of each of the yaw angle irregularities explained in the Methods section.)

Representativeness of the data

The secondary parameters elucidated the differences between the PCTSD and the SHUFO data, and the national data. Many crashes (34%) were reported during the early evening (3.00–8.00 pm) in the latter. However, fewer were present in the SHUFO (29%) and PCTSD (30%) data. In contrast, SHUFO (33%) and PCTSD (30%) had more night-time crashes (12.00–8.00 am) compared to the national data (27%). The distributions for SHUFO and PCTSD did not differ substantially by hour of the day (Fig. E1, Appendix E). Notably, crashes were more frequent when

sunny weather was followed by cloudy and rainy weather in all three datasets. No differences were observed in the distribution pattern of crashes across weather conditions (Fig. E2, Appendix E). Finally, crashes occurred more frequently on a straight road (73%) than at an intersection in the national data, while in the PCTSD and SHUFO data the crashes were more evenly distributed between four-way intersections (37% and 38%) and straight roads (36% and 35%), respectively. Thus it appears that in the SHUFO and PCTSD data, intersection crashes were over-represented and crashes on straight roads were under-represented (Fig. E3, Appendix E).

IV. DISCUSSION

The PCTSD were successfully created with 529 cases in the first version using SHUFO data. The PCTSD, along with in-depth SHUFO data, will help researchers to identify the most appropriate safety systems for Chinese traffic. For example, the data can be used for parametric sensitivity analysis, examining different parameters in a system and their effects, and for studying the possibility of using multiple sensors [8][21-22].

PCTSD data can also improve our understanding of the future crashes that will be left after all the safety interventions are implemented. In the past, future crashes have been analysed by performing case-by-case analysis [6] or by applying system-specific rules to in-depth data [23]. Simulations using the PCTSD can supplement these approaches. The crash characteristics of the future crashes can guide the future development of the testing tools in C-NCAP or C-IASI. For example, if low-speed crashes increases substantially in the future, then more biomechanical research and development to make Anthropomorphic Test Devices (ATDs) suitable for injury assessment in low-speed crashes will be called for.

Further, the duration of pre-crash manoeuvres and the magnitude of velocity or acceleration can also be used to set up finite element simulations (which include pre-crash information) using Active Human Body Models [24] based on real-life data. For all impact types, the median pre-crash braking time (how long before the crash the brakes were first applied) for the car was 1.2 seconds (interquartile range: 0.5–2.3 seconds). Braking (early or late) information from the PCTSD and impact types from SHUFO data can complement specific human body model (HBM) simulations with real-life data. Such simulations could help researchers to study the kinematics of occupants and the biomechanics of injuries in real-world situations.

The PCTSD showed no substantial difference from the SHUFO data in the distribution pattern, indicating the conversion in each of the categories (hour of crash, weather condition, road type) was uniform. However, the PCTSD and SHUFO data distributions differed from the national data, particularly for the hour of the crash and road type. Further crash severity details were unavailable in the national data, so any irregularities in the severity distribution of the PCTSD sample were unknown. Effective use of PCTSD for reducing road traffic deaths in China requires data representative of the country; weighting factors, based on variables such as the hour of the crash and road type, are needed for each case [25-26].

Limitations

The current work differs in some ways from the previous work [13]. The major difference is the method for matching the participant sequences in the reconstruction file and the dataset. In the previous work, the make and model of the vehicle were used, while in the current work the participant's weight was used. The resulting limitation is that the weight of the participants was missing for approximately 25% of the data (all these were VRUs: powered two- and three-wheelers, bicycles, and pedestrians). However, these VRUs were often crashed into by comparatively heavy participants (cars or trucks; Fig. 3) whose weights were available. It was thus possible to match and assign the participant numbers appropriately using only the heavy participant's weight.

It is evident from Fig. 6 that the difference in the participant's weights between the reconstruction file and the dataset was the major reason for dropping most of the cases. Ideally, the weights in the reconstruction file should match the vehicle specification, which was also recorded in the dataset. In practice, however, they did not always match. This information can guide future updates: minimising the differences between the reconstruction file and the dataset would allow a considerable portion (64%) of the dropped cases to be available in PCTSD format. Alternatively, a lookup of the vehicle's make and model could be performed. With practice (in the dataset and in the reconstruction file) the process could be standardised, and the previous method [13] of creating PCTSD data could be applied.

