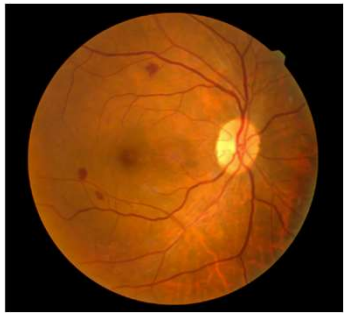
The METRIC-Framework: Vertrauenswürdige KI-Daten für die Medizin

Katinka Becker, Physikalisch-Technische Bundesanstalt (PTB), Berlin

PTB

Einsatz von KI im Medizinbereich

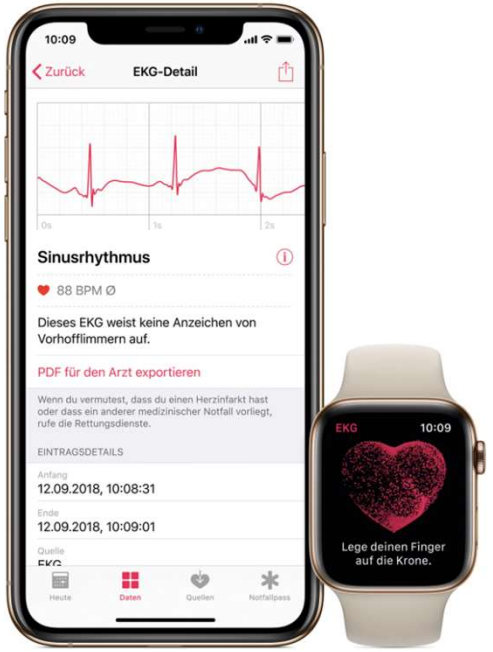


Diabetische Retinopathie

Graphic: aerzteblatt.de

Kardiovaskuläres Ultraschallsystem

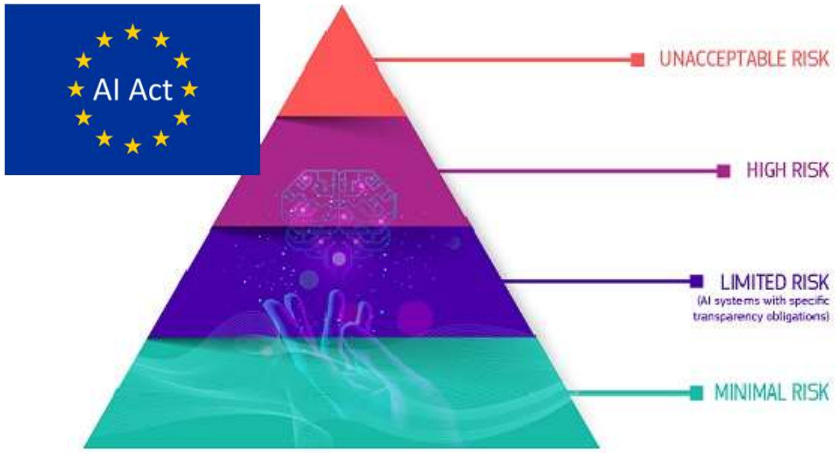
Graphic: philips.de



Apple Watch und Health App

Graphic: apple.com

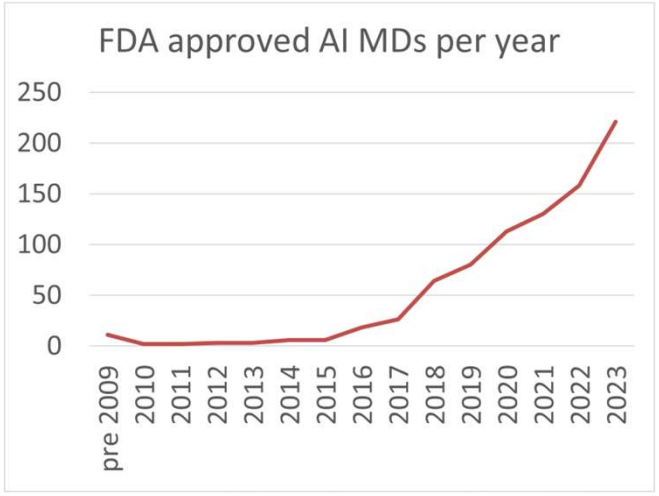
Konformitätsbewertung für KI Medizinprodukte

