EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly
financed by the European Union
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Applying the extended Total Survey Error approach to statistics based on MNO data sources: the case of Mobile Network Operators data

Gabriele Ascari, Giorgia Simeoni - ISTAT

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

The conference is partly financed by the European Union

Introduction

As previous experiences have shown, from a quality perspective it is difficult to study big data as a single category.

The purpose of this work is to understand the risks and errors associated with MNO data as they are included in a process and progressively transformed into statistical objects.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Introduction

The study is made up by two components

- Total Survey Error approach
- Two-phase life cycle model

Both components have been extended to be applied outside the traditional scope of sampling surveys.

The conference is partly financed by the European Union

The original data lifecycle models by Groves et al (2004) and Zhang (2012) allow

- for the identification of errors affecting the records/units and the variables
- for the mapping of the error to specific steps of the process

These are the two main objectives of this work applied to the use of MNO data in statistical processes.

The conference is partly financed by the European Union

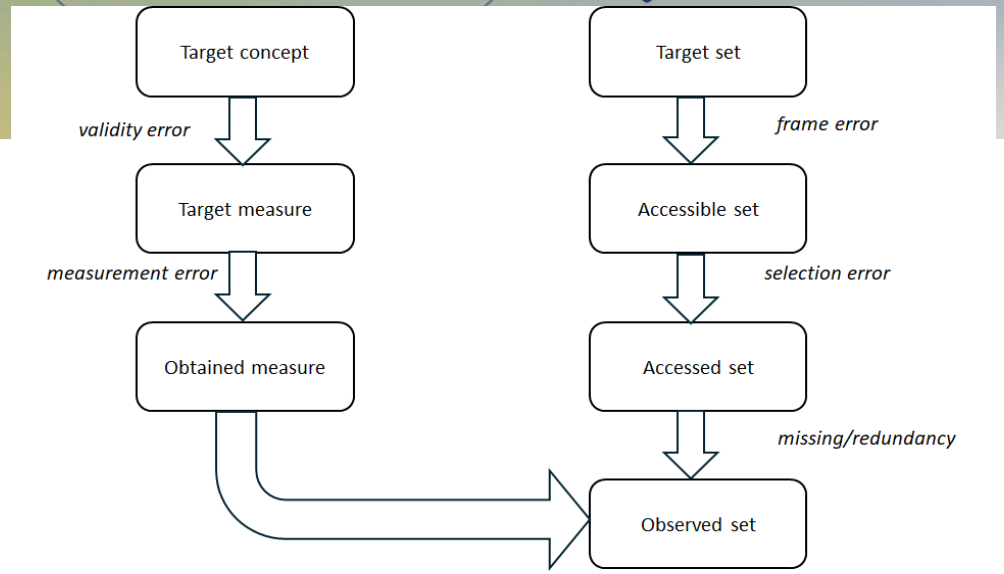
However, those models were developed for sample surveys and processes using administrative data.

The solution was to split phase 1 of the original two-phase lifecycle model into two separate phases, describing event and device data respectively.