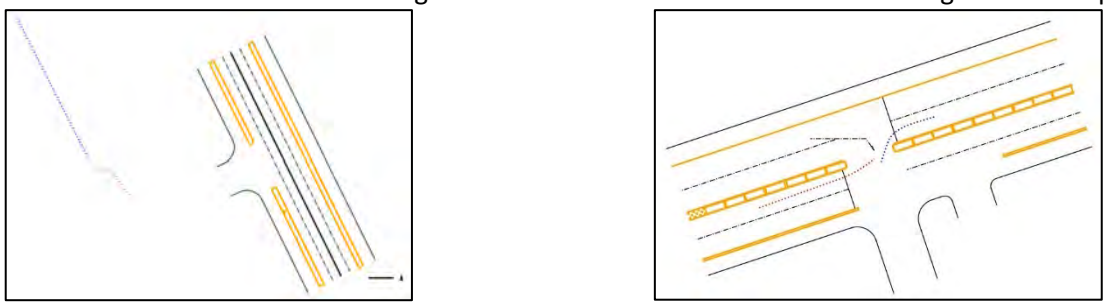
Overall, there were 192 cases in which the participants did not collide. Of these, 84% were corrected by forward extrapolation. The remaining 16% cases (N=31) were dropped. Forward extrapolation did not work in cases where

the rigid body was switched to the multibody. In future, it is recommended that the switch between rigid body and multibody happen only at the collision point at exact coordinates, not during the pre-crash phase.

Forward extrapolation was not even attempted for single-vehicle cases (vehicle-to-object), as there could be multiple objects in a crash, and detecting which object was contacted by the vehicle would be difficult. For example, there could be two poles in the object data, but the vehicle only collided with the first pole in the original case. The forward extrapolation in the current work does not check which of the two objects should be considered. Thus, the vehicle might collide with the second pole in the extrapolation instead of the first. This alteration cannot be detected, making it impossible to validate the extrapolated data.

In the current version, a single variable in the participant data describes the track width of the vehicle. For a powered three-wheeler, the rear trackwidth was recorded. In general, a powered three-wheeler configuration has a single wheel in the front and two wheels in the rear. However, in recent years the configuration of the three-wheeler could also be two wheels in front and a single wheel in the rear. This more recent configuration was overlooked and adds to the limitations of the current version of PCTSD. The limitation can be handled in future by having two dedicated variables for trackwidth: trackwidth for the front and trackwidth for the rear.

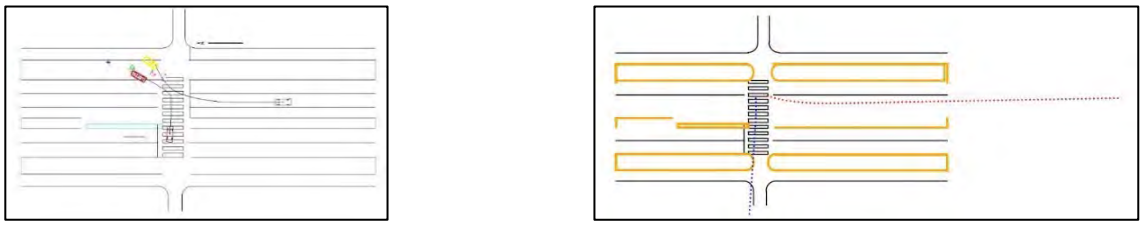
The manual calculations of the offset values are a limitation, as they may include human errors (missing a case or entering the wrong values). Fig. 7(a) shows a pre-2019 case with an inaccurate offset calculation: the dynamics data (red and blue dotted lines) were not even on the road. For comparison, Fig. 7(b) shows a case with perfectly calculated offset values. However, this vulnerability to human error will be eliminated in the future, as since 2019 the SHUFO researchers have used the drawing file in the reconstruction instead of an image or a bitmap file.



(a) (b)

Fig. 7. PCTSD sample case before 2019 with (a) inaccurate offset calculation and (b) accurate offset calculation.

The linear backward extrapolation performed on the data provided enough pre-crash simulation time to simulate any safety system in the future. However, assumptions in the extrapolations add to the limitations. For example, the interference of the participant's path with an object or the deviation from the pathway was not validated. Fig. 8(a) shows a sample case with a pre-crash time of less than 5 seconds, and Fig. 8(b) shows that, after backward extrapolation, the pre-crash time was 5 seconds. However, in the latter, the path of the vehicle (blue dotted line) passes through an object (amber-coloured polygon), which in the real world is not possible. Performing a backward extrapolation that eliminates this issue is beyond the scope of the current work (as well as a tedious undertaking). However, the backward extrapolation can be eliminated if the researcher extends the vehicle path sufficiently backwards while reconstructing. For better accountability between the data based on investigation and the data based on this backward extrapolation, the researcher should document this step during the reconstruction.