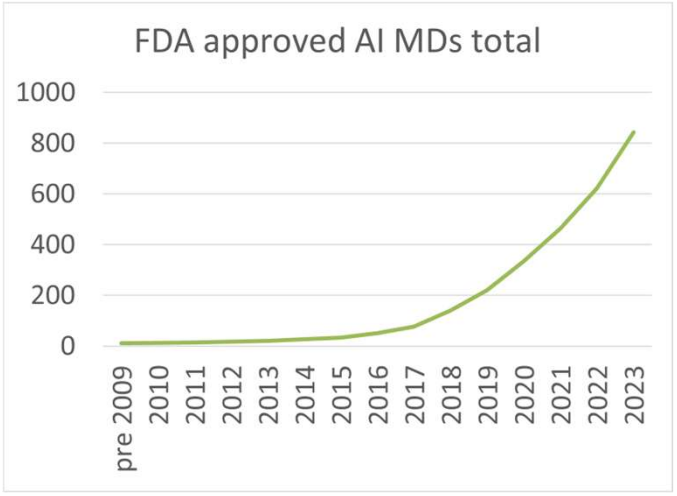


KI Medizinprodukte

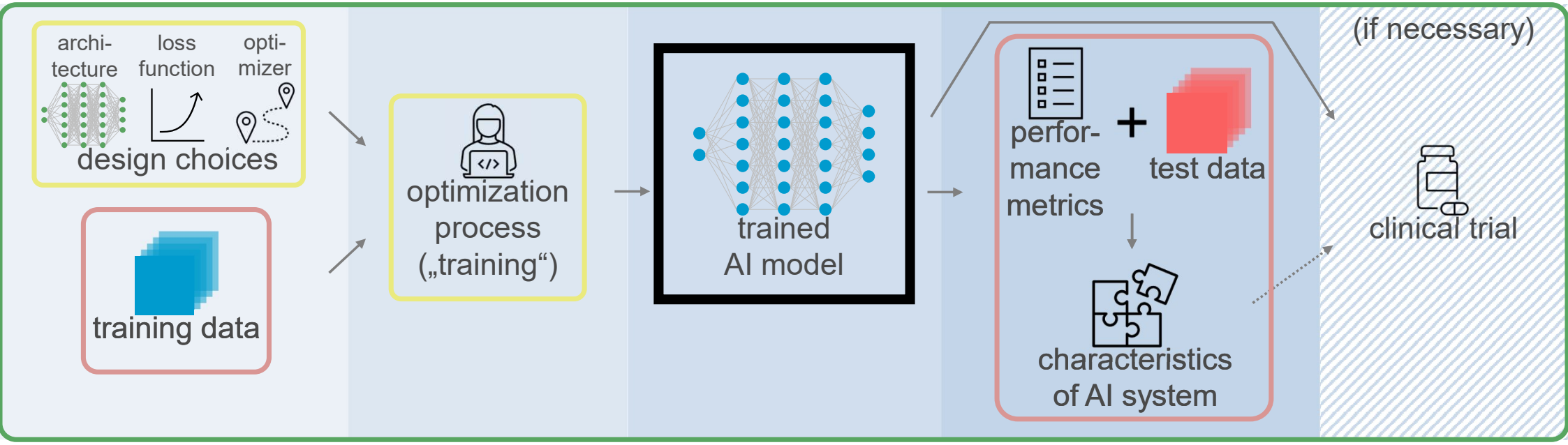
MDR

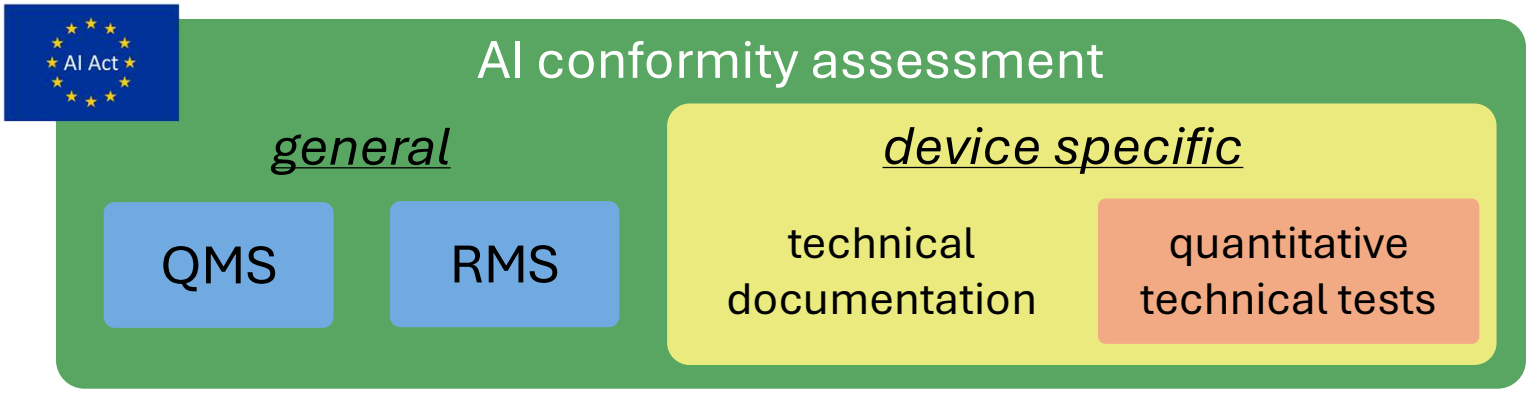
KI Konformitätsbewertung





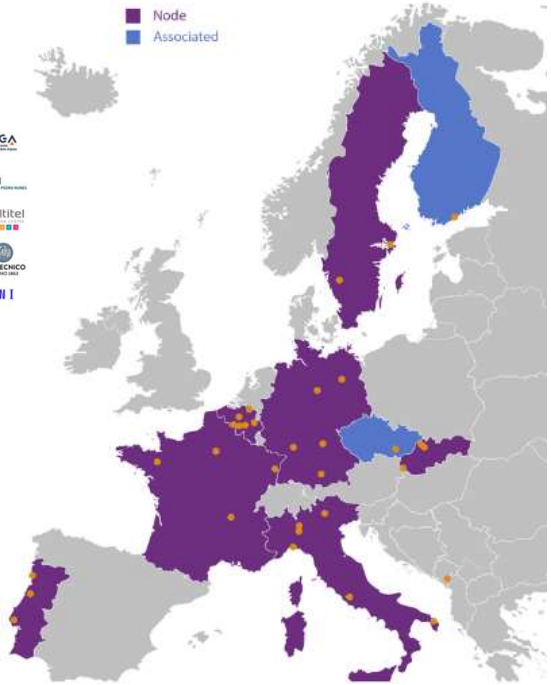
Quantitatives Testen von KI im Medizinbereich





Das EU Projekt TEF- Health

- „Testing and Experimentation Facility for Health AI and Robotics“
- 52 Partner in der EU
- Leitung: Prof. Petra Ritter, Charité
- PTB leitet das Workpackage „Standards and Quality“

