Data Handling

Good Practices in the Context of Naturalistic Driving Studies

Al Haddad, Christelle; Alam, Md Rakibul; Papadimitriou, Eleonora; Brijs, Tom; Antoniou, Constantinos

DOI
10.1016/j.trpro.2024.02.013

Publication date
2024

Document Version
Final published version

Published in
Transportation Research Procedia

Citation (APA)
Al Haddad, C., Alam, M. R., Papadimitriou, E., Brijs, T., & Antoniou, C. (2024). Data Handling: Good Practices in the Context of Naturalistic Driving Studies. *Transportation Research Procedia*, *78*, 95-102. https://doi.org/10.1016/j.trpro.2024.02.013

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 78 (2024) 95–102

Transportation Research Procedia

www.elsevier.com/locate/procedia

25th Euro Working Group on Transportation Meeting (EWGT 2023)

Data Handling: Good Practices in the Context of Naturalistic Driving Studies

Christelle Al Haddad[a,]*, Md Rakibul Alam[a], Eleonora Papadimitriou[b], Tom Brijs[c], Constantinos Antoniou[a]

[a] *Chair of Transportation System Engineering, Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany*
[b] *Delft University of Technology, Delft, the Netherlands*
[c] *School for Transportation Sciences, Transportation Research Institute, UHasselt, Diepenbeek, Belgium*

Abstract

Naturalistic driving studies (NDS) have recently gained attention as a way of instrumenting vehicles in an unobtrusive way and collecting driving data over long periods of time. Aiming at eventually modeling driving behavior, NDS are often a part of larger scale studies. These studies involve several stakeholders who are responsible for different components of the data collection and analysis, and thus are inevitably confronted with challenges in the data management pipeline. The aim of this paper is to develop standard protocols that could be used as guidelines for data handling in the context of NDS. In the development of these protocols, we first review data handling strategies used in previous studies, focusing on data collection, preparation, storage, as well as ethical and legal considerations. This review helps us draw lessons, based on which methods are developed to answer the gaps and challenges arising from handling NDS data. We then introduce a case study, the i–DREAMS project, to show the applicability of the data handling framework. Finally, we showcase standard protocols for data handling, that could serve as data handling guidelines for future studies.

Keywords: Naturalistic driving studies; data handling; standard protocols; safer roads; efficient mobility; transportation technology.

---

1. Introduction

Road crashes take millions of lives across the world every year and as a result, understanding factors contributing to these crashes has been at the forefront of road safety research. These factors may arise from distinct sources of risk such as vehicle factors, environmental factors, and behavioral (driver-related) factors (Afghari et al., 2018). Among these factors, driver behavior and human factors have been identified as primary in crash causation (Afghari,

---

* Corresponding author.
E-mail address: christelle.haddad@tum.de

2352-1465 © 2024 The Authors. Published by ELSEVIER B.V.

10.1016/j.trpro.2024.02.013

2019), and therefore key to model driving behavior. Several methods have been used to model human factors, such as investigating road crashes and relating human factors with various events, simulating the driving environment and research safety-critical scenarios that would be impossible to otherwise study in real road conditions, but also by conducting real-road experiments or driving tests. The latter have taken different forms including field operational tests, where road tests are administered at specific road sections, in a rather more confined environment, and more recently through naturalistic driving studies (NDS).

As their name indicate, these studies are conducted in a natural unobtrusive way in which the participants drive as they normally do, without being asked to drive specific roads, or change their driving patterns. The only difference is that their vehicles are instrumented with data collection devices. With advances of technology and sensory equipment, NDS are increasing in popularity, but also in challenges. Driving for longer periods of time inevitably leads to large amounts of data, and therefore creates challenges in terms of data management, data sharing, and data handling in general. While previous projects and studies have indeed followed some guidelines in handling data, there are currently no comprehensive protocols or guidelines for handling data in NDS.

To address this gap, this paper aims to contribute to research, and particularly to future NDS projects, by reviewing previous studies and focusing on relevant aspects of data collection, preparation, storage, as well as other ethical and legal considerations. Based on the findings and lessons learned, a methodology for data handling is developed, and is then applied to a case study, for which standard protocols for data handling are defined. The authors argue that such protocols could be dynamic in that they may be updated along the course of a project, and serve as checklists, for quality control, wherein the defined goals and followed guidelines could be cross-checked for validity purposes. The contributions of this research can be accordingly summarized as:

- Review of previous NDS focusing on relevant aspects of data collection, preparation, storage, and ethical and legal considerations. The outcome would be a summary of lessons learned and an identification of existing gaps.
- Development of a methodology for data handling in the context of NDS, focusing on the protocols for the above–mentioned aspects.
- Application of the methodology to an on–going large–scale multi–modal European NDS.