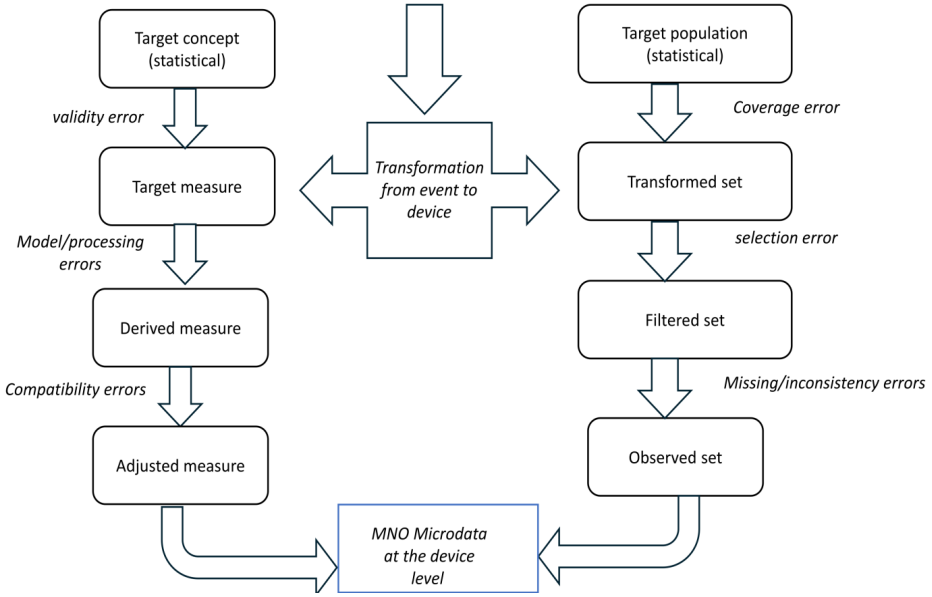
Indeed, the flow assumed here is

The conference is partly financed by the European Union

Phases 1a and 1b

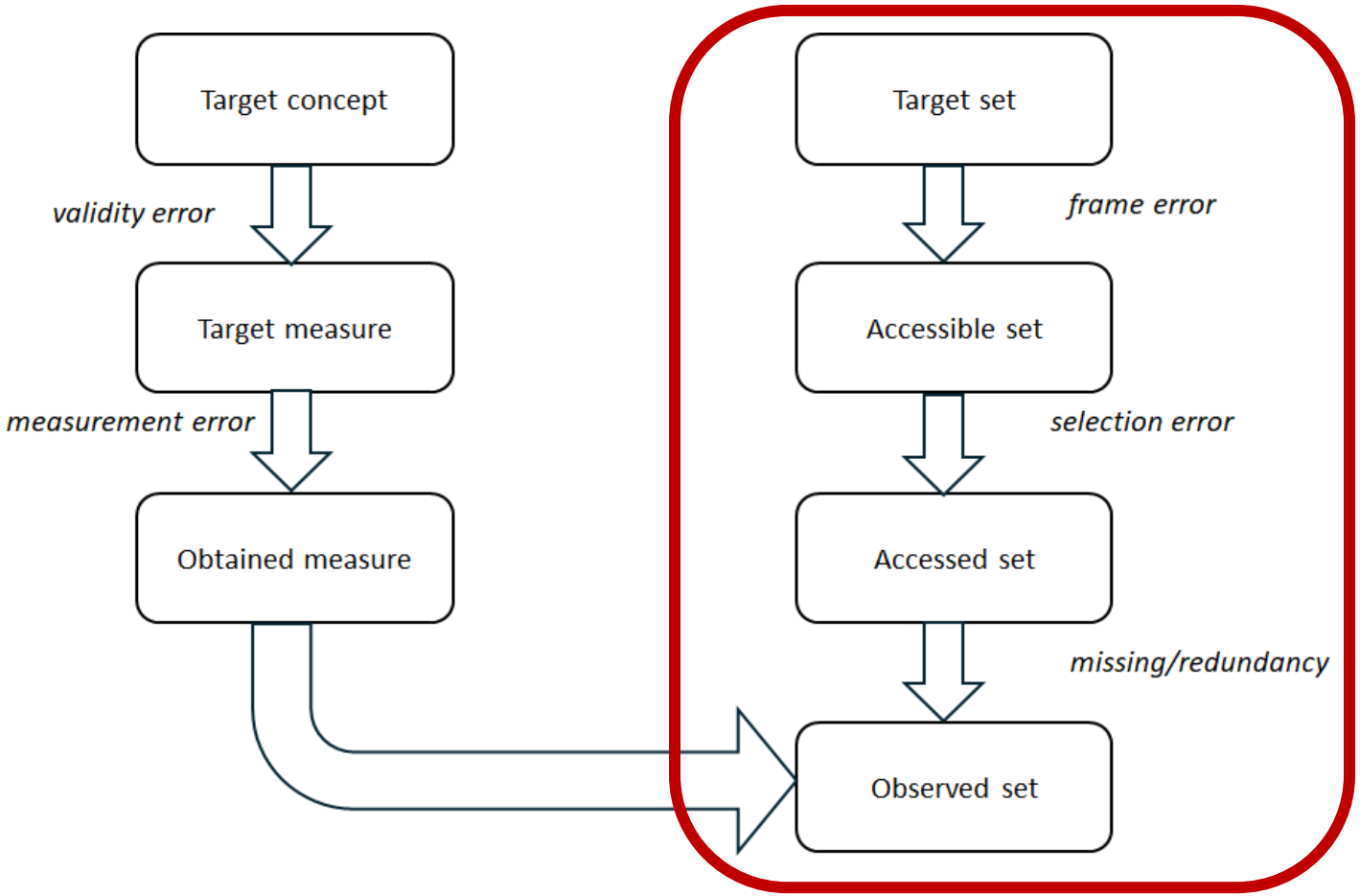


event data

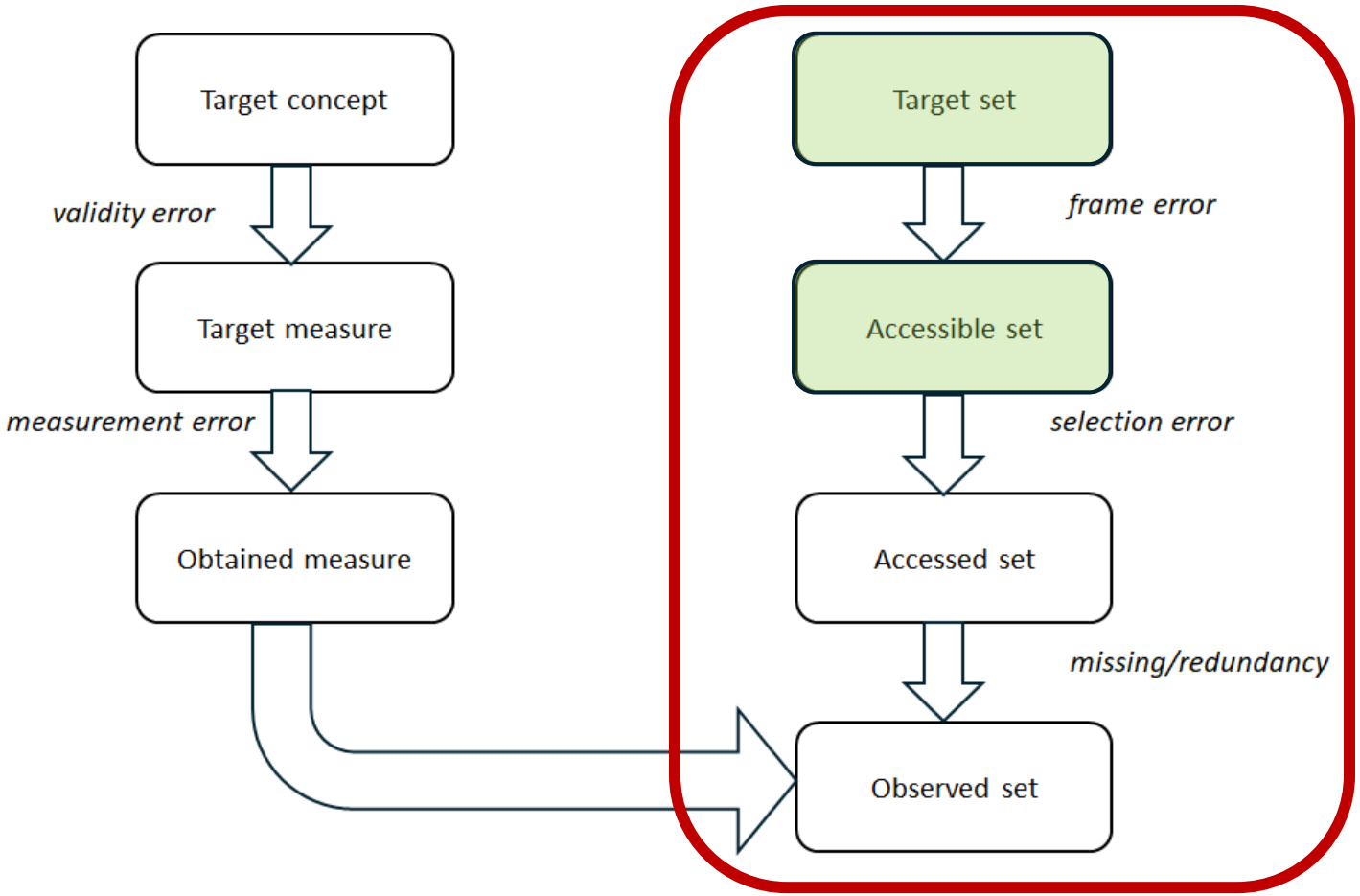


device data

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

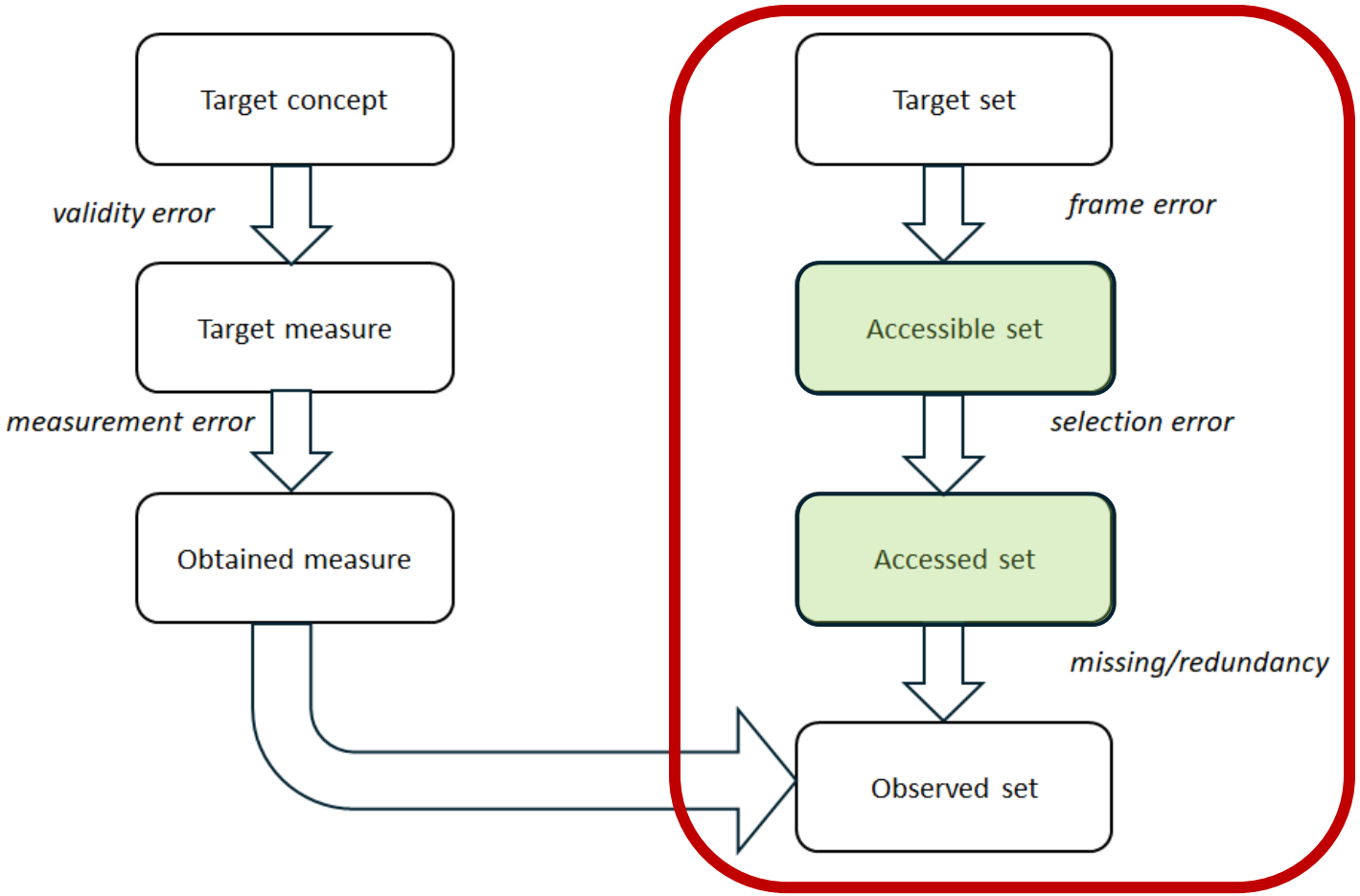


The conference is partly financed by the European Union



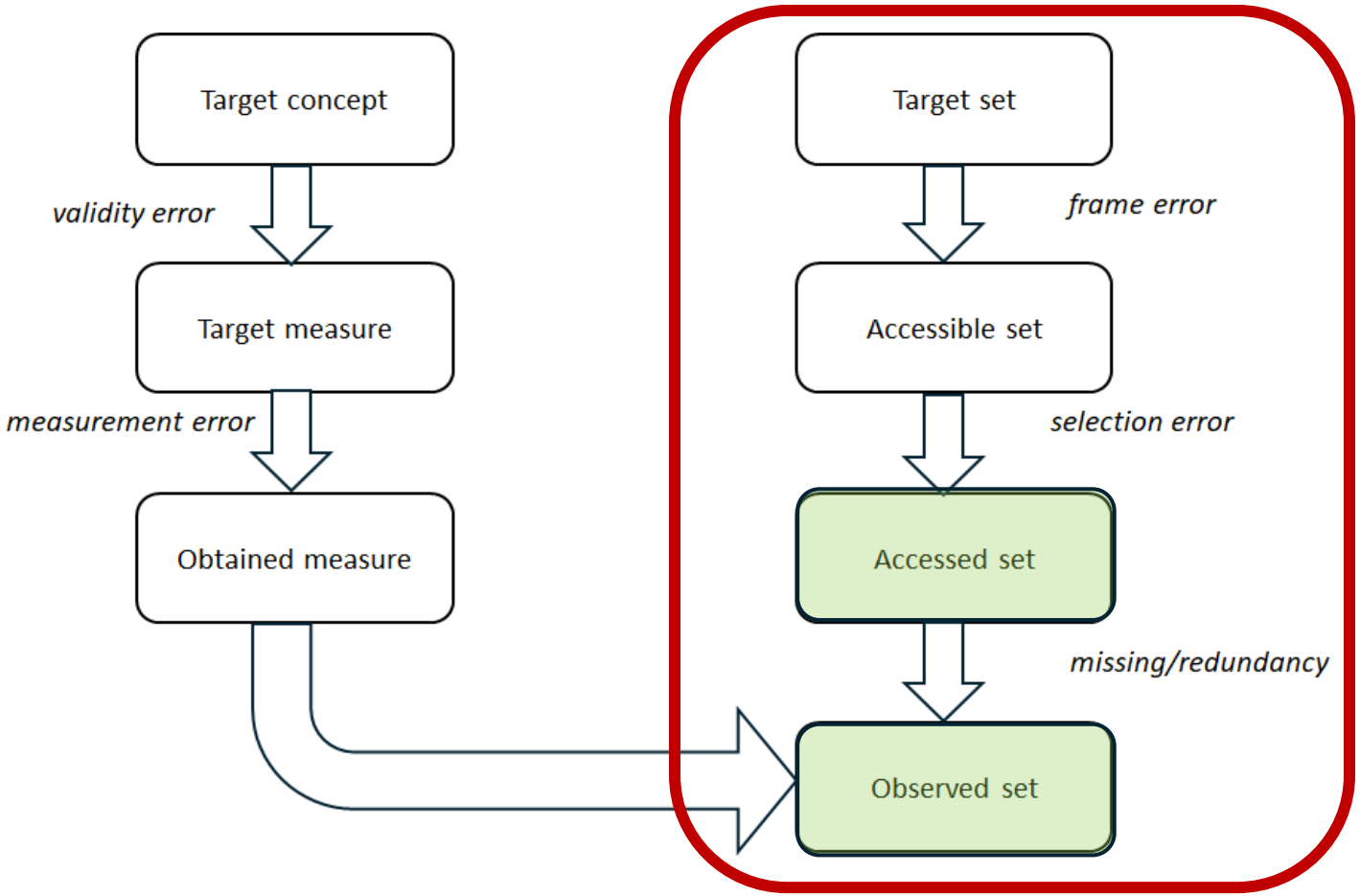
