Reference Data and Benchmarktest for AI

17-October-2024

Tobias Schäffter

Physikalisch-Technische Bundesanstalt
The National Metrology Institute

PTB

Federal Ministry
for Economic Affairs
and Climate Action

Nobel Prizes in Physics and Chemistry 2024

The Nobel Prize in Physics 2024

Ill. Niklas Elmehed © Nobel Prize Outreach

John J. Hopfield

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Geoffrey E. Hinton

Prize share: 1/2

The Nobel Prize in Physics 2024 was awarded jointly to John J. Hopfield and Geoffrey E. Hinton "for foundational discoveries and inventions that enable machine learning with artificial neural networks"

MLA style: The Nobel Prize in Physics 2024. NobelPrize.org. Nobel Prize Outreach AB 2024. Thu. 10 Oct 2024. https://www.nobelprize.org/prizes/physics/2024/summary/

The Nobel Prize in Chemistry 2024

Ill. Niklas Elmehed © Nobel Prize Outreach

David Baker

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

Prize share: 1/4

Ill. Niklas Elmehed © Nobel Prize Outreach

John M. Jumper

Prize share: 1/4

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"

MLA style: The Nobel Prize in Chemistry 2024. NobelPrize.org. Nobel Prize Outreach AB 2024. Thu. 10 Oct 2024. https://www.nobelprize.org/prizes/chemistry/2024/summary/

Generative Pre-trained Transformer (GPT) - ChatGPT

PTB

Chat GPT - Productivity

Science article

ChatGPT gives an extra productivity boost to weaker writers

Abstract

We examined the productivity effects of a generative artificial intelligence (AI) technology, the assistive chatbot ChatGPT, in the context of midlevel professional writing tasks. In a preregistered online experiment, we assigned occupation-specific, incentivized writing tasks to 453 college-educated professionals and randomly exposed half of them to ChatGPT. Our results show that ChatGPT substantially raised productivity: The average time taken decreased by 40% and output quality rose by 18%. Inequality between workers decreased, and concern and excitement about AI temporarily rose. Workers exposed to ChatGPT during the experiment were 2 times as likely to report using it in their real job 2 weeks after the experiment and 1.6 times as likely 2 months after the experiment.

NEWS CAREERS COMMENTARY JOURNALS | COVID-19 Science LOG IN BECOME A MEMBER

NEWS | TECHNOLOGY

AI learns to write computer code in 'stunning' advance

DeepMind's AlphaCode outperforms many human programmers in tricky software challenges

8 DEC 2022 · 2:00 PM · BY MATTHEW HUTSON

Chat GPT- „Dementia“

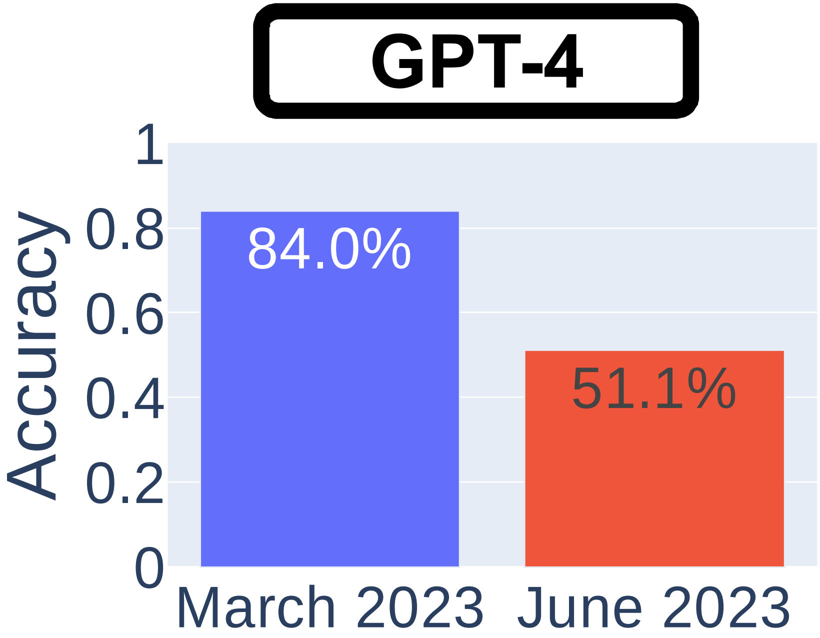
How Is ChatGPT's Behavior Changing over Time?

Lingjiao Chen[†], Matei Zaharia[‡], James Zou[†]

[†]Stanford University [‡]UC Berkeley

Benchmarking

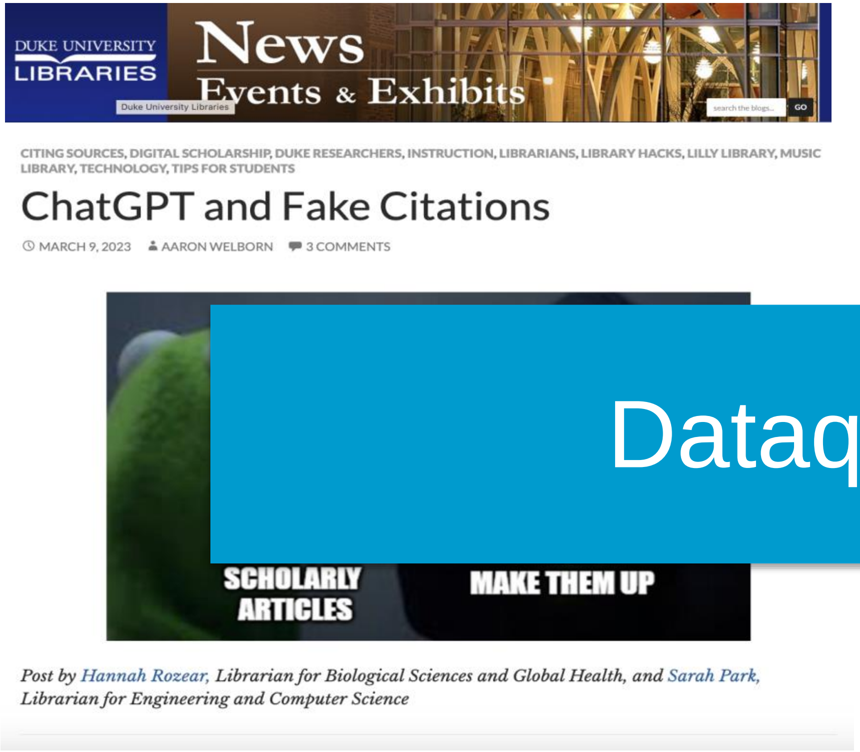
GPT-3.5 and GPT-4
However, when and how
March 2023 and June 2
problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4's amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5's performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. Overall, our findings show that the behavior of the "same" LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.



Is 17077 a prime number? Think step by step and then answer [Yes] or [No].

arXiv:2307.09009; Jul 2023

ChatGPT - „Hallucination“



Dataquality

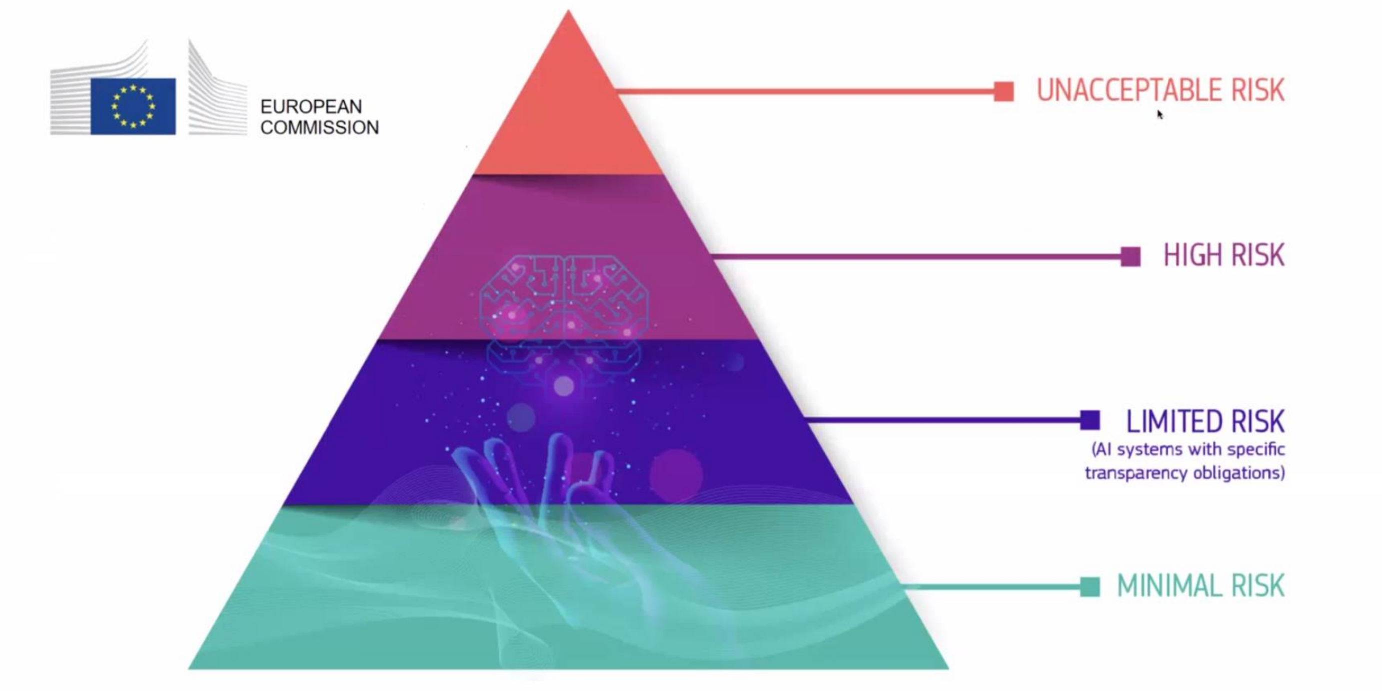
Misleading information due to overfitting, high model complexity and training data quality.

EU-AI Act for Trustworthy AI

Trust in "algorithms" that are not fully understood ("black box"), especially for high risk applications

The trustworthy AI strongly depends on

high quality trainings data



Certification of **AI-Quality** requires:

- robustness,
- accuracy,
- security,
- Explainability.

Data Quality

EU Artificial Intelligence Act

Article 10: Data and Data Governance

Date of entry into force: **2 August 2026**

According to: **Article 113**

See here for a full implementation timeline.

SUMMARY –

This article states that high-risk AI systems must be developed using high-quality data sets for training, validation, and testing. These data sets should be managed properly, considering factors like data collection processes, data preparation, potential biases, and data gaps. The data sets should be relevant, representative, error-free, and complete as much as possible. They should also consider the specific context in which the AI system will be used. In some cases, providers may process special categories of personal data to detect and correct biases, but they must follow strict conditions to protect individuals' rights and freedoms.

Generated by CLaiRK, edited by us.

1. High-risk AI systems which make use of techniques involving the training of AI

Accuracy Robustness, Security

EU Artificial Intelligence Act

Article 15: Accuracy, Robustness and Cybersecurity

1. High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and perform consistently in those respects throughout their lifecycle.

   1a. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 of this Article and any other relevant performance metrics, the Commission shall, in **cooperation with relevant stakeholder and organisations such as metrology and benchmarking authorities**, encourage as appropriate, **the development of benchmarks and measurement methodologies.**

2. The levels of accuracy and the relevant **accuracy metrics** of high-risk AI systems shall be declared in the accompanying instructions of use.

3. High-risk AI systems shall be as resilient as possible regarding **errors, faults or inconsistencies** that may occur within the system ......

