



Advancing Road Safety Through Data: Challenges, Solutions, and Policy Insights from EU Road Safety Projects

Lucchesi, S. T.^a; Schmalholz, N. ^a; Vounasis, D.^b; Hula, A. ^b; Papadakaki, M.^c; Vagionaki, K.^c; Chapman, S. ^a; Burke, M. ^a; Oliveira, G. ^a; Op den Camp, O.^d; van Gils, S. ^d; Leitgeb W.^e; Lucio J. Kolionis G.^f; Inglese G. ^f; Lazarova Y. ^f; Sfyridis, A.^g; Stoilova, R.^g; Bapaume, T. ^g; Ameli, M. ^g; Elizondo, E. ^g; Georgiou, M. ⁱ; Gkountoumas, F.ⁱ; Roungas, B.^h; Amditis, A. ^h; Quintero, K.^j; Loupos K. ^k; Brasinika D. ^k; Despotopoulou A.M. ^k; Katelaris L.^k; Laoudias C.^k; Panayiotou C.^k

^aPHOEBE, ^bCAMBER, ^cSAFETeen, ^dV4SAFETY, ^eProtact-us, ^fFRODDO, ^giDRIVING Project,
^hEVENTS, ⁱAI4CCAM, ^jSOTERIA, ^kEvoRoads

Introduction

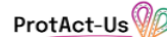
Reliable and comprehensive data is the cornerstone of effective road safety strategies. It enables authorities and practitioners to move from reactive approaches, responding only after crashes occur, to proactive interventions that prevent accidents before they happen, or reduce the severity of their consequences. High-quality data from diverse sources such as crash records, (connected) vehicles, and multiple types of (infrastructure or mobile) sensors allows for accurate risk identification, predictive modelling, and targeted investments in infrastructure improvements.

Nevertheless, obtaining road safety data that is truly comprehensive, accurate, ethically managed, thoroughly documented, and adaptable to changing requirements presents significant challenges. Issues related to data collection, processing, and accessibility are frequently encountered by researchers and practitioners who aim to forecast crash events and their consequences, as well as to analyse the underlying causes contributing to accidents (Mannering & Bhat, 2014).

Against this background, the evolution of road safety data can be understood as a transition from fragmented and reactive datasets towards integrated, predictive, and governance-ready information ecosystems. The European Commission has devoted considerable efforts to stimulate this transition through initiatives like the [European Road Safety Observatory](#), the [EU Road Safety Exchange Project](#) and numerous research and innovation projects, funded under the Horizon Europe framework.

EU-funded road safety projects have joined forces in the EU Road Safety Cluster, aiming to tackle complex urban challenges, create predictive safety frameworks, and enhance decision-making through improved data. These projects focus on different areas such as automated mobility, interaction between drivers, and vulnerable road users' risk (VRUs). One cluster initiative is this joint paper, which promotes knowledge sharing within and beyond the Road Safety Cluster by identifying key challenges in road safety data. Objectives include mapping the road safety data landscape, analysing common challenges within the project and showcase strategies to ensure data is not the obstacle for proper identification and mitigation of road safety risk. It covers traditional data sources (e.g., crash records, survey data, loop sensors, counting flows) and new data sources (e.g., connected vehicle telemetry, and smartphone sensing).

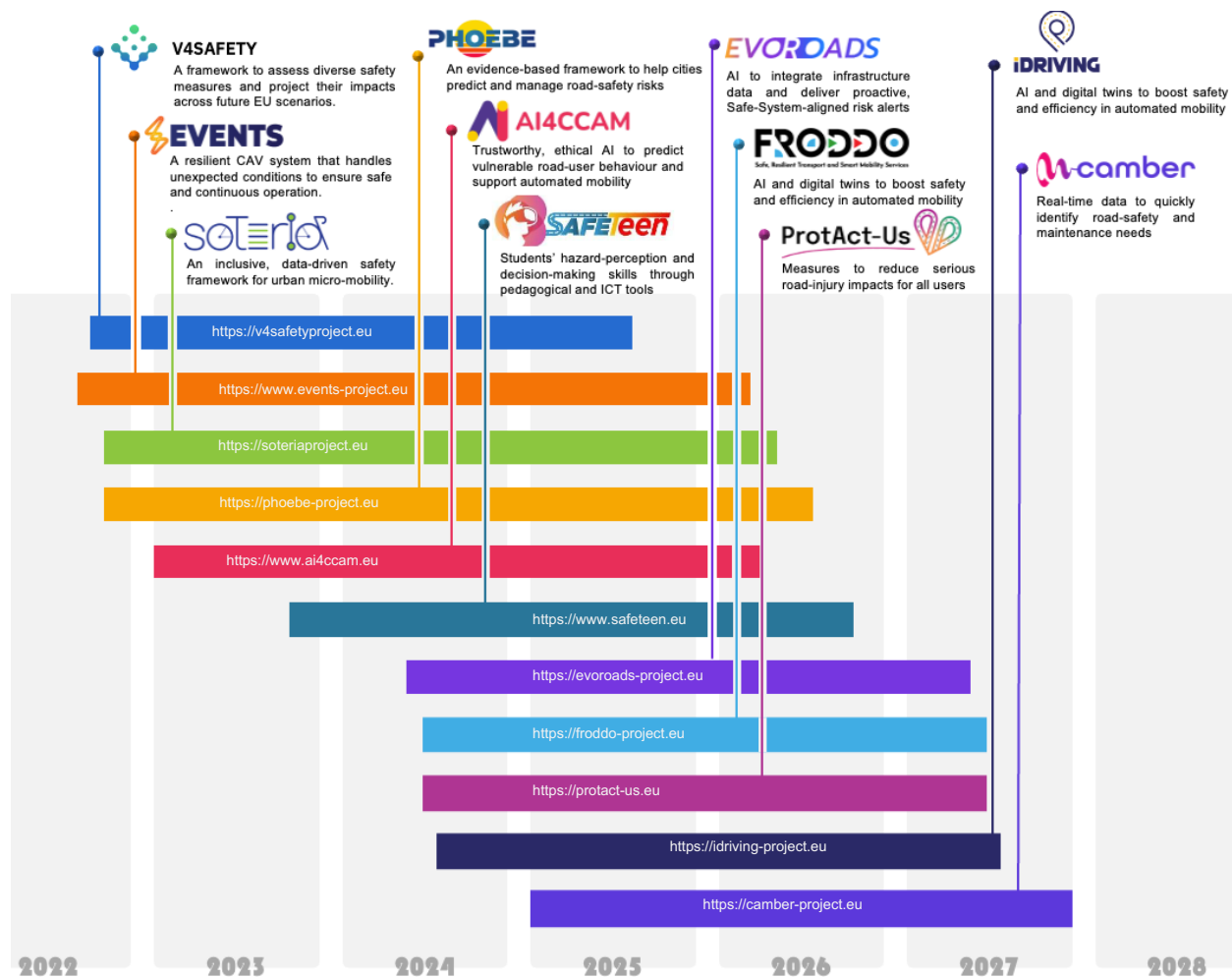
The dialogue within and among the projects engages stakeholders, such as public authorities, industry, and research organisations, focusing on governance, interoperability, and technical capacity, mainly in Europe but with global relevance. Furthermore, the cluster joint effort also presents project solutions to overcome data challenges, emphasising best practices of data management and the use of advanced analytics and harmonisation methodologies and tools. Case studies from these Horizon Europe projects demonstrate practical applications, and the paper concludes with policy recommendations for standardisation, funding, and collaboration to foster more effective and data-driven road safety decisions.



This paper positions EU-funded research projects as key drivers of this transition, demonstrating how advances in data integration, analytics, and governance can support proactive road safety interventions and informed policymaking.

The EU Road Safety Cluster

The EU Road Safety Cluster aims to promote a safe, inclusive, sustainable and efficient mobility system that is resilient, trustworthy, and road user centric. By uniting efforts across the integrating projects, this initiative is set to transform European transport research and establish new standards for road safety and automated driving. The composition of the cluster evolves over time, as projects enter or exit in line with their project duration, and availability after completion, as shown in the Gantt chart below, which also includes short project descriptions, while detailed information is available on the dedicated project websites or as part of the EU Road Safety Cluster [LinkedIn](#) site.



Data sources

Robust road safety analysis requires a foundation built on rigorous methodologies combined with data secondary data sources rather than primary data collection. This reliance stems from the multidisciplinary nature of road safety, which demands insights from diverse domains such as engineering, urban planning, behavioural science, and policy. Consequently, projects often integrate multiple secondary datasets originating from various organisations across both the public and private sectors.