Frame error: this error emerges when we lose information on the **target set** of all the potential events due to device/user's behaviour

The conference is partly financed by the European Union



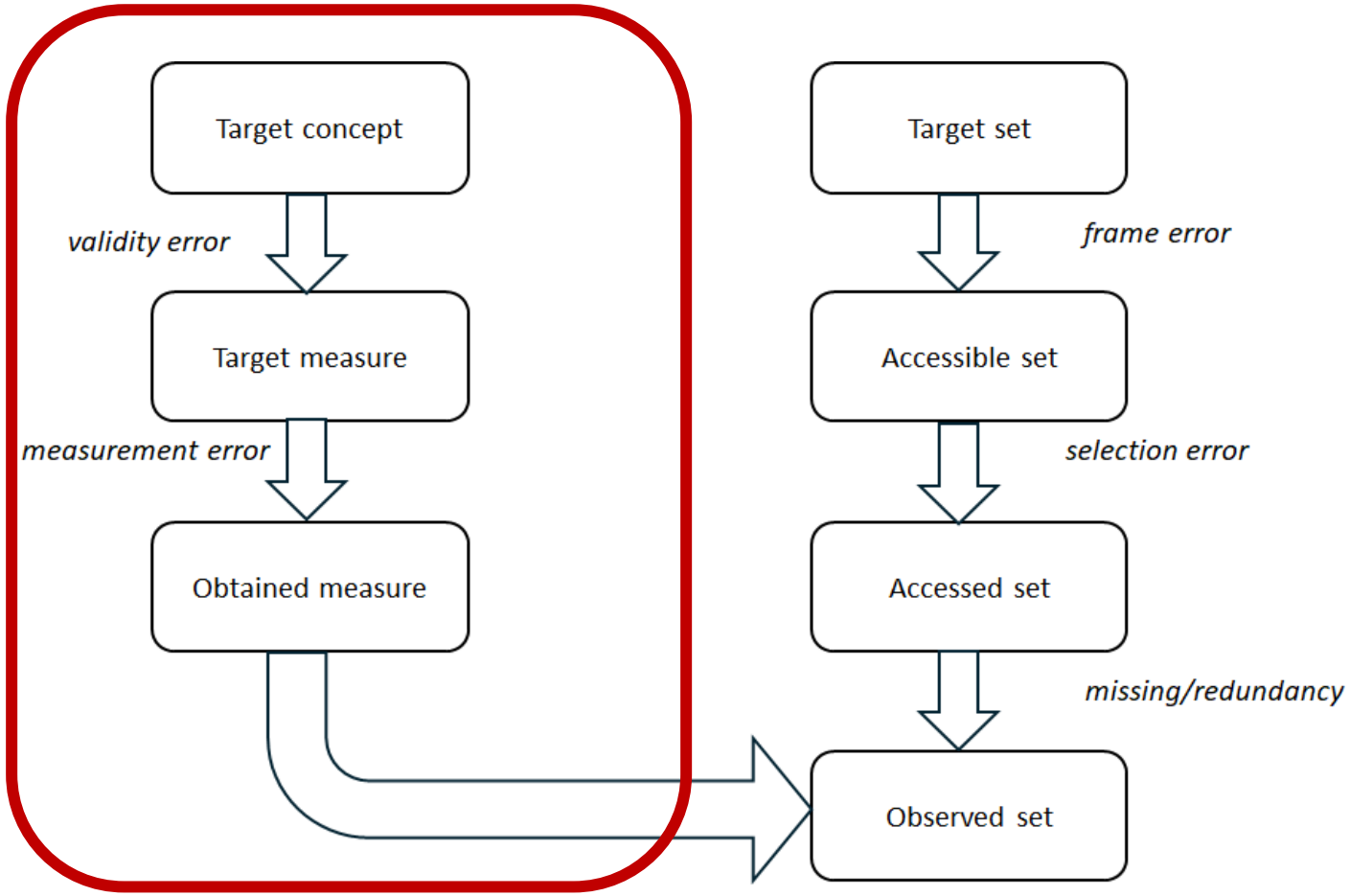
Selection error: the **accessible set** of all the events the instruments may be able, in theory, to observe is reduced to the accessed set due to technical errors on the network's part.