2. Previous NDS: Lessons Learned and Existing Gaps

Several components are crucial for handling data in NDS. The FESTA handbook (Section 7) defines the guidelines for data acquisition, including storage and analysis tools, emphasizing the importance of laws and regulations in such protocols (FOT-Net and CARTRE, 2018). Based on these recommendations, the review of previous NDS highlights findings and lessons learned from previous projects, focusing on various aspects of data handling, paving the way to the methods proposed in this paper. The reviewed projects are: 100-Car Naturalistic Driving Study (Dingus et al., 2006), SeMiFOT (Victor et al., 2010), INTERACTION (FOT-Net WIKI, 2015), 2BeSAFE (2BESAFE, 2012), OBMS (Federal Motor Carrier Safety Administration, 2016), UDRIVE (UDRIVE, 2017), Canada NDS & Canada Truck NDS (Klauer et al., 2018), Track & Know (Track Know, 2021). For each of these projects, we reviewed previous practices for data collection, data preparation, data storage, and legal and ethical issues related to NDS data. In terms of data collection, the collected data was classified by transport mode, sensor frequency, and data type (in–vehicle data, survey data, or sensitive data). When it comes to data preparation practices, the focus was on the different pre–processing, enrichment, and synchronization techniques. Going further to the data storage methods, a distinction was made between on–board storage, and remote storage, including both online and offline methods. Finally, when considering legal and ethical aspects, the most prominent methods were described including legal protocols, anonymization, disclosure permissions, and access restrictions.

Accordingly, based on the gaps, we mapped out the lessons learned into standard protocols which could serve as a blueprint of methods to be followed in the implementation of data handling for these studies. A framework for data handling is presented in Figure 1.

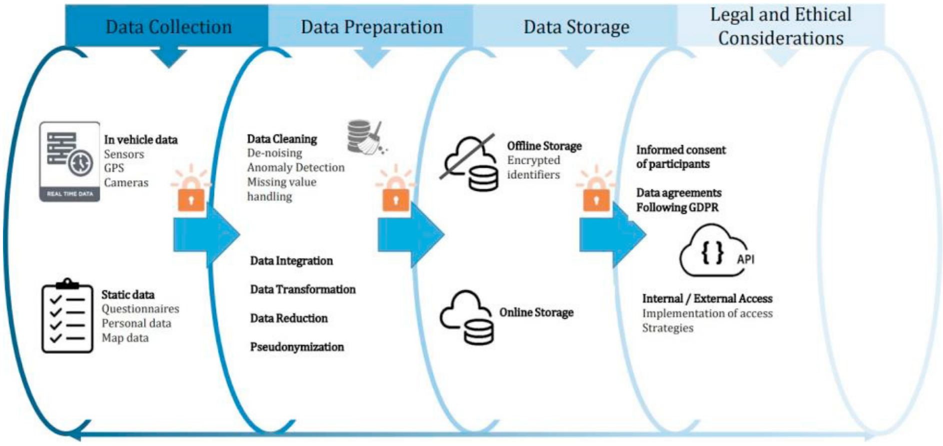

Fig. 1. Generalized data handling framework (own illustration)

3. Case Study: the i-DREAMS Project

In this study, we applied the developed framework to a case study of a European NDS: i–DREAMS. The project aimed to develop, test, and validate a context aware safe driving platform, taking into account driver-related background factors, risk related real-time physiological indicators and driving task complexity, to determine if a driver is within the boundaries of safe driving operation. This NDS collected data from various sources (in–vehicle sensors, and survey and driver–related data), across five EU countries (Germany, the UK, Portugal, Belgium, and Greece), and four modes (cars, trucks, buses, and rail). Accordingly, it resulted in huge amounts of highly heterogeneous data, which need to be adequately handled. In this NDS, various partners played different roles in the data collection and processing pipeline. Project partners can be classified as follows:

