Metrology for AI in medicine

Background, strategy and implementation recommendations

Metrology is a recognised anchor of trust and an essential part of the quality infrastructure. It encompasses the characterisation of measurement technology and measurement methods, the evaluation of the quality of measurement data and the development of new measurement procedures. This also applies in particular to methods, processes and measuring devices in the medical field. As in all fields, medicine is also increasingly affected by the advancing digitalisation and consequently poses new challenges for PTB. In principle, the legal mandate of PTB within the framework of the German Units and Time Act (EinhZG § 6 (3)), the German Measurement and Calibration Act (MessEG § 45) and the German Medical Devices Act (MPG § 32 (2)) is formulated in a very technology-open manner. PTB itself is therefore also continuously required to evaluate and question its own role within the meaning of the legal mandate in view of technological developments. At the same time, with new technological developments it can always be assumed that there is considerable expectation for PTB to fulfil its legal mandate in a competent manner.

A long-standing challenge is the use of software as a (partly independent) product in the medical field. The EU Medical Device Regulation (MDR) already provides for the evaluation of software-only products in the healthcare sector (Medical Device Software - MDSW). If this software is an essential component of a medical device, the hardware and software must be evaluated together. In any case, approval requires clinical studies that demonstrate medical added value.

Artificial intelligence (AI) plays an increasingly catalytic role in the course of digitalisation and accelerates the development of digital products and services enormously. As the use of AI grows, so does the need for clear rules and consideration in the quality infrastructure. This explicitly includes the healthcare sector, where the complexity of the interrelationships, the technical progress in measurement technology, the high benefit for patients and the tremendous economic potential will bring about a rapid increase in AI applications.

In line with the Federal Government's AI strategy, *Artificial Intelligence* is used here to refer to an algorithm system for solving concrete application problems based on methods from mathematics and computer science, with the developed systems being capable of self-optimisation.

The Federal Government's Enquete Commission sees AI as the next stage of digitalisation driven by technological progress. An essential element here is the way in which these algorithms are developed. A classical algorithm usually implements a previously described process in software. The process is based on mathematical, statistical or other assumptions, theories and rules. In contrast, the algorithm of an AI method is trained with the help of data. These algorithms are characterised by a high complexity and a very high-dimensional parameter space. Another feature is the very high adaptability of AI methods. However, this can result in the unintentional inclusion of unrecognised features of the training data in the algorithm. This makes it almost impossible to check the algorithm on the basis of the source code alone, in contrast to other software.

Especially the application of AI methods in the medical field has so far lacked a generally recognised anchor of trust in the quality infrastructure, even though trust in AI methods is indispensable for their acceptance. In general, both the Federal Government and the EU Commission are pursuing the principle of human-centred AI. This means that AI should benefit individuals and society whilst enhancing human capabilities. This model is also referred to as "European-style AI" in distinction to commercially oriented AI development, for example in the USA or China.

Many research institutions, standardisation bodies and companies are developing methods, guidelines and standards on AI. This document addresses the question of what the role of metrology, and PTB in particular, should be in this context.

(1) PTB is systematically expanding its competences in the field of AI and establishing cooperations with external AI experts in order to also be able to fulfil its legal mandate in the future.

(2) The current legal mandate can be seen to indicate that PTB must engage intensively with the topic of AI. In the medium term, efforts will be made to enshrine PTB's mandate and responsibility in the field of AI more firmly in law and to integrate PTB from the outset in the revision of a legal framework for AI.

(3) Work in the field of AI is initially centred on concrete applications. At the same time, fundamental theoretical work is being pursued in order to establish a general AI competence at PTB beyond specific applications in the long term.

1. Background and differentiation from other actors

There is currently virtually no area in which the topic of AI is not being discussed. Numerous publications on new methods, models and application examples are being published with increasing frequency. Various Fraunhofer Institutes, the DFKI and the DLR are substantially expanding their activities in the field of AI research. A total of 100 AI professorships are set to be established in the coming years. In standardisation, new working groups are being set up not only at ISO and IEC. DIN is also very active with the "Standardisation Roadmap AI" [17] and the already partially published DIN SPEC 92001. In some cases, committees with direct relevance to medicine and health have also emerged, such as the ITU/WHO FG-AI4H on AI methods in the healthcare sector. PTB must find its role and place in this environment.