The conference is partly financed by the European Union

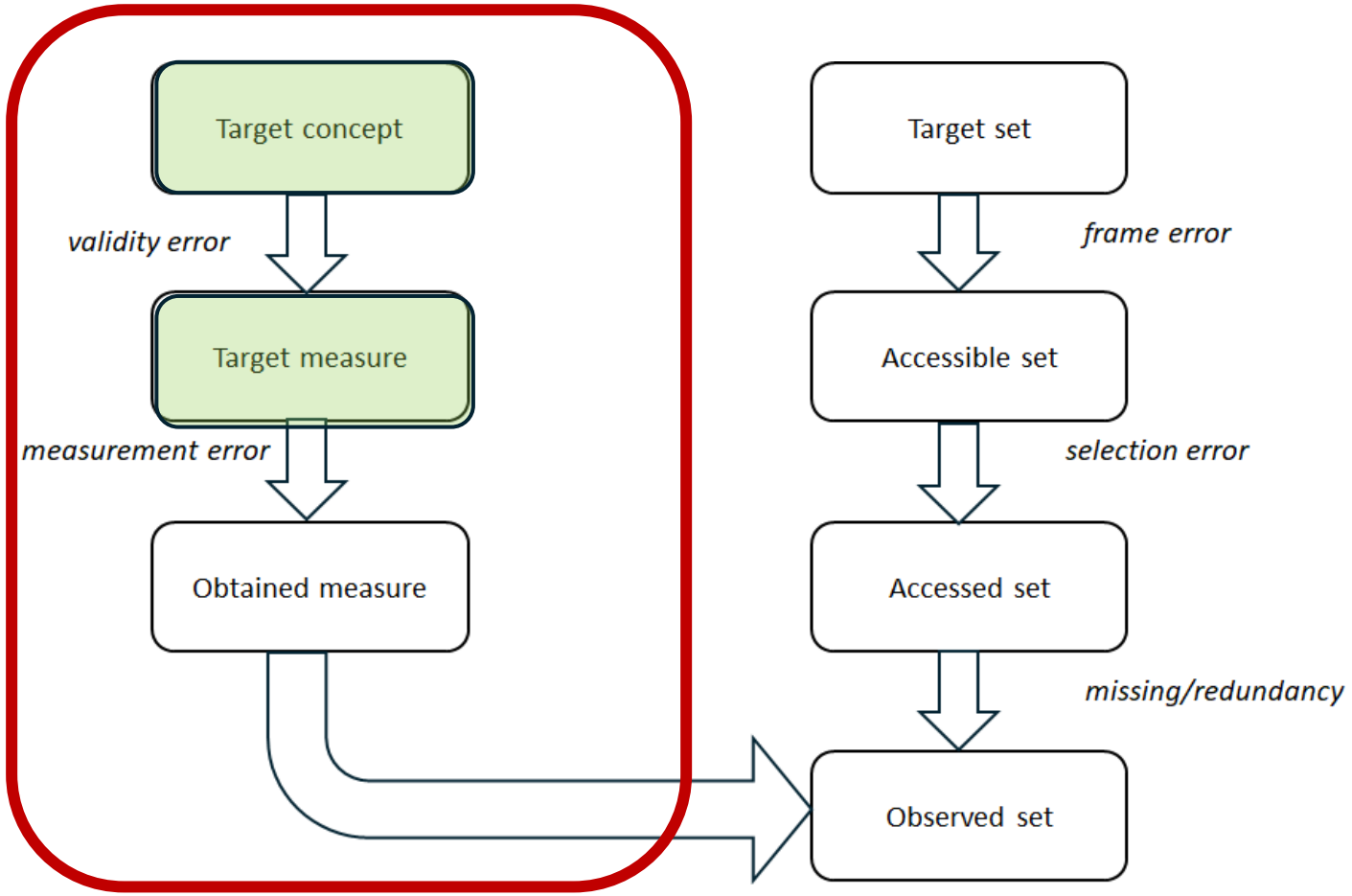


Missing error: along with the many errors that may affect the accessed set, missing values within it are assumed to be purged to obtain the **observed set** of event data.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

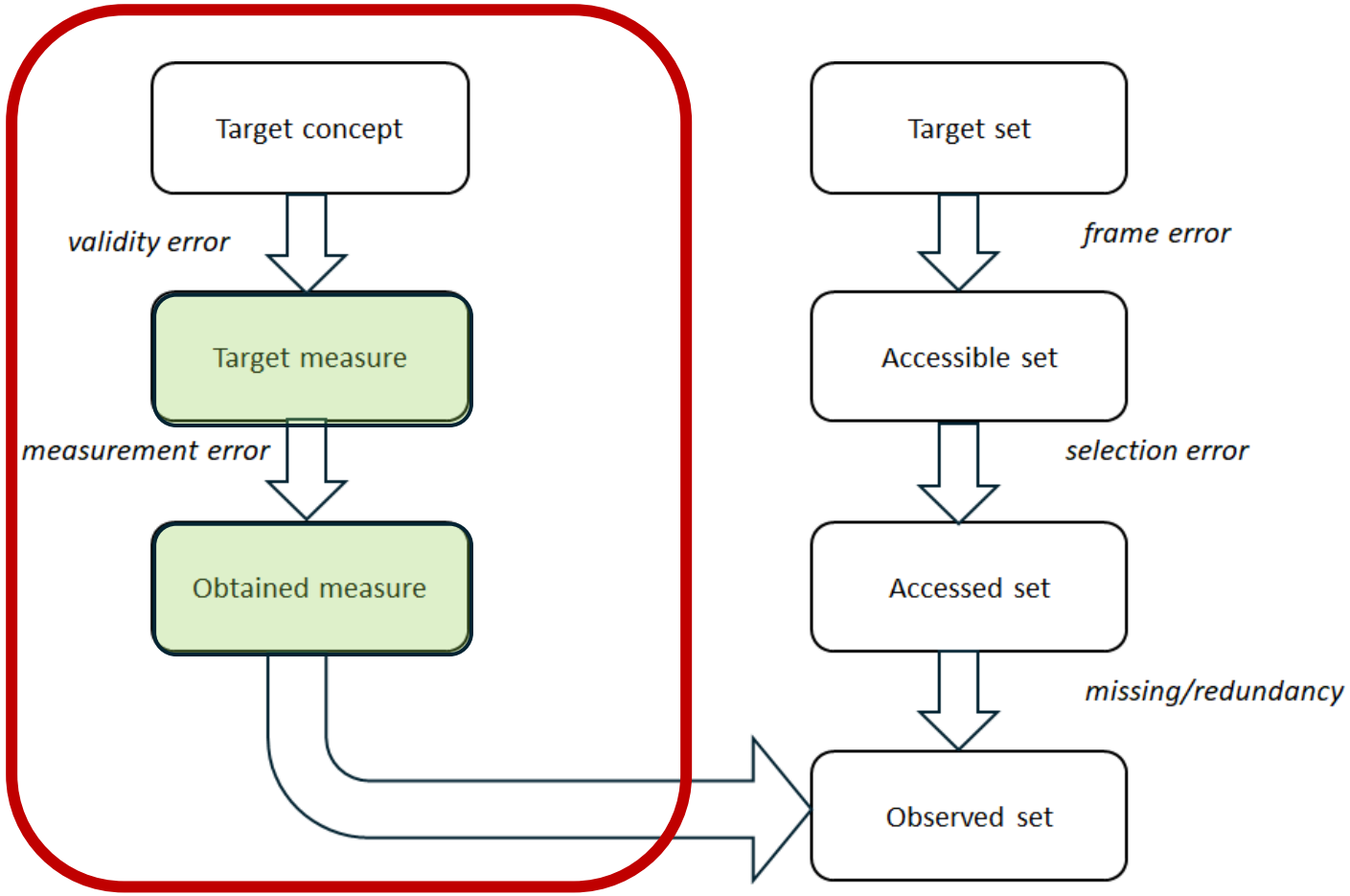


The conference is partly financed by the European Union



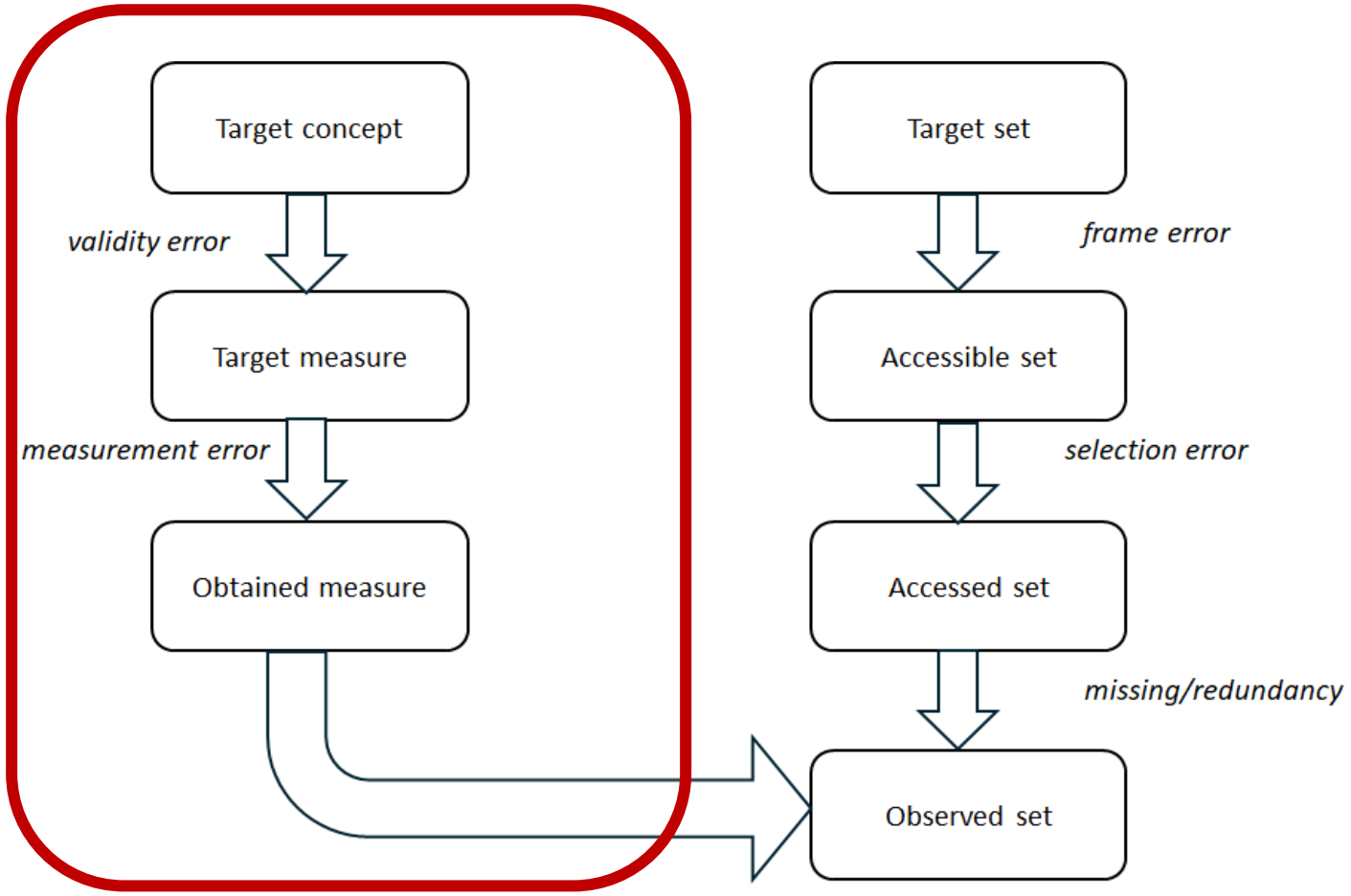
Validity error: the interest in the correct geographical placement of the event (**target concept**), which is measured through the position of the cell and the related information (**target measure**)

