Baseline

Data Quality Control

Version 2.0, November 11, 2021

Baseline

Belgium | Austria | Bulgaria | Cyprus | Czech Republic | Finland | Germany | Greece | Ireland | Latvia | Lithuania | Luxembourg | Malta | Netherlands | Poland | Portugal | Slovakia | Spain | Sweden

baseline.vias.be

Project:	This document has been prepared in the framework of the BASELINE project, for which a grant has been awarded by the European Commission. Information on this project can be found on the website www.baseline.vias.be
References:	(1) Grant agreement under the Connecting Europe Facility (CEF) N° MOVE/C2/SUB/2019-558/CEF/PSA/SI2.835753 collection of Key Performance Indicators (KPIs) for road safety (2) Consortium agreement among the 19 partners of the Baseline project
Authors:	Bart van den Broek, Frits Bijleveld, Letty Aarts, Niels Bos, (SWOV)

Any comments or feedback regarding these guidelines, should be sent to baseline@vias.be.

Version history

Version	Date	Changes
1.0	15-10-2021	First draft – For internal discussion
2.0	11-11-2021	Second draft, including feedback from Vias and NTUA

Contents

1 Introduction and aims

The Communication of the European Commission "Europe on the Move – Sustainable Mobility for Europe: safe, connected and clean" of the 13th of May 2018 confirmed the EU's long-term goal of moving close to zero fatalities in road transport by 2050 and added that the same should be achieved for serious injuries. It also proposed new interim targets of reducing the number of road deaths by 50% between 2020 and 2030 as well as reducing the number of serious injuries by 50% in the same period. To measure progress, the most basic – and important – indicators are of course the result indicators on deaths and serious injuries.

In order to gain a much clearer understanding of the different issues that influence overall safety performance, the Commission has elaborated, in cooperation with Member State experts, a first set of key performance indicators (KPIs). The KPIs relate to main road safety challenges to be tackled, namely: (1) infrastructure safety, (2) vehicle safety, (3) safe road use including speed, alcohol, distraction and the use of protective equipment, and (4) emergency response. The aim of the KPIs is connected to EC target outcomes.

The aim of the BASELINE project, funded partially by the European Commission, is to assist participating Member States' authorities in the collection and harmonized reporting of these KPIs and to contribute to building the capacity of Member States which have not yet collected and calculated the relevant data for the KPIs. The outcomes of this project will be used to set future European targets and goals based on the KPIs.

The purpose of this document is to describe the criteria and process that we advise to use for checking the data quality of the KPI-data that will be gathered. Also further actions of the coordinator team will be described briefly.

2 The process of quality control

The Baseline project team proposed the following data gathering process in which several phases of quality control are foreseen (Yannis & Folla, 2021):

Action	**Actor**
1. Development of structure, templates and guidelines for the KPI datafiles	CT[1]
2. Sending of the templates and guidelines to the Member States	CT
3. Collection of KPI data	MS[2]
4. First data quality control	**MS**
5. Aggregation and KPI calculations	MS
6. Datafile transfer from MS to CT	MS
7. Second data quality control	***CT***
8. Integration of KPI data in central datafile	CT
9. Production of multi-country summary datafile	CT
10. Make datafiles available for public use	CT

These two stages of quality control will be described in more detail below:

2.1 First stage of quality control: Member States

The first stage of data quality control is up to the Member States and other entities that will deliver data to the Baseline project. In this stage, they are requested to check a number of issues and they are advised to follow a particular process.

[1] Baseline Coordination Team: VIAS, SWOV, NTUA

[2] Member States and other entities that will deliver data for this project

2.1.1 Issues to be checked

1. Is the correct datafile format used for those KPI's that will be delivered to the project?
2. Is the datafile format completely filled with the available data of each KPI that will be provided by the MS?
3. Optional: are additional data or disaggregations provided if available?
4. Are the data correct? (see paragraph 3.1)
5. Optional: are clarifications of any outliers or otherwise unexpected values in the data provided, if applicable?
6. Is the metadata of each KPI provided in the Excel file sheet on metadata?
7. Optional: are any additions to the metadata provided, e.g. regarding extra data or extra disaggregations?
8. Is the provided metadata correct? (see paragraph 3.2)
9. Is the datafile named as *Baseline-DataFile-Country abbreviation*[3]

A more elaborated version of this list of items that have to be checked is available in Annex A.