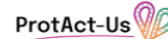
Data types can be grouped into three main groups: behavioural data, road infrastructure data, as well as crash and hospitalisation data. Some projects also demand other types of data for more holistic analysis of intervention impacts like mobility and environmental data. The methods for obtaining this data vary significantly. While certain data types are accessible through public databases, others come from emerging sources such as sensors, telematics, and vehicle-to-everything (V2X) communications.

- **Behavioural Data:** Behavioural data refers to information capturing how road users act and respond within traffic environments under various (weather, lighting, and surface) conditions. It focuses on actions, decisions, and patterns that influence crash risk and injury severity. Several projects, including **EVORoads**, **FRODDO**, **iDRIVING**, **PHOEBE**, **SOTERIA**, **V4SAFETY**, **CAMBER**, and **SAFETEEN** rely on this data type.
- **Road Infrastructure Data:** Road infrastructure data includes information about physical road elements and their characteristics that affect traffic safety. This data supports risk assessment, safer road design, and the prioritisation of safety improvements. Most listed projects, including **EVORoads**, **EVENTS**, **FRODDO**, **iDRIVING**, **PHOEBE**, **SOTERIA**, **V4SAFETY**, **CAMBER**, and **SAFETEEN** also rely on this type of data, with broad coverage across the same group of initiatives.
- **Crash and Hospitalisation Data Linked to Road Traffic Crashes:** This data refers to records and statistics from road traffic crashes, including crash databases and information about injuries requiring medical treatment or hospitalisation. It is essential for understanding the severity and impact of traffic incidents. Like the other data types, it is used across nearly all the projects in the cluster, including **EVORoads**, **FRODDO**, **PHOEBE**, **PROTACTUS**, **SOTERIA**, **V4SAFETY** and **CAMBER**, also rely on this type of data, with broad coverage across the same group of initiatives.

Projects have been able to gather this data due to collaboration with key stakeholders. At the European level, institutions such as **CARE** and **IRTAD** provide centralised crash and safety data for monitoring and benchmarking. National authorities play a vital role in collecting and analysing country-specific data. However, the projects' engagement has shown that while data are available and reported, access is sometimes limited and coordination between institutions could be improved.

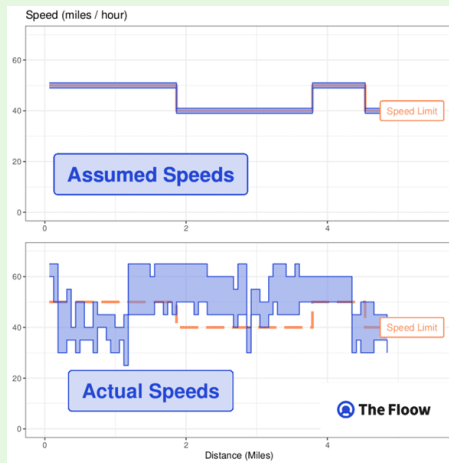
Modern road safety strategies increasingly harness advanced technologies to shift from reactive crash analysis toward proactive risk prediction. These innovations deliver high-resolution, real-time data that complement traditional datasets, enabling smarter, data-driven interventions and more effective safety measures. Emerging approaches increasingly rely on federated data architectures and Digital Twin concepts to integrate these heterogeneous sources without requiring full centralisation of sensitive or proprietary data. Within this context, projects such as **FRODDO** explore how federated Digital Twins can act as abstraction layers that translate diverse road safety and mobility data streams into consistent, safety-relevant indicators usable by operators and authorities.

Here are some examples of how projects are exploring new ways of fulfilling the gaps of road safety related data.



PHOEBE Next-Generation Risk Insights from Smartphone-in-Traffic Data

Smartphone use while driving is widely recognised as a major contributor to road risk. Statistics collected in the United States (FARS, collated by NHTSA) and in the UK (Road Safety Data, DfT) suggest that the contribution to fatal crashes by distraction due to cell-phone use is around 1% of all fatal incidents. However, naturalistic driving studies (Dingus et al. 2016, 2019) have demonstrated that the relative risk of a collision increases by a factor of 2-3 while a driver is using a hand-held phone. Very little work has been carried out to assess the extent to which certain locations or environments might be preferentially favoured by drivers for using a smartphone. Fixed-location monitoring offers limited coverage and fail to provide the comprehensive insights needed to understand when, where, and why risks occur. To address



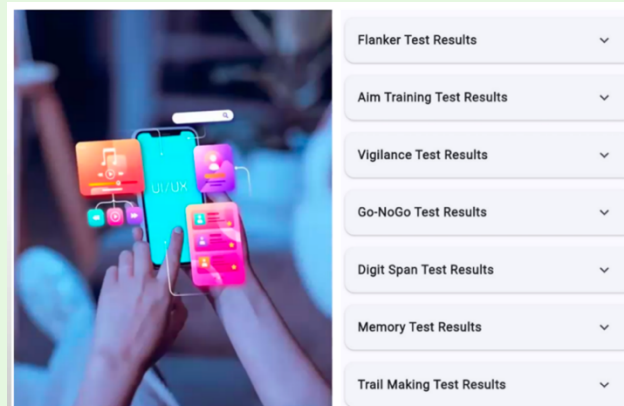
this gap, **PHOEBE** is developing new large-scale, anonymised telematics-based approaches to analyse phone-use patterns across different environments, locations, and times.

Through advanced statistical methods and new data-handling techniques, **PHOEBE** has demonstrated early proof of value for measuring real-world smartphone use during driving and understanding the contexts in which it occurs. Findings show that phone use varies significantly by geography, road type, time of day, and driving conditions, and that driving behaviour noticeably changes when drivers are on calls. These insights enable fine-grained modelling of risk across large areas, highlighting strong potential for locally targeted interventions and scalable policy tools.

SAFETeen apps: Smart Self-Assessment for data collection

The **SAFETeen** Emergency Aid Tool app helps teenagers to quickly assess whether they are fit to commute. It raises awareness of hidden risks, such as mental health, physical condition, cognitive impairments, road hazards, and substance use, and begins with a baseline questionnaire on health, habits, and emotional state.

On subsequent logins, users complete short mood and fitness checks, followed by cognitive tasks like memory, attention, and reaction-time tests. They then receive personalised feedback



featuring test scores compared to norms, constructive remarks linking their baseline profile to mood and fitness responses, commuting safety recommendations, and relevant resources.

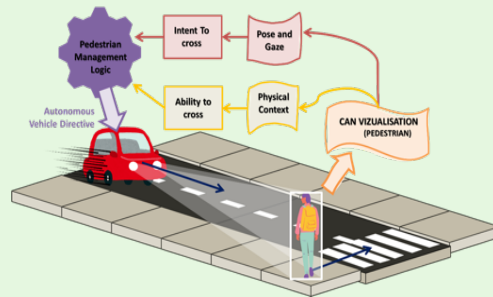
The project also includes a VR/AR education tool that teaches road-safety skills through immersive, everyday scenarios. Users can replay scenarios to learn from past mistakes and omissions. They face key point dilemmas that shape scenario progress, choosing between safe and unsafe options often with hidden risks. The tool educates teenagers to recognise and assess hidden risks and potentially harmful situations. It encourages making the right choices, aligning with VR's strengths in interactive, experiential learning for road safety.

AI4CCAM simulation and VR experiments to model (among others) pedestrian behaviour

Pedestrians decide whether to cross based on perceived ability rather than objective physical measures. This ability reflects a person’s judgement of whether they can safely complete a crossing, shaped by motor and cognitive capacity as well as situational factors like vehicle speed, distance, visibility, weather, and road geometry. Research shows that people rely on perceived time-to-contact (TTC), which varies between individuals and is systematically biased by factors like speed, attention, luminance, and age.

AI4CCAM use case used Virtual Reality experiments to model crossing pedestrians’ and AV approaches. Data was acquired from participants performing this crossing task and they pressed and held a controller when they would initiate crossing.

In the VR model described, ability is represented by an internal safety threshold. The model aims to estimate this threshold from sensor-observable variables to assess crossing feasibility in real time, while intention cues are used secondarily to reduce false alarms without missing genuine crossing attempts. The evaluation of transfer to real data was done using a strict test-only protocol, the model was trained exclusively on VR trials and applied, without any fine-tuning, to three public datasets: PIE, LOKI, and nuScenes-mini. Overall, these results confirm that a threshold-centric model, calibrated under controlled VR conditions and deployed with a conservative bias, transfers effectively to diverse real-world scenes.