Standardisation and regulation of AI

The basic outlines of the requirements and terminologies for the evaluation of AI methods have already been developed. For example, DIN SPEC 92001-1 [1] defines three essential requirements for the quality of AI:

- *Functionality and performance* as an expression of the AI's ability to fulfil its intended task under stated conditions
- *Robustness* as the ability of AI to cope with erroneous, noisy, unknown and adversarial input data
- *Comprehensibility* represents the degree to which the causes of an AI module's output can be understood

DIN SPEC 92001-2 [7] specifies the term robustness in more detail and distinguishes between *adversarial robustness* (AR) and *corruption robustness* (CR). The former refers to robustness to adversarial changes in the input data, the latter to robustness to noise or changes in the statistical properties of the input data. For the development of robust AI methods, [7] recommends a risk analysis-based approach. Methods for the targeted testing of AI methods are also mentioned (Fast Gradient Sign Method; Projected Gradient Descent). In general, scenario-based testing is recommended, which takes into account the later intended use and its characteristics. It is considered important [7] that the risk assessment of an AI method must actually be carried out continuously.

In a recent discussion paper [3], the US Food and Drug Administration (FDA) also assumes the necessity of a "Total Product Lifecycle Regulatory Approach" (TPLC) for AI applications. In the course of the certification of an AI-based medical device, the company's quality management in particular is to be assessed with regard to

- quality assurance in software development
- testing and performance monitoring of the products

The FDA stipulates the following basic principles in this context

- Establishment of recognised good machine learning (ML) practices
- Consideration of the product life cycle in the approval of AI-based medical devices
- Expectation that manufacturers will implement a risk-based approach to monitoring their AI-based medical devices for the entire product life cycle
- Transparent statements for customers and testers on real-world performance and behaviour of AI-based medical devices by manufacturers.

For the FDA, this also includes documenting the planned area of use (Software as a Medical Device (SaMD) Pre-Specification - SPS) as well as the (further) development in an "Algorithm Change Protocol" (ACP): data management, re-training, performance evaluation, update procedures. SPS and ACP are then essential aspects in the approval of new products.

An important part of the quality assurance and management system, according to the FDA [3], is the selection of training and test data and selection of user data for re-training. When it comes to data management, the FDA therefore requires

- plans for data collection
- quality assurance systems for the data
- determination of a reference standard
- auditing and sequestration of test and training data

by the manufacturers. The catalogue of questions by the German "Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland" (IG-NB) for the approval of AI for medical devices also addresses many issues related to the selection and assessment of the data used [19]. At the same time, there is a lack of corresponding norms and standards as a basis for the assessment of AI and the underlying data in [19], particularly with regard to the important questions.

AI certification

The Fraunhofer IAIS white paper [4] discusses a certification for AI applications that can be implemented operationally by accredited auditors. Accordingly, a certificate for AI should

- attest to a certain quality standard,
- help to make AI applications verifiably legally compliant and
- make AI applications comparable.

The IAIS notes in this context that domain knowledge and mathematical-statistical expertise are necessary for the assessment of reliability [4].

EUROLAB distinguishes clearly between indirect conformity assessment (ICA) and direct conformity assessment (DCA) of AI methods [5]. With ICA, the AI method is used to support decision-making. Here, the example of an AI method for the evaluation of an X-ray measurement is used: The AI method uses the measurement data to generate a qualitative statement, e.g. about the patient's state of health. In this case, an accreditation would not require an assessment of the AI method itself, rather the competence of the staff of the body seeking accreditation would have to be determined. This would correspond to the approach already practised today for any other non-linear numerical methods. However, if the AI method presents the result as part of the measurement (e.g. as an overlay) and thereby gives the appearance of a "real result", the AI method itself should be understood as an "authority" and should be included in the accreditation. EUROLAB currently recommends using DCA only in very non-critical areas (e.g. assessing music quality) until the methods are more refined. The EUROLAB position paper [5] also calls for some form of calibration for AI methods, as is required for common measuring equipment. This should make the reliability of the AI method assessable by recording the ability of the method to reproduce the result of a "standard".