Testing und Experimentation Facilities

TEFs are **specialised large-scale reference sites open to all technology providers across Europe.** Their objective is to **support AI developers to bring trustworthy and secure AI to the European market.**

Co-funding between the European Commission (through the Digital Europe Programme) and the Member States will support the TEFs for five years with budgets between EUR 40-60 million per project. TEFs will focus on four high-impact sectors:

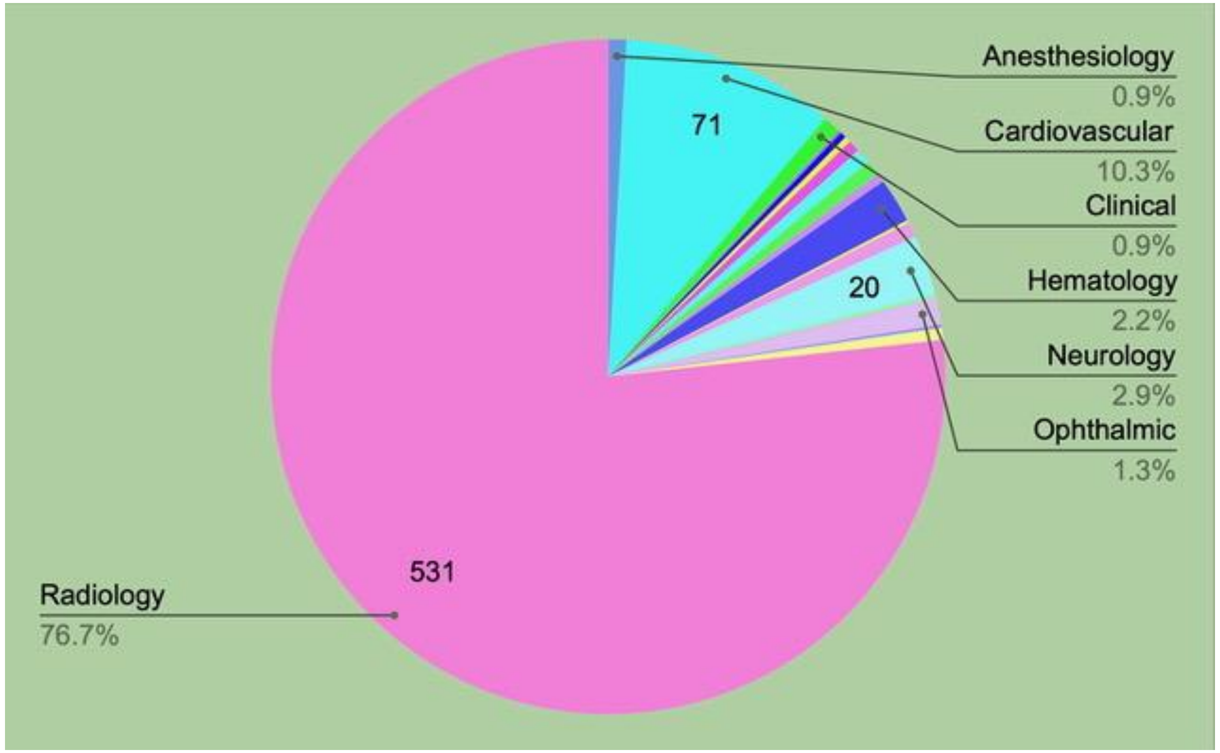
- Agri-Food: project “agrifoodTEF”
- Healthcare: project “TEF-Health”
- Manufacturing: project “AI-MATTERS”
- Smart Cities & Communities: project “Citcom.AI”

FDA-approved Medical Products with AI

FDA's new list (Oct, 2023) with 692 devices:

- 77% are in Radiology: 531 devices
- 10% are in Cardiovascular: 71 devices
- 3% are in Neurology: 20 devices
- 2% are in Hematology: 15 devices

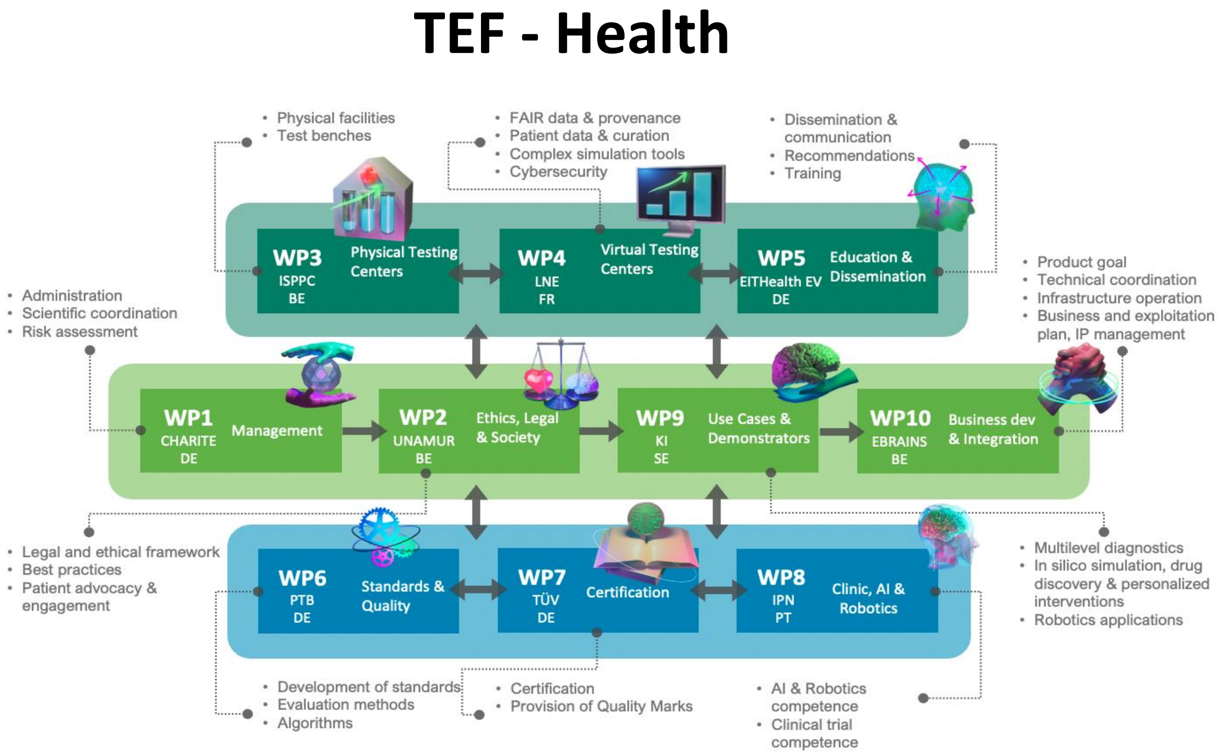
Additional 171 medical devices in 2024 (33% increase)

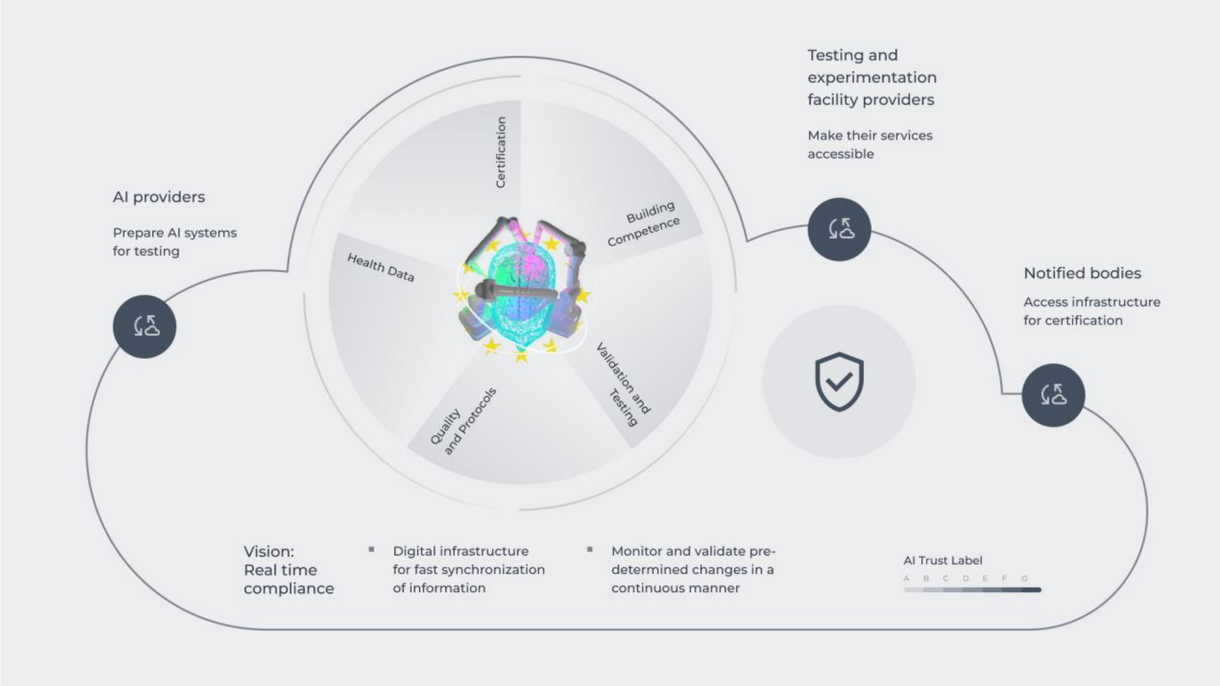


692 authorized AI-enabled devices by specialty.
Image source Margaretta Colangelo; 2023

https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

Testing und Experimentation Facility - Health





WP6 & 7: Agile Certification
(PTB, Fraunhofer, TÜV, LNE, KTH, Charité)

Data is the base of AI

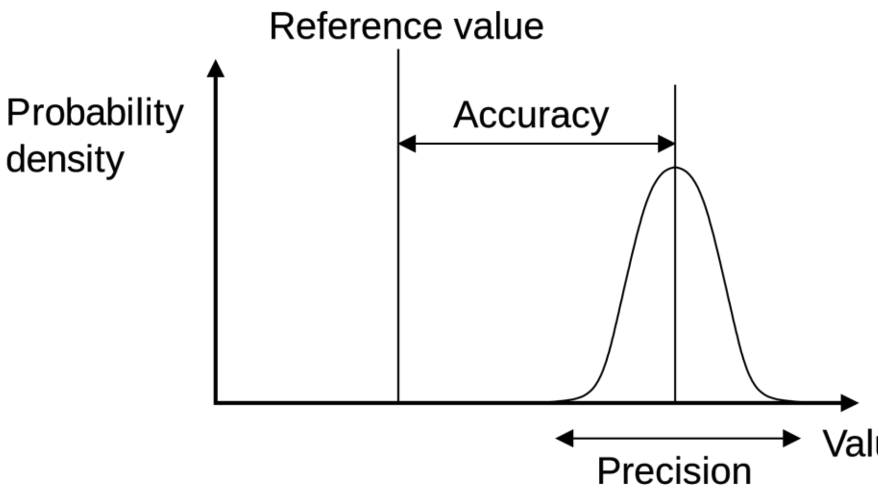
The quality of AI strongly depends on

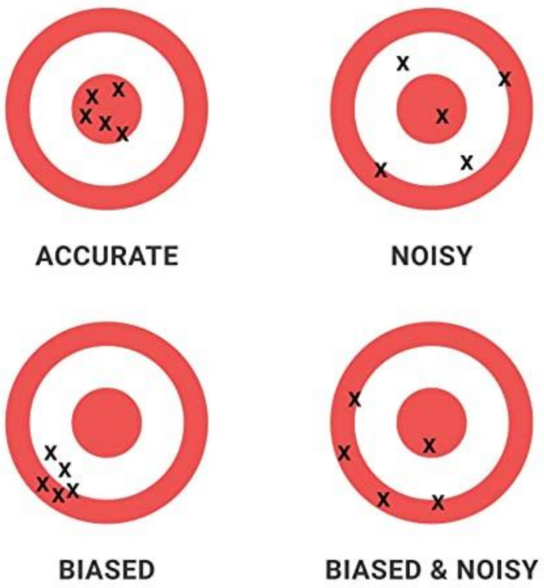
- Data uncertainty ("noise", "bias")
- Annotation inconsistencies ("label noise")

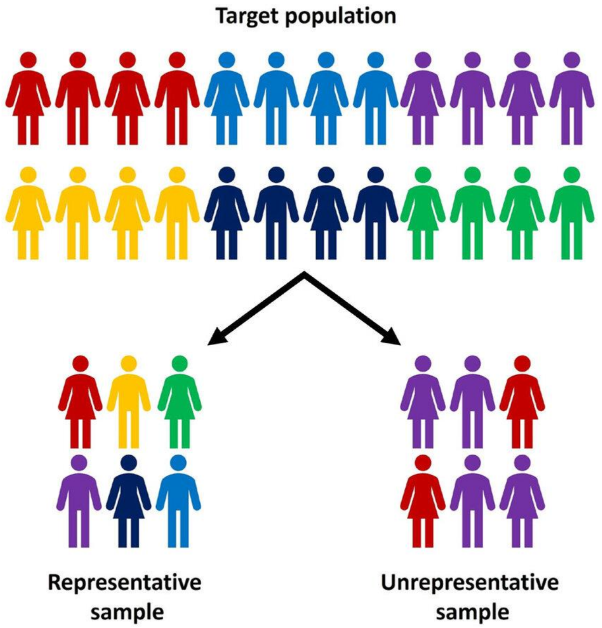
Standards for Data Quality

Uncertainty and Representativeness

- **Precision (Variability)** – closeness of measurements to each other
- **Accuracy *(Bias)*** - closeness of measurement results to a reference;
- **Representativeness** - accurate conclusions about a population from sample