Taken together, these case studies illustrate a shared methodological trend across projects: the use of advanced sensing, simulation, and behavioural modelling to compensate for gaps in traditional road safety datasets. While the specific tools vary, from video analytics to VR-based experiments - the underlying objective is consistent: generating richer, safety-relevant evidence that supports proactive risk identification and intervention in complex urban environments.

Challenges in road safety-related data

The integration of diverse datasets in road safety and mobility research presents a series of complex challenges that span technical, operational, and ethical dimensions. While projects identify that the required data sources exist during the data mapping phase, significant challenges often emerge when attempting to work with this data. This chapter explores the primary obstacles encountered by these projects and highlights how they can inform the development of practices and policies aimed at improving data accessibility. By making road safety data more readily available, these measures can help ensure that road safety risks become easier to identify and address effectively.

Data Homogeneity and Coverage

Inconsistent definitions and technological barriers represent additional challenges in data collection (Imprialou & Quddus, 2019; Karimi et al., 2024; Mannering & Bhat, 2014; Sander et al., 2024; Soltani et al., 2024). Similarly, inconsistencies in injury severity definitions and reporting protocols across European

Union countries hinder cross-national comparisons and policy development (Amoros et al., 2014; Short & Caulfield, 2016; Yannis et al., 2014). In spatial analysis, the use of arbitrarily defined spatial units can lead to substantial variability in statistical inferences due to the Modifiable Areal Unit Problem (MAUP), limiting analysis interpretability and transferability (Wang et al., 2013; Ziakopoulos & Yannis, 2020).

A significant challenge in harmonising road safety and injury data lies in the differences between classification systems used in the medical and engineering domains. The **PROTACTUS** and **SOTERIA** identified that medical coding frameworks such as ICD-10 focus on clinical diagnoses and health conditions significantly different from engineering-oriented systems like ISS (Injury Severity Score) and AIS (Abbreviated Injury Scale). These divergent approaches create inconsistencies when integrating datasets for comprehensive analysis. Additionally, while the EU Injury Database (EU-IDB) and Common Accident Data Set (CADaS) provides a structured framework for injury reporting, its adoption and interpretation vary across countries, further complicating standardization efforts. Compounding these issues is the general lack of universally accepted definitions for key concepts such as “serious injury” and “long-term consequences (LTC)”, which leads to ambiguity in reporting and hinders comparability across studies and jurisdictions. The classification of injury severity (e.g. “slight,” “serious,” “fatal”) varies across countries. Some datasets include additional intermediate categories or use medical-based definitions, while others rely on police-assessed severity. Finally, while harmonisation improves comparability, it inevitably involves a degree of information loss or generalisation due to incompatible or missing attributes in the source data.

Technological barriers also reduce data quality and usability. The absence of standardised data collection and reporting procedures, as well as limited data-sharing agreements, APIs, and supporting infrastructure, creates significant challenges to the integration of multiple datasets—a key requirement for improving completeness through data linkage (Karimi et al., 2024; Soltani et al., 2024). Inconsistencies across agencies make accurate record linkage difficult, while limited data-sharing mechanisms and infrastructure hinder the implementation of linkage systems. Beyond linkage, the processing of emerging data sources such as Autonomous Vehicle (AV) crash reports requires substantial computational resources and specialised expertise to manage and integrate them with traditional datasets (Ding et al., 2025; Wen et al., 2021).

FRODDO, **PHOEBE** and **SOTERIA** also faced difficulties due to inconsistent definitions and classifications, which are expected but particularly challenging in approaches that aim to integrate heterogeneous datasets. For example, traffic KPIs (e.g., travel times, emissions, safety indicators) are defined differently depending on local infrastructure data sources, while human-centred measures of trust, comfort, and user experience varied across questionnaires and sensor modalities. Road functional classes and road type labels also differ significantly between datasets. Weather, lighting, and surface conditions were often recorded using different code lists or free-text values, requiring extensive normalisation.

Finally, data homogeneity and coverage remain a challenge in emerging data sources. The diversity of data formats and collection methods across different sources introduces significant challenges for integration, requiring extensive harmonization efforts to ensure consistency and usability. **FRODDO** have learned that positioning and navigation data relied on multiple standards, which complicates direct comparison across systems. **EVORADS**, on the other hand, identified several limitations on video data collection that affect its reliability and completeness. First, the video data does not always capture all types of defects, leaving certain issues undetected. Coverage is particularly limited in rural or low-traffic regions, where drone deployments are less frequent, resulting in gaps in spatial representation. Environmental factors such as weather conditions, dense vegetation, and physical obstructions further compromise image quality and reduce the reliability of defect detection. Additionally, historical data on defects is often unavailable, making it difficult to perform temporal comparisons or trend analyses.

Gaps in data representation and record-keeping

Gaps in regional data coverage and missing explanatory variables, such as vehicle speed, manoeuvring responses, and traffic volume, represent additional challenges in road safety modelling and inference (Ali et al., 2024; Dong et al., 2020; Eskandari Torbaghan et al., 2022; Imprialou & Quddus, 2019; Lord & Mannering, 2010; Mannering & Bhat, 2014; Mannering et al., 2020; Ziakopoulos, 2024).

While there is an abundance of vehicle-related data at certain levels, data concerning Vulnerable Road Users (VRUs) remains limited. For the **PHOEBE** use cases, behavioural data pertaining to pedestrians, bicyclists, and micromobility users, particularly those representing non-compliant behaviours or travel patterns, could not be sourced. Consequently, the project team had to employ traditional data collection methods such as surveys.

This challenge is mirrored in other initiatives, including **IDRIVING**, where comprehensive data sets are unavailable, necessitating manual data collection on multiple occasions. The **PROTACTUS** project also points out the limited data available regarding injuries from "new mobility" options such as e-scooters and pedelecs, as well as injuries involving bus and shuttle passengers. Some of these records may not be digitised or could only be accessible in local languages, challenges faced by **PHOEBE** and **EVORoads**. Additionally, certain behavioural patterns develop slowly, making them difficult to identify consistently over time.

For the **SOTERIA** project, one identified concern was that the datasets concentrate exclusively on crash outcomes and do not include critical contributing behavioural or environmental factors, which are essential for the development of effective interventions. Within the **CAMBER** project, which aims to develop and demonstrate improved safety monitoring across urban and secondary road networks through real-time data feedback, unequal and uneven data coverage was identified as a key challenge affecting the consistency and representativeness of the analyses. To address this issue, the project focuses on secondary and urban road networks, incorporates VRU metrics and micromobility data, procures 3rd party processed vehicle dynamics data, and complements traditional data sources with citizen reporting and AiRAP-based validation.

In projects focused on CCAM analysis, such as **FRODDO**, the availability of smart card ridership and traffic flow data can vary significantly by location. Additionally, certain types of information, like near misses and user attitudes, are not systematically gathered, which limits the capacity to benchmark safety outcomes and assess impacts directly. Although human-centred datasets are important, they have limitations: physiological and eye-tracking data are generally gathered from limited groups of participants. Because such data often come from small, controlled studies, it is usually not possible to establish broad reporting standards or ensure comparability across different locations. This illustrates a broader convergence between CCAM research and traditional road safety analysis, where automated mobility projects increasingly depend on exposure, behavioural, and near-miss data typically associated with road safety studies. **FRODDO** exemplifies this shift by highlighting how CCAM safety performance cannot be assessed in isolation from wider road safety data ecosystems and governance structures.

These gaps highlight the need for enhanced data collection mechanisms, such as self-reporting tools, sensor-based monitoring, and standardised formats across regions to ensure comprehensive and comparable road safety analysis.

Data quality and accuracy

Data accuracy is a cornerstone of reliable safety assessments and transportation modelling. In the context of automated and connected mobility systems, the integrity of data directly influences the precision of models such as mode choice and induced demand, and ultimately the credibility of policy and design decisions. However, achieving and maintaining high data quality presents significant challenges.

Many of the challenges previously mentioned also affect data quality and accuracy. As mentioned before, variability in terminology and categorization across jurisdictions and data sources introduces bias and hampers comparability. Additionally, as identified by **PROTACTUS** and **SOTERIA**, there is a lack of completeness for minor injuries in crash databases and hospitals, and police reports vary in quality and standards across regions which compromise quality.