The conference is partly financed by the European Union



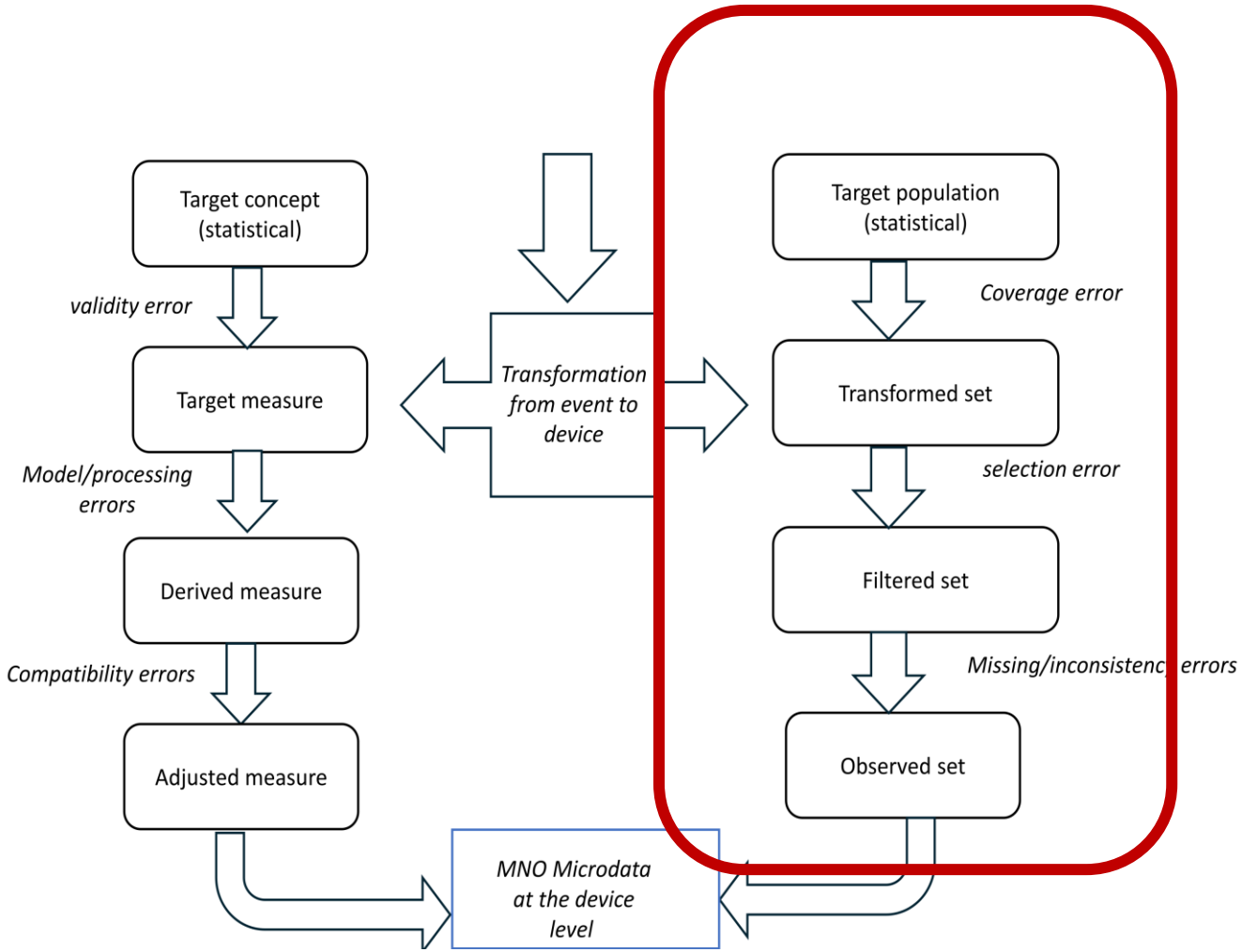
Measurement error: malformed data and errors of similar nature may happen at this step, giving origin to the **obtained measure**.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

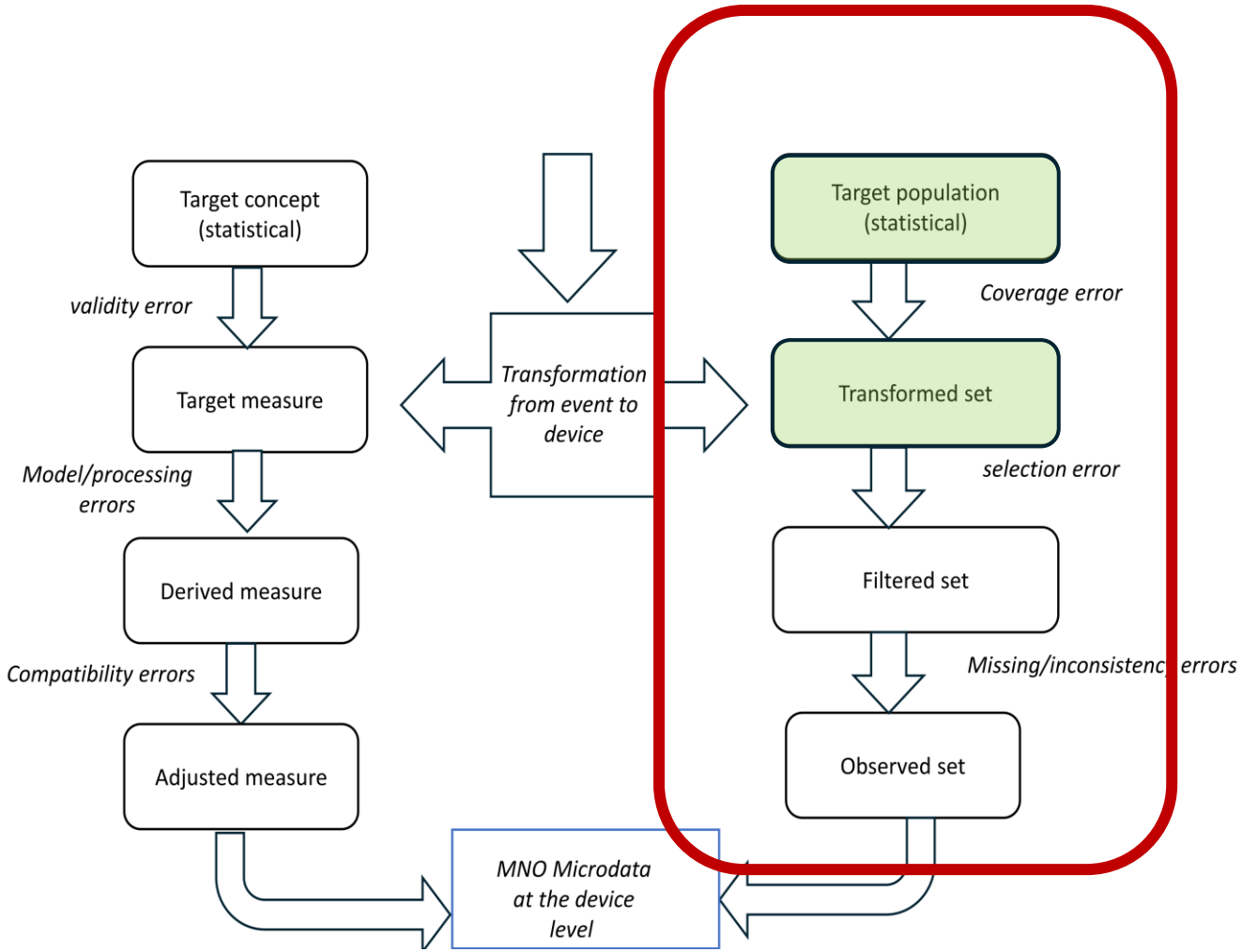


In the assumption that no processing other than record elimination is done on these data, the resulting set corresponds to the observed set in the representation line.