2.1.2 Advice for the process of quality control

In order to reach optimal effectivity, the CT recommends to organize the quality control process as follows:

- **Database management**: commission a database manager who is responsible for the organization of the actions related to the datafile process and checks whether the requested actions are performed in time.
- **Data processing**: assign the tasks of filling the datafile to (a) staff member(s) that is/are familiar with the particular data or data in general. This task could be combined with database management.
- **Data quality control**: assign the task of quality control to (a) staff member(s) who has/have knowledge of the project and expertise in the field of data, but preferably has/have not been part of the data collection and filling of the datafile.

It is advised that the database manager checks that all actions of filling the database and metadata and quality control are performed and that the datafile is sent within the requested time to baseline@vias.be.

2.2 Second stage of quality control: Baseline Coordination Team

In the second stage of data quality control, the quality control staff of the CT will check the received datafiles. This is meant as a double check on the data provided by the MS, to support their quality control, and to consider the homogeneity of all the data. These checks will consist of 3 stages:

1. Check on outliers and correctness of calculations within each datafile.
2. Check on strange values taking into account the method used (metadata) and related to the other datafiles that will be received on a particular KPI.
3. Conclusion on homogenous groups of KPIs.

2.2.1 Check on outliers and incorrect calculations

The quality control staff of the CT will check which data (per KPI, country and disaggregation) can be regarded as correct in terms of correct calculations and of absence of outliers.

- Check on correct calculation of the KPIs, including correct calculation of any intermediate quantities used to derive the KPIs.
- Check on outliers within each datafile.

For those variables where unexpected values are encountered by the quality control staff of the CT, the delivering MS will be contacted for further information or confirmation of the values that were found.

2.2.2 Check on strange values in relation to several references

The quality control staff of the CT will check which data can be regarded as correct in terms of absence of any strange values in relation to other data.

- Per country and KPI, check if values make sense with respect to those of other KPI and metadata.

---

[3] Example: *Baseline-DataFile-BE* (for Belgium)

For those variables where the quality control staff of the CT encounters values that are unexpected, the delivering MS will be contacted for further information or confirmation of the values that were found.

2.2.3 Conclusion on data homogeneity

The quality control staff of the CT will check which data (per KPI, country and disaggregation) can be regarded as homogenous within a certain group of KPI measurements over countries and which data meets the methodology that has been set for each KPI.

- Check on strange values, absolute and relative within and between MS datafiles, also related to the metadata delivered. E.g. per KPI does any MS have exceptionally high or low values with respect to the other MS? The methods used by each MS will be taken into account, and consideration will be given to how comparable resulting KPI data of each MS are.
- Check on correct calculation of aggregated values.
- Check on aggregation per KPI over all MS, taking into account the methods used by each MS. Consideration will be given to how and to what extent values obtained through different methods can be combined.

For those variables where unexpected values are encountered by the quality control staff of the CT, the delivering MS will be contacted for further information or confirmation of the values that were found.

3 Structure of the data

Templates for the gathering of data by MS for each KPI have been provided by the CT. The resulting data files will contain KPI data and metadata. The disaggregated data, KPI data and metadata are expected to have a certain structure.

3.1 Disaggregated data

The disaggregated KPI data as to be provided in the datafiles is the cleaned and (semi)aggregated result of raw measured data. What can be expected from the structure of the disaggregated data?

- Unusual values in the disaggregated data may sometimes be the result of an error code, and in that case should be treated with care. Some measurement devices are programmed such that in the case of an error they return a value that is supposed to be unusual, typically at the upper boundary of the value range, e.g. from 0 to 255 ($= 2^8 - 1$) or 999. High frequencies of such unusual values may be an indication of an error code. These values should be treated with care, e.g. by leaving them out or correcting them to more likely values depending on the meaning of the error code.
- Outlier values in the disaggregated data may be detected using statistical methods; various methods for outlier detection can be found in the literature, e.g. in the survey on outlier detection methodologies by Hodge & Austin (2004).

3.2 KPI data

What can be expected from the structure of the KPI data?

- KPI values may differ over disaggregation levels. For example, the share of drivers driving under the influence of alcohol may be higher on weekend nights than on week days. The CT advises each MS to have an expert with domain knowledge judge whether values per disaggregation level are as expected.
- KPIs that are defined as a percentage are often expected to have typical values. For example, the share of drivers driving within the speed limit may typically have values in between and not close to 0% and 100%, whereas the share of vehicle occupants using a seat belt may typically be close to 100%. The CT advises each MS to have an expert with domain knowledge judge whether values are as expected.