For emerging data sources, it is critical to recognize that automated extraction processes, such as those relying on image detection algorithms, often involve approximations and are vulnerable to labelling subjectivity. **AI4CCAM** reinforced that representativeness of the training data that strongly affects the reliability of the AI models. These factors can introduce inconsistencies in data interpretation and reduce overall reliability, particularly when annotation quality depends on operator expertise or when algorithms lack robust validation mechanisms. For data provided by in-vehicle technologies, **FRODDO** identified that there may be some automated vehicle or automated driving system manufacturers that underreport or misclassify incidents submitted to regulatory bodies, such as NHTSA, to safeguard financial performance or stock value.

When discussing subjective data, inconsistencies can have an even greater impact on data quality and reproducibility. **SAFETeen** project identifies there is no universally accepted benchmark to delineate "safe" versus "unsafe" commuting readiness for adolescents, as developmental variability is high. Cognitive performance norms differ markedly by age (e.g., 11–13 vs. 15–17), pubertal stage, and neurodiversity (e.g., ADHD, autism spectrum traits, or learning differences), which are unevenly represented in existing normative datasets. Sparse, inconsistent record-keeping in school or community settings further hinders derivation of reliable, evidence-based thresholds tailored to this population. Additionally, high intra-individual and contextual variability such as environmental distractions (noise, fatigue, peer influence, weather, time pressure) can substantially alter performance on cognitive tasks within the tool, causing score fluctuations large enough to shift a student across a provisional cut-off on different occasions, even within the same day. This undermines the tool's reliability for consistent, binary "safe/unsafe" classifications and limits its utility for real-time parental, school, or self-monitoring decisions.

Low-quality data propagates uncertainty into predictive models, reducing their precision and reliability. This compromises decision-support systems and safety evaluations, particularly when models are used to inform infrastructure investments or regulatory frameworks.

Data accessibility and privacy issues

Accessibility barriers limit the usability, transparency and reproducibility of road safety data and analysis. These challenges are commonly classified into privacy and legal restrictions, and limited public and research access. Privacy and legal restrictions arise from ethical, legal, and cultural issues surrounding the use of personal data, including concerns about authorisation, transparency, and potential risks to individuals or organisations (Kalkman et al., 2019; Mitchell et al., 2014). Privacy legislation such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States can restrict access to sensitive information or require extensive data de-identification (Karimi et al., 2024). Likewise, the General Data Protection Regulation (GDPR) in the European Union (EU) restricts the use of informed and related connected car guidance from the EU sources or data gathered in the EU space, such as placing strict controls on mobility data usage. This prevents direct reuse of individual telemetry data without anonymised aggregations or direct consent for the purpose.

Limited public and research access, on the other hand, arises from proprietary, institutional, or non-regulatory restrictions on data distribution (Soltani et al., 2024). Data providers and researchers may be unable to disclose methodology details due to commercial or intellectual reasons (Ziakopoulos, 2024). Poor or inconsistent metadata, such as the missing information about data sources, collection procedures, and processing methods, further reduces transparency and restricts reproducibility (Soltani et al., 2024). These

challenges are exacerbated by the lack of standardised data schemas and metadata across jurisdictions, which hampers cross-border analysis (Imprialou & Quddus, 2019; Karimi et al., 2024; Soltani et al., 2024).

Open datasets are available, but sensitive data remain restricted due to GDPR, IP, and commercial sensitivities. Road safety related data is generally not publicly accessible due to its sensitive nature. Access is limited and strict agreements with data providers need to be in place to give projects access to the required data. Additionally, compliance with GDPR and local regulations introduces constraints on data collection, sharing, and processing and many times, the only accessible information is aggregated and anonymised data shared through statistical platforms or under license agreements that prevent raw data to be shared.

These limitations suggest that policy measures supporting trusted data intermediaries, standardised metadata practices, and harmonised access frameworks could play a crucial role in improving the usability and reproducibility of road safety research while respecting legal and commercial constraints.

For instance, the **FRODDO** project identified that EU data protection laws significantly hinder the sharing of detailed crash information, creating legal and operational challenges for data-driven safety research. In a similar vein, the **PHOEBE** project encountered significant limitations during its survey planning, as it was sometimes not possible to obtain valid informed consent for gathering location-specific data or identifiable information. Furthermore, additional GDPR requirements, such as minimising data collection, proportionality, and risk reduction, restricted what data could be gathered. As a result, the team had to make considerable changes to data collection forms to remain compliant. Moreover, **PHOEBE** highlighted that privacy limitations reduce access to demographic and behavioural data, which are crucial for robust analysis. To overcome these gaps, the project adopted alternative strategies such as complementary surveys validated against anonymized telematics datasets. Likewise, **CAMBER** identified a risk in implementing the project related to data accessibility and limitations in sharing data between pilots, due to challenges with data-sharing agreements and privacy requirements, potentially leading to delays or incomplete analyses. Ultimately, all the cases underscore the need for proactive planning, legal oversight, and innovative methodologies to balance data utility with regulatory compliance in mobility and safety projects.

There is also a concern that data collected by projects may unintentionally include sensitive information. For example, drone footage captured by the **EVORoads** project can inadvertently record identifiable details such as faces, license plates, or private property. In the case of **FRODDO** specific concerns include the collection of biometric and behavioural data in pilots, the potential re-identification of individuals in mobility traces, and the secure handling of real-time V2X communications. These risks highlight the need for robust privacy safeguards, including automated blurring or redaction, careful flight planning to avoid sensitive areas, and strict compliance with data protection regulations.

Lessons from EU Projects: Solutions and Best Practices

While the previous chapter detailed significant obstacles, the collaborative efforts within the EU Road Safety Cluster have generated a robust suite of solutions designed to bridge the gap between data collection and proactive safety management. This chapter explores the best practices and strategic methodologies developed across the eleven participating projects, providing a roadmap for transitioning from fragmented, reactive data usage to an integrated and resilient road safety ecosystem. As the Cluster is dynamic, the practices and methodologies described here reflect the current composition of participating projects and are intended to evolve as new projects join and others conclude.

Data Management plans

Across the projects, the Data Management Plans (DMPs) reveal a strong convergence in strategic approach, reflecting best practices in EU-funded research. All DMP are set as a living document, subject to regular updates and continuous improvement, ensuring that data handling procedures remain aligned with evolving project needs, regulatory requirements, and technological advancements. This iterative governance is supported by clear assignment of roles and responsibilities, with dedicated leaders and teams overseeing data management, ethics, and compliance, and by the allocation of specific resources and review mechanisms.

A central pillar in each plan is the rigorous application of the FAIR principles—making data Findable, Accessible, Interoperable, and Reusable. Each project details structured metadata schemes, versioning strategies, and the use of persistent identifiers (such as DOIs), ensuring that datasets are discoverable and traceable. Data is stored in standardized, shareable formats and deposited in trusted repositories, both for internal collaboration and for public dissemination when appropriate. At the same time, all projects emphasize the importance of privacy and ethics, with robust GDPR compliance, anonymisation of personal data, informed consent procedures, and, where relevant, alignment with the EU AI Act and international standards such as ISO 27001.

Finally, the DMPs share a common lifecycle model for data: from exploration and collection, through preprocessing and normalization, to evaluation, publication, and long-term preservation. Access controls and licensing are clearly defined, balancing open science with necessary restrictions for privacy, intellectual property, or contractual obligations. This harmonized approach not only facilitates collaboration and transparency within each project but also supports interoperability and data reuse across the broader research community, ultimately advancing the goals of safe, ethical, and impactful innovation in urban mobility and road safety.

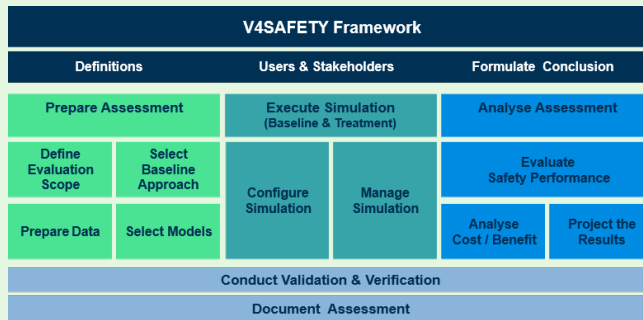
Some of the projects Data Management Plans (initial or final version depending on current project stage) are public materials and can be found in the project libraries. It is the case of [CAMBER](#), [EVENTS](#), [EVORoads](#), [iDRIVING](#), [PHOEBE](#), [SOTERIA](#) and [V4SAFETY](#).