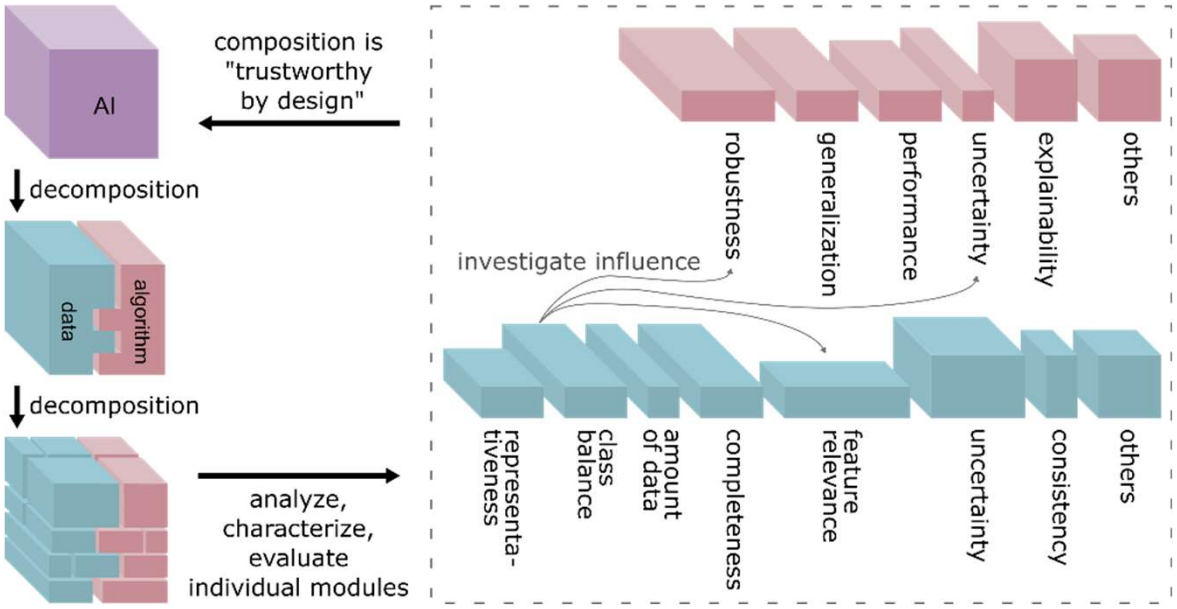


Ziel

Vertrauenswürdige KI im Medizinbereich:
Markteintritt erleichtern
⇕
Qualität sichern

© Boris Nalbach

Vertrauenswürdigkeit per Design



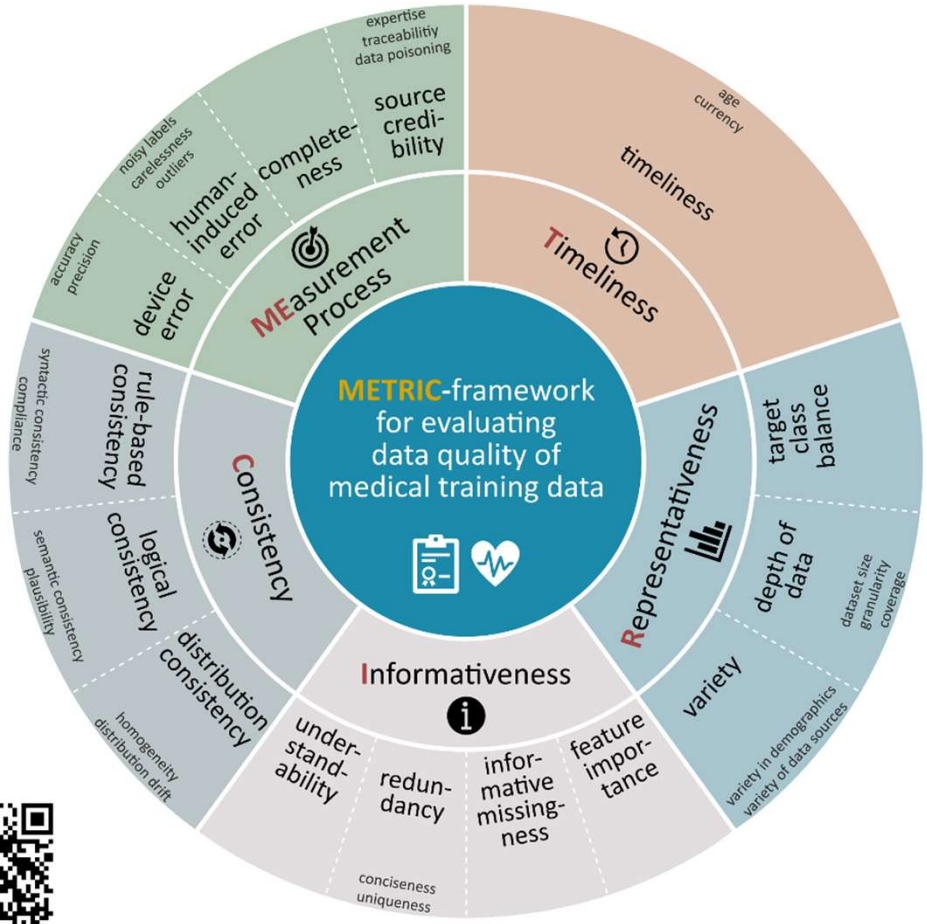
- KI System unterteilt sich in zwei entscheidende Komponenten:
  - **Daten** und **Algorithmus**

Datenqualität

Idee

Qualitätssicherung des Systems durch Qualitätssicherung der Komponenten

Das METRIC-framework



Systematic Review von ~5400 Artikeln (120 included)