(a) (b)

Fig. 8. A car-to-car crash at an intersection with (a) the original pre-crash time of 1.58 seconds and (b) backward extrapolation with a total pre-crash time of 5 seconds.

The yaw angle irregularities observed in 58 cases must be either treated separately (smoothening) or discarded from the analysis. As there is a possibility to treat the data before use, these 58 cases were not dropped. However, to avoid such irregularities in future data, it is recommended to have a smooth vehicle path and to ensure there are no unstable yaw conditions present during the pre-crash phase in the reconstruction file.

Another limitation was due to database management systems such as inconsistencies in file naming nomenclature and in folder structure due to which 11 cases were missed. Recommendations for improving the database management system based on the lessons learned in the current work have been provided to SHUFO.

Future outlook

Finally, some recommendations for the future based on the current work are summarized below:

- The weights of the participants in the reconstruction file and the datasets should be the same.
- A common lookup for make and model could be used in the reconstruction file and in the dataset.
- The participants in the reconstruction file should be loaded in the same sequence as the participant numbers assigned in the dataset.
- Participants should be switched from rigid body to multibody during the in- (or post-) crash phase and at precisely the same coordinates.
- The reconstruction file should undergo a stringent quality control process.
- When the pre-crash simulation time is less than five seconds, the vehicle path should be extrapolated backwards to achieve a time of five seconds.
- Weight factors should be derived against each case, to make the PCTSD representative of China.

V. CONCLUSION

The current work is the first step towards successfully creating PCTSD by modifying the previous method. Overall, the quality of the data is promising, and after addressing the recommendations and calculating the weight factors, the data can be used to assess driver assistance and crash-avoidance systems in virtual settings for Chinese crash scenarios.

VI. ACKNOWLEDGEMENTS

The authors want to thank Da Wang, Mohankumar Jayathirtha, and Ekant Mishra for reviewing the technical contents of the paper. The authors also want to thank Guorong Li and Dr K. Mayberry for the language review.

VII. REFERENCES

[1] World Health Organization, (2023), Global Status Report on Road Safety 2023.

[2] Zhang, X., *et al.*, (2013) Basic characteristics of road traffic deaths in China. *Iranian Journal of Public Health*, **42**: pp. 7–15.

[3] Institute for Health Metrics and Evaluation (IHME) (2020) GBD Compare Data Visualization, http://vizhub.healthdata.org/gbd-compare (accessed 15 February 2024).

[4] Cicchino, J. B. (2017) Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accident Analysis and Prevention*, **99**: pp. 142–152, doi: 10.1016/j.aap.2016.11.009.

[5] Yang, X. (2019) Characterizing car to two-wheeler residual crashes in China Application of AEB in virtual simulation.

[6] Strandroth, J., *et al.* (2012) A new method to evaluate future impact of vehicle safety technology in Sweden. *Stapp Car Crash Journal*, **56**: pp. 1–13.

[7] Erbsmehl, C. (2009) Simulation of real crashes as a method for estimating the potential benefits of advanced safety technologies. *Proceedings of the 21st International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2009, Stuttgart, Germany.

[8] Rosen, E. (2013) Autonomous Emergency Braking for Vulnerable Road Users. *Proceedings of IRCOBI conference*, 2013, Gothenburg, Sweden.

[9] Sander, U. (2017) Opportunities and limitations for intersection collision intervention—A study of real world 'left turn across path' accidents. *Accident Analysis and Prevention*, **99**: pp. 342–355, doi: 10.1016/j.aap.2016.12.011.