The current white paper "Certification of AI systems" [18] by the *Plattform Lernende Systeme* states:

> *"Before a successful certification of AI systems can be established, there are thus still unresolved issues to be clarified. These concern the subject matter of certification, the test criteria, the timing and necessity of certification, the level of detail of certification, and how to approach learning systems."*

Furthermore, [18] also refers to the AI High Level Expert Group of the EU Commission (COM), which recommends the following criteria in the regulation of AI

> *"[Priority of] human agency and oversight, technical robustness and safety, transparency and accountability, diversity, non-discrimination and fairness, societal and environmental well-being and respect [for] the established principles of privacy, and data governance and data protection."*

Especially for high-risk AI applications (such as in healthcare), it is recommended that the conformity assessment should consider

- whether the training data are adequate for the intended use;
- that the results do not result in discrimination in use;
- whether data protection and privacy are respected;
- that relevant records on datasets, training methods and programming methods are available.

As such, these recommendations essentially follow those of the FDA [3], which (like the Federal Government in its statement on the COM white paper) also considers repeated testing of these criteria to be necessary for learning, in other words changing, AI systems.

Conclusions for the competence of PTB

The current publication of the *Plattform Lernende Systeme* states very clearly:

> *"The conformity assessment should be based on existing national structures and processes. Where no such authorities exist, there should be a duty to establish such an authority or to establish responsibilities in existing authorities." [18]*

PTB, as an essential part of the quality infrastructure and with a high degree of neutrality, is predestined to take on this role. In concrete terms, today PTB is already called upon to fulfil its legal tasks even for measuring devices with AI components by continuously building up the corresponding competences. In its statement on the COM white paper on AI, the Federal Government makes the following recommendations:

- Binding legal requirements should be considered for training data from AI systems. For this, requirements for test and evaluation data should also be consistently considered.

- Legal requirements for quality parameters and requirements for training, test and evaluation data are also needed so that appropriate AI systems are developed with quantitatively sufficient and high-quality datasets.
- Robustness and accuracy requirements should cover recognisable and realistic scenarios.
- If suitable processes are in place to check the AI results for representativeness and balance, access to the training/testing data can also be dispensed with.
- Central to this is ensuring the confidentiality, integrity and availability of the AI system throughout its entire life cycle.
- High-risk AI systems should be required to undergo an objective conformity assessment procedure. Here, it is necessary to carry out repeated assessments of evolving adaptive AI systems.
- The accuracy and relevance of the data are the decisive factors. It is furthermore necessary to ensure the provision of reference data, benchmark tests and the verification of algorithms using quality-assured, trustworthy reference data. In principle, the statement that data quality must be guaranteed throughout the entire period of use is supported. It should be noted, however, that this can only be done by the operator, who in turn is not an economic operator in the sense of product safety law.
- In the event that a new product has been created as a result of a software change, this product must fully comply with the state of the art, as there is a new placing on the market. This must be taken into account when considering a software change.

PTB's research activities in the field of AI must therefore ensure that these requirements and expectations can be met. Consequently, the focus should be less on the development of new AI methods and more on the development of evaluation methods. An important unique selling point of PTB is its domain knowledge and its neutrality.

Research cooperation

Even in the long term, PTB will not be able to compete with the scope of work and the capacities of other large research associations - nor will it have to. In order to use the available resources as effectively as possible, PTB will therefore enter into targeted cooperations with research partners.

A good overview of the current research landscape in the field of AI is provided by the "AI Map" of the *Plattform Lernende Systeme*[1]. Some prominent examples are

- German Research Center for Artificial Intelligence (DFKI)
- Fraunhofer Society
- German Aerospace Center (DLR)
- Helmholtz Association (especially Helmholtz AI Cooperation Unit, see below)
- Max Planck Institute for Intelligent Systems
- Mevis Medical Solutions (Bremen)
- University hospitals (e.g. Charité) and medical faculties, e.g. of the University of Duisburg-Essen and University Hospital Essen (Institute for Artificial Intelligence in Medicine)
- Facilities in Cyber Valley in Tübingen as Europe's largest research consortium in the field of AI with partners from science and industry
- Robert Koch Institute (RKI) (establishment of a centre for the validation of AI algorithms in health research)

[1] https://www.plattform-lernende-systeme.de/ki-landkarte.html

With regard to organisational and structural decisions, PTB is also partly guided by large research institutions. For example, the Helmholtz Association has founded the Helmholtz Artificial Intelligence Cooperation Unit (Helmholtz AI). This is one of five platforms initiated by the Helmholtz Incubator for Information and Data Science.