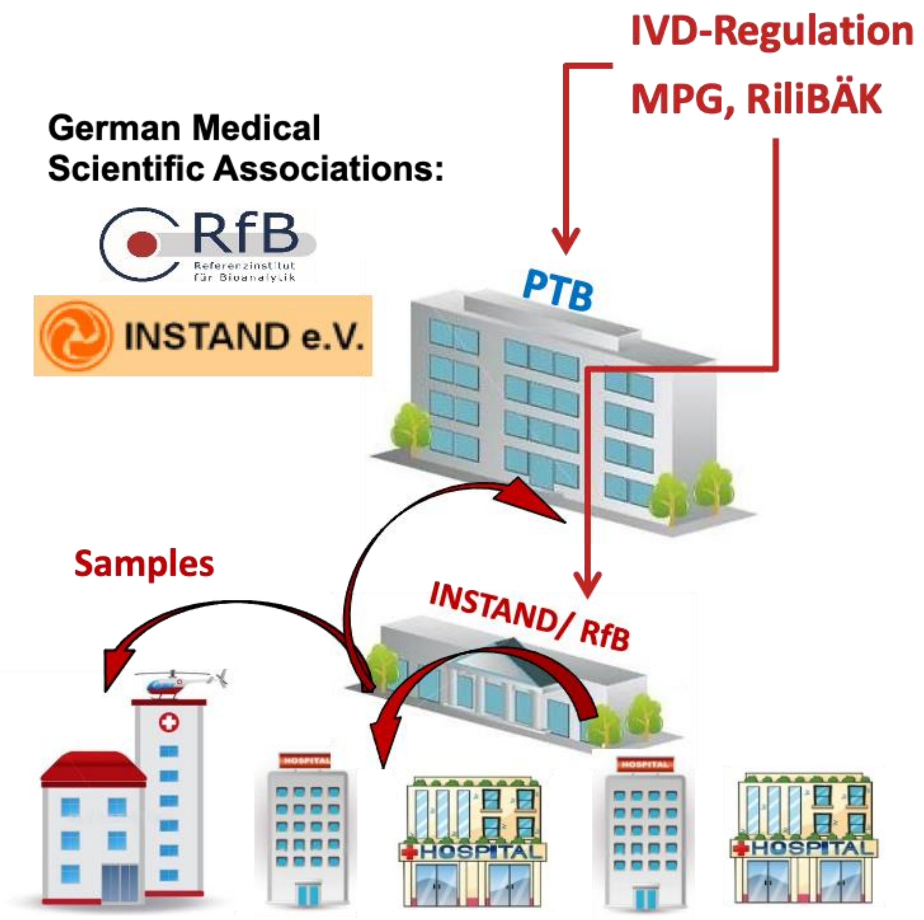


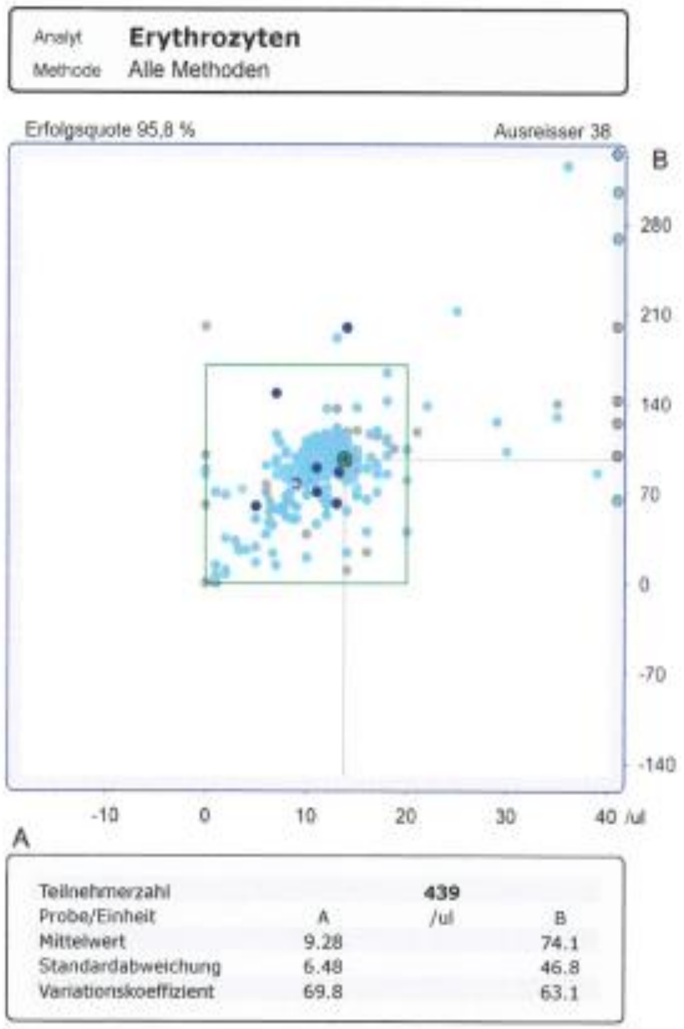


https://en.wikipedia.org/wiki/Accuracy_and_precision

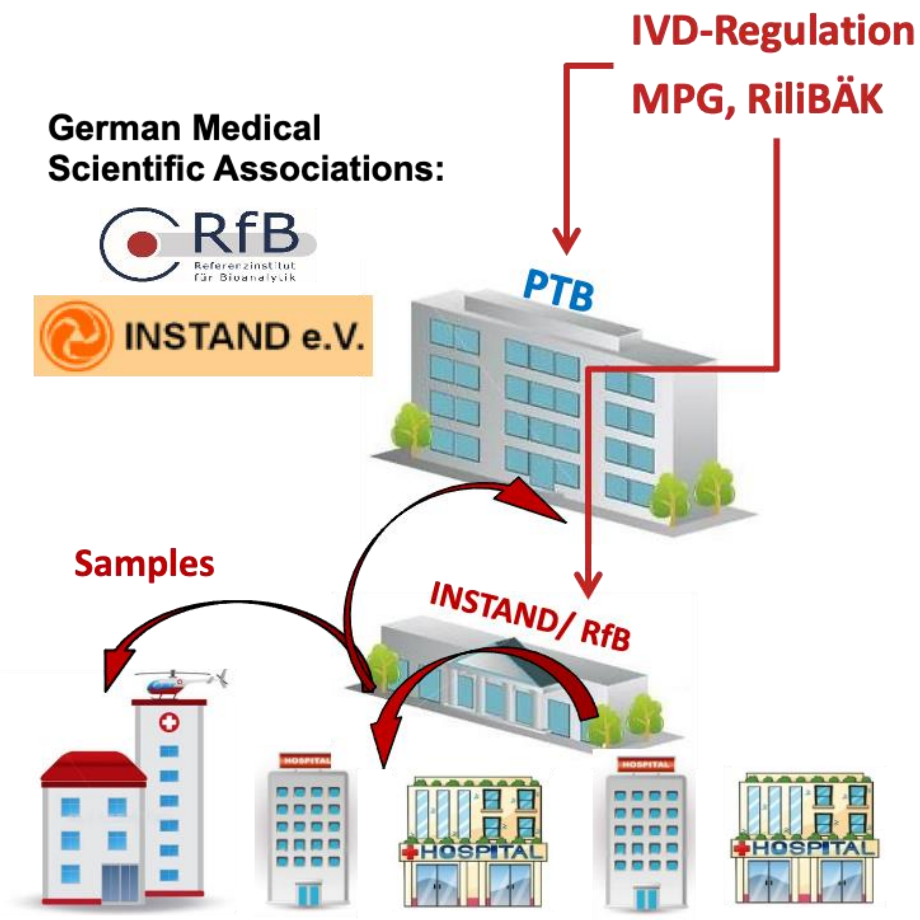
Koçak B. DOI:10.5152/dir.2022.211297

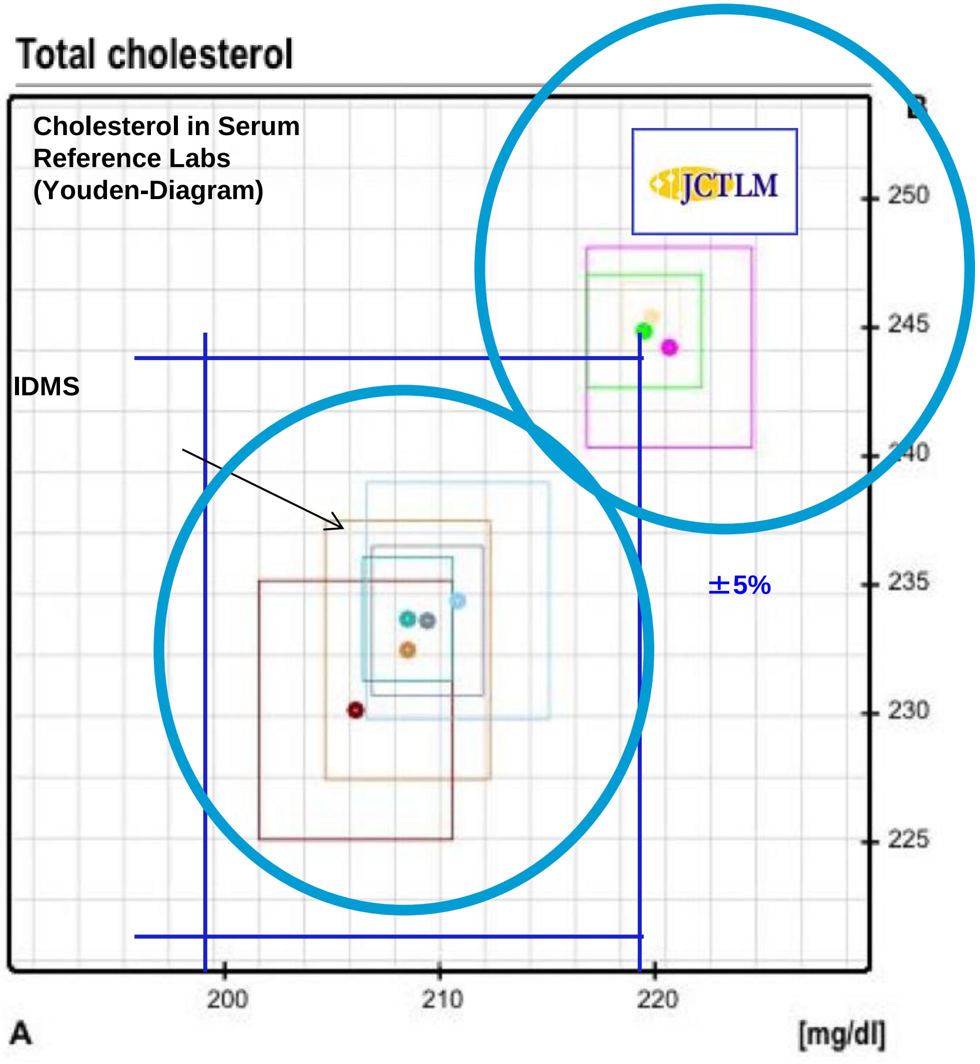
Data Quality relies on Metrology



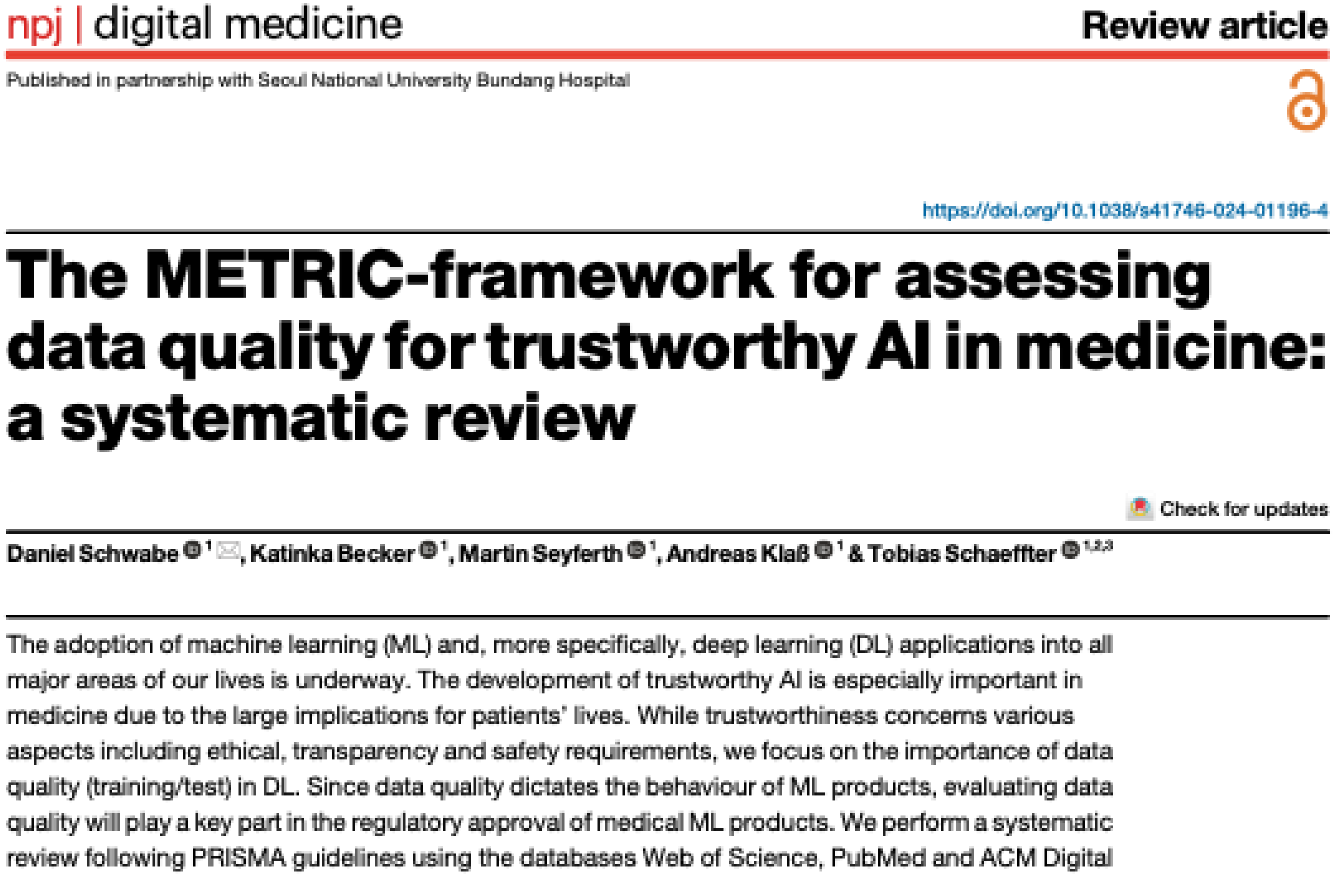


Data Quality relies on Metrology





Data Quality Dimensions – METRIC Framework



npj | digital medicine

Review article

Published in partnership with Seoul National University Bundang Hospital

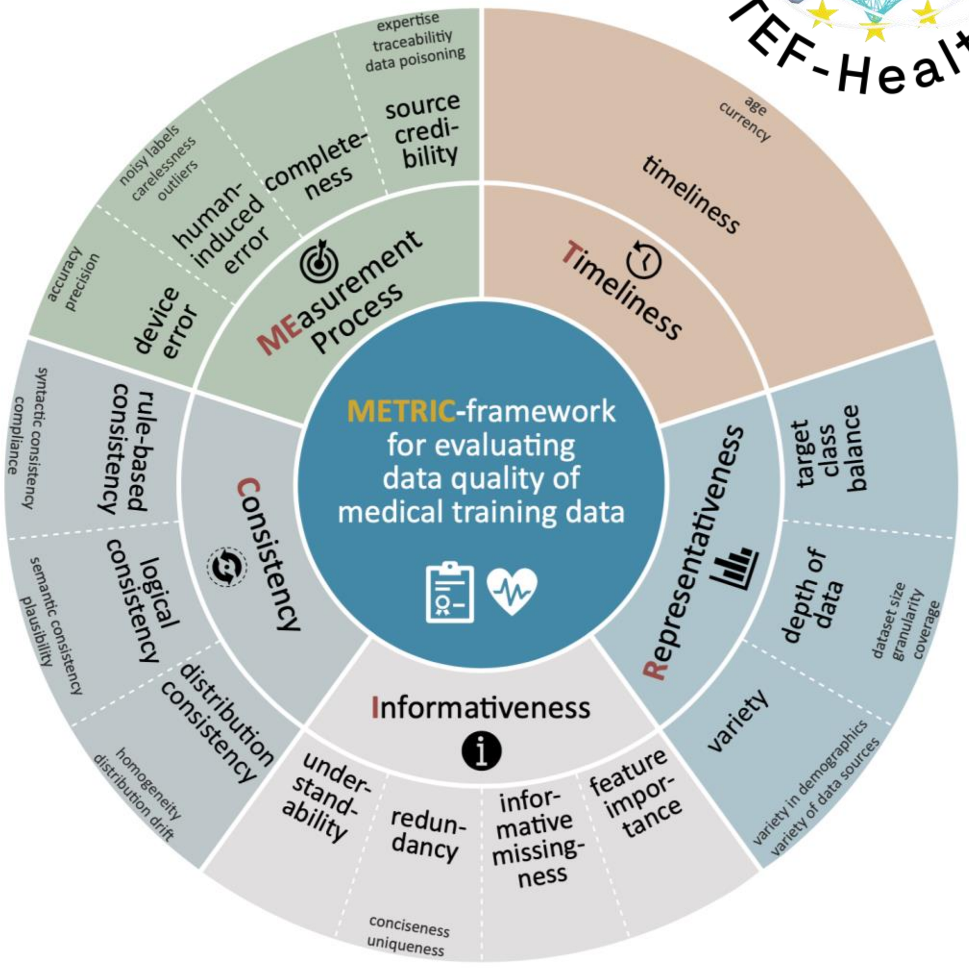
https://doi.org/10.1038/s41746-024-01196-4

The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review

Check for updates

Daniel Schwabe [1], Katinka Becker [1], Martin Seyferth [1], Andreas Klaß [1] & Tobias Schaeffter [1,2,3]

The adoption of machine learning (ML) and, more specifically, deep learning (DL) applications into all major areas of our lives is underway. The development of trustworthy AI is especially important in medicine due to the large implications for patients' lives. While trustworthiness concerns various aspects including ethical, transparency and safety requirements, we focus on the importance of data quality (training/test) in DL. Since data quality dictates the behaviour of ML products, evaluating data quality will play a key part in the regulatory approval of medical ML products. We perform a systematic review following PRISMA guidelines using the databases Web of Science, PubMed and ACM Digital



Schwabe D et al. NPJ Digit Med. 2024 Aug 3;7(1):203.

ECG-Reference-Dataset: PTB-XL

Application:

- Over 300 Mill. ECGs per year
- Strong application of AI for automatic analysis of ECG (arrhythmia, infarkt, hypertrophy,..)

Reference-Data

- EKG-Quality
- Defined Training-, Validation and Testdata
- Distribution within pathologies „taking representativeness into account“

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

PTB-XL+, a comprehensive electrocardiographic feature dataset

Nils Strodthoff[1] ✉, Temesgen Mehari[2,3], Claudia Nagel[4], Philip J. Aston[5,6], Ashish Sundar[5], Claus Graff[7], Jørgen K. Kanters[8], Wilhelm Haverkamp[9], Olaf Dössel[4], Axel Loewe[4], Markus Bär[2] & Tobias Schaeffter[2,10,11] ✉