The conference is partly financed by the European Union

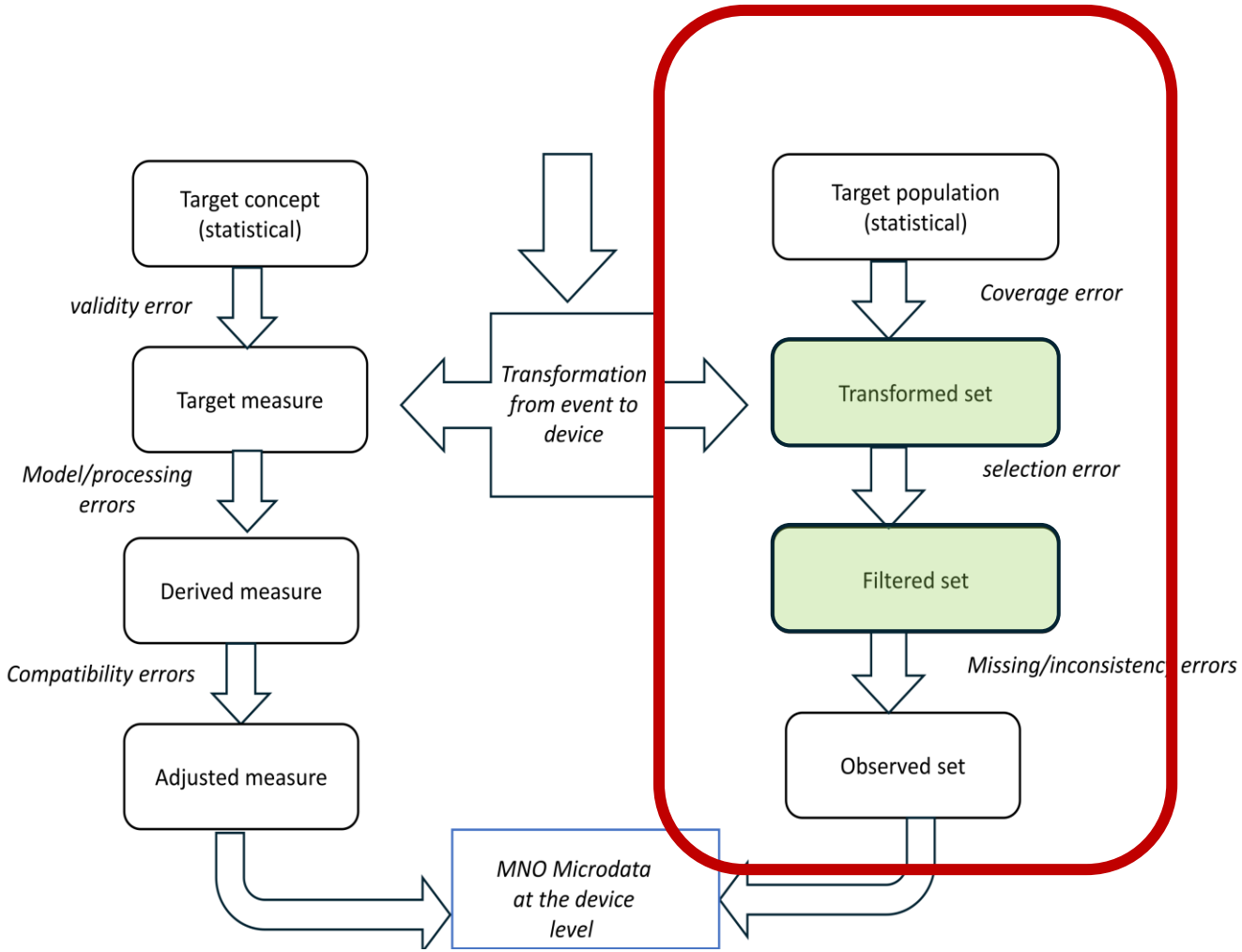


The conference is partly financed by the European Union



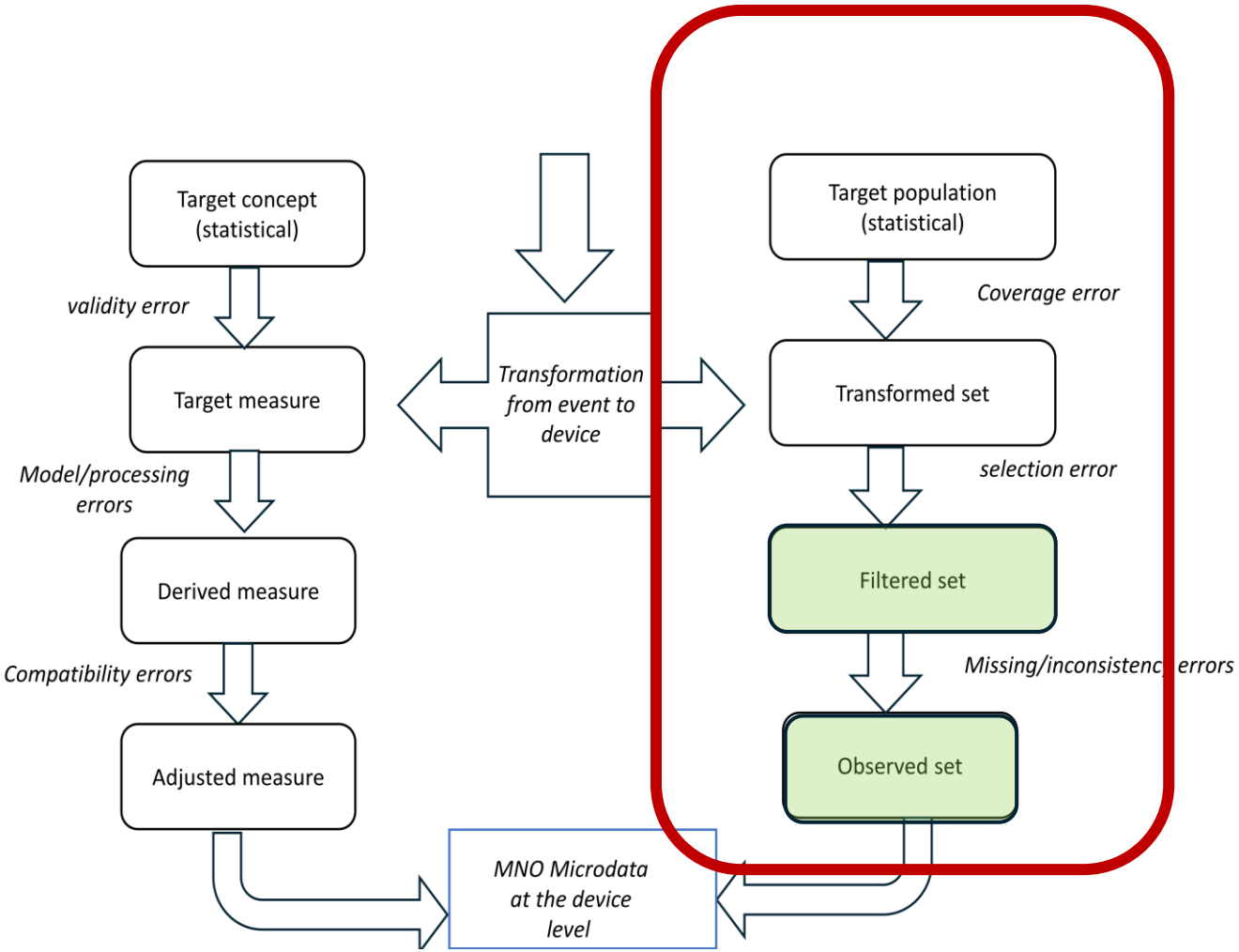
Coverage error: referring to the statistical **target population**, there will be issues concerning both under-coverage and over-coverage in the set of devices that can actually be observed (**transformed set**).

