

On the path to quality assurance of trustworthy AI in medicine: The METRIC-framework for systematically assessing medical training and test data

Daniel Schwabe¹, Katinka Becker¹, Martin Seyferth¹,
Andreas Klaß¹, Tobias Schäffter^{1,2,3}

Introduction

The adoption of machine learning (ML) applications into all major areas of our lives is underway. However, the race for implementation and utilisation is currently outpacing comprehension of the technology. The complex and black box character of AI applications entails safety, security and privacy concerns. A lack of confidence in the transparency of AI prevents its utilisation for society and

economic growth. It can lead to a slowed adoption of innovations and discourage innovators from unlocking the technology's full potential. Hence, the demand for an improved understanding and validation of AI systems is ever increasing. Metrological evaluation of AI supports this effort, aiming to ensure reliability, trust and safety.

EU Project „TEF-Health“: Services for Trustworthy AI



The EU has established large-scale reference testing and experimentation facilities for different sectors, so-called TEFs, to support the **development of trustworthy AI** within the EU. TEFs develop platforms where services that concern experimentation and testing of trustworthy AI solutions are offered mainly to SMEs.

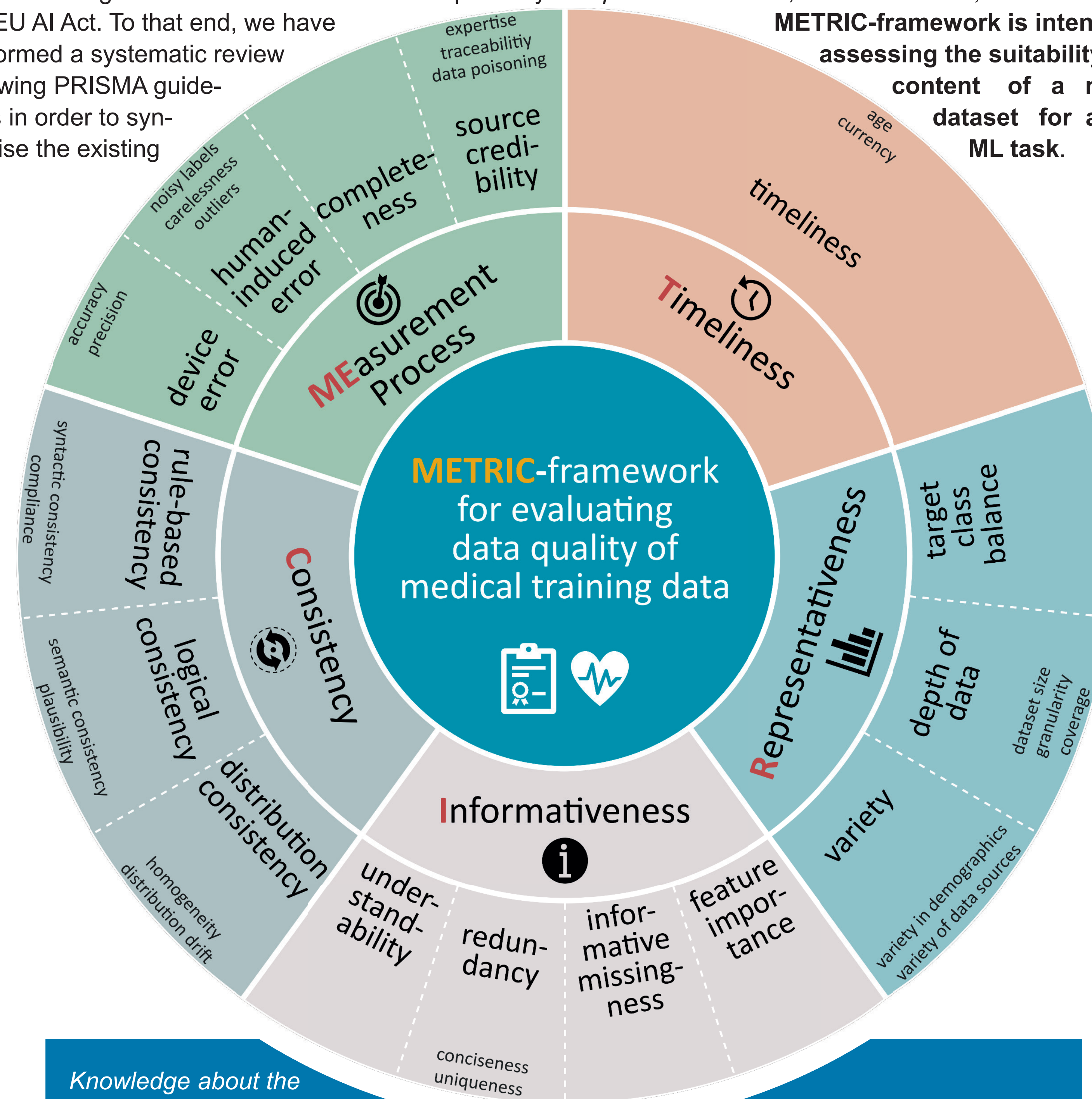
Since trustworthy AI is especially important in medicine due to the large implications for patients' lives, the TEF-Health project operates in the healthcare domain. TEF-Health is an EU-wide network of 52 partners consisting of leading EU hospitals, medical universities, NMIs, research institutes, certification bodies, as well as innovation clusters. Their aim is to facilitate market access for AI solutions in healthcare, remove innovation hindrances and **advance the field of quality assurance and certification** of AI in order to make usage of AI trustworthy. While trustworthiness concerns various aspects including ethical, transparency and safety requirements, we first focus on the importance of data quality (training/test) in ML.