Das „Rad der Datenqualität“

- 5 Cluster mit 15 „Awareness Dimensions“
- Grundlage für **systematische Evaluierung von Qualität von medizinischen Daten** für z.B.:
  - Referenzdatensätze
  - Design von Testdatensätzen
  - Validierung von Trainingsdatensätzen

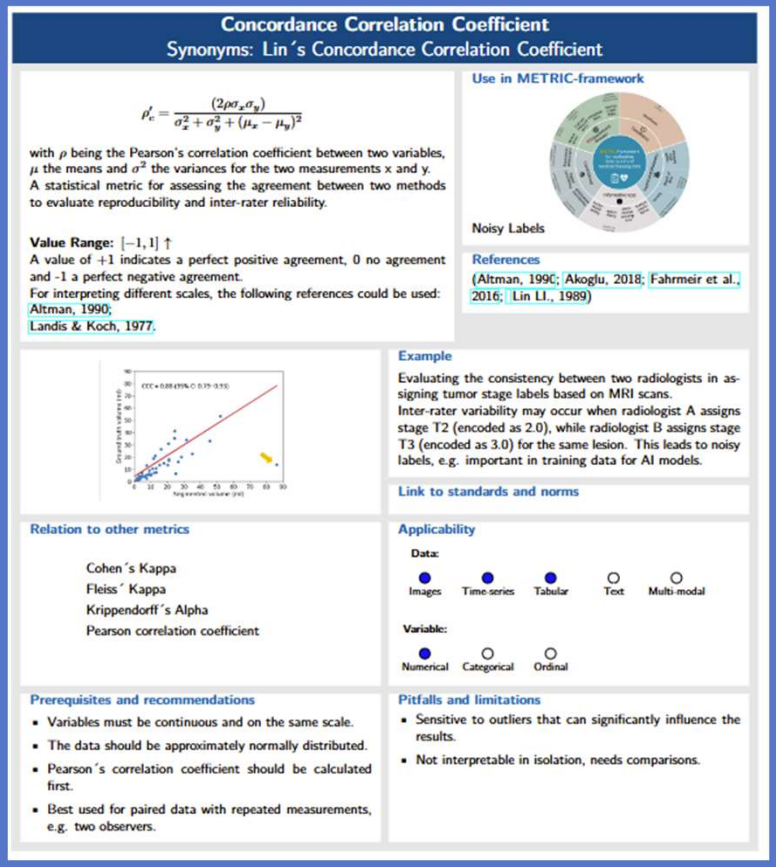
AI Act Mapping

AI Act Article 10, Paragraph 3, fordert:

„...data sets shall be...“:

- „relevant“
- „sufficiently representative“
- „to the best extent possible, free of errors”

Von der Theorie in die Praxis



Concordance Correlation Coefficient
Synonyms: Lin´s Concordance Correlation Coefficient

$$\rho_c' = \frac{(2\rho\sigma_x\sigma_y)}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

with ρ being the Pearson's correlation coefficient between two variables, μ the means and σ^2 the variances for the two measurements x and y.
A statistical metric for assessing the agreement between two methods to evaluate reproducibility and inter-rater reliability.

Value Range: $[-1, 1]$ ↑
A value of +1 indicates a perfect positive agreement, 0 no agreement and -1 a perfect negative agreement.
For interpreting different scales, the following references could be used: Altman, 1990; Landis & Koch, 1977.

Use in METRIC-framework

Noisy Labels

References
(Altman, 1990; Akoglu, 2018; Fahrmeir et al., 2016; Lin LI., 1989)

Example
Evaluating the consistency between two radiologists in assigning tumor stage labels based on MRI scans.
Inter-rater variability may occur when radiologist A assigns stage T2 (encoded as 2.0), while radiologist B assigns stage T3 (encoded as 3.0) for the same lesion. This leads to noisy labels, e.g. important in training data for AI models.