- Technology providers: in i-DREAMS, these are CardioID[1], OSeven[2] , and DriveSimSolutions[3]. They provided the data collection equipment for the different countries and modes that are part of the i-DREAMS project. Particularly, CardioID provided the sensory equipment, OSeven developed the android app, DSS built the simulator and coded the scenarios for various trial partners. These partners (technology providers) needed to therefore ensure proper and consistent data collection and make it accessible to the rest of the partners. For field trials, this happened automatically from the technology providers to their servers. For the simulator trials, the data collected was logged and stored locally in the simulator PC.
- Trial partners: they are the simulator and field trial partners, and were responsible for organizing the experiments at their premises including the collecting of questionnaire data, running the experiments, and managing the logistical aspects at their own premises. The different experiments (both simulator and field or on-road experiments) covered the different countries (Belgium, Germany, Greece, Portugal, the UK) and modes (cars, buses, trucks, and rail) that were part of the i-DREAMS project.
- Data processors. They are partners who contributed to data analysis and processing. They have access to the data and test various hypotheses derived from the research questions.

The collected data can be divided into two broad categories of in-vehicle data and survey data. These came from several sources including in-vehicle instruments and technologies, smartphone applications, and driving simulator

---

[1] https://www.cardio-id.com/
[2] https://www.oseven.io
[3] https://drivesimsolutions.com/

technologies. The survey data came from various questionnaires and were collected at different points of the experiments timeline (Pilkington-Cheney et al., 2020). The in-vehicle data collected from in-vehicle instruments and technologies were mostly vehicle-related driving characteristics (e.g., headway, acceleration, lateral position, etc.), driving environment (e.g., time of the day, weather), and driver attributes (e.g., heart rate variability, mobile phone use, etc.).

The devices collecting such data were connected to a gateway (provided by CardioID) which gathered and centralized information from other components and handled data connectivity and transmission; this component was constant across partners. Additionally, the in-vehicle technologies also provided video data generated from a dashcam. The primary objective of collecting such video data was to visualize the real scenario on road, and to understand why a warning (i.e., event) was generated by the system. Moreover, data were collected from a smartphone application that also monitored and collected driving behavior of individuals using a variety of parameters. The app was also part of the data collection equipment used by drivers recruited for this project.

Data collected in-vehicle includes date and time, GPS data, angles formed by the local axes of the phone to the North and horizontal planes, rate of change of these angles, accelerometer data, gyroscope data, activity data (walking, stopping, driving), screen state (for mobile use), smartphone device data, while processed trip data from this app includes (but is not limited to) number of trips, distance travelled, trip duration, number of harsh brakes, number of harsh accelerations, driving over the speed limit, average speed, mobile phone use, and distance travelled.

4. Standard Protocols for Data Handling

4.1. Protocols for Data Collection

For country-specific trials, local partners from each country were responsible for the logistics of setting up the scenarios, leading to the collected data. However, data acquisition should be done through the same mechanisms (servers, communication protocol, code etc. should be similar, if not same) to ensure consistency of processes and quality of data, even for country-specific scenarios. This is ensured by having common technology providers, who deliver the hardware equipment for the in-vehicle data collection. The frequency of collection should be decided a priori, given the fact that each sensor has a different frequency rate, and each sample has an associated time stamp for appropriate synchronization. Each data-collecting system should be conceptually tied to a vehicle, not a specific driver. Data are acquired within a trip session, which is defined from the moment the vehicle is turned on until it is turned off. A grace period (5 minutes) during which a quick turn off and back on should be considered as the same trip.

4.2. Protocols for Data Preparation

Data pre-processing should be primarily done locally at the gateway and in the tech partners databases. The pre-processing may include:

- Handling missing data: with sensor failure, a trigger and alarm can be sent to the driver to ensure that nothing was disconnected (equipment in-vehicle). For real-time interventions for which input data is missing, the algorithms have a procedure to deal with this, by data interpolation, using the last known value or default value. For communication failure however, data are logged so off-line synchronization is possible even without any real-time communication. Finally, missing data can occur by the non-collaboration of the driver.
- Ensuring temporal order in case of time-series data.
- Handling the time zone information carefully.
- Rectifying incorrect GPS data caused by reporting incorrect latitudes and longitudes when there are momentary losses of GPS signals. A filtering procedure may be implemented to remove these positional jumps. Moreover, raw GPS signals could be better managed when cleaned and simplified, using for instance the Ramer–Douglas–Peucker algorithm (Muckell et al., 2010).
- Pre-processing video data in a way to reduce data volume without compromising the quality of the videos. Metadata of the videos (event, timestamps, trip info etc.) should also be attached with each video for ease of

future analysis. Video data pre-processing aims at obfuscating sensitive information from videos (e.g., faces and number plates of surrounding vehicles).