Machine learning (ML) methods for the analysis of electrocardiography (ECG) data are gaining importance, substantially supported by the release of large public datasets. However, these current datasets miss important derived descriptors such as ECG features that have been devised in the past hundred years and still form the basis of most automatic ECG analysis algorithms and are critical for cardiologists' decision processes. ECG features are available from sophisticated commercial software but are not accessible to the general public. To alleviate this issue, we add ECG features from two leading commercial algorithms and an open-source implementation supplemented by a set of automatic diagnostic statements from a commercial ECG analysis software in preprocessed format. This allows the comparison of ML models trained on clinically versus automatically generated label sets. We provide an extensive technical validation of features and diagnostic statements for ML applications. We believe this release crucially enhances the usability of the *PTB-XL* dataset as a reference dataset for ML methods in the context of ECG data.

Background & Summary

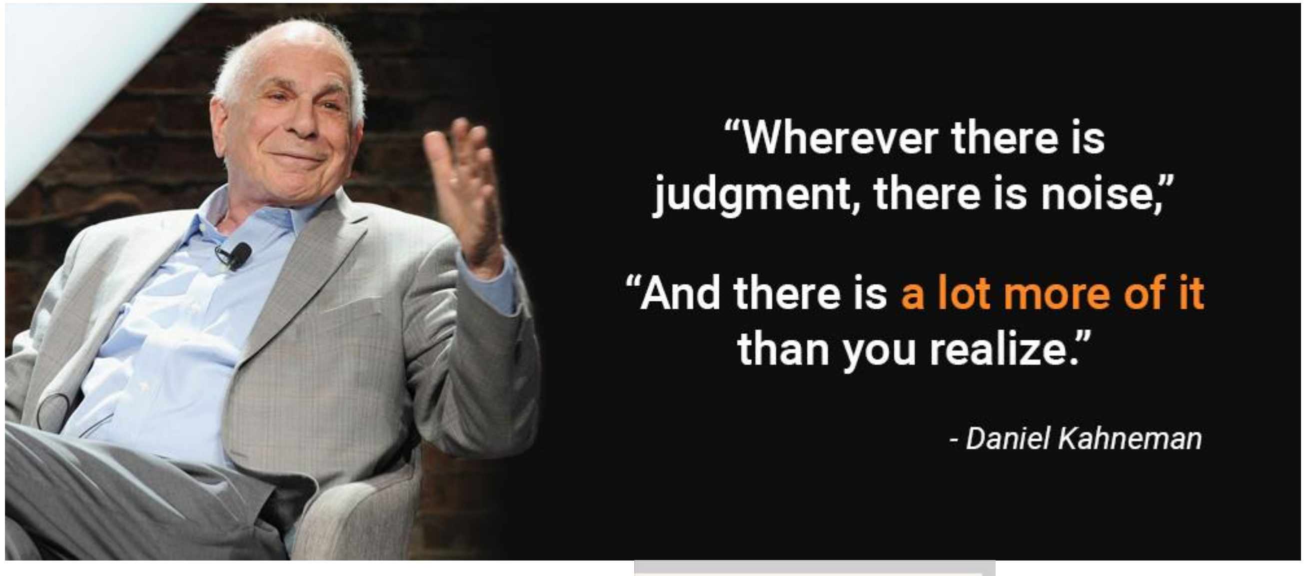
Cardiovascular diseases continue to be one of the largest burdens for the population worldwide[1]. Due to its simplicity, non-invasive nature, widespread use and diagnostic value, the electrocardiogram (ECG) is one of the primary tools for the first assessment. However, it requires the analysis of a huge amount of time-series ECG-data. Therefore automatic analysis tools have become standard. The recent developments in machine learning/AI have demonstrated its potential in this direction[2–5]. Large freely available ECG databases[6,7] are crucial for the development and benchmarking of AI algorithms for automatic classification. Consequently, they

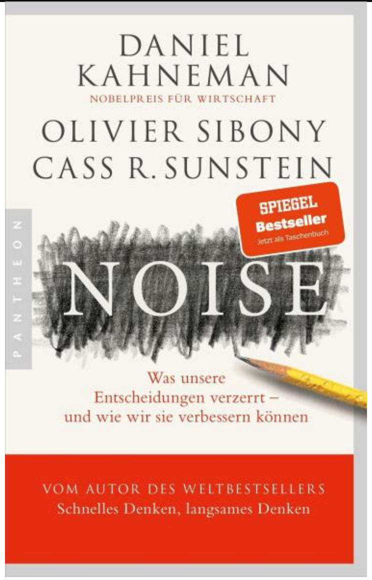
Fraunhofer HHI

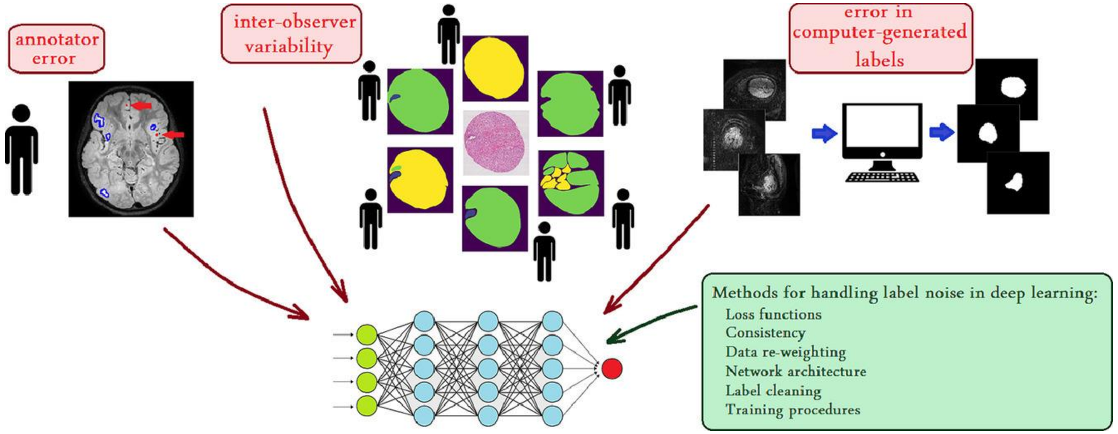
PTB

Figure 1: Graphical summary of the *PTB-XL* dataset in terms of diagnostic superclasses and subclasses, see Table 5 for a definition of the used acronyms.