Link to standards and norms

Relation to other metrics
Cohen´s Kappa
Fleiss´ Kappa
Krippendorff´s Alpha
Pearson correlation coefficient

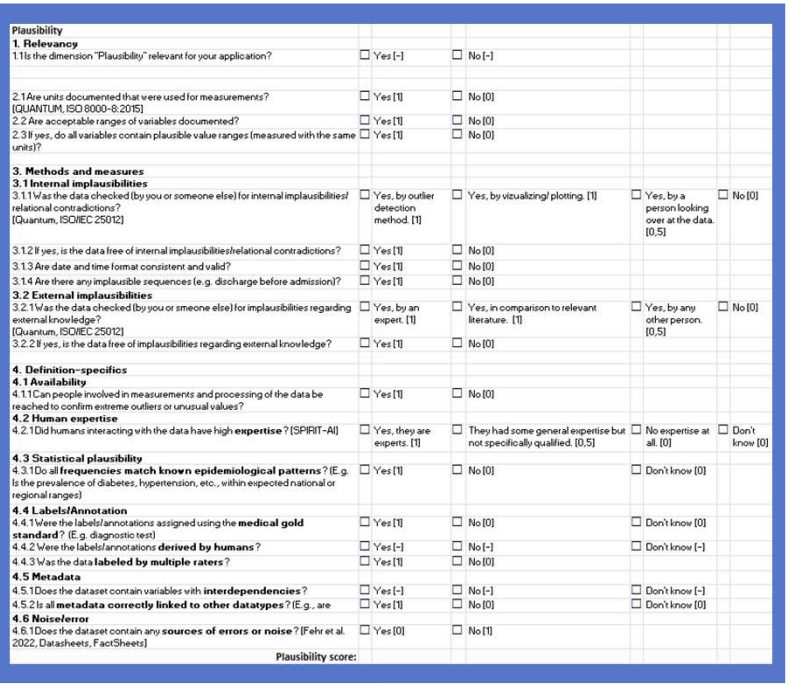
Applicability
Data: Images, Time-series, Tabular, Text, Multi-modal
Variable: Numerical, Categorical, Ordinal

Prerequisites and recommendations
- Variables must be continuous and on the same scale.
- The data should be approximately normally distributed.
- Pearson´s correlation coefficient should be calculated first.
- Best used for paired data with repeated measurements, e.g. two observers.

Pitfalls and limitations
- Sensitive to outliers that can significantly influence the results.
- Not interpretable in isolation, needs comparisons.

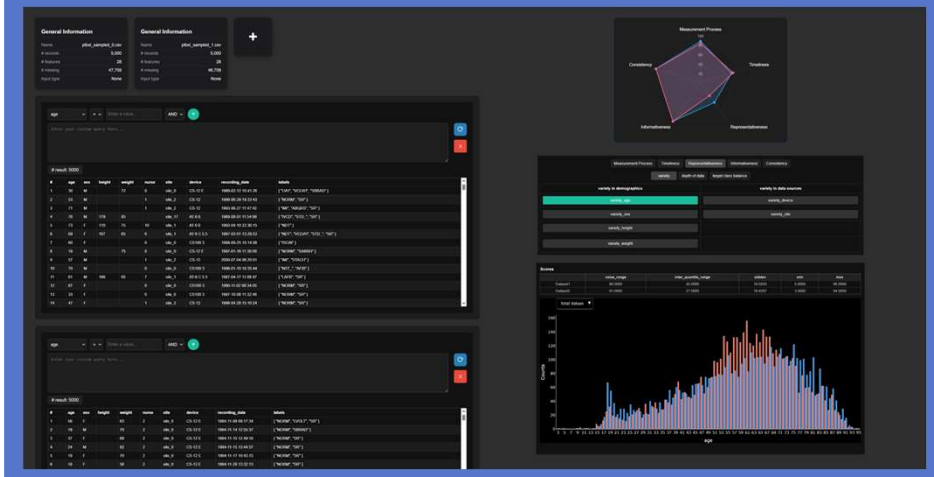
- **Quantitative Dimensionen „messbar machen“** mithilfe einer Metrikensammlung
- Website in Planung