- Detecting outliers and anomalies to ensure quality of data. Detection processes should be done at the source of collection when possible.
- Verifying data to minimize errors during the communication process. Such verification may include validation at the end of a trip session, ensuring temporal order of the data points, and verifying that repeating sample points are filtered out.
- Minimizing data loss at the retrieval/upload and verifying that data are consistent before deleting them from the vehicle. In case inconsistencies are identified, the vehicle data logger should be checked as soon as possible so that any issues can be recognized and fixed.
- Deleting vehicle data after the data have been backed-up and verified.
- Providing a description of the data variables (either from driving simulators, instrumented vehicles, or from questionnaires) by the technical partners generating the data. Having an understandable data format ensures consistency, completeness, integrity, and timeliness. Although survey data is static, a good practice would be for the related information to be attached to each instance including data and time of start and end, the unique identifier, and if applicable a reference to the file name and location.

4.3. Protocols for Data Storage

Partners have the freedom of choosing their preferred storage engines (databases, file systems) for local storage facilities. Nevertheless, the data should be automatically stored locally, via automatic transmission (WIFI, wireless, Bluetooth). Data can be stored in two types: onboard and remote storage (offline and online). Offline refers to storage systems which are not accessible through standard API to external world (other partners and/or third parties). Online storages refer to storage systems which are accessible through standard API to the external world (other partners and/or third parties). This may also include third-party cloud storage. Before being uploaded to the cloud, data needs to be pseudonymized. Data storage type is relevant in terms of data bandwidth (e.g., in the vehicle, the data are sampled from sensors at a very high rate, but usually only a portion of it is uploaded for analysis, or videos are continuously recorded but only a buffer is kept and stored whenever an event takes place, etc), but also in terms of sensitive data, e.g., in i-DREAMS electrocardiogram (ECG) data is processed locally in the vehicle to compute the Karolinska sleepiness scale (KSS) score (Shahid et al., 2011).

The ECG data are not uploaded to the cloud server as they are too sensitive; only the derived indicators such as KSS score or heart pulse are uploaded. Once the data are uploaded, they are deleted in the vehicle to avoid misuse. To ensure proper handling of the data in offline storages, following requirements should be met:

- Persistence: data should be stored for at least till the end of the experiments.
- Reliability: periodic backups should be taken carefully. Deletion/modification of operations should be handled properly (consistency and validity).
- Availability: data shall be sent to the online data storage by uploading through the available API of the online storage system. Once available in the storage, data should be immediately available to the authorized user, preferably via an application programming interface (API).
- Serviceability: data may not be available up to a certain period during storage server maintenance (server down-time).

After transmission, data are downloaded (from the servers of the data collection equipment providers), and then saved to an online back-end server, which saves different components of the integrated and processed data.

4.4. Protocols for Legal and Ethical Considerations

An important aspect as well is the definition and understanding of risks to participants associated with experiments, which is often an integral part of the obtained ethical approval. Accordingly, trial partners might need to subscribe to (third party of liability) insurances, to guarantee to the participants that damages due to participation in the project would be compensated by the organizer.

In the informed consent forms, participants give data collection partners permission to collect and process their data during experiments, including details of which data are to be shared. Their personal information may then be collected, after which a unique identifier is assigned to the participants which is a cross-reference between the experiment data and their personal data (pseudonymization). The personal data should be encrypted to ensure security, should be placed in an offline file system and have limited access.

Servers and hard drive encryption (following the GDPR recommendation: article 34, recital 83) should ensure that all data (including non-personal) are protected (including local storage in the vehicle), as a mitigation against breaches, even if the data are pseudonymized. After the agreed time after the end of the project, the following procedures can be applied for anonymization and for making the data accessible in an open-source platform according to the project objectives:

- The unique identifier that connects the participants' data and their personal data is to be replaced with a random number. The process would then be irreversible.
- In case the primary data (including location data) relates to the Driver ID, the Driver ID is replaced by a random code for each trip. This process is irreversible and there is (i) no longer any possibility of linking the primary data of the trips (including location data) to the personal data of the driver and (ii) no longer any correlation between the trips of a user which is then anonymized.