The datafiles for several KPIs require values for certain auxiliary variables, such as the number of measurement locations, the number of observed vehicles, the traffic volume and weights. These variables have an expected structure:

- The number of measurement locations should be nonnegative and integer valued. Furthermore, per KPI a minimum number of locations per stratum is suggested, such as at least 10 locations per road type for the KPI safety belts, hence much larger values are not to be expected, but could valid.
- The number of observed vehicles, occupants, riders, drivers and/or passengers should be nonnegative and integer valued. Furthermore, per KPI a minimum number of observations is suggested, such as at least 500

observations per road type for the KPI speed, hence much larger values are not to be expected, but could be valid.
- Weights should have values greater than zero. A guideline on weighting of data is provided by Boets, Silverans & Bijleveld (2021).

For those variables where the quality control staff of the CT encounters values that are unexpected, they will contact the delivering MS for further information or confirmation of the values that were found.

3.3 Metadata

What can be expected from the structure of the KPI metadata. This concerns the data that is to be provided by the MS in the datafile templates, under "metadata", and will contain information regarding data collection methodology and analysis.

- Variables such as the data collection method and the sampling units can only have predefined values that are to be selected from a drop-down menu in the data file.
- Other variables have no designated format, most will have a text value providing content ranging from a date to a description of a road type, and their processing will require a level of interpretation. Their content should also agree with the provided KPI data. For example, the method used for the sampling of locations, or the sampling units used, should agree with the provided data. The CT advises the MS to have this checked by an expert with domain knowledge.
- The metadata should support and clarify the KPI data. For example, longer traffic count durations are expected to result in higher traffic counts. The CT advises the MS to have this checked by an expert with domain knowledge.
- The metadata provides quantities such as measurement durations and the number of measurement locations. Any minimum requirements on those quantities should be met. The CT will give consideration to what it means for the data quality when minimum requirements are not met, and how to deal with this.
- The metadata should tell which methods are used. The CT will divide the MS into groups, depending on which methods they used, and consider how and to what extent KPI values from different groups can be combined.

For those variables where the quality control staff of the CT encounters values that are unexpected, they will contact the delivering MS for further information or confirmation of the values that were found.

References

Hodge, V.J. & Austin, J. (2004). *A survey of outlier detection methodologies*. Artificial Intelligence Review, 22 (2). pp. 85-126.

Boets, S., Silverans, P. & Bijleveld, F. (2021) *Considerations for sampling weights in Baseline*.

Yannis, G. & Folla, K. (2021). *Baseline Datafiles*. Training Session on Datafiles, August 26, 2021

Annex A – Checklist for first stage data quality control by MS

☐ Is the correct datafile format used for those KPI's that will be delivered to the project?

☐ Is the datafile format completely filled with the available data of each KPI that will be provided by the MS? Cells in the datafile that cannot be filled, e.g. because they belong to a variable at a disaggregation level for which no data has been collected, can be left empty.

☐ Optional: are additional data or disaggregations provided if available?

☐ Are the data complete and correct?
- **Check on range**: Check if values lie within the right range (e.g. nonnegative speed, integer number of locations, percentages between 0% and 100%, etc.).
- **Check on outliers**: Use quantitative statistical methods, e.g. based on the interquartile range, to check on outliers.
- **Check on strange values**: It is advised that an expert with domain knowledge judge whether values in the disaggregated data and per aggregation level are as expected.
- **Check on minimum requirements**: Are the minimum requirements satisfied, such as on the minimum number of measurement locations?

☐ Regarding specific KPIs:
- **Alcohol**: do the number of observed drivers below and above the legal limit add up to the number of observed drivers?
- **Distraction**: do the number of observed drivers making use or no use of a mobile device add up to the number of observed drivers?
- **Vehicle Safety**: do the number of 4-star and 5-star passenger cars both not exceed the number of passenger cars?
- **Post-crash Care**: do the number of fatalities on arrival at the crash location and on arrival at the hospital both not exceed the number of casualties?

☐ Optional: are clarifications provided for any empty cells in the datafile or any values in the data that are outliers or that the MS for any other reason believes may be considered as unexpected by the CT, if applicable?

☐ Is the metadata of each KPI provided as it is requested in the datafile under "metadata"? Details on how the metadata should be provided is described in the methodological guidelines.

☐ Optional: are there any additions to the metadata provided, e.g. regarding extra data or extra disaggregations?

☐ Is the provided metadata complete and correct?
- **Check on completeness**: Is all metadata provided? Are values from drop-down menus selected, such as for the data collection method?
- **Check on soundness**: It is advised that an expert with domain knowledge judges the content of the meta data.
- **Check on supporting value**: Does the metadata support and clarify the KPI data?
- **Check on minimum requirements**: Are all minimum requirements on the metadata met, e.g. regarding number of measurement locations?
- **Check on methods used**: Check whether the metadata explains which methods were used.

☐ Is the datafile named as *Baseline-DataFile-Country abbreviation*[4]

[4] Example: *Baseline-DataFile-BE* (for Belgium)