- **Qualitative Dimensionen „messbar machen“** mithilfe von Fragebögen



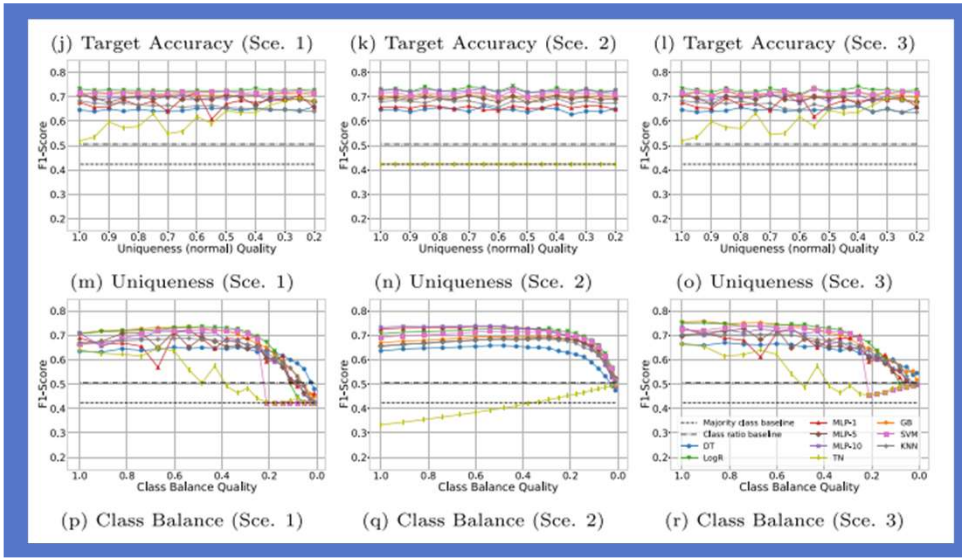
Plausibility					
1. Relevancy					
1.1 Is the dimension "Plausibility" relevant for your application?	☐ Yes [-]	☐ No [-]			
2.1 Are units documented that were used for measurements? [QUANTUM, ISO 8000-8:2015]	☐ Yes [1]	☐ No [0]			
2.2 Are acceptable ranges of variables documented?	☐ Yes [1]	☐ No [0]			
2.3 If yes, do all variables contain plausible value ranges (measured with the same units)?	☐ Yes [1]	☐ No [0]			
3. Methods and measures					
3.1 Internal implausibilities					
3.1.1 Was the data checked (by you or someone else) for internal implausibilities/ relational contradictions? [Quantum, ISO/IEC 25012]	☐ Yes, by outlier detection method. [1]	☐ Yes, by vizualizing/ plotting. [1]	☐ Yes, by a person looking over at the data. [0,5]	☐ No [0]	
3.1.2 If yes, is the data free of internal implausibilities/relational contradictions?	☐ Yes [1]	☐ No [0]			
3.1.3 Are date and time format consistent and valid?	☐ Yes [1]	☐ No [0]			
3.1.4 Are there any implausible sequences (e.g. discharge before admission)?	☐ Yes [1]	☐ No [0]			
3.2 External implausibilities					
3.2.1 Was the data checked (by you or smeone else) for implausibilities regarding external knowledge? [Quantum, ISO/IEC 25012]	☐ Yes, by an expert. [1]	☐ Yes, in comparison to relevant literature. [1]	☐ Yes, by any other person. [0,5]	☐ No [0]	
3.2.2 If yes, is the data free of implausibilities regarding external knowledge?	☐ Yes [1]	☐ No [0]			
4. Definition-specifics					
4.1 Availability					
4.1.1 Can people involved in measurements and processing of the data be reached to confirm extreme outliers or unusual values?	☐ Yes [1]	☐ No [0]			
4.2 Human expertise					
4.2.1 Did humans interacting with the data have high **expertise**? [SPIRIT-AI]	☐ Yes, they are experts. [1]	☐ They had some general expertise but not specifically qualified. [0,5]	☐ No expertise at all. [0]	☐ Don't know [0]	
4.3 Statistical plausibility					
4.3.1 Do all **frequencies match known epidemiological patterns**? (E.g. Is the prevalence of diabetes, hypertension, etc., within expected national or regional ranges)	☐ Yes [1]	☐ No [0]	☐ Don't know [0]		
4.4 Labels/Annotation					
4.4.1 Were the labels/annotations assigned using the **medical gold standard**? (E.g. diagnostic test)	☐ Yes [1]	☐ No [0]	☐ Don't know [0]		
4.4.2 Were the labels/annotations **derived by humans**?	☐ Yes [-]	☐ No [-]	☐ Don't know [-]		
4.4.3 Was the data **labeled by multiple raters**?	☐ Yes [1]	☐ No [0]			
4.5 Metadata					
4.5.1 Does the dataset contain variables with **interdependencies**?	☐ Yes [-]	☐ No [-]	☐ Don't know [-]		
4.5.2 Is all **metadata correctly linked to other datatypes**? (E.g., are	☐ Yes [1]	☐ No [0]	☐ Don't know [0]		
4.6 Noise/error					
4.6.1 Does the dataset contain any **sources of errors or noise**? [Fehr et al. 2022, Datasheets, FactSheets]	☐ Yes [0]	☐ No [1]			
Plausibility score:					