The METRIC-Framework for Systematic Evaluation of Medical ML Data

Due to their black box nature, one of the main ways to establish quality assurance for AI systems is to evaluate and curate training and test data. Moreover, **data quality fundamentally influences the behaviour of ML products**. Therefore, evaluating data quality will play a key part in the regulatory approval of high-risk AI medical devices as required by the EU AI Act. To that end, we have performed a systematic review following PRISMA guide-lines in order to synthesise the existing

knowledge on data quality frameworks and combine it with the perspective of ML applications in medicine. As a result, we propose the METRIC-framework, a specialised data quality framework for medical ML data comprising 15 awareness dimensions divided into five clusters: *measurement process*, *timeliness*, *representativeness*, *informativeness*, *consistency*. The

METRIC-framework is intended for assessing the suitability of the content of a medical dataset for a given ML task.



Knowledge about the composition of datasets may:

- Reduce biases
- Increase robustness
- Facilitate interpretability
- Lay the **foundation for trustworthy AI in medicine**

Potential applications:

- Validating training data (conformity assessment, audit)
- Constructing reference datasets
- Designing test data

Mapping to EU AI Act Shows Regulatory Compliance

The mandatory AI-specific conformity assessment described in the EU AI Act (Articles 8 to 15) lays down data quality requirements (Article 10, paragraphs 3 and 4). It tasks, among others,

metrology institutes with developing measurement methodologies for technical validation (Article 15, paragraph 2).

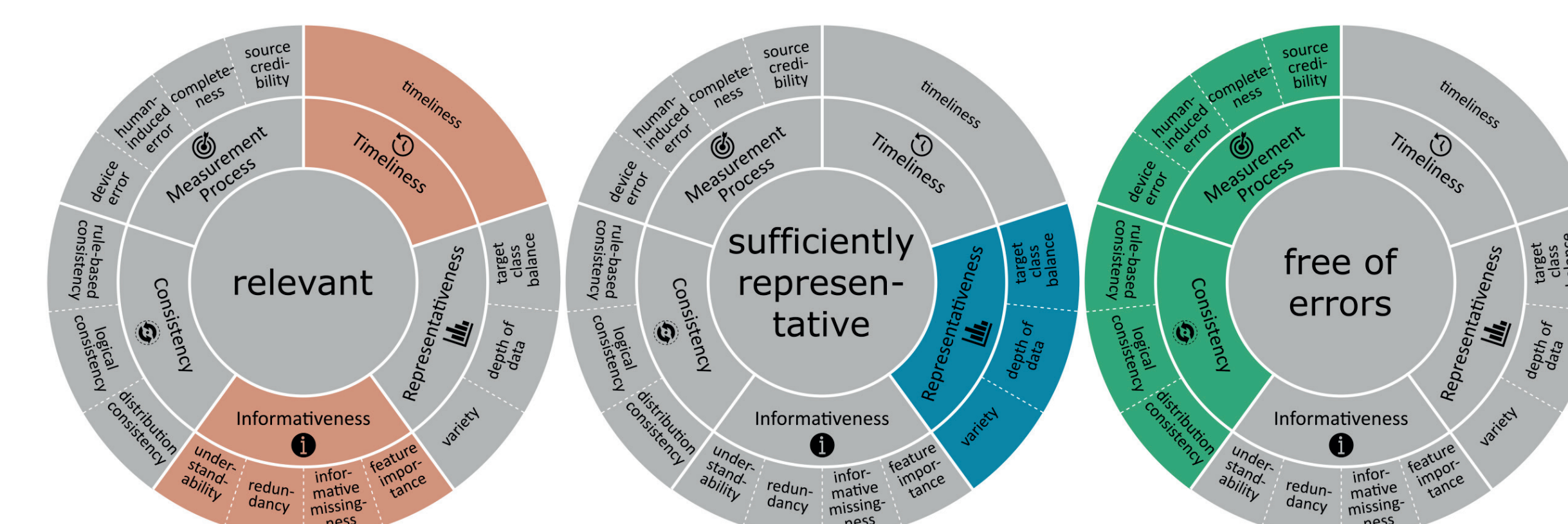


Article 10: Data and data governance

3. Training, validation and testing data sets shall be **relevant, sufficiently representative, and to the best extent possible, free of errors** and **complete in view of the intended purpose**. They shall have the **appropriate statistical properties**, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used.[...]

We find the EU AI Act data requirements to map to the METRIC-framework in a distinct and comprehensive way.

Evaluations of datasets that utilize the **METRIC-framework thus also show compliance with the EU AI Act**.



Conclusion & Outlook

The METRIC-framework is ideally positioned to serve as a **starting point for quality assurance** of AI systems. It forms a base for systematically assessing training datasets during audits, for establishing reference datasets, and for designing test datasets. The harmonization of such evaluations has great **potential to accelerate the approval of AI medical devices** with respect to the EU AI Act.

The next steps of our work include:

- the **development of an automatable tool** for characterising the quality of medical ML data and
- the **quantification of the influence data quality** has on characteristics of ML systems (e.g., predictive performance, robustness, uncertainty quant., explainability, fairness).

References

- [1] Schwabe, Becker, Seyferth, Klaß, Schäffter, *The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review*. npj Digit. Med. 7, 203 (2024). <https://doi.org/10.1038/s41746-024-01196-4>
- [2] Regulation - EU - 2024/1689 - EN - EUR-Lex ("EU AI Act"). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [3] TEF-Health - Testing and Experimentation Facility for Health AI and Robotics. <https://tefhealth.eu/home>