Following the above procedures, the data of the driver would be fully anonymized since it is impossible to connect this data with a natural person. To transfer data efficiently, each partner generating data should either provide API access on their own data or upload the data to a back-office server from where other partners can collect the data. If an API is exposed to transfer data from the responsible partners side, an API specification is also expected from the partner. These APIs should also be secured through an authentication mechanism. Transferring data should therefore take place over HTTPS and hence secured with public/private key encryption mechanism.

Similarly, the data backoffice should also provide an API specification listing out how to access data which are available through its API. API specifications remain confidential among consortium partners. To access the data, different user types should be first defined with different rights of access (e.g., superadmin, admin, user etc.). A list of roles shall be made between data access during the project lifetime, and after the project end. No deletion/modification permission is given to any user of the storage; only reading permission would be provided to the appropriate users. Exceptions can take place in extraordinary circumstances and contingent upon approval of the superadmin (consistency and validity). Data access would follow safe protocols with access points encryption, and should be logged to trace back any problems of data leaks.

Personal data shall not be stored longer than necessary (not longer than five years). After the end of the project, an anonymized portion of the data can be made available and offered to third-parties at the end of the project. A summary of the data handling protocols defined above is given in Table 1.

5. Results and Conclusion

Applying the methodology to the presented case study, we develop a set of standard protocols for data handling for NDS, focusing on the aspects of interest, i.e., data collection, data preparation, data storage, and legal and ethical considerations. Highlights of these protocols include the following aspects:

- Partners who collect data are responsible for the proper collection and handling of data. This can be distinguished between partners who provide the technology, partners who conduct experiments, and partners who process the data. The collection should deal with communication issues, loss of signals, at the source of data collection.
- Storage can be distinguished between on-board storage and remote storage (including online and offline storage).
- Personal data should remain where it was locally collected and separated from the rest of the data. Before being uploaded to a central back-end server, it should be at least pseudonymized. The identifier, which is the unique

Table 1. Implementation of previous findings in i-DREAMS

Previous findings	Implementation in iDREAMS	Remarks
Reliability and validity checks	✓	No delete/modify permissions were given to any users of the storage.
Common data–acquisition system	✓	
Minimizing the number of vehicle models	✓	Choosing the vehicles most compatible with the data collection devices.
Centralizing responsibilities for coding, processing, and analysis	✓	
Data pre-processing prior to storage	✓	Done at the gateway and the tech partners' databases.
Advanced video processing techniques		Done to obfuscate sensitive data.
External data sources		Using weather data, roadway geometry, and maps, where possible.
Ease of access of data	✓	Central back-end API.
Systematic back-ups	✓	
Data well defined and understandable	✓	Data management plan.
Video files stored separately but linked with the rest of the data in file management systems		Possible in i-DREAMS.
Transferring the data should be done automatically	✓	Except for the simulator data (stored locally in the simulator PC).
Storing hard copies for manually extracted files like questionnaires and forms		Paper-based questionnaires (consent forms) were backed-up and hard copies were adequately stored.
Ease of access of data	✓	Using the recommended architectures.
Consent of participants	✓	
Data agreements	✓	
Following GDPR	✓	
Data pseudonymisation	✓	
First and last minutes of driving deleted	✓	
Driving across multiple countries	✓	Based on geofencing, the dashcam was disabled from recording in countries where its use is not allowed.
Non-participant driving the vehicles incidentally	✓	Driver identification at the beginning of each trip. If participant not identified, recording stopped.
Data use after project lifetime	✓	Defined within national ethical and DPO committees, for the use by local partners.

key between the personal data and the corresponding collected data, shall be encrypted and placed in an offline file system with limited access.

- For the implementation of the cloud systems, several decisions need to be made such as what is the level of access to be given to different partners, which need to be defined in data agreements between consortium partners.
- Data agreements (also with professional companies and third parties), the informed consent of participants, along with the adequate approval and consultation with ethical, legal, and data protection authorities are the backbone of the proper handling and collection of data.