Von der Theorie in die Praxis



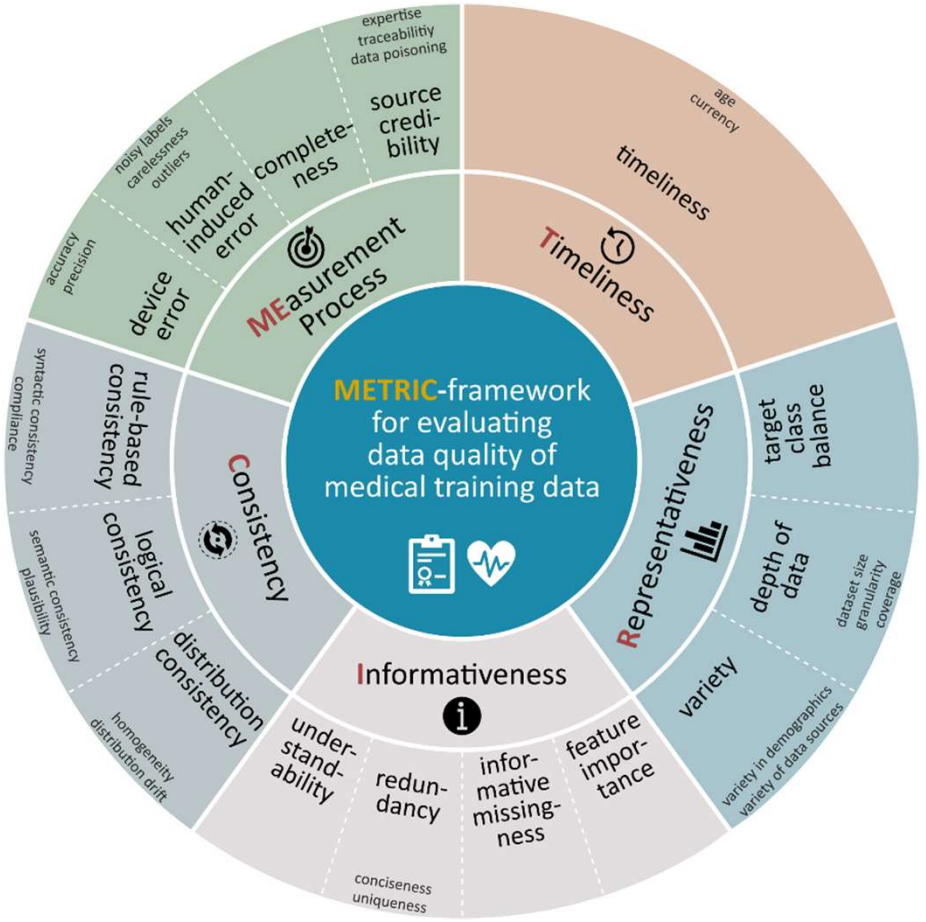
Tool:

- **Evaluierung und Vergleich** von Datensätzen
- **Perspektivisch: Kuratieren** von Datensätzen



- Wie stark wirken sich **einzelne Datenqualitätsdimensionen** auf die Qualität des Systems aus (Performance, Robustheit, Fairness, Erklärbarkeit,...) ?

Danke


METRIC-framework for evaluating data quality of medical training data
MEasurement Process
source credi-bility
expertise
traceability
data poisoning
complete-ness
human-induced error
noisy labels
carelessness
outliers
device error
accuracy
precision
Timeliness
timeliness
age
currency
Representativeness
target class balance
depth of data
dataset size
granularity
coverage
variety
variety in demographics
variety of data sources
Informativeness
feature impor-tance
infor-mative missing-ness
redun-dancy
under-stand-ability
conciseness
uniqueness
Consistency
distribution consistency
homogeneity
distribution drift
logical consistency
semantic consistency
plausibility
rule-based consistency
syntactic consistency
compliance