The conference is partly financed by the European Union



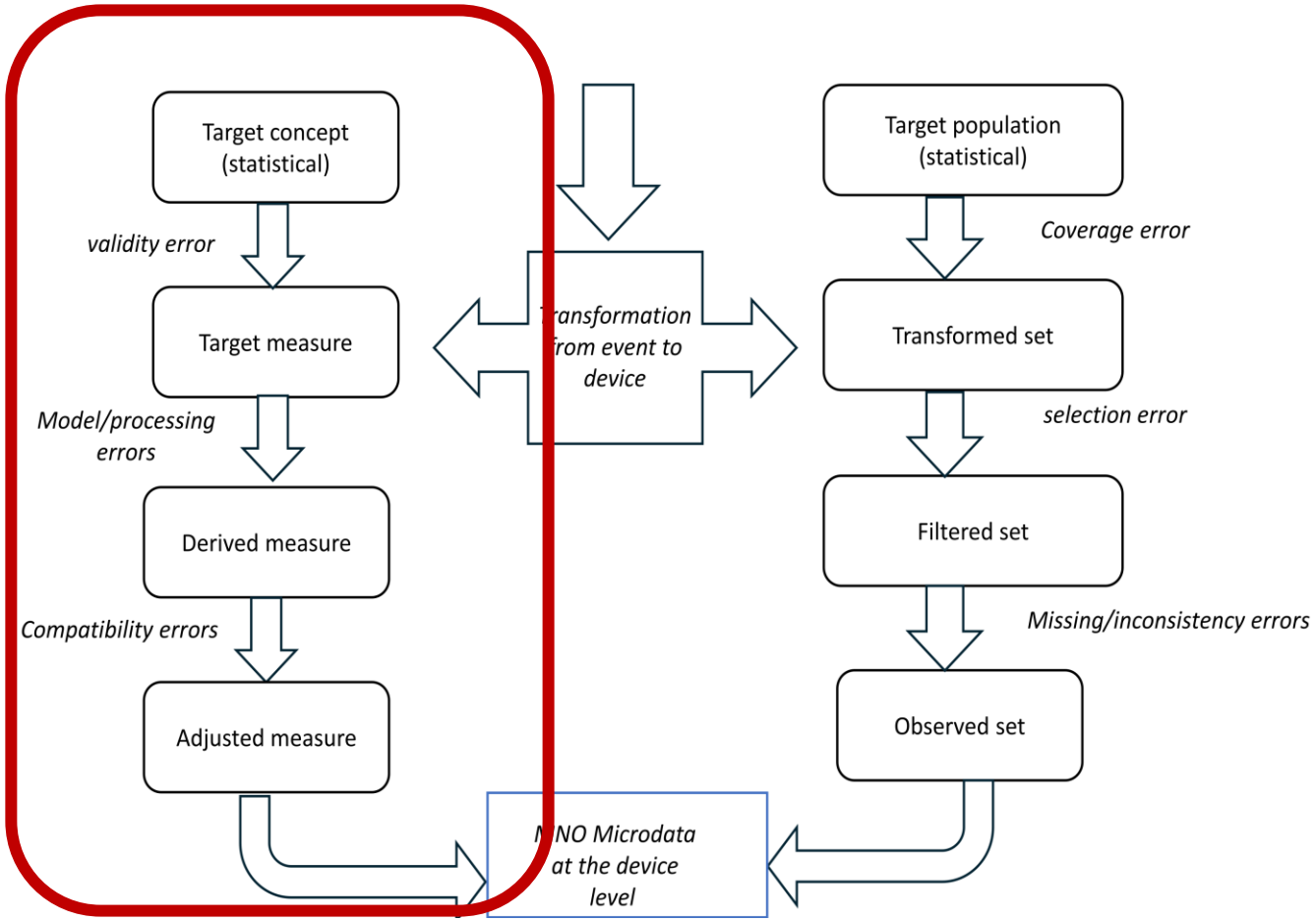
Selection error: this error emerges when expunging non-relevant devices from the data before obtaining the **filtered set**.

The conference is partly financed by the European Union

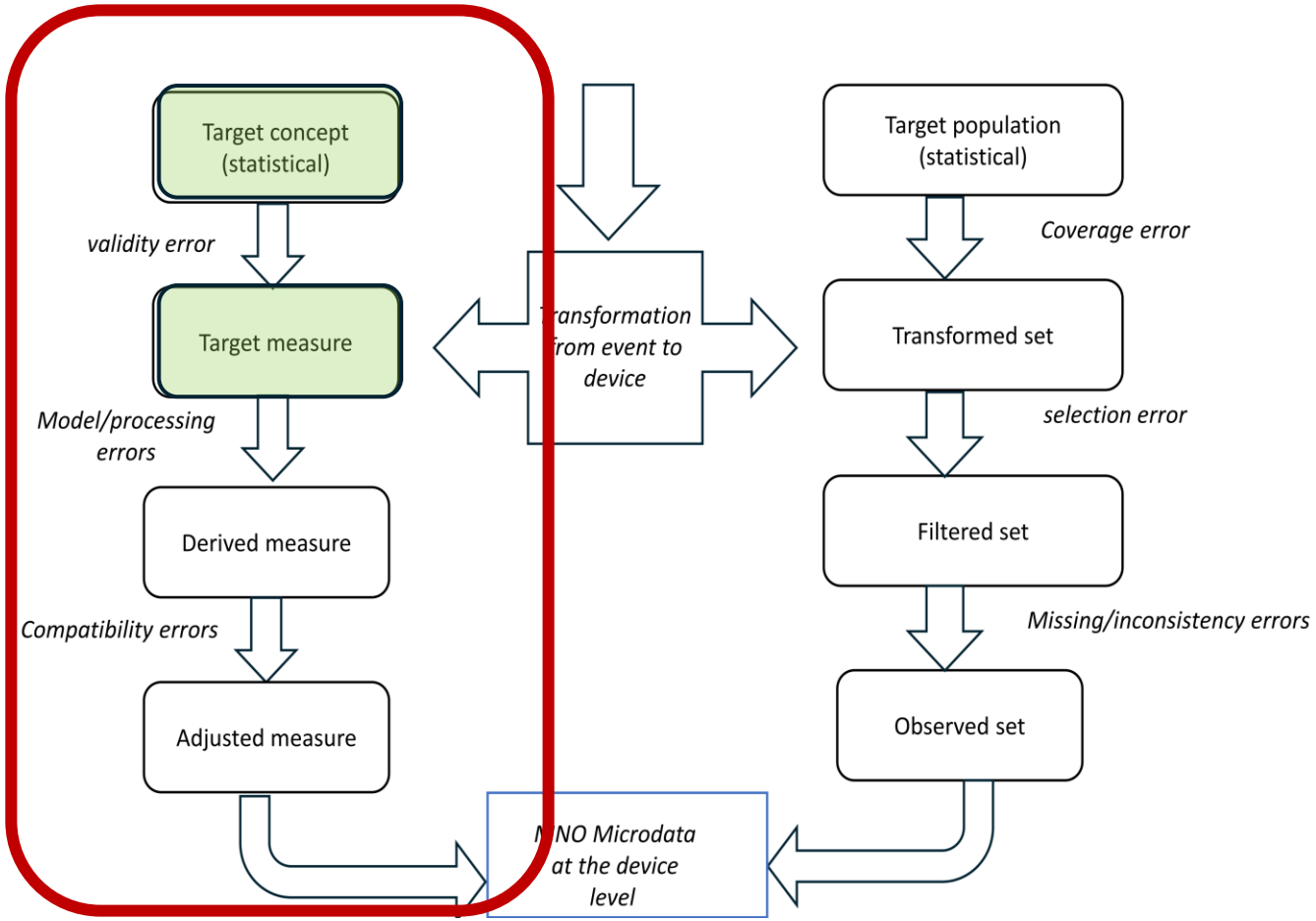


Missing/inconsistency error: specific devices in the filtered set may be affected by inconsistencies in the events they are characterised by; such units should be removed to obtain the **observed set**.