Label Uncertainty – “the human factor”

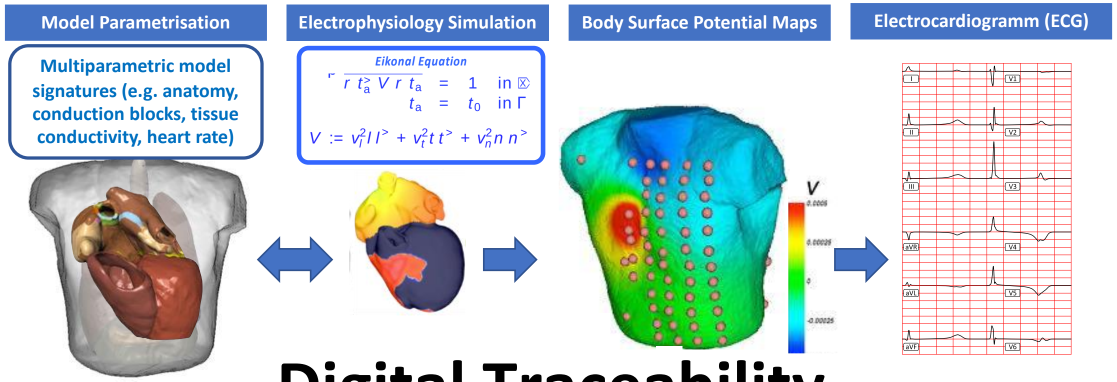




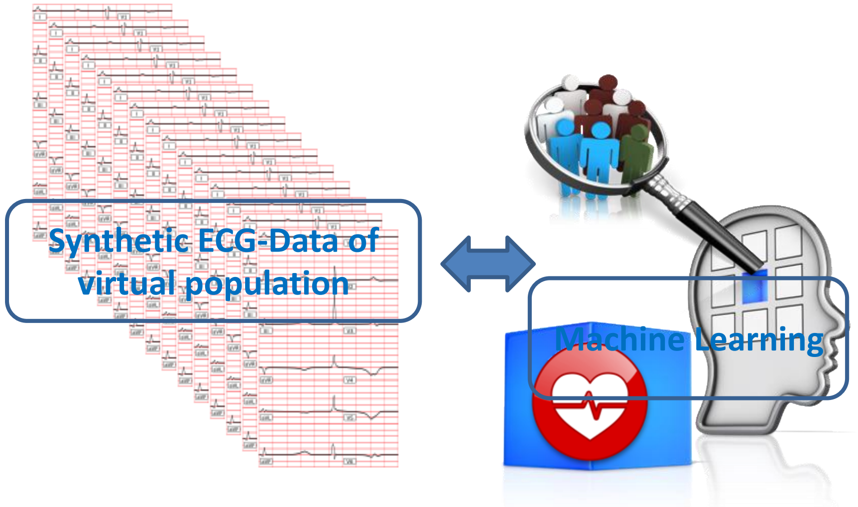


Karimi D et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med Image Anal. 2020

EU-Project: MedalCare Synthetic Reference Data



Digital Traceability



Uncertainty of ML

www.nature.com/scientificdata

scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

MedalCare-XL: 16,900 healthy and pathological synthetic 12 lead ECGs from electrophysiological simulations

Karli Gillette[1,2,8], Matthias A. F. Gsell[1,8], Claudia Nagel[3,8], Jule Bender[3], Benjamin Winkler[4], Steven E. Williams[5,6], Markus Bär[4], Tobias Schäffter[4,5,7], Olaf Dössel[3,8], Gernot Plank[1,2,8] ✉ & Axel Loewe[3,8] ✉

Mechanistic cardiac electrophysiology models allow for personalized simulations of the electrical activity in the heart and the ensuing electrocardiogram (ECG) on the body surface. As such, synthetic signals possess known ground truth labels of the underlying disease and can be employed for validation of machine learning ECG analysis tools in addition to clinical signals. Recently, synthetic ECGs were used to enrich sparse clinical data or even replace them completely during training leading to improved performance on real-world clinical test data. We thus generated a novel synthetic database comprising a total of 16,900 12 lead ECGs based on electrophysiological simulations equally distributed

Medizinische Universität Graz

Fraunhofer HHI

PTB

KIT Karlsruher Institut für Technologie

EU-AI Act for Trustworthy AI

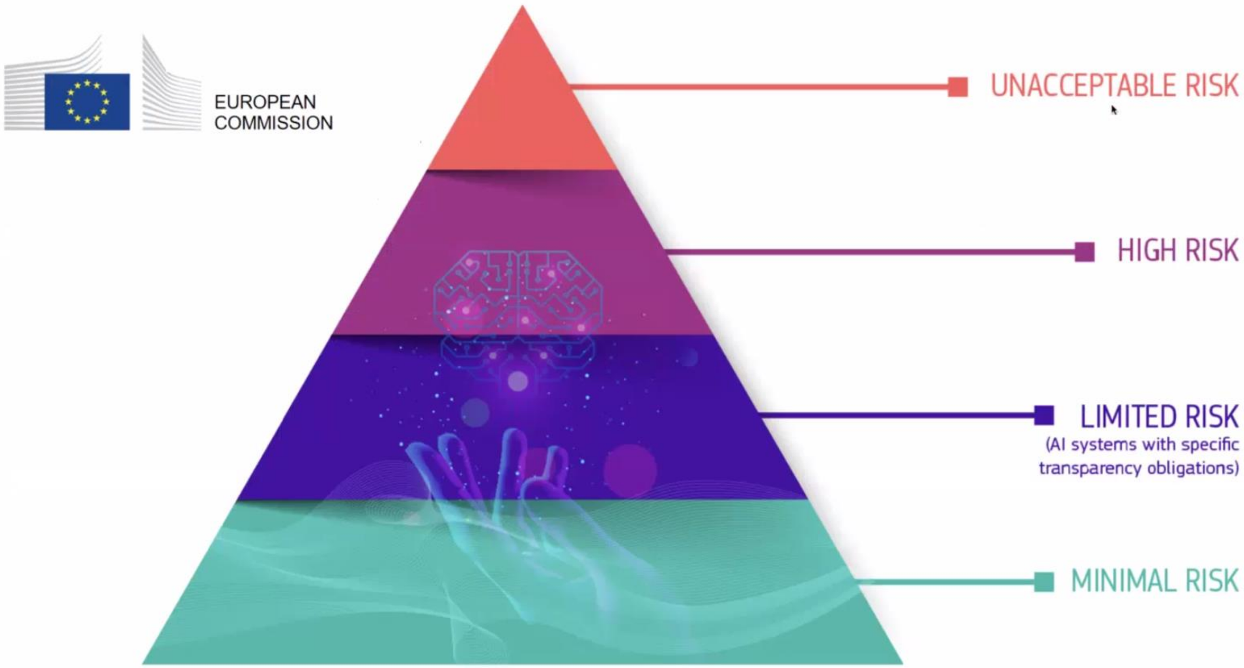
Trust in "algorithms" that are not fully understood ("black box"), especially for high risk applications

The trustworthy AI strongly depends on

high quality trainings data

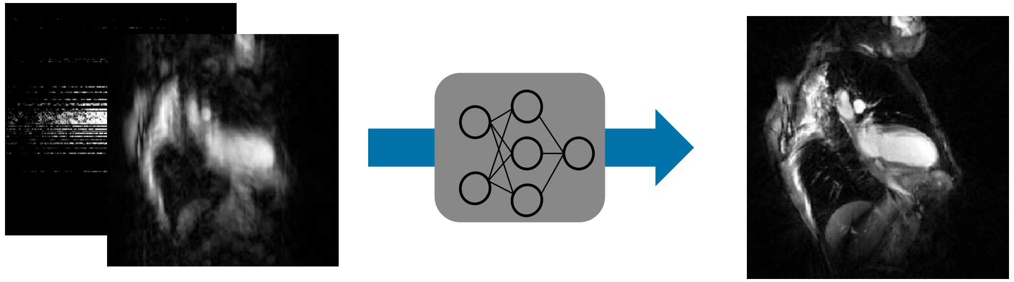
Certification of **AI-Quality** requires:

- robustness,
- accuracy,
- security,
- Explainability.



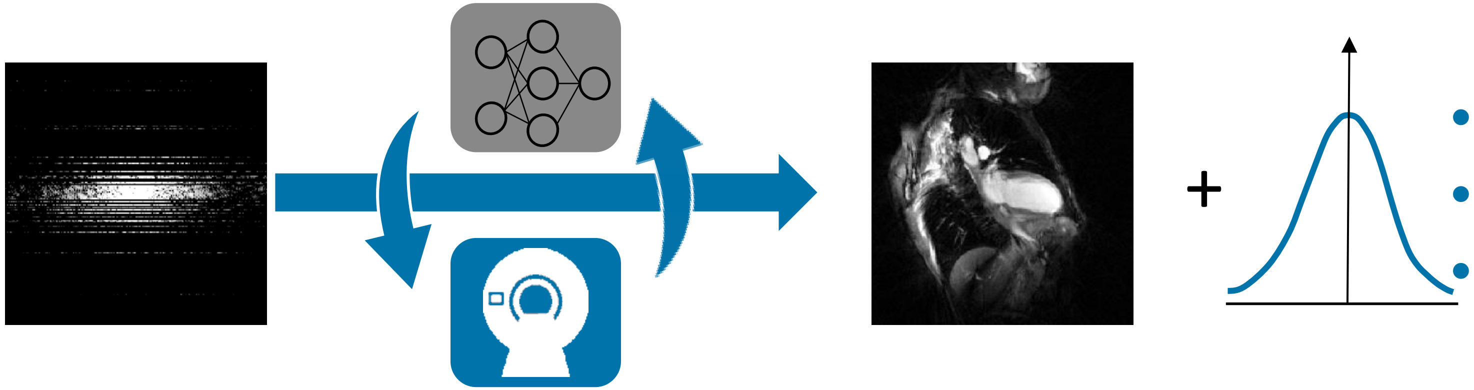
AI for Image Reconstruction

Deep Learning Reconstruction



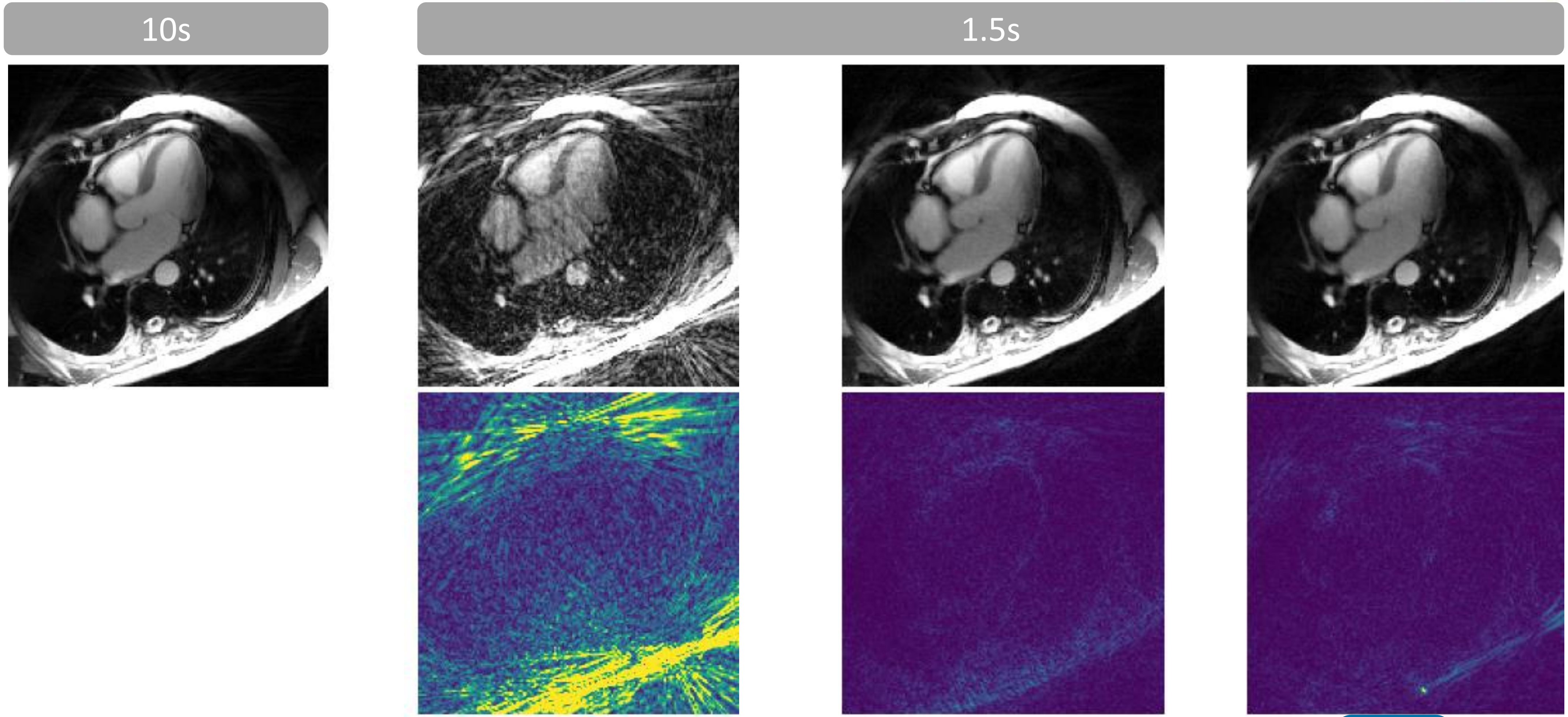
- complex CNN
- High number of parameter
- High amount on trainings data