"Its main goal is to become an engine for applied artificial intelligence (AI) by developing and disseminating AI methods across Helmholtz centres, effectively combining AI-based analytics with Helmholtz's unique research questions and datasets." (Source: https://helmholtz.ai)

This platform brings together scientists from all centres and in this way promotes transdisciplinary research.

2. Research aspects of metrology for AI

With its research activities in the field of AI, PTB puts itself in a position to be able to fulfil its legal mandate also in the future. This includes the specifications for quality features of AI defined by norms and standards as well as requirements from regulations and laws for the certification and conformity assessment of quality assurance methods for training and test data.

AI assessment

PTB already carries out basic investigations and application studies to determine assessment procedures for AI methods. The focus is on the development of quantitative indicators for the assessment of *comprehensibility, uncertainty, generalisability* and *robustness*.

The evaluation of the functionality and performance of AI as a measure of quality, as called for in [1], requires the quantitative evaluation of the uncertainty of the AI's predictions. It is essential that the "measure" with which the uncertainty is measured is standardised, as only then is it possible to compare the uncertainties of the predictions of different AI methods, as demanded in [4]. Methods for the quantitative determination of measurement uncertainties play a central role in metrology, where there is now a globally recognised standard in the form of the GUM [7]. Such standardisation has been lacking in the field of AI, where there are a variety of different approaches to quantifying uncertainty [8-10, 16]. PTB is currently investigating the suitability of current approaches for quantifying the uncertainty of AI methods with the aim of developing a recommendation for possible standardisation. The investigations include basic studies as well as application examples [11]. From the point of view of metrology, it would be desirable if a standardisation of the uncertainty were in line with the principles of uncertainty evaluation in metrology, so that in applications where AI methods and classical methods operate in a similar way, the same uncertainties could also be assigned.

To establish trust in AI methods, it is important to understand their behaviour and ensure that they do not simply react to specific aspects of the training data [12], but use the relevant information in the data. Much like uncertainty, there are now also a variety of approaches to comprehensibility, see e.g. [13, 14] and the references therein. One of PTB's goals in this area is ultimately also to establish a standardised measure for the quantification of comprehensibility. Close cooperation between PTB and HHI is planned in this area, starting in 2021. The cooperation is part of a project conducted at PTB to investigate AI methods in medical imaging from a metrology perspective.

The robustness and generalisability of AI methods with respect to input data that deviates from the data used to train the method plays a major role, especially in medical technology. Significant here are, for example, "out-of-distribution" errors, which arise because certain features are not represented in the training data. A great deal of importance is also attached to so-called "adversarial attacks", in which "benign" input data are deliberately altered slightly in a way that causes an AI method to fail. In order to make the evaluation of robustness with regard to these influencing factors comparable in quantitative terms, several evaluation criteria have been proposed. PTB is investigating these criteria and has developed alternatives based on statistical approaches which have shown excellent properties in previous investigations [15, 16].

Reference data and data quality assessment

All sources for evaluation, certification and conformity assessment of AI applications or products with AI components state the need for reference data as well as generally accepted criteria for data quality and data handling. In order to perform its tasks, PTB must therefore build up competences on these issues. The necessity of domain knowledge, as also mentioned in [4], is important when deciding on suitable research projects. For instance, the representativeness of reference data "in itself" is not possible, but always only in context against the background of a basic population. Statistical criteria (e.g. test for equality of distributions) could be used instead. Here, instructions for the construction of the (synthetic) reference data could also be added as a task for PTB. Metrology already works on the assessment of data, but so far rather at the bottom-up level (GUM-like), based on the understanding of the underlying physics, than top-down via the properties of the data itself. PTB also already provides physical/chemical reference data in some areas. This could in future be further expanded with the aim of developing reference data specifically for the evaluation of AI methods. In this context, the development of methods for synthetic data should also be considered.

There are now first examples of automatic annotation of training data by combining different modalities. For instance, in [18], an ML method was trained in a first step to combine tomography images of the retina and co-registered fundus images to predict retinal thickness. As a result, the trained ML method was used to automatically annotate a database of 120 000 datasets. These in turn are used as a training dataset for ML methods to detect eye damage caused by diabetes with impending blindness. The approval of such an ML method must then no longer only evaluate the pure raw data, but also the entire workflow for the use of this data. Accordingly, PTB would also have to develop competences in the field of data handling in order to be able to map the requirements from [3] and [18], for example.