Central Repositories & Homogenised Platforms

Establishing shared repositories and unified data platforms is essential for ensuring consistent storage, streamlined access, and seamless integration of diverse road safety datasets. Evidence across multiple projects demonstrates that centralised architectures significantly strengthen data quality, comparability, and usability. The [SOTERIA](#) project's Safe Mobility Data Space exemplifies this through its large-scale harmonisation of multi-country datasets using automated transformation pipelines, common spatial standards, and the CADaS model, enabling coherent analysis across regions. Similarly, the [FRODDO](#) project uses a Digital Twin-based Data Fusion Bus as a homogenised layer to integrate high-resolution mobility, behavioural, and infrastructure data while ensuring FAIR-compliant governance. [EVORoads](#) and [PHOEBE](#) reinforce this need by advocating for EU-wide open data sharing platforms that support secure exchange of behavioural, crash, and mobility data. Establishing such shared platforms is therefore crucial to enabling robust analytics, supporting evidence-based policy, and advancing the broader European road safety ecosystem.

The [CAMBER](#) project contributes to improving data interoperability and reliability across sectors by promoting the adoption of common standards for the collection and management of infrastructure, traffic, and road safety data. By enhancing consistency across authorities and systems, [CAMBER](#) enables the seamless integration of information from multiple data sources, including RAM, RISM, and CCAM technologies.

V4SAFETY introduces a comprehensive framework for prospective safety assessment



The framework improves traffic safety by using simulation to evaluate safety measures, focusing on interactions between vehicles and vulnerable road users. It helps stakeholders understand the long-term impact and cost-effectiveness of interventions in evolving traffic and automation systems. A key element is identifying- and collecting the input data needed to run simulations and interpret results.

The framework distinguishes between primary data - directly observed real world information- and secondary data, which is processed or derived from primary sources. Primary data includes five categories: real world crash reports, continuous or event based real world data collection, controlled environment incidents, real world exposure data, and real-world infrastructure information. Examples of event-based data include driving manoeuvres, near crashes, crashes, eyegance patterns, or phone use events. Because data is essential for safety assessment, the project gave an overview of the most relevant considerations and recommendations to help users (i.e. a large variety of stakeholders from industry, governmental organisations, and academia and research institutes) evaluate the processes for data source identification and suitability determination. This not only shows the large need for data but also emphasises that data and data sources need to be documented well to be able to use the data correctly and trust the results that are based on the selected data.

SOTERIA's Safe Mobility Data Space

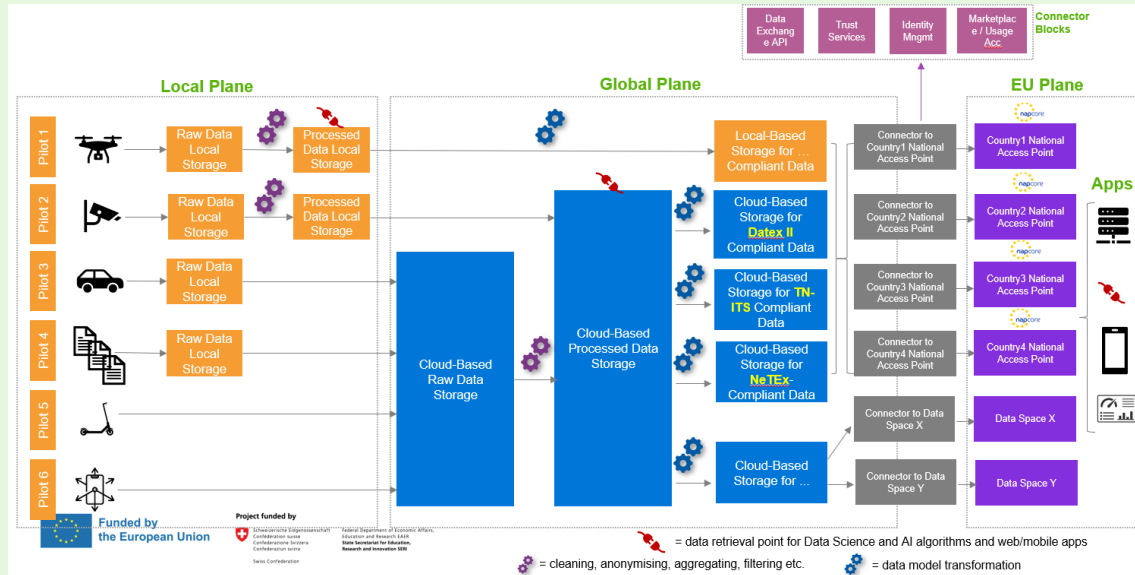
SOTERIA Safe Mobility Data Space (SMDS) is designed to support harmonized, interoperable and governance-ready management of road safety related data across multiple European pilot sites. The SMDS acts as the main data management backbone of the project, enabling the ingestion, documentation, transformation and harmonization of heterogenous datasets related to crashes, infrastructure changes, mobility and VRUs. It is designed according to the latest IDSA standards and is also aligned with FAIR principles to allow data reusability across analytics and decision-support tools such as SOTERIA's safety intelligence service layer.

A key contribution of the SMDS is the CADaS model of the CARE database. Through automated transformation pipelines, crash data originated from different countries and reporting schemes are mapped to a common semantic structure to include consistent definitions of accident types, injury severity, road user categories, environmental conditions and geospatial references.

To increase usability, ease of access and transparency, the SMDS integrates both Eclipse based data space connectors and a CKAN-based data catalogue that provides a user-friendly web interface for dataset publication, discovery, and metadata management, allowing both technical and non-technical partners to upload, document and access datasets through role-based permissions, without requiring direct interaction with backend services. By combining robust backend interoperability with an accessible front-end interface, the SMDS bridges the gap between technical data infrastructures and collaborative research workflows, demonstrating how shared, standardised European Data Spaces can enable advanced analytics and evidence-based road safety interventions at a European scale.

EVORoads Federated Data Space Architecture

The **EVORoads** project adopts a federated data space approach to support road safety assessment and evidence-based decision-making across Europe. Instead of a centralised repository, it uses a three-plane architecture (local, cloud and European) to balance local data sovereignty, scalable analytics and cross-border interoperability.



At the local plane, data are generated and processed close to their source, including roadside sensors, connected vehicles, infrastructure monitoring and citizen reporting tools. Edge processing ensures data quality, anonymisation and GDPR compliance, while local authorities retain control over data ownership, access and operational use.

The cloud plane forms the integration and analytics backbone of **EVORoads**, ingesting heterogeneous data via APIs and event-driven pipelines. Using shared semantic models, it harmonises infrastructure, traffic and environmental data and supports scalable services such as KPI computation, risk assessment and digital twins. It complements local systems by enabling comparative and longitudinal analysis while preserving decentralised control.

At the European plane, **EVORoads** aligns with the emerging European Mobility Data Space and deployEMDS governance principles. Selected datasets and indicators are made discoverable through interoperable catalogues based on Mobility DCAT-AP, extended to better represent dynamic, safety-related data and KPIs, enabling cross-project reuse without centralised data replication.

Compatibility with International Data Spaces concepts supports controlled interaction with external stakeholders and third-party data spaces. Overall, the **EVORoads** data space separates operational data handling, analytical integration and European-level exposure, offering a practical implementation tailored to the challenges of safety-critical data across multiple governance levels.

Digital Twin Architectures with Interoperability Models

Digital Twin frameworks play a pivotal role in enabling the real-time integration, harmonisation, and simulation of heterogeneous datasets for advanced analysis and decision support. The interoperability is achieved through a combination of FAIR-aligned metadata practices, open standards such as ITS-G5 and OpenStreetMap, common licensing frameworks, and structured governance measures, ensuring that data from multiple domains can be securely combined and operationalised. While **V4SAFETY**, **SOTERIA** and **PHOEBE** do not employ full Digital Twin systems, they introduce compatible interoperability mechanisms,

including harmonised data models, standardised geospatial references, automated transformation pipelines, integrated behavioural-simulation frameworks, and traffic optimization to support decision-support tools within a unified operational ecosystem.