Physics-informed deep learning



- efficient training
- **robustness**
- uncertainty

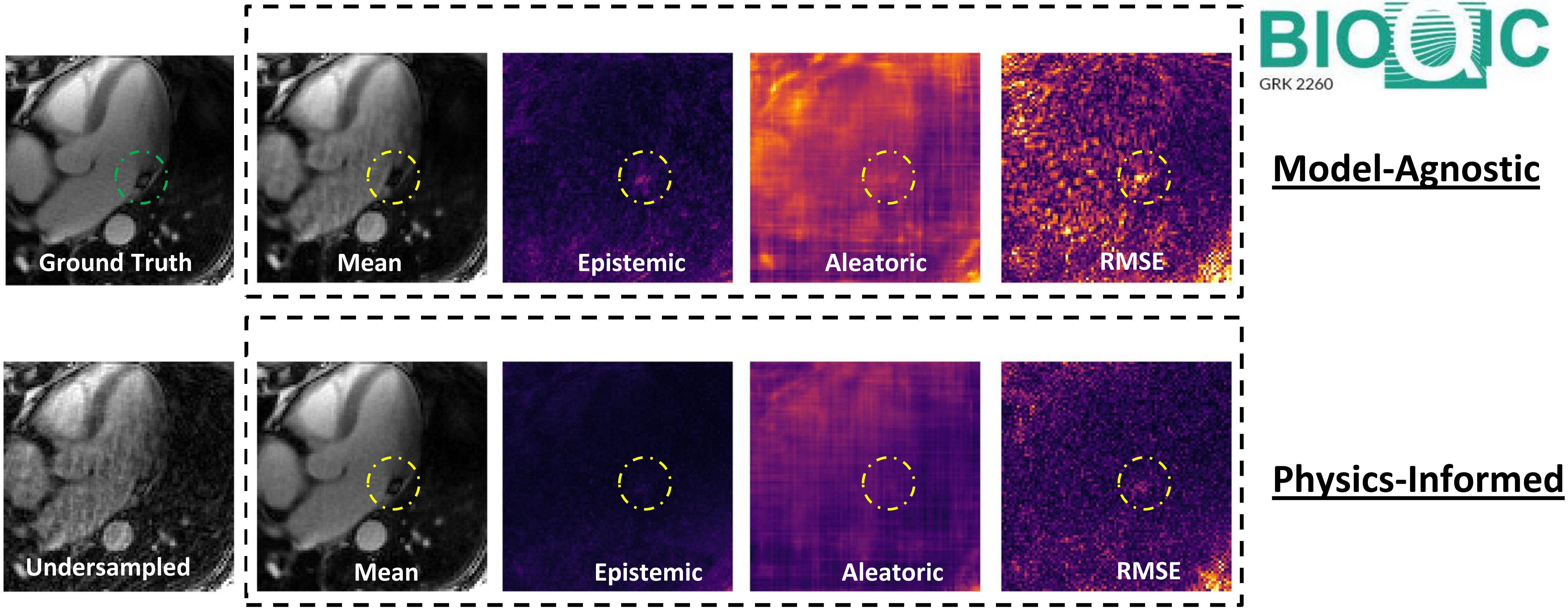
Robustness through physics input



Number of Parameter: >300000 <3000

Kofler et al., Med Phys, 2021; Kofler et al., ISBI, 2022

Reducing „ML-Uncertainty“: „Physics-Informed Learning“

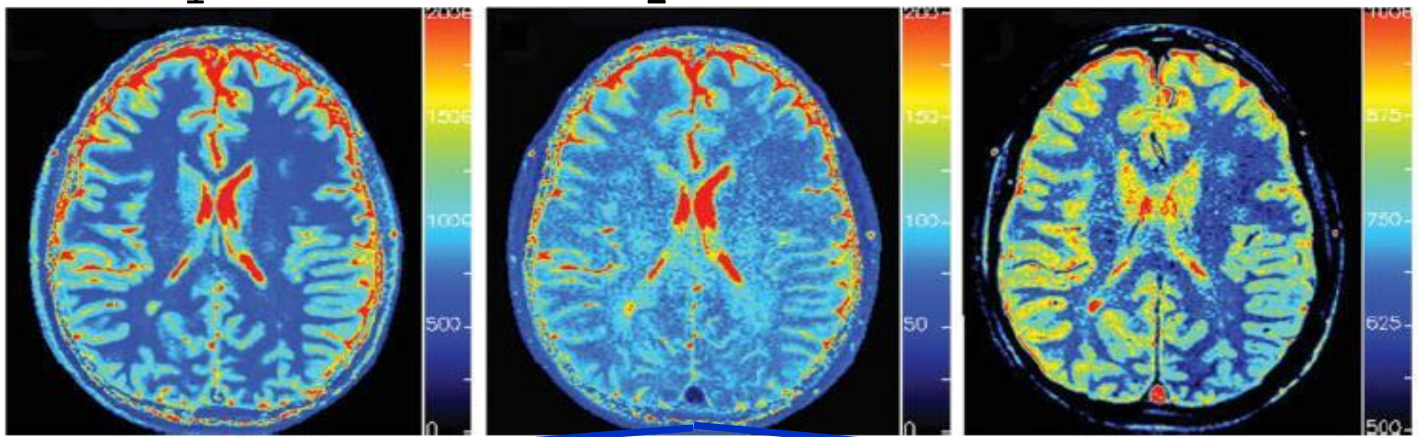


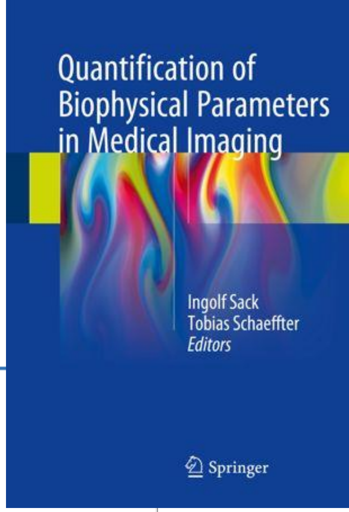
Brahma et al., Med Phys, 2023

Quantitative Imaging

- Parameter-based objective diagnosis
- Comparability
- Detection of diffuse disease
- Contrast agent quantification
- …

T_1 map | T_2 map | PD map





Machine Learning for Quantitative Magnetic Resonance Image Reconstruction

Andreas Kofler, Felix Frederik Zimmermann, and Kostas Papafitsoros

Abstract

In the last years, the design of image reconstruction methods in the field of quantitative Magnetic Resonance Imaging (qMRI) has experienced a paradigm shift. Often, when dealing with (quantitative) MR image reconstruction problems, one is concerned with solving one or a couple of ill-posed inverse problems that require the use of advanced regularization methods. An increasing amount of attention is nowadays put on the development of data-driven methods using Neural Networks (NNs) to learn meaningful prior information without the need to explicitly model handcrafted priors. In addition, the available hardware and computational resources nowadays offer the possibility to learn regularization models in a so-called model-aware fashion, which is a unique key feature that distinguishes these models from regularization methods learned in a more classical, model-agnostic manner. Model-aware methods are not only tailored to the considered data, but also to the class of considered imaging problems and nowadays constitute the state-of-the-art in image reconstruction methods. In the following chapter, we provide the reader with an extensive overview of methods that can be employed for (quantitative) MR image reconstruction, also highlighting their advantages and limitations from both a theoretical and computational point of view.

9.1 Introduction

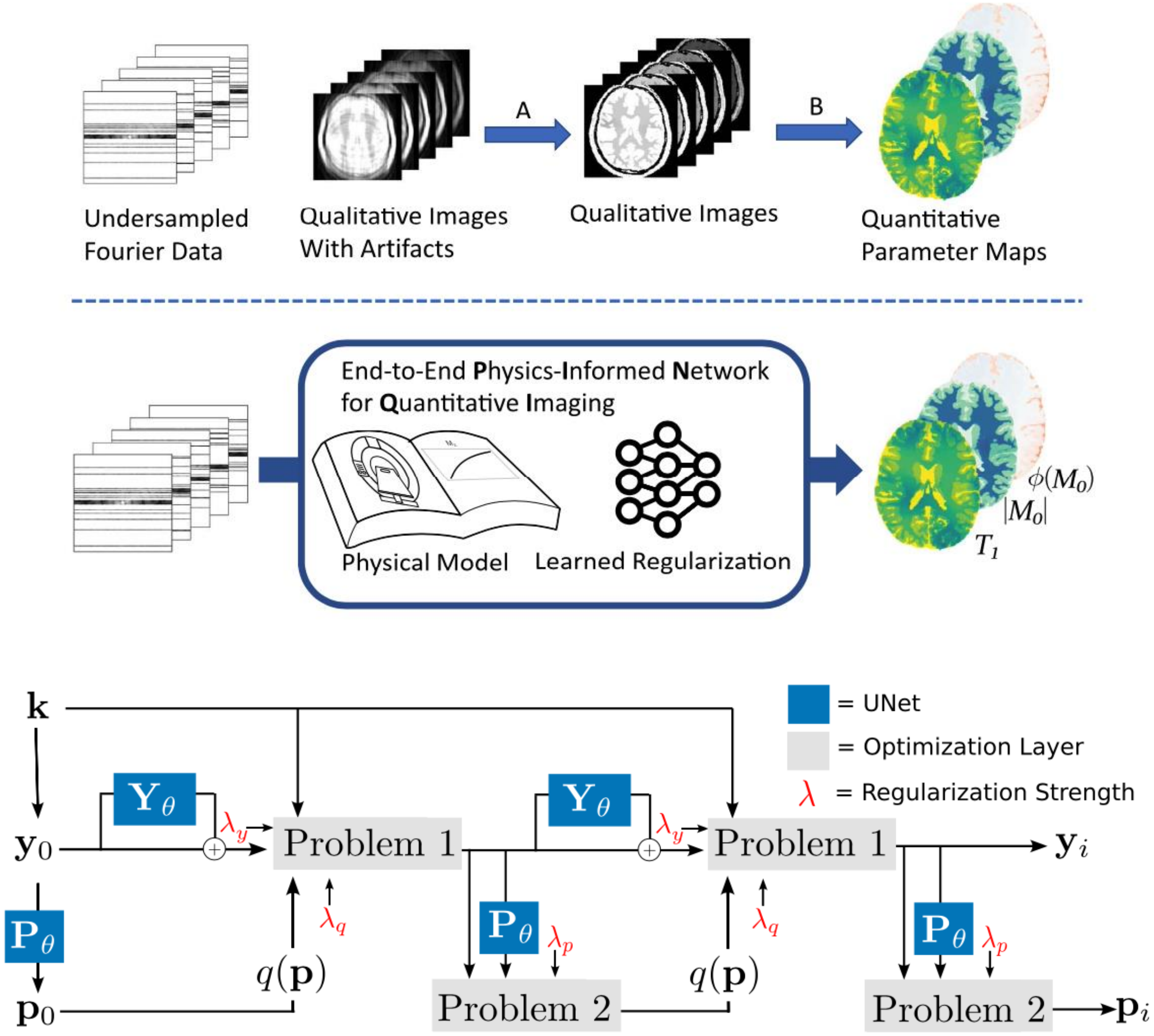
Magnetic Resonance Imaging (MRI) is one of the most important medical imaging tools in nowadays clinical practice. MRI allows for the imaging of organs and joints, parallelly exhibiting excellent soft tissue contrast. Unfortunately, the data acquisition process in MRI is inherently slow. In addition, opposed to other imaging modalities, for example, computed tomography (CT), most MRI scan protocols are not quantitative, i.e., the values in the acquired images do not have a physical and/or biophysical correspondence, which represents a challenge for the comparability of images between different scans, scanners, patients, or institutions. Quantitative MRI (qMRI) can overcome these limitations by the design of data acquisition protocols that allow

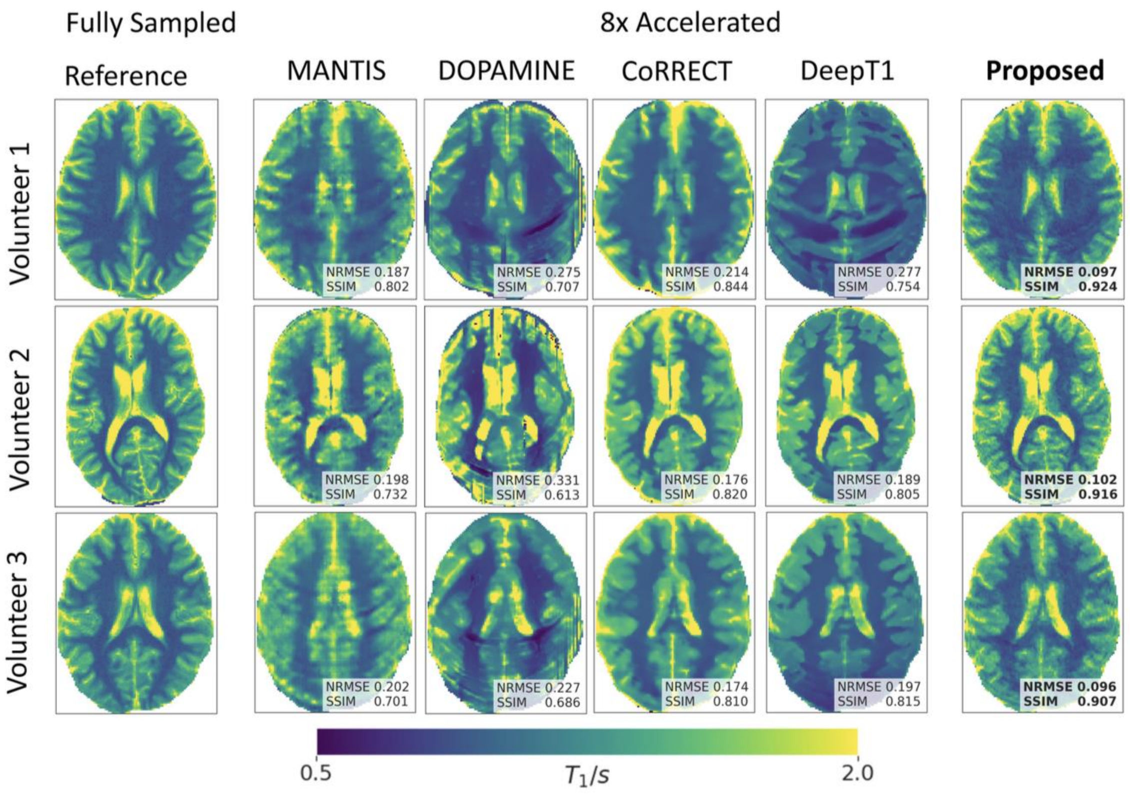
A. Kofler (✉) · F. F. Zimmermann
Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany
e-mail: andreas.kofler@ptb.de; felix.zimmermann@ptb.de