[10] Schubert, A., *et al.* (2012) Standardized Pre-Crash-Scenarios in Digital Format on the basis of the VUFO Simulation. *Proceedings of ESAR*, 2012, Hannover, Germany
[11] Schubert, A., *et al.* (2016) The GIDAS pre-crash-matrix 2016 : Innovations for standardized pre-crash-scenarios on the basis of the VUFO simulation model VAST. *Proceedings of ESAR*, 2016, Hannover, Germany.
[12] Spitzhüttl, F., *et al.* (2015), Creation of pre-crash simulations in global traffic accident scenarios based on the iGLAD database. *Proceedings of Future Active Safety Technology Towards zero traffic accidents (FASTzero)*, pp. 427–433, 2015, Gothenburg, Sweden.
[13] Shaikh, J. and Sander, U. (2018) Creation of Pre-Crash Time-Series Database for Evaluation of Active Safety Systems using RASSI data. *Proceedings of ESAR*, 2018, Hannover, Germany.
[14] Deng, B., *et al.* (2013) Traffic Accidents in Shanghai - General Statistics and in-Depth Analysis. *Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2013, Seoul, Republic of Korea.
[15] Sui, B., *et al.* (2021) Evaluating automated emergency braking performance in simulated car-to-two-wheeler crashes in China: A comparison between C-NCAP tests and in-depth crash data. *Accident Analysis and Prevention*, **159**: doi: 10.1016/j.aap.2021.106229.
[16] Traffic Management Research Institute of the Ministry of Public Security (2018) Annual Report of Traffic Accidents in China.
[17] Bennett, J. (2015) AutoIt v3.
[18] Guthrie CAD/GIS Software Pty Ltd. (2017) CAD2Shape.
[19] VUFO GmBH, Pre-Crash-Matrix (PCM), https://www.vufo.de/pcm (accessed 15 February 2024).
[20] R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria.
[21] Zhao, M., *et al.* (2017) Method to Optimize Key Parameters and Effectiveness Evaluation of the AEB System Based on Rear-End Collision Accidents. *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, **10**: pp. 310–317, doi: 10.4271/2017-01-0112.
[22] Saadé, J., *et al.* (2019) Prospective evaluation of the effectiveness of autonomous emergency braking systems in increasing pedestrian road safety in France. *Proceedings of IRCOBI*, 2019, Florence, Italy.
[23] Ostling, M., *et al.* (2019) Passenger Car Safety Beyond ADAS: Defining Remaining Accident Configurations As Future Priorities. *Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2019, Eindhoven, Netherlands.
[24] Mishra, E., *et al.* (2023) Repositioning forward-leaning passengers by seatbelt pre-pretensioning. *Traffic Injury Prevention*, **24**: pp. 716–721, doi: 10.1080/15389588.2023.2239408.
[25] Hautzinger, H., *et al.* (2004) Expansion of GIDAS Sample Data to the Regional Level : Statistical Methodology and Practical Experiences. *Proceedings of ESAR*, 2004, Hannover, Germany.
[26] Padmanaban, J., *et al.* (2017) Methodology to Derive National Estimates of Injuries and Fatalities in Road Traffic Crashes in India. *SAE Technical Papers*, pp. 1–5.

VIII. APPENDIX

Appendix A: Estimating coefficient of friction using surface type and surface condition

TABLE AI
COEFFICIENT OF FRICTION BASED ON THE COMBINATION OF SURFACE TYPE AND SURFACE CONDITION

Surface Type	Surface Condition	Coefficient of friction
Concrete	Dry	0.85
Concrete	Damp	0.68
Asphalt	Dry	0.75
Asphalt	Damp	0.60
Stone, Pavement	Dry	0.70
Stone, Pavement	Damp	0.56
Sand, Gravel	Dry, Damp	0.50
Concrete, Asphalt, Stone, Pavement, Sand, Gravel	Wet	0.55
Concrete, Asphalt, Stone, Pavement, Sand, Gravel	Snow	0.30
Concrete, Asphalt, Stone, Pavement, Sand, Gravel	Ice	0.18

Appendix B: Description and other details of variables in each record in the PCTSD data

TABLE BI
VARIABLE DESCRIPTIONS FOR GLOBAL DATA

Variables	Description	Values
FALL	Case numbers	
PARTICIP	Total participants involved in the crash	