All collectively support the development of interoperable digital infrastructures across road-safety and CCAM ecosystems. While **CAMBER** and **iDRIVING** are still in early development stages, both projects aim to use Digital Twins as flexible, modular frameworks that integrate diverse real-world, sensor, and simulation data to support better decision-making in road safety and infrastructure management. They share a focus on adaptable, configurable systems that enhance interoperability across stakeholders, and both aim to improve safety outcomes through forward-looking assessment methods, enabling more accurate evaluation of risks, conditions, and future scenarios. **CAMBER** views digital twins as flexible data integration and evaluation frameworks, to be tailored to the local needs of Road Asset Management and Road Safety Management and thus need to be modular and highly configurable. Within **iDRIVING** project, Digital Twins aim to design dynamic virtual representations of road infrastructure and traffic environments, continuously updated through real-time sensors, historical records, and simulation data.

These examples underline the importance of Digital Twin architectures and interoperability protocols as enablers of scalable, adaptive, and evidence-driven safety management. The **FRODDO** project demonstrated a convincing and strong case of digital twins' implementation.

EVORoads Digital Twin - Data Space Coupling

Within **EVORoads**, the digital twin is conceived as an analytical construct that is tightly coupled with the project's federated data space, rather than as a standalone simulation or visualisation artefact. Its primary role is to provide a coherent and evolving representation of road infrastructure and safety conditions by drawing directly on the data streams exposed through the data space architecture.

The digital twin ingests both static and dynamic datasets, including infrastructure inventories, monitoring outputs, connected vehicle observations and environmental data. These inputs are provided by a range of monitoring and sensing technologies deployed across pilot contexts. Examples include AI-based tools that detect pavement distress or signage degradation from roadside sensors, vehicles or aerial platforms, as well as low-cost smart road equipment capable of reporting infrastructure status and anomalies in near real time. Before being integrated into the digital twin, these heterogeneous data streams are harmonised through the data space using common semantic descriptions and metadata.

This coupling is particularly visible in the computation of safety indicators. KPIs are derived within the digital twin using harmonised datasets sourced from multiple technologies, ensuring consistency between data provenance, analytical models and resulting indicators. Conversely, aggregated indicators, inferred risk states or infrastructure condition assessments generated by the digital twin can be exposed back through the data space as derived data products, subject to governance and access constraints.

By aligning the digital twin with the data space in this way, **EVORoads** treats data management and analytical representation as interdependent. The digital twin functions as a data-driven analytical layer that both relies on, and contributes to, the wider data space, supporting coordinated safety assessment across local, cloud and European contexts.

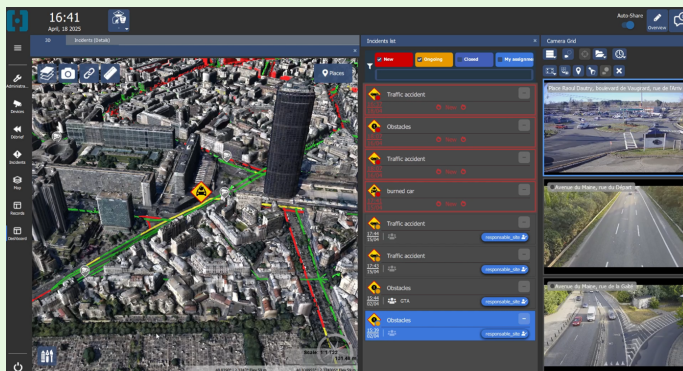
iDRIVING project digital twin framework

In **iDRIVING**, the Digital Twin framework integrates and harmonises real time and historical data from UAVs, ground cameras, OSM, and city archives into a single interoperable environment. Realtime streams support predictive insights, while historical and OSM data are used for calibration and simulation setup. The system uses a modular architecture, linking traffic, maintenance, and visualisation modules through shared interfaces, to transform multidomain data into actionable indicators.

Rather than relying on a single system component, **iDRIVING** adopts a modular architecture in which traffic, maintenance, and visualization modules interact through shared interfaces and governance mechanisms. This structure enables multi-domain data to be securely combined and transformed into actionable indicators, supporting predictive maintenance, proactive safety management, and adaptive traffic control. As a result, the **iDRIVING** Digital Twin functions as an operational decision-support ecosystem that bridges heterogeneous data streams and analytical modelling with real-time control centre decision-making tools.

FRODDO project digital twin

The **FRODDO** Digital Twin (DT) is a federated, AI-assisted cyber-physical platform designed to provide a synchronised, multi-scale virtual representation of CCAM ecosystems, integrating vehicles, infrastructure, traffic systems, and environmental context. It is architected in alignment with the Digital Twin Consortium Platform Stack and implemented on top of the modular Crimson DT platform, ensuring interoperability, scalability, and security. The DT ingests heterogeneous real-time and historical data streams through standardized interoperability services and a dedicated Data Fusion Bus (DFB), which aggregates, harmonises, and semantically aligns inputs from sensors, simulators, traffic management systems, and external data sources. The internal data model is structured around three core abstractions: (i) virtual replicas and temporal histories of physical assets and entities, (ii) a multi-resolution geospatial model supporting 2D/3D representations from regional to indoor scale, and (iii) an interoperability model enabling seamless data exchange via APIs and standardised formats. This design allows continuous synchronisation between the physical and digital domains while preserving data provenance, access control, and compliance with security and privacy constraints.



On top of this core, the **FRODDO** DT integrates advanced simulation, predictive analytics, and decision-support capabilities to support transport management and ODD continuity. AI-driven traffic management simulations enable the evaluation of alternative strategies (e.g., adaptive signal control, dynamic lane management, CAV penetration scenarios) within high-fidelity virtual environments, with results

visualised geospatially and benchmarked through KPI dashboards. A Complex Event Processing–based Early Warning System continuously analyses live data streams to detect hazards and anomalous conditions, while a Decision Support System embeds predictive ML models for traffic, public transport demand, passenger trust, and discomfort. These components are exposed through a secured, role-based user interface that supports real-time monitoring, incident management, historical analysis, and what-if scenario exploration, enabling human-in-the-loop operational decision-making for authorities and road operators across federated organisational boundaries.

FAIR Principles & GDPR Compliance

Ensuring ethical, transparent, and reusable data practices requires the integration of FAIR principles (Findable, Accessible, Interoperable, Reusable) with strict GDPR-compliant governance frameworks. The **FRODDO** project exemplifies this approach through a comprehensive privacy-protection structure, including DPIAs, anonymisation pipelines, a Data Protection Officer, and an Ethics and Legal Committee, ensuring that sensitive data such as video, physiological signals, and behavioural observations are handled responsibly while remaining interoperable and reusable for research and decision support. Complementary efforts in projects such as **SOTERIA** and **PHOEBE** demonstrate GDPR-driven controls through restricted data access, legally mandated sharing agreements, anonymised crash and behavioural datasets, and privacy-preserving storage procedures. While not always explicitly labelled as FAIR, their harmonisation strategies, use of standardised data models, and structured metadata practices support the broader FAIR objectives of consistency, comparability, and long-term reusability. Similarly, the **CAMBER** project establishes clear and robust data-sharing agreements, ensuring full compliance with ethical and GDPR requirements, including data minimisation, security, anonymisation, standardised data formats, and comprehensive data management practices, while also drawing on lessons learned from other relevant projects and respective data protection processes already in place by the experienced data collecting partners.

For the accessibility component, projects are dedicated to ensuring that all data generated is made openly accessible. Data is contributed to a repository where metadata can be accessed, read, and downloaded. Most projects utilize Zenodo as their primary data repository.

FAIR Data Release and Archiving on Zenodo

The European Commission strongly recommends the use of [Zenodo](#) as a trusted repository to support FAIR data sharing across EU projects, as it provides persistent identifiers (DOIs), robust metadata standards, long-term preservation, and open access mechanisms that operationalise the FAIR principles in practice. Through initiatives such as the EU Open Research Repository and the HORIZON-ZEN programme, the Commission has further embedded Zenodo into its Open Science infrastructure, offering beneficiaries a streamlined, compliant, and Commission-endorsed solution for making publicly funded research outputs findable, accessible, interoperable, and reusable

iDRIVING efforts to comply with GDPR

The **iDRIVING** project follows GDPR by testing its detection, simulation, and surveillance tools in controlled environments where informed consent is obtained. Joint controllership agreements define shared data responsibilities, and ethical approvals are secured when required. End user applications collect personal data—such as location—only with explicit consent. A data breach procedure and documentation template are in place for all partners.