K. Papafitsoros
School of Mathematics, Queen Mary University of London, London, UK
e-mail: k.papafitsoros@qmul.ac.uk

 171
I. Sack, T. Schaeffter (eds.), *Quantification of Biophysical Parameters in Medical Imaging*,
https://doi.org/10.1007/978-3-031-61846-8_9

Physics-Informed Learning - Quantitative MRI

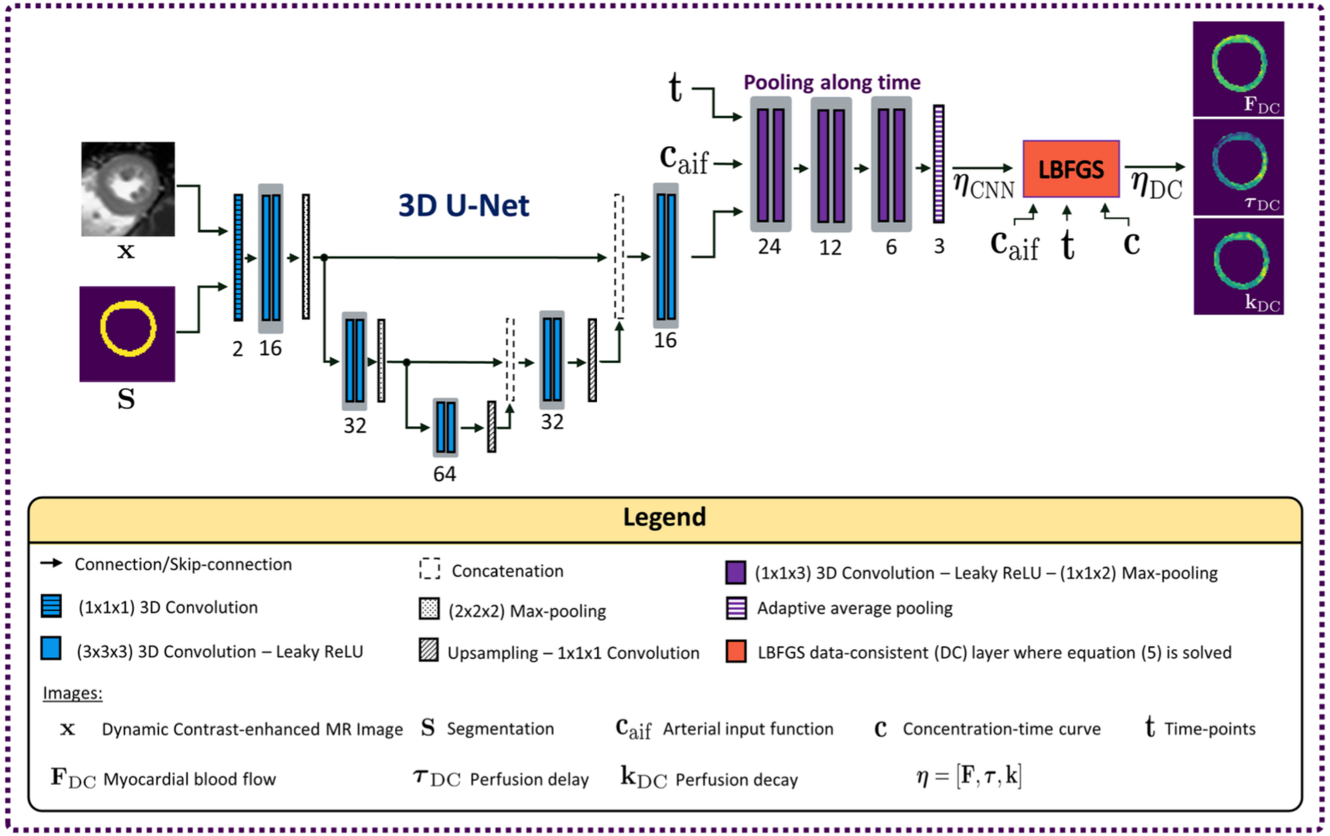


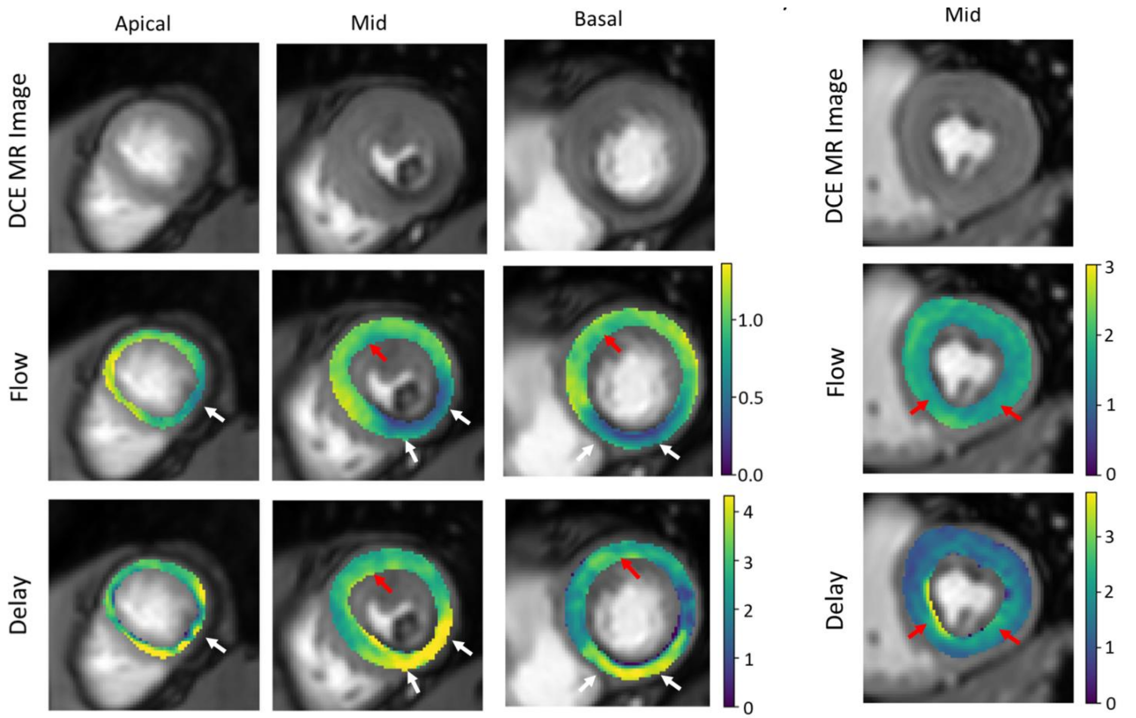


Zimmermann F. et al.
IEEE Transactions on Computational Imaging, 2024,

AI for Quantitative Perfusion

$$\mathbf{C}_t(\eta) := \int_0^t r_{\mathrm{f}}\,(\eta, t')\, c_{\mathrm{aif}}(t-t')\mathrm{d}t'.$$





Brahma et al., IEEE Trans Biomed. accepted

Benchmark-Test

- Comparison Studies
- Ranked-List of Algorithms
- https://grand-challenge.org

nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-022-00559-4

Developing robust benchmarks for driving forward AI innovation in healthcare

Received: 1 June 2022

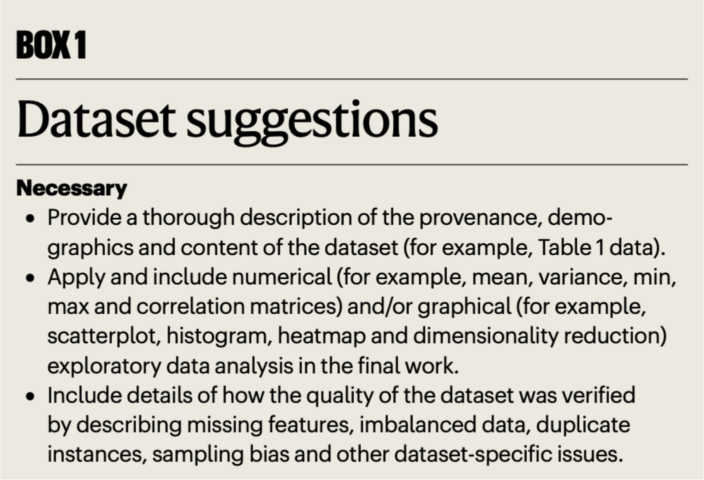
Accepted: 7 October 2022

Published online: 15 November 2022

Check for updates

Diana Mincu & Subhrajit Roy

Machine learning technologies have seen increased application to the healthcare domain. The main drivers are openly available healthcare datasets, and a general interest from the community to use its powers for knowledge discovery and technological advancements in this more conservative field. However, with this additional volume comes a range of questions and concerns – are the obtained results meaningful and conclusions accurate; how do we know we have improved state of the art; is the clinical problem well defined and does the model address it? We reflect on key aspects in the end-to-end pipeline that we believe suffer the most in this space, and suggest some good practices to avoid reproducing these issues.



BOX 1

Dataset suggestions

Necessary

- Provide a thorough description of the provenance, demographics and content of the dataset (for example, Table 1 data).
- Apply and include numerical (for example, mean, variance, min, max and correlation matrices) and/or graphical (for example, scatterplot, histogram, heatmap and dimensionality reduction) exploratory data analysis in the final work.
- Include details of how the quality of the dataset was verified by describing missing features, imbalanced data, duplicate instances, sampling bias and other dataset-specific issues.

nature methods

Perspective

https://doi.org/10.1038/s41592-023-02151-z

Metrics reloaded: recommendations for image analysis validation

Received: 9 February 2023

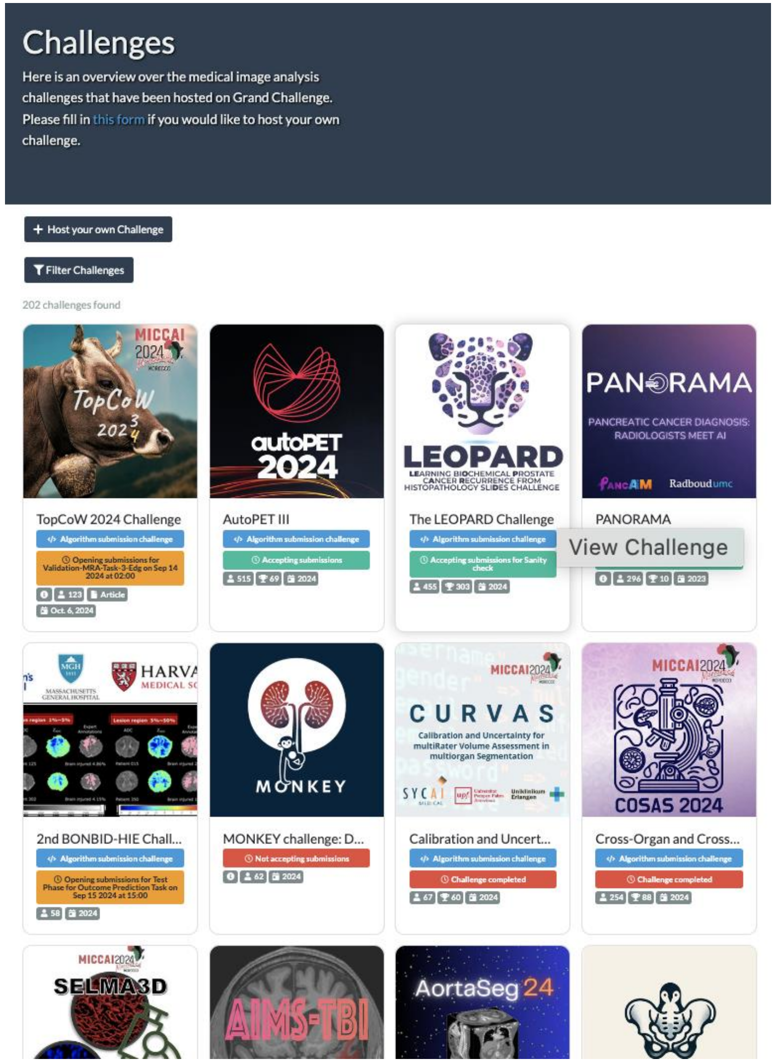
Accepted: 12 December 2023

Published online: 12 February 2024

Check for updates

A list of authors and their affiliations appears at the end of the paper

Increasing evidence shows that flaws in machine learning (ML) algorithm validation are an underestimated global problem. In biomedical image analysis, chosen performance metrics often do not reflect the domain interest, and thus fail to adequately measure scientific progress and hinder translation of ML techniques into practice. To overcome this, we created Metrics Reloaded, a comprehensive framework guiding researchers in the problem-aware selection of metrics. Developed by a large international consortium in a multistage Delphi process, it is based on the novel concept of a problem fingerprint—a structured representation of the given problem that captures all aspects that are relevant for metric selection, from the domain interest to the properties of the target structure(s), dataset and algorithm output. On the basis of the problem fingerprint, users are guided through the process of choosing and applying appropriate validation metrics while being made aware of potential pitfalls. Metrics Reloaded