These guidelines were applied to the i–DREAMS NDS, which to date, has collected more than **100,000 trips**, **200,000 hours**, and an equivalent of **three million kilometers** of multi–modal driving data.

Beyond the protocols and the highlights defined, there are limitations and challenges that were extracted. This includes the management of privacy and consent forms (paper forms) across the different partners, different languages, but also the agreements for data access and use beyond the project end, for open access by other researchers (anonymized data). In particular, which portions of the anonymized data can be accessible for who and the procedures to be followed to make this data open.

Despite the mentioned challenges and limitations, it remains the responsibility for each data collector to document the handling of their data, mostly for quality control, but also to monitor goals achieved against what was expected or proposed. However, it is important to note that the analyzed projects were constrained by the technology and regulations that were available at the time. As a result, additional technologies/regulations may need to be considered for future projects in addition to the insights obtained from the reviewed NDS projects in this study. For example, the European Union has recently issued new guidelines for the processing of personal data in the context of connected vehicles and mobility-related applications (European Data Protection Board, 2020). The protocols drafted in this document therefore aim to be guidelines in the creation of a living and dynamic document on how to best handle data generated throughout NDS, which could serve as checklists for quality control, wherein the defined goals and followed guidelines could be cross-checked for validity purposes.

Acknowledgments

The research was funded by the European Union's Horizon 2020 i-DREAMS project (Project Number: 814761), by the European Commission under the MG-2-1-2018 Research and Innovation Action (RIA) and by the Ethics for the Smart City project funded by the Institute of Ethics in Artificial Intelligence at TUM (IEAI).

References

2BESAFE, 2012. 2besafe. URL: https://www.2besafe.eu/home/.

Afghari, A.P., 2019. Detecting motor vehicle crash blackspots based on their underlying behavioural, engineering, and spatial causes. Ph.D. thesis. University of Queensland.

Afghari, A.P., Washington, S., Haque, M.M., Li, Z., 2018. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Analytic methods in accident research 18, 1–14.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., et al., 2006. The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment. Technical Report. United States. Department of Transportation. National Highway Traffic Safety. URL: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/100carmain.pdf.

European Data Protection Board, 2020. Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related applications. URL: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/ guidelines-012020-processing-personal-data-context_en.

Federal Motor Carrier Safety Administration, 2016. Naturalistic Driving Study (OBMS Data Analysis). URL: https://www.fmcsa.dot.gov/safety/research-and-analysis/naturalistic-driving-study-obms-data-analysis.

FOT-Net, CARTRE, 2018. FESTA Handbook. URL: https://wiki.fot-net.eu/index.php/FESTAHandbook.

FOT-Net WIKI, 2015. Interaction. URL: https://wiki.fot-net.eu/index.php/INTERACTION.

Klauer, C., Pearson, J., Hankey, J., 2018. An Overview of the Canada Naturalistic Driving and Canada Truck Naturalistic Driving Studies. Technical Report. URL: https://www.vtti.vt.edu/PDFs/ndrs-2018/s4/Klauer.pdf.

Muckell, J., Hwang, J.H., Lawson, C.T., Ravi, S.S., 2010. Algorithms for compressing GPS trajectory data: An empirical evaluation, in: GIS'10: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 402–405. doi:10.1145/1869790.1869847.

Pilkington-Cheney, F., Talbot, R., Hancox, G., Filtness, A., Cuenen, A., Polders, E., Al Haddad, C., 2020. Experimental protocol. deliverable 3.4 - ec h2020 project i-dreams.

Shahid, A., Wilkinson, K., Marcu, S., Shapiro, C.M., 2011. Karolinska sleepiness scale (kss), in: STOP, THAT and One Hundred Other Sleep Scales. Springer, pp. 209–210.

Track Know, 2021. Track Know. URL: https://trackandknowproject.eu/.

UDRIVE, 2017. European Naturalistic Driving Study. Technical Report. URL: https://cordis.europa.eu/docs/results/314/314050/final1-udrive-final-publishable-summary-report.pdf.

Victor, T., Bärgman, J., Hjälmdahl, M., Kircher, K., Svanberg, E., Hurtig, S., Gellerman, H., Moeschlin, F., 2010. Sweden–Michigan naturalistic field operational test (SeMiFOT) phase 1. Technical Report. URL: https://www.saferresearch.com/library.