Data collection follows GDPR principles of lawfulness, transparency, data minimisation, and purpose limitation. Only data necessary for objectives like traffic monitoring or safety assessment is collected, with retention policies ensuring timely deletion or anonymisation. Security measures such as encryption, access control, and secure storage protect personal data, while documentation and audits ensure accountability and compliance.

CAMBER FAIR principles

The **CAMBER** project applies the FAIR principles by ensuring that datasets are findable through clear descriptions, standard naming conventions, versioning, and DOIs, and by providing public data in open formats such as CSV and XML. Data accessibility is supported through GDPR compliant review processes, controlled partner access, pilot site data sharing, and the use of open repositories like Zenodo, along with common formats and shared scripts. Interoperability is achieved using standard file formats, consistent naming, shared metadata approaches, and harmonised data structures. To ensure reusability, CAMBER uses well structured, fully documented data formats, provides clear guidance for understanding and reuse, and applies open licences that allow broad redistribution and application of the datasets.

Open Standards & Licensing

Adoption of widely recognised standards (e.g., ITS-G5, OpenStreetMap) and Creative Commons licensing promotes interoperability and legal clarity for data sharing. The **FRODDO** project reports the use of standards such as ITS-G5 for V2X communications, OpenStreetMap and USGS terrain data for geospatial referencing, and a broad suite of open, documented formats (e.g., CSV, JSON, MAT, PCAP). Alongside these technical standards, **FRODDO** partners employ Creative Commons licensing—most frequently CC-BY and CC-BY-4.0—together with open-government data licences, thereby ensuring transparent reuse conditions, consistent attribution practices, and legally robust data governance across partners.

In the case of **CAMBER**, the project makes extensive use of available GIS resources (OpenStreetMap) to plan pilot testing and ground truth data collection and discuss data sources and findings. Findings are communicated in open and widely formats as much as possible. Conversely, the processing of sensor data and indeed the sensor data itself, as well as data procured from third parties, present limitations with regards to Licensing and Open Standards.

AI-Based Video Analytics & Telematics Integration

Across the reviewed projects, AI-based video analytics and telematics integration emerged as key enablers for understanding real-world road-user behaviour and supporting advanced safety modelling. The **PHOEBE** project provides the most comprehensive implementation, employing state-of-the-art computer-vision techniques—including YOLO-based object detection, DeepSort tracking, and automated violation identification—to extract behavioural indicators from video and integrate them with telematics, simulation outputs, and infrastructure data for predictive safety assessments. **FRODDO** similarly combines multiple sensing modalities, including high-resolution video, eye-tracking, physiological monitoring, GNSS/INS positioning, and V2X telematics data, all harmonised within a Digital Twin architecture and Data Fusion Bus to enable rich behavioural and mobility insights. **SOTERIA** contributes complementary telematics-style data through multimodal sensors (GPS, environmental sensing, gyroscope) and applies AI-based risk scoring within its Safe Mobility Data Space, while maintaining strict GDPR-compliant governance. **IDRIVING** and other projects also use video or sensor data.

More is to come. **CAMBER** will, through its pilots and interventions, use the AI-based video and telematics analytics. Telematics will be used to address road condition and road user safety, by procuring vehicle dynamics data and safety (harsh-brakings) and road condition events (including feedback by road users with mobile phone support to report problematic infrastructure) both from 3rd party providers and directly through solutions by the project partners. Furthermore, camera-based test cases with AI-based detection are also included in **CAMBER**, using a camera-based safety system close to a school area in one of the project pilots. Together, these initiatives demonstrate growing maturity in combining AI-enabled video understanding with telematics sources to produce scalable, data-driven road-safety intelligence.

PHOEBE AI video analytics to collect behavioural pedestrian data



The **PHOEBE** project explores automated detection of traffic violations in urban environments using video analysis and machine learning. Focusing on eight high-risk locations in Athens, Greece, the team collected 64 hours of video footage during peak and off-peak hours. Advanced algorithms such as YOLOv8 for object detection and ResNet50 for feature

extraction were employed to identify and classify vehicles and pedestrians. Additional techniques, including ReID features, Kalman filters for predictive tracking, and Savitzky-Golay filters for noise reduction, ensured accurate trajectory mapping and minimized data inconsistencies.

The system successfully detected illegal pedestrian crossings, speeding incidents, and calculated time-to-collision metrics to assess potential risks. Homograph transformations enabled conversion to a top-down view for precise speed and trajectory analysis within defined regions of interest. Despite challenges like ground-level camera placement and occlusions, the approach demonstrated strong potential for improving urban traffic management. Future enhancements include real-time processing, elevated camera setups, and calibration improvements to refine accuracy. This study highlights the value of integrating computer vision and machine learning for proactive road safety interventions in complex urban settings.

Policy recommendations from the EU Road Safety Cluster Projects on Road Safety Data

To achieve the Vision Zero goal of no road deaths by 2050, policy frameworks must transition from recording trauma to proactively managing risk through integrated data ecosystems. The following recommendations, derived from the collective experience of the EU Road Safety Cluster, provide a strategic foundation for this transformation. Central to these solutions is the evolution from traditional, retrospective crash analysis to a proactive Safe System approach, which seeks to eliminate serious injuries by anticipating human error and managing physical forces through advanced sensing and speed management.

Furthermore, these projects leverage emerging technological architectures, such as Digital Twin frameworks and AI-based video analytics, to enable real-time simulation and the integration of heterogeneous datasets, including fused police, hospital, and telematics data. By adopting sophisticated methodologies like probabilistic record linkage and advanced machine learning for surrogate safety measures (SSMs), these initiatives address long-standing issues of under-reporting and data quality.

Based on the evidence synthesised from the Cluster projects, the following detailed policy actions are recommended:

- **Institutionalising the Proactive Safe System Approach:** Policy frameworks must facilitate a structural move away from reactive "blackspot" management toward a preventative design model. This requires the mandatory adoption of proactive and holistic analyses methods and inclusion of SSMs as valid performance indicators to identify system failures before fatalities occur.
- **Strengthening Data Governance and Interoperability:** Publicly-funded road safety research must mandate compliance with FAIR principles to break down existing "data silos".

- **Address Under-reporting:** Acknowledging that police-reported data significantly under-represents VRUs and minor injuries, policy should mandate the use of probabilistic record linkage across police, hospital, and insurance datasets. This is essential for an accurate burden assessment and the subsequent prioritisation of interventions for high-risk cohorts.
- **Deploying Emerging Technological Architectures:** Support the integration of Digital Twin frameworks and AI-based analytics into urban planning and traffic management. Real-time data from Advanced Driver Assistance Systems (ADAS) and smartphone telematics should be harnessed to provide dynamic feedback and identify high-risk spatial units proactively.
- **Balancing Privacy with Research Requirements:** To mitigate GDPR restrictions, policy must support the development of automated anonymisation protocols. This ensures high-resolution data remains available for safety analysis without compromising personal privacy.
- **Ensuring transparency in reporting safety critical incidents:** Establish a methodology and governance mechanism for monitoring and reporting vehicle data related to safety-critical situations (i.e., crashes, near-misses, collision mitigation actions, etc.) to support data transparency and to foster broader safety innovations. It is recommended to study how data resulting from for instance Event Data Recorders (EDR), Data Storage Systems for Automated Driving (DSSAD), or In-Service Monitoring and Reporting (ISMR) can be made available for research, development and innovation purposes.

Conclusions

The collective work of the EU Road Safety Cluster demonstrates that road safety can only advance meaningfully when data ecosystems become more integrated, interoperable, and ethically governed. Across the projects, it is evident that fragmented datasets, inconsistent definitions, and limited accessibility continue to restrict the development of predictive and proactive safety interventions. Yet the initiatives also show that emerging technologies, ranging from AI video analytics to Digital Twins and advanced telematics, are already reshaping research, enabling richer insights into behaviour, infrastructure conditions, as well as risk patterns.

At the same time, the projects reveal that addressing data gaps requires both technological innovation and institutional reforms. FAIR-aligned Data Management Plans, secure shared repositories, harmonised data models, and clear governance protocols are central to ensuring data remains usable across borders, systems, and research domains. As projects experiment with multimodal sensing, federated architectures, and GDPR-compliant data handling, they collectively demonstrate workable pathways toward more robust, privacy-conscious and scalable data infrastructures for road safety and CCAM systems.