TABLE BII
VARIABLE DESCRIPTIONS FOR PARTICIPANT DATA

Variables	Description [units]	Values
FALL	Case numbers	
BETNR	Participant number in line with SHUFO data	
TYPEPCTSD	Type of participant in the PCTSD data	0 – car 1 – pedestrian 2 – motorcycle 3 – bicycle 4 – truck 14 – three-wheeler
LENGTH	Length of the participant [m]	
WIDTH	Width of the participant [m]	
HEIGHT	Height of the participant [m]	
WEIGHT	Weight of the participant [kg]	
TRACKWIDTH	Trackwidth of the participant [m]	
WHEELBASE	Wheelbase of the participant [m]	
DISTCGFA	Distance between centre of gravity and front axle [m]	
CGFRONT	Distance between centre of gravity and front leading edge [m]	
HEIGHTCG	Height of centre of gravity from ground [m]	
WIDTHRATIO	The ratio to define the front leading edge of the car (only applicable for car)	0.6 – car 99999 – others
MUE	Coefficient of friction	
DISTHF	The ratio to define the distance of handlebar to the leading edge (only for two-wheelers)	0.4 – motorcycle 0.27 – bicycle 99999 - others
IXX, IYY, IZZ	Moment of inertia to: roll, pitch, yaw [kgm2]	

TABLE BIII
VARIABLE DESCRIPTIONS FOR DYNAMICS DATA

Variables	Description [units]	Values
FALL	Case numbers	
BETNR	Participant number in line with SHUFO data	
STEP	Simulation step [s]	
XPOS, YPOS	Global X-, Y- coordinates of vehicle's centre of gravity [m]	
VX, VY	Velocity of vehicle in local X-, Y- coordinates [m/s]	
AX, AY	Acceleration of vehicle in local X-, Y- coordinates [m/s2]	
PSI	Yaw angle of vehicle in global coordinate system [rad]	
BRAKING	Sequence considered for the given simulation step	-1 – no braking and acceleration 0 – no braking and constant speed 1 – brakes applied and deceleration

RECON	Details for the given simulation step availability due to reconstruction or extrapolation	0 – Details available due to extrapolation 1 – Details available due to reconstruction
TTC	Time to collision [s]	

TABLE BIV
VARIABLE DESCRIPTIONS FOR OBJECTS AND ENVIRONMENT DATA

Variables	Description [units]	Values
FALL	Case numbers	
OBJTYPE	Object classification type	101, 201 – roadside edge 102, 103, 207 – centre line marking 105, 204, 206 – median lines 199 – interrupted lines 202 – stop line 214 – zebra crossing 300, 399 – others 501 – tree 524 – wall 520 – object (not defined)
LINENO	Sequential line numbers assigned to respective objects	
POINTNO	Sequential point numbers assigned to each line numbers	
X, Y, Z	Global X-, Y-, Z- coordinates [m]	

Appendix C: Predefined values for the participants loaded as a multibody

TABLE CI
PREDEFINED VALUES USED IN PARTICIPANT DATA FOR POWERED TWO-WHEELER WHEN LOADED AS A MULTIBODY

Variables	Values	Remarks
LENGTH	2.5 m	
WIDTH	0.5 m	
HEIGHT	0.75 m	
WEIGHT	259 kg	
TRACKWIDTH	0.1 m	
WHEELBASE	1.599 m	
DISTCGFA	0.75 m	
FOVERHANG	0.01 m	FOVERHANG is the front overhang, the distance between the front axle and the front leading edge. Front overhang is added to the distance between the centre of gravity and the front axle to get the value for CGFRONT.
HEIGHTCG	0.65 m	
IXX	104 kgm2	
IYY	130 kgm2	
IZZ	130 kgm2	

TABLE CII
PREDEFINED VALUES USED IN PARTICIPANT DATA FOR BICYCLE WHEN LOADED AS A MULTIBODY

Variables	Values	Remarks
LENGTH	1.7 m	
WIDTH	0.2 m	
HEIGHT	1.3 m	
WEIGHT	108 kg	
TRACKWIDTH	0.01 m	
WHEELBASE	1.47 m	
DISTCGFA	0.74 m	
FOVERHANG	0.1 m	
HEIGHTCG	0.9 m	
IXX	10.3 kgm2	
IYY	34.2 kgm2	
IZZ	34.2 kgm2	

TABLE CIII
PREDEFINED VALUES USED IN PARTICIPANT DATA FOR PEDESTRIAN

Variables	Values	Remarks
LENGTH	0.4 m	
WIDTH	0.8 m	
HEIGHT	1.8 m	
WEIGHT	99999	
TRACKWIDTH	99999	
WHEELBASE	99999	
DISTCGFA	99999	
CGFRONT	0.2 m	
HEIGHTCG	99999	
IXX, IYY, IZZ	99999	