The conference is partly financed by the European Union

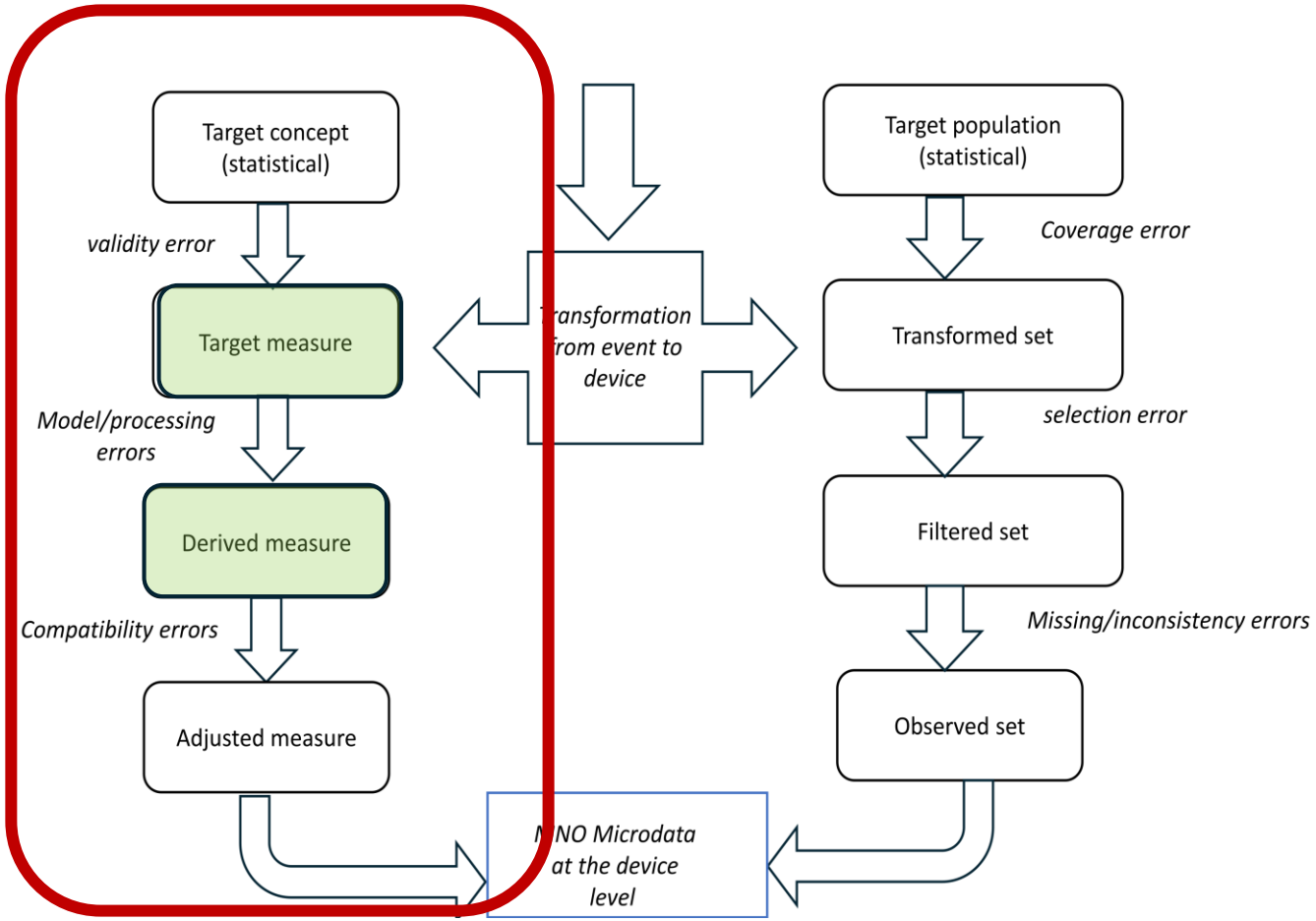


The conference is partly financed by the European Union



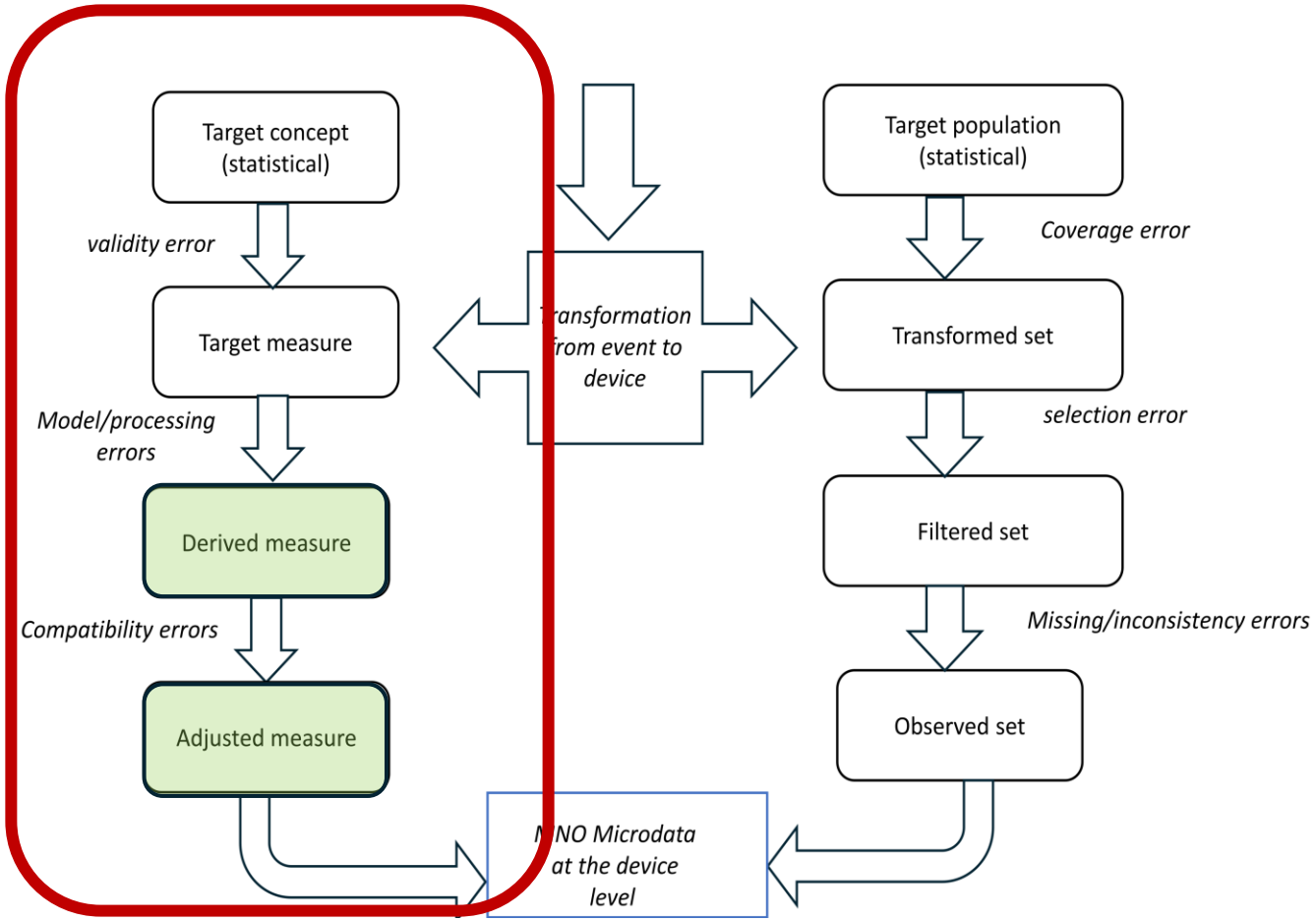
Validity error: information on the device data is a proxy of the **target concept**, which is statistical. By definition a validity error arises, affecting the **target measure**.

The conference is partly financed by the European Union



Model/processing errors: processing of data at this step is done through algorithms that usually focus on measurement of time periods involving the position of the device: the errors caused by such procedures affect the **derived measure**.

The conference is partly financed by the European Union



Compatibility errors: further adjustments using probability models and information provide the **adjusted measure**.

The conference is partly financed by the European Union

Limitations of the model and simplifications:

- Integration with other MNO information (e.g. topology data) assumed as already carried out
- Integration with other categories of data not considered
- Multiple operators not considered
- Ownership of the procedures not considered (“who does what”)

The conference is partly financed by the European Union

Potential applications of the model:

- Developing a comprehensive theory of the TSE and its components concerning MNO data (interactions between errors, main errors related to use cases)
- Identification/development of quality indicators specific for each set or error
- Identification of the risks at each step and responsibilities

The conference is partly financed by the European Union

References

Groves R., Fowler F., Couper M., Lepkowski J., Singer E., Tourangeau R (2004). *Survey Methodology*. Wiley.

Zhang L.C. (2012). *Topics of statistical theory for register-based statistics and data integration*. Statistica Neerlandica, vol. 66, n. 1.

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Thank you for your attention!

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union