Ultimately, the Cluster's findings underline that achieving Europe's Vision Zero goals will depend on a shift from reactive crash reporting to proactive, evidence-based safety management. Policymakers, industry, and researchers must embrace harmonisation, strengthen accountability frameworks, and invest in shared digital infrastructures that support simulation, prediction, and behavioural insights. By institutionalising the Safe System approach and promoting transparent, high-quality data ecosystems, Europe can significantly accelerate its transition toward safer, more resilient, and user-centred mobility.

Acknowledgments

The cluster projects have received funding from several funding frameworks, including European Union's Horizon Europe research and innovation programme, including **PHOEBE** (GA 101076963), **Ai4CCAM** (GA 101076911), **CAMBER** (GA 101146800), **EVENTS** (GA 101069614), **EVORoads** (GA 101147850), **FRODDO** (GA 101147819), **iDRIVING** (GA 101147004), **PROTACT-US** (GA 101147445), **SOTERIA** (GA 101077433), and **V4SAFETY** (GA 101075068). Projects supported through UK Research and Innovation (UKRI) include **PHOEBE** (GA 10038897) and **CAMBER** (GA 10139277). The **SAFETeen** Project is Co-funded by the Erasmus+ Programme of the European Union. (GA 2023-1-EL01-KA220-SCH-000160385). The **EVORoads** project is Co-funded by the Swiss

Confederation (Federal Department of Economic Affairs, Education and Research EAER, State Secretariat for Education, Research and Innovation SERI).

The European Commission's support for the activities of the EU Road Safety cluster, which also includes the production of this publication, does not constitute an endorsement of the contents of this document, as it solely reflects the views of the authors. Thus, the European Commission cannot be held responsible for any information shared in the joint document.

All authors listed in this publication are direct contributors and are named in their capacity as representatives of their respective project consortium.

References

- Abay, K. A. (2015). "Investigating the nature and impact of reporting bias in road crash data." *Transportation Research Part A: Policy and Practice* **71**: 31-45.
- Abdel-Aty, M., J. Lee, C. Siddiqui and K. Choi (2013). "Geographical unit based analysis in the context of transportation safety planning." *Transportation Research Part A: Policy and Practice* **49**: 62-75.
- Ali, Y., F. Hussain and M. M. Haque (2024). "Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review." *Accident Analysis & Prevention* **194**: 107378.
- Alzaffin, K., S.-A. Kaye, A. Watson and M. M. Haque (2023). "A data fusion approach of police-hospital linked data to examine injury severity of motor vehicle crashes." *Accident Analysis & Prevention* **179**: 106897.
- Amoros, E., J.-L. Martin and B. Laumon (2006). "Under-reporting of road crash casualties in France." *Accident Analysis & Prevention* **38**(4): 627-635.
- Benavente, M., M. A. Knodler and H. Rothenberg (2006). "Case Study Assessment of Crash Data Challenges: Linking Databases for Analysis of Injury Specifics and Crash Compatibility Issues." *Transportation Research Record* **1953**(1): 180-186.
- Brubacher, J. R., H. Chan and S. Erdelyi (2019). "Injury severity in police collision reports correlates poorly with requirement for hospital admission." *Journal of Transport & Health* **14**: 100606.
- Ding, H., Z. Liu, H. Fu, X. Fu, T. Chen and J. Zhao (2025). "Can AV crash datasets provide more insight if missing information is supplemented? Employing Generative Adversarial Imputation Networks to Tackle Data Quality Issues." *Transportation Research Part C: Emerging Technologies* **176**: 105154.
- Dong, N., F. Meng, J. Zhang, S. C. Wong and P. Xu (2020). "Towards activity-based exposure measures in spatial analysis of pedestrian-motor vehicle crashes." *Accident Analysis & Prevention* **148**: 105777.
- Elvik, R. and A. Mysen (1999). "Incomplete accident reporting: meta-analysis of studies made in 13 countries." *Transportation research record* **1665**(1): 133-140.
- Eskandari Torbaghan, M., M. Sasidharan, L. Reardon and L. C. W. Muchanga-Hvelplund (2022). "Understanding the potential of emerging digital technologies for improving road safety." *Accident Analysis & Prevention* **166**: 106543.
- Ferenchak, N. N. and R. B. Osofsky (2022). "Police-reported pedestrian crash matching and injury severity misclassification by body region in New Mexico, USA." *Accident Analysis & Prevention* **167**: 106573.
- Hosseinzadeh, A., A. Karimpour, R. Kluger and R. Orthober (2022). "Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records." *Journal of Safety Research* **81**: 21-35.
- Imprialou, M. and M. Quddus (2019). "Crash data quality for road safety research: Current state and future directions." *Accident Analysis & Prevention* **130**: 84-90.
- Kalkman, S., J. van Delden, A. Banerjee, B. Tyl, M. Mostert and G. van Thiel (2022). "Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence." *Journal of Medical Ethics* **48**(1): 3-13.
- Kamaluddin, N. A., C. S. Andersen, M. K. Larsen, K. R. Meltofte and A. Várhelyi (2018). "Self-reporting traffic crashes – a systematic literature review." *European Transport Research Review* **10**(2): 26.
- Karimi, S., A. Hosseinzadeh, R. Kluger, T. Wang, R. Souleyrette and E. Harding (2024). "A systematic review and meta-analysis of data linkage between motor vehicle crash and hospital-based datasets." *Accident Analysis & Prevention* **197**: 107461.
- Lord, D. and F. Mannering (2010). "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives." *Transportation Research Part A: Policy and Practice* **44**(5): 291-305.
- Lord, D. and S. Washington (2018). *Safe Mobility: Challenges, Methodology and Solutions*, Emerald Publishing Limited.
- Mannering, F., C. R. Bhat, V. Shankar and M. Abdel-Aty (2020). "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis." *Analytic Methods in Accident Research* **25**.

- Mannering, F. and C. R. Bhat (2014). "Analytic methods in accident research: Methodological frontier and future directions." *Analytic Methods in Accident Research* **1**: 1-22.
- Mitchell, R. J., C. M. Cameron and M. R. Bambach (2014). "Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential." *Australian and New Zealand Journal of Public Health* **38**(3): 275-280.
- Sander, U., Ek, P., Sander, D., Breunig, S., Bärghman, J., Menzel, T., . . . Hay, J. (2024). Data Sources for Baseline Generation - Overview, Grading, and Recommendations. Deliverable D4.1 of the Horizon Europe project V4SAFETY.
- Short, J. and B. Caulfield (2016). "Record linkage for road traffic injuries in Ireland using police hospital and injury claims data." *Journal of Safety Research* **58**: 1-14.
- Soltani, A., J. Edward Harrison, C. Ryder, J. Flavel and A. Watson (2024). "Police and hospital data linkage for traffic injury surveillance: A systematic review." *Accident Analysis & Prevention* **197**: 107426.
- Tarko, A. and M. S. Azam (2011). "Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data." *Accident Analysis & Prevention* **43**(5): 1689-1695.
- Wang, C., M. A. Quddus and S. G. Ison (2013). "The effect of traffic and road characteristics on road safety: A review and future research direction." *Safety Science* **57**: 264-275.
- Watson, A., B. Watson and K. Vallmuur (2015). "Estimating under-reporting of road crash injuries to police using multiple linked data collections." *Accident Analysis & Prevention* **83**: 18-25.
- Wegman, F. (2017). "The future of road safety: A worldwide perspective." *IATSS Research* **40**(2): 66-71.
- Wen, X., Y. Xie, L. Jiang, Z. Pu and T. Ge (2021). "Applications of machine learning methods in traffic crash severity modelling: current status and future directions." *Transport Reviews* **41**(6): 855-879.
- Xu, P., H. Huang and N. Dong (2018). "The modifiable areal unit problem in traffic safety: Basic issue, potential solutions and future research." *Journal of Traffic and Transportation Engineering* **5**(1): 73-82.
- Yannis, G., E. Papadimitriou, A. Chaziris and J. Broughton (2014). "Modeling road accident injury under-reporting in Europe." *European Transport Research Review* **6**(4): 425-438.
- Ye, F. and D. Lord (2011). "Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit, and Mixed Logit." *Transportation Research Record* **2241**(1): 51-58.
- Ziakopoulos, A. (2024). "Analysis of harsh braking and harsh acceleration occurrence via explainable imbalanced machine learning using high-resolution smartphone telematics and traffic data." *Accident Analysis & Prevention* **207**: 107743.
- Ziakopoulos, A. and G. Yannis (2020). "A review of spatial approaches in road safety." *Accident Analysis & Prevention* **135**: 105323.