Appendix D: Example cases for irregularities in the yaw angle of participants

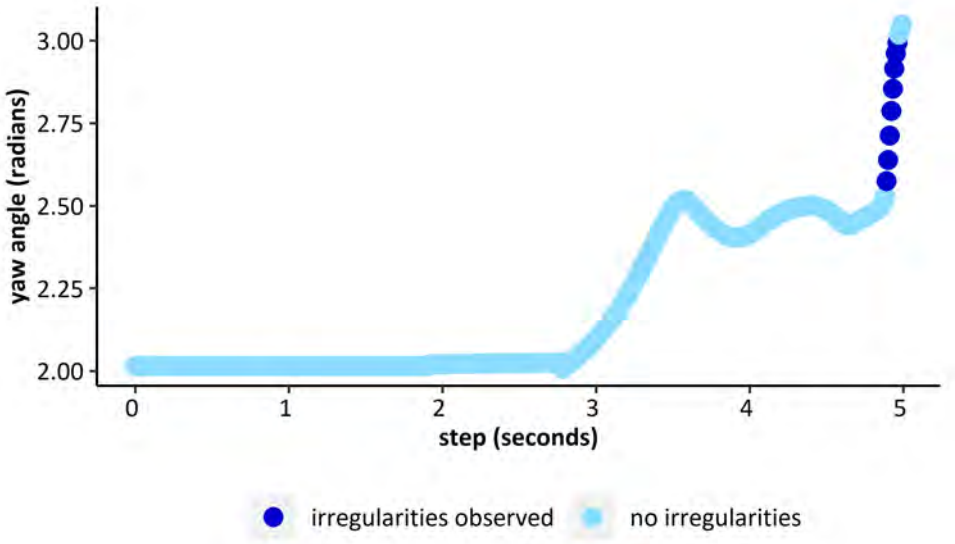


Fig. D1. A case where there was an abrupt change in the direction – dark blue dots indicate the change in yaw angle greater than the threshold value while the light blue dots represent the change in yaw angle within the threshold value.

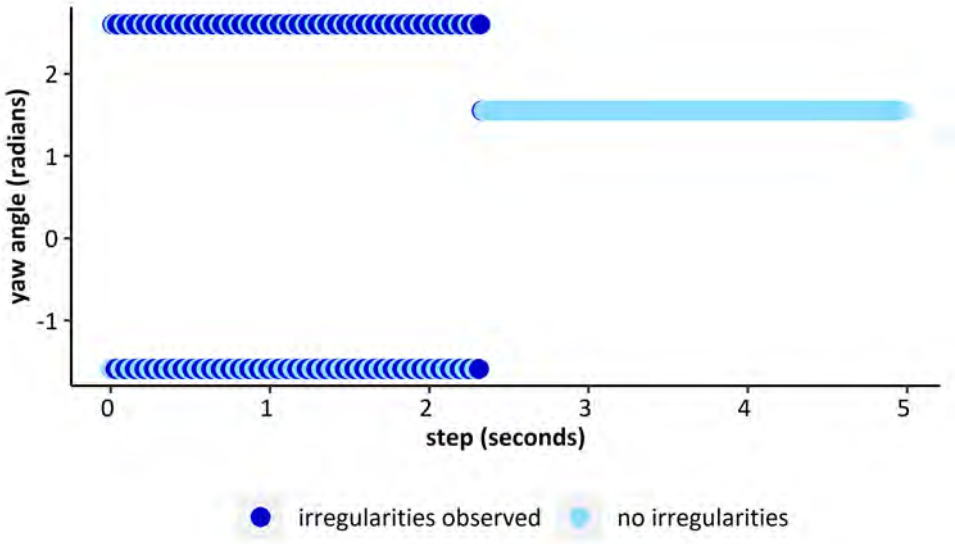


Fig. D2. A case where a constant fluctuation in the yaw angle was observed in the pre-crash phase until 2.3 seconds due to the difference between velocity direction and participant's heading angle – dark blue dots indicate the change in yaw angle greater than the threshold value, while the light blue dots represent the change in yaw angle within the threshold value.