Benchmarktests and Metrics

IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 25, NO. 5, MAY 2021 1519

Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL

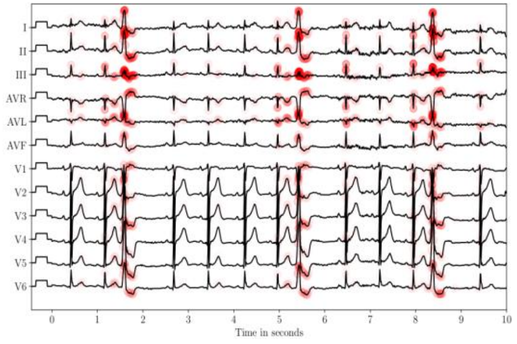
Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek, *Member, IEEE*

Metrics

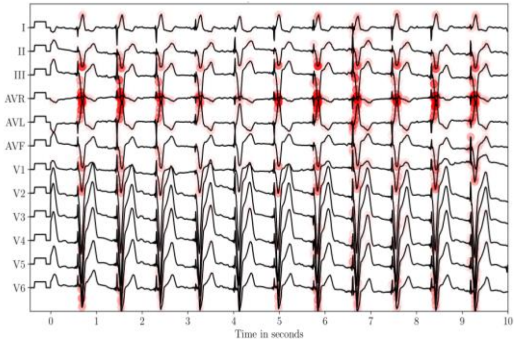
Method	all		diag.		sub-diag.		super-diag.		form		rhythm	
	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax	AUC	Fmax
lstm_bidir	.902(11)	.749(10)	.922(12)	.729(14)	.928(09)	.756(12)	.929(06)	.817(12)	.845(17)	.605(22)	.947(10)	.908(09)
lstm	.893(12)	.745(08)	.905(12)	.724(13)	.912(16)	.753(10)	.928(06)	.819(11)	.813(17)	.596(25)	.948(09)	.907(10)
fcn_wang	.911(10)	.754(08)	.922(10)	.731(14)	.920(14)	.752(11)	.927(07)	.815(12)	.875(18)	.625(23)	.928(10)	.899(11)
resnet1d_wang	.912(11)	.764(08)	.932(08)	.741(15)	**.932(09)**	.760(12)	**.932(06)**	**.825(12)**	.877(14)	.620(23)	.945(09)	.908(09)
xresnet1d101	**.920(08)**	**.765(08)**	**.935(08)**	**.743(13)**	.927(09)	.759(10)	.931(06)	.819(11)	**.885(13)**	**.629(20)**	**.957(20)**	.915(08)
Wavelet+NN	.811(14)	.678(10)	.823(19)	.627(15)	.845(17)	.654(14)	.870(10)	.731(13)	.798(21)	.526(22)	.857(52)	.866(13)
inception1d	.919(08)	**.765(07)**	.929(13)	.737(12)	**.932(08)**	**.763(10)**	.930(06)	.819(11)	**.885(14)**	.627(20)	**.957(14)**	**.917(09)**
ensemble	**.923(09)**	**.767(08)**	**.935(07)**	.740(12)	.928(11)	**.764(11)**	**.937(06)**	**.827(12)**	**.891(12)**	**.638(23)**	**.970(08)**	.916(08)
naive	*.500(00)*	*.557(11)*	*.500(00)*	*.440(18)*	*.500(00)*	*.440(18)*	*.500(00)*	*.448(09)*	*.500(00)*	*.365(19)*	*.500(00)*	*.797(13)*

Strodthoff et al. IEEE Journal of Biomedical and Health Informatics 2020.

Explainability



(a) PVC

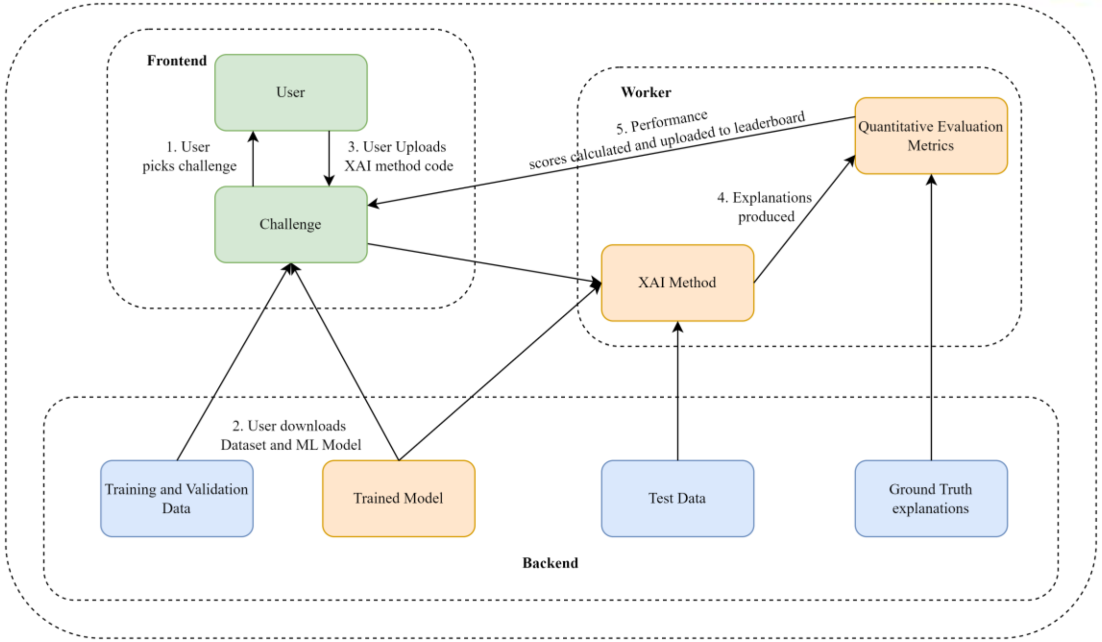


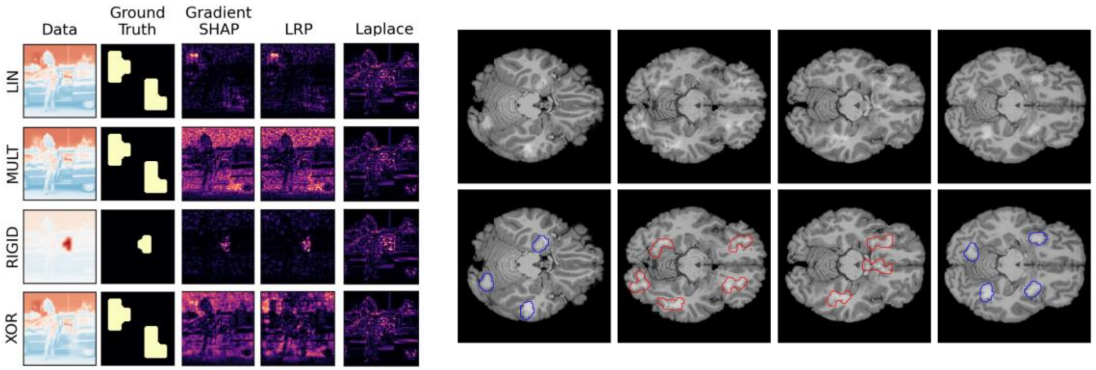
(b) PACE

EXACT: Digital Explainability Testplattform

Aim:

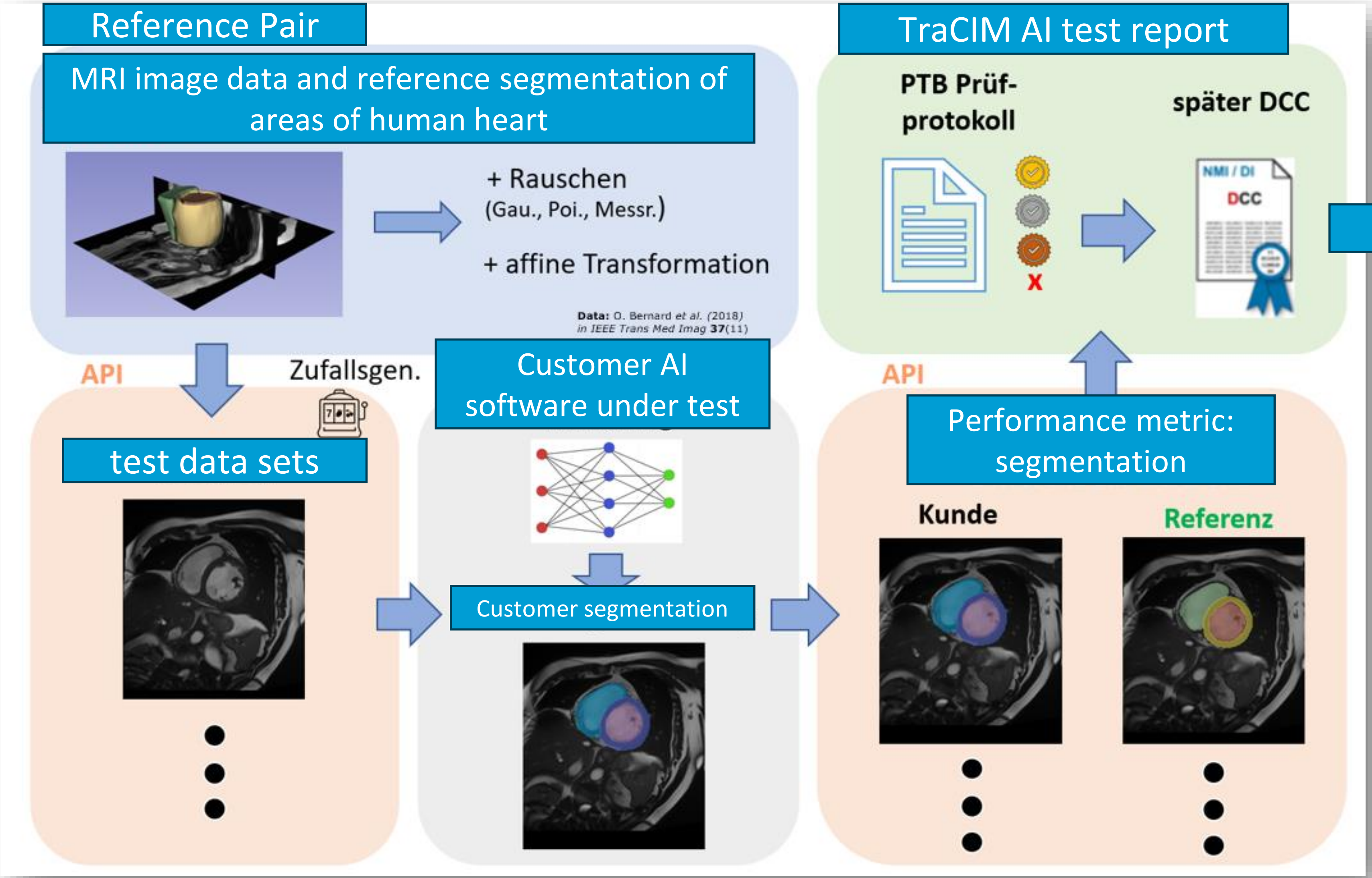
- Proof-of-Concept-Platform in "Challenge-Style" for Benchmarking of „explainable“ AI Methods
- Usage of Reference-Datasets with ground truth
- Application of Performance-Metrics
- Methods for Quality Assurance





Reference XAI Datasets

Digital AI Testplatform



Conclusion

- Data is the base of AI
- Quality of training data is instrumental and can be described by 15 dimensions in 5 clusters (METRIC)
- AI-quality relies on robustness, uncertainty, explainability
- Physics-informed learning can improve robustness and reduce uncertainty
- Benchmarktests require
  - reference datasets (synthetic and real)
  - metrics (use-case specific)

Accuracy

Objectivity

Passion

Physikalisch-Technische Bundesanstalt
The National Metrology Institute

PTB