Appendix E: Distributions of secondary variables (hour of the crash, weather condition, and road type) for Chinese national data, SHUFO database, and PCTSD

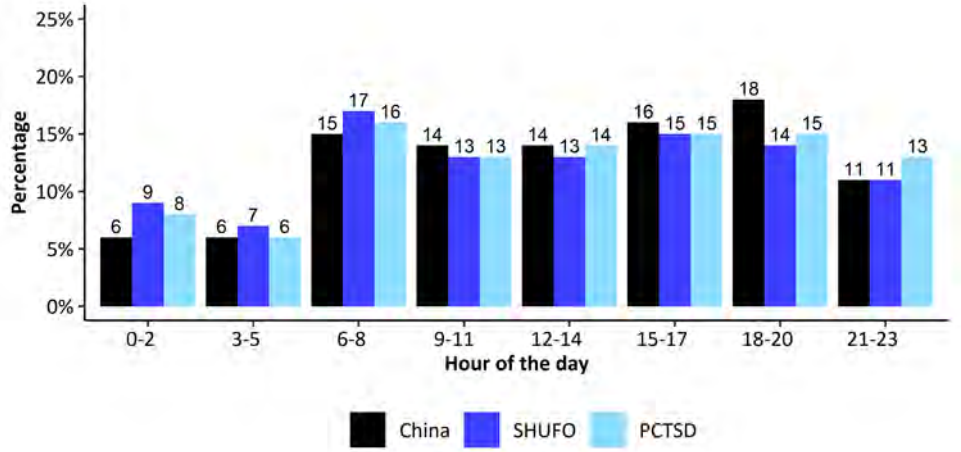


Fig. E1. Distribution of the hour of the crash (time of day) observed in the Chinese national data, SHUFO database and PCTSD.

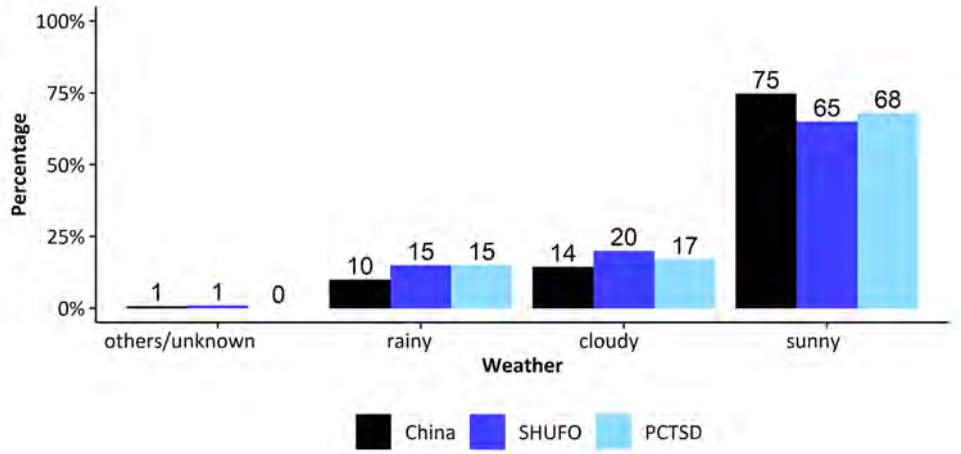


Fig. E2. Distribution of weather conditions observed in the Chinese national data, SHUFO database and PCTSD.

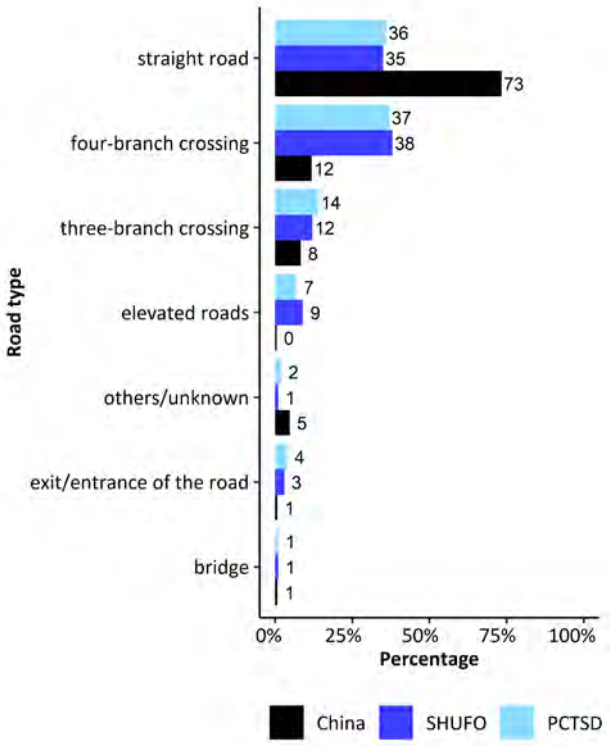


Fig. E3. Distribution of road type in the Chinese national data, SHUFO database and PCTSD.